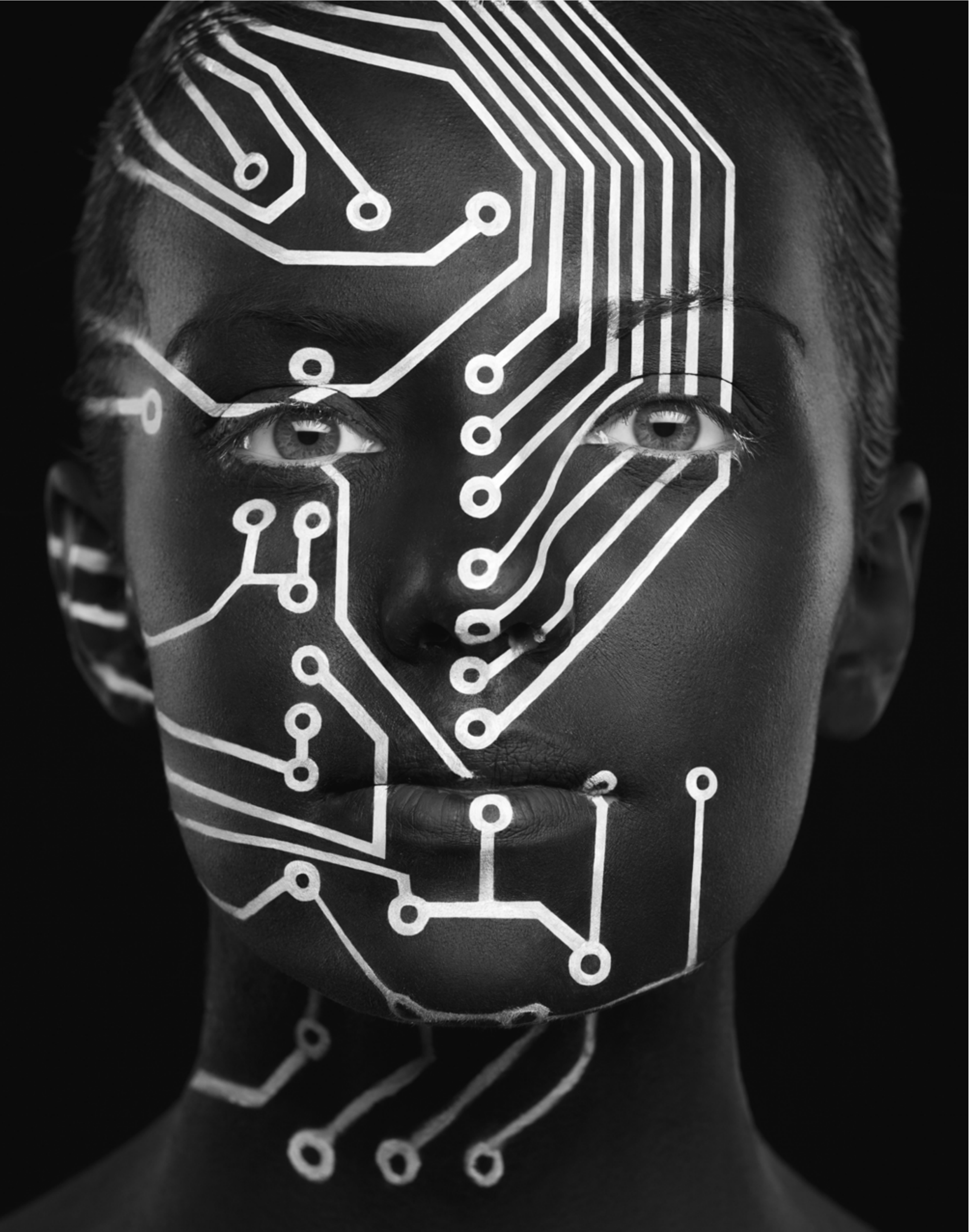

Aspects of Object Recognition

Tim Christian Kietzmann

Aspects of Object Recognition: Sampling, Invariance, and Plasticity

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
im Fachbereich Humanwissenschaften der
Universität Osnabrück

vorgelegt von

Tim Christian Kietzmann

April, 2014

1st Supervisor:	**Professor Peter König** University of Osnabrück, Osnabrück, Germany
2nd Supervisor:	**Professor Frank Tong** Vanderbilt University, Nashville, TN, USA

Cover Photo by Alexander Khokhlov (http://www.alexanderkhokhlov.com/), reprinted with permission.

Publication List[1]

Journal Articles

Kietzmann, T.C., Gert, A.L., & König, P. (in preparation). Representational dynamics of facial viewpoint encoding: head orientation, viewpoint symmetry, and eye contact

Kietzmann, T.C., Ehinger, B.V., Porada, D., Engel, A.K., & König, P. (submitted). Extensive Training Leads to Temporal and Spatial Shifts of Cortical Activity Underlying Visual Category Selectivity

Kietzmann, T.C., & König, P. (2015). Effects of Contextual Information and Stimulus Ambiguity on Overt Visual Sampling Behavior, *Vision Research*, 110, p.76-86

Kietzmann, T.C., Swisher, J., König, P., & Tong, F. (2012). Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways, *Journal of Neuroscience*, 32 (34), p. 11763-11772

Kietzmann, T.C., Geuter, S., & König, P. (2011). Overt Visual Attention as a Causal Factor of Perceptual Awareness, *Plos One*, 6 (7), p. 1-9

Wilming, N., Betz, T., Kietzmann, T.C., & König, P. (2011). Measures and limits of models of fixation selection, *Plos One* 6 (9), p. 1-19

Kietzmann, T.C., & König, P. (2010). Perceptual learning of parametric face categories leads to the integration of high-level class-based information but not to high-level pop-out, *Journal of Vision*, 10 (13), p.1-14

Conference Contributions

Tim C Kietzmann, Sam Ling, Sonia Poltoratski, Peter König, Randolph Blake, & Frank Tong (2014). The Occipital Face Area is Causally Involved in Viewpoint Symmetry Judgments of Faces. *Vision Science Society Meeting 2014*

Kietzmann, T.C., Ehinger, B., Porada, D., Engel, A.K., & König. P. (2013). Perceptual Learning Leads to Category Selectivity 100ms after Stimulus Onset. *European Conference on Visual Perception 2013*

Kietzmann, T.C., Wahn, B., König, P., & Tong, F. (2013). Face selective areas in the human ventral stream exhibit a preference for 3/4 views in the fovea and periphery. *European Conference on Visual Perception 2013*

[1] This list contains only publications, which are part of this dissertation. Please refer to my CV or to http://www.timkietzmann.de for a complete and up-to-date list.

Kietzmann, T.C., Ehinger, B., Porada, D., Engel, A.K., & König. P. (2013). From stimulus onset to category selectivity in 100ms: category-selective visually evoked responses as a result of extensive category learning. *Vision Science Society Meeting 2013*

Kietzmann, T.C., Swisher, J., König, P., & Tong, F. (2012). Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Streams of the Human Visual System. *Vision Science Society Meeting 2012*

Kietzmann, T.C., & König, P. (2010). Parametric Faces in a Pop-Out Paradigm - When Class Information Becomes a Feature. *KogWis 2010*

Book Chapters

König, P., Kühnberger K.U., & Kietzmann, T.C. (2013). A Unifying Approach to High- and Low-Level Cognition. In *Models, Simulations, and the Reduction of Complexity*. (pp. 117-141). De Gruyter

Abstract

We humans are visual creatures, constantly extracting information from the world around us. The source of our ability to understand the visual world is an intricate arrangement of multiple areas in our brains: the visual system. It enables us to recognize our friends and family in diverse conditions, to focus our attention on important aspects of a scene and performs invariant object categorization on multiple levels of abstraction. Vision has been in the focus of scientific interest for many decades and yet our knowledge of the cortical mechanisms involved is only limited. I here describe a series of experiments, in which we investigated how the visual system robustly and efficiently extracts meaning from the environment. In particular, I will focus on thee aspects of object recognition: sampling the environment, visual invariance, and categorization and plasticity.

Starting with the selection of visual information, three eyetracking experiments are described in which we investigate the interplay of overt visual attention and object recognition. We show that overt visual attention and object recognition exert a bi-directional influence on each other. Whereas initial patterns of overt visual attention causally affect the outcome of the later recognition, briefly presented contextual information leads to substantial changes in the attentional sampling behavior, which can be best understood in terms of a shifting exploration-exploitation bias.

Following this, we turn to visual processing within the system and ask how invariant object recognition is accomplished despite large variation in retinal input. As an exemplary case, we focus on changes introduced by rotations in depth. Using a variety of techniques, ranging from fMRI to TMS and EEG, we show that viewpoint symmetry, i.e. the selectivity to mirror-symmetric viewing angles, is a prevalent feature of visual processing across a wide range of higher-level visual regions. These findings jointly suggest that viewpoint-symmetry constitutes a key computational step in achieving full viewpoint invariance.

On the next level of abstraction, we investigate how visual categories are represented at different levels of experience, from novice to expert. By combining

training of novel visual categories with psychophysical measures, we demonstrate a change in the underlying type of category representation. Following this, we combine the training paradigm with electrophysiological measurements. In line with our behavioral results, these data reveal a spatiotemporal shift in category selectivity: from late and frontal to early occipitotemporal activity. These results suggest that novel and re-occurring categories rely on partially separate cortical networks, allowing the brain to balance robust and fast recognition with considerable flexibility and plasticity.

The results of all experiments presented are unified by the concept of a system that has evolved efficient mechanisms for robust performance in a large variety of conditions. Using dynamic sampling strategies, computational shortcuts and a division of labor, the visual system is optimally equipped to support higher-level cognitive function in a complex and constantly changing environment.

Contents

List of Figures

List of Tables

1

General Introduction

Vision is the most dominant of our senses. More than 50% of the cerebral cortex is directly or indirectly involved in visual processing (Felleman & Van Essen, 1991), and alone the striate cortex, the first cortical area to receive input from the retina through subcortical areas, contains around 140 million neurons, in each of the two hemispheres (Leuba & Kraftsik, 1994). The task of this intricate system is no less than to give meaning to the visual information entering our eyes at any given moment, to guide behavior, and to support higher-level cognitive function.

The complexity of the system is in stark contrast to our subjective experience of vision. Vision feels simple. Recognizing your friends and family just happens, seemingly without effort, even in the distance, from different angles, in different lighting conditions, and even when they are partly occluded. We perceive the world around us as continuous and stable, even though we constantly move our eyes and every movement drastically changes the two-dimensional projection to our retinas. The integration of these snapshots into a coherent, seamless percept remains largely unnoticed and poses no subjective difficulty.

Yet, the apparent simplicity of vision is deceptive, obscuring the enormous computational complexity of the task. Indicative of this complexity is the amount of cortical resources and therefore energy our brain devotes to vision.

Moreover, no machine, however powerful, is currently able to match the versatile performance of the biological visual system. Tasks as basic as the categorization of visual input, or the segmentation of a scene into distinct components, pose extreme challenges to computer vision but little difficulty to us. This observation alone should make us wonder at how the human visual system accomplishes all of these feats with such remarkably speed and accuracy.

Vision science is the endeavor to understand exactly this. What are the neuronal mechanisms underlying visual function? How is our knowledge about objects represented in the brain? How are novel, previously unseen categories learned so quickly, and re-occurring categories recognized so robustly? What are the rules underlying the sampling of the environment through eye movements and how are the visual impressions from every single fixation of the eyes integrated into our holistic perception of the world?

Finding answers to these questions is not driven only by scientific curiosity. Understanding vision potentially allows us to understand general principles that underlie higher-level cognitive phenomena and cortical function (König et al., 2013). Visual representations are closely linked to perception, memory, and action and have thus been termed "the currency of cognition" (DiCarlo, 2011). In terms of practical applications, understanding biological vision has the potential to drastically improve computer vision (Kietzmann et al., 2009) and with this to improve robotics, for which vision still constitutes a performance bottleneck. Understanding vision also has significant implications for our understanding and treatment of a wide range of visual impairments (including acute and chronic vision loss, hemispatial neglect, prosopagnosia, apperceptive and associative agnosia and alexia), as well as a large range of neurological diseases, which oftentimes also affect visual function. Finally, advancing our understanding of visual processing and learning could ultimately improve how we teach reading and writing to our children, a major stepping stone for their self-determined participation in social life and social interaction.

Within the field of vision, I will focus on the topic of invariant object and face recognition. I ask: (a) whether and how the sampling of our environment interacts with object recognition, (b) how viewpoint invariance, i.e. invariance to rotation in depth, is accomplished, and (c) how the visual system categorizes visual information and represents novel categories in novices and experts. Albeit using a variety of experimental approaches and techniques, the different subprojects are unified by the overall question of how the visual system copes with the demand to robustly extract information from the noisy and ever-changing visual environment.

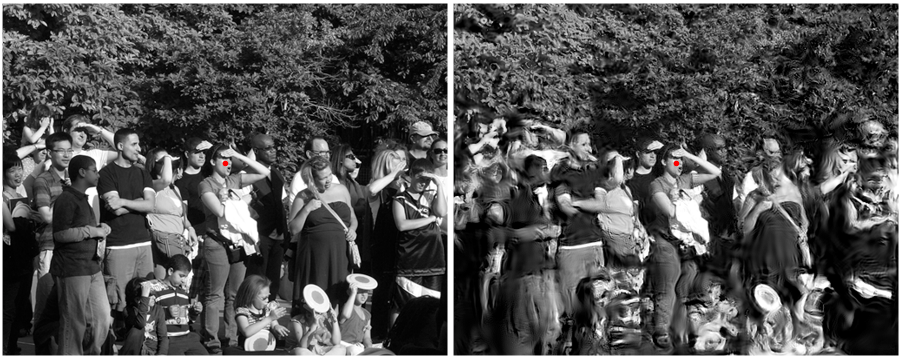

Figure 1.1: Visual Metamers. When fixating on the red dot in the center, both images look almost identical to experimental participants. However, the right image is largely degraded in the outer parts of the image. This manipulation remains largely unnoticed due to the poor visual accuracy in peripheral vision. Figure reprinted with permission.

The following three sections of this introduction will lay the basis for the in-depth experimental chapters. I will provide a more detailed motivation for the respective research questions as well as the chosen experimental approaches and, wherever necessary, provide background information by describing the relevant literature and required terminology.

1.1 Sampling the Environment

Although we usually perceive the visual scene in front of us without gaps and with uniformly high accuracy, the information entering our brains through the retinas is far from perfect. Firstly, part of the visual field is projected onto the optic disc of the retina, which contains no light-sensitive receptors. As a result, we are literally blind in the corresponding part of the visual field. In addition to this usually unnoticed blind spot, the visual acuity is not uniformly distributed across the visual field. The region of 100% (20/20) visual acuity is surprisingly small: it is limited to 2 degrees of visual angle at the center of our gaze. This region is called the fovea (Latin for pit, due to its anatomical shape). The lack of detailed vision in the periphery can be nicely illustrated with visual metamers (Freeman & Simoncelli, 2011). When fixating on the central red dot, both versions of the image shown in Figure 1.1 appear identical. Only upon closer inspection of the outer parts of the images, does it become clear that the right version of the Figure contains drastic manipulations in the periphery.

In order to cope with these limitations, we quickly and constantly move our eyes, at a rate of around 3-5 times a second (Schumann et al., 2008), and focus on, or foveate, different aspects of the visual scene (for example right now while reading this text). With every movement of our eyes, we sample high-accuracy visual information from only a subpart of our environment. Hence, these movement decisions determine which parts of a scene are advanced to the visual cortex for detailed processing. This underlines the decisive contribution of eye movements to our visual experience.

Even more generally, eye movements are among the most frequent actions in the human repertoire. With approximately 230400 eye movements on a given day, they occur twice as often as heartbeats[1]. However, despite this large number of fixations it is not possible to fixate every aspect of a scene, especially when considering that we live and move in a constantly changing environment. As a result, a subset of fixation locations has to be selected. This selection process is not random. As demonstrated by Yarbus (1967), observers consistently fixate informative subparts of a scene (Figure 1.2).

In addition to understanding the selection of elements for focus, studies of eye movements are implicitly also studies into mechanisms of attentional selection, because eye movements are usually aligned with shifts of attention (Yarbus, 1967; Hoffman, 1998; Deubel & Schneider, 1996). Although attention can be shifted without moving the eyes (covert visual attention), it is generally assumed that when eye-movements do occur, they are aligned with attentional focus (Hoffman, 1998), a process termed overt visual attention.

Following the more descriptive work of Yarbus and also Buswell (1935), research advanced along two mainly separated lines to improve our understanding of which aspects of a scene are selected for in-depth processing and which are not. One such line of research was triggered by the observation that patterns of overt visual attention can differ substantially, even in cases of an identical stimulus, when different tasks are being performed (Yarbus, 1967; Hayhoe et al., 2003; Tatler et al., 2006; Betz et al., 2010). These effects are known as 'top-down', since observed differences are not based on the stimulus itself but on higher-level aspects. In contrast to this, other researchers have focused on the effects of the stimulus, studying the relationship between low-level properties and fixation probabilities. They reported elevated stimulus features, such as luminance contrast, at fixated locations (Mannan et al., 1996; Reinagel & Zador,

[1] For the two estimates, I assume an average heart rate of 75bpm and a fixation frequency of 4Hz on a day with 16 awake hours, not including eye movements during REM sleep

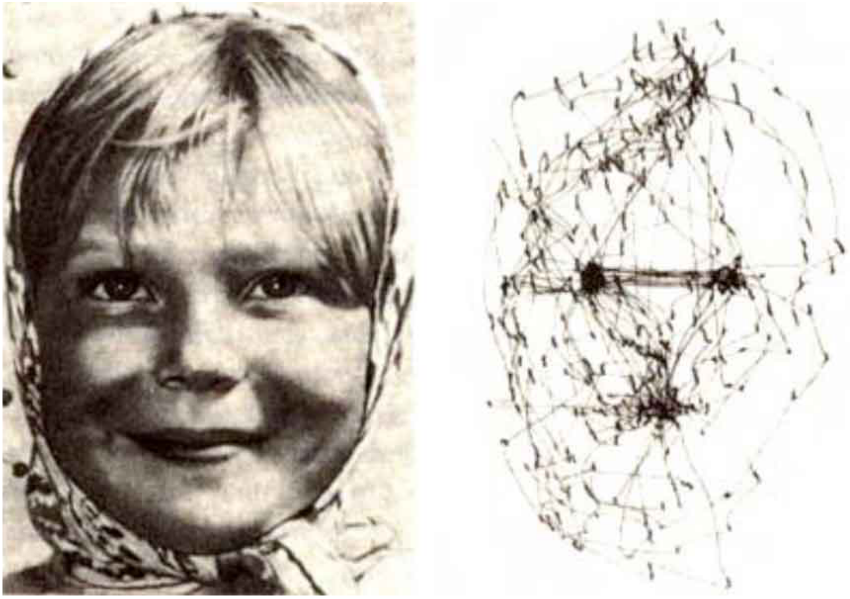

Figure 1.2: Visual Sampling Behavior. When being presented with a face stimulus (left), participants exhibit stereotypical viewing behavior (right), concentrating on some aspects of the face (mostly eyes and mouth), while neglecting others. This example illustrates that eye movements are not random, but systematically focus on informative regions.

1999; Krieger et al., 2000), giving rise to bottom-up, or stimulus driven, accounts of overt visual attention. Extending this view to a wide range of stimulus features, computational models have been proposed which linearly integrate different feature dimensions into a summarizing saliency map to depict the feature-based likelihood of fixations at any point in an image (Koch & Ullman, 1985a; Itti & Koch, 2001a). The rationale of this approach is that a successful prediction of fixation behavior can help reveal the underlying mechanism, as the relative weights of the dimensions in the model can resemble the importance of the respective visual features in the selection process. However, it has been demonstrated that a successful prediction of fixated locations does not necessarily imply a causal relationship (Einhäuser & König, 2003) and that great care must be taken in estimating the success of such models (Wilming et al., 2011).

As a more recent development, objects have moved into the focus of attention research (Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Xu et al., 2014). This line of research suggests that objects, rather than local low-level stimulus properties, constitute the elementary units of attentional selection. In agreement with this hypothesis, it has been demonstrated that knowledge about ’interesting’ objects can lead to improved predictions of overt visual attention (Einhäuser et al., 2008). In parallel to this work, research on object recognition has indicated that the reverse might hold also. Computational vision systems that extend object recognition with saliency-based attentional selection exhibit much enhanced performance, especially in cluttered environments (Walther et al., 2002; Walther & Koch, 2006). Finally, fMRI experiments have shown clearly that selective attention can alter the processing mode of large parts of visual cortex, tuning the system to task- and category-specific features (Cohen & Tong, 2013; Cukur et al., 2013; Tong & Pratte, 2010; Jehee et al., 2011). Taken together, these results advocate an integrated view in which the processes of attention and recognition are intertwined and not strictly separate, as was previously assumed.

Despite these important advances, however, current models and experiments do not explicitly investigate causal interactions of these two processes and typically neglect differences in fixation behavior within the objects themselves. Consequently, it is not known whether distinct patterns of overt visual attention on a given object precede its subsequent recognition or whether the direction of causality is in fact reversed, and distinct patterns of eye movements follow the recognition of an object. In other words, it is unknown whether overt visual attention is a cause or a consequence of object recognition. Framed in a more general context, it is unclear whether the mechanism of attentional selection dynamically changes the sampling behavior upon altered information content in the system or whether the sampling is largely independent. For this, not only the spatial patterns of eye movements are of relevance: temporal aspects, such as fixation durations, need to be considered also, as they can provide valuable insights into changes of the exploration-exploitation behavior.

To address this issue, we performed a series of eye-tracking experiments in which we investigated patterns of eye movements as well as fixation durations in the context of object and face recognition. Central to all studies is the use of ambiguous or bistable figures, which are an extreme form of the ambiguity our visual system faces every day. Although appearing like mere visual peculiarities at first, ambiguous figures provide the clear advantage over regular stimuli that the stimulus can be held constant, while the percept (or recognition) switches between two competing alternatives. This aspect is important, as it prevents experimental effects from being ascribed to simple differences in the underlying stimulus. Moreover, although the perception of ambiguous stimuli can oscillate between the two alternatives given prolonged exposure, it is unique and singular at any given point in time. In investigations of object recognition and attention, this is a necessary prerequisite, as it allows experimental data to be assigned to individual perceptual outcomes. These characteristics make ambiguous stimuli an essential component of vision science.

Apart from explicitly demonstrating that vision is often intrinsically ambiguous, bistable stimuli can provide unique insights into the inner workings of the system, which would otherwise remain unnoticed in studies based on regular stimuli. Consequently, research on ambiguous figures has a longstanding tradition (Necker, 1832; Boring, 1930). Today, we know that the initial percept of such figures can be influenced by a large variety of bottom-up and, more relevant in the current context, top-down aspects (Toppino & Long, 2005). The latter include strong effects of priming and memory (Leeper, 1935; Bugelski & Alampay, 1961), context (Bruner & Minturn, 1955; Bar & Ullman, 1996), and even motivational aspects (Balcetis & Dunning, 2006). Because all of these factors can have a strong impact on the initial perception, they prevent an unbiased investigation of overt attention in the recognition process. It is therefore essential to study the initial perception of naïve observers, which have no prior knowledge of the stimuli.

This particular aspect separates the current set of experiments from previous work on well-known ambiguous stimuli, in which eye movements were mainly investigated during prolonged presentation times, inducing frequent perceptual switches (Pomplun et al., 1996; Ito et al., 2003; Einhäuser et al., 2004). Although the application of such steady-state paradigms has the clear advantage of expediting data-collection, it does not allow for conclusions about the initial interplay of attention and recognition of a previously unseen object, which is our current focus. Moreover, steady-state paradigms complicate the investigation of the direction of causality, as any found differences may either precede the current or follow the previous percept.

In our experiments, we circumvented these issues by measuring eye movements preceding the initial perception of ambiguous stimuli in naïve observers. In the first experiment, we tested whether the fixation patterns that preceded the initial recognition were systematically different for the two possible interpretations of the ambiguous figure. This was accomplished by comparing the experimental data with distributions of eye movements recorded while subjects viewed disambiguated versions of the stimuli. We then tested whether patterns of eye movements would allow us to predict the later percept of our subjects. In a follow-up experiment, we explicitly tested the causal interaction of the two processes by manipulating the initial locus of visual attention on the ambiguous stimulus, and testing whether this minute experimental manipulation affected the subsequent percept of our participants.

Having investigated the impact of eye movements on the subsequent recognition of ambiguous stimuli, we focused on the reverse, testing whether information that affects object recognition would also lead to changes in the attentional sampling process. To do so, we introduced contextual information to the experimental display. Two context frames were designed for each ambiguous image, such that each was congruent with one of the two interpretations. To display the context, a gaze-contingent paradigm was used, which allowed us to present contextual information only in the brief moment preceding the first voluntary saccade. As a result, the context is only shortly visible in the periphery and cannot be fixated. Once the context is hidden, the visual display is identical to the first two experiments, in which no context was shown. As before, we recorded fixation behavior preceding the initial percept of naïve participants. In the analysis, we concentrated on fixation durations to study changes in the exploratory behavior in the presence or absence of preceding contextual information.

In summary, we performed a series of three experiments to advance our understanding of visual sampling and its interplay with object recognition. In the first two, we investigated whether initial eye movements exhibit significant predictive power and causally affect the later recognition. In the third, we tested whether briefly presented contextual information would lead to changes in the overall sampling behavior. Taken together, our results indicate a close and bidirectional interaction of attentional sampling and recognition.

1.2 Visual Invariance

In the previous section I described the necessity of studying eye movements and introduced our experiments in which we explore a more integrated view of overt visual attention and object recognition. In this section, we will turn to problems of visual invariance, which adds considerable complexity to the problem of robust object and face recognition, even when no saccadic eye movements occur.

The core of the problem lies in the fact that identical objects can lead to largely different retinal input in different settings. These include, among others, differences in illumination, size, and position. Although we are usually not aware of it, invariance to these changes it is an essential aspect of vision. As an example, imagine a scenario in which you no longer recognize your conversation partner as soon as a cloud moves in front of the sun, or that a friend becomes unrecognizable while approaching you, because of the increasing size of their image projected onto your retina. Perhaps the most extreme challenge, however, is caused by 3D rotations in depth. Such changes in viewpoint can drastically alter the pattern of retinal input despite the fact that the identity of the object remains unchanged. In all above cases, the visual system is extremely robust to such identity-preserving variation.

How the system achieves this remarkable feat is in the focus of intense research. Following the work of Marr & Nishihara (1978), Biedermann suggested that invariant recognition is achieved by matching simple 3D shapes, or geons, to the surfaces of the seen object. According to his recognition-by-components theory (RBC), object representations consist of combinations of different, change-invariant geons, and the overall shape of an objects is encoded by the spatial relationships of its constituent elements (Biederman, 1987). A mechanistically different approach is rooted in the work by Fukushima (1980), who developed an artificial neuronal network for pattern recognition: the Neocognitron. Inspired by the hierarchical structure of the visual system and extending earlier considerations of Hubel & Wiesel (1962), the Neocognitron consists of alternating layers of increasing specificity and invariance (called S and C layers respectively, reminiscent of the model of simple and complex cells described by Hubel & Wiesel (1962)). Starting from an S-layer of Gabor-like edge-detectors, a layer of C-cells pools information of similar features across space and thereby achieves partial position invariance for these features. The next layer, consisting again of S-cells, then recombines the output of these C-cells to obtain more complex feature selectivity. Ascending through the hierarchy, the iterative combination of simpler features leads to increasing complexity and therefore specificity, while visual invariance is accomplished by combining information from multiple feature-detectors.

Although Neocognitron has predominantly been tested for position invariance, the overall approach can straight-forwardly be extended to other types of invariance, such as 3D viewpoint invariance. With this, it directly matches the view-based theory of object recognition, which suggests that viewpoint invariance is accomplished by combining information from a potentially small set of 2D views, without the requirement for three-dimensional internal representations such as geons. Initiated by computational models, which provided a first proof of concept (Poggio & Edelman, 1990; Ullman & Basri, 1991), the theory soon gained experimental support from behavioral data in monkeys (Logothetis et al., 1994) and humans (Bülthoff & Edelman, 1992; Bülthoff et al., 1994; Tarr & Bülthoff, 1995; Tarr et al., 1998), as well as electrophysiological data in monkeys (Perrett et al., 1991; Logothetis et al., 1995). In the latter experiments, the majority of cells in the macaque temporal cortex was found to be view-specific. However, smaller populations of viewpoint-invariant cells in the macaque have been reported, too (Logothetis et al., 1995; Perrett et al., 1998), providing evidence for a view-invariant representation, potentially at a later stage of processing. More recently, single-cell recordings performed in the medial temporal lobe of pharmacologically intractable epilepsy patients showed large degree of viewpoint-invariance also in humans (Quiroga et al., 2005).

Extending the computational aspects of the theory, Riesenhuber and colleagues combined the idea of a Neocognitron-like neural network of alternating S and C layers with a view-based account of object representations. This ultimately lead to the development of HMAX (Riesenhuber & Poggio, 1999, 2000), which, after adjustments and extensions to match receptive field properties of cells in macaque visual cortex and natural scene statistics (Serre et al., 2007b), now consists of four S and four C layers and an additional classification layer (Figure 1.3). In its most recent version, HMAX exhibits robust recognition performance in a variety of tasks, en par with human performance in a rapid categorization task (Serre et al., 2007a). While the selectivity of simulated neurons in HMAX is based on static images, other approaches exploit the temporal structure of sequences of natural input to optimize receptive field properties in an unsupervised fashion (Wallis & Rolls, 1997; Einhäuser et al., 2005; Wyss et al., 2006; Franzius et al., 2008; Rolls, 2012). Underlying this family of computational hierarchies is the idea that an optimal neuronal code would reflect the fact that the identity of a seen object varies on a considerably slower timescale than the rapidly and constantly changing retinal signals. This does not imply that neurons with corresponding receptive fields should respond slowly. Rather, neurons should respond robustly and quickly to slowly changing aspects of a scene, because these elements are most likely to contain the important semantic information.

Despite the computational and experimental advances described above, the specifics of how invariant object information is represented in the brain remain unclear. While there is growing experimental consensus that cells in the inferotemporal cortex (IT) provide a neuronal activation pattern, or code, that supports complete viewpoint-invariant representations of objects (Booth & Rolls, 1998; Logothetis et al., 1995) and faces (Freiwald & Tsao, 2010; Perrett et al., 1991), the exact type of code remains a matter of debate. One possibility is that extremely sparse subsets of neurons, potentially even single cells, in anterior

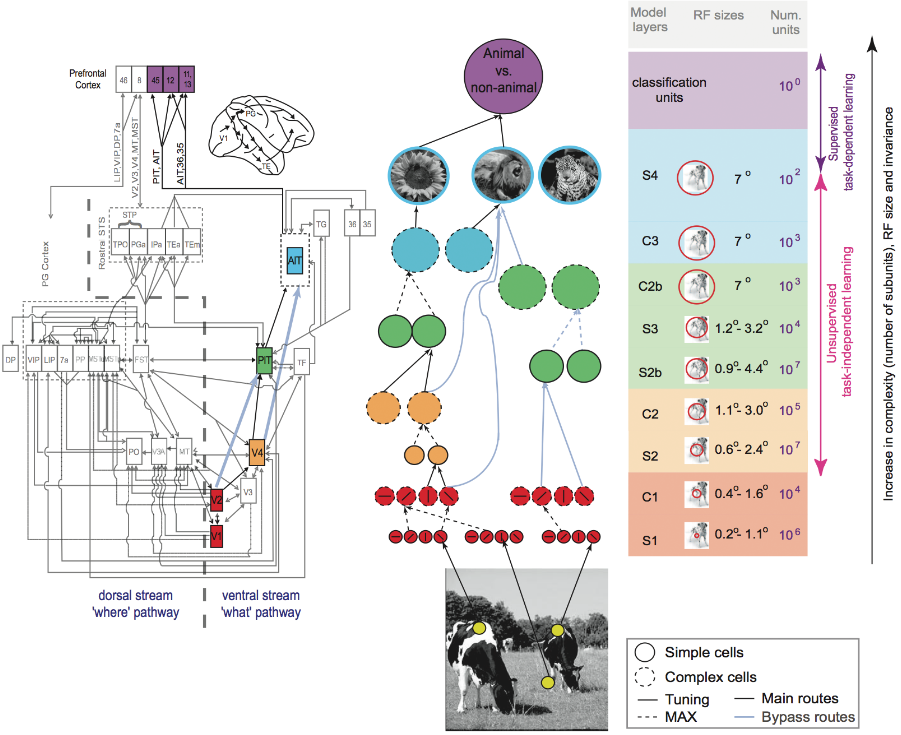

Figure 1.3: The HMAX Model. Similar to the Neocognitron Architecture, HMAX alternates between layers of specificity (S) and invariance (C). Whereas the S1 layer consists of Gabor-like receptive fields, resembling V1 simple cells, higher level selectivity layers are randomly imprinted based on natural image input. According to the model, task-dependent learning affects only the highest levels of the hierarchy. Reprinted with permission (Serre et al., 2007a)).

parts of the temporal lobe or prefrontal cortex (PFC) pool information from view-selective units and achieve full viewpoint invariance by interpolating between these views (Riesenhuber & Poggio, 2002b; Quiroga et al., 2005; Bowers, 2009). An alternative hypothesis states that viewpoint independence is achieved with a population-based and distributed code (Young & Yamane, 1992; Perrett et al., 1998) rather than through single, fully invariant cells. As an extension of this view, it has been proposed that single IT cells do not need to exhibit complete invariance to changes in size, position, and viewpoint, as long as the respective object representations, or manifolds (Edelman, 1999), remain separable in the high-dimensional feature space spanned by the set of neurons (DiCarlo & Cox, 2007). Such coding scheme has the clear advantage that multiple types of information can be independently decoded from the same population, including view-invariant object identity and position information, simply by using different readout strategies (Hung et al., 2005). At first sight, these two approaches, single-cell vs population-based, seem incompatible. However, if neuronal selectivity is understood in terms of sparseness (Connor, 2005), they can be seen as lying on a continuum (Quiroga & Kreiman, 2010) with massively distributed representations on one end, and maximally sparse 'grandmother cells' serving as theoretical upper limit.

Independent of the true form of view-invariant representations in the visual cortex, another related aspect deserves consideration. If cells with view-specific selectivity form the basis of viewpoint invariance, then the question arises how many, and which, views are required to allow for full invariance. Evidence for a non-uniform sampling of the space of all viewpoints, i.e. for the hypothesis that only a small number of views are stored, comes from behavioral studies in humans (Bülthoff et al., 1994; Tarr, 1995) and monkeys (Logothetis et al., 1994). These results are in accordance with suggestions by Perrett et al. (1998), who were able to explain the prolonged reaction times, observed when presenting objects from previously unseen views, in terms of slower activity accumulation due to lower firing rates for non-preferred views. The most extreme version of the view-based theory suggests that recognition could be accomplished by a single canonical view, to which a seen object is internally aligned for recognition(Palmer et al., 1981; Ullman, 1989). Although the latter was shown to be at odds with experimental data (Bülthoff & Edelman, 1992), a more general concern is that the notion of a universal number of required viewpoints, identical for different objects, is itself highly unlikely. First, Edelman & Bülthoff (1992) observed that canonical-preference effects were reduced with extended experience. This suggests that frequently observed objects are represented in more detail, based on a larger, and therefore changing, number of viewpoints. Second, different objects can differ vastly in the complexity of their three-dimensional structure. While some objects inherently exhibit larger rotational invariance, others are structurally more complex (compare for instance the complexity of a rotating apple and a rotating face) (Tjan & Legge, 1998). Finally, spontaneously selected viewpoints, despite being clearly non-accidental, can differ with different tasks (Blanz et al., 1999). As a result of these considerations, a more dynamic recruitment of stored object views is a more likely coding scheme. Evidence for the feasibility of such an account was again provided by computational modeling (Kietzmann et al., 2008). It was shown that a recognition system, when pressed for both, representational efficiency and performance, will assign different numbers of viewpoints to different objects, depending on their underlying complexity. The resulting system achieved a fourfold reduction in the total number of required views, with no observable decrease in performance. Interestingly, the selected views closely matched the ones generally assumed to be canonical. This model was later extended to a hierarchical setting by building on the C2 layer of the HMAX hierarchy Kietzmann et al. (2009). The resulting system automatically extracts different numbers of viewpoints, while at the same time focusing on subparts of the C2 feature space, comparable to a multiplicative attentional weighting of relevant features. Experiments with this biologically more plausible setup showed that the number of required views, as well as the selected features changed in a task-specific manner. Taken together, it is rather unlikely that a single, fixed number of viewpoints is applicable to all objects. Instead, experimental and computational work suggests that the most likely answer as to how many views are required to span the full range of viewpoints of a given object lies again on a continuum spanning from a single canonical view to a dense uniform sampling.

Independently of the final number of views required to achieve full viewpoint invariance, all current view-based theories (Bülthoff et al., 1994; Bülthoff & Edelman, 1992; Logothetis et al., 1995; Tarr & Bülthoff, 1995; Tarr et al., 1998) and computational models (Poggio & Edelman, 1990; Ullman & Basri, 1991; Riesenhuber & Poggio, 1999; Kietzmann et al., 2009, 2008) posit that the entire space of views needs to be covered to allow for robust performance. Curiously, the seemingly unrelated effect of mirror-image confusion suggests that the visual system might be able to exploit the axial symmetry of many objects, including faces, to reduce the number of required viewpoints. Problems in distinguishing lateral mirror images, or enantiomorphs, can be observed in many situations, but are most notably in children who currently learn to read and write. Many children confuse the letters 'd' and 'b', but not 'd' and 'q', although both cases are based on a simple mirror transformation (Dehaene et al., 2010). Even more strikingly, many children spontaneously write from right to left, a phenomenon known as mirror writing, when not enough space is provided to the right (Cornell, 1985) (Figure 1.4). Such effects are, however, not limited to children: adults also regularly confuse images and their mirrored counterparts. For example, when being presented with a famous painting, such as the Mona Lisa shown in Figure 1.5, a large proportion of observers fail to correctly judge whether they are looking at the original or a mirror-reversed version (Walsh, 1996). As a first electrophysiological study of this peculiar effect, Rollenhagen & Olson (2000) found that responses of neurons in IT were more similar for lateral

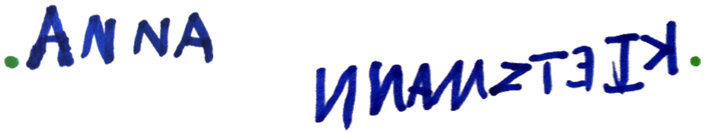

Figure 1.4: Spontaneous Mirror Writing. Early in their experience of learning to read and write, children spontaneously write complete words from right to left, mirroring individual letters, when not enough space is provided to the right. (image curtesy of Anna Kietzmann)

mirror images than for vertical mirror images, in accordance with the confusion of ’d’ and ’b’ in school children. Similar observations have been made in the human temporal cortex using functional magnetic resonance imaging and images of natural scenes (fMRI) (Dilks et al., 2011).

But how does mirror-image confusion aid in acquiring viewpoint invariance? An intermediate, mirror-invariant visual representation would substantially reduce the computational complexity, particularly for objects with axial symmetry, such as faces, animals and many objects (Vetter et al., 1994). For this type of input, an equally large leftward and rightward rotation away from the front view would result in the same pattern of neural activity, thereby considerably reducing the overall problem. An essential prerequisite for the claim of reduced complexity, however, is that invariance to mirror reversals is not the result of already achieved viewpoint invariance. A possible test of this is to show an object or face from different viewpoints and to determine whether viewpoints that are mirror images of each other, such as +60° and −60°, lead to more similar response patterns than viewpoints that are not mirror-images. Indeed, Freiwald & Tsao (2010) demonstrated this property in the anterior lateral face patch of the macaque temporal lobe using single-cell recordings. What remains unknown is whether such viewpoint symmetry can also be found in humans. Moreover, single-cell recordings are often restricted to a very focal region. It is therefore an open question whether the effect is a local phenomenon, specific to a single region in the face-processing network, or whether viewpoint symmetry is a more prevalent representational aspect of visual processing that can be observed in other visual areas of the ventral and dorsal visual cortex, including regions that prefer other types of stimuli such as objects and scenes.

To elucidate these issues, we performed an fMRI study in which we monitored cortical activity while participants were presented with pictures of upright and inverted faces shown from five different viewpoints. The data was

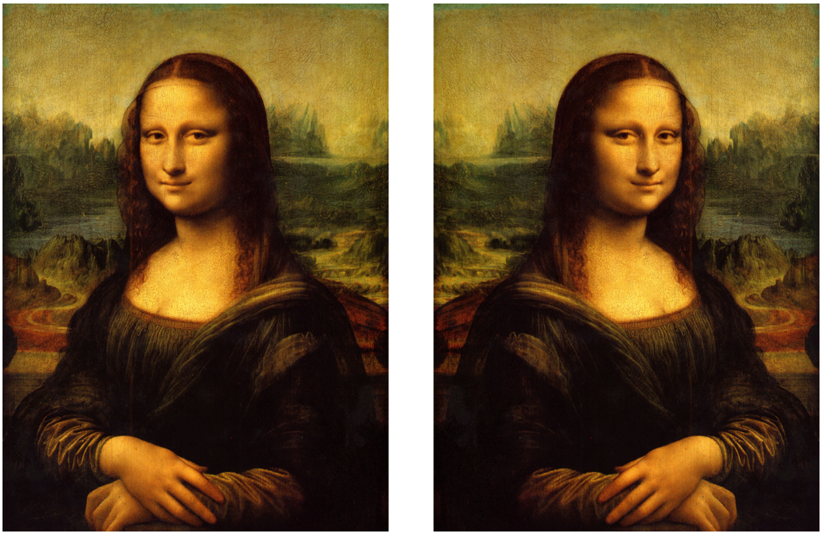

Figure 1.5: Visual Memory and Mirror-invariance. Visual memory is mostly invariant to left and right. Which one is the original?[2]

analyzed using an approach related to Representational Dissimilarity Analysis (Kriegeskorte et al., 2008a), a form of multivariate pattern analysis (MVPA), which focuses on activation patterns instead of univariate changes in overall BOLD signal (Haynes & Rees, 2006; Norman et al., 2006; Tong & Pratte, 2010). Based on these data we tested whether mirror-symmetric viewpoints exhibit more similar activation patterns than non-symmetric ones. Potential effects of low-level similarity were ruled out by a carefully designed stimulus set, which was guided by the results of a biologically plausible model of V1 response properties, comparable to the S1 layer in the HMAX hierarchy. Our analyses were based on predefined regions of interest (ROIs), as well as a spatially unconstrained searchlight analysis (Kriegeskorte et al., 2006) across the occipital, temporal and parietal cortex. In a follow-up experiment, we focused on the causal mechanism underlying viewpoint symmetry by utilizing a transcranial magnetic stimulation (TMS) protocol. Targeting the occipital face area (OFA), we tested whether judgments of viewpoint symmetry are influenced by activity in the hemisphere ipsilateral to the visual stimulation.

In a third experiment, we recorded Electroencephalography (EEG) data, again with visual stimuli comprised of various viewpoints of faces. The com-

[2] The original Mona Lisa is shown on the left

bination of high temporal resolution of EEG and multivariate pattern analyses allowed us to estimate the latency of viewpoint symmetry. In addition to replicating the overall effect, this aspect is of particular interest, as the latency of the effect can be informative of the underlying cortical mechanism. In addition to the focus on viewpoint symmetry, we performed data- and model-driven analyses, revealing dynamical changes in the encoding of face viewpoints with high temporal resolution.

1.3 Categorization and Plasticity: Adapting to an ever-changing environment

While the previous section focused on identity-preserving variation, with emphasis on invariance to rotations in depth, this section will explore an even more general form of invariance: categorization. Instead of considering potential changes in the retinal projection of a single object, successful categorization requires the ability to generalize across large intra-category variability. Generalizing from previous encounters with other objects of the same class enables the recognition of previously unseen objects and the successful prediction of their function and behavior. Classification is thus a fundamental aspect of visual processing and higher-level cognitive function.

Although it is itself sufficiently challenging, categorization is additionally complicated by the fact that we live in a constantly changing environment. The visual system must not only continually adapt existing category representations, but also learn entirely novel categories, often with very limited exposure (one-shot learning). In addition to rapid learning speeds, re-occurring categories need to be robustly and quickly recognized. This poses a computational challenge, as the system must retain sufficient plasticity to allow for the rapid learning of novel categories, and at the same time optimally tune to re-occurring categories to achieve high processing speeds.

To better understand the cortical mechanisms underlying the efficient yet adaptable categorization capabilities of the brain, we need to identify the mechanism of category learning and investigate how category information is represented at different levels of experience, from novice to expert. Similarly to the temporal aspects of viewpoint symmetry, knowledge of how quickly category information is extracted after the onset of visual processing can impose constraints on possible models of the neural mechanisms. Finally, to understand visual processing on a more general level we need to understand how high-level category information is integrated with lower-level information during categorization and perceptual decision making.

In the field of psychology, a variety of mechanisms have been suggested to account for human categorization behavior. Proposals for the basis of categorical representations have included: the explicit storage of category exemplars (Hitzman, 1986; Nosofsky, 1986; Nosofsky & Palmeri, 1997); prototypes representing either the most clear category exemplar (Rosch, 1973) or the central tendency of all exemplars, which does not need to match any encountered exemplar (Posner & Keele, 1968); as well as an abstract decision boundary (Ashby & Gott, 1988). A combination of these different approaches has also been proposed. Mixture models suggest that the number of stored elements can vary depending on the complexity of the stimulus set, category, or task (Tjan & Legge, 1998; Rosseel, 2002). These models were in part proposed as an answer to the claim that exemplar-based representations are inefficient at storage and retrieval. This criticism builds on the intuition that the retrieval of a category label is highly uneconomical if the visual input needs to be compared to every single object exemplar ever encountered. Mixed models regain efficiency by combining different types of category representation. If members of a category never need to be identified on a subordinate level, it is conceivable that a more general representation based on a small set of exemplars, prototypes, or an abstract decision-boundary suffices, whereas categories whose members need to be individually separated would require an exemplar-like format. Finally, the type of representation can be subject to change and there is no principled reason why multiple category representations cannot be implemented simultaneously.

Although a review of the extensive psychological literature on the topic is beyond the scope of this introduction, one final aspect is worth noting. Exemplar-based models of recognition have been described as complementary to the view-based theory of view-invariant object recognition described in the previous section (Palmeri & Gauthier, 2004). In line with this proposal, (Ullman, 1998) suggested that storing views of different objects from the same class allows for generalization to other category exemplars. In line with this proposal, we demonstrated that a single computational architecture could be used for view-invariant object recognition and object classification Kietzmann et al. (2009). Similar to mixture models, our model assigns different numbers of instances to individual classes depending on the underlying category structure and task complexity. At the same time, the relevance of the different features is dynamically reweighted to allow for optimal category separation.

Independently of the type of category representation, scientists of different disciplines have asked where category selectivity is encoded in the brain and which areas are involved in category learning. Originating from computational neuroscience, HMAX and its successors (Riesenhuber & Poggio, 1999; Serre et al., 2007a) are implemented around the idea that all levels of categorization, subordinate, basic and superordinate, and all visual tasks are based on the same visual feature space, provided by the IT, and that category selectivity is performed only on the highest level, e.g. in the PFC. The PFC is assumed to contain task-related units that can dynamically read out and combine patterns of visual features. Put differently, response patterns in IT serve mainly as a complex feature set, which does not in itself provide category selectivity, and categorization is interpreted as a late perceptual decision process rather than an aspect of visual selectivity. In analogy to this, the two-stage model of perceptual category learning hypothesizes that areas in the ventral stream acquire sharper tuning to re-occurring stimulus properties whereas nonvisual prefrontal areas perform the actual categorization in a task-dependent manner. Experimental evidence in support of such separation of feature- and category-selectivity in humans was provided by (Jiang et al., 2007), who showed that although category training affected both IT and PFC, only PFC responded differentially to changes in category membership. On the other hand, training effects in IT were best explained in relation to changes in physical appearance. Converging evidence for this view was provided by monkey electrophysiology (Freedman et al., 2001; Thomas et al., 2001; Freedman et al., 2006).

That being said, the validity of the two-stage model relies in large part on the negative results regarding category-related learning effects in IT. However, there is evidence that IT can indeed encode category-relevant information following category training in macaques (De Baene et al., 2008; Sigala & Logothetis, 2002) and human (Li et al., 2009), even if stimuli are not actively categorized (Folstein et al., 2012a), and in cases of categorical perceptual decisions (McKeeff & Tong, 2007). Moreover, areas in the ventral stream usually considered face-selective can adapt their selectivity with prolonged experience, as demonstrated for car and bird experts (Gauthier et al., 2000a). Related to this, although not explicitly formulated in the setting of categorization, the reverse hierarchy theory of perceptual learning (RHT) (Ahissar & Hochstein, 2004) proposes that learning only starts at the highest levels of the hierarchy. If more detailed information is required to fulfill the task, practice-induced learning effects progress down to lower levels in a top-down guided manner. Plasticity is therefore not limited to the highest levels, but is assumed to follow a gradient from high- to low-level.

The theory rests on psychophysical measurements in humans using an odd-one-out paradigm and can account for the observation that the learning of easy perceptual tasks usually precedes the learning of more complex ones.

A different approach to understanding cortical categorization mechanisms rests on investigations of temporal rather than spatial aspects of categorization. Knowing the earliest time-point at which category information is present in the system can provide valuable cues as to which types of processing are involved. Short latencies imply only few synaptic transmissions, and are therefore usually interpreted as the result of feed-forward processing. Following this approach, psychophysical and electrophysiological experiments have provided ample evidence that naturally occurring categories can be extracted after about 100ms in humans (Thorpe et al., 1996; Liu et al., 2002; Kirchner & Thorpe, 2006; Liu et al., 2009; Carlson et al., 2011; Cichy et al., 2014) and from macaque IT (Hung et al., 2005).

In light of the current focus on 'pure' category selectivity, however, naturally occurring categories can only provide limited insight, because they can exhibit systematic differences in low-level stimulus properties, which prevent stronger claims about the underlying signal (VanRullen, 2011; Crouzet & Serre, 2011). Moreover, the phenomenon of category learning cannot be investigated with natural categories because subjects will almost always have extensive prior experience. A solution to these limitations is the use of artificial categories defined in a feature space that allows for a close control of low-level properties. In addition to allowing for exact control of the subjects' exposure to the categories, this approach has the advantage that it allows for baseline measurements prior to category learning. Such measurements are very important to ensure that found category effects are due to learning and not the result of residual low-level properties of the stimulus space.

The above considerations reveal the need to further investigate how novel and re-occurring categories are represented in the human brain and how high-level categorical information is integrated with lower-level stimulus properties. Moreover, it is currently unknown whether the latency of category selectivity changes and whether different cortical networks are involved at different stages of category training. To investigate these issues, we conducted a series of experiments in which we collected behavioral data and recorded functional neuroimaging data using Magnetencephalography (MEG) while extensively training our subjects to distinguish two artificial visual categories of faces embedded in four-dimensional parametric feature space. This approach allowed us to control the low-level properties of the two trained categories, as well as the subjects' experience with them. Furthermore, testing subjects in multiple sessions across training enabled us to monitor changes in the cortical representations of the trained categories.

In a first behavioral experiment, we focused on how high-level categories are cortically represented and how low-level stimulus properties and high-level category information are integrated in perceptual decisions. Across training sessions, the subjects were tested in an odd-one-out paradigm in which they had to indicate whether a target stimulus, presented with varying numbers of distractors, was located in the left or right visual field. In this setup, exemplar and more abstract decision-boundary models make distinct predictions about the pattern of results. The experimental approach therefore allowed us to monitor changes in the type of category representation with extended training. Additionally, the experimental paradigm allowed for a direct test of predictions of RHT about the temporal sequence of high-and low-level visual properties in perceptual decisions.

In a second experiment, we again conducted a longitudinal study in which our participants received extensive training with two artificial categories, using a similar feature space as in the previous experiment. As an extension to the previous work, the training was combined with MEG recordings of brain activity. MEG offers high temporal and good spatial resolution and is therefore perfectly suited to estimate temporal and spatial aspects of category learning. To guarantee that any effects observed were the result of category training and not of low-level stimulus characteristics, we included an MEG baseline measurement. All data recorded during training was then tested against this baseline dataset. In the analyses, we first concentrated on temporal aspects of category selectivity, asking for the earliest point in time at which the category of a presented stimulus is extracted. Following this, we performed a source analysis to estimate the cortical areas involved in category selectivity at different stages of category learning.

1.4 About this dissertation

1.4.1 Why faces?

Most of the experiments described in the following are based on stimulus sets containing faces. This highly restricted subset of all possible visual categories was chosen for a number of reasons. For one, faces are among the most important visual categories we must distinguish, providing essential information for our everyday behavior and forming the foundation of our social interactions. As a result, the highly developed face-recognition system of the brain extracts facial information from a given stimulus with unparalleled accuracy and robustness. The signals evoked by the brain in response to these stimuli are strong, which makes them prime candidates for experimental studies of visual processing. Additionally, faces are a rich stimulus set, enabling a large variety of interesting research questions including many general aspects of vision. Faces can be detected (face vs. no face), classified (sex, ethnicity, known vs. unknown), and identified. Face recognition is robust even in the most diverse situations, and faces can be learned, often with extremely limited exposure. Moreover, studies focusing on social aspects of vision find rich information about emotions and shared attention in the display of a face. Finally, unlike many objects, faces almost constantly change, for instance through the display of emotions or during speech production, and therefore combine static and dynamic aspects of visual processing.

Faces are an especially diverse stimulus category, but it is still worth considering whether findings based on faces allow for more general claims about visual processing and other visual categories. While faces undeniably lead to specific spatiotemporal activation patterns in the brain (Kanwisher et al., 1997a; Bentin et al., 1996), different lines of research now suggest that faces are not an inherently special category. On the behavioral side, some effects previously thought to hold uniquely for faces have been shown either to exist equally strongly for non-face objects if the correct controls are performed (Wong et al., 2010), or to be based on low-level confounds (VanRullen, 2006). In the area of neuroimaging, visual selectivity to faces has been at the center of a fierce debate. One of the most prominent theories holds that faces and few other categories are represented in a modular fashion, with each module being uniquely selective to a single category. Among others, this view is based on the observation that faces strongly activate only a small part of the fusiform gyrus, namely the fusiform face area (FFA), which suggests that this area is mainly face selective (Kanwisher et al., 1997a; McKone et al., 2007). Support for the face-selectivity in this region is substantial. In addition to fMRI experiments, it includes findings from monkey electrophysiology (Tsao et al., 2006), lesion studies (Moscovitch et al., 1997; Sergent & Signoret, 1992) and electrical stimulation experiments (Parvizi et al., 2012). Still, the modular theory of face representations has been challenged on multiple grounds. For one, it was suggested that high-level visual areas follow a distributed rather than modular organization, which encompasses a large variety of categories (Haxby et al., 2001; Huth et al., 2012). As a middle ground between these extreme views, the seemingly modular representation of faces has been ascribed to the co-activation of multiple 'selectivity maps' overlapping in visual cortex (Op de Beeck et al., 2008; Cohen & Tong, 2001). While each of such feature maps is only weakly activated by itself, faces jointly activate multiple features that co-occur in a focal region of cortex: the FFA. From this view it follows that all object categories activate the underlying feature maps, but that faces just so happen to activate cortically co-aligned features, thereby creating a modular activation pattern.

As a second, more substantial challenge, the face-selectivity of the FFA has been questioned. The process-map hypothesis (Gauthier et al., 2000a; Tarr & Gauthier, 2000) describes the selectivity of FFA as the result of visual expertise, which is not limited to faces. By this account, the FFA represents any extensively foveated element (Hasson et al., 2002), for which detailed visual processing is required. Hence, faces and categories of expertise share the same cortical regions (Mcgugin & Gauthier, 2013; McKeeff et al., 2010). In line with this proposal of a dynamic recruitment of resources, it has been shown that the organization of the FFA itself is subject to change. As one example, it has been shown to increase in size from children to adults (Golarai et al., 2007; Scherf et al., 2007). Compelling behavioral evidence for experience-dependent changes in the representation of faces was provided by researchers focusing on face recognition in different age groups. They found that their participants were significantly better at recognizing others in their own age range as compared to other age groups, the so-called own-age effect (Anastasi & Rhodes, 2005; Hills & Lewis, 2011). These results were corroborated by experiments with preschool teachers (de Heering & Rossion, 2008) who, compared to an age-matched group of adults with no regular experience with children, exhibited stronger recognition performance when presented with faces of children.

Taking this approach one step further, the results of self-organizing computational models have recently been used to argue against a principled division between the cortical mechanisms underlying face and object recognition (Wallis, 2013). It was shown that holistic perceptual effects and the cortical arrangement of face- and object-selective areas are naturally emerging properties of a hierarchical, feature-based recognition system. Based on these simulations it was concluded that the "face recognition system can be viewed as a carbon copy of the object recognition system in terms of competitive mechanisms involved in its construction" (Wallis, 2013). Taking the presented computational and experimental evidence together, it is safe to assume that studies of face processing can provide valuable insights into the more general cortical mechanisms underlying visual processing.

Finally, the experiments described in this dissertation do not solely rely on any face-specific spatial or temporal components and all closely control for low-level stimulus properties. We therefore expect our findings to generalize well to other stimulus categories.

1.4.2 Chapter Overview

Having now motivated the centrality and importance of the three investigated aspects on a more general level, the following chapters will provide in-depth descriptions of the performed experiments and results. Chapter 2 describes the eyetracking experiments, which we performed to investigate the topic of 'Sampling the Environment'. We will show that overt visual attention significantly precedes the recognition of ambiguous stimuli and that contextual information can have strong effects on the ongoing sampling process, arguing in favor of a more integrated view of the two processes. Chapter 3 will report the results of the fMRI, TMS and EEG experiments conducted to study visual invariance. Focusing on rotations in depth, it will be demonstrated that effects of viewpoint symmetry can be observed in a large range of higher-level visual areas and at a comparably early stage of processing. Moreover, we observed a cascade of different viewpoint encoding schemes, each supporting a distinct function in face processing. Chapter 4 centers on the topic of 'Categorization and Plasticity', describing the psychophysical and electrophysiological (MEG) experiments performed. Our behavioral data suggest that high-level category information is integrated into the perceptual decision process only after extensive training. In line with these late high-level category effects, we found furthermore that extensive training can result in a spatiotemporal shift of the cortical activity that underlies visual categorization. Finally, following these in-depth experimental chapters, the discussion will provide an overall summary of the individual chapters and explore their relationship in the context of efficient and robust processing. Based on the experimental results presented, I will argue that the visual system developed highly efficient mechanisms to cope with the constantly changing visual environment and the requirements of the embedding organism. These include adaptive sampling, cortical plasticity and computational shortcuts to visual invariance, such as viewpoint symmetry, even if they imply a non-veridical representation of the external world.

2

Sampling the Environment

2.1 Overt Visual Attention as a Causal Factor of Perceptual Awareness[1]

Abstract Our everyday conscious experience of the visual world is fundamentally shaped by the interaction of overt visual attention and object awareness. Although the principal impact of both components is undisputed, it is still unclear how they interact. Here we recorded eye-movements preceding and following conscious object recognition, collected during the free inspection of ambiguous and corresponding unambiguous stimuli. Using this paradigm, we demonstrate that fixations recorded prior to object awareness predict the later recognized object identity, and that subjects accumulate more evidence that is consistent with their later percept than for the alternative. The timing of reached awareness was verified by a reaction-time based correction method and also based on changes in pupil dilation. Control experiments, in which we manipulated the initial locus of visual attention, confirm a causal influence of overt attention on the subsequent result of object perception.

[1] Part of the experimental work presented in this section was started as a MSc project of Tim Kietzmann and Stephan Geuter. It was later extended, and published together with Stephan Geuter and Peter König as a peer reviewed article in PLOS One. See Publication List for details.

The current study thus demonstrates that distinct patterns of overt attentional selection precede object awareness and thereby directly builds on recent electrophysiological findings suggesting two distinct neuronal mechanisms underlying the two phenomena. Our results emphasize the crucial importance of overt visual attention in the formation of our conscious experience of the visual world.

2.1.1 Introduction

Conscious object recognition and overt visual attention belong to the most essential capabilities of the human visual system and cognition. Because both substantially contribute to our everyday experience of the world, they have moved into the center of scientific interest. Although there is evidence for their interconnection on a behavioral level (Einhäuser et al., 2004; Zhaoping & Guyader, 2007), the two phenomena were recently shown to rely on distinct neuronal mechanisms (Wyart & Tallon-Baudry, 2008; Fernandez-Duque et al., 2003), and are most often investigated in isolation (Itti & Koch, 2000; Kawabata & Mori, 1992; Georgiades & Harris, 1997; Toppino & Long, 2005; Tse et al., 2005). As a result, the exact roles and temporal dynamics governing the interplay of the two processes remain unclear. In this context, one of the most fundamental questions is whether overt visual attention has a causal impact on the perceptual outcome of the recognition process (also named object perception hereafter), or whether the direction of causality is in fact reversed and that the awareness of an object's identity guides subsequent patterns of eye-movements.

These two views can be characterized by two hypotheses. The first hypothesis sees overt visual attention as following the perceptual outcome. According to this view, fixations are guided towards crucial local features of the object only after the subjects are aware of its identity (*action follows perception* hypothesis) (Pheiffer et al., 1956). The competing hypothesis suggests that features that are attended to prior to recognizing the object substantially contribute to the perceptual outcome (*action precedes perception* hypothesis) (Holm et al., 2008). In this scenario, fixation patterns are in line with the upcoming percept prior to the actual awareness of the object identity.

To probe these two hypotheses, we conducted two eye-tracking experiments based on a set of twelve ambiguous stimuli. In order to provide a baseline of viewing behavior corresponding to the different perceptual outcomes, two unambiguous stimuli were created from every ambiguous one that bias the initial perception towards one of the two interpretations (Figure 2.1).

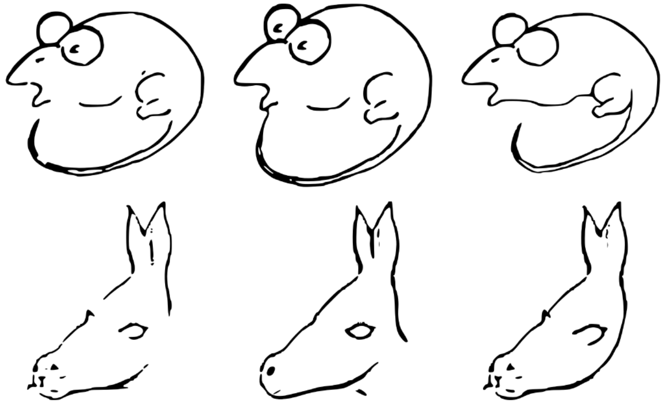

Figure 2.1: Exemplary Stimuli. Two out of the 12 stimulus sets used in the experiment. Each row shows an ambiguous image in the first column with the two unambiguous versions next to it.

In the main experiment, experiment 1, we first investigate whether distinct patterns of overt visual attention precede different perceptual outcomes during the presentation of ambiguous stimuli. To ensure that only fixations prior to object perception are taken into account, we apply two correction methods. The first is based on the minimum reaction time of individual subjects during the complete experiment. The second is based on percept-related changes in pupil dilation that were recently shown to significantly precede perceptual events (Einhäuser et al., 2008; Hupé et al., 2009). Following this, we explore the possibility that object awareness has an effect on the subsequent patterns of overt visual attention. This is accomplished by comparing the viewing behavior of the subjects before and after the perceptual event.

In experiment 2, we investigate whether changes in the initial locus of visual attention, induced by shifting the initial gaze position of the subjects to different parts of the stimulus, would have a causal effect on the later perceived object identity.

2.1.2 Experiment 1: Materials and Methods

Participants

Seventy-eight subjects, recruited via university mailing lists, took part in the experiment. The data of five subjects was discarded due to insufficient calibration accuracy. Forty-nine of the remaining 73 subjects were female. All subjects had normal or corrected to normal visual acuity and were informed of their right to withdraw from the experiment at any time without the need to state a reason and gave written informed consent to participate. Furthermore, all subjects were informed of the experimental procedure and were naive to the purpose of the study. Upon completion of the overall experiment, the subjects were debriefed. The study, including experiments 1 and 2, was approved by the Ethics Committee of the University of Osnabrück.

Stimuli

The used stimulus set consisted of 12 ambiguous stimuli, for each of which two additional disambiguated versions were created (leading to 36 stimuli in total). To create the disambiguated versions, the ambiguous stimuli were altered such that they would favor either one of the two percepts (see Figure 2.1 for an example). This was accomplished by manually adding or deleting small elements of the original ambiguous images. The 12 ambiguous stimuli included a version of the ambiguous donkey/seal figure (Fisher, 1968a), an image allowing for the percept of a woman's face or a saxophone player by Sara Nader, the man/mouse figure (Bugelski & Alampay, 1961), an ambiguous stimulus showing a duck and a rabbit (Tsal & Kolbet, 1985), the squirrel/swan figure of G. H. Fischer, "My Wife and Mother-in-law" (Boring, 1930), "My Husband and Father-in-law" (Botwinick, 1961), an instance of the images used in Fisher's hysteresis experiments (Fisher, 1967), an ambiguous image showing either a couple or a rose, and, finally, an image showing a hand and a dancer. Included, but not used for analyses because all subjects reported the same initial percept of the ambiguous stimulus versions, was an image showing either a fist or a mother with her child, and a two-interpretation version of the Fisher family (Fisher, 1968b). The complete set of stimuli is presented in Figure S2.1.

The stimulus presentations were interleaved with 36 black and white filler images showing animals and everyday objects. Each subject saw all 72 stimuli during the course of the experiment.

Apparatus

Eye tracking data were recorded using an Eyelink II system (SR Research Ltd., Mississauga, Ontario, Canada). It is capable of tracking both eyes; however, only the eye that gave a lower validation error after calibration was recorded at 500 Hz. No headrest was used. Stimulus presentation and response logging were programmed in python. For stimulus presentation, we used a 30-inch Apple Cinema Display (Apple Inc., Cupertino, CA, USA) with a native resolution of 2560x1600px and an average response time of 14 ms. The stimuli, which were scaled to a size of 1000x1000px, were presented centrally and in front of a white background. The distance to the screen was 60 cm such that the stimuli covered approximately 23.8° of the subject's visual field.

Task and Procedure

Subjects were individually tested in a dimly lit eye-tracking laboratory. After filling out a standard demographic questionnaire, the subjects were verbally introduced to the experimental procedure and by on-screen instructions. After successfully completing the calibration procedure (defined by reaching a validation error below 0.3°), the experiment was started. If required, the system was re-calibrated during the experiment.

Each trial started with a drift correction, requiring subjects to fixate the screen center. After manual acceptance by the experimenter, the stimulus appeared. Subjects were asked to press a response button as soon as they recognized the identity of the shown object. Following the button-press, the stimuli stayed visible for 4 more seconds. Although subjects were not explicitly informed of the potential ambiguity of the stimuli, they were asked to indicate changes in perception through additional button presses. After each stimulus had disappeared, subjects were asked to verbally report the perceived identity of the object. If multiple interpretations were reported, they were asked to give their reports in chronological order. Since the main interest of this study is an investigation of naïve, initial perceptual processing, the subjects were then asked to report whether they had ever seen the stimulus prior to the experiment.

The randomization of stimuli was accomplished as follows. Each subject saw one version of each stimulus during the first 24 trials. The presented stimuli contained four ambiguous and eight disambiguated stimulus versions, interleaved with a total of 12 fillers. The order of stimulus appearance was pseudo-randomized and counterbalanced across groups of four subjects out of which two were presented with the ambiguous version, one saw the disambiguated version A and one saw version B. This procedure was implemented in order to yield approximately the same amount of data for the ambiguous stimuli, which allowed for two distinct interpretations, as well as the unambiguous versions of the images.

Data Pre-Processing

Due to the objective of the current experiment, the data pre-processing procedure was rather restrictive. First, we discarded ambiguous-stimulus trials in which the reported percept did not match one of the two interpretations. For the unambiguous-stimuli, we excluded the trials in which the perceptual outcome was inconsistent with the intended interpretation (more than 80% of the percepts on the disambiguated stimuli were as intended, illustrating the efficacy of the stimulus manipulations). Additionally, trials in which the subjects indicated prior knowledge of the presented stimulus were excluded. Also trials that were either interrupted by the experimenter (no response after 20 seconds) (2,4%) or whose corresponding button press was outside the range of two standard deviations around the mean, were discarded (5,2%). After these steps, a set of n = 470 trials was left for further analyses.

In order to be able to investigate fixation behavior during the time of initial percept formation in a non-oscillatory setting, only fixations prior to subjects' object perception (i.e. the awareness of the perceived object identity) were selected for further analyses. For this, the subject's button press marks the upper limit of the time window of interest, because it also includes response preparation and execution in addition to the perceptual process. To exclude these response-related components, we identified the individual minimum reaction time (RT) across all trials for each subject and subtracted these minimum RTs from the recorded button press time points. Only fixations starting prior to this RT correction were used in further analyses. This method is quite conservative, as the correction estimate includes both, perceptual processing and the time required for the motor response of the shortest trial. Using this method, the fixation dataset was reduced by 28.6% (average minimum reaction time across subjects was 645 ms). As a control, we applied a second cleaning procedure based on changes in pupil dilation. As previously demonstrated in the literature (Einhäuser et al., 2008; Hupé et al., 2009), the average pupil diameter significantly deviates from baseline prior to the perceptual reports. Using the pupil dilation method of fixation selection, for which dilation changes upon initial object recognition were compared to data collected in a follow-up experiment in which the subjects were asked to freely push a button without visual stimulation (see Supplemental Analyses: Pupil Dilation and Figure S2.3 for more details), the estimate of the average time ascribed to the motor-response was 528 ms prior to the button-press. Using this method, 26.1% of the data are discarded, rendering it less conservative than the RT-based approach. Because of this and because a subject individual procedure is clearly preferable to an experiment-wide cut-off, the following analyses were based on the RT method.

Finally, on the level of individual fixations, single data points that had no overlap with any others in a range of one degree of visual angle were treated as noise (2.1% of the fixations).

2.1.3 Experiment 1: Analyses

The following analyses are based entirely on data collected during the subjects' first encounter with the experimental stimuli. As basis for analyzing viewing behavior, the recorded fixation data were first transformed into fixation density maps (FDM). These maps are created by first calculating 2D histograms of fixations across the stimuli, followed by a convolution with a Gaussian Kernel equivalent to 1° of visual field (FWHM = 42 pixels). The resulting maps were smoothed and normalized to a sum of one. For every set of stimuli (containing the ambiguous an the two disambiguated stimuli), FDMs were created for: ambiguous stimulus-percept A, ambiguous stimulus-percept B, disambiguated stimulus version A-percept A, disambiguated stimulus version B-percept B.

Comparing Viewing Behavior in Different Conditions

As a first analysis step, we compared the FDM's from the two disambiguated conditions against each other as well as the two perceptual conditions of the ambiguous stimuli. For this, we used a symmetric extension of the Kulback-Leibler (KL) divergence as a difference metric:

$$D_{KL}(P||Q) = \sum_i P(i) * log(\frac{P(i)}{Q(i)}) \tag{2.1}$$

$$D_{KLSymmetric}(P, Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{2.2}$$

To assess statistical significance of the found differences between FDMs we applied a separate bootstrapping analysis for each of the twelve stimulus sets. Using KL divergence as the test statistic, all subjects belonging to the two conditions to be compared were first pooled into one combined set. Resampling was then performed on the level of subjects, as resampling of individual fixations would violate the independence assumption of the bootstrapping analysis. In detail, two new sets of subject-data were randomly drawn with replacement from the overall set. It was ensured that the novel sets were identical in size, compared to the original ones. The resulting data was then used to calculate two new fixation density maps, for which the KL divergence was computed. The repetition of this procedure for 5000 times then leads to a distribution of KL divergences. This distribution describes the divergences that can be expected by chance, given the data. It can therefore be used as statistical distribution to which the original KL divergence can be compared. If the KL value of the original data falls into the highest 5% of values in this distribution, the null hypothesis of equal distributions can be rejected. To analyze statistical significance on the group level including all tested stimuli, the distribution of calculated p-values was then tested for uniformity (H_0). If the FDMs from two conditions do not differ across all stimulus sets, a uniform distribution of p-values would be expected.

Predicting Perceptual Outcomes from Single Fixations

To assess whether it is possible to predict the later perceptual outcome of our subjects based on single fixations made prior to recognition, we trained stimulus-individual Support Vector Machines using the SVMlight implementation (Joachims, 1999). For each ambiguous stimulus, the raw fixation coordinates prior to recognition were used as input space. Prediction performance was evaluated with a leave-one-subject-out cross validation, i.e. the individual fixations of each subject were once excluded from training and used as test set for the classifier. Prediction performance was then assessed based on the average accuracy gained from classifying single fixations of the test subject. Averaged across subjects, we then yield the stimulus-individual prediction performance. Finally, the grand total predictability of the perception of the subjects based on singular fixations made prior to the actual recognition is obtained via subsequent averaging across the stimulus performances.

Alignment of Viewing Behavior on Ambiguous & Unambiguous Stimuli

In order to examine whether equal perceptual outcomes on the ambiguous and the corresponding unambiguous stimuli would be preceded by similar viewing behavior, a similarity index δ was defined. It is positive if the differences in viewing behavior between the ambiguous stimuli with different percepts are in the same direction as the differences on the unambiguous stimuli. δ is computed as follows. First, a difference map (D) is created for each stimulus set by subtracting the two unambiguous fixation density maps from each other. Then, the cosines of the angles between this difference map and the fixation density maps of all four conditions, one for each unambiguous stimulus ($uFDM_{A/B}$)

and one for each possible percept on the ambiguous stimulus ($aFDM_{A/B}$), are calculated. The final similarity index (δ), is then defined as the quotient between the difference of ambiguous cosines and the difference between the corresponding unambiguous cosines:

$$cos(D, FDM) = \frac{D \cdot FDM}{||D||_2 \; ||FDM||_2} \tag{2.3}$$

$$\delta = \frac{cos(D, aFDM_A) - cos(D, aFDM_B)}{cos(D, uFDM_A) - cos(D, uFDM_B)} \tag{2.4}$$

Eq. (2.4) therefore expresses the differences between ambiguous conditions as fraction of the maximal possible difference, as estimated from the unambiguous reference FDMs. For the statistical analysis of the similarity index, a randomization analysis was performed. The approach was similar to the described bootstrapping in case of the KL divergence, but resampling was done without replacement and the index δ was used as a test statistic. As before, the original δ-value was compared to the distribution of resampled δ-values to obtain the p-value.

Temporal Analysis

Based on the similarity index δ, it is possible to analyze the temporal development of viewing behavior on the ambiguous stimuli with regard to different perceptual outcomes. For this, we first aligned the fixation data to the button press. Starting at the button press, we then shifted a time window of 200 ms backwards over the fixation data of each image set. Based on the fixations that fell into this time window, the similarity index was calculated. The final curve was calculated by averaging the δ-indices of all image sets.

Subject Level Analysis of Individually Collected Evidence

The previously described analyses were performed across subjects and therefore on the level of stimuli. However, by using the difference maps (D) created from the FDMs of the unambiguous stimuli as reference, it is also possible to investigate subject individual scan paths on the ambiguous images to check whether the subjects collected more evidence consistent with their later percept than for the alternative one. For this, the difference map is interpreted as depicting evidence for the different perceptual outcomes. Positive values in the difference maps represent evidence for percept A, whereas negative values correspond to evidence associated with percept B. To obtain an estimate of the individually collected evidence of a subject, the recorded scan-path on the corresponding ambiguous stimulus is projected onto the difference map. For each fixation along the trajectory, the collected evidence corresponds to the average values of the difference map within a circular region of two degrees of visual angle. The overall evidence of a scan path is then defined as the sum of evidence across all fixations. This value is positive, if the subject collected more evidence for interpretation A, whereas it is negative, if more evidence was collected for B.

To statistically evaluate whether subjects collected more evidence for their actual rather than the alternative interpretation of the ambiguous image, the individually collected evidence was sorted into two sets, according to the initial perception of the subjects. These sets were then tested with a Mann-Whitney U test.

2.1.4 Experiment 1: Results

Because our data pre-processing procedure is related to the reaction times of the subjects, these were analyzed first. We found that the unambiguous stimuli were significantly faster recognized than on ambiguous ones (median $RT_{unambiguous}$ = 1.51s ±0.15 SEM, $RT_{ambiguous}$ = 1.78s ±0.09 SEM; SEM will be used for each ± hereafter, Mann Whitney U-test Z = 3.46, p<0.001). Moreover, the average minimum reaction time across subjects was found to be 645 ms. After performing the described pre-processing procedure, excluding fixations that were made during the time window associated with the motor response, on average 3 fixations (distribution median) were left for further analyses in the unambiguous case and 4 fixations for the ambiguous stimuli.

Viewing behavior prior to object awareness

As a first analysis of overt visual attention during the time in which no conscious recognition has yet occurred, we assessed whether differences in fixation patterns existed on sets of two unambiguous stimuli corresponding to the two interpretations of an ambiguous one. For this, we computed fixation density maps and used the described symmetric extension of the Kullback-Leibler (KL) divergence as difference metric. Bootstrapping revealed that the viewing behavior on the different unambiguous versions of the stimuli differed strongly and significantly across the stimulus set (see Figure 2.2a for an example). The p-values of the 10 stimulus sets obtained via bootstrapping are nonuniformly distributed, contrary to the tested null hypothesis of similar viewing behavior, and right-skewed ($\chi^2 = 10$, $p<0.01$; Figure 2.3). Nine out of the ten analyzed stimulus sets are individually significant ($p<0.01$, Bootstrapping). In view of the subtle changes in the images, the robust differences in viewing behavior are remarkable. Furthermore, they form an important reference for the differences in viewing behavior that are to be investigated on the ambiguous stimuli.

The most important aspect with regard to the two examined hypotheses, *action follows perception* and *action precedes perception*, is the viewing behavior on the ambiguous stimuli. For this comparison, we grouped the fixation data according to the subjects' perceptual decisions. A comparison of the corresponding fixation density maps revealed that even in the case of an identical stimulus initial viewing behavior recorded prior to object awareness differed significantly for different perceptual outcomes. The distribution of p-values from the KL divergence bootstrapping is nonuniform and right-skewed ($\chi^2 = 6.4$, $p<0.025$). Two stimulus sets were individually significant ($p<0.01$) (Figure 2.2a shows an example, the distribution of p-values can be seen in Figure 2.3). This is a particularly strong case, because only the perceptual formation process and the sampling of stimulus properties, but no differences in the presented stimuli can be associated with the found differences in overt attention. This finding implies that it should be possible to predict the perceptual outcome of our subjects based on their overt visual attention recorded prior conscious recognition. Put differently, it should be possible to predict the subjects' perception based on data that is recorded at a time in which the subjects themselves are not yet aware of their later percept. Indeed, after training radial-basis Support-Vector-Machines on the ambiguous fixations using a leave-one-out cross validation scheme, it was possible to predict the subjective percept with an average accuracy of 70% (±4%). Figure 2.4 shows the prediction accuracy for the individual stimuli.

Following the individual analysis of conditions based either on unambiguous or ambiguous stimuli, we assessed whether the viewing behavior on the ambiguous stimuli was aligned with that on the unambiguous ones upon similar perceptual outcomes. This step is crucial because it implies that the found differences on the ambiguous stimuli are in fact percept-related and not incidental. To assess the similarity, an index δ was defined by projecting fixations on the ambiguous stimuli onto the axis spanned by the differences of viewing behavior on the unambiguous stimuli (see Methods for more details). The index δ is positive if the differences in viewing behavior between the ambiguous stimuli with different percepts are in the same direction as the differences on the unambiguous stimuli. In all stimulus sets, the index was found to be positive ($\bar{\delta} = 0.26 \pm 0.05$; see Figure 2.2b), indicating that overt visual attention on the

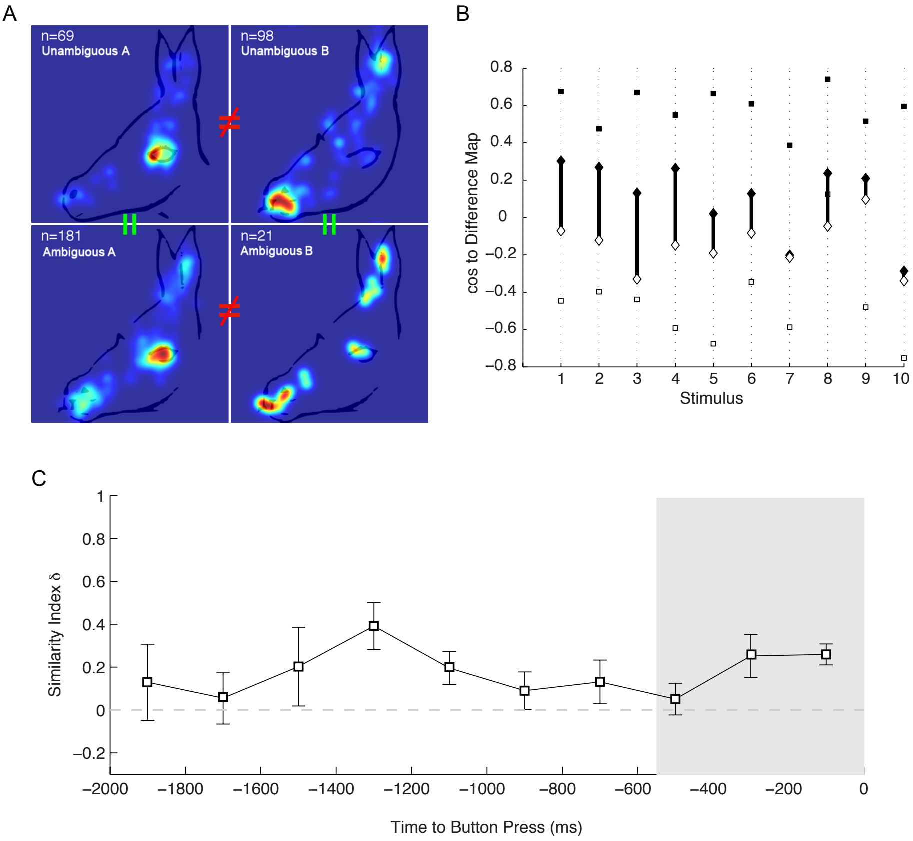

Figure 2.2: Viewing behavior prior to awareness. (a) Examples of viewing behavior prior to object awareness on the ambiguous and unambiguous stimuli with corresponding percepts. There are significant differences between the groups with different percepts (as indicated by the KL divergence analysis), and the differences in the viewing behavior on the ambiguous and unambiguous stimuli are aligned with identical percepts (as shown by the similarity index δ). The shown fixation patterns correspond to the fifth-largest index value out of the ten examined stimulus sets. **(b)** The cosine values underlying the similarity index calculation for the individual FDMs. Filled symbols represent percept A, the empty ones percept B. Squares denote cosines calculated from the unambiguous FDMs; diamonds indicate values calculated from the ambiguous FDMs. Image 1 corresponds to the example shown in (**a**). **(c)** The time-analysis showing the index-peak at about 1330 ms before the button press. Error bars are SEM. The shaded area marks the time during which data would be discarded according to the pupil analysis.

ambiguous stimuli was aligned with fixations on the corresponding unambiguous ones prior to conscious recognition of the shown object. A randomization analysis, analogously to the above Bootstrapping analyses, confirmed the statistical significance of the effect (see Figure 2.3; $\chi^2 = 6.4$, $p<0.025$; five stimulus sets individually significant with $p<0.05$).

Our previous analyses were based on collapsed data taken from the complete period prior to conscious recognition. To analyze the temporal dynamics of overt visual attention in more detail, we performed a sliding window analysis. After aligning the trials to the button press and selecting data according to the current time-window, the mean similarity index δ exhibits a clear peak at about 1300 ms prior to button press (Figure 2.2c). At this time, which is largely before the later report of conscious recognition, the fixation behavior on the ambiguous stimuli is most similar to the corresponding unambiguous ones, and therefore most different for different percepts on the same ambiguous stimuli.

To approach the differences in viewing behavior on a subject instead of stimulus basis, the difference maps can be interpreted as depicting evidence for the different percepts. Now, each individual scan-path on an ambiguous stimulus represents subsequent collecting of evidence for one or the other percept, depending on whether a positive or negative region in the difference map is fixated. This analysis revealed that the evidence collected by subjects prior to recognition significantly differs for subjects with different percepts (median $\overline{evd_{PerceptA}} = 0.026 \pm 0.007$, $\overline{evd_{PerceptB}} = -0.025 \pm 0.007$, Mann Whitney U-test $Z = 6.23$, $p = 4.7 * 10^{-10}$). As the sign of these two numbers shows, subjects collected more evidence for their actual than for the competing, but not perceived, percept.

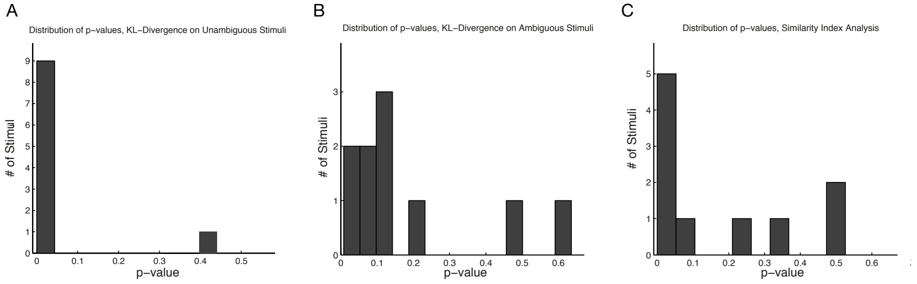

Figure 2.3: Bootstrapping Distributions. Shown are the distributions of p-values for **(a)** KL-Divergence on the unambiguous stimuli, **(b)** KL-Divergence on Ambiguous Stimuli with different percepts, **(c)** the similarity index δ. All of them are nonuniform and right-skewed.

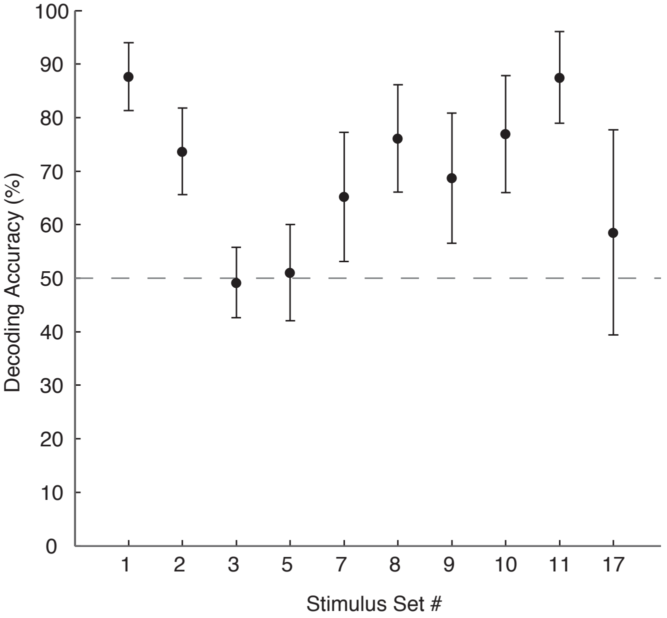

Figure 2.4: SVM performance. Mean Support Vector Machine prediction accuracy for the correct percepts is shown for the ten tested image sets. Accuracy over fixations of one subject was calculated using SVM's trained on the remaining fixation data of the ambiguous stimuli (leave-one-out cross validation). Errorbars depict SEM.

Viewing behavior after a formed percept

By showing that significant and percept-aligned differences in pre-conscious viewing behavior exist on ambiguous stimuli, our results strongly favor the *action precedes perception* hypothesis. However, the results presented so far do not exclude the possibility that also object awareness had an effect on the subsequent overt visual attention. Since the stimuli were shown for four more seconds after the first button press of the subjects, it is possible to test this issue by comparing the viewing behavior before and after the reaction-time corrected perceptual report.

First, we tested only subjects, which did not report a change in perception during the 4-second period. Contrary to the predictions of the *action follows perception* hypothesis, we did not find evidence for an increase in probability to fixate characteristic local features as determined by the similarity index δ (t (9) = 2.1, p>0.05; normality and homoscedasticity verified by Lilliefors tests (p>0.05), and Bartlett's test (p>0.05)). Following this, we analyzed the data of the subjects who identified a second interpretation and therefore a switched perception. For this type of event, the *action precedes perception* hypothesis predicts a drop in similarity index prior to the perceptual switch, since subjects are expected to sample more evidence for the competing percept prior to becoming aware of the alternative interpretation. This test requires subjects who switch during ambiguous trials from an initial percept A to percept B and vice versa for each stimulus set. Despite the large number of subjects, the required data existed only for two ambiguous stimuli (stimulus 2 and 3) and therefore the sample size is not sufficient for detailed statistical analysis. However, we found a correct tendency, as the subjects with a perceptual switch exhibit a smaller index than the non-switching ones during the time between the two button presses ($median_{image2,3}$ $\bar{\delta}_{noswitch} = 0.27$, $\bar{\delta}_{switch} = 0.01$).

Given the results of experiment 1, which illustrated significant differences in overt visual attention prior to conscious recognition, we investigated in a second experiment whether the found correlative relationship also has causal capacities. The experimental reasoning was that if the initially attended information has a causal effect, it should be possible to manipulate the perceptual outcome by means of changing the initial fixation of the subjects. Leaving everything else equal to the first study, we therefore manipulated the position of the fixation cross shown prior to stimulus presentation, and tested whether this would result in a changed perception of our subjects.

2.1.5 Experiment 2: Materials and Methods

Participants

Twenty-six subjects (19 female) took part in the second experiment. None of them had participated in experiment 1. As before, the subjects received either 5Eur or course credit for their participation. Six additional subjects took part in a pre-experiment to assess the altered positioning of the fixation cross.

Stimuli

In the second experiment we used the 10 ambiguous stimuli also included in the analyses of experiment 1 and ten of the previously used fillers.

Apparatus

The used apparatus and experimental setup was identical to experiment 1.

Task and Procedure

The experimental procedure of the second experiment was largely identical to the first. However, to check for the effect of visual attention on the later percept, we introduced an experimental manipulation in which the position of the initial fixation cross, shown before stimulus onset, was altered. For this, the new starting positions were shifted to locations that were expected to be consistent with either one of the two percepts. To determine these positions, six additional subjects that were not involved in any of the eye-tracking experiments, were asked to freely mark the regions of the ambiguous stimuli that "clearly favor one or the other percept". Contrary to the main experiments, these subjects were informed about the two percepts beforehand in order to verify that each subject was able to perceive both versions. From the marked regions of these subjects, clusters with more than 80% congruence across subjects were selected and the cluster centroids were calculated. Following this, a straight line was drawn through both centroids. The new initial fixation points were positioned on this line at 1-1.5 degree of visual angle away from the centroids towards the image borders (see Figure S2.2).

Importantly, the introduced manipulation only changed the subjects' initial locus of visual attention as the fixation cross disappeared with stimulus onset and subjects were then allowed to freely move their eyes. Compared to earlier studies, which either forced the subjects' view onto a specified position during the complete trial (Kawabata & Mori, 1992), or which directly manipulated the stimuli in order to bias the perception towards one or the other outcome (Pomplun et al., 1996), this manipulation is very subtle and allows for very natural viewing behavior on the stimuli.

The starting position was balanced across subjects. Half of the subjects started at the location favoring interpretation A and the other half at position relevant for B. An experimental session lasted approximately 30 minutes.

Data Pre-Processing

Again, we excluded trials for which the subjects had indicated prior knowledge of the presented stimulus. Moreover, trials for which the reaction times fell outside of a 2 standard deviation range around the mean were excluded. On stimulus level, we excluded stimulus 6 (old/young woman), as all of the recorded subjects reported prior knowledge. Finally, we excluded stimulus 8 (man/woman taken from Fisher's hysteresis experiments) as an outlier because the results of the manipulation were more than two standard deviations away from the group average.

2.1.6 Experiment 2: Results

We statistically assessed the efficacy of the manipulation based on two methods. First, we performed a χ^2 cross-tab test and found a significant dependence of the reported percepts on the initial fixation position ($\chi^2 = 5.74$, p = 0.006). On average, 60.3% of the percepts were consistent with the bias induced by the starting position. Importantly, any perceptual biases in the ambiguous stimuli during the first experiment cannot explain this result, as their effect equally affects the result positively for one percept, but negatively for the other. Still, to explicitly account for the found biases of the stimuli in the original experiment, we performed a second analysis. For this, we first calculated the percentage of subjects perceiving A and B for every ambiguous stimulus in experiment 1. Then, we calculated the percentage of subjects who correctly perceived A (and B) in condition A (and B) during experiment 2 and calculated the percentage gained through the experimental manipulation on each stimulus by averaging the subtracted the percentages of experiment 1 from the percentages of experiment 2. As an example, if for a given stimulus in the original experiment A was perceived in 60% and B in 40% of the cases, and in experiment 2, 70% of the subjects perceived A in condition A and 60% perceived B in condition B, then the average percentage gain for this stimulus would be 15%. Once this was calculated for every stimulus, we checked the resulting distribution for a deviation from zero using a t-test (the normality assumption was verified with a Lilliefors test). The test showed that the introduced changes in the initial fixation positions had a significant effect on the perceptual outcome of the subjects (t = 3.45; p = 0.01).

This robust effect, which is in line with earlier studies emphasizing the importance of local features in fixed eye-position setups (Kawabata & Mori, 1992; Georgiades & Harris, 1997; Tsal & Kolbet, 1985; Long & Toppino, 2004), is quite remarkable because the subjects were allowed to freely move their eyes as soon as the stimulus appeared.

2.1.7 General Discussion

The current work aimed at a clarification of the interplay between overt visual attention and object perception. We approached this problem by investigating patterns of viewing behavior preceding the conscious recognition of ambiguous stimuli and show that different percepts (and perceptual switches) are preceded by significant and percept-aligned differences in viewing behavior. In line with this, we demonstrated that eye-movements recorded prior to the conscious recognition are a good predictor for the later perceptual outcome, and that subjects collect more evidence for the later perceived object identity than for the alternative one. In experiment 2, we extended the correlative results from experiment 1 by showing that manipulations of the initially attended positions significantly influence the later perceptual outcome. This finding further clarifies the role of overt visual attention by providing evidence for a causal influence on perception. All of these results are completely compatible with the view that the object awareness follows overt visual attention (*action precedes perception* hypothesis). However, as the results do not exclude the possibility that also the awareness of object identity has an impact on the subsequent overt visual attention, we additionally compared the viewing behavior preceding and following object awareness. No significant difference in the similarity index could be found. Directly related to this, the interplay of hippocampus-dependent memory, in form of awareness of image manipulations, and patterns of overt visual attention were recently investigated (Smith et al., 2006; Smith & Squire, 2008). The authors conclude, that the awareness of image manipulations was reflected in subsequent eye-movements. However, our current results suggest that, in fact, overt visual attention preceded the awareness of the stimulus manipulation.

Our results extend recent experimental and theoretical evidence pointing into the direction of neurally distinct mechanisms for visual awareness and attention (Wyart & Tallon-Baudry, 2008; van Gaal & Fahrenfort, 2008; Koch & Tsuchiya, 2007; Lamme, 2003). For instance, using faint stimuli that reached perceptual awareness in only about 50% of the trials, Wyart & Tallon-Baudry (2008) showed that visual awareness correlated with an increase in mid-frequency gamma-band activity at the contralateral visual cortex, whereas covert visual attention modulated high-frequency gamma-band activity in the same region. These results suggest that the neural correlates of the two processes are in fact distinct. In addition to this, there is electrophysiological evidence suggesting that processes of attentional selection precede visual awareness. Fernandez-Duque et al. (2003) investigated event related potentials (ERPs) related to visual attention and aware vs. unaware changes in a flicker paradigm (Simons & Rensink, 2005). Their data was grouped based on the subjects' awareness of changes, either aware or unaware, in subsequently presented scenes. The results showed early, attention-related components over frontal and parietal sites, followed by a late component that was related to awareness of visual change. The latter component exhibited distinct topography, by being broadly distributed with its center in medial centro-parietal regions. The described attentional regions broadly correspond to earlier results of Beck et al. (2001) and also of Huettel et al. (2001). Comparing fMRI responses in a similar flicker paradigm, they associated the awareness of change with enhanced BOLD responses in parietal and right dorsolateral prefrontal cortex as well as extrastriate visual cortex. Similarly, the data presented in (Koivisto & Revonsuo, 2007; Koivisto et al., 2009; Brascamp et al., 2010) suggests based on behavioral and electrophysiological measurements that attention and consciousness are initially independent, whereas later, higher-level visual awareness is strongly depended on focused attention. Taken together, there is converging evidence for the view that visual awareness and visual attention rely on two distinct neural mechanisms and that patterns of activity correlated with visual attention precede effect of visual awareness. Our results now clarify the interaction of both phenomena on a behavioral level by showing that patterns of overt visual attention have a causal impact on the resulting object awareness.

Converging evidence for our results comes from previous studies investigating the effects of eye-movements on perceptual illusions (Troncoso et al., 2008; Martinez-Conde et al., 2006; Otero-Millan et al., 2008) and on perceptual oscillations (Einhäuser et al., 2004; Kawabata & Mori, 1992; Pomplun et al., 1996; Toppino, 2003). For the latter, informed subjects and prolonged stimulus presentations were used in order to induce regular perceptual oscillations. Although this approach has the clear advantage of comparably easy data collection, it severely complicates an analysis of the direction of causality, as all recorded eye-movements precede but also follow perceptual events. Also, in addition to being a rather artificial setting, the study of perceptual oscillations has the problem that the data is collected from non-naive subjects that become 'stimulus-experts' due to the long presentation time and because they are typically presented with both interpretations prior to the experimental trial. In the current set of experiments, we overcome these limitations by only analyzing data recorded prior to the initial perception of the object's identity and by excluding all subjects with prior knowledge of the stimuli.

The most important difference of our approach is that we investigate overt visual attention occurring prior to the first conscious perception of the subjects (perceptual formation) whereas the data recorded from perceptual oscillations is always accompanied with active perceptual interpretations and perceptual switches. The same argument holds for the previously reported results of perceptual events on pupil dilation, which were always based on the recordings of perceptual switches (Einhäuser et al., 2008). Because of this, it was previously unclear whether the neuronal mechanisms of pupil dilation involving norepinephrine release (see below) followed or lead to the perceptual switch. In the current experiments, we show pupil dilation effects based on the initial perceptual interpretation following a time in which the subjects were not yet aware of any object identity. With regard to the underlying mechanism, Einhäuser et al. (2008) argue that pupil dilation recorded around perceptual switches reflects norepinephrine release from locus coeruleus (LC). LC has been implicated in regulating the balance between exploitation and exploration within the sensory domain and to consolidate perceptual decisions (Aston-Jones & Cohen, 2005; Devilbiss et al., 2006). This exploitation-exploration model is very well in line with our results.

Similar to our disambiguated stimuli, albeit again based on data recorded from perceptual oscillations, Pomplun et al. (1996) showed that changes of ambiguous stimuli can result in perceptual biases, leading the subjects to perceive one interpretation significantly more often than the other. Kawabata & Mori (1992) provided evidence in line with the results of experiment 2 by showing that the perceptual outcome of the subjects can be altered if forced onto one stimulus position. In our case, however, the experimental manipulation is much more subtle, because the attended position is only altered prior to the actual stimulus presentation and not during the complete trial.

An open research question is on what basis the targets of eye-movements are selected during the initial phase in which the object is not yet recognized. Possible mechanisms include bottom-up visual salience (either mediated via low-level features and feature-combinations represented in V1 (Li, 2002; Troncoso et al., 2005, 2007) or determined by a saliency-based approach combining multiple feature maps (Itti & Koch, 2000; Koch & Ullman, 1985b)) and high-level, hypothesis driven attention working in a top-down manner (Renninger et al., 2005; Triesch et al., 2003). In either case, it might be of special importance to differentiate local stimulus properties from the effects of stimulus context. The gist of a scene can provide a strong cue for the object identity and is therefore a promising candidate for future research in this direction. Please note that the current findings do not argue against the task-dependent view of overt visual attention (Triesch et al., 2003; Rothkopf et al., 2007), because attention can be guided towards task-relevant objects without requiring constant and conscious awareness of their identities. Our results are compatible with a constructive view of perception (Chastain & Burnham, 1975) and provide new evidence for the impact of eye movements during the formation object awareness (Hafed & Krauzlis, 2006).

2.1.8 Acknowledgements

The authors would like to thank Sabine König and Paula Vieweg for their assistance during data acquisition, and Alper Acik for valuable discussions during the initial phase of the study. We also thank Konrad Körding, Merav Ahissar, David Leopold, Christoph Kayser, Li Zhaoping and Wolfgang Einhäuser for comments on an earlier draft of this manuscript as well as Kevan A.C. Martin and Wolf Singer for valuable advice.

2.1.9 Supporting Information

Supplemental Analyses: Pupil Dilation

The diameter of the pupil was previously shown to increase at the time of cognitive events (Richer et al., 1983), as for instance switches in perception (Einhäuser et al., 2008; Hupé et al., 2009). However, it has been argued that part of the pupil response is due to response execution (Richer et al., 1983; Hupé et al., 2009) and therefore not related to perception per se. It is argued that especially the peak of the pupil dilation is modulated by this effect. Here, we are interested in the onset of the pupil dilation in relation to the report of conscious recognition of the object identity. Compared to the button press, which of course includes the time required for the motor response, the start of pupil dilation is a better estimate of the completed recognition process because it is either related to the perceptual event or to the response preparation signaling object awareness. Both processes are closer to the time point of object recognition than a pure button press.

For the current pupil analysis, the data from experiment 1 was related to data collected in a control task (conducted directly after experiment 2). Subjects were presented 6 pink noise images and 2 blank screens, resulting in a total of 8 trials. The subjects' task was to press a button as often and whenever they wanted. They were allowed to freely move their eyes during this task. Since the stimuli do not depict any object or meaningful structure, these trials did not include any object awareness, but only a self paced motor preparation and response. To assess differences in perceptual and motor-related processes, the pupil size around these simple motor actions were then compared to the pupil size around motor report of percept formation in experiment 1.

Pupil size was recorded by the Eyelink II system with 500 Hz. In the original experiment, pupil traces from all three blocks of experiment 1 were extracted from stimulation onset to 10 seconds after onset. All trials interrupted by the experimenter and trials in which the button press occurred earlier than 500 ms or later than 10 seconds after stimulus onset were discarded. Additionally, all trials in which the interval between two button presses was smaller than 500 ms were removed. Pupil traces were trial wise z-transformed, periods of 100ms before and after automatically detected blinks were excluded together with manually detected artifacts. The removed data were replaced via linear interpolation. The traces were then aligned to the button press and data within ± 3s around the button press were averaged over all trials (n=790). From the control task, only the last button presses within each trial were considered to allow for maximal delay to the last perceptual change, i.e. the switch from drift correction to pink noise or blank screen. Identical to the experimental data, pupil responses were z-transformed and artifacts were removed. As before, pupil traces from the time window of ± 3s around the last button press were extracted, aligned to the button press and averaged (n=147).

Pupil diameter is plotted for both conditions in Supplemental Figure S2.3. First, we compared the pupil size maxima of individual trials in the two conditions and found that they were significantly higher in the object recognition condition (mean=0.6) as compared to the control condition (mean=0.38; Wilcoxon's rank sum test; $p<0.001$).

In order to estimate the point in time at which the two curves significantly deviate from each other, we tested the pupil diameter traces at each 2ms time samples within the 3-second period around the button press. The analysis revealed a significantly larger pupil diameter in the recognition condition from 528 ms prior to 3000 ms after to the button press (t-tests, $p<0.05$ FDR corrected; with Satterthwaite's approximation for unequal variances).

These results show that the human pupil reacts to motor actions such as button presses in the absence of visual stimulation. However, if the motor action follows a perceptual event, such as object recognition of line drawings, the pupil dilation is facilitated. This additive pupil reaction to percept formation is in line with previous results from experiments investigating perceptual switches on similar stimuli (Einhäuser et al., 2008; Hupé et al., 2009). On an aggregated level, the pupil dilation distinguishes between isolated button presses and button presses after object recognition. Since the object recognition occurs before the button press and contributes to the size of the pupil dilation, the onset of the pupil dilation can be used as a measure of the on average completed object recognition.

Supplemental Figures

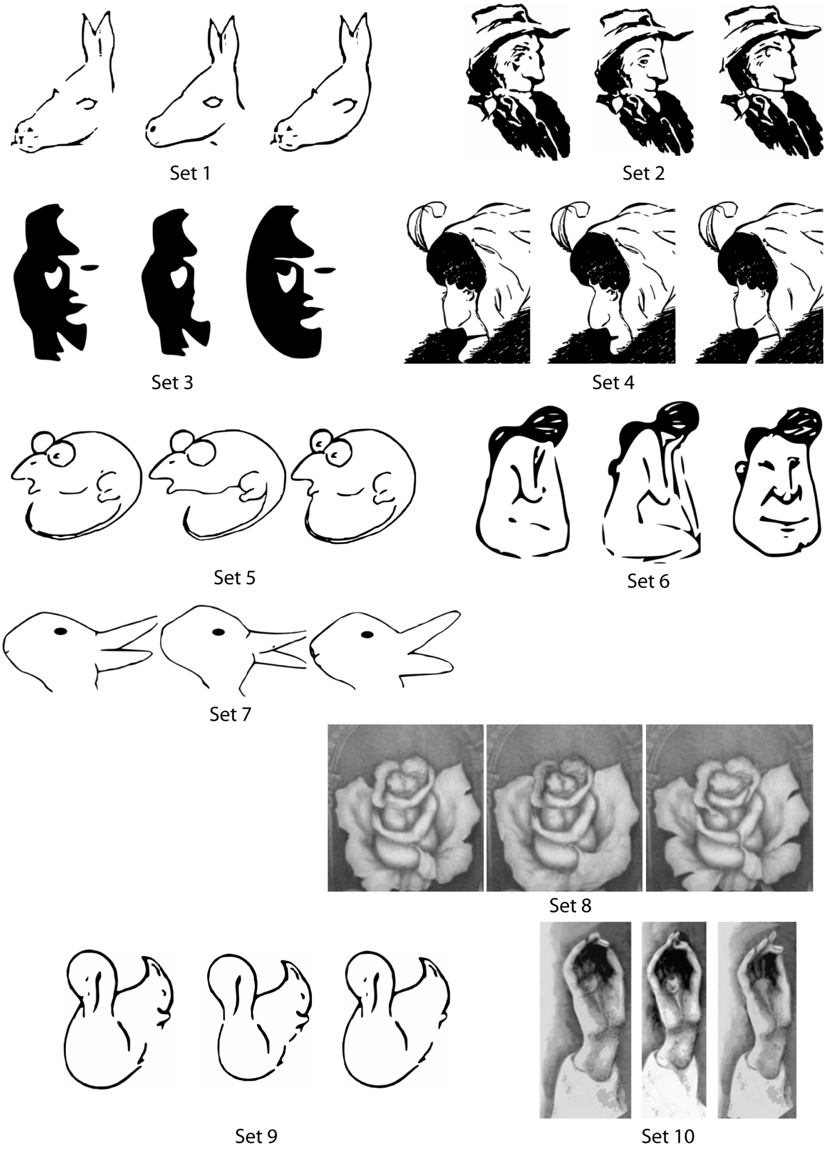

Figure S2.1: Experimental Stimuli. Shown are the ten ambiguous and disambiguated stimuli that were used for the analysis. The first column contains the ambiguous image, the second and third the respective disambiguated versions.

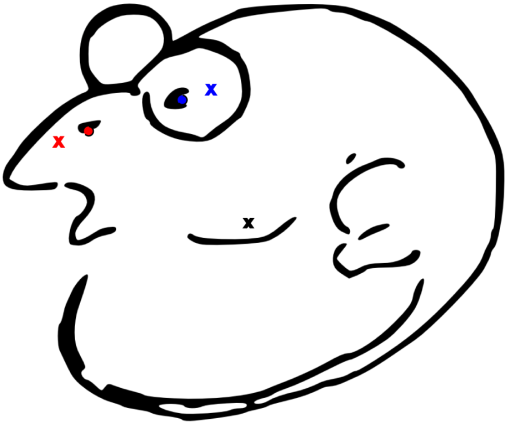

Figure S2.2: Experiment 2 Design. Shown is an example stimulus together with the calculated centroids of the 80% congruency regions (circles), as marked by a set of independent subjects. The colored crosses correspond to the shifted fixation cross positions used in experiment 1, the black cross shows the centered fixation cross used in experiment 1.

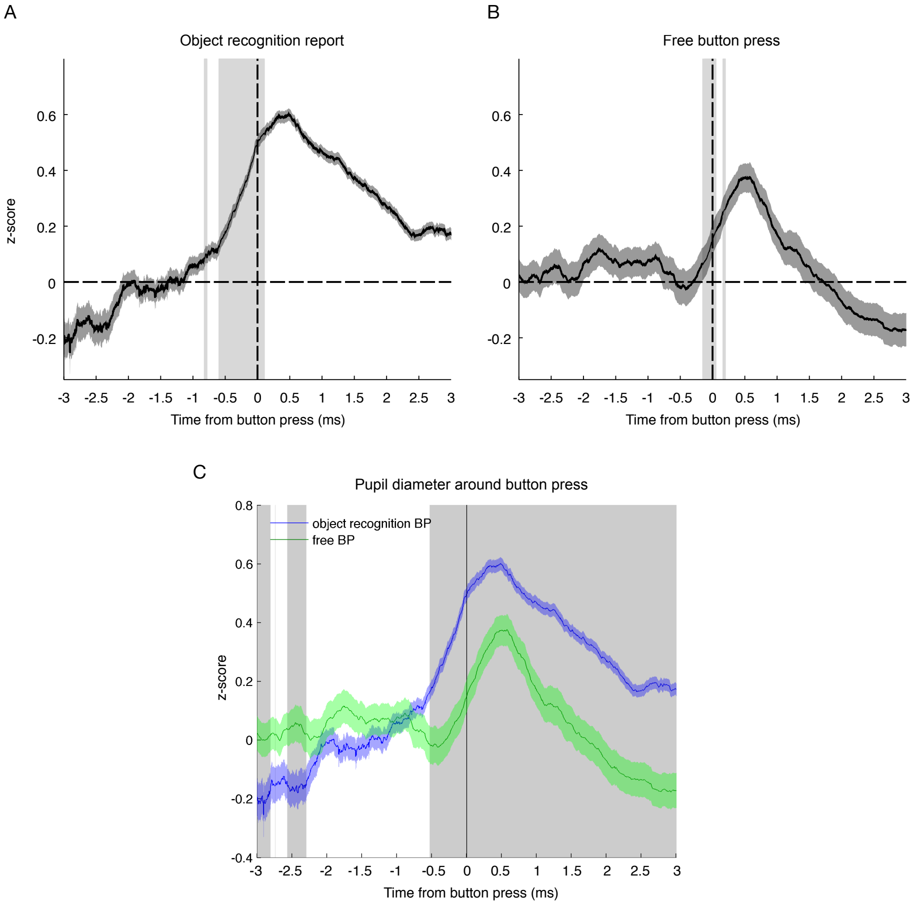

Figure S2.3: Pupil Size Analysis. The averaged pupil size z-scores from the **(a)** percept formation condition (data from experiment 1) and **(b)** the control experiment in which subjects pressed the same keyboard button whenever they wished to do so. The shaded area around the pupil diameter shows the SEM. Time periods with a significant positive slope are marked with a light grey bar. **(c)** A statistical comparison of the perceptual- and motor-task showing significant differences at 528 ms before the button press.

2.2 Effects of Contextual Information on Overt Visual Sampling Behavior

Abstract The sampling of our visual environment through saccadic eye movements is an essential function of the brain, allowing us to overcome the limited accuracy of peripheral vision. Due to the constant changes of our surroundings, however, not all aspects of a given scene can be attended and a subset of visual locations has to be selected for in-depth processing. Understanding which parts of a scene attract attention is subject to intense research and considerable progress has been made in unraveling the underlying cortical functions. In contrast to spatial aspects, however, only little is understood about temporal aspects of the sampling mechanism. At every fixation, the oculomotor system faces the decision whether to keep fixating, and thereby to increase recognition performance of the currently seen object, or whether to move to a different object at a different location - a problem that can be understood in terms of exploration and exploitation. To improve our understanding of the factors involved in this decision, we here use a gaze-contingent paradigm and investigate how scene context changes the subsequent sampling behavior preceding the recognition of ambiguous stimuli. Behaviorally, we find that context, although only presented until the first voluntary saccade, biases the perceptual outcome, reduces reaction times and increases perceptual certainty. Importantly, we find that initially presented visual context significantly increases subsequent fixation durations. These results speak in favor of an unconscious effect of perceptual certainty on visual sampling behavior, biasing the exploration-exploitation strategy towards in-depth analyses with increasing evidence for the identity of the seen object.

Section published as: Kietzmann, T.C., & König, P. (2015). Effects of Contextual Information and Stimulus Ambiguity on Overt Visual Sampling Behavior, Vision Research, 110, p.76-86. doi:10.1016/j.visres.2015.02.023

2.3 Chapter Summary

In this first experimental chapter, we investigated whether the sampling of visual information, in form of overt attention, interacts with processes devoted to recognition and perception. The key element of the experimental design was the use of ambiguous stimuli, which can be recognized in two alternative ways although the underlying stimulus is identical. This allows for a close control of low-level stimulus features. To exclude high-level influences, we furthermore focused on eye-movements preceding the initial perception of naïve participants. Using this setup, we demonstrated that patterns of eye-movements preceding the conscious recognition are predictive of the later perceptual outcome: ’action precedes perception’. In a follow-up experiment, we demonstrated the causality of the effect by showing that a subtle manipulation of the initial locus of attention has strong effects on the perceptual outcome. Having found evidence for the impact of attentional sampling on the forthcoming recognition, we then tested whether the reverse would hold, too. For this, we introduced contextual information during the initial phase of stimulus presentation and examined whether this additional source of information in the system would change the subsequent sampling behavior. Central to this experiment was the use of a gaze-contingent paradigm, which allowed the removal of the context stimulus upon the initial saccade of the participants. As a result, the subsequent stimulus display was identical and therefore comparable to the first experiment without context. Using this design, we found strong differences in the behavior of our participants. The brief presentation of context had profound effects on the perceptual outcome, it lead to shortened reaction times and increased the certainty of the perceptual decision of our participants. Most importantly, however, our data revealed prolonged fixation durations when contextual information had previously been presented, although the actual stimulus display in the time-window analyzed was identical in the contextual and original experiments. This suggests that an increase in perceptual certainty lead to a shift in the exploration-exploitation behavior in favor of a more detailed analysis at a given fixation. These results were corroborated by additional analyses demonstrating longer fixation durations for unambiguous than ambiguous stimuli and a positive relationship between the evidence at a fixation position and fixation durations. Again, these data indicate a relationship of recognition evidence and exploration strategy. Finally, although changes in fixation duration are traditionally not interpreted in terms of a changing exploration bias, we showed that this view is in line with a large body of experimental results, indicating that it might serve as a unifying explanation and as starting point for future work.

Summing up, our experiments demonstrate a bi-directional relationship between the processes of overt visual attention and recognition. Attention has a causal impact on the later recognition, and increased certainty in the recognition process alters the sampling behavior towards more in-depth analyses.

3 Visual Invariance

3.1 Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways[1]

Abstract Although the ability to recognize faces and objects from a variety of viewpoints is crucial to our everyday behavior, the underlying cortical mechanisms are not well understood. Recently, neurons in a face-selective region of the monkey temporal cortex were reported to be selective for mirror-symmetric viewing angles of faces as they were rotated in depth (Freiwald & Tsao, 2010). This property has been suggested to constitute a key computational step in achieving full view-invariance. Here, we measured functional magnetic resonance imaging activity in nine observers as they viewed upright or inverted faces presented at five different angles (-60, -30, 0, 30, and 60°). Using multivariate pattern analysis, we show that sensitivity to viewpoint mirror symmetry is widespread in the human visual system. The effect was observed in a large band of higher order visual areas, including the occipital face area, fusiform face area, lateral occipital cortex, mid fusiform, parahippocampal place area, and extending superiorly to encompass dorsal regions V3A/B and the posterior

[1] This section was published as a peer reviewed article in the Journal of Neuroscience together with Jascha Swisher, Peter König and Frank Tong. See Publication List for details.

intraparietal sulcus. In contrast, early retinotopic regions V1-hV4 failed to exhibit sensitivity to viewpoint symmetry, as their responses could be largely explained by a computational model of low-level visual similarity. Our findings suggest that selectivity for mirror- symmetric viewing angles may constitute an intermediate-level processing step shared across multiple higher order areas of the ventral and dorsal streams, setting the stage for complete viewpoint-invariant representations at subsequent levels of visual processing.

3.1.1 Introduction

People can recognize faces and objects across a wide variety of viewing conditions, despite changes in retinal position, size, and illumination. Changes in viewing angle represent a further challenge, as large rotations of a three-dimensional object can drastically alter the pattern of retinal input. Although people can readily recognize objects from different viewpoints, neurophysiological studies have found that the vast majority of object-selective neurons in the monkey inferotemporal cortex exhibit viewpoint-specific rather than viewpoint-invariant tuning (Perrett et al., 1991; Logothetis et al., 1995). These findings have led to the proposal that object recognition relies on multiple view-specific representations, and that the combined input of several view-specific neurons might be a necessary precursor to obtain fully view-invariant object selectivity (Bülthoff & Edelman, 1992; Logothetis et al., 1994; Perrett et al., 1998; Ullman, 1998).

Recently however, Freiwald & Tsao (2010) reported that neurons in an intermediate region of the monkey face-processing network exhibited the peculiar property of being selective to mirror-symmetric viewing angles of faces. For instance, neurons that responded preferentially to the view of a head rotated 60° to the left were also likely to respond to a rightward rotation of 60°, but not to intermediate near-frontal views. This pattern of viewpoint symmetry can be distinguished from previous neurophysiolological reports of exclusive selectivity for a single viewpoint (Perrett et al., 1991; Logothetis et al., 1995), and has been suggested to represent a key computational step towards achieving full viewpoint invariance. However, these single-unit recordings were restricted to focal regions of interest; thus, it is presently unknown whether viewpoint symmetry is a specific property of a single region in the face-processing network or whether it might be found in other visual or category-selective areas, including regions that prefer non-face stimuli such as objects and scenes.

In the present study, we investigated whether selectivity for mirror-symmetric viewing angles might also be found in the human visual system. We monitored cortical activity using functional magnetic resonance imaging (fMRI) while subjects viewed images of upright or inverted faces, taken from five different viewpoints (Figure 3.1). Using multivariate pattern analysis (Haynes & Rees, 2006; Norman et al., 2006; Tong & Pratte, 2010), we then tested whether activity patterns were more similar between mirror-symmetric viewing conditions (e.g., -60° and +60°) than between viewing angles that lacked this relationship (e.g., -60° and 0°). If so, this would imply cortical selectivity for viewpoint symmetry similar to that recently found in the monkey Freiwald & Tsao (2010). However, a key difference was that our pattern analytic approach did not require a region to be selective for faces or specific facial identities to exhibit mirror-symmetric selectivity.

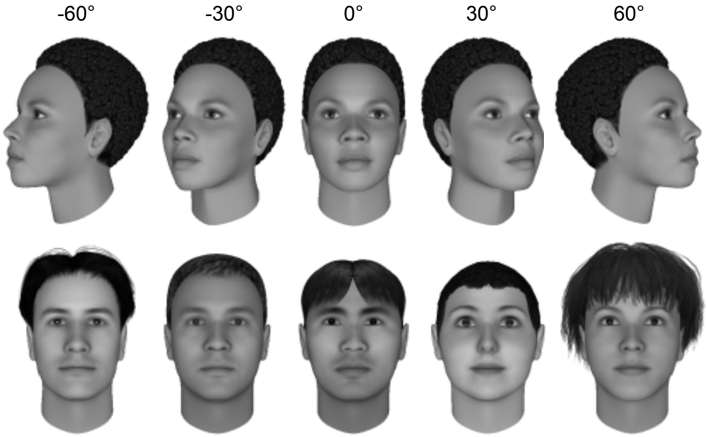

Figure 3.1: Stimuli. The stimuli included five different viewpoints (-60, -30, 0, 30, and 60°, upper row) of six different individuals (lower row).

To rule out potential confounding effects of low-level similarity, we developed an experimental stimulus set guided by the results of a biologically realistic model of V1 neurons (Figure 3.2). We analyzed activity patterns from multiple regions of interest (ROIs) throughout the ventral and dorsal processing streams, and performed a spatially unconstrained searchlight analysis (Kriegeskorte et al., 2006) to uncover any additional areas that exhibited selectivity for viewpoint symmetry.

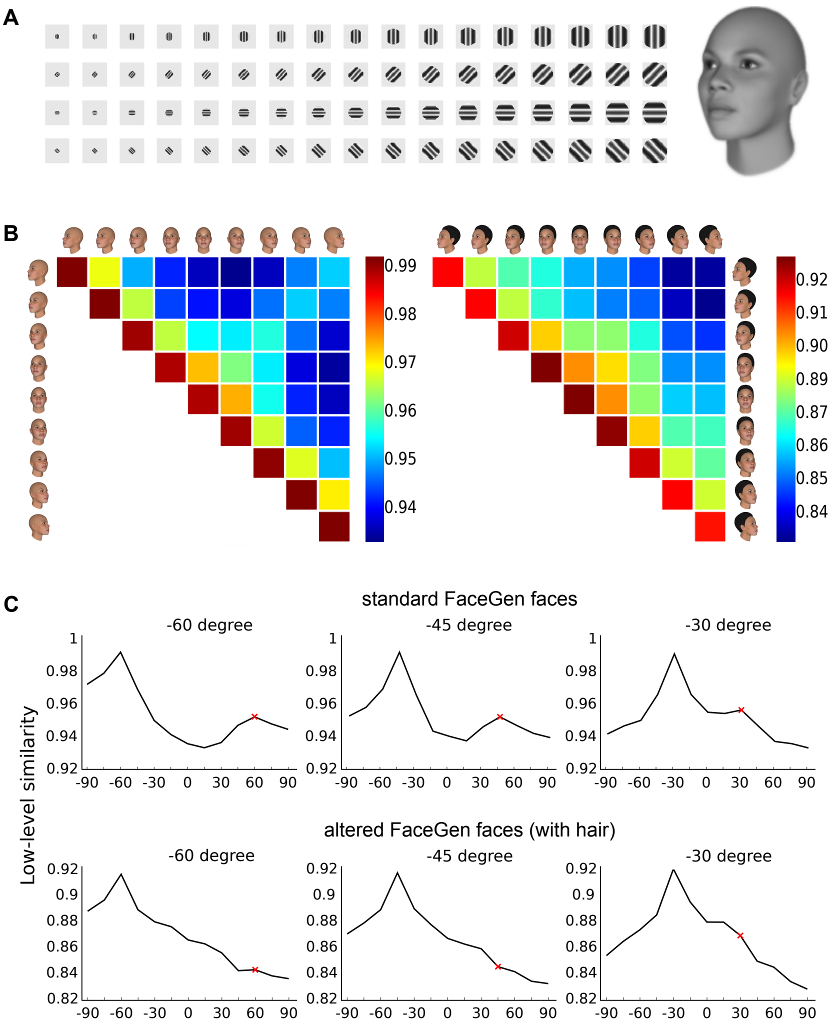

Figure 3.2: Control for low-level confounds. (**a**) To exclude the possibility that low-level features of the stimuli would already lead to patterns of viewpoint mirror symmetry, a biologically realistic model of V1 simple cells was implemented (see Materials and Methods for details). As shown on an exemplary face on the right, the stimuli were spatially filtered (foveated) to account for differences in visual accuracy. The size of the face is proportional to the size of the Gabor filters used in the model. (**b**) The V1 model responses to the standard FaceGen stimuli, as shown on the left, showed increased correlations for mirror-symmetric head orientations. This low-level confound was overcome by the addition of structured hair (shown on the right). (**c**) The low-level similarity tuning curves, as estimated from the model. The red "x" marks the mirror-symmetric viewpoint.

3.1.2 Materials and Methods

Participants

Ten healthy subjects (aged 22-34 years, four female) with normal or corrected-to-normal vision participated in the experiment. One subject had to be excluded from the analyses due to extreme signal dropout in the vicinity of the ear canal. All subjects were informed of their right to withdraw from the experiment at any point in time and gave written consent to participate. The study was approved by the Vanderbilt University Institutional Review Board.

Experimental Design and Procedure

Each experimental run consisted of 12 blocks: fixation blocks at the start and end of a run, and 10 blocks containing two presentations of each of the five viewpoint conditions (-60°, -30°, 0°, 30°, 60°). The order of conditions was pseudo-randomized and it was ensured that no condition was repeated in two consecutive blocks. Each block included three presentations of each of six face identities (pseudo-randomized order) and lasted 20s, leading to a total time of four minutes for every experimental run. Every odd run showed upright faces whereas every even run showed inverted faces. A complete scan session typically included 18 runs (9 for the upright and 9 for the inverted conditions) and lasted 2 to 2.5 hours.

Each stimulus was shown for 800ms, followed by a blank of 311ms in which only a small fixation dot remained visible. Subjects were asked to perform a 1-back detection task for which stimulus repetitions occurred randomly with a probability of 0.15. Furthermore, the horizontal and vertical position of the stimuli was randomly jittered by up to 10 pixels. The display computer was a luminance-calibrated MacBook Pro using Matlab and the Psychophysics Toolbox 3 (Brainard, 1997) for experimental control. The stimuli were projected on a screen and covered 5.5° of visual angle.

Stimuli & V1 Model

The stimuli were created using the face modeling software FaceGen (Singular Inversions Inc). They included six individuals (three female) shown from five different viewpoints (-60°, -30°, 0°, 30° and 60°) on a white background, leading to a total of 30 grayscale stimuli.

To exclude potential low-level explanations, we tested the stimuli prior to the experiment based on a biologically realistic model of V1 simple cells (Serre & Riesenhuber, 2004). This model is based on a set of 2D Gabor functions with 17 different receptive field sizes and 4 orientations, the parameters of which were previously estimated based on data from monkey electrophysiology (Figure 3.2a). Moreover, to account for the effects of decreasing visual acuity in the periphery of the visual field, we added an additional preprocessing step in which the stimuli were 'foveated' such that the model input contained high spatial resolution only in the center of the stimulus and decreasing high spatial frequency content towards the periphery. The output of the model was used to create correlation-matrices depicting the low-level similarity between the different experimental conditions given a set of stimuli. This way, we were able to evaluate whether effects of viewpoint symmetry were evident in this low-level description of the stimuli and to change them accordingly. Interestingly, an analysis of the default FaceGen stimuli, which do not include hair, revealed higher low-level similarity between the mirror-symmetric viewing angles (for instance -60° and 60°) than between the respective angles and the straight-on face (-60° and 0°). We suspected that this low-level confound was due to the relatively homogenous and low-texture posterior part of the head. Because of this, we added structured hair to the face stimuli and were thereby able to overcome this low-level confound (Figure 3.2b and c). In fact, the resulting faces even exhibit decreased correlations between viewpoint-symmetric viewing angles as compared to the correlations with the straight-on faces. An overview of the final faces and face angles can be seen in Figure 3.1.

In addition to allowing us to avoid low-level confounds arising from the stimuli, we used the output of the computational V1 model for the final stimuli as a predictor of low-level similarity in the later multivariate analysis.

MRI Data Acquisition

The experimental data was collected at the Vanderbilt University Institute for Imaging Science using a 3T Philips Intera Achieva magnetic resonance imaging scanner with an eight-channel head coil. The functional data was acquired using standard gradient-echo echoplanar T2*-weighted imaging with 28 slices, aligned approximately perpendicular to the calcarine sulcus and covering the entire occipital lobe as well as the posterior parietal and posterior temporal cortex (TR, 2s; TE, 35 ms; flip angle, 80°; FOV, 192 x 192; slice thickness 3 mm with no gap; in-plane resolution, 3 x 3 mm). In addition to the functional images, we collected a T1-weighted anatomical image for every subject (1 mm isotropic voxels). A custom bite bar system was used to minimize the subject's head motion.

fMRI Analysis

Preprocessing

Preprocessing of the fMRI data was based on Freesurfer, FSL and custom Matlab scripts. The functional data was first motion corrected with respect to the average of one run. This average image was also used to coregister the functional with the structural T1 data. After detrending the functional data and converting to percent signal change, the spatial mean of the individual ROIs were regressed out at every point in time. Following this, we z-transformed the data with respect to the mean and standard deviation of the signal across the whole run. Finally, to extract patterns of voxel activity for the different conditions, we took the average across the corresponding time-series, excluding the first 8 seconds after condition onset to account for hemodynamic lag. For the ROI analyses and the searchlight estimates, no smoothing was applied and the data remained in its native space (subject co-alignment was performed based on individual cortical curvature, as described below). The structural volumes were automatically segmented into gray- and white-matter and flattened/inflated using Freesurfer (Fischl et al., 1999a,b).

Correlation Analysis

After preprocessing the data, we estimated the similarity of response patterns across all viewing angles in the different ROIs. To do this, we used a Pearson correlation measure along with an iterative split-half procedure. In the split-half method, individual functional runs were randomly divided into two sets and the respective average response vectors of the two halves were then used to estimate the correlations between the different conditions. The resulting correlation values were then Fisher z-transformed. This entire procedure was repeated for 2000 random splits of the functional runs for each subject. The average of the resulting correlation values can be represented as a correlation matrix, which depicts the similarity of the voxel response patterns across all pairs of conditions, including the similarity of repeated presentations of one condition with itself. The standardized correlation matrix can then be used as input for subsequent analyses, in which its congruency with different models or predictors are tested (Kriegeskorte et al., 2008a,b).

Here, we applied a two-step analysis procedure. First we estimated the extent to which the similarity of activity patterns across changes in viewpoint could be attributed to low-level similarity. Next, we estimated the degree to which viewpoint symmetry accounted for the response patterns in the region of interest, independent of that region's sensitivity to low-level similarity.

To estimate whether an ROI showed selectivity for low-level similarity, we computed the Spearman correlation between the predicted correlation matrix of the computational V1 model (Figs. 3.2 and 3.3a, left) and the empirical correlation matrices, which were estimated based on the activity patterns in the respective ROIs of both hemispheres of each subject. (For the analysis, only the upper triangular part of the matrices were used, as the correlation matrices are themselves essentially symmetric, with slight deviations from perfect symmetry caused by the finite number of split-half replications. Hence the lower triangular part of the matrix can be neglected in the analysis for reasons of efficiency.) This approach lead to an effect-estimate for each subject and ROI. Thus, to test whether an ROI showed significant effects of low-level similarity, the corresponding distribution of correlation values can be tested against a null result of zero correlation, by applying a t-test to the Fisher-transformed correlation values.

Next, we estimated the effects of viewpoint symmetry by constructing a model correlation matrix that predicted high correlation values for mirror-symmetric views and low correlations for non-symmetric ones (see Figure 3.3a, right). It should be noted that regardless of whether a brain region were selective for low-level similarity or mirror symmetry, both forms of selectivity should lead to high correlation values along the diagonal of the correlation matrix when the same viewpoint is presented. However, to be conservative in our estimates of sensitivity to mirror symmetry, we chose to use a prediction matrix with high correlations only for the mirror-symmetric cells. As a result, the model matrices for low-level similarity (Fig 3.3a, left) and mirror symmetry (Figure 3.3a, right) were nearly orthogonal. To ensure complete orthogonality, we first regressed out the effects of low-level similarity from the empirical matrix and then computed the effect size of viewpoint symmetry based on the residual pattern, again based on a spearman correlation. This allowed us to calculate the partial correlation between the predicted effects of mirror symmetry and the empirical measures of cortical similarity across changes in viewpoint, having partialled out the potential contributions of low-level visual similarity (Kriegeskorte et al., 2008a).

Finally, we compared the correlation matrices of the different ROIs, independently of the two models tested. For this, we first correlated all ROI correlation matrices with each other (taking as basis the average correlation matrix across subjects and inversion condition), thereby forming a second-order similarity matrix. This matrix was then subject to a principal component analysis (PCA).

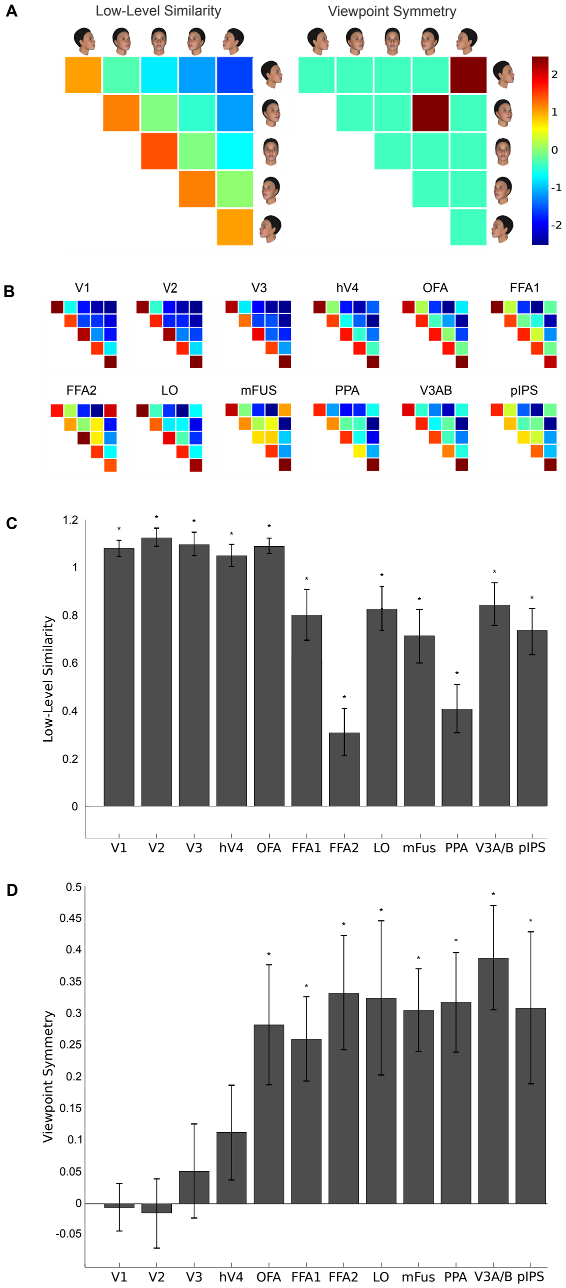

Figure 3.3: Effects of low-level similarity and viewpoint symmetry. (for captions, see next page)

Figure 3.3: (shown on previous page) (**a**) In a given ROI, the effects of low-level similarity were estimated by correlating the upper triangle of the empirical correlation matrices of both hemispheres with a model of low-level similarity, derived from the output of a computational V1 model (shown left). The effects of viewpoint symmetry, which predict higher correlation values for viewpoints with mirror-symmetric viewing angles, were estimated based on the partial correlation between the empirical correlation matrix and the viewpoint-symmetry model (right), after first regressing out the effects of low-level similarity. (**b**) Visualization of the average correlation matrix for the ROIs. Please note that we estimated the effect sizes for every subject individually and not based on these averages. (**c**) The average effect sizes of low-level similarity in the different ROIs (error bars indicate SEM). All regions show significant effects. (**d**) Average effect size of viewpoint symmetry. While higher level ROIs show significant effects of viewpoint symmetry, the early and intermediate-level areas V1-hV4 do not (see text for details and p values).

Similar to multidimensional scaling (MDS), this approach allowed us to visualize the similarity relationship of the individual ROIs in a lower-dimensional space. To directly estimate which parts of the correlation matrix explained most variance across ROIs, we performed an additional PCA in which we used every cell of the upper diagonal of the correlation matrix as an input dimension and the different ROIs as individual observations. This approach has the advantage that the principal components can be visualized in the same manner as the correlation matrices of the ROIs.

Searchlight Analysis

The computation for the searchlight analysis was mostly identical to the correlation analysis performed for the ROIs, but based only on data from cortical voxels falling within the respective searchlight. The underlying functional data was z-transformed based on the mean activity and standard deviation of each voxel across the whole run. Each searchlight included 3x3x3 voxels, from which a correlation matrix was estimated using Pearson correlation and an iterative split-half approach (using 200 random splits), as described above. Once estimated, the correlation matrix was tested for its correlation with a model of low-level similarity and its partial correlation with the viewpoint symmetry model after regressing out the effects of low-level similarity. This again yielded an effect estimate for each of the two models, which were assigned to the voxel in the center of the searchlight. Shifting the searchlight across the whole brain then leads to an effect-map for the two models, low-level similarity and viewpoint symmetry, for every subject. For the group analysis, the results of the searchlight analysis for all subjects were first transformed into a common space (fsaverage, (Fischl et al., 1999b)) via spherical averaging based on cortical surfaces, and smoothed to account for smaller errors due to imperfect intersubject alignment using a 6mm full-width-half-maximum Gaussian kernel. The effects were then modeled by a general linear regression. The resulting significance map was subject to a clusterwise correction for multiple comparisons based on Monte Carlo simulations. Null volumes of normally distributed data were generated on the cortical surface and spatially filtered to match the smoothness of the subject effect size maps, as estimated by a spatial AR1 model. Clusters were defined as contiguous sets of surface vertices exceeding a significance value of p<0.01. Clusters of activation were determined to be significant when their size exceeded that of the largest cluster in 95% of the simulated null volumes, for a clusterwise significance of p<0.05. As an additional analysis, we performed this same procedure using a larger searchlight of 5x5x5 voxels, and observed essentially the same pattern of results.

Functional ROI Definitions

ROIs were defined based on independent sets of localizer data, which was either collected during the experimental session (higher-order visual areas) or in a previous scan (retinotopic visual areas).

- *Retinotopic Areas.* The visual areas V1, V2, V3, hV4, V3A/B, and pIPS were defined based on standard retinotopic mapping (Sereno et al., 1995; Engel, 1997) on a flattened cortical representation. pIPS was defined as the union of areas V7, IPS1, and IPS2, the borders between which could not be clearly delineated in all hemispheres. In 4 hemispheres (3 left, 1 right), the retinotopic maps showed minimal significant activation in this region; for these hemispheres, approximate anatomical ROIs were defined along the medial bank of the posterior IPS. For one subject, no retinotopic mapping data was available. Here a V1 ROI was defined based on automated anatomical criteria (Hinds et al., 2008).

- *Higher-Order Visual Areas.* In addition to the experimental runs, we also included three runs of a functional localizer targeting a number of higher-order visual areas. The localizer included separate blocks showing objects, faces, images of bodies (without heads) and blocks containing scrambled versions of the stimuli used in each of these three categories.

Table 3.1: Overview of the number of subjects for which the RIOs were successfully defined.

V1	V2	V3	hV4	OFA	FFA1	FFA2	LO	mFus	PPA	V3A/B	pIPS
lh:9	lh:8	lh:8	lh:8	lh:6	lh:9	lh:6	lh:7	lh:9	lh:9	lh:9	lh:9
rh:9	rh:8	rh:8	rh:8	rh:9	rh:9	rh:8	rh:9	rh:9	rh:9	rh:9	rh:9

FFA, fusiform face area; LO, lateral occipital cortex; mFus, mid fusiform;
OFA, occipital face area; pIPS, posterior intraparietal sulcus.

As the focus of the main experiment was on the representations of different head orientations, the face localizer contained not only stimuli showing front-on faces but also the other orientations used in the main experiment (30°, -30°, 60° and -60°). To avoid differences in the retinotopic extent of the scrambled and unscrambled versions of the images, we first fitted a 2D Gaussian function to the grayscale image of each stimulus. The resulting parameter estimates were then used to create a corresponding probability density function, which served as basis for the positioning of the scrambled parts of the images. The scrambled images therefore occupied approximately the same region of space as their unscrambled counterparts.

The sequence of localizer blocks was similar to the one used by Freiwald & Tsao (2010), showing a block of fixation, followed by scrambled faces, faces, scrambled objects, objects, scrambled bodies, bodies and finally another block of fixation. The contrasts used for the individual regions are detailed below; an overview of the number of subjects for which the respective region could be defined is given in Table 3.1. Importantly, voxels previously labeled as belonging to one of the retinotopically organized areas (V1 to hV4, V3A/B and pIPS) were excluded from the higher-order visual area definitions. For all of the higher-level ROIs, we selected voxels exhibiting significantly larger activation in the respective contrast (at least p<0.01, uncorrected). Similar to the early visual areas, the ROIs were defined on the flattened cortical representation of every individual subject.

For the fusiform face area (FFA), we localized the voxels in the fusiform gyrus, whose activation was significantly higher for faces than objects (Kanwisher et al., 1997a). Where applicable, we assigned the labels FFA1 and FFA2 to the posterior and anterior patches of FFA, similar to Pinsk et al. (2009). The occipital face area (OFA) was localized based on the same contrast as FFA and restricted to face-selective voxels in the occipital lobe (Puce et al., 1996; Gauthier et al., 2000b). The lateral occipital cortex (LO) was defined as the set of voxels around the LO exhibiting significantly higher activation for complete objects as compared with scrambled ones (Malach et al., 1995; Kanwisher et al., 1997b).

The parahippocampal place area (PPA) responds preferably to images of houses or scenes, as compared with faces ((Aguirre et al., 1998; Epstein & Kanwisher, 1998). As our localizer did not include corresponding stimulus conditions, a second set of localizer data was used to define PPA. These data were collected during a different scan session. PPA was defined as voxels around the posterior parahippocampal gyrus, showing significantly higher activation for houses than faces. Mid fusiform gyrus (mFus) was previously described as an object-selective region located on the medial side of the fusiform gyrus (Grill-Spector, 2003). In line with this definition, we determined mFus as the set of voxels in the fusiform gyrus and intermediate to FFA and PPA that exhibited significantly larger activation for objects, as compared with faces. The definition of mFus excluded voxels previously defined as PPA.

Eye-Tracking Analysis

During the experimental runs, the subjects were asked to remain fixated on the center of the screen, as indicated by a small red dot, and their eye position was monitored using an fMRI compatible 60Hz eye-tracking system (Applied Science Laboratories Eye-Trac 6). To exclude the possibility that any residual differences in eye position could explain our observed fMRI results, the available eye-tracking data of seven of our subjects were used for further analyses. First, we tested whether there were systematic differences in the average eye position in the different condition. For this, the average horizontal and vertical gaze direction of every subject and condition was entered into a repeated-measures ANOVA. In addition to this, we adapted a similar approach to Harrison & Tong (2009) and performed the same analysis on the eye tracking data as we did previously on the fMRI data to see whether the pattern of eye-movement would lead to effects in the direction of the found fMRI results. Accordingly, we estimated empirical correlation matrices based on patterns of eye movements rather than fMRI activation for every subject, and then tested their correlation with the model of low-level similarity and the partial correlation with the viewpoint symmetry model. To estimate the empirical correlation matrices, we first converted the eyetracking data of each subject, run and condition into a probability density function of fixation. Based on these distributions, and analogous to the fMRI analysis, we then computed a correlation matrix for every subject across all conditions by applying an iterative split-half procedure. The resulting correlation matrices, one for every subject, were then used to estimate the effects of low-level similarity and viewpoint symmetry.

3.1.3 Results

Participants viewed upright or inverted faces in separate experimental runs, shown from 5 possible viewpoints using a randomized fMRI block design. Observers were instructed to maintain fixation on a point in the center of the screen and to perform a one-back stimulus repetition detection task (average hit rate 69%, d'=2.99).

Multivariate Pattern Analysis

For each region of interest, we measured the similarity of cortical activity patterns across the five presented viewing angles of faces by dividing each participant's set of fMRI runs into separate halves and measuring the correlation strength (Fisher z-transformed Pearson's *r*) across these independent data sets (see Materials and Methods). All pairwise correlations between face viewpoints can be displayed in a correlation matrix, which can then be used as basis for further analyses. Here, we analyzed the pattern of correlations across the different viewpoints, based on two separate models of visual selectivity. The first model estimated low-level visual similarity, based on a computational V1 simple cell model (Figure 3.3a, left). This model predicts that repeats of the same viewpoint will elicit high correlation values (i.e., similar pattern), nearby viewpoints will elicit moderate correlation values, and distal mirror-symmetric views will elicit low correlation values. By contrast, the second model, viewpoint symmetry, predicts increased correlation values between mirror-symmetric views as compared to non-symmetric ones (Figure 3.3a, right). Our measure of sensitivity to mirror symmetric viewpoints focused specifically on the predicted relationship between +60° and -60° views, and between +30° and -30° views. With this, we ensured that our measures of goodness-of-fit with this model were approximately orthogonal to our measures of sensitivity to low-level similarity. For a given ROI, the contributions of the two effects, low-level similarity and viewpoint symmetry, were assessed based on the agreement of the respective model with the empirical correlation matrices across the individual subjects (see Materials and Methods). As an overview, the average correlation matrices of the ROIs are shown in Figure 3.3b.

Using this approach, we first assessed the effects of low-level similarity and viewpoint symmetry in early visual areas V1 to hV4. Our analysis of these areas revealed significant effects of low-level similarity ($p<0.001$ in all cases, one-tailed t-test), but no significant viewpoint-symmetric effects ($p=0.55$, $p=0.6$, $p=0.64$ and $p=0.08$ for V1, V2, V3, and hV4 respectively, one-tailed t-test). Low-level similarity alone, as predicted by our computational V1 model, accounted for 72, 75, 72 and 65% of the total variance for V1, V2, V3 and hV4 respectively.

Following this, we analyzed higher order face-selective regions including the OFA, as well as the posterior and anterior segments of the FFA (FFA 1 and FFA2). Again, all these regions showed significant patterns of low-level similarity ($p<0.02$ in all cases, one-tailed t-test, Figure 3.3c). In contrast to the early visual areas, however, they also exhibited reliable effects of viewpoint symmetry ($p<0.01$, one-tailed t-tests, Figure 3.3d). Moreover, although LO, mFus and PPA are known to respond maximally to views of objects or scenes, they are nevertheless activated by stimuli showing faces (Ishai et al., 1999). We therefore investigated whether activation patterns in these ROIs might also reveal effects of viewpoint symmetry. All three areas, LO, mFUS and PPA showed reliable effects of viewpoint symmetry ($p<0.02$ in all cases), as well as low-level similarity ($p<0.02$ in all cases). Finally, we concentrated on areas in the dorsal stream of visual processing and tested areas V3A/B and the posterior intraparietal sulcus (pIPS). Both regions showed significant effects of low-level similarity ($p<0.001$) as well as viewpoint symmetry ($p<0.02$).

To summarize, we found statistically significant effects of low-level visual similarity as well as viewpoint symmetry in all tested higher-order visual areas. In contrast, early- and intermediate-level visual areas only showed significant effects of low-level visual similarity, and no signs of viewpoint symmetry.

We substantiated our above conclusions with a series of controls. First, we confirmed that similar results were obtained by using a standardized regression approach in which the empirical correlation matrix was jointly predicted by the model of low-level similarity and viewpoint symmetry. This approach resulted in the same pattern of significant and non-significant ROIs as the two-step (partial) correlation procedure: significant effects of low-level similarity across all ROIs ($p<0.01$ in all cases, one-tailed t-test) and significant effects of viewpoint symmetry for all higher-order visual areas reported above ($p<0.05$, one-tailed t-tests) but no such significant effects in early areas ($p>0.05$, one-tailed t-tests).

Second, we find no evidence for viewpoint symmetry in our computational V1 model as well as the fMRI pattern of responses in early visual areas, indicating that our choice of stimuli effectively avoided low-level stimulus confounds (see Figs. 3.2b right and 3.3d). We observed strong effects in low-level similarity in area V1, consistent with the predictions of our V1 simple cell model, whereas effects of viewpoint symmetry emerged only at higher stages of visual processing. Next, we tested a uniform correlation matrix with additive Gaussian noise in the same protocol as the ROIs to exclude explanations based on a fully viewpoint-invariant representation or a potential bias towards one of the two model predictors. This led to no significant similarity or viewpoint symmetry effects (p=0.76 and p=0.71, respectively, one-tailed t-test). Furthermore, our results also cannot be explained based on overall amplitude changes in the ROIs, as we regressed out the spatial average of each ROI during preprocessing (see Materials and Methods) and because the correlation measure disregards the spatial mean of every condition. To investigate the possibility that residual eye movements of our subjects could explain our effects, we examined the eye tracking data, for which we obtained reliable measurements from seven of our nine subjects. We found no reliable differences in the mean horizontal or vertical fixation position across conditions and subjects (p>0.26 in all cases, repeated-measures ANOVA with condition as factor). In addition, we analyzed the complete set of eye tracking data based on the same approach previously applied to the functional MRI data. However, we now utilized probability density functions of fixation to estimate the empirical correlation matrix instead of fMRI activation patterns (see Materials and Methods). This analysis showed no significant effects of low-level similarity (p=0.2, one-tailed t-test) or viewpoint symmetry (p=0.9, one-tailed t-test) in the residual eye movements. Thus, we have no indication that eye movements can account for the observed effects.

Having found patterns of viewpoint symmetry in the responses to upright faces in both face- and object-selective higher-order areas, we next asked whether similar effects could also be observed following the presentation of inverted faces. Compared to the processing of upright faces, inverted faces have been suggested to rely on distinct cognitive mechanisms, by engaging object- and scene-selective regions in addition to the highly specialized face-processing network (Aguirre et al., 1999; Haxby et al., 1999; Epstein et al., 2006; Pitcher et al., 2011). Hence, it was interesting to test whether the effects of viewpoint symmetry would generalize to the processing of inverted faces. Applying the same analyses as before, we found the same pattern of results for inverted faces (Figure 3.4) as for upright faces (Figure 3.3, compare c,d). A repeated-measures

ANOVA with ROI and inversion as within-subject factors revealed a significant effect of ROI ($p<0.05$), but no significant effects of inversion ($p=0.38$) or interaction effects ($p=0.34$; all p-values Greenhouse-Geisser corrected). When tested individually, the higher-level areas OFA, FFA 1 and 2, LO, mFus, PPA again showed significant viewpoint symmetry effects for inverted faces ($p<0.025$ in all cases, one-tailed t-test). In contrast to this, all of the early visual areas (V1-hV4) again failed to show significant effects of viewpoint symmetry ($p>0.6$ in all cases, one-tailed t-test). V3A/B and pIPS failed to show significant viewpoint symmetry effects for inverted faces ($p>0.05$ in both cases, one-tailed t-test), but there was also no statistically significant difference between the effect-sizes for upright- and inverted faces ($p>0.05$ in both cases, paired t-test). These results indicate that effects of mirror symmetry were also prevalent for face views presented upside-down.

To compare the correlation matrices of the different ROIs with each other, we projected an across-ROI similarity matrix into two dimensions using PCA (see Materials and Methods). In the resulting space, the distances between ROIs resemble the similarity of the respective correlation matrices (Figure 3.5a). Notably, the first principal component, which explained 66% of the similarity structure across ROIs, exhibits a clear separation between regions with and without viewpoint symmetry (the second component explained an additional 29% of variance). Following this, we determined which cells of the correlation matrix accounted for the most variance across ROIs by performing a PCA directly on the entries of the correlation matrices (see Materials and Methods). The resulting first principal component, based on which 63% of the variance across ROIs could be explained, exhibits large weights in the two matrix diagonals. This is in direct agreement with our models of low-level similarity and viewpoint symmetry (Figure 3.5b). It should be noted that the principal component was found by simply maximizing the explained variance across ROIs; this approach is model-free and did not make any assumptions regarding how the correlation matrices should tend to vary across ROIs. The results of the PCA analysis provide further confirmation that sensitivity to mirror symmetry is a prominent functional organization principle that accounts for changes in visual selectivity across the visual hierarchy.

Searchlight Analysis

In addition to the analyses of specific regions of interest, we performed a searchlight analysis to test for viewpoint-symmetric response patterns throughout the whole functional volume (Kriegeskorte et al., 2006). The underly-

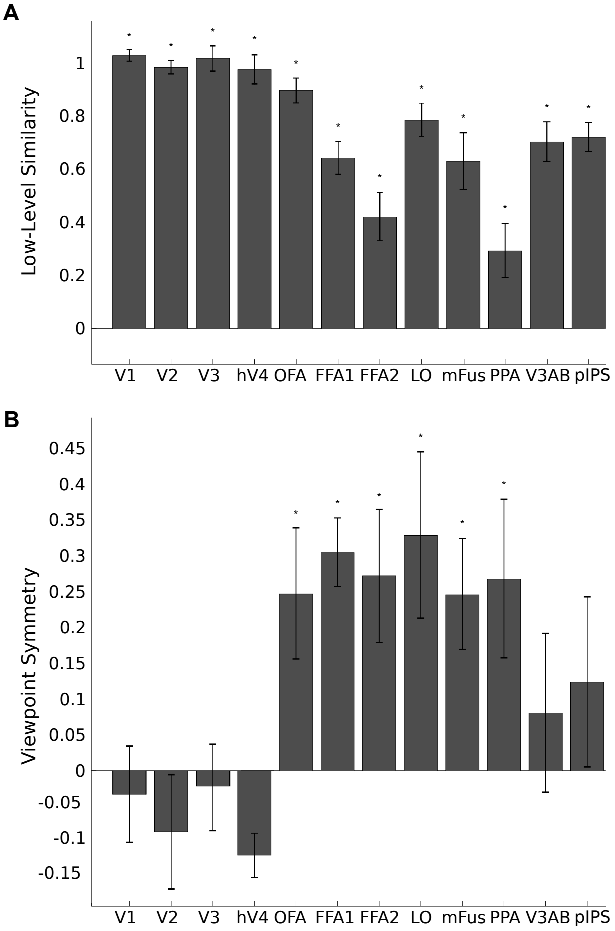

Figure 3.4: ROI effects of low-level similarity and viewpoint symmetry. (**a**) Effects of low-level similarity and (**b**) viewpoint symmetry during the processing of inverted faces. As in the upright condition, all areas show significant effects of low-level similarity, whereas only higher level areas show robust effects of viewpoint symmetry.

ing analysis was identical to the multivariate pattern analysis described above. However, instead of selecting all voxels in an ROI, the activation patterns of a local 3x3x3 neighborhood of voxels were used for the analysis. Shifting this searchlight across the functional volume thus yields an effect map for every subject and model. These individual subject maps were then normalized to a common space, smoothed, and tested for clusters showing significant effects of

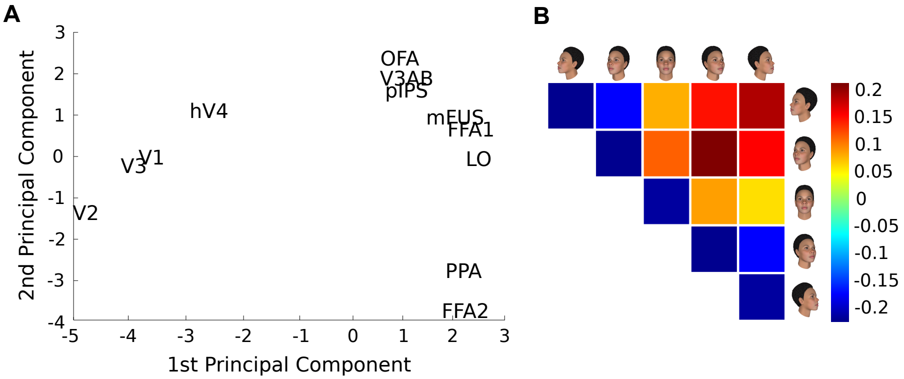

Figure 3.5: Principal component analyses. (**a**) When projected into two dimensions, the similarity of the correlation matrices of the different ROIs can be visualized. The first component, which already explains 81.8% variance, shows a clear separation between ROIs with and without effects of viewpoint symmetry (the second component explains an additional 13.9% of the variance). (**b**) The resulting first component when computing a PCA directly on the entries of the correlation matrices. The component exhibits large weights in the two diagonals, in direct agreement with the effects of low-level similarity, and viewpoint symmetry.

viewpoint symmetry or low-level similarity on the population level (see Materials and Methods).

The searchlight analysis revealed a large band of cortical regions exhibiting significant viewpoint-symmetric response patterns (Figure 3.6). In line with our earlier results, this band of viewpoint symmetry overlapped with the previously tested ROIs, including higher-order visual areas of the ventral and dorsal streams. Early retinotopic areas failed to exhibit symmetry effects and their responses were again found to be best explained based on low-level similarity. Moreover, while our searchlight analysis revealed a cluster of significant viewpoint-similarity effects in posterior superior temporal sulcus (STS), the searchlight approach revealed no significant effects of viewpoint symmetry in this region. Finally, more anterior regions such as the anterior part of the temporal lobe did not show any significant patterns of low-level similarity or viewpoint symmetry.

As a control, we performed the same analysis with a larger searchlight of 5x5x5 voxels. The brain regions implicated were the same as those identified using the 3x3x3 searchlight, except that the band of cortical regions was somewhat larger due to the use of a larger searchlight. This verifies that our results were not dependent on a specific searchlight-size. For this study, we present the

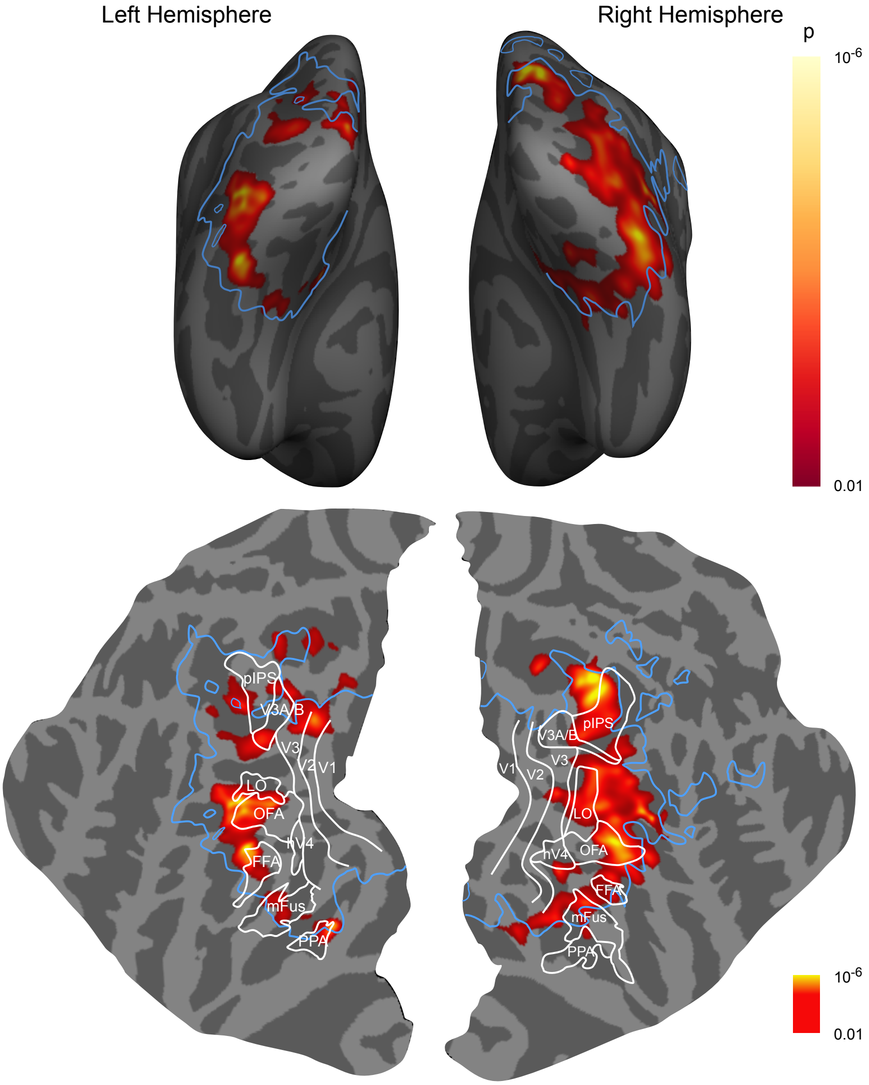

Figure 3.6: Searchlight results. Clusters of significant low-level similarity and viewpoint symmetry across subjects on the inflated (top) and flattened (bottom) standard brain. The light-blue line delineates regions showing significant effects of low-level similarity. Regions of significant viewpoint symmetry are marked in hot colors. They form a band of higher order visual regions, which excludes more posterior (early and intermediate-level) visual areas and more anterior areas. The delineated ROIs are from the localizer results of a representative subject (M051).

results of the 3x3x3 voxel searchlight, as this analysis provided a more conservative and spatially precise measure of the regions that displayed a preference for mirror symmetric views of faces.

3.1.4 Discussion

Our analyses of cortical activity patterns revealed a spatially distributed, yet functionally specific representational property in the human visual system: selectivity for mirror-symmetric viewing angles of faces. This property was not restricted to a single focal region, but instead was found to be prevalent in a large band of higher-order visual areas. In addition to regions typically associated with face processing, OFA, FFA1 and FFA2, we observed effects of viewpoint symmetry in several cortical areas that do not respond preferentially to faces, including object-selective (LO, mFUS) and scene-selective areas (PPA). These effects were equally prevalent for inverted faces as for upright faces, even though stimulus inversion is known to impair face-specific processing (Yin, 1969; Valentine, 1988; Kanwisher et al., 1998) and the robustness of face-specific responses (Freiwald et al., 2009). This suggests that our fMRI measures of sensitivity to viewpoint symmetry do not depend on a cortical specialization for faces.

An unexpected finding was the fact that the dorsal regions V3A/B and the posterior IPS also exhibited selectivity for symmetric views of face stimuli. Object processing is commonly believed to rely on the ventral visual pathway (Mishkin & Ungerleider, 1982; Goodale & Milner, 1992). However, a few studies have demonstrated the presence of shape selectivity and view-invariant object selectivity in the parietal lobe as well (Sereno & Maunsell, 1998; Konen & Kastner, 2008; Króliczak et al., 2008).

Importantly, all of the visual areas we found to be sensitive to mirror-symmetric viewing angles also revealed strong effects of low-level similarity. This indicates that these ROIs failed to show complete viewpoint invariance, and that the acquisition of partial view invariance does not preclude the possibility of maintaining sensitivity to low-level image similarity as well. These findings are in line with earlier fMRI work demonstrating viewpoint-dependent adaptation effects for faces in the posterior fusiform region (PFS), including the FFA (Grill-Spector et al., 1999; Pourtois et al., 2005; Andresen et al., 2009).

In parallel to our work, a different research group has recently reported that face-selective regions including the FFA, right STS, as well as object-sensitive area LO, exhibit effects of viewpoint symmetry (Axelrod & Yovel, 2012), while no symmetry effects were found in the OFA. Here, we specifically aimed at assessing the prevalence and generality of viewpoint symmetry effects while controlling for low-level confounds typically present in standard FaceGen stimuli. Because of this, we tested upright and inverted faces across a multitude of visual areas, including retinotopically defined early visual areas V1-hV4, ventral areas OFA, FFA 1 and 2, LO, mFUS and PPA as well as dorsal areas V3 A/B, pIPS (including V7 as well as IPS 1 and 2). We find positive evidence of selectivity for mirror-symmetric views not only in the FFA, OFA, and area LO, but also in medial ventral temporal areas such as mFus and the PPA, as well as dorsal visual areas in the posterior parietal cortex.

The extensive band of higher-order visual areas, for which we find sensitivity to symmetric 3D viewpoints, overlaps to a considerable extent with cortical areas previously shown to prefer symmetric 2D patterns. In particular, the lateral occipital complex has been found to respond more strongly to symmetric dot patterns (reflected along the vertical axis) than to random dot patterns (Sasaki et al., 2005; Tyler et al., 2005). Because we observed sensitivity to symmetric views of faces rotated away from 0°, our results cannot be explained by a general preference for visually symmetric stimuli. Nevertheless, the overlap of areas raises an interesting question as to whether sensitivity to 3D viewpoint symmetry and 2D visual symmetry might reflect a shared neural mechanism.

Another related visual property is mirror reversal. Previous studies using fMRI adaptation found that ventral visual areas exhibit invariance to mirror reversals of written text, objects and scenes (Eger et al., 2004; Dehaene et al., 2010; Dilks et al., 2011). Consistent with these neuroimaging studies, neurophysiological recordings in monkeys have shown that left-right mirror reversals lead to more similar responses in inferotemporal neurons, when compared to stimulus reversals along the vertical dimension (Rollenhagen & Olson, 2000). Although these studies found evidence of invariance to image reversal in many object-sensitive areas, consistent with the present findings, they did not directly test for selectivity to mirror-symmetric viewing angles. Because of this, it was left as an open possibility that the reported invariance to image reversals could also be explained by fully viewpoint-invariant representations. Ruling out such explanations would require the presentation of the same objects from multiple viewpoints, including views intermediate to those realized by image reversal, as was evaluated in the current work.

Taken together, our findings suggest that selectivity for mirror-symmetric views may constitute an intermediate-level processing step shared across multiple higher-order areas of the dorsal and ventral streams. The prevalence of such representations could set the stage for realizing viewpoint-invariant representations at subsequent stages of visual processing. Indeed, Freiwald & Tsao (2010) found that viewpoint symmetric response properties existed in a lateral region of the monkey temporal lobe (region AL), whereas neurons in a more anterior face-selective region (called AM) exhibited complete viewpoint invariance in their selectivity for different individuals (Tanaka et al., 2007; Freiwald & Tsao, 2010). Interestingly, some of the neurons in AL maintained a preference for a particular facial identity across mirror symmetric views, suggesting that such mirror symmetric coding might serve as an important intermediate step to developing a fully view-invariant representation.

An important point to consider is why the present study found such widespread effects of viewpoint symmetry, whereas Freiwald & Tsao (2010) found these effects to be largely restricted to neurons in the anterior lateral face patch. Neurons recorded from the middle face patches (middle lateral and middle fundus of the superior temporal sulcus) often preferred a single viewpoint and failed to show evidence of viewpoint symmetry, suggesting that this visual property emerges at a relatively late stage of processing in anterior regions of the macaque visual system. Although the precise homologies between monkey and human face-selective areas have yet to be fully determined (Tsao et al., 2008), we observed effects of viewpoint symmetry at earlier processing stages in posterior ventral visual areas, including regions OFA and LO. What factors might account for the differences between studies? One major difference was that our pattern analysis approach could test for sensitivity to viewpoint symmetry without requiring the brain region in question to be selective for face stimuli or sensitive to facial identity. We observed strong effects of symmetry in regions such as the PPA and mFus, which prefer objects more than faces, presumably because these areas contain partially view-invariant representations of features that are common across multiple object classes, including faces. Another factor is that fMRI pattern analysis pools information over much larger regions of cortex, and this too might have facilitated our ability to detect information about viewpoint symmetry in posterior visual areas and cortical regions that respond weakly to face stimuli. Finally, there could also be genuine differences between species, either due to innate factors or differential amounts of experience with symmetrical stimuli. Sasaki et al. (2005) tested for sensitivity to 2D symmetrical dot patterns in both humans and monkeys, and reported finding only a small region of the monkey visual cortex that showed preferential responses to symmetrical stimuli, around areas V4d, V3A and TEO, whereas a much larger cortical region was activated in humans. Future studies might address these issues by performing comparable fMRI pattern analysis studies of viewpoint symmetry in monkeys.

It would also be interesting for future fMRI studies to investigate which regions of the human visual pathway contain view-invariant representations of facial identity. This was not possible here, as our experimental design showed different individuals from a selected viewpoint within a block. Generally, the ability to decode information about facial identity remains a major challenge for fMRI research, and although a limited degree of success has been reported (Natu et al. 2010; Nestor et al. 2011; Kriegeskorte et al. 2007, but see Tsao et al. 2008), the present study illustrates the importance of ensuring that low-level confounds cannot account for the successful discrimination of different face stimuli.

Computationally, there could be advantages to relying on viewpoint-symmetric object representations as an intermediate processing step. View-based theories of invariant object recognition propose that viewpoint invariance can be accomplished by interpolating between a small set of informative 2D views (Bülthoff & Edelman, 1992; Tarr et al., 1998; Poggio & Edelman, 1990; Ullman, 1998; Kietzmann et al., 2009). Within such a framework, viewpoint-symmetric representations could be exploited to allow for a substantial reduction in computational complexity. While selectivity for mirror-symmetric views can itself be regarded as an example of partial viewpoint invariance, it might be particularly beneficial for encoding objects with axial symmetry, such as faces, animals and many objects (Vetter et al., 1994). For this type of input, the number of viewpoint-specific representations required to represent an object could be substantially reduced by relying on representations that incorporate viewpoint symmetry as an intermediate processing step.

3.1.5 Acknowledgments

The authors would like to thank Mike Pratte and Sam Ling for valuable discussions on this project, as well as Elizabeth Counterman for her help with the data acquisition. This work was made possible by NSF grant BCS-064Z633 and NIH grant R01-EY017082 to FT, a Fulbright scholarship to TCK, NIH NRSA Fellowship F32-EY019448 to JS, ERC 269716 Multisense to PK and NIH P30-EY008126 center grant to the Vanderbilt Vision Research Center.

3.2 The Occipital Face Area is Causally Involved in Viewpoint Symmetry Judgments of Faces[2]

Abstract Humans are highly proficient at recognizing individual faces from a wide variety of viewpoints, but the neural substrates underlying this ability remain unclear. Recent work suggests that viewpoint-symmetric responses to rotated faces, found across a large network of visual areas, may constitute a key computational step in achieving full viewpoint invariance. Here, we used transcranial magnetic stimulation (TMS) to examine whether the occipital face area (OFA) causally contributes to the perception of viewpoint symmetry. The experiment followed a 2x2 design with TMS (repetitive vs. sham) and task (symmetry vs. angle judgments) as experimental factors. Subjects underwent 5 minutes of either sham stimulation or true 1Hz rTMS to the right OFA prior to each 4-minute block of behavioral test trials. Visual stimuli were presented ipsilateral to the site of TMS stimulation to avoid retinotopically specific impairments. Subjects reported either which of two consecutively presented pairs of face viewpoints was mirror-symmetric (symmetry task) or which pair of faces had a larger angular difference (angle task). Prior to the experiment, both tasks were titrated by an adaptive staircase procedure (QUEST) to achieve an average of 80% correct performance. Compared to sham, rTMS led to a significant decrease in performance specifically for viewpoint symmetry judgments, whereas no significant differences were found for the angle task. A repeated-measures ANOVA revealed a significant interaction effect, indicating that the effect of rTMS over OFA was larger for symmetry than for angle judgments. Our data provide novel evidence for the causal involvement of OFA in the processing of viewpoint symmetry and provide important restrictions on models of viewpoint symmetry and face perception in general. In particular, the specific effect on viewpoint symmetry judgments after rTMS applied to the ipsilateral OFA provides support for proposals emphasizing the role of inter-hemispheric sharing of information in the perception of viewpoint symmetry.

[2] This section was accepted as a contribution to the Vision Science Society Conference 2014. See Publication List for details.

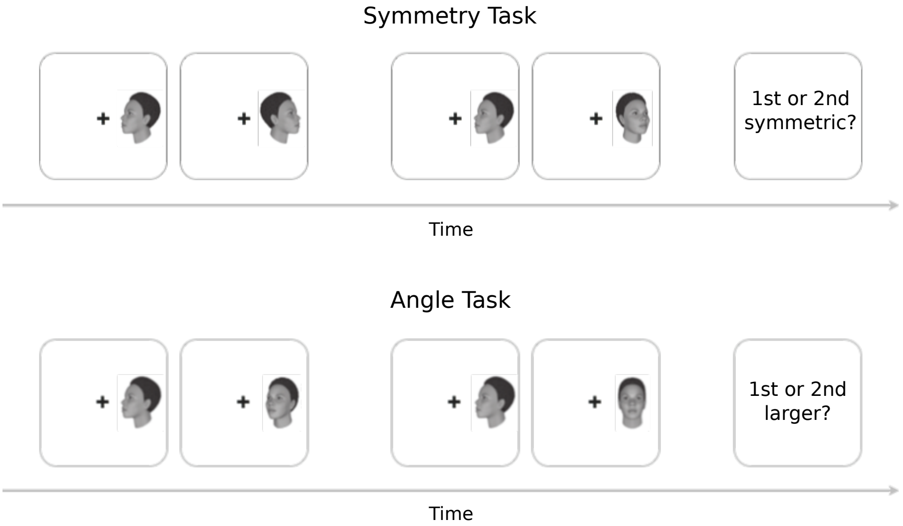

Figure 3.7: Task Design. The experiment included two tasks: symmetry judgment and an angle judgment. The latter served as control condition to exclude explanations based on generally decreased performance as the result of TMS. The stimuli were presented ipsilaterally to the stimulation, to rule out retinotopically specific impairments.

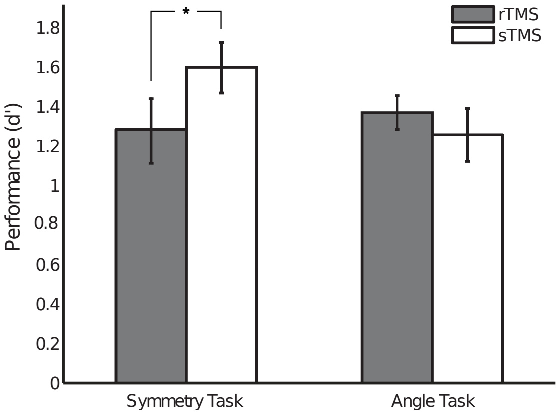

Figure 3.8: Results. Our analyses revealed a significant decrease in performance for the symmetry task, whereas no significant effects were found for the control task (angle). The difference in effect sizes is significant.

3.3 Representational dynamics of facial viewpoint encoding: head orientation, viewpoint symmetry, and eye contact

Abstract The cortical processing of facial viewpoints requires distinct forms of viewpoint encoding, from view-dependent representations, supporting mechanisms of social attention, to a fully viewpoint-invariant code, which subserves face identification. While different cortical areas have been suggested to support either function, little is known about temporal aspects of viewpoint encoding in the human brain. Here we combined electroencephalography data, recorded while subjects viewed faces shown from 37 viewpoints, with multivariate pattern analyses to investigate the representational dynamics of facial viewpoint encoding with high temporal resolution. Our data- and model-driven analyses reveal a distinct temporal sequence of viewpoint encoding. Head orientations are encoded first, potentially driven by low-level stimulus features, followed shortly afterwards by strong effects of viewpoint symmetry, the joint selectivity for mirror-symmetric viewing angles. At a considerably later stage, EEG activation patterns demonstrate a large degree of viewpoint invariance across almost all viewpoints tested with the marked exception of front-on faces, the only viewing angle exhibiting direct eye-contact. Our results indicate that the encoding of facial viewpoints follows a sequence of coding schemes, supporting distinct task requirements at different stages of face processing.

3.3.1 Introduction

Faces are among the most important visual categories, providing diverse information that are essential to a variety of visual tasks. Faces are by far the most reliable way to identify other individuals and lay the foundation for social interactions. One aspect of facial processing that has attracted much research in recent years is the cortical representation of different face viewpoints, which can lead to large changes in the retinal projection.

Viewpoints provide a fruitful, yet challenging research domain, as different computational goals in face processing rely on different sets of viewpoints and require different levels of viewpoint invariance. For example, while face identification requires mechanisms supporting full invariance across a large range of viewpoints, shared attention is in need of the opposite. Here, different head orientations provide strong cues for the recognition of another person's attentional focus and therefore need to be distinguishable (Haxby et al., 2000; Perrett et al., 1992). Finally, direct eye-contact, a powerful social and emotional cue (Senju & Johnson, 2009), frequently co-occurs with front-on viewpoints, whereas seeing a face from all other viewpoints usually signals no eye-contact. Especially in context of social interaction, front-on viewpoints thus play a special role compared to oblique head orientations. The distinct requirements introduced by these tasks allow for substantially different predictions about the underlying cortical representations of face viewpoints. While it is conceivable that different cortical structures subserve these goals, different forms of viewpoint encoding can also be separated across time, mirroring varying computational mechanisms, complexity and importance.

Here we investigate the temporal development of viewpoint encoding in the human brain by recording electroencephalograhy (EEG) data from healthy participants while they were presented with faces shown from 37 different viewpoints (example stimuli are shown in Figure 3.9a) and performed a color-change detection task at fixation. The viewpoints tested span 180°, from left to right profile in steps of five degrees. To analyze how this fine-grained set of facial viewpoints is represented at different stages of processing, we employed a multivariate representational similarity analysis (Kriegeskorte et al., 2008a) on the high-dimensional, visually evoked responses. Crucially, the high temporal resolution of the EEG measurements allowed us to investigate fast changes in the similarity structure over time.

As a first data-driven approach, we visualized the empirical similarity matrices and applied multidimensional scaling (MDS) to project the similarity structure to two dimensions. This enabled us to explore the temporal development of viewpoint encoding and the corresponding changes in the representational similarity structures across time. In addition to this, we defined three models, which were directly derived from the distinct requirements of different face processing tasks.

For the task of face identification, it was recently suggested that viewpoint-symmetry, the joint selectivity to mirror-symmetric viewpoints of faces, might constitute a key computational step in achieving full viewpoint invariance. Indeed, corresponding effects were shown based on single-cell recordings in the macaque (Freiwald & Tsao, 2010), and in human higher-level visual areas using functional magnetic resonance imaging (fMRI)(Kietzmann et al., 2012; Axelrod & Yovel, 2012). Due to the comparably low temporal resolution of fMRI, however, the latency of viewpoint symmetry effects in the human visual system is currently unknown. Because of this, and because mirror-symmetry may constitute a prevalent feature of visual coding, we tested a viewpoint encoding model that predicts increased similarity of mirror-symmetry viewpoints.

Both, viewpoint symmetry and full viewpoint invariance potentially complicate the recognition of the currently observed head and gaze direction, as leftward and rightward rotations lead to highly similar activation patterns. However, both effects were observed mainly in higher-level visual areas. This opens the possibility that lower-level visual representations support the distinction of different head orientations. Alternatively, it is possible that even a higher-level, fully invariant neuronal code also supports the decoding of other stimulus features, such as viewpoint, if a different readout strategy is used (DiCarlo & Cox, 2007). To assess whether, and at what point in time, head orientation can be decoded from visually evoked potentials, we designed a simple model which tests whether leftward and rightward head orientations can be separated based on the visually evoked response.

Finally, only the front-on viewpoint exhibits direct eye contact in the currently used stimulus set. Thus, to estimate the effects of direct vs averted gaze, the third model contrasts the front-on viewpoint with the neighboring conditions showing faces rotated away by only 5°. Although differing in direct vs averted eye-contact, the overall visual differences between the tested conditions are small.

Taken together, by combining data-driven and model-based analyses, we were able to investigate the temporal development of viewpoint encoding. To rule out alternative explanations for the found effects, multiple controls were performed. These include a more detailed investigation of low-level stimulus properties, based on a biologically realistic model of V1 simple cells (Serre & Riesenhuber, 2004), effects of residual eye movements and effects of task complexity.

3.3.2 Results

The basis of all analyses performed was the transformation of the high-dimensional EEG data into representational similarity matrices, which depict the similarity of the visually evoked responses across all conditions. Similarity was defined as the Pearson product-moment correlation between the 128-dimensional EEG response patterns (one dimension for every EEG channel). The similarity matrices were computed for each time-point individually at a temporal resolution of 10ms. This allowed us to investigate fast changes in the

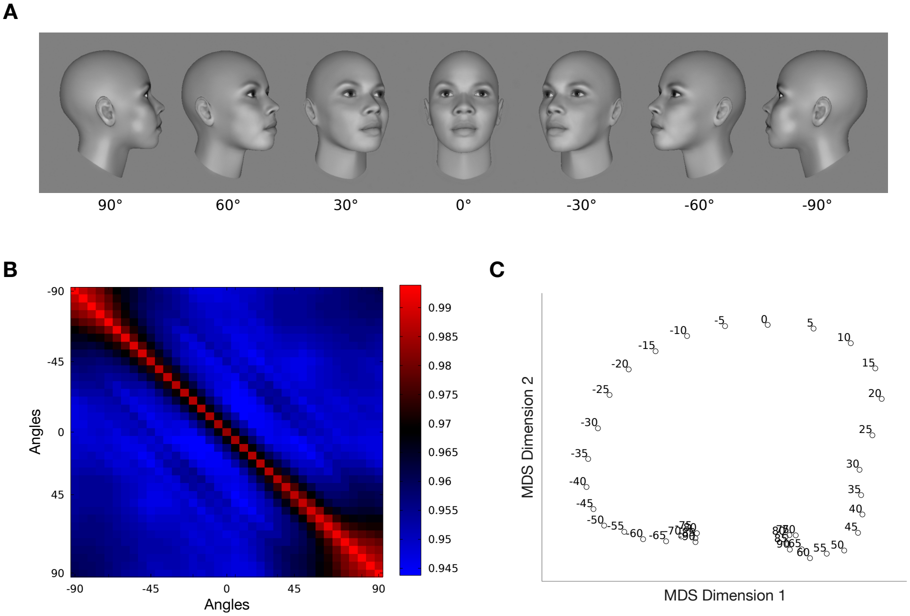

Figure 3.9: Exemplary Stimuli. (**a**) A subset of all stimuli used. In the experiment, 37 angles ranging from -90° to 90° in steps of 5° were used. (**b**) The results of the V1 model, displayed as similarity matrix. (**c**) Same data as in (**b**), displayed in 2D after multidimensional scaling (MDS) was applied.

representational similarity structure across time, indicative of changes in the underlying viewpoint encoding.

Data-driven analyses

To explore the temporal development of viewpoint encoding, we visualized the average similarity matrix at each point in time, and performed an MDS analysis on the data to project it into two-dimensional space. The similarity matrices revealed two prominent, temporally separated effects (Figure 3.10). While the similarity matrices during baseline period were random, as expected, we observed strong effects of viewpoint symmetry around 130ms of processing: mirror-symmetric viewpoints exhibited increased correlations, compared to intermediate viewpoints. At a later point in time, an effect specific to front-on viewpoints of faces was observed. While all other viewpoints exhibited comparable similarity in the evoked response patterns, the cortical responses to front-on views were decidedly different. The same pattern of results can be observed in the MDS projection. The order of conditions during baseline is largely random, whereas the arrangement after about 130ms of processing is almost perfectly ordered with respect to overall rotational angle and furthermore shows close proximity for symmetric viewpoints. At a considerably later point in time, the order of almost all viewpoints seems mostly chaotic with one marked exception: front-on viewpoints, which appear at a large distance to all other viewpoints, including conditions that differ by only small rotation angles, such as ±5°.

The effects observed in the similarity structures are distinctly different to the representational similarity obtained from a model of V1 simple cells, which was visualized in accordance with the EEG data (Figure 3.9b, c). In the V1 model, neighboring viewpoints exhibit strong similarity, which decrease with increasing angular difference. There is an effect of increased similarity for mirror-symmetric viewpoints for more extreme viewpoints, as reported earlier for a similar stimulus set (Kietzmann et al., 2012). However, the effect is only small, compared to the strong effects of viewpoint symmetry observed in the EEG data. Nevertheless, to ensure that low-level similarity was not the driving force behind the observed effects, the V1 similarity model was regressed out of the EEG similarity structures for every participant and time point. After applying MDS to the residual data, we again observed effects of viewpoint symmetry and the previously observed special status of the front-on faces (Figure S3.1). Thus, low-level similarity cannot explain either of the two observed effects.

Model-based Analyses

One of the positive aspects of the representational similarity approach is that it easily allows for different model predictions to be tested against the empirical data. To better understand the dynamics of viewpoint encoding, we therefore extended our data-driven analyses with a model-driven analysis approach. Three models were defined, in close agreement with the predictions of different task requirements: viewpoint symmetry, head orientation, and eye-contact (details of the models are provided in the Materials and Methods section).

The fits of the individual models were tested across time by computing the correspondence of model and EEG similarity matrices for each participant and every time point. To statistically test the resulting model activations, we performed a t-test against zero, the expected model response under the null-hypothesis, across subjects. To correct for the inflated family-wise error introduced by the multiple comparisons performed across time, we performed a cluster-based correction method (Maris & Oostenveld, 2007).

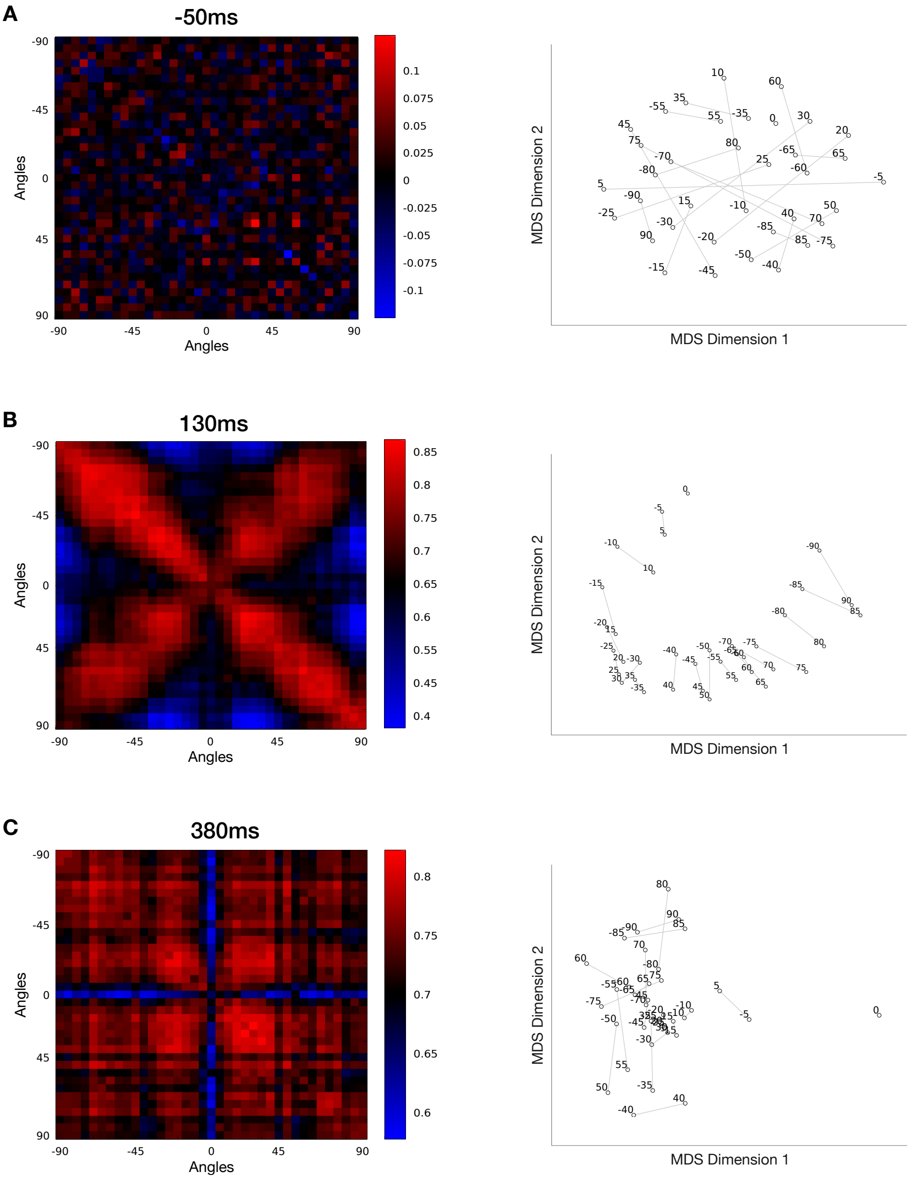

Figure 3.10: Empirical Correlation Matrices and MDS Results. The representational similarity matrices obtained from the high-dimensional EEG data were visualized in their original form and projected into two dimensions using MDS. Panels (**a**), (**b**), and (**c**) show data at timepoints representative of the overall development of the similarity structure during baseline, after 130ms of processing and after 380ms. Symmetric viewpoints are connected via grey lines.

Starting with the head orientation model (Figure 3.11a), which tested whether leftward and rightward head orientations can be differentiated, the analysis revealed multiple significant clusters. The first cluster started around 60ms of processing and remained significant for the whole stimulus duration of 400ms. Two more clusters were found at later latencies (cluster 2 started at 450ms and ended at 550ms, cluster 3 lasted from 570ms to 650ms), potentially mirroring the response to the stimulus offset. These results indicate that information about head orientation is almost constantly present in the visual response. Following this, we tested the EEG similarity matrices for effects of viewpoint symmetry (Figure 3.11b). We observed a slightly later cluster, which started around 100ms, peaked around 120ms and ended around 160ms after stimulus onset. Finally, differences between front-on and directly neighboring views (±5°) were observed considerably later (Figure 3.11c), starting at 280ms of processing and ending around 450ms.

To test in how far low-level properties of the used stimuli had an impact on our experimental results, we again used the results of the V1 model to regress out effects of low-level stimulus similarity from the empirical data, and tested the model fits on the residual data. In line with out previous analyses, the effects of viewpoint symmetry and eye-contact remained significant. However, all effects of head orientation disappeared. Together with the comparably early latency of the effect, this suggests that effects of head orientation are largely driven by low-level stimulus properties in the current setup.

Summing up, our model-based analysis revealed a temporally distinct sequence of viewpoint encoding stages in the human visual system. Effects of left vs right head orientation occurred first, followed by effects of viewpoint symmetry. Contrary to these early effects, differences between front-on and oblique viewpoints, presumably representing effects of eye-contact, were present at later stages of processing.

In addition to low-level similarity, we tested whether residual effects of eye-position, which might have occurred despite the color-change detection task at fixation, would account for the results observed. For this, we estimated two-dimensional probability density functions based on the eyetracking data recorded during the experimental trials. One function was estimated for every subject and condition, leading to an eyetracking representational similarity matrix for each participant (the resulting average similarity matrix and its projection into a two-dimensional space are shown in Figure S3.2). Similar to the control for low-level effects, the individual similarity matrices were used to regress-out effects of different eye-positions from the EEG similarity matrices.

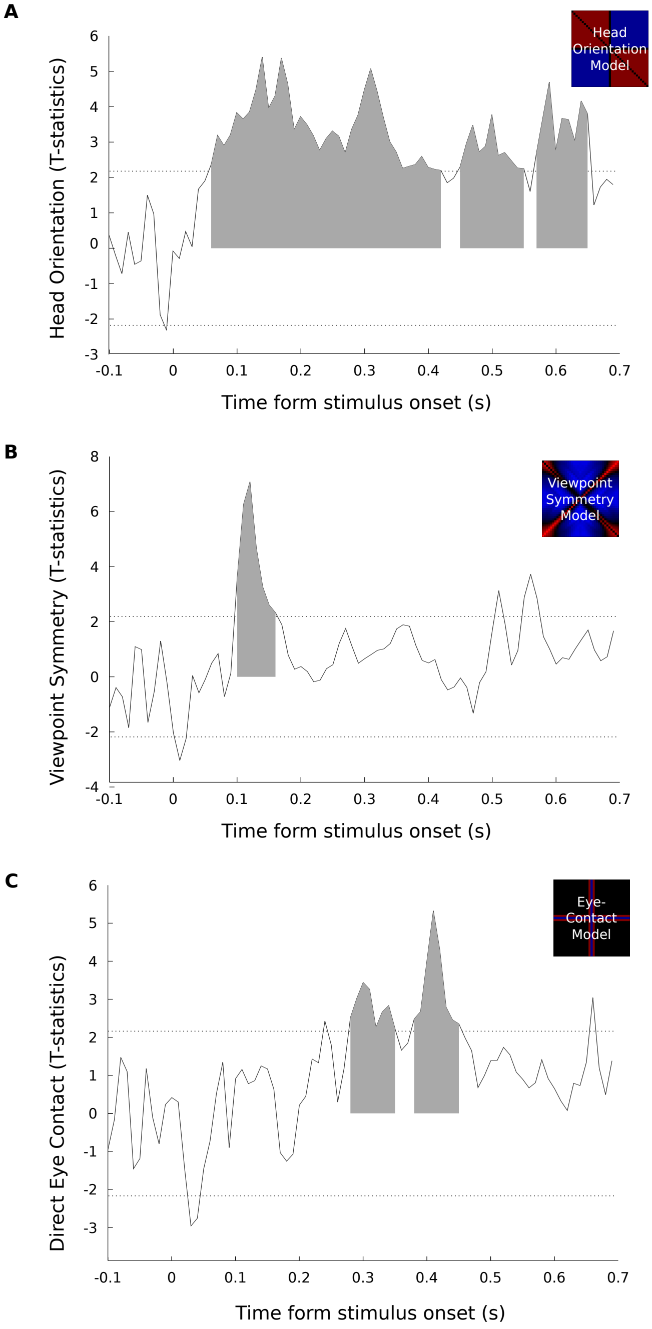

Figure 3.11: Model Results. (for captions see next page)

Figure 3.11: (shown on previous page) (**a**) Results of the Head Orientation Model, which tested whether leftward and rightward head orientations are separable in the visually evoked responses. Shaded areas mark significant effects, corrected for multiple comparisons using a cluster-based correction method. (**b**) The viewpoint symmetry model estimated whether mirror-symmetric viewpoints lead to larger similarity estimates as compared non non-symmetric ones. (**c**) The direct eye-contact model tested whether front-on viewpoints lead to different similarity structures as compared to slightly oblique viewpoints, rotated 5° away from the front-on view.

As before, the residual data was used to fit and evaluate the different models based on a t-test against zero and a cluster-based correction for multiple comparisons. All clusters reported previously remained significant. Residual effects of eye-position are therefore unlikely to contribute to the significance of the found effects.

Finally, to rule out explanations based on differences in attentional load introduced by the distractor task, we tested the behavioral performance of our participants across all viewpoint conditions using a repeated-measures ANOVA. Neither d-prime ($p_{dprime} = 0.42$), nor hitrate or false alarm rate alone ($p_{Hit} = 0.5045$, $p_{FA} = 0.5446$) showed significant effects.

3.3.3 Discussion

Here we investigated the representational dynamics of face viewpoint encoding in the human brain. We recorded EEG data with high temporal resolution and performed a multivariate similarity analysis on the high-dimensional sensor data across time. Our data-driven and model-based analyses revealed a sequence of distinct viewpoint encoding stages in the cortical activation patterns. Effects of head orientation occurred first, followed by effects of viewpoint symmetry, which peaked around 120ms after stimulus onset. At a significantly later processing stage, we observed that front-on viewpoints lead to a different activity patterns compared to all other viewpoints. As the front-on faces are the only ones exhibiting direct eye-contact in the stimulus set, this effect was interpreted as the result of a differential treatment of direct vs averted gaze. Ruling out alternative explanations for the found effects, we showed that viewpoint symmetry and direct eye-contact cannot be explained based on low-level stimulus properties, residual eye-movements or difference in distractor-task performance.

The distinction between left and right head orientation was present early and long-lasting. The short latency of the effect, and the observation that it was strongly affected by a control for low-level stimulus properties suggest that the successful decoding of head orientation was mainly driven by low-level properties. The observation that the effects lasted throughout the whole trial indicates furthermore that low-level stimulus properties remain an important factor in visually evoked responses, even at later stages of visual processing. In the current experiment, head orientation and gaze direction were congruent in all stimuli presented. It is therefore not possible to separate cortical signals related to gaze and head orientation. Evidence in favor of such a separation was provided by Carlin et al. (2011), who demonstrated that the superior temporal sulcus (STS) contains finely graded information about the direction of gaze, independently of head-view and physical image features. This view of a high-level representation of gaze direction is in line with the data from a behavioral adaptation paradigm, in which effects of gaze direction were demonstrated despite changes in size and head direction (Jenkins et al., 2006). Moreover, electrophysiological studies using the same paradigm found late effects, starting around 250ms after stimulus onset (Schweinberger et al., 2007; Kloth & Schweinberger, 2010). In context of the current analyses and results, it would be interesting to combine both methodologies, gaze-adaptation and spatiotemporal pattern similarity, in order to investigate how gaze-adaptation affects the dynamics of viewpoint encoding. Moreover, future work might benefit from the currently used analysis approach to investigate the temporal development and separability of head and gaze direction signals.

Using multivariate analyses of fMRI data, effects of viewpoint symmetry were previously shown to be prevalent across a large range of higher-order visual areas (Kietzmann et al., 2012). In addition to replicating effects of viewpoint symmetry in the human brain, our current findings extend these result by allowing for an estimate of the effect latency. In line with electrophysiological data from macaque monkeys (Freiwald & Tsao, 2010), effects of viewpoint symmetry were found to occur at around 120ms after stimulus onset. This finding constrains possible models of viewpoint symmetry, and rules out explanations based on extended, recurrent processing. Nevertheless, viewpoint symmetry was observed later than effects of head orientation. This result is in line with previous suggestions that viewpoint symmetry might act as an intermediate step in achieving full viewpoint invariance. The latency furthermore directly matches the typically observed latencies of the higher-level visual areas found to exhibit viewpoint symmetry using fMRI.

Finally, our finding of a distinct cortical activation pattern for front-on viewpoints at late stages of processing is potentially driven by effects of direct vs averted gaze. This effect, which is also known as 'eye-contact effect', was previously studied in the context of social attention (Nummenmaa & Calder, 2009) and social neuroscience (Itier & Batty, 2009; Senju & Johnson, 2009). Interestingly, we observed such effects although our participants performed a color-change detection task at fixation. This finding is in line with results of behavioral work, suggesting that effects of direct gaze are present even if attention is drawn away from the face stimulus (Yokoyama et al., 2014). Despite the importance of direct eye contact for social cognitive inferences (Nummenmaa & Calder, 2009), the results of previous electrophysiological studies on the topic are mixed. Whereas one study suggests that effects of eye-contact occur after only 160ms of processing (Conty et al., 2007), others have not found differences between direct and averted gaze at similar latencies (Taylor et al., 2001), in line with univariate analyses performed in developmental studies, which indicate that effects of eye-contact disappear with adulthood (Grice et al., 2005). Finally, data from intracranial recordings suggest that effects of direct vs averted gaze occur during late stages of processing (>200ms)(Pourtois & Spinelli, 2010), in agreement with the current findings. Given the strength of the effects observed here, it would be interesting to revisit studies that produced negative results, and to reanalyze the data based on multivariate pattern analyses.

While front-on viewpoints are different from all other views in terms of direct vs averted gaze, they are also the only stimuli exhibiting reflectional symmetry. This raises the question whether the latter property might have contributed to the effects observed. Indeed, it was recently reported that reflectional symmetry of abstract patterns increases the sustained posterior negativity (Makin et al., 2012). Although previous effects of eye contact were shown based on non-symmetric stimuli (Pourtois & Spinelli, 2010), demonstrating that effects eye contact exist independently of reflectional symmetry, controlling for the effect is an important aspect of future studies to support the current results. One possibility would be to contrast images of front-on faces exhibiting direct eye contact with front-on faces in which the gaze of the shown individual deviates vertically, thereby retaining overall stimulus symmetry but preventing eye contact.

In search for a cortical origin of gaze direction effects, converging evidence from fMRI experiments emphasizes the role of the STS. In addition to the work by Carlin et al. (2011), discussed above, experiments by Hoffman & Haxby (2000) suggest that the processing of facial information is separated between the STS and Fusiform Face Area (FFA). Whereas the FFA was suggested to process information related to facial identity, the STS was found to be involved in the processing of changeable aspects of faces, including the perception of eye gaze (Haxby et al., 2000). In line with this, Calder et al. (2007) used an fMRI adaptation paradigm to study the representation of gaze direction and found that the anterior STS and inferior parietal cortex contain representations that allow for a separation of different gaze directions.

Despite the consistent finding of an involvement of STS in gaze processing, effects of direct vs averted gaze were also shown to alter the functional connectivity between fusiform regions and the amygdala, whereas averted gaze increased correlations between the fusiform cortex and intraparietal sulcus (George et al., 2001). Given these distinct, intra-areal effects, it would be interesting for future studies to exploit the temporal resolution of EEG or MEG in order to test whether effects of oscillatory synchronization (Hipp et al., 2011) contribute to the increased functional connectivity between these areas, potentially binding different features of facial processing.

3.3.4 Materials and Methods

Participants

19 healthy participants (aged 19-29 years, 7 female) took part in the experiment. All subjects had normal or corrected-to-normal visual acuity. They were informed about their right to withdraw from the experiment at any time and gave written informed consent to participate. Participants were paid for their participation. Due to poor task-performance and technical error, three subjects had to be excluded from the analyses.

Stimuli

The stimulus set was created using FaceGen (Singular Inversions Inc). It included four individuals (two female), shown from 37 different angles ranging from -90° to 90° in steps of 5°. The stimuli were presented in greyscale on grey background. The luminance histograms were matched across all faces using the SHINE toolbox (Willenbockel et al., 2010). Because face-selective ERP components were investigated in addition to the analyses presented here, the face models did not include hair, matching the existing literature (example stimuli are shown in Figure 3.10a).

Experimental Setup and Design

Each experimental session consisted of 16 blocks, with 312 trials each. Within each block, each identity and viewpoint was shown twice in pseudo-randomized order, to prevent the direct repetition of identical viewpoints. To keep the participants' attention on the stimulus display, a distractor task was used, which is orthogonal to the experimental question. In each target trial, the color of the fixation cross changed during the presentation of the face stimulus and the task of the subjects was to report these rare events via button presses. 16 target trials were included at random positions in every block. The presentation of the experimental blocks was self-paced by the subjects.

During the EEG measurements, the subjects were seated in a dark room. Stimuli were presented at a 24 inch BenQ screen (Model 2420T) running at a resolution of 1920x1080px and a refresh rate of 120Hz. The latency between EEG trigger and stimulus onset was measured to be 7.5ms. The data shown was not corrected for this delay. The distance to the screen was 80cm, leading to a stimulus display size of $17.7° \times 15{,}7°$ of visual angle. The stimuli were each presented for 400ms, with a random inter stimulus interval (ISI) of 300 to 500ms.

In addition to EEG, we measured the eye movements of our participants using a remote eye tracker (EyeLink) at a sampling frequency of 500Hz. The eye tracker was calibrated prior to the experiment. Recalibration was conducted after every fourth experimental block. The eyetracking error was kept below an average of 0.5° of visual angle.

Data Acquisition, & Preprocessing

Electrophysiological data were recorded using a 128 Ag/AgCl-electrode system by ANT, with electrodes placed on a Waveguard cap according to the five percent electrode system (Oostenveld & Praamstra, 2001). The data were recorded with a sampling rate of 1024Hz and it was ensured pre-measurement that the impedances of all electrodes were below 10kohm.

The data of every subject were preprocessed according to the following procedure. As a first processing step, the data were downsampled to 512Hz. Following this, the data were filtered using a 1Hz high-pass and a 120Hz low-pass filter, and a notch filter around 50 and 100Hz. Channels exhibiting either excessive noise or strong drifts were removed. The remaining, continuous data were manually cleaned, rejecting data sequences including jumps, muscle artifacts and other sources of noise (on average 17% of data were rejected this way). To remove eye-related artifacts, an ICA (AMICA) was computed on the cleaned data and the corresponding independent components were manually selected and removed before transforming the data back into the original sensor space. The initially removed channels were then interpolated based on the activity of their neighboring channels, selected via triangulation. Subsequently, the continuous data were epoched according to the stimulus triggers including data from 100ms prestimulus to 700ms post stimulus, using the time window between -100ms and stimulus onset for baseline correction. As a final preprocessing step, the data were low-pass filtered and downsampled to 100Hz for reasons of computational efficiency. The resulting temporal accuracy of the MVPA analysis is 10ms.

Data Analysis

Multivariate Analyses of EEG Data

After preprocessing the data, we performed a multivariate pattern analysis on the visually evoked potentials. We estimated the similarity between all viewing angles by computing the Pearson product-moment correlation between the respective ERP activation patterns across all 128 EEG channels. The resulting correlation matrices were ordered based on the rotational angle of the shown faces, ranging from -90° to 90°. Neighboring cells in each similarity matrix therefore represent neighboring viewing angles. Applied to every point in time, a spatiotemporal similarity matrix was computed for every subject, which represents the temporal development of representational similarity across all experimental conditions. Although frequently used in fMRI analyses, the representational similarity approach has recently been successfully applied to MEG data with high temporal resolution (Carlson et al., 2013; Cichy et al., 2014).

The representational similarity matrices computed on the EEG data allow for straight forward comparisons with model predictions, such as low-level similarity, control conditions and other behavioral measures (Kriegeskorte et al., 2008a). Here, we performed multiple data- and model-driven analyses on the spatiotemporal correlation matrices in order to investigate the dynamics of viewpoint encoding across time. First, we applied multidimensional scaling (MDS) to the average similarity matrix at each point in time in order to project the data into a two-dimensional space. The resulting two-dimensional arrangement of conditions closely resembles the similarity structure in the original data, but is visually more easily accessible. To account for the temporal smoothness of the underlying data, each MDS optimization was seeded with the MDS result computed for the previous time-point (Carlson et al., 2013). The initial MDS result was based on a randomly selected seed.

As noted earlier, different tasks performed during facial processing require different viewpoint encoding schemes, which provide distinct predictions about the optimal representational similarity structure supporting the respective function. Because of this, we extended our data-driven approach and created three models in agreement with pre-defined task demands. These models allowed for statistical evaluations and latency estimates of the respective effects. The effects investigated included viewpoint symmetry, suggested to support invariant face identification, head orientation (left vs right) and direct eye-contact, tested by contrasting front-on vs slightly oblique viewpoints.

- *Viewpoint Symmetry.* Viewpoint symmetry was suggested to be a key computational step in achieving full viewpoint invariance, which is required for robust face identification. The corresponding model was based on a V1 similarity matrix, described in more detail below. However, to account for the prediction of high similarity across mirror-symmetric viewpoints, the matrix was partly mirrored to exhibit the typical x-shape. The resulting model predicts similar viewpoints and mirror-symmetric viewpoints to exhibit high similarity in the cortical response compared to viewpoints that lack this relationship. The prediction matrix was standardized using a z-transformation.

- *Head Orientation.* A different aspect of face processing is the support of joint attention. In order to be able to differentiate and follow the current direction of another person's attention, the head orientation is a valuable indicator. Effects of viewpoint symmetry, however, suggests that corresponding viewpoints facing left and right are represented similarly. To test whether the viewing direction is nevertheless decodable from the empirical similarity matrix, we tested whether the representational similarity of all viewpoints facing leftward (and rightward) was on average higher than the similarity across left- and rightward viewpoints. Put differently, the model tests whether activity patterns in response to a single viewing direction are more similar than responses to different viewing directions, thereby implicitly testing for the possibility of viewpoint decoding.

- *Eye Contact.* Another aspect of face processing with important implications for social interaction and communication is direct eye contact. To test for this effect in the current dataset, we compared the front-on viewpoint, which exhibits direct eye-contact, to its two neighboring views ($\pm 5^\circ$). The respective model matrix contains negative weights for all similarity estimates involving front-on viewpoints and positive weights for correlations of slightly leftward- and rightward facing viewpoints. All other elements of the correlation matrix were set to zero. The weights in the final model matrix were scaled to sum to zero. Given this setup, the eye contact model tests whether the similarity of front-on viewpoints with all other views is on average smaller than the similarity of slightly oblique viewpoints with all other views.

For all models, elements along the diagonal were not considered. Moreover, the empirical correlation matrices were Fisher-z transformed and standardized in line with the model definitions. To statistically evaluate the model fits, we computed the dot-product of the model matrices and the empirical matrices for every point in time and every subject. Since all models were normalized to a sum of zero, the model fits will on average result in a dot-product of zero given random data. Hence, to test the predictive performance of the models, we tested the model fits against zero, using a t-test across subjects for every point in time.

To control for multiple comparisons performed, one for each time-point, we used a cluster-based permutation test (Maris & Oostenveld, 2007). All connected time-points exhibiting a p-value of $p<0.05$ were considered as empirical candidate clusters. The null-distribution was computed by randomly shuffling the conditions underlying the empirical correlation matrix. This randomization was kept stable across time to preserve the temporal smoothness of the original data. In every bootstrap iteration, we estimated positive and negative clusters to be expected under the null-hypothesis, keeping only the strongest positive and negative clusters (max-sum T-statistic) in the null-distribution. After the bootstrapping procedure, only empirical clusters with a cluster-statistic larger than 2.5% of the positive and smaller than 2.5% of the negative parts of the null-distribution were considered significant.

Control for Low-Level Stimulus Properties: V1 Model

To exclude the possibility of low-level explanations for the found effects, we followed our previous approach (Kietzmann et al., 2012) and computed the low-level stimulus similarity based on a realistic model of V1 simple cells, consisting of a set of 2D Gabor functions with 17 spatial scales and 4 orientations (Serre & Riesenhuber, 2004). Based on the model responses, we estimated the similarity across all experimental conditions, leading to a V1 similarity matrix (shown in

Figure 3.10b). In accordance with the empirical EEG data, the data was projected into a two-dimensional space using MDS (Figure 3.10c). To exclude low-level effects from the EEG similarity matrices, we used the V1 similarity matrix to regress out the low-level stimulus similarity from the empirical similarity matrices of each subject. In line with the previous analyses, we performed MDS (Figure S3.1) and estimated the statistical significance of the model fits on the residual data.

Eye-tracking control

Although our subjects were asked to remain fixated and performed a color-change detection task at fixation, we tested whether eye-movements contribute to the found effects by computing a similarity matrix based on the recorded eyetracking data. For this, we first computed a smoothed, two-dimensional probability distribution (Fixation Density Map) for every subject and condition. Similar to the EEG activity and the V1 model, we then correlated the data of each condition with every other condition. This resulted in a correlation matrix, which describes the similarity structure of the eyetracking data for each individual participant. As a next step, we again projected the average similarity matrix into two dimensions using MDS (Figure S3.2), and used the subject-individual eyetracking similarity matrices to regress out effects of different eye-positions from the EEG similarity matrices. Similar to the V1 model, the residual data was used to estimate and test the significance of the different model responses.

3.3.5 Supporting Information

Supplemental Figures

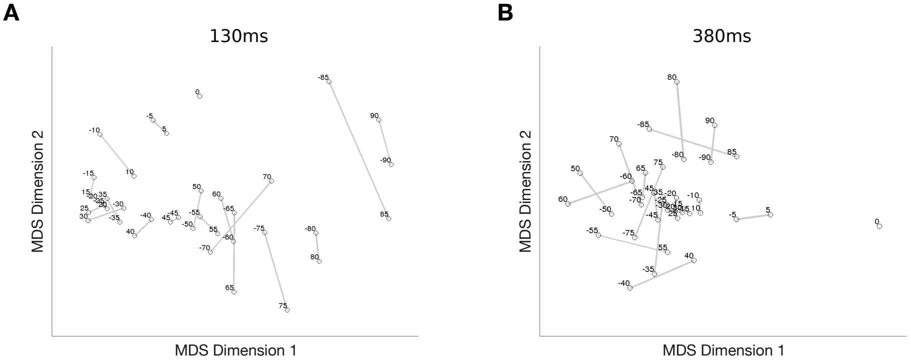

Figure S3.1: MDS Results on Residual Similarity Data. To account for effects of low-level stimulus similarity, the results of a V1 model were used to create a low-level similarity matrix. After regressing out the effects from the EEG similarity matrices, MDS was applied to project the data into a two-dimensional space. Panels (**a**) and (**b**) show the similarity structure 130ms and 380ms after stimulus onset.

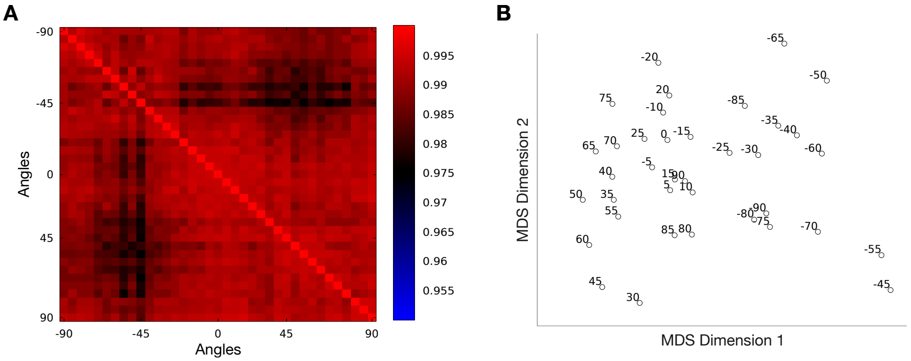

Figure S3.2: Eyetracking Control. Although our participants performed a color-change detection task at fixation, it is possible that residual eye-movements occur. To test whether these data can explain the found effects, we computed similarity matrices based on the average fixation distribution of each subject in each condition. (**a**) The result of the analysis revealed only minor effects which run counter the found symmetry effect observed in the EEG data. (**b**) Same data as in (**a**), but after MDS was applied to project the similarity structure into a two-dimensional space.

3.4 Chapter Summary

Having investigated the interplay of overt visual attention and recognition in the previous chapter, we here explored how invariant recognition is achieved despite significant changes in the retinal projection. In addition to changes caused by eye-movements, there are multiple 'external' reasons for such variation, including, but not limited to, changes in luminance, size and position. Perhaps the most drastic example, however, is based on three-dimensional rotations in depth. To improve our understanding of the neural mechanisms involved in achieving viewpoint invariance, we conducted a series of experiments in which we explored the representation of face viewpoints in the visual cortex. In our analyses, we focused on a potential computational shortcut, viewpoint symmetry, which was previously suggested to constitute a key computational step in achieving full viewpoint invariance in the macaque Freiwald & Tsao (2010). Viewpoint symmetry suggests that mirror-symmetric viewpoints of faces, such as $+60^\circ$ and -60° rotations away from a front-on view, should give rise to similar cortical activation patterns.

In the first experiment we collected fMRI data with high spatial resolution while participants viewed faces seen from different viewpoints. Our analyses revealed that viewpoint symmetry is a widely distributed property of the human visual cortex, including a large variety of higher-level, but not lower-level visual areas. In addition to face-selective ROIs, the list of areas exhibiting viewpoint symmetry also included regions with preferences for other visual categories in the ventral and also areas in the dorsal stream. Following the demonstration of the overall effect, we investigated the underlying mechanism by applying TMS to the right OFA of our participants. This manipulation revealed a causal involvement of OFA in judgments of viewpoint symmetry. Moreover, since the stimulation was applied ipsilateral to the visual presentation, our results suggest that inter-hemispheric interactions play a significant role for viewpoint symmetry. After focusing on spatial and causal aspects of the effect, we estimated its temporal latency by recording EEG data with high temporal resolution in an event-related paradigm. Our multivariate analyses of the EEG data revealed that effects of viewpoint symmetry occur at a comparably early latency of about 120ms after stimulus onset.

Taken together, the prevalence, strength and latency of the effect indicate that viewpoint symmetry might be a general coding principle in the visual cortex. Moreover, the causal involvement of the ipsilateral hemisphere points towards an inter-hemispheric interaction, driving the overall effect. Although our findings have advanced our understanding of the underlying mechanisms, further experimental and computational modeling work will be required to fully understand how viewpoint-symmetry is computed in higher-level visual areas of the brain. As a promising starting point for further investigations, it should be noted that viewpoint-symmetry and effects of mirror-reversals are closely related in stimulus space. Due to the axial symmetry of faces, symmetric viewpoints are almost identical to mirror-reversals. Thus, findings from both fields of research are highly informative of one another and can provide important model constraints. Taken one step further, it currently seems likely that both effects are in fact caused by one and the same cortical mechanism.

4

Categorization and Plasticity

4.1 Perceptual learning of parametric face categories leads to the integration of high-level class-based information but not to high-level pop-out[1]

Abstract To date, the relative contribution of the different levels of the visual hierarchy during perceptual decisions remains unclear. Typical models of visual processing, with the reverse hierarchy theory (RHT) as a prominent example, strongly emphasize the role of higher levels and interpret lower levels as sequence of simple feature-detectors. Here, we investigate this issue based on two analyses. Using a novel combination of perceptual learning based on two classes of parametric faces and a subsequent odd-one-out paradigm, we first test a vital prediction of RHT: high-level pop-out. With this experimental approach, we overcome the low-level confounds of previous studies while still introducing distinct high-level representations. Contrary to previous findings, our

[1] This section was published as peer-reviewed article in the Journal of Vision together with Peter König. See Publication List for details.

analyses show that there is no high-level pop-out, despite very early, near-perfect classification accuracy and extensive training of our subjects. Secondly, we explore the underlying form of category representation during subsequent stages of perceptual training. This is accomplished by including class-external and class-internal target-distractor combinations. Whereas the subjects' responses during the first sessions are best explained instance-based and dependent on low-level metric differences, later patterns exhibit the inclusion of high-level, class-based information that is independent of target-stimulus similarity. Finally, we show that the utilized level of information is highly task-dependent.

4.1.1 Introduction

The search for the cortical mechanism underlying conscious object recognition and classification has been a longstanding focus of scientific interest. Although we are far from a clear picture, one aspect has become central to our understanding: the hierarchical nature of the visual system. Starting at low levels of the ventral pathway, which contains cells with relatively simple receptive field properties, representations become increasingly complex in subsequent levels of the hierarchy. Together with this shift in complexity, representations of the ventral pathway become more abstract and thus invariant to smaller changes in scale, viewpoint and translation (Tanaka, 1996).

Despite the evolving consensus on the hierarchical view of visual processing, it is still an open question how the different levels of the hierarchy contribute to the overall perceptual process. Prominent models of visual processing strongly emphasize the role of high-level representations. That is, lower-levels are interpreted as a sequence of feature-detectors with the purpose to provide input to the higher-levels (Bar, 2004; Serre et al., 2007a). The latter are then associated with the actual perception and are expected to contribute to remote processes including planning and action execution. This view is taken to an extreme by the reverse hierarchy theory (RHT) (Hochstein & Ahissar, 2002; Ahissar & Hochstein, 2004). It argues that visual processing starts with an implicit pass of information through the lower levels of the hierarchy, leading to a first, broad categorization of the input based on the properties of high-level populations of neurons. Then, if required, the high levels access the information present at lower levels, where the receptive field sizes are smaller and more detailed information can be found. Importantly, both types of visual inference attribute the decisive part of processing to high-level representations and view lower-levels as passive feature detectors. This view parallels earlier work ascribing visual awareness to high- rather than lower levels of the visual hierarchy (Crick & Koch, 1995). Recently, however, it has been subject to intensive debate as studies using transcranial magnetic stimulation (TMS) closely linked visual awareness and recognition abilities to activity in V1, the area of lowest-level representational complexity in the cortical hierarchy (Pascual-Leone & Walsh, 2001; Silvanto et al., 2005b; Tong, 2003).

In order to advance our understanding of this issue, we performed two main analyses based on data collected from psychophysical experiments in which we combined perceptual learning with a subsequent odd-one-out task. First, we tested an important prediction of RHT: high-level pop-out. Pop-out is the ability to immediately spot a target stimulus in an array of distractors (spotting the 'odd one out'), independently of the actual number of distractors. Contrary to earlier theories, RHT associates pop-out with high-level neuronal properties. This is based on the view that the initial feed-forward sweep pre-attentively activates abstract, translation invariant representations on high levels, with generalized processing. Based on these properties, the odd-one-out can be immediately spotted. Thus, instead of interpreting pop-out as the source of parallel processing on lower levels (Treisman & Gelade, 1980), RHT explains the effect based on the large, translation invariant receptive fields on higher levels. As a result, features that pop-out are conceived to reflect properties present at high-levels, rather than low-level ones. A direct consequence of this high-level interpretation of the phenomenon is the prediction that also high-level, conceptual information should pop out if high-level representations exist that can differentiate between target and distractor (Hershler & Hochstein, 2005; Hochstein & Ahissar, 2002).

High-level pop-out has been intensively investigated using faces as targets in arrays of distractors taken from other basic level object categories (Figure 4.1). However, despite its conceptual simplicity, the phenomenon is heavily debated (Brown et al., 1997; Hershler & Hochstein, 2005, 2006; Purcell et al., 1996; VanRullen, 2006). While the existence of the overall effect is generally undisputed (a face seems to pop-out of an array of distractors), it is argued that low-level differences can provide a sufficient explanation: although target and distractor belong to different high-level categories, they could additionally differ systematically in features that are explicitly represented on lower levels. These low-level confounds presently prohibit a pure high-level interpretation of the results, as required for RHT.

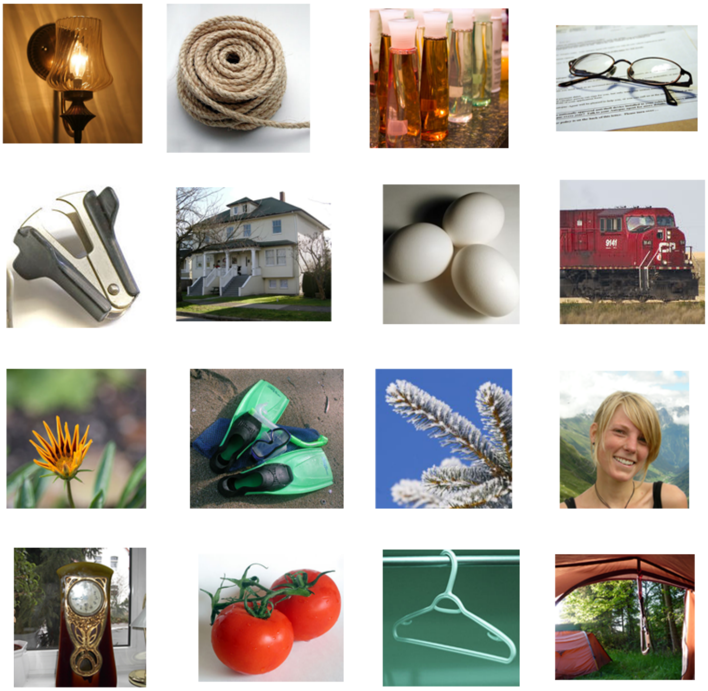

Figure 4.1: A typical stimulus array in which faces pop out among nonface objects. Despite the undisputed existence of the phenomenon, it is still a matter of debate whether it is based on high- or low-level properties in the visual hierarchy.

An unambiguous way of demonstrating high-level pop-out has to meet two conditions. First, the low-level properties of target and distractor need to be controlled, i.e. two categories of stimuli need to be used that are not distinguishable based on simple low-level properties. Second, target and distractor need to belong to different high-level classes, with respective, separate high-level neuronal representations. Recently, Sigala & Logothetis (2002) have illustrated both of these properties for a stimulus set consisting of two subordinate face categories with parametrically varying features. In parameter space, two categories were defined such that only complex conjunctions of feature properties, but no single feature, could be used for categorical decisions. After training macaque monkeys, single cell recordings in inferotemporal cortex (IT) revealed neurons which were specifically responsive to the diagnostic features and therefore discriminated the two arbitrarily defined categories. Using comparable category training, other studies have provided additional support for the existence of high-level representational changes in sensory and prefrontal cortex. Baker et al. (2002) report an enhanced selectivity to learned stimuli of neurons in IT, whereas Jiang et al. (2007) attribute the underlying changes to the lateral occipital cortex as well as prefrontal cortex. As RHT predicts high-level pop-out independently of whether the representations are located in IT or PFC (Hochstein & Ahissar, 2002), all of the above studies show that the training paradigm fulfills both requirements: the two classes of parametric stimuli are similar in low-level features (only complex compositions of these features allow for a categorization), but nevertheless lead to different high-level representations.

In addition to the first analysis addressing high-level pop-out, our second analysis investigates the underlying form of category representation. It provides an estimate on how much low-level metric and high-level class-based information, that is 'stimulus similarity' and 'class membership', contribute to the perceptual decision and is therefore directly related to the question for the relative contribution of the different levels of the visual hierarchy. To be able to distinguish between the two options, we used the same experimental paradigm as in the case of high-level pop-out, but extended it with a condition in which target and distractor were taken from the same category instead of different ones. The reasoning was that if the representations were purely instance-based, then the response times should be independent of the class-membership of the distractor but dependent on the low-level similarity of the two. However, if a more abstract class-membership was used, we should find significant differences between the two conditions. Importantly, since the experimental setup included successive sessions of perceptual learning of the two parametric categories, it is possible to monitor the different modes of category representation during the different stages of training.

Finally, we included a perceptual similarity judgment task and tested the subjects' responses to intermediate stimuli. The recorded similarity data was tested for converging evidence with regard to high-level class effects and utilized as a measure of perceptual similarity in other analyses. The responses to intermediate stimuli are of interest, as they test the results of the pop-out paradigm in a rather different task and can provide further insights into the utilized representation and level of processing that underlie the categorical dissociation.

Based on above considerations, the current study combines several sessions of training of two categories of parametric face-stimuli with subsequent tests for high-level pop-out. This allows for a close monitoring of the subjects' categorization- and high-level pop-out performance during the different phases of learning. The training procedure was kept identical to the one described by Sigala & Logothetis (2002), showing one stimulus at a time and providing audio-feedback indicating the validity of the classification (see Tasks and procedure for more details). In the test phase, a target face was presented together with different numbers of distractor faces, the distractors being all identical. The distractor could either be taken from the same or the other category (class-internal and class-external conditions), as required for the analysis of the underlying form of representation.

4.1.2 Methods

Participants

Four volunteers took part in the experiment and completed 16 sessions each (3 female, 1 male, ages ranging from 23 to 49). All subjects were informed of their right to withdraw from the experiment at any time without the need to state a reason and gave written informed consent to participate. Furthermore, all subjects were informed of the experimental procedure and were naïve to the purpose of the study. Upon completion of the overall experiment the subjects were debriefed.

Stimuli

The experiments were based on parameterized line drawings of faces, as also used in previous studies (Sigala & Logothetis, 2002; Sigala et al., 2002). The faces varied in four dimensions: eye separation, eye height, mouth height and nose length. As can be seen in Figure 4.2b, only the combination of two of the dimensions was diagnostic for the category membership. No single low-level feature could be used to distinguish between the two categories. Five stimuli of each category were used during the standard training and test phases. In addition to these ten, eight intermediate stimuli were created, which were used in a later ’intermediate stimuli’ test (Figure 4.7a). These stimuli were selected to contain all combinations of close/far from the trained instances and close/far from the abstract decision boundary. By this, we expected to be able to gain further insights into the nature of the representation used for the perceptual decisions.

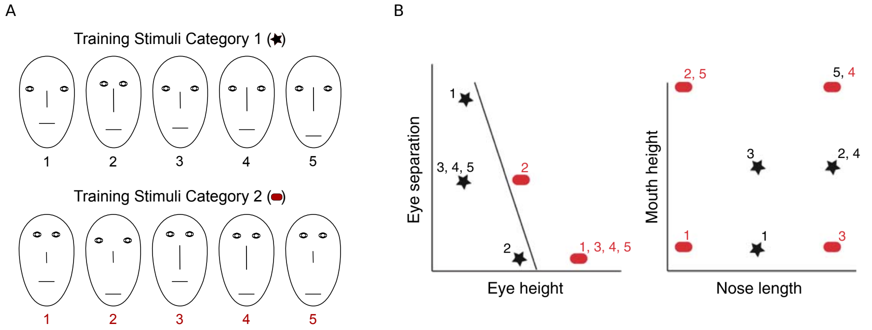

Figure 4.2: The parametric face stimuli. (**a**) The training set consisted of 10 stimuli, five belonging to each class. (**b**) The face stimuli were parametric in four dimensions; no single bottom-up feature can be used to distinguish the two classes. Neither the existence of these dimensions nor the decision boundary was revealed to the subjects. (Figure reprinted by permission from Macmillan Publishers Ltd: Nature (Sigala & Logothetis, 2002), copyright (2010).).

Apparatus

The experiment was conducted on an Apple Mac Pro (4x2.66 GHz, 4 GB Ram) running Linux. The distance to the screen was 60 cm. Stimuli were presented on a 19-inch flat screen monitor (Sync Master 971p, Samsung Electronics, Seoul, South Korea) with a native screen resolution of 1280x1024 pixels and a refresh rate of 75 Hz. Each face was presented at a width of 2° of visual angle, with the entire display being 35.56° x 28.4°. The subject responses were recorded using the arrow keys and num pad of the keyboard.

Tasks and procedure

Each of the 16 experimental sessions contained a training phase and a test phase. In addition, sessions 1 and 16 contained an initial perceptual similarity task and sessions 2 and 15 contained an intermediate stimulus categorization task. Each session was conducted on a separate day, with 3-4 sessions per week. After session 8, there was a two-week pause. Each of the different phases (training, test, similarity and intermediate stimuli) was preceded by on-screen instructions and 5 preliminary trials in which the subjects could get used to the upcoming procedure.

Training Phase

Each session contained a training phase in which the subjects learned the class-membership of the individual faces. For this, the faces were shown separately, and the subjects were asked to indicate, via left or right button press, to which category the presented face belonged. The association of classes and buttons was randomized across subjects. The stimulus presentation ended immediately after the button press. If the answer was correct, a high-pitch feedback tone was provided. If the answer was incorrect, a low-pitch feedback tone was played, followed by a pause of 2 seconds with an empty screen. The pause was used to motivate the subjects to give correct responses and to ensure that the setup was identical to the previous study by Sigala & Logothetis (2002). Specifically, at no point in time the assignment of stimuli to categories was defined or explained. Each training session contained three blocks, with intermediate pauses. Each block contained 240 stimulus presentations, such that each experimental training phase contained 720 trials. The order of stimulus presentation was randomized, and each stimulus was shown equally often in each block.

Test Phase

In order to test for pop-out effects, an odd-one-out paradigm with different numbers of distractors (3, 7 and 11) was used following the training phase of each experimental session. Each trial started with a fixation cross in the middle of the screen. After fixating it, the subjects started the trial by pressing the arrow-up key. In each trial, a target stimulus was shown together with different numbers of distractors. The distractors were either taken from the same category (class-internal condition), or the other category (class-external condition). The distractors were homogeneous, such that each display contained only two different stimuli: one being the target and the other being the distractors. The task of the subject was to indicate via a corresponding button press whether the odd face was present at the left or right half of the display. They were asked to respond as fast and accurately as possible, staying at about 90% accuracy (in line with (VanRullen, 2006)).

In the visual array, target and distractors were randomly positioned on a 4x3 (horizontal x vertical) grid, with 1° spacing between the stimuli and 0.5° random jitter (see Figure 4.3 for an example). The width of the individual stimuli was 2° and therefore identical to the training presentations. The grid-position of the target was selected randomly and it was ensured that an equal amount of stimuli was presented on both sides of the display. In order to allow for reliable

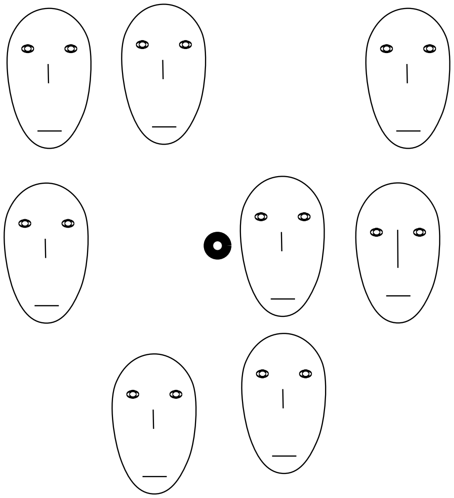

Figure 4.3: The odd-one-out paradigm. A target was presented with different numbers of identical distractors (3, 7, 11). The subjects task was to press a button to indicate in which half of the screen, left or right, the target was shown.

left/right dissociation even in cases with small numbers of distractors, the fixation cross stayed visible in the center of the screen during the complete trial. Moreover, the combination of target and distractor was randomized across trials. For each session, it was ensured that all stimuli were used as target in the same amount of trials. Moreover, each target was shown equally often in the class-external and class-internal conditions. Each test session contained 60 trials.

Similarity Judgment

In addition to the training and test phase, sessions 1 and 16 contained an initial block, in which subjects 2 to 4 were asked to rate the perceptual similarity of two concurrently presented stimuli. The rating was given on a scale from 1 to 5, with 1 being dissimilar and 5 being highly similar. Before each trial, the subjects were asked to fixate a central fixation cross, but they could freely explore the stimuli upon trial onset. Each combination of two stimuli was shown twice, such that each of the two compared stimuli appeared once at each position (left and right of the fixation cross). Each similarity judgment block contained 90 trials in total. No feedback and no category-information were provided.

Intermediate Stimuli

After sessions 2 and 15, subjects 2 to 4 were asked to classify the trained plus previously unseen intermediate stimuli. In parameter space, the latter stimuli were positioned such that all combinations of 'close/far from boundary' and 'close/far from instance' were covered (grey circles in Figure 4.7). Similar to the training phase, the stimuli were presented individually, and the subjects were asked to indicate the category of the shown instance via a left or right button-press, corresponding to the previously learned mapping. However, no feedback was provided. All stimuli were shown twice in randomized order. This resulted in 36 trials for each intermediate-stimuli block.

Data Analysis

Data Cleaning

The analyses were only conducted on valid trials, i.e. trials in which the correct response was given. In order to remove outliers that could occur in cases in which the subjects would pause or talk, only trials with reaction times within 2 standard deviations around the subject-mean were considered. After these two steps, 89.3% of training trials, 79.4% of test trials and 81.5% of trials of the intermediate stimulus test remained for further analyses.

Pop-out analysis

To check whether the performance of the subjects could be interpreted as pop-out, the slope of reaction time per distractor was computed separately for data from the class-external, and the class-internal conditions. In the literature, a reaction time slope below 10ms per item is regarded as pop-out (VanRullen, 2006).

Analysis of the underlying form of representation and implied contributing levels

In addition to the test for pop-out, we used the odd-one-out setup to infer whether the abstract class boundary was used for classification or whether the decision was only based on the similarity to the trained category instances. The first option, a decision using abstract, high-level class information predicts a reaction time difference between class-internal and class-external distractors and an independence of low-level metric differences in the class-external condition. The alternative, a representation that is purely based on instances with their own decision boundaries predicts no differences between the conditions. Specifically, it should not matter whether a target belongs to the same or different category as long as it is a different instance.

Although it was ensured that no single low-level feature could be used as basis for category discrimination, it is still the case that the similarity of stimuli within a category is on average higher than the similarity across categories. Although this does not affect the test for high-level pop-out, which should only occur in the class-external condition, it must to be taken into account before differences between the class-internal and class-external conditions can be attributed to high-level differences because reduced reaction times in the class-external condition could simply be due to the on average decreased similarity between target and distractor. To control for this, a measure of low-level similarity is required. Since the stimuli are parametric, a straightforward option is to use a Euclidean distance metric in the underlying four-dimensional space. In addition to this, we included the subject responses from the similarity judgment task as a measure of perceptual similarity during a later analysis. To check for a reaction time difference in the two conditions, while accounting for low-level similarity effects, we used an analysis of covariance (ANCOVA) with 'class-membership' (external/internal) as main factor, 'stimulus distance' as covariate variable and a two-way interaction between main factor and covariate. For the statistical analysis, the 16 experimental sessions were divided into four analysis blocks, with four sessions each (four ANCOVAs with $p<0.05$, Bonferroni corrected). To verify the validity of this grouping, we conducted an ANOVA for each group with the main factors 'class-membership' and 'session'. None of the four blocks exhibited a significant interaction of session and class-membership ($p>0.45$).

4.1.3 Results

Fast category learning, long-lasting improvements

The analysis of the training performance shows that the two categories are learned very quickly (see Figure 4.4). Already after the first session, subjects correctly classify the stimuli in more than 80% of the cases. After session three, the subjects have successfully inferred and learned the categories and reach more than 95% accuracy. During the following sessions, performance increases even further, reaching 99% accuracy in the mean over sessions 6 to 16. In addition to the near-perfect classification accuracy, the reaction times continue to decrease up to the final session (starting at 772ms in the first session, and reaching 530ms in the last). This indicates continuing improvements until the end of the experiment.

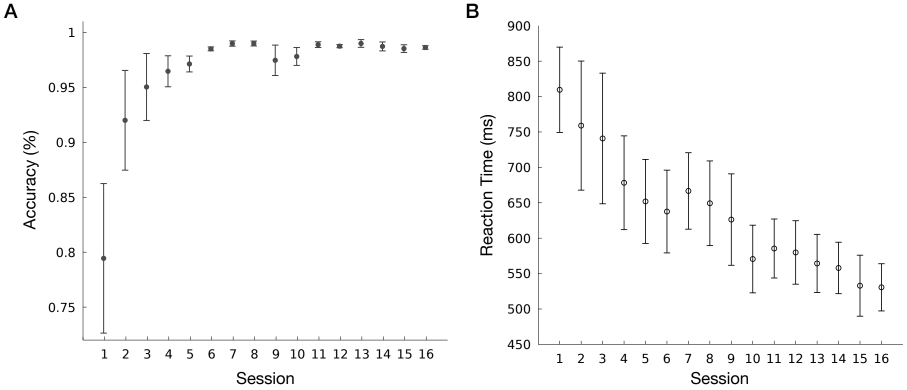

Figure 4.4: Training results. The average training accuracy and reaction times in the different training sessions are shown. The subjects reach near-perfect accuracy very early and their reaction times improve until the end. After session 8, there was a 2-week pause for the subjects.

No high-level pop-out for parametric faces

The high training accuracy indicates that a high-level discrimination is possible rather early. This is in line with the results of the human subjects in the study by Sigala et al. (2002), in which the subjects learned to categorize the stimuli in about 500-1000 trials (Sigala, personal communication). This corresponds roughly to the training phase of a single session in our experiment. Do the implied high-level representations also lead to high-level pop-out effects, as predicted by RHT? For this, the required result would be an independence from the number of distractors for the class-external condition, and a dependence in the case of class-internal targets. As can be seen in Figure 4.5a-d, the reaction times for both conditions, class-internal as well as class-external, increase across conditions and are therefore not independent of the number of distractors. This picture is further clarified by looking at the reaction time slopes (Figure 4.5e). In the literature, values below 10ms per item are interpreted as pop-out. With a mean slope across sessions of 61ms/distractor, the slopes are far beyond this threshold until the end of the experiment. To statistically verify the dependence on the number of distractors, we performed an ANOVA with 'condition' and 'analysis block' as main factors. The results show that both main effects are significant ($p<0.001$), indicating the dependence on the number of distractors and the overall performance improvement with training in the pop-out task. Nevertheless, no significant interaction could be found ($p>0.2$), which implies that the dependence on the number of distractors does not change with training (i.e. there are no significant changes in reaction time slopes with training), despite steadily improving classification performance (see also Figure S4.1). Thus, we do not find any evidence of high-level pop-out.

Late inclusion of abstract category information: when class information becomes a feature

Because the class-internal condition was included in the odd-one-out setup, the recorded data allow for further investigations of the type of representation underlying the perceptual decision. The results of the performed ANCOVA for the different analysis blocks can be seen in Table 4.1. As main factor, class-membership (external/internal) was included and the Euclidean distance between the stimuli in parameter space was taken as covariate. The log-transformed reaction times were included as dependent variable (please note that the resulting patterns of significance remain unchanged when the reaction time data is directly used). Normality and homoscedasticity were verified using the Kolmogorov-Smirnov and Hartley's F_{max} tests respectively. In the first three analysis blocks, the only significant factor is stimulus distance, no significant class-membership and interaction effects can be found. Thus, despite near perfect classification accuracy in the first three analysis blocks, the high-level class-membership does not play a significant role in the odd-one-out task. This allows for the interpretation that the underlying form of representations is instance-based and that the metric-based target-distractor similarity is the important aspect governing performance.

Table 4.1: Significance results of the ANCOVA analysis. In blocks 1-3, only the covariate "distance" is significant, in line with an instance-based interpretation. In block 4, the group membership (class-internal vs. class-external) becomes a significant factor. Finally, there is a significant interaction effect, indicating a different slope of the linear regression fit in the different group-membership conditions.

	Block 1	Block 2	Block 3	Block 4
Main effect: Group	$p > 0.10$	$p > 0.10$	$p > 0.10$	$p < 0.01$**
Main effect: Distance	$p < 0.01$**	$p < 0.01$**	$p < 0.01$**	$p < 0.01$**
Interaction effect	$p > 0.10$	$p > 0.10$	$p > 0.10$	$p < 0.01$**

In analysis block 4 (corresponding to the final four experimental sessions), things change. In addition to the significant effect of stimulus distance, the main factor class-membership and the interaction between the distance and class-membership become significant. First and foremost, this indicates that, after

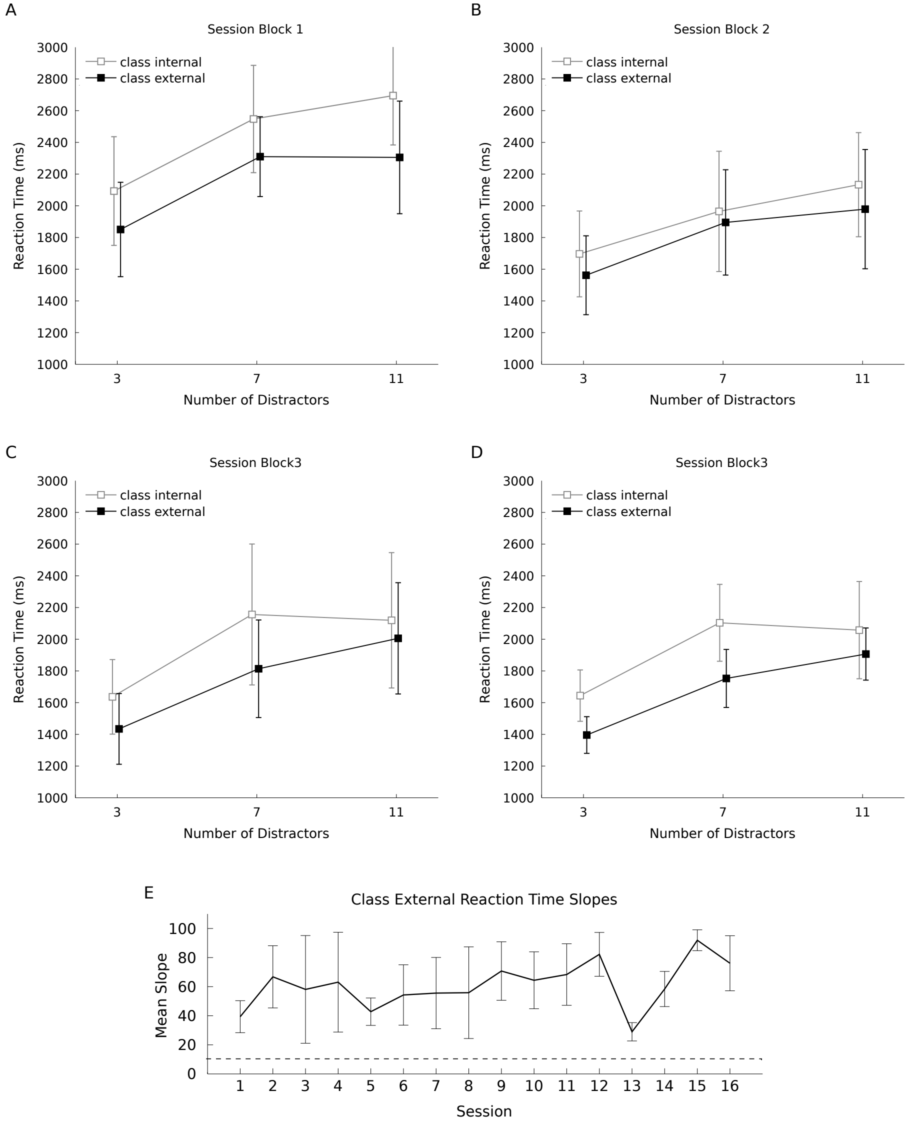

Figure 4.5: Pop-out results. (**a-d**) The reaction times in the different conditions, with targets being either class internal or class external. (**e**) Mean slope for class-external targets across sessions. The resulting slopes are far from the threshold of pop-out at 10 ms/item, indicating that no pop-out occurred during the experiment.

correcting for low-level similarity, a significant difference in reaction times in the pop-out paradigm exists, depending on the relative class-memberships of target and distractor. On average, spotting the target is significantly faster if it is taken from a different category than if it belongs to the same category as the distractor (mean $RT_{internal} = 1947ms \pm 44ms$ SEM; $RT_{external} = 1694ms \pm 30ms$ SEM). Still, due to the significant interaction effect, which implies that the dependence on target-distractor similarity differs in the two class-membership conditions, it is important to also analyze the resulting regression models (Figure 4.6). In the current case, the slopes of the regression functions depict the dependence of reaction times on low-level similarity. The underlying data is pooled across all conditions and therefore provides information that cannot be inferred from the previous pop-out analyses and figures. In blocks 1 to 3, the two functions have a negative slope, indicating that increased stimulus similarity (decreased stimulus distance) correlates with increased reaction times. This shows the reaction time dependence on metric-based target-distractor similarity. Also, the two lines are mostly identical in slope and intercept, which would be expected from an instance-based decision process that is ignorant to abstract, high-level class-information. In block 4, the class-external fit is mostly below the class-internal fit. This is the visualization for the significant main effect of class membership. Moreover, the slope of the two fits is different, visualizing the interaction effect. While the slope of the class-internal fit remains negative, indicating an unchanged dependence on target-distractor similarity, the class-external fit becomes rather flat. This shows that, if the target is taken from a different class than the distractor, the dependence of the reaction times on stimulus similarity is significantly reduced. Please note that this is not a sign of pop-out, which is defined as an independence of reaction times and the number of distractors. Instead it signifies an independence from the metric-based target-distractor similarity and is a sign of a high-level distinction. Finally, the difference between the class-internal and class-external conditions is most prominent in cases of high similarity (small Euclidean distances between target and distractor). If target and distractor are very dissimilar, there is no difference between the two conditions. In order to exclude the possibility that the found class membership effects are due to the similar motor-responses during the training and pop-out parts, we performed an ANOVA with the factors 'class membership' and 'congruency'. Congruent trials were the ones, in which the target was presented at the side, which was also associated with the categorical response during training. As expected, there was a significant effect of class-membership ($p<0.001$), but no significant main effect of congruency ($p>0.5$) and no significant interaction-effect (p>0.5). This indicates that the found class-membership effects did not originate from the selected type of motor-responses.

Thus, despite not finding any sign of high-level pop-out, the results of the ANCOVA argue that, after extensive training, the abstract class information becomes a vital part in the perceptual decision - a change in processing, which can also be interpreted as an effect of a developing expertise of the subjects. Importantly, the class-information is only used if applicable, whereas a dependence on low-level similarity remains, if the class-feature is not expressive.

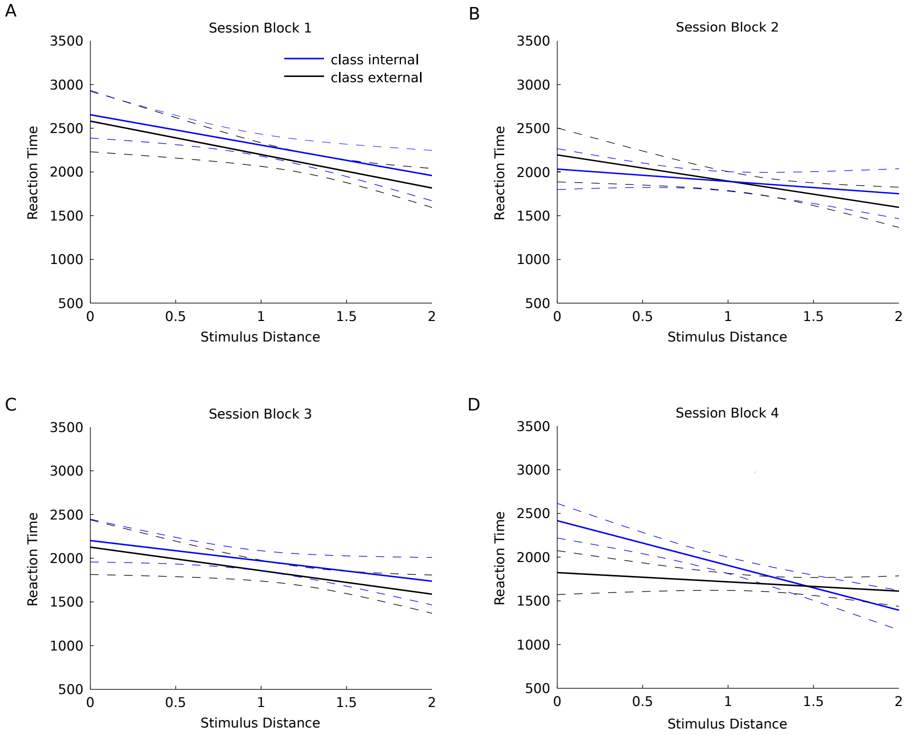

Figure 4.6: Linear regression fits. The panels show regression models created for the four analysis blocks. With a 4-dimensional, normalized parameter space, the maximal distance between stimuli is $\sqrt{4}$. During blocks 1-3, the fits for the class-internal and class- external data are highly similar. The negative slope is a visualization of the significant effect of stimulus distance. In block 4, the slope of the class-external fit becomes less steep, indicating a decreased dependence on low-level similarity. When target and distractor are from the same class, however, the low-level similarity is still a significant factor. Finally, in cases of high stimulus similarity, the fit for the class- external condition is below the fit for the class-internal condition.

Similarity judgments underline the found high-level effects

The results of the ANCOVA indicated that high-level class information is an integrated part of the perceptual process. From this, it can be expected that it also influences the subjective judgment of perceptual similarity. To test this hypothesis, we asked the subjects before sessions 1 and 16 to judge the similarity of two concurrently presented stimuli. With training, the class-external similarity judgment decreased significantly ($s\_ext_{early} = 2.71 > s\_ext_{late} = 2.38$; p<0.01 Wilcoxon paired Test). For the judgment of the class-internal data, a reverse tendency could be found ($s\_int_{early} = 3.38 < s\_int_{late} = 3.58$; p=0.118 Wilcoxon paired Test). A decrease in similarity for class-external judgments and an increase for class-internal judgments is an implicit prediction of an integrated high-level class-feature in the perceptual process. To ensure that the similarity test had no effect on the later pop-out performance, we compared the accuracy and reaction times of the three subjects who performed the similarity test with the data from the subject for which the test was not included. Neither the accuracy nor the reaction times differed significantly indicating that the similarity judgment did not affect the later pop-out performance.

As a measure of low-level stimulus similarity, we so far used Euclidean distances in parameter space, which assign identical weights to the different parameter dimensions. This selection is valid as a measure of physical differences between stimuli. With regard to perceptual differences, however, it cannot be excluded that the subjects do not attribute equal weights to the different parameter dimensions. To exclude this possibility as an explanation for the found effects, we reanalyzed the data collected during the test phases of analysis blocks 1 and 4 and used the average perceived similarities for the stimulus pairs, recorded during the similarity judgment task, as a measure of stimulus distance. The results of the ANCOVA are similar to our previous analyses and conclusions. During analysis block 1, only the perceived stimulus similarity had a significant effect on the reaction times, whereas analysis block 4 exhibits significant effects of class-membership and a significant interaction. Again, the underlying model fits showed that the dependence on stimulus similarity was significantly reduced in the class-external condition as compared to the class-internal condition.

Intermediate stimuli exhibit a mixed mode combining exemplar and high-level effects

Using the pop-out paradigm, we compared the levels of processing underlying two distinct cases: class-internal and class-external targets and distractors.

Table 4.2: Reaction times and accuracies for the different intermediate stimuli depicted in Figure 4.6. In session 2, the accuracy seems to be dependent on the proximity to the training instances. In session 15, the only misclassified stimulus is far from a training instance and close to the decision boundary. The accuracy patterns are therefore in line with the reaction time results.

		Session 2	Session 15
Category 1	1	681.9 (100%)	696.4 (100%)
	2	629.2 (100%)	406.6 (100%)
	3	2382.7 (67%)	917.7 (83%)
	4	726.0 (100%)	712.8 (100%)
Category 2	1	2333.8 (83%)	559.5 (100%)
	2	965.9 (83%)	457.3 (100%)
	3	1171.0 (100%)	769.0 (100%)
	4	470.7 (100%)	488.9 (100%)

To address whether similar results can also be observed during the classification of singular stimulus instances, we presented intermediate stimuli to the subjects during an early and a late session. These stimuli were explicitly selected to test for an instance- vs. boundary-based form of perceptual decision and were positioned accordingly. Figure 4.7 shows the resulting reaction time patterns (the reaction times and accuracies are presented in textual form in Table 4.2). In session 2 (Figure 4.7b), the longest reaction times correspond to stimuli far from the category instances (1, 3, 1, 3), whereas the ones close to the trained stimuli (2, 4, 2, 4) are classified faster. This is in line with the significant factor distance and non-significant class-membership results of the ANCOVA analysis in the pop-out paradigm. After session 15 (Figure 4.7c), an additional effect can be observed. Now, not only the proximity to the trained instances, but also the distance to the decision boundary becomes a relevant factor. If stimuli are close to a training instance (2, 4, 2, 4), or far from the decision boundary (1, 4, 1, 4), they are classified faster. If the tested stimulus is close to the boundary and far from the training instances it tends to be classified more slowly. This change of pattern can best be observed in stimuli 1 and 3. After training, stimulus 1, being far from the decision boundary, is classified faster, whereas stimulus 3, close to the boundary, is classified slower than before. A similar pattern can be observed in the accuracy results (Table 4.2). In session 2, most of the errors are being made for stimuli that are far from the training instances, whereas the only errors in session 15 are made for stimulus 3 of category 1, which is far from the training instances and close to the decision boundary.

The late 'intermediate stimulus' response patterns are therefore best described as a 'logical or' interaction in which low-level metric and high-level class-based information play a joint role in the decision process. These results show a development from an instance-based to an integrated high- and low-level form of representation for the classification of singular, intermediate stimuli.

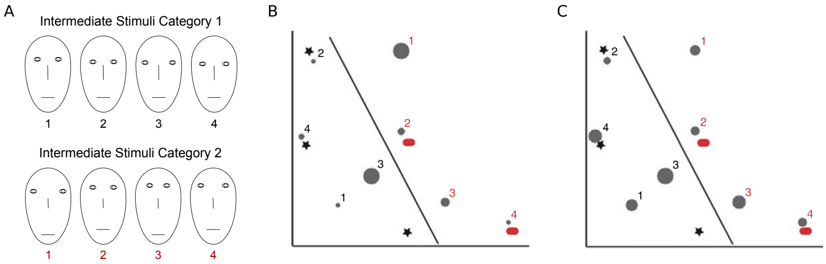

Figure 4.7: Results of the intermediate stimuli test. (**a**) For both categories, additional stimuli were created at intermediate positions in parameter space. Stimuli corresponding to all combinations of close/far from trained instances and close/far from an abstract decision boundary were created. In (**b**) and (**c**), the black stars and red ovals indicate the positions of the training instances. The centers of the gray circles represent the position of the intermediate stimuli. The diameters of the circles encode the mean reaction times of the subjects, normalized in each of the two plots. In addition to the stimulus instances, the linear decision boundary is shown. The boundary is an abstract description of the two classes that is independent of the specific class instances selected for training. (**b**) Reaction times after session 2, resembling an instance-like pattern (low-level metric-based). (**c**) Reaction times after session 15 exhibit a "logical or" combination of low-level metric and high-level class-based effects.

4.1.4 Discussion

In this study, we addressed two important aspects with regard to the relative contribution of different levels during perceptual processing. Our analyses are based on data collected in an experimental paradigm, which combines perceptual learning of parametric stimulus-categories with an odd-one-out task. First, we investigated the existence of high-level pop-out in a controlled setting and found that, despite the extensive training and performance improvements until the end, search slopes remained dependent on the amount of distractors, i.e. we find no evidence for high-level pop-out. Notably, our odd-one-out setup was based on homogenous distractors, which are typically expected to simplify the search and hence favor pop-out (Hershler & Hochstein, 2006).

In response to earlier studies that did not find pop-out for faces (Brown et al., 1997; Kuehn & Jolicoeur, 1994; Nothdurft, 1993), Hershler and Hochstein note that the distractors used were too similar to the target. This way, they argue, one single high-level representation was activated, which could not differentiate target and distractors after the initial feed-forward sweep of information (Hershler & Hochstein, 2005). This 'one population' argument does not apply here, since it was previously shown that the same training setup as the current one lead to distinct high-level representations responsive to the two categories[2]. Additionally, the argument introduces a dependence on low-level dissimilarity between target and distractor, which could again act as explanation for the found effect.

An additional argument in favor of high-level pop-out could be that the perceptual training was not long enough in order to lead to well-separated neuronal representations. For two reasons, this argument does not apply. First, the number of training trials for the fish-stimuli in the study conducted by Sigala & Logothetis (2002) was below 10000 and therefore smaller than the one in the current experiment (Sigala, personal communication). Still the neuronal effects could reliably be shown. Second, a recent study by Hershler & Hochstein (2009) tested car and bird experts for high-level pop-out. Despite years of training, pop-out was only visible for faces, but not for target-stimuli corresponding to their area of expertise (birds or cars). This is a particularly strong case, as comparable experts were previously shown to exhibit significant activation of FFA upon perceptual decisions involving these types of stimuli (Gauthier et al., 2000a).

In our setup, we avoided the low-level confounds of earlier studies, by utilizing two subordinate stimulus categories which cannot be classified on the basis of simple low-level features but only via feature conjunctions. Additionally, perceptual training based on these stimuli was previously shown to give rise to distinct high-level representations. Therefore, they fulfill both requirements of high-level pop-out according to RHT. For future studies, it might nevertheless be interesting to test two basic-level categories of parametric stimuli, with higher (but still similar) inter- and intra category differences.

With the second major analysis, we addressed the underlying form of category representation and implied contributions of the different levels used during the perceptual decisions. Here, we found a shift from an early low-level instance-based to a late, abstract, and class-based form of representation. For the latter, we demonstrated a dependence on stimulus similarity if target

[2] This presupposes that similar representations are present in monkey and human upon identical training procedures. Evidence for the validity of this assumption was provided by Sigala et al. (2002) who showed that monkeys and humans follow the same categorization strategies in the currently used task.

and distractor are taken from the same category, whereas response times were shown to be significantly less dependent of low-level differences if the high-level class-membership of target and distractor is different. Again, this is no case of high-level pop-out, as this would require independence of the number of distractors which is never the case. These results show that, dependent on the available information, different modes of processing are required and used for the perceptual decision.

With regard to an early instance-based representation, as reported here for the first three analysis blocks, converging evidence comes from a recent fMRI-adaptation study that investigated response properties in the lateral occipital cortex (Gillebert et al., 2009). Before the fMRI measurements, the subjects were trained to categorize two classes of parametric stimuli in a paradigm comparable to ours. The subjects were trained for two sessions, 50 minutes each, which is roughly equivalent to our first analysis block. As measure, they used the amount of BOLD adaptation resulting from the presentation of two subsequent stimuli that could either be taken from the same or different categories. The analyses show that there is no significant difference between their class-internal and class-external conditions. In line with more recent work by Sigala, this indicates that the early form of representation is best explained instance-based (Gauthier & Palmeri, 2002; Sigala, 2004). This is in agreement with our analyses. However, one prediction of our class-membership effect in analysis block 4 is that prolonged training should result in a significant difference in adaptation based on class-internal vs. class-external stimulus presentations.

Converging evidence for the integration of categorical information into the perceptual process is provided by the similarity judgment task. With training, the perceived similarity of stimuli from the same class increased whereas the similarity judgment across categories decreased. These results are in line with earlier studies describing the effect of categorization on perceived similarities (Goldstone, 1994; Goldstone et al., 2001)[3]. Moreover, by using the recorded perceptual similarities as a measure of stimulus distance, we could show that the found high-level class effects could not be explained by a differential weighting of dimensions in parameter space. The question of whether the found high-level representations are based on an additionally learned decision hyperplane or rather on the additional storage of appropriate prototypes is beyond the scope of the current study (please see Sigala et al. (2002) for an argument against a

[3] It should be noted that similarity tests involving the judgment of categorical stimuli cannot exclude the possibility that subjects explicitly judge stimuli as being more similar if they realize that the stimuli belong to the same category (Goldstone et al., 2001).

representation based on single prototypes). In any case, both approaches form an abstract representation of the learned categories.

In line with the analyses of the odd-one-out paradigm, the results of the intermediate stimulus classification task showed that early reaction time patterns resemble an instance-based form of category representation with a dependence on the training exemplar distances. However, the later response patterns were shown to resemble a 'logical-or' mixture of both, metric- and class-based information. This further exemplifies the task-dependent contribution from levels of different complexity during the late stage of training. In pop-out, if high-level information is expressive it is used exclusively, whereas low-level target-distractor similarity is predominant in the class-internal condition. In the intermediate stimulus categorization task, in which no distractors were present, we find a mixed mode integrating both types of information.

Having found the two modes of perceptual processing in the pop-out task, being either based on target-distractor similarity or more abstract class-information, the resulting question is which physiological levels in the visual hierarchy might be underlying these effects. Based on previous results in the literature, we expect the abstract class-information to originate from high-level visual areas (Sigala & Logothetis, 2002; Kiani et al., 2007) or from the lateral prefrontal cortex (LPFC) (Freedman et al. 2003, but see Minamimoto et al. 2010). Please note that this distinction does not matter for the predictions of RHT, as both areas were described as being potential sources of the categorization during the initial vision at a glance (Hochstein & Ahissar, 2002). The dependency on stimulus-distance, however, is directly related to metric-based target-distractor similarity. This type of effect is expected to originate from lower-levels of processing, which are mostly stimulus-driven and contain the relevant and more detailed stimulus information. However, a clear decision cannot be based on psychophysical measurements, especially in the case of long perceptual training, and future work is required to fully address this issue.

In our experiments, no high-level pop-out could be observed, despite the near-perfect classification performance of our subjects. Instead, we have found clear evidence for an integrated use of low-level metric- and high-level class-based information, dependent on the task and the information content of the individual levels of representation. As a result of these findings, we suggest that the visual hierarchy only provides a means for increasingly complex and non-linear response properties, or modules, whereas the evaluation of the resulting representations is highly distributed. Although a hierarchical view on visual processing might be applicable with regard to the representational complexity, it does not automatically imply an increasing level of perceptual importance with subsequent levels of processing. The proposed view is in line with the results of recent experiments, which underline the importance of lower level representations during perceptual decisions (Tong, 2003; Kamitani & Tong, 2005; Silvanto et al., 2005a)

4.1.5 Acknowledgements

The authors gratefully acknowledge the support of the Niedersächsisch-Israelische Gemeinschaftsvorhaben "Does the study of simple visual stimuli assess the primitives of natural vision" (Prof. König, Prof. Ahissar), and the Research Training Group "Adaptivity in Hybrid Cognitive Systems" of the Institute of Cognitive Science at the University of Osnabrück. Furthermore, we would like to thank Frank Jäkel, Niklas Wilming, Alper Acik, Orit Hershler, Konrad Körding and Merav Ahissar for helpful discussions and their comments on an earlier version of the manuscript. Finally, we would like to thank Natasha Sigala for valuable advice and help with the stimulus set.

4.1.6 Supporting Information

Supplemental Figures

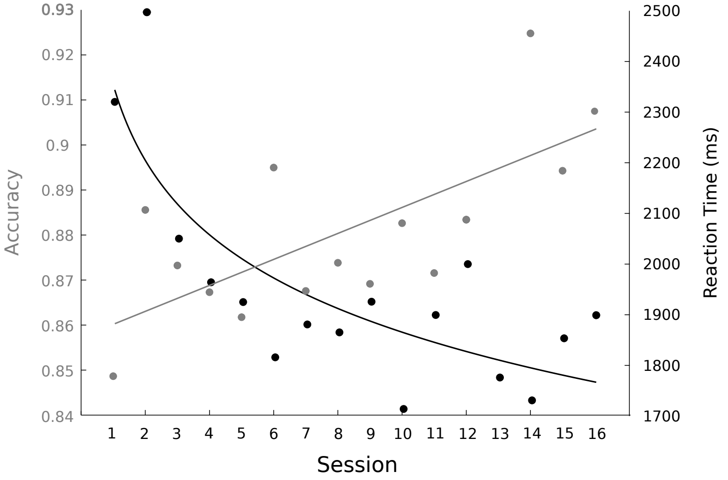

Figure S4.1: Odd-one-out Training Effects. Across training sessions, the performance in the odd-one-out paradigm increased as evident through increasing accuracy (shown in grey), and decreasing reaction times (shown in black).

4.2 Extensive Training Leads to Temporal and Spatial Shifts of Cortical Activity Underlying Visual Category Selectivity

Abstract The human visual system is able to distinguish naturally occurring categories with exceptional speed and accuracy. At the same time, it exhibits substantial plasticity, permitting the seamless and rapid learning of entirely novel categories. Here we investigate the interplay of these two processes by asking how category selectivity emerges and develops from initial to extended category learning. Using an MEG adaptation paradigm, a novel filter approach for analyses of visually evoked responses and source analyses we demonstrate a spatiotemporal shift of cortical activity underlying category selectivity. After initial category acquisition, the onset of category selectivity was observed after 275ms together with stronger activity in prefrontal cortex. Following extensive training, the earliest category effects occurred at a markedly shorter latency of 113ms and were accompanied by stronger occipitotemporal activity. Our results suggest that the brain balances plasticity and efficiency by relying on different mechanisms to recognize new and re-occurring categories.

4.2.1 Introduction

One of the most essential tasks of our visual system is to make sense of the complex signals it receives from the world around us. A central aspect of this is the ability to group objects into various categories, allowing for considerable simplification, generalization and supporting higher cognitive function.

To advance our understanding of the underlying cortical mechanisms, a large body of experimental work focuses on temporal aspects of category selectivity, asking for the earliest point in time at which category information is extracted. As a result, we now have ample psychophysical and electrophysiological evidence that naturally occurring categories can be extracted in only little more than 100ms of processing (Hung et al., 2005; Sugase et al., 1999; Carlson et al., 2011; Liu et al., 2002, 2009; Kirchner & Thorpe, 2006). However, apart from the necessity for fast and robust categorization of re-occurring categories, our ever-changing environment poses the additional challenge to retain considerable plasticity in order to support the rapid learning of entirely novel categories

(Summerfield et al., 2011). Here, the study of naturally occurring categories provides only limited possibilities, as it focuses on categories with which we already have extended experience (for instance, all of us can be considered face- and house-experts, as these categories play a vital role in our everyday behavior). It therefore remains an open question, how cortical representations and the temporal dynamics of category selectivity develop from the initial learning of a category towards category expertise.

To elucidate this issue, we performed a longitudinal study with nine subjects, in which we investigated the impact of extended category training of two artificial visual categories in a parametric feature space on the visually evoked responses using a magnetoencephalography (MEG) adaptation paradigm. Combining perceptual category training and an MEG adaptation paradigm provides multiple advantages. First, the data recorded during a baseline session allowed us to exclude the possibility that potentially found category effects are an inherent low-level property of the utilized feature space. Such differences in low-level statistics have previously lead to considerable challenges in the interpretations of category effects in studies using naturally occurring categories (Thierry et al., 2007; Rossion & Jacques, 2008; Crouzet & Thorpe, 2011; VanRullen & Thorpe, 2001). Second, MEG offers high temporal resolution, which enabled us to resolve changes in the temporal aspects of category processing, indicative of changes in the underlying cortical mechanism. Finally, the relatively poor spatial resolution of MEG can be bypassed with the use of adaptation paradigms (Grill-Spector & Malach, 2001), which have the potential to reveal differences in neuronal selectivity that would otherwise remain unnoticed in more traditional analyses of amplitudes in evoked fields.

We investigated the emergence and development of category selectivity by recording MEG data in a baseline session, prior to any category training, a second time after five training sessions, and a third time after extensive category training in 22 training sessions. Category selectivity was estimated by comparing the visually evoked responses to stimuli that were either preceded by a different stimulus from the same category (category-internal), or by a stimulus of a different category (category-external), while holding low-level stimulus differences constant.

Using this approach, we observed a temporal shift in category selectivity from a latency of 275ms after initial category acquisition to only 113ms following extensive training. This speedup suggests a marked change in the cortical network mediating the categorization of visual input. Indeed, source analysis revealed an anterior-to-posterior shift of cortical activity from initial to extensive category training. While the time-window of category selectivity found after five training sessions exhibited stronger activation in the prefrontal cortex (PFC), the early category effects found after 22 training sessions showed increased activation in occipitotemporal regions.

Previous theories on visual categorization viewed either PFC or regions in the ventral stream as the origin of category selectivity. Our findings now reconcile these contrasting views by suggesting that both processes are likely to contribute to categorization at different stages of category learning. While PFC is involved in the categorization of rather novel, and dynamic categories, extensively used categories seem to obtain a privileged status and are resolved faster relying more heavily on cortical resources in occipitotemporal cortex.

4.2.2 Results

Behavioral Data on Category Training

Subjects were trained to categorize two artificial categories of faces defined in a parametric feature space (Figure 4.8). Training lasted for a total of 22 sessions consisting of 756 training trials each. Classification accuracy reached 89.4% after five training sessions, and 95.3% after training was completed in session 22 (Figure 4.9). At the same time, reaction times continued to decrease with training (from 679ms in session five to 538ms in session 22, $p<0.01$ paired t-test). Thus, although high classification performance was reached already after five training sessions, the behavioral data indicate continued improvements over the whole training period.

Behavior during the Delayed Match-To-Category Task

The electrophysiological correlates of category effects were estimated using an adaptation paradigm in which subjects performed a delayed match-to-category task (Figure 4.10a). During the baseline session, and therefore prior to category training, the behavioral performance of our subjects did not differ significantly from chance (49% accuracy, $p=0.128$, t-test against a chance-level of 50%). This demonstrates that our artificial category structure is not an inherent property of the stimulus space. With training, performance increased to 66.2% in session five and 76.0% in session 22 (Figure S4.2). A repeated measures ANOVA (with session (baseline, five, 22) and category membership (internal, external) as factors) revealed a significant main effect of session ($p<0.01$, all pairwise comparisons are significant at $p<0.01$, t-test, Bonferroni corrected),

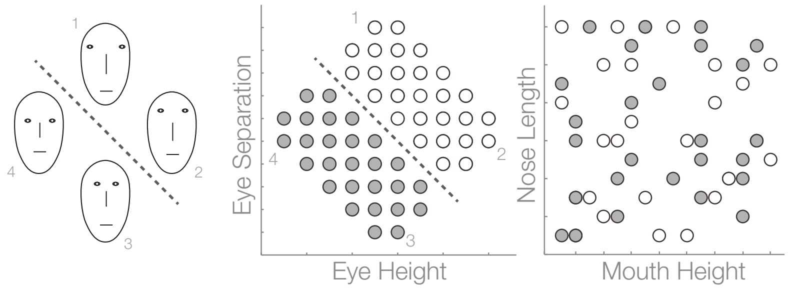

Figure 4.8: Stimulus Space. Subjects were trained to distinguish two artificial categories of faces, defined in a four-dimensional parametric feature space. Two dimensions were category-relevant (eye height and eye separation), and two were irrelevant (nose length and mouth height). No single feature was decisive for the category of a given face, only the combination of features allowed for correct categorization. The category boundary was rotated by 90 degrees for every other subject.

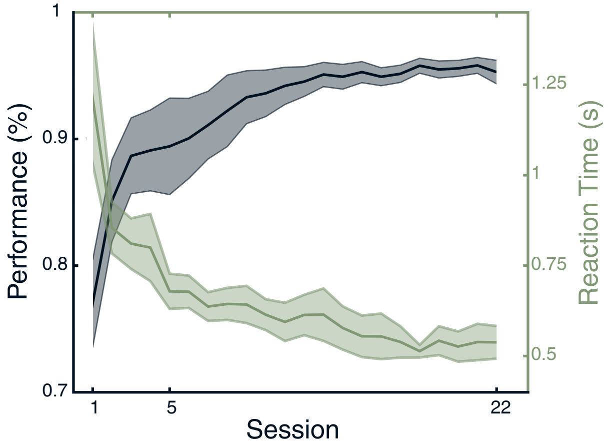

Figure 4.9: Effects of Training on Performance and Reaction Times. Subjects received auditory feedback as a training signal, but no explicit information about the underlying feature space or category structure. Recognition performance and reaction times improved until session 22, illustrating continued training effects.

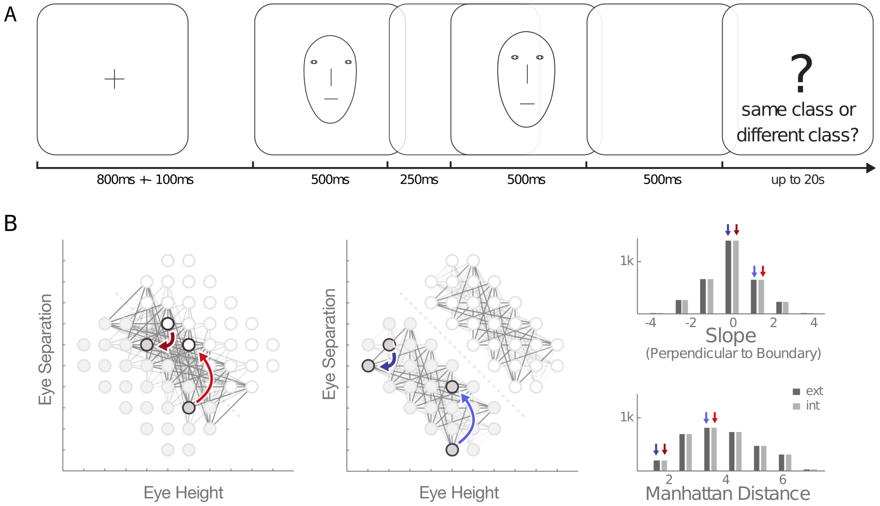

Figure 4.10: Adaptation Paradigm. To test for electrophysiological correlates of category information, an adaptation paradigm was used. Each trial either crossed the category boundary (category-external) or stayed within a category (category-internal). (**a**) Temporal sequence of an adaptation trial. (**b**) Low-level properties of the category-internal and category-external adaptation trials were controlled by matching the distance and slopes of the corresponding stimulus-pairs. Exemplary trials are highlighted in color (category-external in red, category-internal in blue).

but no main effect of category membership (p>0.05) and no significant interaction (p>0.05). Thus, although there was an overall increase in task performance with training, there was no significant difference in the performance of the category-internal and category-external trials indicating that the task was equally demanding in trials of both conditions.

Training-Induced Category Effects in Visual Responses

To test for category effects in the visually evoked responses, we compared the magnetic fields evoked by the second stimulus in the category-internal and category-external adaptation trials in the MEG adaptation paradigm, while controlling the low-level stimulus properties of the two conditions (Figure 4.10b). This indirectly tests for category selectivity, as differences between these two conditions will only be detectable if category-information is encoded in the underlying cortical activity.

For analyses of visually evoked responses, we employed a novel filter approach that allowed us to focus on relevant yet temporally changing cortical processes by projecting the data from its original 271 dimensional sensor space into a highly informative one-dimensional filter space (see Figure 4.11a, and section Data Analyses for more details). This avoids an a priori selection of sensors of interest and the multiple comparison problem occurring when all sensors are analyzed independently. Accordingly, we first computed the filter activations for each session, subject and condition (category-internal and category-external). We then performed a t-test on the filter response at every time-point to test for differences between category-internal and category-external conditions. This provided us with temporal candidate clusters that exhibit significant category effects for every session. To ensure that the observed category effects were indeed the result of category training, it had to be shown that category effects were significantly larger post-training as compared to the baseline session. As a final step, we therefore estimated the effect sizes and confidence intervals of the training interaction for each temporal candidate cluster (Bonferroni corrected at the cluster level, thereby controlling for multiple comparisons). Only temporal clusters surviving this rigorous control will be reported in the following. These clusters will not only exhibit significant category effects, but also show a significant training interaction, indicating that the seen category effects are the result of category training.

Using this procedure, we first analyzed the data from the baseline session. We found no significant category effects (Figure 4.12a), confirming that the category structures used for training were not an inherent property of the used stimulus space. We then analyzed the data of the two post-training sessions five and 22. After five training sessions, the earliest significant, training-induced category effects were evident between 275-293ms (Figure 4.12b, Figure S4.3a). With developing category expertise, however, a temporal shift in category effects was observed. After 22 training sessions, the earliest cluster exhibiting significant training effects occurred already after 113ms (lasting from 113 to 140ms). Additional time-windows of significant category effects were found between 171-175ms and between 220-233ms (Figure 4.12c, Figure S4.3b). The corresponding filter topographies are shown in Figure 4.11b. The observed speed-up of more than 160ms from session five to 22 is remarkable, as our subjects already categorized the stimuli at about 90% accuracy during training session five.

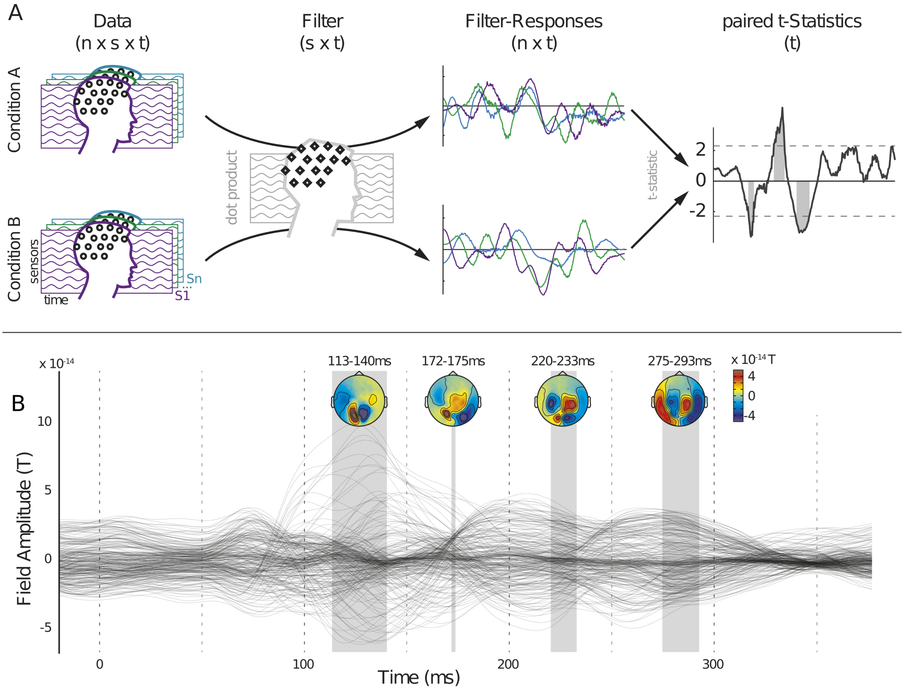

Figure 4.11: Filter Approach. A novel spatiotemporal filter approach was used to analyze the data. It focuses on relevant yet temporally changing cortical activity while avoiding the multiple comparison problem. The data used to define the filter is independent of the experimental question (comparing category-internal and category-external trials evoked by the second stimulus). (**a**) Schematic illustration of the filter approach. The evoked fields of the category-internal and category-external trials were projected to a one-dimensional filter response using a dot product. The resulting filter activations were subject to statistical analyses across time, highlighting temporal candidate clusters that show significant differences between category-internal and category-external trials. Abbreviations: n, trial number; s, sensor number; t, time point. (**b**) Butterfly plot of the used filter. Shaded areas and topographies mark time-points of training-induced category selectivity.

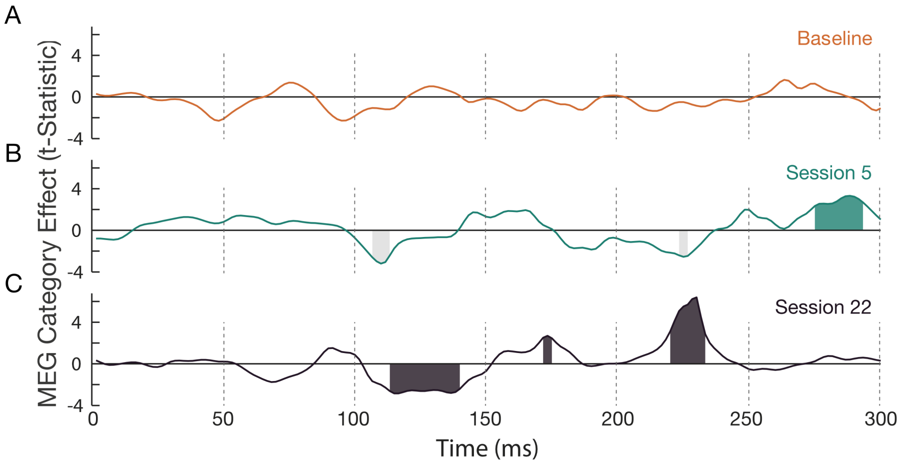

Figure 4.12: Training-Induced Category Effects across Time. The traces represent the T-statistic of the category effects, comparing category-internal and category-external trials, in the different sessions. Candidate temporal windows during which visually evoked responses showed significant category effects and a significant training interaction are shaded in dark colors. Candidate windows exhibiting no significant training effects are marked in light grey. (**a**) During the baseline session, no category effects could be found. (**b**) After five training sessions, the first training-induced window of category selectivity is present from 275-293ms. (**c**) After extended category training in 22 sessions, the earliest training-induced category effects are present from 113-140ms. Additional clusters of significant training-induced category effects were found between 171-175ms and 220-233ms.

Relation of Physiological Category Effects to Behavior

To test whether the observed category effects were behaviorally relevant, we compared the category effect sizes for successful and erroneous trials during the delayed match-to-category task. Again, we estimated the effect sizes and confidence intervals, while Bonferroni correcting for multiple comparisons at the cluster level. This analysis revealed significant differences for the earliest cluster in session 22, indicating the behavioral relevance of the effect. No other cluster in session 22 and five exhibited significant behavioral effects. Considering the absence of significant differences for session five, it should be noted that behavioral errors in the delayed match-to-category task can have various origins. These include incorrectly categorized stimuli, errors in working memory and an incorrect mapping of the perceptual decision to the appropriate behavioral response. These significantly complicate the search for behavioral relevance based on the subjects' performance. Moreover, it is possible that effects of behavioral relevance occurred at an even later point, extending beyond the 300ms analyzed here.

Source Analyses

Following the analyses in sensor space, we tested whether the temporal shift in category selectivity observed between session five (275-293ms) and session 22 (113-140ms) is due to altered neuronal processing in the same cortical areas, or whether different sets cortical areas are activated during these two time windows of interest. To this end, we computed a standardized low resolution brain electromagnetic tomography (sLORETA) (Pascual-Marqui, 2002) on the same data that was also used to define the spatiotemporal filter. We estimated the average source activations during the two time-windows of interest and tested for significant differences based on a T-statistic, while controlling for multiple comparisons using a cluster-based permutation test (Maris & Oostenveld, 2007) on the cortical surface. This analysis revealed that the previously shown temporal shift in category selectivity was accompanied by an anterior-to-posterior shift of cortical activation (Figure 4.13). The time-window of training-induced category effects in session five showed a stronger activation in the ventrolateral and ventromedial parts of the PFC. In contrast, the cortical activation during the earlier time-window of category selectivity in session 22 exhibited significantly stronger activity in more posterior regions, including the occipitotemporal cortex.

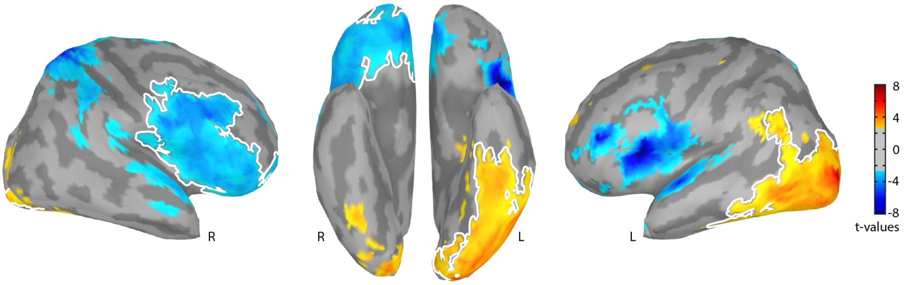

Figure 4.13: Source Localization Results. Source activations of the earliest clusters of category selectivity in sessions five and 22 were contrasted. Shown are uncorrected T-values with a cutoff at p<0.05. Blue regions show larger activity during the category-selective time-window in session five, red regions show larger activity in the early category-cluster in session 22. A white border highlights clusters surviving a control for multiple comparisons (cluster-based permutation test).

4.2.3 Discussion

Previous work on naturally occurring categories has demonstrated that category information can be rapidly extracted from visually presented objects. It remained unclear, however, how the visual system copes with the challenge to reach such rapid recognition speeds while at the same time allowing for sufficient plasticity to encompass the fast learning of entirely new categories. Are the same neuronal mechanisms and structures involved in recognizing reoccurring and newly learned categories, or are they different? And, if they are different, are novel categories implemented differently with prolonged experience? Here we investigated these issues by extensively training nine subjects to categorize two artificial visual categories. During training, we recorded MEG data in an adaptation paradigm to investigate the emergence of category selectivity in visually evoked responses. Additionally, MEG data were recorded prior to category training to serve as a baseline. Using a novel filter approach to analyze the visually evoked responses, we demonstrate the emergence and, following this, a temporal shift in category selectivity. The data recorded in the baseline session did not exhibit any category effects, indicating successful control for low-level stimulus properties. After five training sessions, the earliest training-induced category effects were found around 280ms of processing. With extensive training in 22 sessions, we observed a temporal shift in category selectivity. The first significant differences were now found about 160ms earlier, between 113 and 140ms. We then investigated whether the temporal shift in category selectivity was accompanied by changes in the spatial pattern of the underlying cortical activity. We compared the source activations during the two earliest temporal clusters of sessions five and 22 and found a significant anterior-to-posterior shift. While the cortical activity during the late category effects in session five showed stronger signals in PFC, the early time-window of category selectivity in session 22 exhibited an increased activation in occipitotemporal regions.

Our finding of an early cluster of category selectivity, starting at 113ms and lasting until 140ms, is fully compatible with previous studies of natural categories in macaque and human. In the macaque, Sugase et al. (1999) recorded from inferotemporal cortex (IT) and observed a peak in category information after only 117ms of processing. In line with this, Hung et al. (2005) demonstrated that relatively small numbers of randomly selected neurons in IT allow for reliable category decoding, peaking 125ms after stimulus onset. Interestingly, they also show decoding of low-level properties such as size and position of an object, arguing for residual retinotopic information in the neuronal response. This emphasizes the necessity to control for low-level stimulus properties and underlines the benefits of baseline measurements in category training. Finally, Freedman et al. (2003) applied a receiver operator characteristic approach to recordings from macaque IT and PFC. They showed that IT cells exhibited category selectivity after 127ms. In humans, electrocorticographic recordings provided direct evidence that natural categories can successfully be decoded at a mean latency of 115ms (Liu et al., 2009). Remarkably, decoding was possible based on single trials, allowing for generalization across rotation and changes in scale. In line with this, MEG recordings of human subjects provided evidence that visually evoked responses of houses and faces can be separated already at the time of the M100 component (Liu et al., 2002). In the same study, a positive correlation of response amplitude and categorization performance was shown, indicating the behavioral relevance of the early category signals. Using a multivariate decoding approach, Carlson et al. (2011) showed that it is possible to differentiate two visual categories (faces and cars) after 135ms of processing, even if the trained and tested stimulus positions were different. Finally, electrooculography (EOG) data provided by Kirchner & Thorpe (2006) suggest that category information is present and behaviorally relevant after only 120ms of processing. However, it should be noted that all of the studies mentioned above investigated neuronal responses to naturally occurring categories. Apart from the inherent challenges to differentiate category selectivity from systematic differences in the low-level statistics (Wichmann et al., 2010; Thierry et al., 2007; Rossion & Caharel, 2011; Crouzet & Thorpe, 2011; VanRullen & Thorpe, 2001), these stimuli do not allow for an investigation of emerging category selectivity with increasing category experience, which is the focus of the current study.

Overall, the neuronal mechanisms underlying the categorization of visual input have been in the focus of a lively debate over the recent years. A prominent view centers around the idea that category information is extracted by PFC (Antzoulatos & Miller, 2011; Roy et al., 2010; Cromer et al., 2010; Serre et al., 2007a). Accordingly, neuronal selectivity in temporal regions is seen as merely providing a sufficiently complex vocabulary from which the category information can be flexibly read out. This view is consistent with the predictions of the two-stage model of perceptual category learning (Riesenhuber & Poggio, 2002a), which hypothesized that neurons in IT obtain sharper tuning to re-occurring stimulus features, while regions in frontal cortex learn to associate these features with the corresponding category membership. In humans, experimental evidence supporting such division of labor was provided by Jiang et al. (2007). They showed that category training can lead to an increased shape selectivity in ventral areas whereas category selectivity was found only in the lateral PFC (but see Minamimoto et al. 2010). Moreover, there is evidence for enhanced shape selectivity in ventral areas in human and macaque (Freedman et al., 2006; van der Linden et al., 2010, 2013; Sigala & Logothetis, 2002). Nevertheless, the large body of evidence for rapid category selectivity in IT, as reviewed above, supports a contrasting view according to which category information might already be extracted at the level of the temporal lobe (DiCarlo et al., 2012). Providing a potentially unifying solution to this controversy, we have demonstrated here that prolonged category training can lead to a temporal shift in category selectivity, which is accompanied by an anterior-to-posterior shift in cortical activity. These data provide a cue as to how the brain could balance the need for robust and fast recognition of re-occurring categories while still allowing for considerable flexibility and rapid plasticity. Selectivity for novel categories relies more heavily on PFC and, as indicated by the long latency of the observed effect, potentially recurrent processing. Sufficient expertise with the categories, however, leads to changes in the cortical implementation of the trained categories, thereby allowing for a substantial speedup in processing times and emphasizing cortical processes in occipitotemporal regions.

A comparable view was recently described by Seger & Miller (2010) who proposed that the brain might simultaneously implement fast and slow learning processes. Fast learning provides multiple advantages, such as increased flexibility and rapid adjustments, but at the cost of an increased risk of erroneous classification. Slow learning, on the other hand, is less error-prone but at the cost of extended training requirements. In line with this, our results can also provide a potential explanation as to why some previous studies did not see (early) category selectivity in temporal areas after category training (Jiang et al., 2007; Li et al., 2007; Gillebert et al., 2009; Scholl et al., 2014). Apart from many differences between these experiments and our study, including the type of feature space used (Folstein et al., 2012a,b), our data suggests that the extent of training is a decisive factor. Comparably short training times might only reveal rather late category selectivity in frontal regions, as observed in session five here, whereas prolonged training is required for early occipitotemporal effects.

By contrasting correct and incorrect responses, we demonstrated significant behavioral relevance of the early category effects starting at 113ms in session 22. It has to be noted, however, that the time-points of category selectivity observed in sessions five and 22 do not necessarily mark the end point of the perceptual decision process. Successful performance in the delayed match-to-category task requires the successful completion of additional processing steps, such as the successful comparison of the two shown categories and the mapping of the perceptual decision to the appropriate motor response. Moreover, effects of perceptual certainty (Philiastides & Sajda, 2006) and ongoing evidence accumulation (Donner et al., 2009) can be expected to play a vital role in the perceptual decision process.

While further experiments are required to fully disentangle the contribution of these different factors, we have shown here that the brain is capable of extracting visual categories based on two different modes. Novel categories are recognized late, involving recurrent processing and increased activity in PFC. This pattern of results is consistent with a re-labeling of existing visual features, which would allow the system to flexibly learn new categories. Extended category experience, however, leads to a significant speed-up in category selectivity, accompanied by increased activity in occipitotemporal cortex. This suggests that re-occurring categories are processed differently to allow for quick and reliable recognition. Taken together, our results suggest that the brain balances plasticity for acquisition of new and efficiency in processing of known categories by relying on different brain networks.

4.2.4 Experimental Procedures

Participants

Nine healthy, right-handed subjects (5 female, aged 19-30) participated in the study. All subjects had normal or corrected-to-normal visual acuity, were naŢïve to the purpose of the study and gave written informed consent to participate. The experimental procedures were approved by the ethics committees of the University of Osnabrück and the Ärztekammer Hamburg. Each subject participated in a total of 23 experimental sessions (1 baseline sessions and 22 training sessions). MEG data were recorded during the first baseline session, as well as after training sessions five and 22. The MEG recording from subject nine in session 22 was excluded from the analyses due to excessive noise in the data.

Stimulus Space

Similar to previous work with macaques (Sigala & Logothetis, 2002) and humans (Kietzmann & König, 2010; Sigala et al., 2002), category training was based on two artificial categories of faces defined in a four-dimensional, parameterized stimulus space (Figure 4.8), also known as a factorial morphspace (Folstein et al., 2012b). Two of the dimensions were category-relevant (eye height and eye separation), while the two others (mouth height and nose length) were assigned pseudo-randomly to ensure that no stimulus clusters of the same category existed that could potentially render the task-irrelevant dimensions informative. A linear category boundary split the category space of the two relevant dimensions in half, such that no single stimulus property was decisive for the category membership. The overall stimulus space consisted of 60 stimuli, six of which defined the respective category boundary and were not included in the training and testing. The final two categories comprised 27 stimuli each. Moreover, the category boundary was rotated by 90° for every other subject. The subjects were at no point in time instructed about the design of the stimulus space or the relevant category dimensions. Hence, the category training used here resembles an information integration task (Ashby & O'Brien, 2005). All stimuli shown during training and the MEG adaptation sessions were presented using the Psychophysics Toolbox 3 (Brainard, 1997; Kleiner et al., 2007) running under Matlab 2010a.

Category Training

In order to allow subjects to learn the two artificial categories of faces, they received category training in a total of 22 sessions with 756 trials each. Here, we largely followed our previous procedure (Kietzmann & König, 2010). In each training trial, subjects were presented with a single stimulus and were then asked to categorize it as either category A or B with their index- or middle-finger. Auditory feedback was provided as training signal. A high-pitch tone indicated a correct response, whereas a low-pitch tone and a forced break of two seconds indicated an incorrect response. To prevent a fixed association between the category membership and motor response, the finger used to indicate the category decision was switched three times across each training session. The subjects were notified of the switches.

Adaptation Paradigm

To estimate the time-course of electrophysiological category effects, we used a rapid MEG adaptation paradigm. This approach is similar to the more common fMRI adaptation (Grill-Spector & Malach, 2001), and has only recently been introduced to the field of EEG and MEG (Caharel et al., 2009; Vizioli et al., 2010; Zimmer & Kovács, 2011; Harris & Nakayama, 2007; Scholl et al., 2014). MEG data were recorded in a baseline session, prior to category training, as well as after five and 22 training sessions. The selection of five and 22 training sessions was based on previous work using a similar feature space, in which subjects were able to perform at >90% accuracy after only five training sessions (Kietzmann & König, 2010), while exhibiting high-level category effects only after prolonged training. To test for category effects, two adaptation conditions were compared. Category-internal trials included two different stimuli from the same category (216 trials) and category-external trials included two stimuli from a different category (216 trials). In both cases, the visually evoked fields of the second stimulus were used for the analysis. To control for low-level differences in the two adaptation conditions, all category-internal and category-external trials were matched in distance and direction in the two relevant dimensions of feature space (see Figure 4.10a). This has the additional advantage that no linear re-weighting of the category-relevant dimensions can account for category selectivity, because all category internal and external adaptation trials will be affected likewise. During the randomized adaptation trials, the subjects performed a delayed match-to-category task, indicating via button-press whether the two stimuli were of the same or a different category. To prevent a fixed mapping of experimental condition to motor response, the target keys for the two answers were switched after half of the experiment. The structure of the adaptation trials was as follows. First, a fixation cross was presented for 800ms with an SOA of ±100ms. Then, a first stimulus was presented for 500ms, followed by an inter-stimulus-interval of 250ms. The second stimulus was again shown for 500ms. Finally, after an additional delay of 500ms a question mark was displayed on the screen, indicating to the subject that a response can be given (Figure 4.10b). This timed response was introduced to keep presentation of the second stimulus free of cortical activity related to the motor-execution.

MEG Acquisition

MEG data were acquired at 1200Hz using a CTF whole-head system with 271 axial gradiometers (CTF275, VSM MedTech). The position of the participants' head was continuously recorded using three head localization coils

(NAS/LPA/RPA). Moreover, a bipolar electrocardiogram (ECG) and an electrooculogram (EOG) with three channels were recorded. The EOG electrodes were placed below the eyes and on the forehead. The reference was positioned on the tip of the nose. The experimental stimuli were back-projected on a screen with a LCD projector (Sanyo XP51) at 60Hz refresh rate. The presentations distance was 60cm, leading to a display size of 2°x3.3° of visual angle.

Data Analyses

All analyses were performed using custom code in Matlab R2010a (Mathworks, Natick, MA, USA), fieldtrip (Oostenveld et al., 2011) and Brainstorm (Tadel et al., 2011).

Preprocessing

After downsampling the data to 600Hz, artifacts due to muscle activity, sensor-jumps and extreme noise were first detected automatically using fieldtrip, followed by manual cleaning of the data. To account for sensor drifts, the data were high-pass filtered at 1 Hz. Moreover, frequencies above 100Hz and the artifactual frequency bands around 50Hz and 60Hz were excluded using a zero-phase Butterworth IIR filter. To remove eye-related and cardiac artifacts from the data, we used an automated procedure based on an independent component analysis. The underlying algorithm relies on a correlation-based and a weight-based artifact-metric computed for each independent component. Components surpassing a selected threshold were labeled as artifacts and removed from the data. The optimal thresholds were determined automatically based on a receiver operator characteristic (ROC) analysis applied to a subset of the data for which two experts had classified components as artifacts. The resulting algorithm was able to detect 98.1% of the components tagged by the experts, with only 0.3% false positives (see Supporting Information and Figure S4.4 for more details). Finally, although our fully randomized design prevents systematic effects of head-position, we removed any residual effects form our data. We first extracted a six dimensional description of the head position and direction from the simultaneously recorded localization coils (NAS/LPA/RPA) and used this to regress out the effects of head-position (Stolk et al., 2013). All evoked potentials were baseline-corrected with respect to the 700ms of fixation prior to the presentation of the first stimulus.

Filter Approach and Statistical Analyses

The statistical analysis of multivariate MEG data introduces the problem of multiple comparisons. This problem is traditionally solved by either selecting sensors of interest a priori, or alternatively by testing all sensors individually and afterwards controlling for multiple comparisons, for instance by applying non-parametric cluster-based correction methods (Maris & Oostenveld, 2007). Unfortunately, an a priori selection of sensors is not without problems, either because it is unclear which sensors should be selected or, even more so, because sensors of interest change over time. The second solution, testing all sensors individually while applying a cluster-based correction, allows for tests in all sensors, but potentially at the price of decreased statistical sensitivity. Here we overcome these limitations by using a novel spatiotemporal filter approach, which allowed us to focus on relevant yet temporally changing cortical activity without the need to select sensors of interest and without the problems of controlling for multiple comparisons. The key insight to our approach was the observation that the response to the first stimulus in our adaptation paradigm provides all information necessary to focus on relevant cortical processes during the processing of the second stimulus. Hence, we were able to use the group average response to the first stimulus in the baseline session (low-pass filtered at 35 Hz using a zero-phase Butterworth IIR filter) as spatiotemporal filter on the response to the second stimulus (see Figure 4.11). Importantly, the data used to define the filter are independent of the experimental data in question (comparing category-internal and category-external responses in the second stimulus). In order to allow for comparisons of effect sizes across experimental sessions, the underlying filter was held constant. The use of a different filter for every session would make it impossible to decide whether observed changes in the effect sizes are merely due to changes in the underlying filter. To ensure that this approach was appropriate for the current dataset, we performed a non-parametric cluster test based on an F-statistic in which we compared the responses to the first stimulus across all three sessions (baseline, session five, session 22) within the first 300ms of processing (the cluster-threshold was set to $p<0.05$). No significant differences were found, indicating that the same filter was applicable for all sessions.

By using the described filter approach, the original 271 dimensional data were projected into a highly informative one-dimensional subspace. The approach therefore circumvents the problem of multiple comparisons when considering all sensors individually and avoids the need to select a subset of sensors a priori. The assumption of this, and in fact any localizer approach, is that the same cortical processes are active during the trials used to define the filter and the experimental trials of interest. While many experimental settings meet this assumption, adaptation paradigms are particularly suited for this approach. This is due to the fact that they already presuppose the same cortical regions to be active during the processing of the first and second stimulus. The use of the first-stimulus response as a filter on the evoked response to the second is therefore simply a consequent extension of the experimental paradigm to the analysis methodology. The focus on the same cortical processes active during the first stimulus provides additional benefits. First, it automatically excludes experimentally irrelevant processes, such as motor-preparation, which can interfere with relevant activity evoked by the processing of the category information of the second stimulus. Moreover, as all statistical analyses are based on the filter activations, the cortical sources of found effects can be assumed to be co-localized with the sources underlying the filter. This simplifies the analysis on source level by allowing to cortically localize the filter signal.

In our analyses, we applied the spatiotemporal filter to project the data of every subject, condition, and session to a one-dimensional time-series. To temporally localize time-windows of interest, i.e. time-windows exhibiting significant category effects ($p<0.05$), we performed a paired t-test at every point in time based on these filter responses. Following this, we investigated for every candidate time-window whether the observed category selectivity was the result of category training. This is an important additional prerequisite in investigations of developing category selectivity, as found differences between category-internal and category-external conditions could also be an inherent property of the selected feature space and not the result of category training. As a next step, we therefore tested each candidate time-window for a significant increase in effect size, as compared to the effect observed in the baseline session, by estimating the interaction effect size and its 95% confidence intervals. Corrections for multiple comparisons across time were performed at this final stage by applying a Bonferroni correction at the cluster-level. As a result of this statistical procedure, any cluster reported in the following will not only have shown significant category effects, but also a significant training interaction, verifying that the found effects are indeed caused by category training.

Summing up, by combining a novel filter approach with rigorous statistical analyses, we overcome the need to select sensors and time-windows of interest while controlling for multiple comparisons in space and time. The only free parameter of the overall approach is the p-value for the selection of temporal candidate windows, which was selected to be $p<0.05$. Finally, we here focus on the first 300ms of processing after stimulus onset, as this time-window approximately resembles typical fixation durations during free-viewing of natural scenes (Underwood et al., 1998).

Behavioral Relevance

To estimate the behavioral relevance of the observed category effects, we contrasted the effect size of adaptation trials in which the response of the subject was correct and trials in which an incorrect response was given. The reasoning of this approach was that if the found effects are behaviorally relevant, larger effects should be expected upon correct performance in the delayed match-to-category task. Similar to the statistical analyses of the training-interactions, we focused on clusters that exhibited significant category effects and training interactions, estimated the effect size and bootstrapped the respective upper- and lower bounds of the 95% confidence intervals (with replacement) while applying a Bonferroni correction for multiple comparisons at the cluster level.

Source Analysis

We used the sLORETA algorithm (Pascual-Marqui, 2002), as implemented in the Brainstorm software (Tadel et al., 2011), on the data previously used to define the spatiotemporal filter in sensor space, to compare source activity on the cortical surface. For every subject, we first segmented the individual MRI into white and gray matter using Freesurfer (Fischl et al., 1999a; Dale et al., 1999). We then performed the source reconstruction based on each individual anatomy and aligned the results to MNI space (Colin27) using spherical averaging of the cortical surfaces. To account for smaller errors in subject alignment, the surface data were smoothed with a 10mm full-width-half-maximum Gaussian Kernel. For statistical analyses, we contrasted the average source activity (L2-Norm) during the earliest time-window of category selectivity in session five (275-293ms) and session 22 (113-140ms) at every surface vertex and applied a clusterwise correction for multiple comparisons based on a nonparametric permutation test (Maris & Oostenveld, 2007). Only vertices showing $p<0.05$ were included in the cluster estimates.

4.2.5 Acknowledgements

We would like to thank José P. Ossandón for valuable discussions and input during the data analyses, Guido Nolte for his advice on MEG source localization and Maike Gaca for her help during data acquisition. This work was supported by EU grants ERC-2010-AdG- 269716 and FP7-ICT-270212.

4.2.6 Supporting Information

Supplemental Material: Automated ICA Component Tagging

In order to automatically remove eye-related and cardiac artifacts from the data, an independent component analysis (ICA) was computed on the combined MEG, ECG and EOG data. Two experts were asked to analyze part of the data to classify eye- and cardiac-related components. This labeled dataset was then used to develop an automated procedure, which was applied to the whole dataset as a final step (see Plöchl et al. 2012 for a related approach).

The automated procedure is based on two artifact-related metrics computed for every component: correlation-based and weight-based. The first metric was computed by correlating the activity of each component with the raw ECG signal and with eye-related activity, as computed by the differences between the EOG channels (horizontal = leftEOG-rightEOG and vertical = mean(leftEOG,rightEOG)-foreheadEOG). The idea was that artifact-related components should exhibit high activity at times in which the artifact is most pronounced, as observed most directly in the ECG and EOG channels. Once computed, the distribution of correlations across all components was z-scored to obtain the final metric.

The second metric was based on the weights of each ICA component. As mentioned above, the ICA was computed on the joint MEG, ECG and EOG data. It was therefore possible to define an artifact-related metric based on the relative weight that each component ascribed to the dimensions containing the ECG and EOG data. The reasoning of this approach was that artifact-related components should receive the clearest information from the artifact related ECG and EOG channels and thereby ascribe high weights to them. To compute the final weight-based metric values, we again applied a z-transformation - this time, on the distribution of ICA-weights in each component. Once the two artifact-metrics were defined, we performed a two-dimensional ROC analysis to determine the best combination of artifact thresholds for the two measurements. To define hits and false alarms, the labels provided by the experts were used as ground truth. The resulting thresholds lead to an artifact detection rate of 98.1%, with only 0.3% false positives. They were therefore used to automatically label and reject all artifactual components in the whole dataset. On average, 7.6 components were removed for every subject and session.

Supplemental Figures

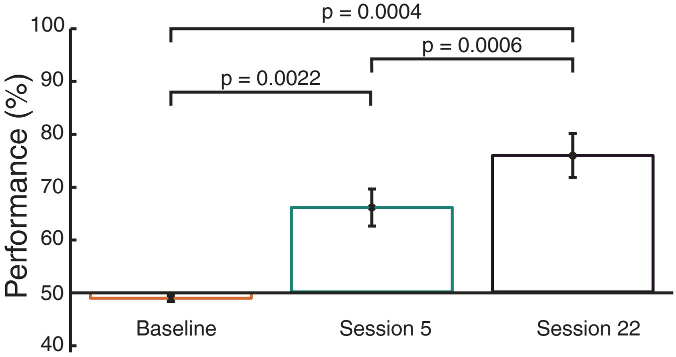

Figure S4.2: Adaptation Task Performance Relative to Chance Level. During the baseline session, the performance of our subjects in the adaptation task was not significantly different from chance, indicating that the category boundary could not be inferred without training. With category training, however, significant performance improvements were observed.

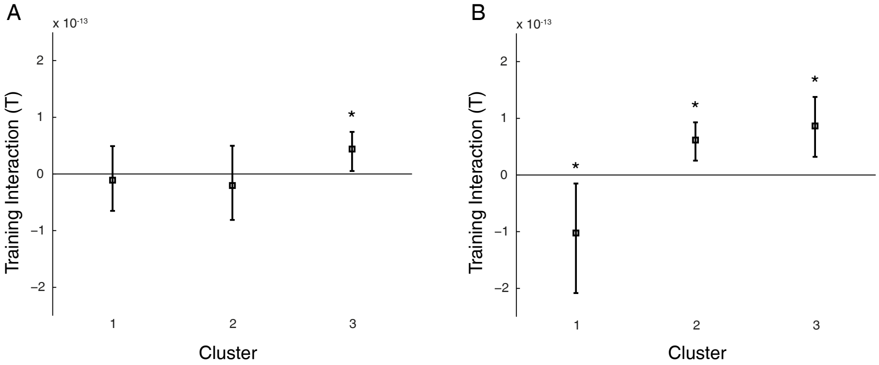

Figure S4.3: Training Interaction. To test whether the found category effects in the temporal candidate clusters are the result of category training, we estimated the effect size and confidence intervals of the training interaction (contrasting the category effects before and after training). Only candidate clusters for which the 95% confidence interval (Bonferroni corrected for multiple comparisons at the cluster level: 98.3% confidence intervals shown) did not include zero were reported as showing significant training effects. (**a**) Interaction effects and confidence intervals for temporal candidate windows in session five. (**b**) Same data for session 22.

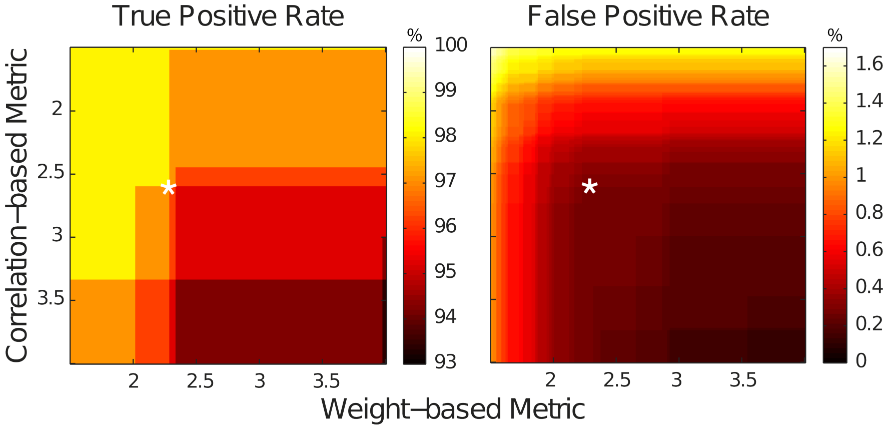

Figure S4.4: ROC Analysis to Determine the Best Artifact Thresholds. Shown are the 2D-ROC plots for the true- and false-positive rates. The optimal combination of thresholds yields a true positive rate of 98.1% and only 0.3% false positives.

4.3 Chapter Summary

Extending the notion of invariance, we here focused on the problem of object categorization and its interplay with cortical plasticity. Previous research based on naturally occurring categories demonstrated that categorical information is extracted at a rather early stage of visual processing. However, it remained unknown how novel categories are spatiotemporally represented in the brain and how these representations change with extended category experience. In addition to problems with category-inherent low-level stimulus statistics, naturally occurring categories cannot be used to answer questions of category learning, as an experimental control for category exposure is impossible due to prior encounters.

To investigate emerging category representations, we trained our participants to categorize two artificial categories of faces in two longitudinal studies. In the first, we investigated behavioral effects of prolonged category experience. In particular, we tested for effects that are indicative of a change in the underlying type of category representation and investigated the interplay of low-level stimulus properties with high-level category information during perceptual decisions. We found that while the initial performance of our participants was best explained by an exemplar-based representation (Gauthier & Palmeri, 2002), the effects observed after extensive category training indicated a more abstract type of category representation. Together with this shift, we observed changes in the integration of high- and low-level information with training. Whereas initial performance in an odd-one-out paradigm was purely dependent on low-level similarity between target and distractor, extensive training lead to a more dynamic use of low- and high-level information.

In the second experiment, we recorded cortical activity at different stages of category training using MEG, while our participants made categorical decisions in a delayed match to category paradigm. Importantly, we ensured that any found category effects were the result of training and not an inherent property of the used stimulus space by including a baseline measurements preceding training. Using this paradigm together with a novel filter approach for the analysis, we observed a spatiotemporal shift in category selectivity. After little training, category effects were observed at a late point in time, together with stronger activity in prefrontal cortex. After extensive training, however, the category membership of the shown stimulus was resolved significantly earlier. This early category selectivity was accompanied by stronger activity in occipitotemporal cortex.

Taken together, we showed in this chapter that extensive learning of novel visual categories can lead to changes in the underlying category representation. In line with the observed shift towards a more abstract category-representation, we provided evidence for a spatiotemporal shift in the underlying network dynamics. These findings answer the question of how the brain robustly recognizes re-occurring categories while allowing for sufficient plasticity to account for rapid learning of novel categories: by relying on partially separate cortical networks that either concentrate on novel and changeable categories or on re-occurring categories for which efficient and timely recognition is essential.

5
General Discussion

In the past chapters, I described a series of experiments, in which we investigated how the visual system robustly and efficiently extracts information from our complex and constantly changing environment. I focused on three main aspects of recognition: sampling the environment, visual invariance, and categorization and plasticity. Since more detailed summaries were already provided in the respective chapters, this section will focus on the principles of efficiency and robustness, which unify all experimental findings presented.

Regarding the first aspect of visual processing investigated, namely the sampling of the environment, three eyetracking experiments were conducted in which we tested whether the sampling of information, in the form of overt visual attention, interacts with cortical processes devoted to recognition and perception. Our analyses revealed a bi-directional influence. Whereas initial patterns of overt visual attention were predictive of (and causally related to) the later conscious recognition, briefly presented contextual information lead to substantial changes in the sampling behavior, as evidenced by prolonged fixation durations. The direction of change suggests a bias shifted away from an explorative and towards a more exploitive analysis strategy. An integration of attentional selection and recognition processes is clearly efficient, as it enables spatial and temporal aspects of attentional sampling to guide but also to be guided by the current information state of the system. A shifting exploration-exploitation bias furthermore allows for dynamic reactions in response to environmental changes. For the temporal aspects of attentional sampling, we have seen that extended in-depth processing is initiated only after sufficient knowledge about the overall scene has been collected. This behavior is also beneficial with respect to an optimal survival strategy. A quick and dirty assessment of the visual scene should precede a more elaborate analysis of local details, to reduce the chance that important aspects are overlooked. If the initial assessment is expedited, as in the case of contextual information, more detailed processing can be initiated earlier. Taken together, our eyetracking experiments provide behavioral evidence that the processes of attentional sampling and object recognition interact to a larger degree than previously thought, and in a dynamic, efficient manner.

Following this, we concentrated on object recognition in a more static scenario to investigate how information, once in the system, is efficiently processed to achieve visual invariance. The experiments presented focused on the potentially most drastic version of such identity-preserving variation: changes in 3D viewpoint. Collecting fMRI data with high spatial resolution, we identified a spatially distributed, yet functionally specific representational property: selectivity for mirror-symmetric viewing angles. This effect was prevalent across a large variety of higher-level visual areas in the ventral and dorsal visual streams, indicating that viewpoint symmetry might be a fundamental property of visual processing. In addition to demonstrating the overall effect, we provided evidence for the causal involvement of the OFA in judgments of viewpoint symmetry in a follow-up TMS experiment. Interestingly, our analyses revealed that stimulation applied to the hemisphere ipsilateral to the visual information lead to specific impairments in symmetry judgments. This hints at the possibility that hemispheric interactions are part of the neural mechanism underlying viewpoint symmetry. In the third experiment in this series, we conducted an EEG study to gain insight into the temporal development of viewpoint processing. Our multivariate analyses revealed that the human brain exhibits effects of viewpoint symmetry at an early stage of processing, around 120ms after stimulus onset. This imposes further constraints on potential neural mechanisms, excluding explanations that postulate extensive recurrent processing. At a later stage of processing, we observed a marked shift in the representational similarity structure. Instead of low-level similarity and symmetry effects, observed earlier, the front-on viewpoint lead to significantly different activity patterns as compared to all other views, potentially related to the social importance of direct eye contact analyzed at later stages of visual processing. Let us again return to the question of how these findings demonstrate the efficient and robust processing capabilities of the visual system. With viewpoint symmetry, a property previously unpredicted by models of viewpoint invariance, the visual system has developed an effective way to reduce the computational complexity of the overall viewpoint invariance problem. Viewpoint symmetry could be implemented by a simple mirroring operation, which is computationally inexpensive compared to the potential benefits of increased viewpoint invariance. Moreover, separating the analyses of different facial properties across time, as observed in our multivariate analyses of the EEG data, is again efficient from a survival point of view because detailed social information, such as shared attention, can be assumed less time-critical compared to the overall recognition of objects in a given scene.

In the third set of experiments, we studied the effects of categorization, which can be considered a more general case of visual invariance, and its interplay with cortical plasticity. To investigate how novel categories are cortically implemented in novices and experts, we conducted two experiments in which we extensively trained our subjects to distinguish two artificial categories of faces. Using behavioral measures, we showed that prolonged training shifts the type of category representation from an exemplar-based to a more abstract form of category information. After training, we observed a dynamic use of the available information. When high-level category membership was uninformative for the given task, our participants relied solely on low-level stimulus properties. However, when differential high-level information was present, the use of low-level stimulus information was strongly attenuated. Following this purely behavioral study, we combined category training with MEG to investigate the underlying cortical mechanisms in more detail. In line with our behavioral results, our analyses revealed a spatiotemporal shift in category selectivity. Whereas category information in novices was resolved at a late stage of processing, involving stronger activity in prefrontal cortex, experts exhibited much earlier category effects, accompanied by stronger occipitotemporal activity. Picking up the unifying theme of efficiency, our behavioral data indicate that perceptual decisions can rely on high- and low-level visual information in a dynamic manner, recruiting low-level information on demand, if high-level information is not informative for the current task. Moreover, our electrophysiological data demonstrate that the initial representation of newly learned categories and the robust recognition of re-occurring categories rely on partially separate cortical networks. This suggests a solution to the question of how a balance is achieved between robust, fast recognition of re-occurring categories and considerable flexibility and plasticity. Changing the existing category representations too quickly with every novel category, whether it might prove to be important in the long run or not, jeopardizes the reliability of the recognition of well-established categories. It is therefore clearly sensible to initially separate novel and re-occurring categories and to integrate a new category into the feed-forward mechanisms only if it proves to be continually important.

All experimental results presented adhere to the unifying view of a dynamic system, which has evolved intricate mechanisms to allow for efficient, yet robust performance in a complex and changing environment. In this respect, the current findings mark important advances to our understanding. An overall focus on efficiency in visual representations is, however, not new. Representational efficiency, for example, is based on the insight that not all visual information offered by the outside world needs to be represented equally by the visual system. Rather than including unnecessary information and detail, a more efficient approach would be to concentrate on relevant statistical regularities in the environment. This view has gained considerable empirical support in more recent years. For instance, Gauthier et al. (2000a) demonstrated expertise effects in otherwise face-selective regions for car and bird experts. When presented with their category of expertise, the experts' FFA exhibited increased activity, indicative of a preferential treatment of the respective category by the visual system. The effects of statistical regularities on visual representations have also been studied more explicitly. Again using fMRI, it was shown that faces and bodies elicit strongest activation in face- and body-selective areas when being presented in commonly experienced configurations in relation to the center of fixation (Chan et al., 2010). Moreover, it was demonstrated that co-occurrence statistics of objects in a large variety of scenes can be used to predict activity levels in non-retinotopic anterior visual cortex (Stansbury et al., 2013). Taken together, these functional imaging experiments indicate that the visual system exploits statistical regularities to specifically represent re-occurring, relevant visual information. Psychophysical evidence for this account comprises the so-called own-age and other-race effects. The own-age effect is the phenomenon that children perform significantly better at recognizing their peers compared to other age groups (Anastasi & Rhodes, 2005; Hills & Lewis, 2011). Consequently, the optimal recognition age changes with subject age, and the visual system seems to constantly focus on the currently relevant peer group. Similar effects have been reported with ethnicity. In the other-race effect, people are significantly worse in distinguishing individuals from another ethnicity. Importantly, the effect is not present from birth, but acquired in the first nine months

(Kelly et al., 2007). Both effects, other-race and own-age, provide crucial evidence that judgment of which information is relevant and cortically represented can change and thereby emphasize the role of cortical plasticity in visual representations. Whereas accounts based on statistical regularities can explain such effects, theories purely based on the influence of genetic factors fail to account for such developmental changes.

Closely related to the account based on statistical regularities, computational approaches based on objective functions propose that visual representations are the result of an optimization process that minimizes a selected criterion contingent on the naturally occurring input to the system. This implies that receptive field properties are the result of the interplay between natural image statistics and unsupervised optimization procedures. In the introduction, we have already touched upon perhaps the most prominent objective function in discussing differences between fully distributed and grandmother-cell-like representations in higher-level visual cortex. Sparse coding (Barlow, 1961) suggests that the receptive fields of neurons should be shaped such that only a small fraction of neurons in a given population respond strongly at any given time. As an important proof of concept, it was shown that an application of the sparse coding principle to natural images can result in feature detectors that closely resemble the receptive field properties of simple cells in the striate cortex (Olshausen & Field, 1996). Experimental evidence for sparse coding in the brain has been provided by findings in early visual areas (Vinje & Gallant, 2000; Ravikumar et al., 2008) and higher-level visual cortex (Young & Yamane, 1992). As an extension of this approach, the closely related concepts of stability (Einhäuser et al., 2002; Körding et al., 2004; Wyss et al., 2006; Einhäuser et al., 2005), slowness (Wiskott & Sejnowski, 2002; Berkes & Wiskott, 2005; Franzius et al., 2008) and temporal coherence (Wallis & Rolls, 1997; Rolls, 2012) have been proposed. These approaches build on the observation that noise and small environmental changes vary on a faster timescale than the identity of meaningful elements in a given scene. As an example, imagine walking past another person in the street. Sensory noise, movement, and saccades cause fast changes on your retina. On a slower timescale, the viewpoint from which you see the other person will change as you walk by. On the slowest timescale, however, is the identity of the person, which stays constant. By exploring the spatiotemporal structure of the input, the family of 'slow' objective functions bootstraps information from the input and learns invariances on multiple levels of description. Independently of the objective function chosen, a hierarchical application of the optimization can lead to increasingly complex features, which respond quickly to increasingly slow features of the visual input. To return to the walking example, a hierarchical application of this principle predicts that identity information should be resolved at a later stage than viewpoint information, because it varies on a slower timescale. When applied hierarchically, these approaches are similar to Neocognitron and its descendants, which also rely on the repeated application of identical selection criteria to achieve increasingly invariant and complex representations across different network layers. However, these approaches, with the notable exception of VisNet (Wallis & Rolls, 1997), do not include temporal information, and do not explicitly optimize a given objective function, but mirror the statistics of natural input through a random representation of static images.

To return to our previous discussion of efficiency, we have now established that an efficient representation implies neither a uniform representation of all available information, nor a static representation. Instead, efficiency and robustness are best understood in terms of a dynamic focus on relevant representational subspaces, which optimally support the survival and the current task of the organism. This focus on subparts of the visual input does not imply that the existing representations are themselves not a veridical, one-to-one portrayal of the external world. We will now, however, go one step further and argue that veridicality is not necessary for efficiency. In fact, there are situations in which a non-veridical representations can be highly beneficial. In the introductory chapter, we have seen that the information entering our brain is far from perfect due to the inhomogeneities of the retina. Extending this view, the results of our experiments suggest that cortical representations, too, are clearly not veridical. As proposed above, non-veridicality can be seen as a feature, rather than a shortcoming, when the overall goal of the system is considered. As a first example, let us consider the case of viewpoint-invariant face identification. If the goal of the system is to recognize a face irrespective of the current viewpoint, then a veridical representation of the exact viewpoint is task-irrelevant. Mechanisms like viewpoint symmetry, which presumably collapse symmetric viewpoints onto similar response patterns, provide a shortcut to invariant recognition, albeit at the cost of veridicality. A similar argument can be constructed based on our visual memory. As illustrated in the introduction, we are astonishingly bad at determining whether a seen picture was mirror-reversed across the vertical axis (Mona Lisa example). This demonstrates that visual memories, too, are not veridical. Ambiguous stimuli, and in fact a large variety of perceptual illusions including binocular rivalry (Tong et al., 2006), provide another argument against a cortical representation that is faithful to the outside world.

The very idea of ambiguous stimuli two different interpretations are possible from exactly the same stimulus. Despite the fundamental ambiguity, the visual system settles on a single interpretation at any given moment in order to allow unique actions and predictions. Of course, this does not imply that the system is or should be blind to the overall ambiguity, as alternative interpretations can provide valuable information and should therefore not be missed. Nevertheless, mirroring the outside world would not solve the overall problem of ambiguity. As a final example, we have seen that cortical representations are not stable but undergo constant re-organization to account for changes in the external world or in the system itself. Although the underlying stimulus is again constant, categorization training can alter the perceptual similarity between categories such that stimuli belonging to different categories are perceived as less similar while stimuli of the same category are perceived as more similar, although the actual differences between the stimuli are identical in both cases. Clearly, a strict veridical form of representation does not predict such perceptual changes. Taken together, we have seen that efficient and robust visual representations do not need to be complete or veridical. "For natural selection does not care about truth; it cares about reproductive success" (Stich, 1990, p. 62).

The view that I defended in the last paragraph suggests that visual representations are tuned to the requirements of the overall system and the statistical regularities of its natural input, even if this implies non-veridicality. This view is strongly coherent with a more recent philosophical position in cognitive science, which strongly emphasizes the role of the complete organism and its action repertoire in understanding cognition (Varela et al., 1993; Clark, 1999; O'Regan & Noë, 2001; Engel et al., 2013). Contrary to the classical representational approach, the key idea of these proposals is that the goal of vision is not a faithful representation of the surrounding world, but rather the extraction of action-relevant aspects from the visual input. Recent advances in functional neuroimaging support the view that action-related properties can shape visual representations in areas predominantly associated with visual processing (Mahon et al., 2007; Bracci & Peelen, 2013). In line with this more integrated view, it has been shown that visual category information is present in cortical areas outside the traditional ventral stream, including the parietal (Fitzgerald et al., 2012) and even early somatosensory cortex (Smith & Goodale, 2013). Potentially unifying these diverse findings, it has been suggested that the overall organization of domain-specific conceptual knowledge in the brain is the result of functional coupling between distinct cortical areas, including but not limited to sensory and motor-areas, which process information about the same domain

(Mahon & Caramazza, 2011). Interpreting neuronal selectivity in light of network dynamics, instead of the traditional stimulus-to-response mapping, has the advantage that it easily allows for the integration of action-related aspects in otherwise purely sensory regions. As a potential mechanism for such large-scale integration, neuronal synchronization was suggested (Varela & Lachaux, 2001; von Stein et al., 2000; Engel et al., 2001; Hipp et al., 2011). Finally, incorporating the action repertoire of the embedding organism into our understanding of visual function provides an elegant answer to the question of how different species can develop different visual receptive field properties, despite being subject to comparable image statistics (König & Krüger, 2006; Weiller et al., 2010).

What should be clear, not only from this general discussion but throughout this dissertation, is that vision research is an interdisciplinary endeavor at the heart of cognitive science, which fascinates and challenges neuroscientists, psychologists, computer scientists and philosophers alike. With this great variety of scientific disciplines comes a largely successful scientific approach that spans multiple levels of explanations: experimentally, from single-cell recordings to neuroimaging of the whole brain; algorithmically, from low-level unsupervised learning to high-level symbolic processing; and conceptually from neuronal plasticity to conscious perception. These are exciting times to be a cognitive scientist, with experimental methods that allow for previously unheard-of temporal and spatial accuracy in our measurements and computational power that supports models of a complexity unthinkable only a decade ago. Importantly, it is now starting to be understood that only by bridging the gaps between all involved disciplines and by embracing a multi-leveled approach will we eventually be able to understand all aspects of one of the most prominent functions of the brain: vision.

6 Appendix

6.1 Measures and limits of models of fixation selection[1]

Abstract Models of fixation selection are a central tool in the quest to understand how the human mind selects relevant information. Using this tool in the evaluation of competing claims often requires comparing different models' relative performance in predicting eye movements. However, studies use a wide variety of performance measures with markedly different properties, which makes a comparison difficult. We make three main contributions to this line of research: First we argue for a set of desirable properties, review commonly used measures, and conclude that no single measure unites all desirable properties. However the area under the ROC curve (a classification measure) and the KL-divergence (a distance measure of probability distributions) combine many desirable properties and allow a meaningful comparison of critical model performance. We give an analytical proof of the linearity of the ROC measure with respect to averaging over subjects and demonstrate an appropriate correction of entropy based measures like KL-divergence for small sample sizes in the context of eye-tracking data. Second, we provide a

[1] This section was published together with Torsten Betz, Niklas Wilming, and Peter König as a peer reviewed article in Plos One. See Publication List for details.

lower bound and an upper bound of these measures, based on image-independent properties of fixation data and between subject consistency respectively. Based on these bounds it is possible to give a reference frame to judge the predictive power of a model of fixation selection. We provide open-source python code to compute the reference frame. Third, we show that the upper, between subject consistency bound holds only for models that predict averages of subject populations. Departing from this we show that incorporating subject-specific viewing behavior can generate predictions which surpass that upper bound. Taken together, these findings lay out the required information that allow a well-founded judgment of the quality of any model of fixation selection and should therefore be reported when a new model is introduced.

6.1.1 Introduction

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

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

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

In the first part of this article, we therefore review different commonly used evaluation measures. We define properties that are desirable for evaluation measures and provide evidence that many commonly used measures lack at least some of these properties. Because no single measure has all of the desirable properties, we argue that reporting both the Area Under the receiver-operating-characteristic Curve (AUC) for discriminating fixated from non-fixated locations, and the Kullback-Leibler divergence (KL divergence) between predicted fixation probability densities and measured fixation probability densities, gives the most complete picture of a model's capabilities and facilitates comparison of different models.

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

Finally, we examine the conditions under which the proposed upper bound holds by turning to a top-down factor that has so far been neglected in the literature. We show that incorporating subject idiosyncrasies improves the prediction quality over the upper bound set by inter-subject consistency. This should be interpreted as a note of caution when using our proposed bounds, but does not call into question their validity in the more general and typical case of modeling the viewing behavior of a heterogeneous group of subjects.

6.1.2 Results

Measures of model performance

In this section, we review commonly used measures for the evaluation of models of fixation selection. Our aim is to investigate, on a theoretical basis, what the advantages and disadvantages of different measures are and to identify the most appropriate measure for model evaluation. To reach this aim, we choose a four step approach. First, we establish a list of desirable properties for evaluation measures. Second, we identify commonly used measures in the literature and describe how they compare model predictions to eye-movement data. Third, we assess how the measures fare with regard to the desirable properties.

Justified by this, we recommend the use of the AUC. Finally, we elucidate the effect of pooling over subjects and conclude that in some circumstances, KL-divergence is a more appropriate measure.

Desirable properties for evaluation measures

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

- Few parameters: The value of an evaluation measure ideally does not depend on arbitrary parameters, as this can make the comparison of models difficult. If parameters are needed, meaningful default values or a way of determining the parameters are desirable.
- Intuitive scale: A good measure should have a scale that allows intuitive judgment of the quality of the prediction. Specifically, a deviation from optimal performance should be recognizable without reference to an external gold standard.

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

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

distributed but, due to its sparseness, dominated by the presence of many very unlikely events.

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

Existing measures

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

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

- *Chance-adjusted salience (S_a)* (Parkhurst et al., 2002) is the difference between the mean salience value of fixated locations on an image and the mean salience value of the viewed image. Thereby, if values are larger than 0, salience values at fixated locations are above average.

- *Normalized scan-path salience (NSS)* (Peters et al., 2005) is the mean of the salience values at fixation locations on a salience map with zero mean and unit standard deviation.

- The *ratio of medians* (Parikh et al., 2010) compares the salience values at fixated locations to the salience at random control points. The salience value of a location is determined by finding the maximum of the salience map in a circular area of radius 5.6 degree around that location. The median salience at fixated locations and the median salience of a set of random control points on the same image are computed for each image. The ratio of both medians is used as evaluation measure.

- The 80^{th} *percentile measure* (Torralba et al., 2006) reports the fraction of fixations that fall into the image area that is covered by the top 20% of salience values. It therefore reports the true positive rate of a classifier that uses the 80^{th} percentile of the salience distribution as a threshold. The selected area covers, by definition, 20% of the image, which is therefore the expected value for a random prediction.

- The *area under the receiver-operating-characteristics curve (AUC)* (Tatler et al., 2005) describes the quality of a classification process. Here, the classification is based on the salience values at fixated and non-fixated image locations. All locations with a salience value above a threshold are classified as fixated. The AUC is the area under the curve that plots the true positive rate against the false alarm rate for all possible thresholds (the receiver operating characteristic). As the threshold is continuously lowered from infinity the number of hits and false alarms are both increasing. When the salience map is useful, the hits will increase faster than the false alarms. With still lowering threshold the latter will catch up and the fraction of hits and false alarms both reach 1 (100%). The AUC gives an estimate of this trade-off. An area of 1 indicates perfect classification, 100% hits with no false alarms. An area of 0.5 is chance performance. See Fawcett (2006) for an introduction to ROC analysis.

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

- The *Kullback-Leibler divergence (D_{KL})* (Itti & Baldi, 2005a,b) is a measure of the difference between two probability distributions. In the discrete case it is given by:

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

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

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

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

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

Evaluation of measures with respect to the described properties

Having proposed a list of desirable properties and introduced a number of different measures, we can now examine how these measures cope with the requirements and what aspect of the prediction they evaluate. For an overview, please see Table 6.1.

- *Few parameters:* There are three measures that do not have parameters: S_a, NSS and AUC. The ratio of medians is dependent on the radius that is used for selecting a salience value for a fixation. Although there may be reasons for choosing one value over another, this parameter is essentially arbitrary. The percentile chosen for the 80^{th} percentile measure is completely arbitrary; it might as well be the 82^{nd} percentile. For the naïve Bayes classifier, the correlation and the KL-divergence, it is necessary to estimate probability distributions, which in the simplest case depends on the binning used. The naïve Bayes classifier furthermore requires the specification of the number of cross-validation runs.

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

- *Low data demand:* The three methods that require probability density functions, KL-divergence, correlation and naïve Bayes classifier, require a lot of data to form accurate estimates of the necessary probability distributions. In contrast, all other methods use only the fixated locations as positive instances and can in principle be computed on very few data points.

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

Table 6.1: Summary of described evaluation measures. The table shows a summary of the evaluation measures and their performance with regard to the desirable properties described above. '+' indicates that the measure exhibits the property, while '0' and '-' indicate that the measure is neutral w.r.t. to the property or does not exhibit it.

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

In summary, our evaluation shows that there are large differences in the suitability of the different measures when it comes to evaluating models of fixation selection. S_a, NSS and the ratio of medians are not intuitive to interpret and/or not robust. From the three classification measures, the AUC appears to be most favorable. It improves on the 80^{th} percentile measure by removing the arbitrary parameter and by including false alarms into the analysis. The naïve Bayes approach needs more data than is often available and the estimation of probability density maps is non-trivial. Correlation and KL-divergence need much data and require the estimation of density functions. Additionally, KL-divergence is not easy to interpret, but has a sound theoretical basis when the comparison of probability densities is concerned. The AUC stands out on the properties we have outlined. Based on our defined requirements, the AUC seems to be the best choice for evaluating models of fixation selection.

The effect of pooling over subjects

The selection of an appropriate measure is only one aspect of the evaluation process. Additionally, properties of the data against which the model is evaluated are of importance. Usually, when devising models of fixation selection, we are interested in the combined viewing behavior of several subjects, i.e. fixation data is pooled across subjects. The model should preferably predict those locations that are fixated by many subjects, because these fixations are most likely caused by salience or other factors that are stable across subjects, and not causes of fixations that are irrelevant to understanding information selection mechanisms. As a consequence of this, models that are trained to predict the joint-subject viewing behavior should perform better in predicting fixations from a set of subjects than in predicting the individual subjects from that set. This important property of model quality is not captured by the AUC and NSS . Figure 6.1 shows an example, where the quality of prediction as measured by AUC or NNS for the combined smooth fixation density map is just as good as the average quality of prediction of the individual subjects. That this is a general property of the NSS measure is easy to see: it takes the mean salience values at fixated locations, and for the mean it does not make a difference whether we take it for subsets individually and then average over the resulting value, or take the mean of the complete set directly. The linearity of AUC under decomposition of positive observations into subsets is less obvious, but proven in Materials and Methods: Proof of AUC linearity. In contrast, KL-divergence and correlation yield better values for predicting the joint viewing behavior, because they operate on fixation density map estimates, which take the spatial relation between fixations into account, and they are thus able to give non-linearly more weight to those locations that have been looked at by many subjects (see Figure 6.1). This non-linear weighting can be a good reason to consider the KL-divergence or correlation for model evaluation, despite their computational difficulties mentioned above. Deciding which of the two measures to use when one wants to exploit the effect of pooling over subjects is a difficult questions. Both measures are not robust, and both have the potentially disadvantageous property of being sensitive to non-linear monotonic transformations of the prediction. Correlation has the advantages of boundedness and being slightly less sensitive to some rescalings of the model output. However, the intuitive interpretation of its scale breaks down and becomes misleading if the dependence that is being measured is not really linear. KL-divergence is extremely sensitive to low (close to zero) predictions for locations that get a higher empirical salience, but is conceptually more appropriate for comparing probability distributions. In the end, both measures are not optimal, but because of its sound theoretical basis, we recommend using the KL-divergence when one wants to capture the ability of the model to exploit similarities in the viewing behavior within a group of subjects. In practical applications of this measure, one should also be aware of an additional complication: KL-divergences are dependent on the number of fixations used to compute the fixation density maps (Materials and Methods: Fixation density map estimation). As a result, values which are estimated from different numbers of fixations are not directly comparable. For example, when the average fixation duration in an experiment with fixed viewing time per stimulus is dependent on image category, this can con-

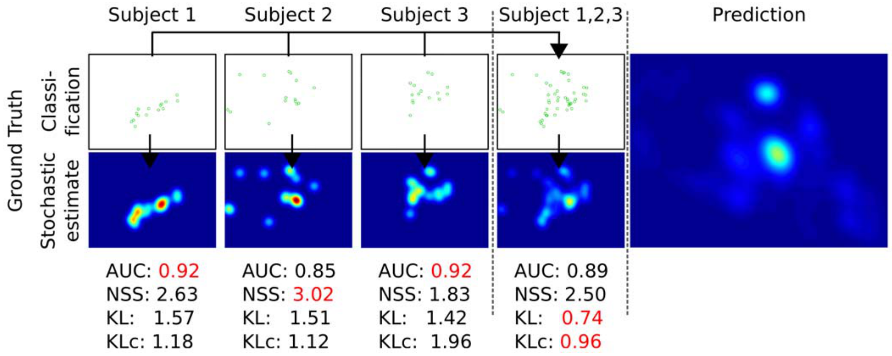

Figure 6.1: Predicting the joint fixation selection process of several subjects vs. predicting individual subjects. The prediction in this case was generated not from a model but from the fixations of several independent subjects. It therefore captures the joint process of a group of subjects. When treated as a classification problem (top row), only the fixation locations are important. In this case, the mean of the AUC or NSS scores for the individual evaluations are identical to the AUC or NSS score of evaluating the joint process. When treated as a stochastic process (bottom row; see Materials and Methods: Fixation density map estimation for computational details of fixation density map estimation), locations that were fixated by one but not all subjects are less important to predict. KL-divergence, which evaluates not individual fixations but the prediction of the stochastic process, yields a better score for the evaluation of the joint process. This also holds true when it is corrected for the number of fixations in the data (KLc).

found a comparison between categories. In Materials and Methods: Correction of KL divergence for small samples we investigate this dependency and describe a method for correcting KL-divergence scores for the bias introduced by limited data by exploiting the measure's relation to information entropy. In summary, the linearity of AUC under decomposition into subsets and the sensitivity of KL-divergence and correlation for joint-viewing versus single-subject behavior are both relevant whenever a model of fixation selection is evaluated against fixation data. KL-divergence is especially appropriate when fixation data from a group of subjects are the target of a prediction.

Intermediate Summary

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

Properties of fixation data

The aim of the second part of this work is to investigate the upper and lower bounds on the prediction performance of fixation selection models. To this end, we examine the image and subject independent spatial bias on the one hand, and image-specific inter-subject consistency on the other hand. We use data from an eye tracking study carried out previously in our group (see Materials and Methods: Description of the eye-tracking study for details and Figure 6.2 for some examples of stimuli). We first analyze what kind of predictions can be achieved purely from the spatial bias without any knowledge of the image that is being viewed, and evaluate how this lower bound is influenced by the number of subjects and images available for its estimation. Secondly, we describe a method for computing an upper bound for model performance that is based on 'inter-subject consistency' and investigate in how far it depends on the number of subjects used for its computation.

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

Estimating the lower bound for fixation selection models

A way to estimate the lower bound for performance of fixation selection models is to compute the predictive power of the spatial bias. This prediction does not exploit information specific to the image or subject whose fixations

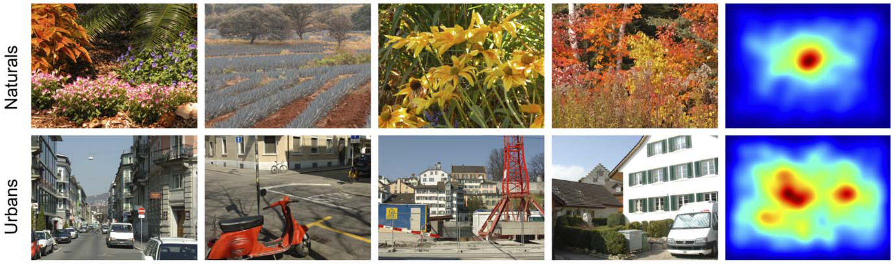

Figure 6.2: Four representative exemplary stimuli from each category used in the eye-tracking study. The top row shows natural scenes, the bottom row shows examples from the urban scenes. The right-most panels depict the spatial distribution of the first 15 fixations across all 64 images and 48 subjects in the two categories. On the natural scenes, there is a rather strong central fixation bias, while on the urban images fixations are more spread out.

are being predicted. Thus it has to be surpassed by any valuable model of fixation selection. Here, we take into account that the spatial bias varies between different image classes (Figure 6.2). We estimate the lower bound for NSS and AUC as the best representatives of central tendency measures and classification measures. As the results for AUC and NSS are qualitatively very similar, only the former is further considered here. More details on NSS results can be found as reference values in Materials and Methods: Reference values for spatial bias and inter-subject consistency. Since we explicitly wish to consider small data sets, KL divergence is not suitable here (but see Materials and Methods). To obtain a better understanding of the reliability of the lower bound, we investigate the dependence of the estimation quality on the number of subjects and images used. Specifically, we compute the lower bound by predicting fixation patterns of one subject on one image (the test set) with fixation data from other subjects on other images (the training set). To predict fixations in the test set, we construct an FDM from the training set and interpret it as a prediction for fixations in the test set. To quantify the quality of this prediction, we compute the AUC and NSS between the calculated FDM and fixations in the test set. To assess the dependence of the spatial bias estimation quality on data set size, we vary the number of images and subjects used to create the FDM. In detail, we individually increase the number of subjects and images in the training set exponentially from 1 to the maximum in seven steps ($N_{img} \in \{1, 2, 4, 8, 16, 32, 63\}$; $N_{sub} \in \{1, 2, 4, 7, 13, 25, 47\}$). For each of the 49 combinations, we use every image and subject combination as the test set 47 times such that each of the repetitions is one random sample of images and subjects for the training set.

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

The spatial bias depends on the image category (Figure 6.3, naturals and urban scenes left and right respectively, $p < 0.0001$). Furthermore, an increasing number of subjects (Figure 6.3, rows of large matrix, $p < 0.0001$) and images (Figure 6.3, columns of large matrix, $p < 0.0001$) significantly increase the predictive power of the spatial bias estimate (three factorial ANOVA, category X number of subjects X number of images). For natural scenes (left) the increase is steeper than for urban scenes (right) and thereby suggests that eye-movement patterns across subjects and stimuli are more similar during the viewing of natural scenes. The predictive power of the spatial bias estimate reached for the maximum number of subjects is surprisingly high (AUC of 0.729, 0.673 for naturals and urban scenes respectively) and poses a challenging lower bound for prediction performance. The predictive power of the spatial bias estimate increases extremely slowly when more than 32 images and 25 subjects are used, implying that the estimation becomes reliable at this point. A smaller number of subjects can be compensated by a larger image set and vice versa. However, using too few data leads to a danger of underestimating the lower bound and thereby overestimating one's model quality. In conclusion, the reliability of the lower bound estimation depends on the size of the data set; for all practical purposes, 32 images and 25 subjects seem to be sufficient for a reliable estimate.

Estimating the upper bound for fixation selection models

To derive the upper bound for fixation selection models, we estimate the inter-subject consistency analogously to the spatial bias reliability. The rationale is that, due to variance across subjects, models that do not account for individual idiosyncrasies cannot perform perfectly. Therefore, comparing model scores to a score obtained by predicting fixations from one subject with other subjects provides an intuitive normalization. If the model score and inter-subject consistency are equal, the model predicts a new subject's fixations as well as other subjects' fixations would. In the following, we investigate the dependence of inter-subject consistency on the number of subjects used for the prediction. To estimate inter-subject consistency, we first separate subjects into a test and a training set and compute an FDM from the training set. Then, we measure how

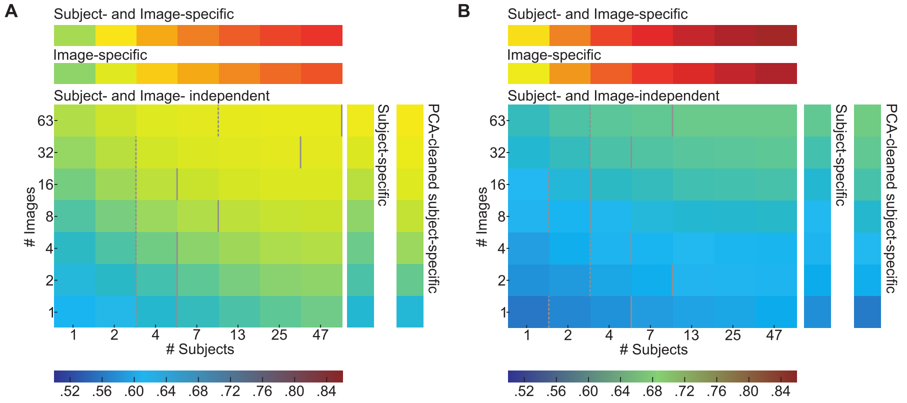

Figure 6.3: Estimation of lower and upper bounds for natural (a) and urban scenes (b). All data shown are AUC values averaged over all predictions of single subjects on single images in a given parameter combination. The predictions are based on a spatial bias (large matrix, 'Subject and Image independent'), a subject-specific bias (column next to the matrix, 'Subject-specific'), a PCA-cleaned subject-specific bias (rightmost column), an image-specific bias (row above the matrix, 'Image-specific', also referred to as inter-subject consistency) and the combination of image and subject-specific bias (topmost row). The 'Subject and Image independent' scores depend on the number of subjects and images used for the prediction and represent a lower bound for fixation selection models. The 'Image-specific' scores also depend on the number of images and yield an upper bound for fixation selection models. Comparing 'Subject-specific' and Subject and Image independent reveals the effect of using a subject-specific bias. The dashed lines indicate at what subject group size the subject-specific bias stops being significantly better than the spatial bias (paired t-test, $p > 0.05$). The subject-specific bias is not significantly different from the spatial bias between the dashed and solid lines. See main text for more detailed descriptions.

well this FDM predicts the one subject in the test set. In contrast to above, the images in test and training sets are identical. To obtain a maximally accurate estimate of the training set size for which inter-subject consistency saturates, the number of subjects in the training set is increased in steps of one. Similar to the procedure above, we use every subject and image combination 47 times as test set for every possible number of subjects in the training set. For each of the 47 repetitions a random set of training subjects is drawn. The cases where only one or two training subjects are drawn are explicitly balanced across test subjects. In the following, we report the mean AUC over test subjects, test images and repetitions as a measure of inter-subject consistency. As expected, the inter-subject consistency increases with the number of subjects in the training set (Figure 6.3, second row from top 'image-specific' in panels A and B, $p < 0.0001$; one factorial ANOVA with number of subjects as factor; additional datapoints omitted for clarity). With the maximum number of training subjects, AUC is 0.802 for naturals and 0.846 for urbans. In contrast to the pure spatial bias predictions, predictability is higher for urbans than for naturals. This results in a dynamic range of the AUC between lower and upper bound of 0.073 and 0.173 for naturals and urbans respectively. Looking at the development of inter-subject consistency with increasing subject set size, it is reasonable to assume that further increasing the training set would not have a strong effect. The second derivative of the curve is always negative, suggesting that the curve saturates. For example, from 20 to 21 subjects, the increase is 0.001, from 40 to 41 it is only 0.0002. Thus, for all practical purposes, the inter-subject consistency of about 20 subjects constitutes an upper bound for generic models of fixation selection in free viewing tasks.

Subject-specific spatial bias

To investigate the importance of subject idiosyncrasies for the prediction of fixation locations, we examine whether knowledge of a subject-specific spatial bias is more valuable than knowledge of the bias of other subjects. To that end, we estimate how well a subject-specific spatial bias predicts fixations of the same subject on other images. We proceed as before and predict fixations in the test set with an FDM based on fixations in the training set. For every combination of the number of predicting images, test subject, and test image, we use 63 different training sets. The images in the different training sets are randomly sampled and the subject is the same in training and test set. The random samples are balanced explicitly if there are only one or two images in the training set. Analogous to the generic spatial bias, the subject-specific spatial bias's predictive power is dependent on the number of images used for estimation (Figure 6.3; vertical bar 'subject-specific' directly to the right of the large matrix in panel A and B, $p < 0.001$, ANOVA with number of images as the only factor). For any number of images, the subject-specific spatial bias is more predictive than the predictive power of a single independent subject (Figure 6.3 compare left-most column in the central square to vertical column directly to the right). However, it is not higher than the predictive power of the best spatial bias, obtained from a set of 47 independent subjects (Figure 6.3 compare right-most column in the central square to vertical column directly to the right). With the exception of 63 images from the 'natural' category, the bias from a large number of subjects achieves better performance than the subject-specific bias. The exact number of subjects that is needed to achieve better performance than the subject-specific spatial bias depends on the number of images (see dashed lines in Figure 6.3). The improvement in AUC over a generic prediction based on a single independent subject ranges from 0.009 on urbans and 0.021 on naturals for a single image to 0.017 on urbans and 0.029 on naturals for 63 images. The increase in predictive power of the spatial bias achieved through incorporating subject-specific information might appear small, but it is a sizable fraction of the dynamic range between lower and upper limit (0.073 and 0.173 naturals and urbans respectively), and significant for all numbers of training images (paired T-tests over 48 subjects, $p < 0.001$).

Combining the positive effects of knowing the correct subject and knowing many subjects

We have seen that the prediction of the spatial bias from one independent subject can be improved on in two ways. By incorporating information from more independent subjects (see Properties of fixation data), reducing the uncertainty in the estimation of the true spatial bias, or by using subject-specific information (see Properties of fixation data). Both improvements have effects of similar sizes. It seems possible that combining both methods would allow an even better prediction. We hypothesize that the spatial bias of a large set of subjects consists of certain identifiable components, to which individual subjects contribute with different strengths. In that case, it should be possible to express an individual subject's spatial bias as a combination of these components. Such an approach would be more reliable, because the components can be estimated from many different subjects, effectively reducing the noise in the estimate. To identify these components, we compute the spatial bias for all training subjects on a given number of images, and perform a principal components analysis (PCA) on these biases. Figure 6.4 A,B shows the first 12 principal components of an exemplary case, which are the directions where the spatial bias varies most over subjects. Importantly, the amount of variance explained by the components drops rapidly (see Figure 6.4C). Hence the first few components explain the larger part of variance of the data and the remainder is increasingly noisy and uninformative. To enhance the reliability of the estimate, we only keep the first 5 components. We incorporate subject-specific traits by finding subject-individual weights for the components. These weights are computed by regressing the components onto the subject-specific bias, which is computed on all images in the training set. Figure 6.4D illustrates the subject-specific weighting of the multi-subject spatial bias. This combines the subject specific information and the statistical reliability of a large data base.

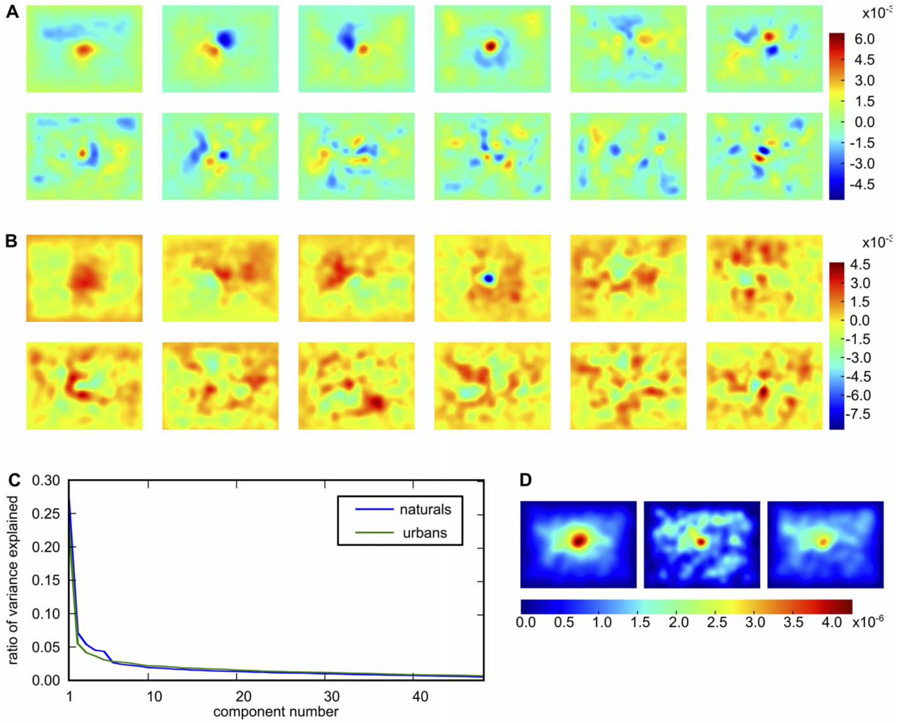

Figure 6.4: PCA-based cleaning of a subject-specific spatial bias. Panels A and B show the first 12 principal components respectively for naturals and urbans. For demonstration purposes, the underlying subject biases were computed with fixation data from all images and all subjects. Please note that the sign of the principal components is arbitrary. Panel C shows that the variance explained by each component drops dramatically. This, and the fact that the first 5 components carry some interpretable meaning, led us to choose the first five components for the cleaning of the subject-specific bias. Panel D shows an example of this. The left plot shows the spatial bias of all other subjects, the center one the subject-specific bias and the right plot shows the result of reconstructing the subject-specific bias with the first five principal components.

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

First we investigate the improvement in predictive power in comparison to the subject specific spatial bias (Figure 6.5). In case a single natural image is used to compute the principal components no improvement is observed. For an intermediate number of images a significant improvement (paired t-test, 48 subjects, significance level indicated by number of asterisks) compared to the subject specific spatial bias is demonstrated (Figure 6.5B upper row, significant deviation of blue dots from the horizontal axis that was the main diagonal in the original scatter plot). Testing subjects on even larger numbers of natural images leads to a smooth distribution of the spatial bias and no further improvement by PCA-cleaning is achieved. In the case of urban images an improvement is observed in a range from 4 to 63 images (Figure 6.5B lower row), which is shifted by a factor of two compared to naturals. Hence, in comparison to the subject specific spatial bias PCA-cleaning boosts performance by a modest degree for the case of testing with an intermediate number of images. Second, we compare prediction performance of PCA-cleaned individual spatial bias to the average obtained by a large number of subjects. Here we observe a small but significant improvement only for a larger number of images (Figure 6.5B significant deviation of red dots from the horizontal axis). The small effect size might be expected because there is already so little noise in the spatial bias for one subject. Thus, the predictive power of the generic spatial bias is already very high, leaving little room for improvement. On the other hand, the results for a small number of images illustrate that the PCA cleaning requires a certain amount of data to work properly. There is a possibility that the subject specific weights do not contribute to the observed effect, but that PCA-cleaning is only effective by removing noisy components. To control for this we repeated the same analysis but omitted the subject-specific weighting and instead weighted the components with their eigenvalues obtained from the PCA. This does not lead to a change in predictive power compared to the pure spatial bias (paired T-test, $p > 0.2$; data not shown). In summary, the PCA cleaned subject-specific spatial bias estimate combines the positive effects of reliable bias estimation and exploiting subject-specific traits.

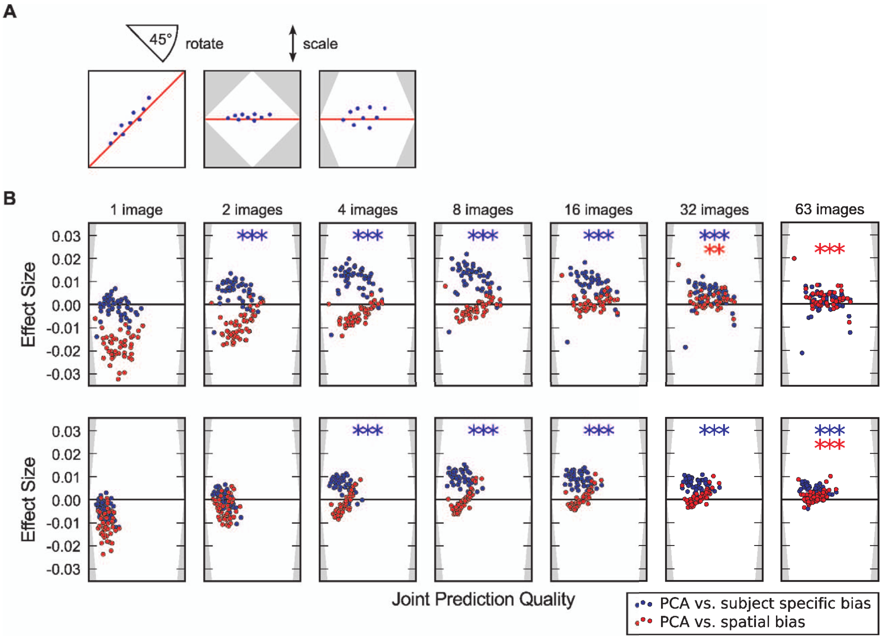

Figure 6.5: The effect of using a subject-specific PCA cleaned bias for prediction. Panel A explains how the plots in B come about. We scatter the AUC score for predicting individual subjects averaged over images and repetitions with the PCA-cleaned bias against either the scores for the subject-specific or average spatial bias. For better visibility we rotate the plot by 45° degrees and scale it. This causes the x-axis to become a measure of how well a subject can be predicted with either method and the y-axis becomes a measure of effect size, i.e. how much the prediction improves by application of the PCA. Please note, the y-axis is labeled such that it indicates the difference between the two scores and not the distance to the diagonal. To make the effects more visible we scale the y-axis to include the relevant range. The blue dots compare the effect of using PCA-cleaning to the subject-specific bias. It can be seen that in both categories the effect of the PCA depends on the number of images. The asterisks indicate that the effect size is significantly larger than zero (paired t-test, $* \widehat{=} p < 0.05$, $** \widehat{=} p < 0.01$, $*** \widehat{=} p < 0.001$).

Predicting better than perfect: combining subject- and image-specific biases

The previous section showed that subject-specific predictions can improve the already good prediction of a large group of subjects in the domain of the spatial bias. After estimating the upper bound for fixation selection models, we established that the inter-subject consistency marks an upper bound for prediction quality of subject independent models. Given these observations, the question arises whether subject-specific models can surpass the inter-subject consistency bound. As a proof of concept, we combine inter-subject predictions with the subject-specific spatial bias as a simple form of subject-specific information, and analyze if this procedure can lead to a better prediction. We assume that viewing behavior on an image is driven partly by a subject-specific spatial bias and by image properties, i.e. the inter-subject prediction contains both components. The idea is to replace the general spatial bias in the inter-subject fixation density map with a subject-specific spatial bias while keeping the image dependent part. To achieve this, we first compute the fixation density map of all training subjects on the image in question, i.e. the inter-subject prediction. Second, we remove the general spatial bias by dividing the inter-subject prediction point-wise through the training subjects' spatial bias computed on all other images. To arrive at a prediction, we multiply the resulting image-specific bias point-wise with the spatial bias of the predicted subject. Finally, we normalize the resulting map to unit mass and evaluate how well it predicts the fixations of our test subject. We use the same cross-validation procedure as for the generic inter-subject predictions, but limit the computations to the logarithmically increasing training set sizes used for the spatial bias evaluation. Inter-subject consistency is recomputed for these new training sets to allow paired tests between subject-specific and generic predictions. The results show a small but significant effect on naturals ($p < 0.001$, paired t-test for 4 or more subjects. See Figure 6.6). For example, the improvement for 47 subjects is a mean AUC increase from 0.809 to 0.815. There is no significant effect on urbans (paired t-test, $p > 0.2$ for all numbers of subjects). The difference between categories can probably be explained by the fact that the spatial bias has less predictive power for urbans and that the inter-subject consistency is already higher in urbans. We conclude that the combination of subject-specific information and image-specific information can surpass the inter-subject consistency upper bound on natural but not on urban images.

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

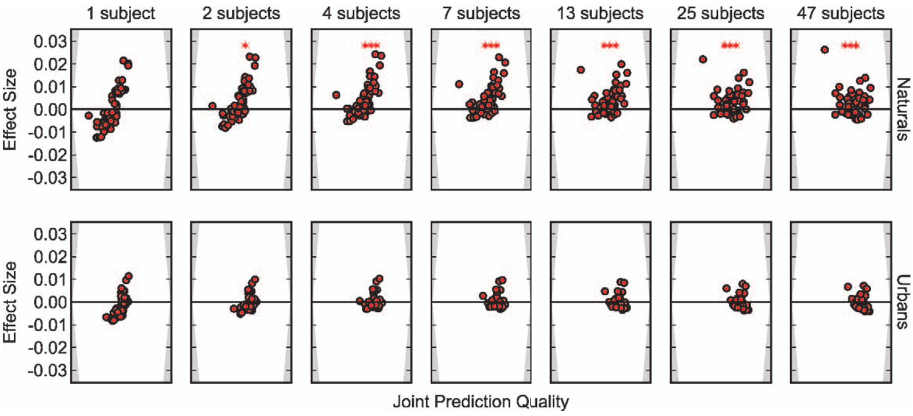

Figure 6.6: Combining a subject-specific and image-specific spatial bias for a better than perfect prediction. The plots are produced as in Figure 6.5. The effect depends on the number of subjects that enter the bias estimation and the image category. For natural scenes, a statically significant effect (paired t-test, *** = $p < 0.001$) can be seen when four subjects or more are used. The effect cannot be seen for urban scenes, which might be explained by the low predictive power of the subject-specific bias compared to the high predictive power of the image-specific bias on urbans.

6.1.3 Discussion

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

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

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

Secondly, in our data we found that the category where the spatial bias is weaker (urbans) has a stronger inter-subject consistency. This double-dissociation has important consequences for the evaluation of fixation selection models. One and the same model, incorporating both spatial bias and image statistics, may score higher on naturals than on urbans, because of the predictive power of the spatial bias. On the other hand, if a model is almost optimal and comes close to the predictive power of other subjects' fixations, it will score higher on urbans. Thus, the type of dataset the model is evaluated on will have an effect on one's judgment of model quality. As a result of this, a comparison of different models is nearly impossible if they were evaluated on different data sets, unless the upper and lower bounds for the specific datasets are explicitly given.

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

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

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

The inter-subject consistency has been used before as an upper bound for model performance, which allows for a direct comparison of our values and the ones provided in the literature. Interestingly, we found that on first sight not all values were in line with our results (Figure 6.7). However, there seems to be a consistent explanation for the deviations: All values of inter-subject consistency that lie above those found in our data were computed on data where there was an explicit task during the eye-tracking experiment (object naming (Einhäuser et al., 2008) or pedestrian search (Ehinger et al., 2009; Kanan et al., 2009)), or the stimuli material contained a wealth of high-level information (web-pages (Betz et al., 2010)). On the other hand, Hwang et al. (2009) explicitly designed their experiment to minimize high-level information by rotating the images by 90° or 180°. They report lower inter-subject consistency, but the effect of group size is in line with our results. Finally, Cerf et al. (2009) use a free viewing task similar to our experiment and obtain values almost identical to ours. We conjecture that inter-subject consistency is strongly influenced by the subjects' task and the availability of high-level information. This is also in line with the category differences found in our data (urbans > naturals), since the urban scenes provide more high-level information (e.g. man-made objects, people), as well as with category differences reported by Frey et al. (2008). Interestingly, high inter-observer consistency is not related to a large influence of the spatial bias. In our dataset, the former is higher for urban scenes while the latter is higher on natural images. A speculative explanation of this finding is that when high-level information is present in an image, it will guide the eye movements of many subjects to locations that are not necessarily in the center of the image, increasing inter-subject consistency and decreasing the influence of the spatial bias. In the absence of high-level information, subjects tend to look more towards the center of the screen, but in a less homogenous fashion. This fallback strategy leads to an increased spatial bias and decreased inter-subject consistency. Further evidence for this hypothesis comes from eye-tracking studies with pink-noise stimuli, which are completely devoid of high-level information and where the influence of the spatial bias is comparatively large (Açik et al., 2010). Our analyses of subject idiosyncrasies relative to our established bounds showed that the increase in performance, although statistically significant, is very small. In the case where data from 63 images are used, knowing the spatial bias of a specific subject is as good as knowing more than 7 other subjects on naturals, or knowing more than 2 other subjects on urbans. The smaller effect for urbans fits the observation that inter-subject consistency is higher in that category, making knowledge about a specific subject less unique. This relates to a possible reason for the small overall effect size in both categories: Açik et al. (2010) show that different demographic subject groups have remarkably different viewing behavior. Specifically, explorativeness, a property that is closely related to the spatial bias, decreases with increasing age. Our subject group consisted exclusively of university students between 19 and 28 years of age. Thus it can be expected that the effect of knowing the subject to be predicted would be much larger in a more heterogeneous subject group with lower inter-subject consistency. In such a scenario, the improvement caused by PCA-cleaning demonstrated in the present study could become more relevant. In general, the PCA-cleaning requires fixation data on a fair number of images for a good signal to noise ratio. In practice, the principal components could be determined from a large set of subjects and images recorded in a baseline study. It may then be possible to tailor a clean subject-specific spatial bias based on fixations from the subject of interest on few images. This technique may be useful in a modeling context, when the goal is to fine-tune a generic model for predicting individual subjects' fixations.

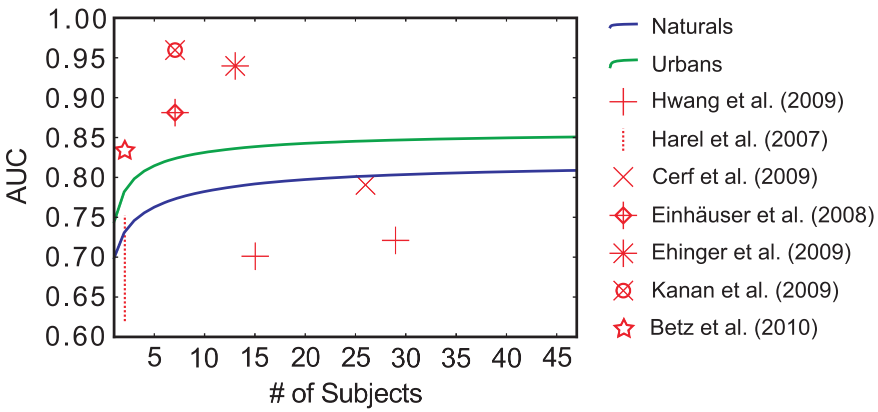

Figure 6.7: A comparison of inter-subject consistency AUC in different studies. Green and blue lines show the dependence of inter subject consistency on the number of subjects in our data. The symbols show inter-subject consistency values reported in other studies. All studies that reported higher values used either stimuli that contained a wealth of high-level information or employed a specific task. Cerf et al. (2009) also use a free viewing task and are compatible with our findings. Harel et al. (2007) only report a range of values (read from a figure). Notably, Hwang et al. (2009) use image rotations to diminish top-down influences and observe lower inter-subject consistency

The spatial bias is of course only one feature of viewing behavior where subject idiosyncrasies can play a role. There are possibly many different ways to incorporate these into a model of fixation selection. An obvious candidate would be the relative importance of different image features in a bottom-up model. Whether subject-specific modeling of feature weights has a positive effect is an interesting question for further research, but goes beyond the scope of this article.

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

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

6.1.4 Materials and Methods

Theoretical maximum value for AUC

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

changing the bounds

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

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

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

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

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

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

Proof of AUC linearity

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

$$AUC(P_1 \uplus P_2, N) = \frac{|P_1|}{|P_1 \uplus P_2|} \cdot AUC(P_1, N) + \frac{|P_2|}{|P_1 \uplus P_2|} \cdot AUC(P_2, N) \tag{6.1}$$

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

$$AUC(P, N) = \sum_{i=2}^{n} \frac{TPR(t_i) + TPR(t_{i-1})}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \tag{6.2}$$

$$TPR(t) = \frac{|\{x | x \in P \wedge x \geq t\}|}{|P|} \tag{6.3}$$

$$FPR(t) = \frac{|\{x | x \in N \wedge x \geq t\}|}{|N|} \tag{6.4}$$

$$t_1 = \infty,\ i < k \Rightarrow t_i > t_k,\ t_n = -\infty \tag{6.5}$$

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

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

That implies for any set $T \subseteq S$

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

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

The Lemma reduces (6.1) to

$$AUC(P,N) = \frac{|P|-1}{|P|} \cdot AUC(P \setminus \{p\}, N) + \frac{1}{|P|} \cdot AUC(\{p\}, N) \tag{6.6}$$

From (6.3) follows

$$\forall p \in P, \quad TPR_{P \setminus \{p\}}(t) = \begin{cases} \frac{TPR_P(t) \cdot |P| - 1}{|P|-1} & \text{if } t \leq p \\ \frac{TPR_P(t) \cdot |P|}{|P|-1} & \text{if } t > p \end{cases} \tag{6.7}$$

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

$$\begin{aligned}
AUC(P \setminus \{p\}, N) &= \sum_{i=2}^{k} \frac{(TPR_P(t_i) + TPR_P(t_{i-1})) \cdot |P|}{2 \cdot (|P|-1)} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&\quad + \sum_{i=k+1}^{n} \frac{(TPR_P(t_i) + TPR_P(t_{i-1})) \cdot |P| - 2}{2 \cdot (|P|-1)} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&= \frac{|P|}{|P|-1} \cdot \sum_{i=2}^{k} \frac{TPR_P(t_{i-1}) + TPR_P(t_i)}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&\quad + \frac{|P|}{|P|-1} \cdot \sum_{i=k+1}^{n} \frac{TPR_P(t_{i-1}) + TPR_P(t_i)}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\
&\quad - \frac{1}{|P|-1} \cdot \sum_{i=k+1}^{n} FPR(t_i) - FPR(t_{i-1}) \\
&= \frac{|P|}{|P|-1} \cdot AUC(P,N) - \frac{FPR(t_n) - FPR(t_k)}{|P|-1} \\
&= \frac{|P|}{|P|-1} \cdot AUC(P,N) - \frac{1 - FPR(t_k)}{|P|-1}
\end{aligned} \tag{6.8}$$

and

$$\begin{aligned} AUC(\{p\},N) &= \sum_{i=2}^{n} \frac{TPR(t_{i-1}) + TPR_P(t_i)}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\ &= \sum_{i=2}^{k} 0 + \sum_{i=k}^{n} \frac{1+1}{2} \cdot (FPR(t_i) - FPR(t_{i-1})) \\ &= FPR(t_n) - FPR(t_k) \\ &= 1 - FPR(t_k) \end{aligned} \tag{6.9}$$

Using (6.8) and (6.9) it is easy to see that (6.6) is true, proving (6.1).

Computational details of AUC analysis

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

Fixation density map estimation

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

Correction of KL divergence for small samples

The KL-divergence can be expressed in terms of Information Entropy, and for Information Entropy it is known that it systematically depends on the sample size (Hausser & Strimmer, 2009; Miller, 1955; Nemenman et al., 2002). These observations lead us to suspect that the KL-divergence is also biased, which is problematic when different models are evaluated against densities estimated from different sample sizes. We carry out two simulations to investigate the size of this potential confound. First, we treat the overall spatial bias as our prediction. We then take a random sample of fixations from the set that constitutes the spatial bias and repeatedly calculate the KL-divergence between the FDM of our sample and our prediction. If the sample gives a perfect estimate of the distribution it was drawn from the KL-divergence should be zero. We increase the number of fixations per sample from 6 to 800 in steps of 2, and draw 1000 samples of every size. Since discrete Entropy estimates are also strongly influenced by the binning of the probability density function, we do not use our standard procedure for computing fixation density maps. Instead, we sort the data into a grid of 16x12 bins (leading to N=192). The number of grid cells was selected such that the area of each bin is equal to the area of a circle of diameter two degrees of visual angle. These FDMs are not smoothed, since they already have a coarse resolution. In a second simulation, we take a normal distribution with specified parameters ($\mu = 0$, $\sigma = 1$) as our prediction and sample our data from a different normal distribution ($\mu = 2$, $\sigma = 1$). In this case the true KL-divergence can be determined analytically and the KL-divergence computed from different sample sizes can be compared to this target value. We proceed in the same way as before and increase the sample size from 6 to 800 in steps of 2 and draw 1000 samples of every size. Densities are estimated as histograms with 100 bins. In both cases the estimated KL-divergence was higher than the analytical value. The difference between mean estimated KL-value and analytical value decreased with increasing sample size (the results for simulation 1 are depicted in Figure 6.8A; results for simulation 2 were similar). Thus, comparing models evaluated on different data set sizes is difficult. One approach to cope with the sample size dependence of the estimate is to keep the sample size constant in every comparison by randomly sampling as many fixations from each data set as are available from the smallest one. However, if the size of a novel data set is comparably small and previous model evaluations were performed on a larger and inaccessible data set, it is not possible to reduce the larger data set. Thus, to foster comparisons between different studies, it would be advantageous to be able to directly correct for the bias introduced by sample size. There are multiple methods that try to improve the estimate of entropy values (recall that KL-divergence is directly dependent on the Entropy estimates), as compared to the typically-used maximum likelihood approach. We therefore investigate the applicability to fixation data of several methods (Chao & Shen, 2003; Hausser & Strimmer, 2009; Holste et al., 1998; Krichevsky & Trofimov, 2002; Miller, 1955; Nemenman et al., 2002; Schürmann & Grassberger, 1996; Trybula, 1958), for which Hausser & Strimmer (2009) provide an implementation. To compare the efficacy of the different approaches, we carried out simulations in which we estimated the entropy of differently sized samples from the general spatial bias. In addition to the direct relevance for the calculation of KL-divergence, an important advantage of an unbiased entropy estimate is that entropy can be used to characterize viewing behavior (Açik et al., 2010; Gilland, 2008; Recarte & Nunes, 2000). It is therefore relevant to have an unbiased estimate, e.g. for comparing different experimental conditions with different amount of fixations. The simulations follow the pattern that we used for determining the sample size dependence in KL-divergence. Due to the large number of different correction methods compared, we only draw 200 samples of each size to reduce computational load. We compare estimates for different sample sizes to the entropy of all fixations in one category ($N_{naturals} = 43295$,$N_{urbans} = 44753$), assuming that the estimate is nearly unbiased with such a large sample size. The simulations show that it is in principle possible to improve the entropy estimate. However even in the best case, the number of samples required for a reasonable estimate is approximately half the number of bins of the fixation density map. This is a large improvement over uncorrected Entropy, which requires the number of data points to be at least equal to the number of bins squared. The fixation densities in our simulations were down sampled to 169 bins. Considering that FDMs are typically smoothed with a 2deg FWHM Gaussian kernel, the effective resolution of a FDM is already much lower than the number of pixels suggests, making the down sampling tenable. Overall the correction methods proposed by Chao & Shen (2003) and Jeffreys Krichevsky & Trofimov (2002) work best of all tested methods. To yield a correction method for the KL-divergence, its Entropy and cross-Entropy terms have to be corrected. Starting with Chao-Shen, the pure entropy term can straightforwardly be corrected. Moreover, if we presuppose that a model output corresponds to a correct probability density (Q), we can also apply Chao-Shen to correct the cross Entropy $H(P,Q)$. Here, we use

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

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

to compute the corrected KL-divergence, where pcs and Coverage are the two Chao-Shen correction terms (see Chao & Shen (2003)). The Jeffreys correction can simply be applied by adding 1/2 to the cell counts of the FDM before it is normalized to unit mass. To validate applicability of Chao-Shen in the case of KL, we repeated the simulations for the maximum likelihood KL-divergence estimation but used the Chao Shen and Jeffreys corrected estimation. As shown in Figure 6.8B, the correction substantially improves the KL estimates as compared to the maximum likelihood version. The Jeffreys correction works well on our data, which is in part due to the fact that our distribution does not deviate too much from the uniform prior assumed by the correction method. If there are strong reasons to believe that one's data deviate much from a uniform distribution, one should therefore be careful with this correction. The Chao

Shen correction is very close to the true KL-divergence between the underlying distributions at a sample size of about $N/2$.

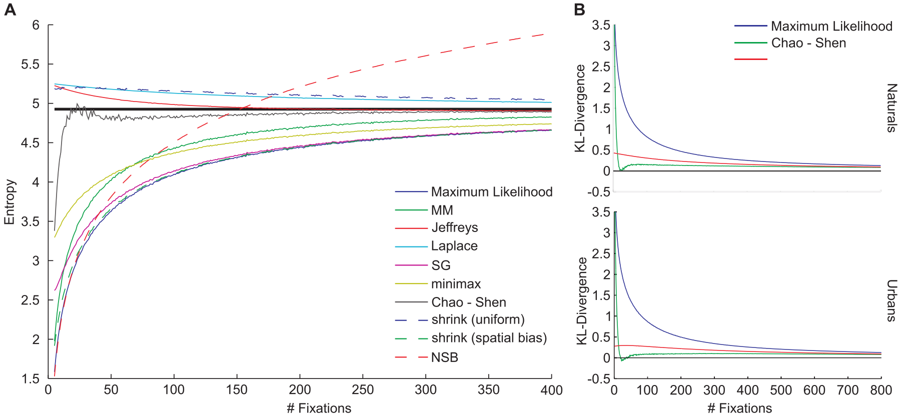

Figure 6.8: The effect of sample size on the KL-divergence. A. Performance of different methods to remove the sample size bias from entropy estimates in a simulation using eye-tracking data. The bold line shows the maximum likelihood entropy estimate computed on the entire data set ($N > 40000$)and can be interpreted as ground truth. The Chao-Shen and Jeffreys correction methods approach the target value with the lowest number of samples. Descriptions of the individual methods can be found in Chao & Shen (2003) (Chao-Shen), Hausser & Strimmer (2009) (shrink), Holste et al. (1998) (Laplace), Krichevsky & Trofimov (2002) (Jeffreys), Miller (1955) (MM), Nemenman et al. (2002) (NSB), Schürmann & Grassberger (1996) (SG), Trybula (1958) (minimax). B. Sample size dependence of different KL-divergence estimation methods. The standard maximum likelihood method shows a strong positive bias for small samples, both correction methods tested can reduce this problem for sample sizes of ca. half the number of bins in the estimated distributions or larger.

Description of the eye-tracking study

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

Reference values for spatial bias and inter-subject consistency

Here we report numeric AUC (Table 6.2 and 6.3) and NSS (Table 6.4 and 6.5) values for predicting fixations of one subject on one image with a subject and image independent spatial bias (estimated lower bound, see Properties of fixation data) and with an image-specific bias (inter-subject consistency, estimated upper bound, see Properties of fixation data). All reported values are means across cross-validation runs, as described in Properties of fixation data. So far we omitted the computation of upper and lower KL-divergence boundaries. Testing the estimation reliability by changing the number of subjects and images in the training set would be confounded by the different numbers of fixations in the training set (our correction methods are intended for controlling the test set and thus do not apply here). To nevertheless be able to report sensible reference bounds, we restrict ourselves to a large training set size such that the influence of different amounts of fixations in the training set is small. In detail, we pick out one row (63 images, varying the number of subjects for prediction) and one column (25 subjects, varying the number of images for prediction) of the subject and image independent predictions. This leaves either many images or many subjects in the training set, such that there are at least 375 fixations in the training set. To furthermore minimize the effect of different amounts of fixations in the training set, we bin the screen into 12x16 squares. The test set always contains fixations from 23 subjects, we omit the case where more than 25 subjects are in the training set, such that the number of fixations is constant at 345 fixations. The evaluation of the entropy correction methods has shown that with this amount of fixations and dimensionality of the probability density map, no correction for different amounts of fixations is needed. We also compute the inter subject consistency for predicting 23 subjects with data from the remaining 25 subjects for every image and 48*63 random assignments of subjects into test and training set. Table 6.6 and 6.7 report the mean over images and random assignments.

Table 6.2: AUC values for natural scenes

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

Table 6.3: AUC values for urban scenes

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

Table 6.4: NSS values for natural scenes

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

Table 6.5: NSS values for urban scenes

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

Table 6.6: KL-divergence values for natural scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25	47
Image-specific						0.424	
63	0.900	0.763	0.707	0.684	0.670	0.662	
32						0.678	
16						0.707	
8						0.757	
4						0.850	
2						1.037	
1						1.467	

Open-source python toolbox

To foster model comparison and ease reproduction of our results we provide a free open-source python toolbox. It allows to conveniently represent fixation data and can be used to estimate the lower and upper bound for fixation selection models on a given data set. Implementations of AUC and KL-divergence, as well as a few other measures, are also contained in the toolbox. The toolbox

Table 6.7: KL-divergence values for urban scenes

Nr. of subjects → Nr. of images ↓	1	2	4	7	13	25	47
Image-specific						0.364	
63	1.274	1.190	1.153	1.141	1.137	1.139	
32						1.153	
16						1.201	
8						1.298	
4						1.501	
2						1.981	
1						3.280	

can be accessed at https://github.com/nwilming/ocupy. Furthermore, the data used in the current work is available from the authors upon request.

6.1.5 Acknowledgments

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

6.2 A unifying approach to high- and low-level cognition[2]

Abstract In our endeavor to better understand the mind, separate research areas concentrate on low- and high-level cognitive processes. This tradition is often understood as being rooted in a principled division of functions and algorithms supported by different brain areas. Indeed, a rapidly growing number of studies report functional specializations of different cortical regions and are therefore interpreted as supporting this conclusion. Here, we challenge this view with a three-step argument that relies on a critical analysis of prime examples from low-level cognition (object recognition) and high-level cognition (predictive analogies): First, we argue that a homogeneously structured cortical module may subserve different functions contingent on the properties of its afferent signals. Specifically, optimizing statistical properties of sensory representations, such as sparseness, stability, and predictability, provides a normative model of sensory areas at different levels. Second, this optimization process naturally leads to the emergence of invariant representations and a gradual change of sensory to action-related representations. From a bottom-up view, cortical modules convey information on increasingly invariant sensory representations. Information flowing top-down, however, supports invariant action representations and predictions of afferent changes induced by the afforded actions. Third, to bring together low- and high-level cognition, we argue that our previously introduced invariant actions are at the core of predictive analogies and therefore function as a general cognitive mechanism. Together, these three steps establish the view of 'optimally predictive active representations' as a unified description and postulate a uniform cortical substrate and functional mechanisms for low-level and high-level cognitive processes.

Section published as: König, P., Kühnberger, K.U., & Kietzmann, T.C. (2014). A unifying approach to high- and low-level cognition. In Models, Simulations, and the Reduction of Complexity. (pp. 117-141). De Gruyter. ISBN: 978-3-11-031360-4

[2] This section was published as a book chapter together with Peter König and Kai-Uwe Kühnberger. See Publication List for details.

Acknowledgments

First and foremost, I wish to express my deep gratitude to Professor Peter König and Professor Frank Tong, my PhD supervisors, for their continuous and encouraging support, outstanding advice, and for sharing their never-ending enthusiasm and scientific curiosity. Without your guidance, this dissertation would not have been possible. I will be forever grateful for the lessons learned. I would also like to thank you for giving me complete scientific freedom and financial security, despite the fact that not all my experiments were at the core of your own research agenda.

Part of the experimental work was performed in collaboration with Professors Andreas Engel, and Randolph Blake, to whom I would like to extend my gratitude for their valuable advice and for their support of the MEG measurements in Hamburg (Prof. Engel) and TMS experiments in Nashville (Prof. Blake).

During the years of my PhD, I had the pleasure of working in the Neuro-BioPsychology Group in Osnabrück and the Tonglab in Nashville. I would like to thank the members of these two wonderful labs for providing a scientific environment in which new thoughts are appreciated and in which scientific work is as much fun as it should be. In particular, I wish to thank Niklas Wilming, Robert Muil, Jose Ossandon, Benedikt Ehinger, Torsten Betz, and Saskia Nagel from Osnabrück, as well as Sam Ling, Jascha Swisher, and Mike Pratte from Nashville for your (y'alls) encouragement, enthusiasm, critical thoughts and friendship.

I would furthermore like to thank all co-authors, students, and helping hands who contributed to the different projects, as well as Niklas Wilming and Robert Muil for proof reading parts of this document. My special thanks go to the technical administrators of the IKW for providing a highly functional and powerful computational infrastructure, and for helping out instantly whenever problems occur. I would also like to thank the Graduate Program of the Institute of Cognitive Science, and the Fulbright Commission for their financial support.

Finally, I would like to thank my family: my parents Ingolf and Christine, my brothers Jan and Sven Ole, and my Julia for their endless encouragement, support, and patience so dearly needed. I cannot thank you enough for the love you are giving me.

Disclaimer

All experiments reported in this thesis conform with the Declaration of Helsinki and have been approved by the ethics committees of the respective institution (University of Osnabrück, University Medical Center Hamburg Eppendorf, Vanderbilt University). 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. Furthermore, this thesis was neither published in Germany nor abroad, except the parts indicated above, and has not been used to fulfill any other examination requirements.

Bibliography

Açik, A., Onat, S., Schumann, F., Einäuser, W., & König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision research*, 49(12), 1541–1553. Cited on page 164

Açik, A., Sarwary, A., Schultze-Kraft, R., Onat, S., & König, P. (2010). Developmental changes in natural viewing behavior: bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology*, 2(0). Cited on pages 185, 194

Aguirre, G.K., Singh, R., & D'Esposito, M. (1999). Stimulus inversion and the responses of face and object-sensitive cortical areas. *Neuroreport*, 10(1), 189–94. Cited on page 70

Aguirre, G.K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron*, 21(2), 373–83. Cited on page 67

Ahissar, M. & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10), 457–64. Cited on pages 18, 102

Anastasi, J. & Rhodes, M. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, 12(6), 1043–7. Cited on pages 22, 154

Andresen, D.R., Vinberg, J., & Grill-Spector, K. (2009). The representation of object viewpoint in human visual cortex. *NeuroImage*, 45(2), 522–36. Cited on page 75

Antzoulatos, E.G. & Miller, E.K. (2011). Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. *Neuron*, 71(2), 243–9. Cited on page 135

Ashby, F.G. & Gott, R.E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of experimental psychology. Learning, memory, and cognition*, 14(1), 33–53. Cited on page 17

Ashby, F.G. & O'Brien, J.B. (2005). Category learning and multiple memory systems. *Trends in cognitive sciences*, 9(2), 83–9. Cited on page 138

Aston-Jones, G. & Cohen, J. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Neuroscience*, 28(1), 403. Cited on page 44

Axelrod, V. & Yovel, G. (2012). Hierarchical processing of face viewpoint in human visual cortex. *The Journal of neuroscience*, 32(7), 2442–52. Cited on pages 75, 82

Baddeley, R.J. & Tatler, B.W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision research*, 46(18), 2824–33. Cited on page 164

Baker, C., Behrmann, M., & Olson, C. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, 5(11), 1210–1216. Cited on page 105

Balcetis, E. & Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *Journal of personality and social psychology*, 91(4), 612–25. Cited on page 7

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5, 617–629. Cited on page 102

Bar, M. & Ullman, S. (1996). Spatial Context in Recognition. *Perception London*, 25, 343–352. Cited on page 7

Barlow, H.B.H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge. Cited on page 155

Beck, D.M., Rees, G., Frith, C.D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. *Nature Neuroscience*, 4(6), 645–50. Cited on page 43

Bentin, S., Allison, T., & Puce, A. (1996). Electrophysiological Studies of Face Perception in Humans. *Journal of Cognitive Neuroscience*, 8(6), 551–565. Cited on page 21

Berkes, P. & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5, 579–602. Cited on page 155

Betz, T., Kietzmann, T.C., Wilming, N., & König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision*, 10(3), 15. Cited on pages 4, 164, 185

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2), 115–47. Cited on page 9

Blanz, V., Tarr, M.J., & Bülthoff, H.H. (1999). What object attributes determine canonical views? *Perception*, 28(5), 575–600. Cited on page 12

Booth, M.C. & Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8(6), 510–23. Cited on page 10

Boring, E. (1930). A new ambiguous figure. *The American Journal of Psychology*, 42(3), 444–445. Cited on pages 7, 28

Botwinick, J. (1961). Husband and father-in-law: a reversible figure. *The American Journal of Psychology*, 74, 312–313. Cited on page 28

Bowers, J.S. (2009). On the Biological Plausibility of Grandmother Cells : Implications for Neural Network Theories in Psychology and Neuroscience. *Psychological Review*, 116(1), 220–251. Cited on page 11

Bracci, S. & Peelen, M.V. (2013). Body and object effectors: the organization of object representations in high-level visual cortex reflects body-object interactions. *The Journal of Neuroscience*, 33(46), 18247–58. Cited on page 157

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–6. Cited on pages 59, 138

Brascamp, J., van Boxtel, J., Knapen, T., & Blake, R. (2010). A dissociation of attention and awareness in phase-sensitive but not phase-insensitive visual channels. *Journal of Cognitive Neuroscience*, 22(10), 2326–2344. Cited on page 43

Brown, V., Huey, D., & Findlay, J. (1997). Face detection in peripheral vision: do faces pop out? *Perception London*, 26, 1555–1570. Cited on pages 103, 119

Bruce, N.D.B. & Tsotsos, J.K. (2009). Saliency, attention, and visual search: an information theoretic approach. *Journal of Vision*, 9(3), 5.1–24. Cited on page 164

Bruner, J. & Minturn, A. (1955). Perceptual identification and perceptual organization. *The Journal of General Psychology*, 53(1), 21–28. Cited on page 7

Bugelski, B. & Alampay, D. (1961). The role of frequency in developing perceptual sets. *Canadian Journal of Psychology*, 15(4), 205–211. Cited on pages 7, 28

Bülthoff, H.H. & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 89(1), 60–4. Cited on pages 9, 12, 13, 56, 78

Bülthoff, H.H., Edelman, S.Y., & Tarr, M.J. (1994). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5(3), 247–60. Cited on pages 9, 12, 13

Buswell, G. (1935). *How people look at pictures: a study of the psychology and perception in art.* University of Chicago Press, Chicago. Cited on page 4

Butko, N.J. & Movellan, J.R. (2008). I-POMDP: An infomax model of eye movement. In *7th IEEE International Conference on Development and Learning*, 1, pages 139–144. Ieee. Cited on page 164

Caharel, S., D'Arripe, O., Ramon, M., Jacques, C., & Rossion, B. (2009). Early adaptation to repeated unfamiliar faces across viewpoint changes in the right hemisphere: evidence from the N170 ERP component. *Neuropsychologia*, 47(3), 639–43. Cited on page 139

Calder, A.J., Beaver, J.D., Winston, J.S., Dolan, R.J., Jenkins, R., Eger, E., & Henson, R.N.a. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. *Current Biology*, 17(1), 20–5. Cited on page 92

Carlin, J.D., Calder, A.J., Kriegeskorte, N., Nili, H., & Rowe, J.B. (2011). A Head View-Invariant Representation of Gaze Direction in Anterior Superior Temporal Sulcus. *Current Biology*, pages 1–5. Cited on pages 90, 91

Carlson, T., Hogendoorn, H., & Kanai, R. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, 11(10), 1–17. Cited on pages 19, 125, 135

Carlson, T., Tovar, D., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10), 1–19. Cited on page 94

Cerf, M., Frady, E.P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9, 1–15. Cited on pages 161, 162, 164, 185, 186

Cerf, M., Harel, J., Einäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems*, 20, 241–248. Cited on pages 161, 164

Chan, A., Kravitz, D., Truong, S., Arizpe, J., & Baker, C. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience*, 13(4), 417–418. Cited on page 154

Chao, A. & Shen, T. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4), 429–443. Cited on pages 194, 195, 196

Chastain, G. & Burnham, C. (1975). The first glimpse determines the perception of an ambiguous figure. *Perception and Psychophysics*, 17(3), 221–224. Cited on page 44

Cichy, R.M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17, 455–462. Cited on pages 19, 94

Clark, A. (1999). An embodied cognitive science? *Trends in cognitive sciences*, 3(9), 345–351. Cited on page 157

Cohen, E.H. & Tong, F. (2013). Neural Mechanisms of Object-Based Attention. *Cerebral Cortex*. Cited on page 6

Cohen, J.D. & Tong, F. (2001). The face of controversy. *Science*, 293(5539), 2405–7. Cited on page 22

Connor, C.E. (2005). Friends and grandmothers. *Nature*, 435, 1036–1037. Cited on page 12

Conty, L., N'Diaye, K., Tijus, C., & George, N. (2007). When eye creates the contact! ERP evidence for early dissociation between direct and averted gaze motion processing. *Neuropsychologia*, 45(13), 3024–37. Cited on page 91

Cornell, J.M. (1985). Spontaneous mirror-writing in children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 39(1), 174–179. Cited on page 13

Crick, F. & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature*, 375(6527), 121–123. Cited on page 103

Cromer, J.A., Roy, J.E., & Miller, E.K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron*, 66(5), 796–807. Cited on page 135

Crouzet, S.M. & Serre, T. (2011). What are the Visual Features Underlying Rapid Object Recognition? *Frontiers in Psychology*, 2, 326. Cited on page 19

Crouzet, S.M. & Thorpe, S.J. (2011). Low-level cues and ultra-fast face detection. *Frontiers in Psychology*, 2, 342. Cited on pages 126, 135

Cukur, T., Nishimoto, S., Huth, A.G., & Gallant, J.L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6), 763–70. Cited on page 6

Dale, A.M., Fischl, B., & Sereno, M.I. (1999). Cortical surface-based analysis. I. Segmentation and Surface Reconstruction. *NeuroImage*, 9(2), 179–194. Cited on page 143

De Baene, W., Ons, B., Wagemans, J., & Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learning & Memory*, 15(9), 717–27. Cited on page 18

de Heering, A. & Rossion, B. (2008). Prolonged visual experience in adulthood modulates holistic face perception. *PloS one*, 3(5), e2317. Cited on page 22

Dehaene, S., Nakamura, K., Jobert, A., Kuroki, C., Ogawa, S., & Cohen, L. (2010). Why do children make mirror errors in reading? Neural correlates of mirror invariance in the visual word form area. *NeuroImage*, 49(2), 1837–48. Cited on pages 13, 76

Deubel, H. & Schneider, W.X. (1996). Saccade Target Selection and Object Recognition: Evidence for a Common Attentional Mechanism. *Vision Research*, 36(12), 1827–1837. Cited on page 4

Devilbiss, D., Page, M., & Waterhouse, B. (2006). Locus ceruleus regulates sensory encoding by neurons and networks in waking animals. *Journal of Neuroscience*, 26(39), 9860. Cited on page 44

DiCarlo, J.J. (2011). Untangling Object Recognition: Which neuronal population codes can explain human object recognition performance? In *Neural Computation: Population Coding of High-Level Representations*. Cited on page 2

DiCarlo, J.J. & Cox, D.D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8), 333–41. Cited on pages 11, 83

DiCarlo, J.J., Zoccolan, D., & Rust, N.C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–34. Cited on page 136

Dilks, D.D., Julian, J.B., Kubilius, J., Spelke, E.S., & Kanwisher, N. (2011). Mirror-image sensitivity and invariance in object and scene processing pathways. *Journal of Neuroscience*, 31(31), 11305–12. Cited on pages 14, 76

Donner, T.H., Siegel, M., Fries, P., & Engel, A.K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, 19(18), 1581–5. Cited on page 137

Edelman, S. (1999). *Representation and recognition in vision*. MIT Press, Cambridge, MA. Cited on page 11

Edelman, S. & Bülthoff, H.H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32(12), 2385–400. Cited on page 12

Eger, E., Henson, R.N.A., Driver, J., & Dolan, R.J. (2004). BOLD repetition decreases in object-responsive ventral visual areas depend on spatial attention. *Journal of Neurophysiology*, 92(2), 1241–7. Cited on page 76

Ehinger, K.A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling Search for People in 900 Scenes: A combined source model of eye guidance. *Visual Cognition*, 17(6-7), 945–978. Cited on pages 162, 164, 185

Einhäuser, W., Martin, K., & König, P. (2004). Are switches in perception of the necker cube related to eye position? *European Journal of Neuroscience*, 20(10), 2811–2818. Cited on page 43

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 1–26. Cited on pages 5, 6, 162, 164, 185

Einhäuser, W., Stout, J., Koch, C., & Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5), 1704–1709. Cited on pages 27, 30, 43, 44, 46, 47

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, 93(1), 79–90. Cited on pages 10, 155

Einhäuser, W., Kayser, C., König, P., & Körding, K.P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience*, 15(3), 475–86. Cited on page 155

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

Einhäuser, W., Martin, K.A.C., & König, P. (2004). Are switches in perception of the Necker cube related to eye position? *European Journal of Neuroscience*, 20(10), 2811–8. Cited on pages 7, 26

Elazary, L. & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3), 1–15. Cited on page 164

Engel, A.K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2(10), 704–16. Cited on page 158

Engel, A.K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, 17(5), 202–209. Cited on page 157

Engel, S. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, 7(2), 181–192. Cited on page 65

Epstein, R. & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601. Cited on page 67

Epstein, R.A., Higgins, J.S., Parker, W., Aguirre, G.K., & Cooperman, S. (2006). Cortical correlates of face and scene inversion: a comparison. *Neuropsychologia*, 44(7), 1145–58. Cited on page 70

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–874. Cited on page 165

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

Fernandez-Duque, D., Grossi, G., Thornton, I., & Neville, H. (2003). Representation of change: Separate electrophysiological markers of attention, awareness, and implicit processing. *Journal of Cognitive Neuroscience*, 15(4), 491–507. Cited on pages 26, 42

Fischl, B., Sereno, M.I., & Dale, A.M. (1999a). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2), 195–207. Cited on pages 61, 143

Fischl, B., Sereno, M.I., Tootell, R.B., & Dale, A.M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4), 272–84. Cited on pages 61, 65

Fisher, G. (1967). Preparation of ambiguous stimulus materials. *Perception and Psychophysics*, 2, 421–422. Cited on page 28

Fisher, G. (1968a). Ambiguity of form: Old and new. *Perception and Psychophysics*, 4, 189–192. Cited on page 28

Fisher, G. (1968b). Mother, father, and daughter: a three-aspect ambiguous figure. *American Journal of Psychology*, 81(2), 274–7. Cited on page 28

Fitzgerald, J.K., Swaminathan, S.K., & Freedman, D.J. (2012). Visual categorization and the parietal cortex. *Frontiers in Integrative Neuroscience*, 6, 18. Cited on page 157

Folstein, J.R., Palmeri, T.J., & Gauthier, I. (2012a). Category Learning Increases Discriminability of Relevant Object Dimensions in Visual Cortex. *Cerebral Cortex*. Cited on pages 18, 136

Folstein, J., Gauthier, I., & Palmeri, T. (2012b). How category learning affects object discrimination: Not all morphspaces stretch alike. *Journal of Experimental Psychology-Learning Memory and Cognition*, 38.4(807). Cited on pages 136, 138

Franzius, M., Wilbert, N., & Wiskott, L. (2008). Invariant object recognition with slow feature analysis. In *Artificial Neural Networks-ICANN 2008*, pages 961–970. Springer, Berlin Heidelberg. Cited on pages 10, 155

Freedman, D.J., Riesenhuber, M., Poggio, T., & Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502), 312–6. Cited on page 18

Freedman, D.J., Riesenhuber, M., Poggio, T., & Miller, E.K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience*, 23(12), 5235–46. Cited on pages 122, 135

Freedman, D.J., Riesenhuber, M., Poggio, T., & Miller, E.K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex*, 16(11), 1631–44. Cited on pages 18, 136

Freeman, J. & Simoncelli, E.P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, 14(9), 1195–201. Cited on page 3

Freiwald, W.A. & Tsao, D.Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005), 845–51. Cited on pages 10, 14, 55, 56, 57, 66, 77, 82, 90, 99

Freiwald, W.A., Tsao, D.Y., & Livingstone, M.S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, 12(9), 1187–96. Cited on page 75

Frey, H., Honey, C., & König, P. (2008). What's color got to do with it? The influence of color on visual attention in different categories. *Journal of Vision*, 8(14), 1–17. Cited on page 185

Fukushima, K. (1980). Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 46, 193–202. Cited on page 9

Gauthier, I., Skudlarski, P., Gore, J.C., & Anderson, a.W. (2000a). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2), 191–7. Cited on pages 18, 22, 120, 154

Gauthier, I., Tarr, M.J., Moylan, J., Skudlarski, P., Gore, J.C., & Anderson, A.W. (2000b). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience*, 12(3), 495–504. Cited on page 67

Gauthier, I. & Palmeri, T.J. (2002). Visual Neurons : Categorization-Based Selectivity. *Current Biology*, 12, 1–3. Cited on pages 121, 148

George, N., Driver, J., & Dolan, R.J. (2001). Seen gaze-direction modulates fusiform activity and its coupling with other brain areas during face processing. *NeuroImage*, 13(6 Pt 1), 1102–12. Cited on page 92

Georgiades, M. & Harris, J. (1997). Biasing effects in ambiguous figures: Removal or fixation of critical features can affect perception. *Visual Cognition*, 4(4), 383–408. Cited on pages 26, 41

Gilland, J. (2008). *Driving, eye-tracking and visual entropy: Exploration of age and task effects*. Ph.D. thesis, The University of South Dakota. Cited on page 194

Gillebert, C.R., Op De Beeck, H.P., Panis, S., & Wagemans, J. (2009). Subordinate categorization enhances the neural selectivity in human object-selective cortex for fine shape differences. *Journal of Cognitive Neuroscience*, 21(6), 1054–64. Cited on pages 121, 136

Golarai, G., Ghahremani, D.G., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J.L., Gabrieli, J.D.E., & Grill-Spector, K. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience*, 10(4), 512–22. Cited on page 22

Goldstone, R.L., Lippa, Y., & Shiffrin, R.M. (2001). Altering object representations through category learning. *Cognition*, 78(1), 27–43. Cited on page 121

Goldstone, R. (1994). Influences of Categorization on Perceptual Discrimination. *Journal of Experimental Psychology: General*, 123(2), 178–200. Cited on page 121

Goodale, M.A. & Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1), 20–5. Cited on page 75

Grice, S., Halit, H., Farroni, T., & Baron-Cohen, S. (2005). Neural Correlates of Eye-Gaze Direction in Young Children. *Cortex*, 141, 342–353. Cited on page 91

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology*, 13(2), 159–166. Cited on page 67

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, 24(1), 187–203. Cited on page 75

Grill-Spector, K. & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta psychologica*, 107(1-3), 293–321. Cited on pages 126, 139

Hafed, Z. & Krauzlis, R. (2006). Ongoing eye movements constrain visual perception. *Nature Neuroscience*, 9, 1449–1457. Cited on page 44

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 19, 545–552. Cited on pages 162, 164, 186

Harris, A. & Nakayama, K. (2007). Rapid Face-Selective Adaptation of an Early Extrastriate Component in MEG. *Cerebral Cortex*, 17(1), 63–70. Cited on page 139

Harrison, S. & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238), 632–5. Cited on page 67

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. *Neuron*, 34(3), 479–90. Cited on page 22

Hausser, J. & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10, 1469–1484. Cited on pages 193, 194, 196

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–30. Cited on page 22

Haxby, J.V., Ungerleider, L.G., Clark, V.P., Schouten, J.L., Hoffman, E.A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, 22(1), 189–99. Cited on page 70

Haxby, J., Hoffman, E., & Gobbini, M. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6), 223–233. Cited on pages 81, 92

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, 3, 49–63. Cited on page 4

Haynes, J.D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523–34. Cited on pages 15, 57

Hershler, O. & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*, 45(13), 1707–1724. Cited on pages 103, 120

Hershler, O. & Hochstein, S. (2009). The importance of being expert: Top-down attentional control in visual search with photographs. *Attention, Perception, & Psychophysics*, 71(7), 1478. Cited on page 120

Hershler, O. & Hochstein, S. (2006). With a careful look: still no low-level confound to face pop-out. *Vision Research*, 46(18), 3028–35. Cited on pages 103, 119

Hills, P.J. & Lewis, M.B. (2011). The own-age face recognition bias in children and adults. *Quarterly journal of experimental psychology (2006)*, 64(1), 17–23. Cited on pages 22, 154

Hinds, O.P., Rajendran, N., Polimeni, J.R., Augustinack, J.C., Wiggins, G., Wald, L.L., Diana Rosas, H., Potthast, A., Schwartz, E.L., & Fischl, B. (2008). Accurate prediction of V1 location from cortical folds in a surface coordinate system. *NeuroImage*, 39(4), 1585–99. Cited on page 65

Hipp, J.F., Engel, A.K., & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron*, 69(2), 387–96. Cited on pages 92, 158

Hitzman, D. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review*, 93(4), 411–428. Cited on page 17

Hochstein, S. & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 791–804. Cited on pages 102, 103, 105, 122

Hoffman, E.A. & Haxby, J.V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, 3(1), 80–4. Cited on page 91

Hoffman, J.E. (1998). Visual Attention and Eye Movements. In H. Pashler, editor, *Attention*, chapter 3, pages 119–154. Psychology Press, Hove UK. Cited on page 4

Holm, L., Eriksson, J., & Andersson, L. (2008). Looking as if you know: Systematic object inspection precedes object recognition. *Journal of Vision*, 8(4), 1–7. Cited on page 26

Holste, D., Grosse, I., & Herzel, H. (1998). Bayes' estimators of generalized entropies. *Journal of Physics A: Mathematical and General*, 31, 2551. Cited on pages 194, 196

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *The Journal of Physiology*, 160, 106–154. Cited on page 9

Huettel, S., Güzeldere, G., & McCarthy, G. (2001). Dissociating the neural mechanisms of visual attention in change detection using functional mri. *Journal of Cognitive Neuroscience*, 13(7), 1006–1018. Cited on page 43

Hung, C.P., Kreiman, G., Poggio, T., & DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749), 863–6. Cited on pages 12, 19, 125, 134

Hupé, J., Lamirel, C., & Lorenceau, J. (2009). Pupil dynamics during bistable motion perception. *Journal of Vision*, 9(7), 1–19. Cited on pages 27, 30, 46, 47

Huth, A.G., Nishimoto, S., Vu, A.T., & Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210–24. Cited on page 22

Hwang, A., Higgins, E., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, 9, 1–18. Cited on pages 161, 162, 164, 166, 185, 186

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., & Haxby, J.V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16), 9379–84. Cited on page 69

Itier, R.J. & Batty, M. (2009). Neural bases of eye and gaze processing: the core of social cognition. *Neuroscience and Biobehavioral Reviews*, 33(6), 843–63. Cited on page 91

Ito, J., Nikolaev, A.R., Luman, M., Aukes, M.F., Nakatani, C., & Leeuwen, C.V. (2003). Perceptual switching, eye movements, and the bus paradox. *Perception*, 32(6), 681–698. Cited on page 7

Itti, L. & Baldi, P. (2005a). A principled approach to detecting surprising events in video. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 631–637. IEEE. Cited on page 166

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12), 1489–1506. Cited on pages 26, 44

Itti, L. & Koch, C. (2001a). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3), 194–203. Cited on page 5

Itti, L. & Koch, C. (2001b). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3), 194–203. Cited on pages 160, 161, 164

Itti, L. & Baldi, P. (2005b). Bayesian surprise attracts human attention. *Vision Research*, 49(10), 1295–306. Cited on pages 160, 161, 164, 166

Jehee, J.F.M., Brady, D.K., & Tong, F. (2011). Attention improves encoding of task-relevant features in the human visual cortex. *The Journal of Neuroscience*, 31(22), 8210–9. Cited on page 6

Jenkins, R., Beaver, J.D., & Calder, A.J. (2006). I thought you were looking at me: direction-specific aftereffects in gaze perception. *Psychological Science*, 17(6), 506–13. Cited on page 90

Jiang, X., Bradley, E., Rini, R.A., Zeffiro, T., Vanmeter, J., & Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron*, 53(6), 891–903. Cited on pages 18, 105, 136

Joachims, T. (1999). Making large scale SVM learning practical. In A. Schölkopf, B., Burges, C., Smola, editor, *Advances in kernel methods: support vector learning*, pages 169 – 184. The MIT Press, Cambridge, MA. Cited on page 32

Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5), 679–685. Cited on page 123

Kanan, C., Tong, M., Zhang, L., & Cottrell, G. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition*, 17(6), 979–1003. Cited on pages 160, 161, 162, 164, 185

Kanwisher, N., McDermott, J., & Chun, M.M. (1997a). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302–11. Cited on pages 21, 66

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, 68(1), B1–11. Cited on page 75

Kanwisher, N., Woods, R.P., Iacoboni, M., & Mazziotta, J.C. (1997b). A Locus in Human Extrastriate Cortex for Visual Shape Analysis. *Journal of Cognitive Neuroscience*, 9(1), 133–142. Cited on page 67

Kawabata, N. & Mori, T. (1992). Disambiguating ambiguous figures by a model of selective attention. *Biological Cybernetics*, 67(5), 417–425. Cited on pages 26, 40, 41, 43, 44

Kelly, D.J., Quinn, P.C., Slater, A.M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: evidence of perceptual narrowing. *Psychological Science*, 18(12), 1084–9. Cited on page 155

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, 97(6), 4296–309. Cited on page 122

Kienzle, W., Franz, M., Schölkopf, B., & Wichmann, F. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 9(5), 1–15. Cited on pages 160, 161, 164

Kietzmann, T.C. & König, P. (2010). Perceptual learning of parametric face categories leads to the integration of high-level class-based information but not to high-level pop-out. *Journal of Vision*, 10(13), 1–14. Cited on pages 138, 139

Kietzmann, T.C., Lange, S., & Riedmiller, M. (2008). Incremental GRLVQ: Learning relevant features for 3D object recognition. *Neurocomputing*, 71(13-15), 2868–2879. Cited on pages 12, 13

Kietzmann, T.C., Swisher, J.D., König, P., & Tong, F. (2012). Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways. *Journal of Neuroscience*, 32(34), 11763–11772. Cited on pages 82, 85, 90, 96

Kietzmann, T.C., Lange, S., & Riedmiller, M. (2009). Computational object recognition: a biologically motivated approach. *Biological Cybernetics*, 100(1), 59–79. Cited on pages 2, 13, 17, 78

Kirchner, H. & Thorpe, S.J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Research*, 46(11), 1762–76. Cited on pages 19, 125, 135

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, 36, 14. Cited on page 138

Kloth, N. & Schweinberger, S. (2010). Electrophysiological correlates of eye gaze adaptation. *Journal of Vision*, 10(12), 1–13. Cited on page 90

Koch, C. & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in Cognitive Sciences*, 11(1), 16–22. Cited on page 42

Koch, C. & Ullman, S. (1985a). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4, 219–227. Cited on page 5

Koch, C. & Ullman, S. (1985b). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4), 219–227. Cited on page 44

Koivisto, M., Kainulainen, P., & Revonsuo, A. (2009). The relationship between awareness and attention: Evidence from ERP responses. *Neuropsychologia*, 47(13), 2891–2899. Cited on page 43

Koivisto, M. & Revonsuo, A. (2007). Electrophysiological correlates of visual consciousness and selective attention. *NeuroReport*, 18(8), 753. Cited on page 43

Konen, C.S. & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, 11(2), 224–31. Cited on page 75

König, P., Kühnberger, K.U., & Kietzmann, T.C. (2013). A unifying approach to high- and low-level cognition. In U. Gähde, S. Hartmann, & J.H. Wolf, editors, *Models, Simulations, and the Reduction of Complexity*, Kaas, pages 117–140. De Gruyter, Berlin. Cited on page 2

König, P. & Krüger, N. (2006). Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics*, 94(4), 325–34. Cited on page 158

Kootstra, G., de Boer, B., & Schomaker, L.R.B. (2011). Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, 3(1), 223–240. Cited on pages 164, 166

Körding, K.P.K., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 91(1), 206–212. Cited on page 155

Krichevsky, R. & Trofimov, V. (2002). The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2), 199–207. Cited on pages 194, 195, 196

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, 13(2-3), 201–14. Cited on page 5

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51), 20600–5. Cited on page 78

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10), 3863–8. Cited on pages 15, 57, 71

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November), 4. Cited on pages 15, 61, 62, 82, 94

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P.A. (2008b). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60, 1–16. Cited on page 61

Króliczak, G., McAdam, T.D., Quinlan, D.J., & Culham, J.C. (2008). The human dorsal stream adapts to real actions and 3D shape processing: a functional magnetic resonance imaging study. *Journal of Neurophysiology*, 100(5), 2627–39. Cited on page 75

Kuehn, S. & Jolicoeur, P. (1994). Impact of quality of the image, orientation, and similarity of the stimuli on visual search for faces. *Perception*, 23(1), 95–122. Cited on page 119

Lamme, V. (2003). Why visual attention and awareness are different. *Trends in cognitive sciences*, 7(1), 12–18. Cited on page 42

Leeper, R. (1935). A study of a neglected portion of the field of learningâĂŤthe development of sensory organization. *The Pedagogical Seminary and Journal of Genetic Psychology*, 46(1), 41–75. Cited on page 7

Leuba, G. & Kraftsik, R. (1994). Anatomy and Embryolo and total number of neurons of the human primary visual cortex. *Anatomy and Embryology*, 190(4), 351–366. Cited on page 1

Li, S., Mayhew, S.D., & Kourtzi, Z. (2009). Learning shapes the representation of behavioral choice in the human brain. *Neuron*, 62(3), 441–52. Cited on page 18

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible coding for categorical decisions in the human brain. *The Journal of Neuroscience*, 27(45), 12321–30. Cited on page 136

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1), 9–16. Cited on page 44

Liu, H., Agam, Y., Madsen, J.R., & Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2), 281–90. Cited on pages 19, 125, 135

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nature Neuroscience*, 5(9), 910–6. Cited on pages 19, 125, 135

Logothetis, N.K., Pauls, J., Bülthoff, H.H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, 4(5), 401–14. Cited on pages 9, 12, 56

Logothetis, N.K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5), 552–63. Cited on pages 10, 13, 56

Long, G. & Toppino, T. (2004). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin*, 130(5), 748–768. Cited on page 41

Mahon, B.Z. & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, 15(3), 97–103. Cited on page 158

Mahon, B.Z., Milleville, S.C., Negri, G.a.L., Rumiati, R.I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. *Neuron*, 55(3), 507–20. Cited on page 157

Makin, A., Wilton, M., Pecchinenda, A., & Bertamini, M. (2012). Symmetry perception and affective responses : A combined EEG / EMG study. *Neuropsychologia*, 50, 3250–3261. Cited on page 91

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., & Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 92(18), 8135–9. Cited on page 67

Mannan, S.K., Ruddock, K.H., & Wooding, D.S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10, 165–188. Cited on page 4

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of neuroscience methods*, 164(1), 177–90. Cited on pages 85, 96, 133, 141, 143

Marr, D. & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B*, 200(1140), 269–94. Cited on page 9

Martinez-Conde, S., Macknik, S.L., Troncoso, X.G., & Dyar, T.a. (2006). Microsaccades counteract visual fading during fixation. *Neuron*, 49(2), 297–305. Cited on page 43

Mcgugin, R.W. & Gauthier, I. (2013). Face Recognition. In S. Kosslyn & K. Ochsner, editors, *The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges*, pages 165–181. Oxford University Press USA. Cited on page 22

McKeeff, T.J., McGugin, R.W., Tong, F., & Gauthier, I. (2010). Expertise increases the functional overlap between face and object perception. *Cognition*, 117(3), 355–60. Cited on page 22

McKeeff, T.J. & Tong, F. (2007). The timing of perceptual decisions for ambiguous face stimuli in the human ventral visual cortex. *Cerebral Cortex*, 17(3), 669–78. Cited on page 18

McKone, E., Kanwisher, N., & Duchaine, B.C. (2007). Can generic expertise explain special processing for faces? *Trends in cognitive sciences*, 11(1), 8–15. Cited on page 22

Miller, G. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods II-B*, II, 95–100. Cited on pages 193, 194, 196

Minamimoto, T., Saunders, R.C., & Richmond, B.J. (2010). Monkeys Quickly Learn and Generalize Visual Categories without Lateral Prefrontal Cortex. *Neuron*, 66(4), 501–507. Cited on pages 122, 136

Mishkin, M. & Ungerleider, L. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioral Brain Research*, 6(1), 57–77. Cited on page 75

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What Is Special about Face Recognition? Nineteen Experiments on a Person with Visual Object Agnosia and Dyslexia but Normal Face Recognition. *Journal of Cognitive Neuroscience*, 9(5), 555–604. Cited on page 22

Natu, V.S., Jiang, F., Narvekar, A., Keshvari, S., Blanz, V., & O'Toole, A.J. (2010). Dissociable neural patterns of facial identity across changes in viewpoint. *Journal of Cognitive Neuroscience*, 22(7), 1570–82. Cited on page 78

Necker, L. (1832). Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *The London and Edinburgh Philosophical Magazine and Journal of Science*, 1(5), 329–337. Cited on page 7

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. *Advances in Neural Information Processing Systems*, 1(14), 471–478. Cited on pages 193, 194, 196

Nestor, A., Plaut, D.C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24), 9998–10003. Cited on page 78

Norman, K.A., Polyn, S.M., Detre, G.J., & Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424–430. Cited on pages 15, 57

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, 115(1), 39–61. Cited on page 17

Nosofsky, R.M. & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104(2), 266–300. Cited on page 17

Nothdurft, H. (1993). Faces and facial expressions do not pop out. *Perception*, 22(11), 1287–1287. Cited on page 119

Nummenmaa, L. & Calder, A.J. (2009). Neural mechanisms of social attention. *Trends in cognitive sciences*, 13(3), 135–43. Cited on page 91

Nuthmann, A. & Henderson, J.M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8), 1–19. Cited on page 5

Olshausen, B.A. & Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13), 607–609. Cited on page 155

Oostenveld, R. & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical neurophysiology*, 112(4), 713–9. Cited on page 93

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 1–10. Cited on page 140

Op de Beeck, H.P., Haushofer, J., & Kanwisher, N.G. (2008). Interpreting fMRI data: maps, modules and dimensions. *Nature Reviews Neuroscience*, 9(2), 123–35. Cited on page 22

O'Regan, J.K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *The Behavioral and Brain Sciences*, 24(5), 939–73; discussion 973–1031. Cited on page 157

Otero-Millan, J., Troncoso, X.G., Macknik, S.L., Serrano-pedraza, I., & Martinez-conde, S. (2008). Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *Journal of Vision*, 8, 1–18. Cited on page 43

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J.B. Long & A.D. Baddeley, editors, *Attention and Performance IX*, pages 135–151. L. Erlbaum Associates. Cited on page 12

Palmeri, T.J. & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, 5(4), 291–303. Cited on page 17

Parikh, N., Itti, L., & Weiland, J. (2010). Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering*, 7(1), 16006. Cited on pages 164, 165

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1), 107–123. Cited on pages 160, 161, 164, 165

Parvizi, J., Jacques, C., Foster, B.L., Witthoft, N., Withoft, N., Rangarajan, V., Weiner, K.S., & Grill-Spector, K. (2012). Electrical stimulation of human fusiform face-selective regions distorts face perception. *The Journal of Neuroscience*, 32(43), 14915–20. Cited on page 22

Pascual-Leone, A. & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, 292(5516), 510–2. Cited on page 103

Pascual-Marqui, R. (2002). Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. *Methods & Findings in Experimental & Clinical Pharmacology*, 24(D), 5–12. Cited on pages 133, 143

Perrett, D.I., Hietanen, J.K., Oram, M.W., & Benson, P.J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 335(1273), 23–30. Cited on page 81

Perrett, D.I., Oram, M.W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition*, 67(1-2), 111–45. Cited on pages 10, 11, 12, 56

Perrett, D.I., Oram, M.W., Harries, M.H., Bevan, R., Hietanen, J.K., Benson, P.J., & Thomas, S. (1991). Viewer-Centered and Object-Centered Coding of Heads in the Macaque Temporal Cortex. *Experimental Brain Research*, pages 159–173. Cited on pages 10, 56

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18), 2397–2416. Cited on pages 160, 161, 164, 165

Pheiffer, C., Eure, S., & Hamilton, H. (1956). Reversible figures and eye-movements. *American Journal of Psychology*, 62, 452–455. Cited on page 26

Philiastides, M.G. & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, 16(4), 509–18. Cited on page 137

Pinsk, M.A., Arcaro, M., Weiner, K.S., Kalkus, J.F., Inati, S.J., Gross, C.G., & Kastner, S. (2009). Neural representations of faces and body parts in macaque and human cortex: a comparative FMRI study. *Journal of Neurophysiology*, 101(5), 2581–600. Cited on page 66

Pitcher, D., Duchaine, B., Walsh, V., Yovel, G., & Kanwisher, N. (2011). The role of lateral occipital face and object areas in the face inversion effect. *Neuropsychologia*, pages 6–11. Cited on page 70

Plöchl, M., Ossandón, J.P., & König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, 6, 278. Cited on page 145

Poggio, T. & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343(6255), 263–266. Cited on pages 9, 13, 78

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, 25(8), 931. Cited on pages 7, 40, 43, 44

Posner, M. & Keele, S. (1968). On the Genesis of Abstract Ideas. *Journal of experimental psychology*, 77(3). Cited on page 17

Pourtois, G., Schwartz, S., Seghier, M.L., Lazeyras, F., & Vuilleumier, P. (2005). Portraits or people? Distinct representations of face identity in the human visual cortex. *Journal of Cognitive Neuroscience*, 17(7), 1043–57. Cited on page 75

Pourtois, G. & Spinelli, L. (2010). Modulation of face processing by emotional expression and gaze direction during intracranial recordings in right fusiform cortex. *Journal of Cognitive Neuroscience*, 22(9), 2086–2107. Cited on page 91

Puce, A., Allison, T., Asgari, M., Gore, J.C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of Neuroscience*, 16(16), 5205–15. Cited on page 67

Purcell, D., Stewart, A., & Skov, R. (1996). It takes a confounded face to pop out of a crowd. *Perception*, 25(9), 1091–1120. Cited on page 103

Quiroga, R.Q. & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review*, 117(1), 291–297. Cited on page 12

Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045), 1102–7. Cited on pages 10, 11

Ravikumar, P., Vu, V., Yu, B., Naselaris, T., Kay, K., & Gallant, J. (2008). Nonparametric sparse hierarchical models describe v1 fmri responses to natural images. *Advances in Neural Information Processing Systems (NIPS)*, 21, 1–8. Cited on page 155

Recarte, M. & Nunes, L. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology Applied*, 6(1), 31–43. Cited on page 194

Reinagel, P. & Zador, A.M. (1999). Natural scene statistics at the centre of gaze. *Network (Bristol, England)*, 10(4), 341–50. Cited on page 4

Renninger, L.W., Verghese, P., & Coughlan, J. (2007). Where to look next? eye movements reduce local uncertainty. *Journal of Vision*, 7(3), 6. Cited on page 164

Renninger, L., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in neural information processing systems*, 17, 1121–1128. Cited on page 44

Richer, F., Silverman, C., & Beatty, J. (1983). Response selection and initiation in speeded reactions: A pupillometric analysis. *Journal of Experimental Psychology*, 9(3), 360–370. Cited on page 46

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1019–25. Cited on pages 10, 13, 18

Riesenhuber, M. & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, 3 Suppl, 1199–204. Cited on page 10

Riesenhuber, M. & Poggio, T. (2002a). How Visual Cortex Recognizes Objects: The Tale of the Standard Model. In J.S.W.E. L. M. Chapula, editor, *The visual neurosciences*, pages 1640–1653. Cambridge, MA: MIT Press. Cited on page 135

Riesenhuber, M. & Poggio, T. (2002b). Neural mechanisms of object recognition. *Current opinion in neurobiology*, 12(2), 162–8. Cited on page 11

Rollenhagen, J.E. & Olson, C. (2000). Mirror-Image Confusion in Single Neurons of the Macaque Inferotemporal Cortex. *Science (New York, N.Y.)*, 287, 1506–1508. Cited on pages 13, 76

Rolls, E.T. (2012). Invariant Visual Object and Face Recognition: Neural and Computational Bases, and a Model, VisNet. *Frontiers in Computational Neuroscience*, 6(June), 35. Cited on pages 10, 155

Rosch, E. (1973). Natural categories. *Cognitive psychology*, 4(3), 328–350. Cited on page 17

Rosseel, Y. (2002). Mixture Models of Categorization. *Journal of Mathematical Psychology*, 46(2), 178–210. Cited on page 17

Rossion, B. & Caharel, S. (2011). ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Research*, 51(12), 1297–311. Cited on page 135

Rossion, B. & Jacques, C. (2008). Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170. *NeuroImage*, 39(4), 1959–79. Cited on page 126

Rothkopf, C., Ballard, D., & Hayhoe, M. (2007). Task and context determine where you look. *Journal of Vision*, 7(14), 12. Cited on page 44

Roy, J.E., Riesenhuber, M., Poggio, T., & Miller, E.K. (2010). Prefrontal cortex activity during flexible categorization. *The Journal of Neuroscience*, 30(25), 8519–28. Cited on page 135

Sasaki, Y., Vanduffel, W., Knutsen, T., Tyler, C., & Tootell, R. (2005). Symmetry activates extrastriate visual cortex in human and nonhuman primates. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8), 3159–63. Cited on pages 76, 77

Scherf, K.S., Behrmann, M., Humphreys, K., & Luna, B. (2007). Visual category-selectivity for faces, places and objects emerges along different developmental trajectories. *Developmental science*, 10(4), F15–30. Cited on page 22

Scholl, C., Jiang, X., Martin, J., & Riesenhuber, M. (2014). Time Course of Shape and Category Selectivity Revealed by EEG Rapid Adaptation. *Journal of Cognitive Neuroscience*, 26(2), 408–421. Cited on pages 136, 139

Schumann, F., Einhäuser-Treyer, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision*, 8(14), 1–17. Cited on page 4

Schürmann, T. & Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos*, 6(3), 414–427. Cited on pages 194, 196

Schweinberger, S.R., Kloth, N., & Jenkins, R. (2007). Are you looking at me? Neural correlates of gaze adaptation. *Neuroreport*, 18(7), 693–6. Cited on page 90

Seger, C.A. & Miller, E.K. (2010). Category learning in the brain. *Annual review of neuroscience*, 33, 203–19. Cited on page 136

Senju, A. & Johnson, M.H. (2009). The eye contact effect: mechanisms and development. *Trends in cognitive sciences*, 13(3), 127–34. Cited on pages 82, 91

Sereno, A. & Maunsell, J. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature*, 395(6701), 500–3. Cited on page 75

Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., Rosen, B.R., & Tootell, R.B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889–93. Cited on page 65

Sergent, J. & Signoret, J.L. (1992). Varieties of functional deficits in prosopagnosia. *Cerebral Cortex*, 2(5), 375–88. Cited on page 22

Serre, T., Oliva, A., & Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15), 6424–9. Cited on pages 10, 11, 18, 102, 135

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3), 411–26. Cited on page 10

Serre, T. & Riesenhuber, M. (2004). Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. *Technical Report CBCL Paper 239 / AI Memo 2004-017*, (July). Cited on pages 59, 83, 96

Sigala, N., Gabbiani, F., & Logothetis, N.K. (2002). Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience*, 14(2), 187–98. Cited on pages 106, 112, 120, 121, 138

Sigala, N. (2004). Visual categorization and the inferior temporal cortex. *Behavioural Brain Research*, 149, 1–7. Cited on page 121

Sigala, N. & Logothetis, N.K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415(6869), 318–20. Cited on pages 18, 104, 106, 107, 108, 120, 122, 136, 138

Silvanto, J., Cowey, A., Lavie, N., & Walsh, V. (2005a). Striate cortex (V1) activity gates awareness of motion. *Nature neuroscience*, 8(2), 143–4. Cited on page 123

Silvanto, J., Lavie, N., & Walsh, V. (2005b). Double Dissociation of V1 and V5/MT activity in Visual Awareness. *Cerebral Cortex*, 15(11), 1736–1741. Cited on page 103

Simons, D. & Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1), 16–20. Cited on page 42

Smith, C., Hopkins, R., & Squire, L. (2006). Experience-dependent eye movements, awareness, and hippocampus-dependent memory. *The Journal of Neuroscience*, 26(44), 11304. Cited on page 42

Smith, C. & Squire, L. (2008). Experience-dependent eye movements reflect hippocampus-dependent (aware) memory. *The Journal of Neuroscience*, 28(48), 12825. Cited on page 42

Smith, F.W. & Goodale, M.A. (2013). Decoding Visual Object Categories in Early Somatosensory Cortex. *Cerebral Cortex*, pages 1–12. Cited on page 157

Stansbury, D.E., Naselaris, T., & Gallant, J.L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5), 1025–34. Cited on page 154

Stich, S. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation.* MIT Press, Cambridge, MA. Cited on page 157

Stolk, A., Todorovic, A., Schoffelen, J.M., & Oostenveld, R. (2013). Online and offline tools for head movement compensation in MEG. *NeuroImage*, 68, 39–48. Cited on page 140

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400(6747), 869–73. Cited on pages 125, 134

Summerfield, C., Behrens, T.E., & Koechlin, E. (2011). Perceptual classification in a rapidly changing environment. *Neuron*, 71(4), 725–36. Cited on page 126

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., & Leahy, R.M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational intelligence and neuroscience*, 2011, 879716. Cited on pages 140, 143

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19, 109–39. Cited on page 102

Tanaka, Y., Miyauchi, S., Misaki, M., & Tashiro, T. (2007). Mirror symmetrical transfer of perceptual learning by prism adaptation. *Vision Research*, 47(10), 1350–1361. Cited on page 77

Tarr, M.J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic bulletin & review*, 2(1), 55–82. Cited on page 12

Tarr, M.J. & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of experimental psychology. Human perception and performance*, 21, 1494–1505. Cited on pages 10, 13

Tarr, M. & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, 3(8), 764–769. Cited on page 22

Tarr, M., Williams, P., & Hayward, W. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, 1(4), 275–277. Cited on pages 10, 13, 78

Tatler, B.W., Baddeley, R.J., & Gilchrist, I.D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5), 643–59. Cited on pages 161, 164, 165, 184

Tatler, B.W., Baddeley, R.J., & Vincent, B.T. (2006). The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, 46(12), 1857–62. Cited on page 4

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7, 1–17. Cited on page 161

Tatler, B. & Vincent, B. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6), 1029–1054. Cited on pages 161, 164, 166

Taylor, M.J., Itier, R.J., Allison, T., & Edmonds, G.E. (2001). Direction of gaze effects on early face processing: eyes-only versus full faces. *Cognitive Brain Research*, 10(3), 333–40. Cited on page 91

Thierry, G., Martin, C.D., Downing, P., & Pegna, A.J. (2007). Controlling for interstimulus perceptual variance abolishes N170 face selectivity. *Nature Neuroscience*, 10(4), 505–11. Cited on pages 126, 135

Thomas, E., Van Hulle, M.M., Vogels, R., Hulle, M.V., & Vogel, R. (2001). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience*, 13(2), 190–200. Cited on page 18

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of Processing in the Human Visual System. *Nature*, 381, 520–522. Cited on page 19

Tjan, B.S. & Legge, G.E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research*, 38(15-16), 2335–50. Cited on pages 12, 17

Tong, F. (2003). Primary visual cortex and visual awareness. *Nature Reviews Neuroscience*, 4(3), 219–29. Cited on pages 103, 123

Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in cognitive sciences*, 10(11), 502–11. Cited on page 156

Tong, F. & Pratte, M.S. (2010). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, 63(1), 110301102248092. Cited on pages 6, 15, 57

Toppino, T. (2003). Reversible-figure perception: Mechanisms of intentional control. *Perception and Psychophysics*, 65(8), 1285–1295. Cited on page 43

Toppino, T. & Long, G. (2005). Top-down and bottom-up processes in the perception of reversible figures: Toward a hybrid model. In Nobuo Ohta, C.M. MacLeod, & B. Uttl, editors, *Dynamic cognitive processes*, pages 37–58. Springer, Tokyo. Cited on pages 7, 26

Torralba, A., Oliva, A., Castelhano, M.S., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4), 766–86. Cited on pages 161, 164, 165, 183

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1), 97–136. Cited on page 103

Triesch, J., Ballard, D., Hayhoe, M., & Sullivan, B. (2003). What you see is what you need. *Journal of Vision*, 3(1), 86–94. Cited on page 44

Troncoso, X.G., Macknik, S.L., & Martinez-Conde, S. (2005). Novel visual illusions related to vasarely's 'nested squares' show that corner salience varies with corner angle. *Perception*, 34(4), 409–420. Cited on page 44

Troncoso, X.G., Macknik, S.L., Otero-Millan, J., & Martinez-Conde, S. (2008). Microsaccades drive illusory motion in the enigma illusion. *Proceedings of the National Academy of Sciences of the United States of America*, 105(41), 16033–8. Cited on page 43

Troncoso, X.G., Tse, P.U., Macknik, S.L., Caplovitz, G.P., Hsieh, P.J., Schlegel, A.A., Otero-Millan, J., & Martinez-Conde, S. (2007). Bold activation varies parametrically with corner angle throughout human retinotopic cortex. *Perception*, 36(6), 808–820. Cited on page 44

Trybula, S. (1958). Some problems of simultaneous minimax estimation. *The Annals of Mathematical Statistics*, 29(1), 245–253. Cited on pages 194, 196

Tsal, Y. & Kolbet, L. (1985). Disambiguating ambiguous figures by selective attention. *The Quarterly Journal of Experimental Psychology Section A*, 37(1), 25–37. Cited on pages 28, 41

Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., & Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761), 670–4. Cited on page 22

Tsao, D., Moeller, S., & Freiwald, W. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105(49), 19514. Cited on pages 77, 78

Tse, P.U., Martinez-Conde, S., Schlegel, A.A., & Macknik, S.L. (2005). Visibility, visual awareness, and visual masking of simple unattended targets are confined to areas in the occipital cortex beyond human v1/v2. *Proceedings of the National Academy of Sciences of the United States of America*, 102(47), 17178–83. Cited on page 26

Tyler, C.W., Baseler, H.A., Kontsevich, L.L., Likova, L.T., Wade, A.R., & Wandell, B. (2005). Predominantly extra-retinotopic cortical response to pattern symmetry. *NeuroImage*, 24(2), 306–14. Cited on page 76

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, 67(1-2), 21–44. Cited on pages 17, 56, 78

Ullman, S. & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10), 992–1006. Cited on pages 9, 13

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32, 193–254. Cited on page 12

Underwood, G., Henderson, J.M., & Hollingworth, A. (1998). Eye guidance in reading and scene perception. In G. Underwood, editor, *Eye movements during scene viewing: An overview.*, volume 1, chapter 12, pages 269–293. Elsevier. Cited on page 143

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, 79(4), 471–491. Cited on page 75

van der Linden, M., van Turennout, M., & Indefrey, P. (2010). Formation of category representations in superior temporal sulcus. *Journal of Cognitive Neuroscience*, 22(6), 1270–82. Cited on page 136 van der Linden, M., Wegman, J., & Fernández, G. (2013). Task- and Experience-dependent Cortical Selectivity to Features Informative for Categorization. *Journal of Cognitive Neuroscience*, pages 1–15. Cited on page 136

van Gaal, S. & Fahrenfort, J. (2008). The relationship between visual awareness, attention, and report. *The Journal of Neuroscience*, 28(21), 5401. Cited on page 42

VanRullen, R. & Thorpe, S.J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of Cognitive Neuroscience*, 13(4), 454–61. Cited on pages 126, 135

VanRullen, R. (2006). On second glance: still no high-level pop-out effect for faces. *Vision Research*, 46(18), 3017–27. Cited on pages 21, 103, 108, 110

VanRullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Frontiers in Psychology*, 2, 365. Cited on page 19

Varela, F. & Lachaux, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2, 229–239. Cited on page 158

Varela, F.J., Thompson, E., & Rosch, E. (1993). *The Embodied Mind: Cognitive Science and Human Experience*, volume 1992. MIT Press. Cited on page 157

Vetter, T., Poggio, T., & Bülthoff, H.H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology*, 4(1), 18–23. Cited on pages 14, 78

Vinje, W.E. & Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456), 1273–6. Cited on page 155

Vizioli, L., Rousselet, G.A., & Caldara, R. (2010). Neural repetition suppression to identity is abolished by other-race faces. *Proceedings of the National Academy of Sciences of the United States of America*, 107(46), 20081–6. Cited on page 139

von Stein, A., Chiang, C., & König, P. (2000). Top-down processing mediated by interareal synchronization. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26), 14748–53. Cited on page 158

Wallis, G. & Rolls, E. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51, 167–194. Cited on pages 10, 155, 156

Wallis, G. (2013). Toward a unified model of face and object recognition in the human visual system. *Frontiers in Psychology*, 4, 497. Cited on pages 22, 23

Walsh, V. (1996). Neuropsychology: Reflections on mirror images. *Current Biology*, 6(9), 1079–1081. Cited on page 13

Walther, D., Itti, L., & Riesenhuber, M. (2002). Attentional selection for object recognition - a gentle way. In H.H. Bülthoff, C. Wallraven, S.W. Lee, & T.A. Poggio, editors, *Biologically Motivated Computer Vision*, pages 472–479. Springer, Berlin Heidelberg. Cited on page 6

Walther, D. & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19(9), 1395–407. Cited on page 6

Weiller, D., Märtin, R., Dähne, S., Engel, A.K., & König, P. (2010). Involving motor capabilities in the formation of sensory space representations. *PloS one*, 5(4), e10377. Cited on page 158

Wichmann, F., Drewes, J., Rosas, P., & Gegenfurtner, K. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision*, 10(4), 1–27. Cited on page 135

Willenbockel, V., Sadr, J., Fiset, D., Horne, G.O., Gosselin, F., & Tanaka, J.W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior research methods*, 42(3), 671–84. Cited on page 92

Wilming, N., Betz, T., Kietzmann, T.C., & König, P. (2011). Measures and Limits of Models of Fixation Selection. *PloS one*, 6(9). Cited on page 5

Wiskott, L. & Sejnowski, T.T.J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4), 715–770. Cited on page 155

Wong, Y.K., Twedt, E., Sheinberg, D., & Gauthier, I. (2010). Does Thompson's Thatcher Effect reflect a face-specific mechanism? *Perception*, 39(8), 1125–1141. Cited on page 21

Wyart, V. & Tallon-Baudry, C. (2008). Neural dissociation between visual awareness and spatial attention. *The Journal of Neuroscience*, 28(10), 2667. Cited on pages 26, 42

Wyss, R., König, P., & Verschure, P.F.M.J. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology*, 4(5), 836–843. Cited on pages 10, 155

Xu, J., Jiang, M., Wang, S., Kankanhalli, M.S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, 14(1), 1–20. Cited on page 5

Yanulevskaya, V., Marsman, J.B., Cornelissen, F., & Geusebroek, J.M. (2011). An image statistics-based model for fixation prediction. *Cognitive Computation*, 3(1), 94–104. Cited on page 164

Yarbus, A. (1967). *Eye Movements and Vision*. Plenum Pres, New York, vol. 2 edition. Cited on page 4

Yin, R. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1), 141–145. Cited on page 75

Yokoyama, T., Sakai, H., Noguchi, Y., & Kita, S. (2014). Perception of direct gaze does not require focus of attention. *Scientific reports*, 4, 3858. Cited on page 91

Young, M.P. & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256(5061), 1327–31. Cited on pages 11, 155

Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 1–32. Cited on pages 160, 161, 164

Zhaoping, L. & Guyader, N. (2007). Interference with bottom-up feature detection by higher-level object recognition. *Current Biology*, 17(1), 26–31. Cited on page 26

Zimmer, M. & Kovács, G. (2011). Position specificity of adaptation-related face aftereffects. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366(1564), 586–95. Cited on page 139