Object recognition put in context

Exploring scene segmentation in real-world vision

Seijdel, N.

Publication date
2021

Document Version
Final published version

Link to publication

Citation for published version (APA):
Seijdel, N. (2021). *Object recognition put in context: Exploring scene segmentation in real-world vision*. [Thesis, fully internal, Universiteit van Amsterdam].

Object recognition put in context

Noor Seijdel

OBJECT RECOGNITION PUT IN CONTEXT

Exploring scene segmentation in real-world vision

Noor Seijdel

Object recognition put in context

The investigations in this thesis were supported by a European Research Council Advanced Grant (#339374) awarded to E.H.F. de Haan

This thesis was typeset using the amsterdown[1] template developed by Leon Reteig, using (R) Markdown, LaTeX and the `bookdown` R-package
ISBN: 978-94-6421-289-1
Printed by Ipskamp Printing B.V., Enschede
Cover design by author using VGG19 and neural style transfer in Keras

[1] https://lcreteig.github.io/amsterdown/

Object recognition put in context

exploring scene segmentation in real-world vision

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen
op vrijdag 16 april 2021, te 13:00 uur

door

Noortje Seijdel

geboren te Amsterdam

Promotiecommissie:

Promotores:	dr. H.S. Scholte	Universiteit van Amsterdam
	prof. dr. E.H.F. de Haan	Universiteit van Amsterdam
Overige leden:	dr. I.I.A. Groen	Universiteit van Amsterdam
	dr. R.M. Cichy	Freie Universität Berlin
	prof. dr. C. G. M. Snoek	Universiteit van Amsterdam
	prof. dr. M.A.J. van Gerven	Donders Instituut voor Brein, Cognitie en Gedrag
	prof. dr. V.A.F. Lamme	Universiteit van Amsterdam
	prof. dr. P.R. Roelfsema	Netherlands Institute for Neuroscience

Faculteit: Faculteit der Maatschappij- en Gedragswetenschappen

Contents

Chapter 1

General introduction

If this introduction were written ten years ago, it could begin with a hymn of praise for the human visual system and its incredible achievement in object recognition. It could emphasize how the brain is the only known system that is able to accurately and efficiently perform this task. Back then, visual object recognition was an extremely difficult computational problem; so difficult that even the most advanced computer models and algorithms were not sufficiently advanced to simulate humans' incredible performance. Nowadays however, boosted by larger datasets and growing computational power, advances in artificial neural networks have led to vision systems that are starting to rival human accuracy in basic object recognition tasks. This chapter will still start with a description of the human visual system and its praiseworthy performance. Next, it provides a brief overview of the relevant psychological, neuroscientific and computational work on object recognition. Throughout, it outlines how the work presented in this thesis aims to contribute to our understanding of object recognition in the human brain, and how recent advances in a class of computational models called deep convolutional neural networks might mimic and elucidate the processes underlying human recognition. Finally, a brief overview of the remaining chapters in this thesis is provided.

Seemingly without any effort, our brain makes sense of the light that is projected onto our retina, and in the blink of an eye we recognize the objects surrounding us. Unlike a camera, which simply stores raw visual information projected from the physical world, we understand our environment in terms of a constellation of vivid visual features, structures and objects. This performance is especially impressive given that object recognition is a computationally demanding process. A single object, for example a banana, can generate a virtually infinite number of different retinal projections based on many factors, such as its viewpoint, lighting or its ripeness. Moreover, objects from the same category may vary in color, size, texture and other features. In addition to the objects that we already know, recognition often generalizes easily to new, previously unseen exemplars. To make it even more complicated, objects in the real world rarely appear in isolation. We see the world in scenes, where objects are embedded and often partially occluded in rich and complex surroundings. How does the brain extract and transform diagnostic low-level visual features into robust visual representations, whilst there are so many factors that affect the appearance of natural object categories? And how are these representations mapped onto high-level structure, meaning and memory? Visual perception is the organization, identification and interpretation of visual information in order to represent and understand the outside world. The visual cortex is one of the best characterized areas of the human brain and decades of research have focused on these questions.

From raw sensory input to complex feature constellations

It all starts when light falls onto the lens of our eyes, and is projected onto our retina. Light is composed of photons, that make up electromagnetic waves. The retina is tiled with four different types of light sensitive photoreceptors that each respond differentially to electromagnetic frequencies. Activation is then fed forward through bipolar cells and ganglion cells, via the optic nerve, of which half of the axons crosses over in the optic chiasm and a synaps in the lateral geniculate nucleus (LGN) to the cortex, with most of the fibers terminating in the primary visual cortex (V1). Apart from V1, many other cortical visual areas have been identified. Early studies with patients with damage to certain areas (lesions) have provided strong evidence that damage to different cortical regions may lead to selective impairments (Bodamer, 1947; Dejerine, 1892; Jackson, 1876; Kleist, 1934; Lewandowsky, 1908; Newcombe, 1969; Wilbrand, 1892; Zeki, 1993), including, for example, achromatopsia (impairment of color vision) and prosopagnosia (impairments in the recognition of faces). Findings from electrophysiological and neuroimaging studies have subsequently identified areas that are specialized to represent different types of information. Some areas preferably respond to simple stimuli such as oriented contrasts (Hubel & Wiesel, 1968) or textures; other areas only become active when a specific (more complex) object is visible, such as a face or a bicycle (Downing et al., 2006; Epstein & Kanwisher, 1998). In the last few decades, researchers have identified more than 40 separate maps in the brain that are selectively tuned to specific visual features, such as color (Zeki et al., 1998) or motion (Zihl et al., 1983). Why has the brain evolved with so many visual areas? The classic hypothesis is that the areas form a hierarchy or a 'pathway', in which each area incrementally expands on the representation derived by computations and processing in earlier areas, each time representing the sensory input in a different way. The ventral pathway (also referred to as the "what" pathway) has been shown to play a key role in the computations underlying the identification and recognition of objects (Goodale et al., 1992). This pathway extends from V1 through a series of stages, V2, V3, V4, to the inferior temporal (IT) region. Going forward in this ventral pathway, cell responses gradually become size and position invariant, as well as selective for increasingly complex features. An alternative hypothesis is that for each particular visual feature that we can perceive (e.g. colors or textures), there is a dedicated system in the brain containing several visual maps. These different visual maps designed for various visual tasks such as color or face perception, are then distributed over the posterior brain, involving both more basic processes and higher-order processing (De Haan & Cowey, 2011).

Feed-forward vs. feedback processing in object recognition

Following a pathway from 'simple' to 'more complex' feature constellations, object recognition has traditionally been seen as the result of a processing pipeline in which, after detection of low-level features (e.g. edges or orientations), grouping and segmentation of the relevant features form the basis on which higher level operations like object recognition can operate (Riesenhuber & Poggio, 1999; Serre et al., 2005). In this pipeline, grouping of information may occur based on relatively simple cues like motion, orientation or collinearity (Mack et al., 1990). By sequentially building up increasingly complex feature conjunctions, a single feed-forward sweep through this pipeline may suffice to perform 'core object recognition'[1] (DiCarlo et al., 2012; Serre et al., 2007) (Figure 1.1A).

Indeed, the speed and efficiency of behavior suggest that a fast feed-forward buildup of perceptual activity should be sufficient for object recognition. For example, rapid categorization experiments showed that humans produced extremely fast and reliable behavioral responses about the category of objects in natural scenes (Thorpe et al., 1996). If participants only needed to make an eye movement towards the relevant category, this could even be done within 120 ms (Kirchner & Thorpe, 2006). Looking at the brain, electroencephalography (EEG) measures of visual responses to different object categories already started to diverge at 150 ms after stimulus onset (Thorpe et al., 1996; VanRullen & Thorpe, 2001). In this feedforward account of object recognition, the role of the early retinotopic visual cortex (i.e., V1/V2) is limited to performing basic computations of the visual input and feeding the output to higher areas for more complex processing. This seems more probable than the recurrent processing account of object recognition: with roughly ten synapses from the retina to high-level visual areas, that each take around 10-20 ms to transfer information, behavioral responses occurring within 250 ms after stimulus onset are likely to be produced by processing that is largely feed-forward because it is too early for global recurrent processing to play a role (Lamme & Roelfsema, 2000).

However, the visual system is clearly not a strict feed-forward hierarchy: it contains an abundance of horizontal and feedback connections that support recurrent computations (Felleman & Van Essen, 1991). Several studies have suggested that information coded in early visual areas still remains functionally relevant for categorical representations at later time points (Cichy et al., 2014). Moreover, there have been studies showing that the disruption of visual processing, beyond feed-forward stages (e.g. >150 ms after stimulus onset, or after object-selective activation of higher-order areas) can lead to decreased object recognition performance. For example, TMS studies have shown that categorization (Camprodon et al., 2010) and detection (Koivisto et al., 2011) of objects in natural scenes is affected when activity in early visual areas is disrupted after the feed-forward sweep. This suggests that activity in early visual areas (V1/V2) remains

[1] The ability to rapidly recognize objects despite substantial appearance variation

functionally important for categorization, even after the ‘first round’ of computations.

Contrary to the classical view of the visual hierarchy, it has been proposed that a rapid, global percept of the input (gist) precedes a slow and detailed analysis of the scene (Biederman, 1972; Hochstein & Ahissar, 2002; Oliva, 2005; Oliva & Schyns, 1997), or accompanies detailed feature extraction (Rousselet et al., 2005; Wolfe et al., 2011).

Recently, a growing body of literature suggests that while feed-forward activity may suffice to recognize isolated objects that are easy to discern, the brain employs increasing feedback or recurrent processing for object recognition under more ‘challenging’ natural conditions (Groen, Jahfari, et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Tang et al., 2018). One way in which feedback is thought to facilitate object recognition is through ‘visual routines’ such as curve tracing and figure-ground segmentation. When performing a visual object recognition task, the visual input (stimulus) elicits a feed-forward drive that rapidly extracts basic image features through feedforward connections (Lamme & Roelfsema, 2000). For simple tasks, such as recognizing whether an object is animate or inanimate, or for recognizing clear, isolated objects in sparse scenes, this set of features might be informative enough for successful recognition. While for more detailed tasks or for detecting objects embedded in more complex scenes, the jumble of visual information (‘clutter’) may forward an inconclusive set of features. For those images, extra visual operations (‘visual routines’), such as scene segmentation and perceptual grouping are required. These processes require feedback activity, because they rely on the integration of line segments and other low-level features that are encoded in early visual areas (Crouzet & Serre, 2011; Epshtein et al., 2008; Hochstein & Ahissar, 2002; Lamme & Roelfsema, 2000; Petro et al., 2014; Roelfsema et al., 1999; Self et al., 2019; Zheng et al., 2010). But when is an image considered complex? And how does the brain pick up that an image is complex and that it should therefore employ more extensive processing?

Natural scene statistics index complexity

The scenes that we encounter in our everyday environment do not contain randomly sampled pixels, but they adhere to specific low-level regularities called natural scene statistics. Natural scene statistics have been demonstrated to carry diagnostic information about the visual environment: for example, slopes of spatial frequency spectra estimated across different spatial scales and orientations (‘spectral signatures’) are informative of scene category and spatial layout (Greene & Oliva, 2009b, 2009a; Oliva & Torralba, 2001). Similarly, the width and shape of histograms of local edge information estimated using single- and multi-scale non-oriented contrast filters have been shown to systematically differ with scene category and complexity (Brady & Field, 2000; Tadmor & Tolhurst, 2000). Scene complexity reflected in local contrast distributions can be estimated using an early visual receptive field model that outputs two parameters, contrast energy (CE) and spatial coherence (SC), approximating the scale and shape of a Weibull fit to the local contrast distribution (Figure 1.1C). Earlier studies have shown that visual activity evoked by natural scenes can be well described by scene complexity, suggesting that the brain is adapted or tuned to those statistical regularities and potentially using them during visual perception (Brady & Field, 2000; Ghebreab et al., 2009; Scholte et al., 2009).

Importantly, CE and SC are computed in a biologically plausible way, using a simple visual model that simulates neuronal responses in one of the earliest stages of visual processing. Specifically, they are derived by averaging the simulated population response of LGN-like contrast filters across the visual scene (Ghebreab et al., 2009). Similar to other models of representation in early vision (e.g. Rosenholtz et al. (2012)), these two-parameters thus provide a compressed representation of a scene. Given that CE and SC inform about the strength and coherence of edges in a scene, they are possibly involved in the formation of the initial coarse percept of a scene. In turn, by providing the system with a measure of the 'inherent segmentability' of the scene, they could serve as a complexity index that affects subsequent computations towards a task-relevant visual representation. Indeed, prior work reported effects of scene complexity on both neural responses and behavior (Groen, Jahfari, et al., 2018), indicating enhanced recurrent activity for more complex scenes.

In Chapter 2 of this thesis, we formally modeled the influence of natural scene complexity on perceptual decision-making in an animal detection task. Differences in complexity were task-irrelevant, i.e. not diagnostic of the presence of the target. Our results indicated that scene complexity modulates perceptual decisions through the speed of information processing and evidence requirements, suggesting that the brain is sensitive to low-level regularities even when the task goal is to extract high-level object category information. Having gathered additional evidence for the behavioral relevance of scene complexity, we attempted to dissociate the contributions of the two different axes describing the image complexity 'space' (contrast energy and spatial coherence). In chapter 3, we evaluated whether the effects of complexity on neural responses and behavior could be attributed to the computation of SC and CE directly, as a general measure of complexity, or indirectly, as diagnostic information to estimate other task-relevant scene properties. Using EEG measurements and backward masking, we systematically investigated whether scene complexity influenced the involvement of recurrent processing in object recognition.

To summarize, a growing body of literature suggests that we do not necessarily need to always 'segment' the object from the scene, enabling very rapid object recognition. Only when the image is cluttered and chaotic or when we perform specific tasks that rely on explicit encoding of spatial relationships between parts, we need to employ extra 'visual routines'. While this view seems to mostly emphasize that object recognition relies on the integration of features that exclusively belong to the object, there is also evidence that humans can exploit scene regularities to efficiently search for target objects (Castelhano & Heaven, 2010; Malcolm et al., 2014; Torralba et al., 2006) and use these predictions to facilitate object recognition.

Interaction between scenes and objects

Much of what we know about object recognition emerged from the study of simple, isolated objects and the evaluation of corresponding behavior and neural activity. Visual objects in the real world, however, rarely appear in isolation; they co-vary with other objects and environments. In turn, a scene often holds clues about the object identity or where to search for it (Auckland et al., 2007; Bar, 2004; Bar & Ullman, 1996; Biederman et al., 1982; Greene et al., 2015; Joubert et al., 2008; Oliva & Torralba, 2007; Sun et al., 2011; Võ et al., 2019; Zimmermann et al., 2010). The real world is structured in predictable ways on multiple levels: object-context relations are usually coherent in terms of their physical and semantic content, and they generally occur in typical configurations. The human visual system seems to be sensitive to this structure. For example, participants show enhanced performance for objects appearing in typical locations (e.g. shoes in the lower visual field), and neural representations are sharper as compared to objects appearing in atypical locations (Kaiser et al., 2019). When multiple scene elements are arranged in typical relative positions, cortical processing is more efficient (Kaiser et al., 2019; Kaiser & Cichy, 2018). Additionally, objects appearing on a congruent background (e.g. a toothbrush in a bathroom) are detected more accurately and quickly than objects in an unexpected environment (Davenport & Potter, 2004; Greene et al., 2015; Munneke et al., 2013). Overall, results show that visual processing is tuned to the properties of the real world. So how does scene information influence object recognition? Different accounts of object recognition in scenes propose different 'loci' for contextual effects (Oliva & Torralba, 2007; Võ et al., 2019). It has been argued that a bottom-up visual analysis is sufficient to discriminate between basic level object categories, after which context may influence this process in a top-down manner by priming relevant semantic representations, or by constraining the search space to the most likely objects (e.g. Bar (2003)). Recent studies have also indicated that low-level features of a scene (versus high-level semantic components) can modulate object processing (Lauer et al., 2018; Võ et al., 2019) by showing that seemingly meaningless textures with preserved summary statistics contribute to the effective processing of objects in scenes. In chapter 4 of this thesis, we manipulated information from objects and their backgrounds to better understand how information of the background affects the recognition of objects (Figure 1.1D). Linking human visual processing to performance of models from computer vision (that are described in more detail in the next section) we evaluated what type of computations might underlie the segregation of objects from their backgrounds and the interaction between them.

Probing cognition with deep convolutional neural networks

As already mentioned in the first paragraph of this chapter, the 'problem' of object recognition in natural scenes has not only occupied researchers interested in human behavior. Object recognition in natural scenes is one of the most studied problems in computer vision, in which decades of research have been spent on the development of models that could recognize an object by first segmenting the relevant 'region' from the background and tracing its outline. Despite these efforts, such models never reached human-level performance in general object recognition tasks. In 2012, the success of a model called 'AlexNet' led to a paradigm shift in the field of computer vision (Krizhevsky et al., 2012). Since then, this class of computational models called deep convolutional neural networks (DCNNs), inspired by the hierarchical architecture of the ventral visual stream, have become the most popular approach to object recognition problems. Compared to earlier object recognition algorithms, DCNNs need relatively little preprocessing. While prior models required, for example, hand-engineering of certain filters and preprocessing steps, DCNNs can learn these things by extensive experience with many different examples. This independence from prior knowledge and human effort in feature design was a major advantage, and led to a huge jump in performance.

DCNNs as models of human visual processing

DCNNs and the visual system share two important characteristics. First, both have receptive fields that increase in size along the hierarchy. DCNNs employ receptive fields that act like filters covering the entire input image to exploit the fact that natural images contain strong spatially local correlations. This mimics how the primate visual cortex efficiently accomplishes visual recognition tasks. Second, the further information progresses in the network, the more complex the features become. Various studies have found striking similarities between the respresentations within the artificial neural networks and the cascade of processing stages in the human visual system. In particular, it has been shown that internal representations of these models are hierarchically similar to neural representations in early visual cortex (V1-V3), mid-level (V4), and high-level (IT) cortical regions along the visual ventral stream. For example, neural activity in early areas of visual cortex, as measured with BOLD-MRI, show the highest correspondence with the early layers of DCNNs while higher cortical areas show the highest correspondence with later DCNN layers (Eickenberg et al., 2017; Güçlü & Gerven, 2015; Seeliger et al., 2018; Seibert et al., 2016; Wen et al., 2018). MEG/EEG studies have furthermore shown that early layers of DCNNs explain more of the variance in neural activity early in time, whereas later layers seem to better explain late activity (Cichy et al., 2016; Ramakrishnan et al., 2016). In addition, DCNNs have been shown to predict neural responses in IT, much better than any other computational model (Khaligh-Razavi & Kriegeskorte, 2014; Kubilius et al., 2018; Schrimpf et al., 2018; Yamins et al., 2014). Based on all these previous findings, it has been argued that DCNNs could function as computational models for biological vision (Kietzmann, McClure, et al., 2019; Kriegeskorte, 2015; Lindsay, 2020).

There are, of course, many limitations and non-trivial differences between DCNNs and the human visual system. For example, while the brain is abundant with lateral and feedback connections (Felleman & Van Essen, 1991), most DCNNs are generally feed-forward. The backpropagation algorithm is not considered biologically-plausible enough to be an approximation of how the visual system learns, and there is a stark difference between the simplicity of the model neurons in neural network models and the complexity of real neurons (to name a few). DCNNs often make different types of errors (e.g. Baker et al. (2018) or Geirhos, Rubisch, et al. (2018)). Moreover, several studies have shown how DCNNS are often overly sensitive to changes in the image that would not fool a human observer (Szegedy et al., 2013), and how adding various types of noise, occlusion or blur or even one pixel to standard images leads to a decrease in recognition performance for DCNNS, while leaving human performance intact (Geirhos, Temme, et al., 2018; Ghodrati et al., 2014; Su et al., 2019). Unlike humans, DCNNs are not robust to simple manipultations and tend to generalize poorly beyond the dataset on which they were trained (Serre, 2019).

Still, DCNNs are capable of solving many of the computational problems during visual processing, and on this merit alone one might argue that they deserve our attention as computational models for the human visual system. Possibly, some of the solutions DCNNs provide are similar to biological vision; at the very least, they can be used to explore and generate new hypotheses about the computational mechanisms (Cichy & Kaiser, 2019). Importantly, DCNNs give us the ability to test and compare their performance to humans. By investigating these models, we can perform experiments that would not be possible otherwise, and subsequently gain new insights into how actual neural networks work.

A zoo of artificial animal models

One way of using deep convolutional neural networks to understand human cognition is by using them in a way similar to how we might utilize animal models (Scholte, 2018). An animal model for cognition typically emerges when an animal shows behavior that can be studied in a systematic fashion; such a model becomes even more interesting when the underlying anatomy, architecture and/or physiology of the animal can be linked to the behavior of interest. Current implementations of DCNNs are being studied mostly because they demonstrate an impressive performance on both object and scene recognition. Since these are computer models, we can easily change the architecture or compare different architectures to evaluate the mechanisms that produce this behavior. For example, we can compare different functional architectures or differences in information flow (e.g. feed-forward vs. recurrent; Kietzmann, Spoerer, et al. (2019)) and compare the performance (Figure 1.1B). Leveraging DCNNs as animal models, in chapter 3, we systematically investigated whether recurrent processing is required for figure-ground segmentation during object recognition, by comparing recognition performance in different feed-forward and recurrent DCNN architectures.

In addition to manipulating the structure of the DCNNs, we can also manipulate visual input and evaluate how different models deal with variations in sensory input ('psychophysics') during training (e.g. Xu et al. (2018)) and testing (e.g. Wichmann et al. (2017); Kubilius et al. (2016), Ghodrati et al. (2014); Geirhos et al. (2017)), just like in experiments with human participants. In chapter 4, we controlled the information in objects and backgrounds, as well as the relationship between them to manipulate object-background congruence. We found that with an increase in network depth, there is an increase in the distinction between object- and background information. Importantly, we also found that less deep networks benefited from training on images with objects without a background, while this benefit was decreased or even absent for deeper networks. Overall, our results indicate that scene segmentation, the isolation of an object from its background, is implicitly performed by a network of sufficient depth, without dedicated routines or operations.

Interestingly, DCNNs also allow for the simulation of different 'learning environments' or experimental paradigms, by changing the methods and material used during training. For example, by changing initial conditions prior to training (Mehrer et al., 2020), evaluating the influence of training (Storrs et al., 2020), or by training them on different tasks (e.g. to categorize different scene categories, rather than objects (Zhou et al., 2014) to learn more about potential task- or category-related differences in the internal representations and behavior. Groen, Greene, et al. (2018), conducted a series of analyses to assess the contribution of DCNN layers to fMRI responses in scene-selective cortex, comparing DCNNs that were trained using either object or scene labels. While they did not observe strong differences in terms of their ability to explain fMRI responses, the correlation between layers of both networks decreased for higher/later layers. In chapter 5, we show how training models on different goals (manmade vs. natural scenes, or animate vs. inanimate objects) can elucidate the role of perceptual demands during different experiments.

Finally, with these models, we can even 'damage' or 'lesion' certain regions in the network and evaluate how this influences the model's performance. In chapter 5, we evaluated object and scene categorization in a brain-injured patient with severe object agnosia and category-specific impairments. By removing connections to later layers in our artificial network, we 'mimicked lesions' to higher-order areas in the visual processing stream, and showed an overlap in response patterns.

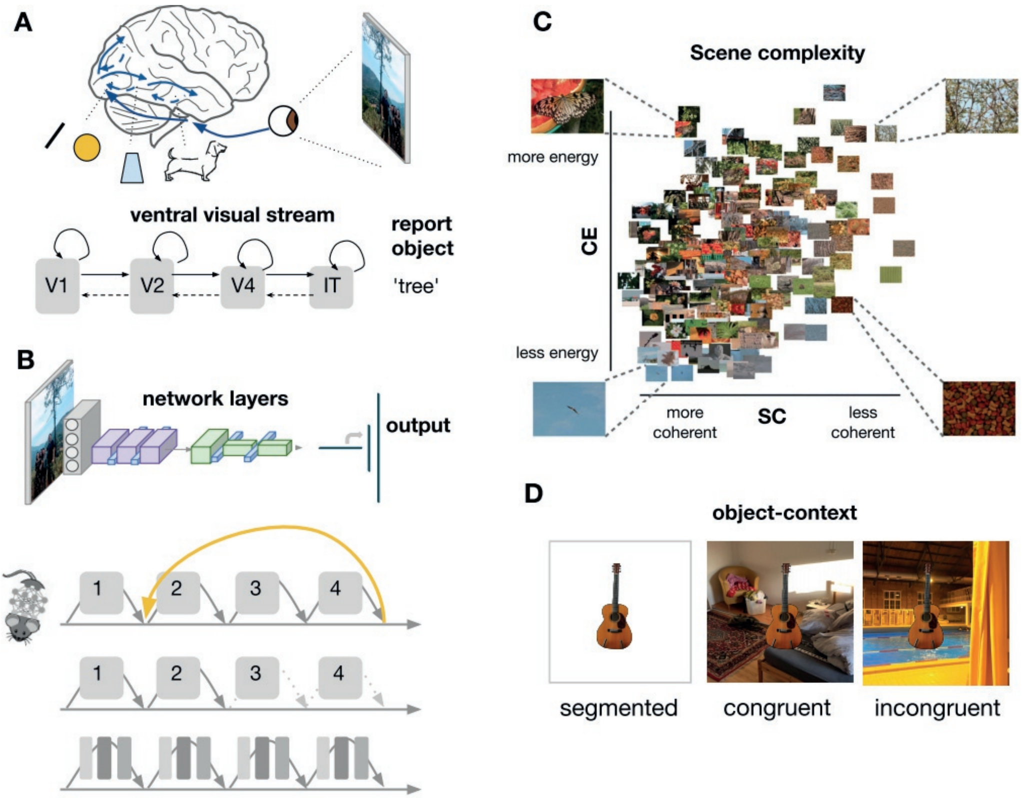

Figure 1.1: **Overview of the approach taken in this thesis. A)** Schematic representation of visual processing **B)** Schematic overview of a Deep Convolutional Neural Network (ResNet-10; He et al. (2016)). By comparing feedforward and recurrent DCNNs architectures, 'lesioning' or manipulating the network, we can explore the underlying computations that produce behavior. **C)** scene complexity as indexed by two parameters. Spatial Coherence (SC) describes the shape of the contrast distribution: it varies with the amount of scene fragmentation (scene clutter). Contrast Energy (CE) describes the scale of the contrast distribution: it varies with the distribution of local contrasts strengths.Figure adapted from Groen et al. (2013). **D)** Example of object-context manipulation by placing objects onto homogenous (segmented), congruent and incongruent backgrounds.

Aims and outline of this thesis

The main focus of this thesis is to study to what extent the human brain is influenced by real-world properties during object perception. More specifically, it sets out to investigate how different functional architectures or differences in information flow (feedforward vs. recurrent) extract information from objects and their backgrounds during object recognition. Additionally, it showcases a potential role for DCNNs as artificial animal models of human visual processing.

In **Chapter 2**, participants performed an animal detection task on low, medium or high complexity scenes as determined by two biologically plausible natural scene statistics, contrast energy (CE) or spatial coherence (SC). While prior work already reported effects of scene complexity on neural responses and behavior, these effects were not formally modeled using perceptual decision making models. In addition, there was no in-depth attempt to dissociate the contributions of the two different axes describing the image complexity 'space' (CE and SC). Diffusion modeling on the reaction times showed that the speed of information processing was affected by low-level scene complexity. Separate manipulation of the two parameters refined these observations by showing that isolated manipulation of SC resulted in weaker but comparable effects, with an additional change in response boundary, whereas the variation of only CE had no effect.

In **Chapter 3**, we evaluated whether these behavioral effects were directly based on the computation of SC and CE, as a general measure of image complexity, or indirectly, as diagnostic information to estimate other task-relevant scene properties. Our results suggest the former, as we show that how object recognition is resolved depends on the complexity of the context: for objects presented in isolation or in 'simple' environments, object recognition appears to be mostly dependent on the object itself, resulting in a situation that can likely be solved within the first feed-forward sweep of visual information processing. When the environment is more complex, recurrent processing appears to be necessary to group the elements that belong to the object and segregate them from the background.

In **Chapter 4**, we investigated the extent to which object and context information is represented and used for object recognition in different deep convolutional neural networks. We show that more layers (i.e. a deeper network) are associated with 'more' or better segmentation, by virtue of increasing selectivity for relevant constellations of features. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and might be one of the ways in which scene segmentation is performed in the brain using recurrent computations.

In **Chapter 5**, we examined what happens when visual information can no longer be reliably mapped onto existing conceptual knowledge. In this study, we evaluated object and scene categorization in a brain-injured patient MS, with severe object agnosia and category-specific impairments. We show that category-specific effects, at least for patient MS, cannot be explained by a purely semantic disorder (i.e. by category membership only). Using Deep Convolutional Neural Networks as ‘artificial animal models’ we further explored the type of computations that might produce such behavior. Overall, DCNNs with ‘lesions’ in higher order areas showed similar response patterns, with decreased performance for manmade (experiment 1) and living (experiment 2) things.

Finally, in **Chapter 6**, we summarize the main findings of this thesis and discuss their implications for our understanding of object recognition in natural scenes.

Chapter 2

Low-level image statistics in natural scenes influence perceptual decision-making

Abstract A fundamental component of interacting with our environment is gathering and interpretation of sensory information. When investigating how perceptual information influences decision-making, most researchers have relied on manipulated or unnatural information as perceptual input, resulting in findings that may not generalize to real-world scenes. Unlike simplified, artificial stimuli, real-world scenes contain low-level regularities that are informative about the structural complexity, which the brain could exploit. In this study, participants performed an animal detection task on low, medium or high complexity scenes as determined by two biologically plausible natural scene statistics, contrast energy (CE) or spatial coherence (SC). In experiment 1, stimuli were sampled such that CE and SC both influenced scene complexity. Diffusion modeling showed that both the speed of information processing and the required evidence were affected by low-level scene complexity. Experiment 2a/b refined these observations by showing how isolated manipulation of SC resulted in weaker but comparable effects, whereas manipulation of only CE had no effect. Overall, performance was best for scenes with intermediate complexity. Our systematic definition quantifies how natural scene complexity interacts with decision-making. We speculate that CE and SC serve as an indication to adjust perceptual decision-making based on the complexity of the input.

This chapter is published as: Seijdel, N., Jahfari, S., Groen, I.I.A. & Scholte, H.S. (2020). Low-level image statistics in natural scenes influence perceptual decision-making. *Scientific Reports 10*, 10573. https://doi.org/10.1038/s41598-020-67661-8

Introduction

During decision-making, observers extract meaningful information from the sensory environment in a limited amount of time. In recent computational accounts of perceptual decision-making, sensory evidence for a decision option is integrated and accumulates over time until it reaches a certain boundary (Gold & Shadlen, 2007; Heekeren et al., 2008). Across these computational accounts, the speed of evidence accumulation is thought to depend on the quality or strength of sensory information available (the drift rate, as formalized with the well-known drift diffusion model (Ratcliff & McKoon, 2008).

In the current study, we aimed to investigate how decision-making processes are influenced by low-level image properties, diagnostic of scene complexity. While multiple studies have shown that specific image properties (such as spatial frequency, or stimulus strength) interact with decision-making, they manipulate visual information into "unnatural" stimuli. For example, we recently showed that image quality modulates response inhibition, and decision-making processes (Jahfari et al., 2015), by manipulating the spatial frequencies of images. Ultimately, however, our goal is to understand how decision processes are influenced by information in natural scenes (Malcolm et al., 2016). The scenes that we encounter in our everyday environment do not contain randomly sampled pixels, but adhere to specific low-level regularities called natural scene statistics. Natural scene statistics have been demonstrated to carry diagnostic information about the visual environment: for example, slopes of spatial frequency spectra estimated across different spatial scales and orientations ('spectral signatures') are informative of scene category and spatial layout (Greene & Oliva, 2009b, 2009a; Oliva & Torralba, 2001). Similarly, the width and shape of histograms of local edge information estimated using single- and multi-scale non-oriented contrast filters have been shown to systematically differ with scene category and complexity (Brady & Field, 2000; Ghebreab et al., 2009; Scholte et al., 2009).

Earlier studies have shown that visual activity evoked by natural scenes can be well described by scene complexity, suggesting that the brain is adapted or tuned to those statistical regularities (Ghebreab et al., 2009; Scholte et al., 2009), and potentially using them during visual perception. Scene complexity reflected in local contrast distributions can be estimated using an early visual receptive field model that outputs two parameters, contrast energy (CE) and spatial coherence (SC), approximating the scale and shape of a Weibull fit to the local contrast distribution, respectively (see Supplementary section 1). CE and SC reflect different aspects of the local contrast distribution: CE approximates the scale parameter of the Weibull fit and reflects the average local contrast strength in an image. SC approximates the shape parameter of the Weibull fit and reflects to what degree the contrast distribution resembles a power law or Gaussian distribution. Cluttered or complex scenes, with high CE/SC values, have more Gaussian (bell-shaped) distributions compared to sparse or simple scenes with low CE/SC values (power-law shaped), that often contain one or a few salient objects (Figure 2.1; adapted from Groen et al. (2013)).

Importantly, CE and SC are computed using a simple visual model that simulates neuronal responses in one of the earliest stages of visual processing. Specifically, they are derived by averaging the simulated population response of LGN-like contrast filters across the visual scene (Scholte et al., 2009). Similar to other models of representation in early vision (e.g. Rosenholtz et al. (2012)), these two-parameters thus provide a compressed representation of a scene. In turn, they could serve as a complexity index that affects subsequent computations towards a task-relevant visual representation.

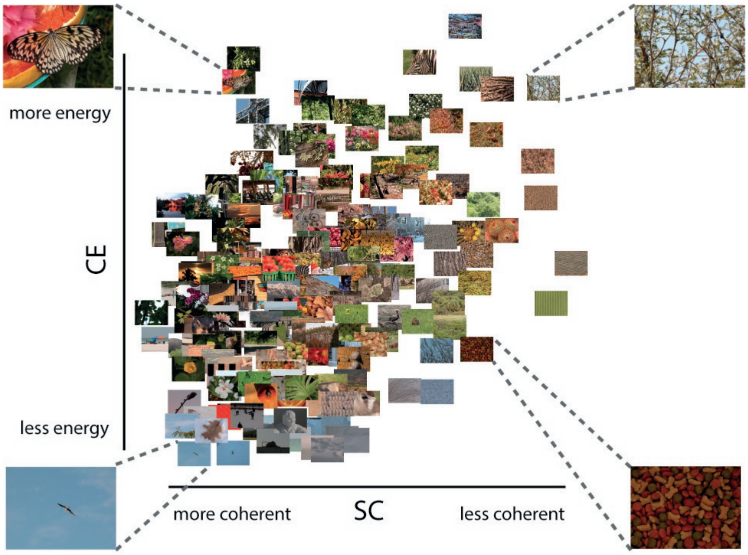

Figure 2.1: **A Subselection of stimuli plotted against their CE and SC values.** Figure adapted from Groen et al. (2013). SC (the approximation of the γ parameter of the Weibull function) describes the shape of the contrast distribution: it varies with the amount of scene fragmentation (scene clutter). CE (the approximation of the β parameter of the Weibull function) describes the scale of the contrast distribution: it varies with the distribution of local contrasts strengths. Four representative pictures are shown in each corner of the parameter space. Images that are highly structured (e.g., a street corner) are found on the left, whereas highly cluttered images (e.g., a forest) are on the right. Images with higher figure-ground separation (depth) are located on the top, whereas flat images are found at the bottom.

Here, we investigated whether task-irrelevant manipulations of SC and CE interact with perceptual decision-making by using the drift-diffusion model (DDM). By considering response time distributions for both correct and incorrect choices, the DDM models the speed of evidence accumulation, as well as the amount of evidence required to make a decision. In experiment 1, stimuli were selected such that both CE and SC co-varied with scene complexity, with increasing values representing more complex natural scenes. This is the 'natural situation', since SC and CE are typically correlated within our natural environment. To refine the observations in experiment 1, in experiment 2a and 2b, we also selected stimuli in such a way that the effects for both parameters could be evaluated separately.

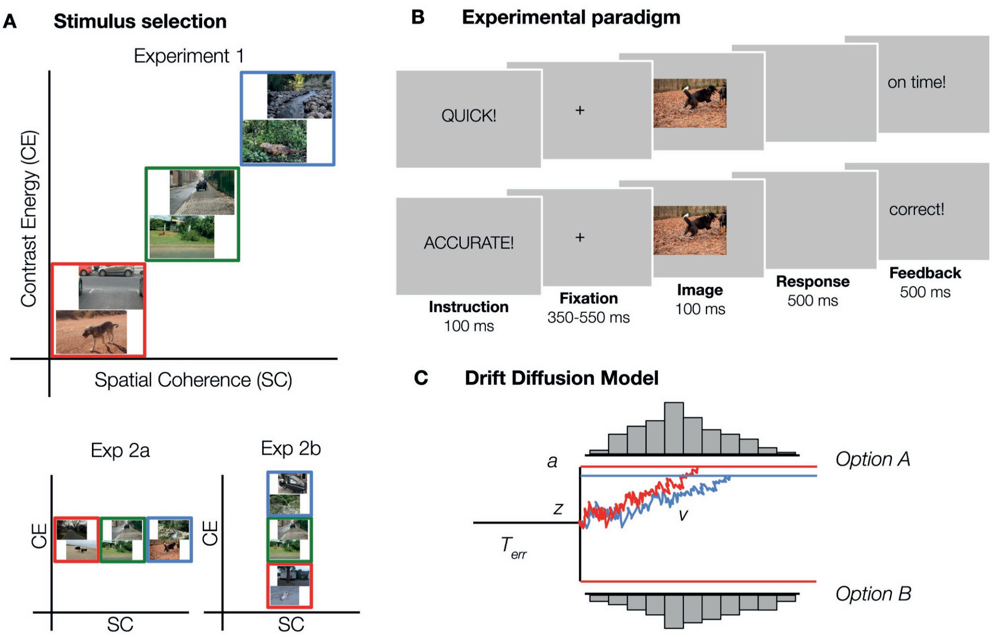

Figure 2.2: **Experimental design and methods. A)** Examples of the stimuli used in experiment 1, 2a and 2b. Images varied both in SC and CE (red = low, green = medium, blue = high) in experiment 1. To investigate whether it is meaningful to differentiate between SC and CE, the two parameters were manipulated separately in experiments 2a (SC) and 2b (CE). For each condition, 80 animal and non-animal scenes were selected. **B)** Experimental paradigm. Participants categorized scenes based on the presence or absence of an animal. On half of the trials, participants were asked to respond as quickly as possible (“speed trials”), as indicated by a pre-cue. On the other half of the trials, participants had to respond as accurate as possible (“accurate trials”). **C)** Schematic representation of the Drift Diffusion Model. From a starting point z, information begins to accumulate in favor of one of the options with drift rate v until it reaches a boundary a, and the decision is made. Non-decision time $Terr$ captures the processes that are unrelated to decision-making, such as response execution.

Materials and methods

Experiment 1

In experiment 1, we investigated the combined influence of SC and CE on decision-making. As SC and CE are generally highly correlated, varying them together provides the strongest manipulation of information. We expected the drift rate to decrease with increased scene complexity, with an additional shift in the amount of evidence required (boundary) reflecting potential strategic adjustments to the complexity of the scene.

Participants

Twenty participants (7 males) aged between 18 and 25 years (M = 21.9, SD = 1.9) with normal or corrected-to-normal vision, gave written informed consent prior to participation and were rewarded with research credits or monetary compensation. The ethics committee of the University of Amsterdam approved the experiment. All experimental protocols and methods described below were carried out in accordance with the guidelines and regulations of the University of Amsterdam.

Stimuli

480 images (640*480 pixels, full-color) were obtained from a previous study by Groen et al. (2010). The complete image set contained 7200 scenes from online databases, including the INRIA holiday database (Jegou et al., 2008), the GRAZ database (Opelt et al., 2006), ImageNet (Deng et al., 2009) and the McGill Calibrated Color Image Database (Olmos & Kingdom, 2004). For each scene, we computed CE and SC values using the model described in Ghebreab et al. (2009) and Groen et al. (2013), and selectively sampled scenes for three conditions: low, medium and high (Figure 2.2). Each condition contained 160 scenes, half of which contained an animal. Importantly, within conditions, animal and non-animal scenes were matched in CE and SC values such that these two categories did not differ from each other in mean or median values (mean: all $t(158) < 1.14$, all $p > 0.26$, median: all $z < 1.08$, all $p > 0.28$).

Procedure

Participants performed an animal/non-animal categorization task (Thorpe et al., 1996) (Figure 2.2). Scenes were presented in randomized sequence, for a duration of 100 ms. Between trials, a fixation-cross was presented with a semi-randomly duration (350, 400, 450, 500 or 550 ms), averaging to 450 ms. There were two trial instructions, that appeared on screen before every trial in randomly alternating blocks of 20 trials: on "speed trials", participants were asked to respond as fast as possible, whereas on "accuracy trials", they responded as accurately as they could. While instruction influences both the accuracy and duration of decisions, the ease of evidence accumulation (drift rate) should remain constant (Ratcliff, 2014). Using a Speed-Accuracy manipulation allows for a stronger and more sensitive test of the influence of scene complexity on perceptual decision-making. If animal detection in more complex scenes is indeed associated with more cautious or elaborate processing, performance in the high condition should be most affected for "speed trials, in which extensive visual processing is potentially limited by time constraints. Therefore, we aimed to specify how the processing of natural scenes can modulate decision-making processes when participants emphasize accuracy - and allow ample time for processing - or speed. Every scene was presented once for both instructions (960 trials in total). Keyboard buttons were switched halfway (based on a simultaneous EEG study). Comparing % errors in blocks before and after the switch did not indicate switch costs: Mbefore = 0.13, SD = 0.15;

Mafter = 0.16, SD = 0.11, only taking participants into account for which the same instruction was repeated before and after the switch, averaged across experiments. On speed trials, participants received feedback on their response time (”on time” < 500 ms > “too slow”). On accuracy trials, participants were presented with “correct” and “incorrect” feedback. When participants didn’t respond, “miss” appeared on screen. Par-ticipants were seated ~90 cm from the monitor such that stimuli subtended ~10x14° of visual angle. Images were presented at eye-level on a 23-inch Asus LCD display (sRGB, 2.27 gamma, 1.31 dE) with a spatial resolution of 1080*1920 pixels, at a refresh rate of 60Hz, using Presentation (version 17.0, Neurobehavioral Systems, Inc.). The ambient illumination in the room was kept constant across different participants.

Hierarchical Drift Diffusion Model

We fitted a hierarchical version of the DDM (HDDM; Wiecki et al. (2013)) using the RT distributions of correct and incorrect responses. HDDM uses a hierarchical Bayesian estimation, that uses MCMC sampling to estimate the joint posterior distribution of all model parameters, and has been described as method of preference in estimating drift rates for a small number of observations (in the order of 100-20; Ratcliff & Childers (2015)). HDDM assumes that during decision-making, information begins to accumulate from a starting point z, in favor of one of the options with drift rate v until it reaches a boundary a, and the decision is made. Non-decision time $Terr$ captures the processes that are unrelated to the decision-making, such as response execution. (Figure 2.2).

First, we evaluated five models in which drift rate (v) and boundary (a) were either fixed or varied across trial type (speed, accurate) and/or scene complexity (low, medium, high). Using the Deviance Information Criterion (DIC) for model selection we established that, next to varying response boundary across trial type (ΔDIC = -3404 compared to fixed), varying both parameters across scene complexity was justified to account for the data (Spiegelhalter et al., 2002). This fit produced lower DIC values compared to a fit in which the drift rate (ΔDIC = -133.3), response boundary (ΔDIC = -40.4) or both (ΔDIC = -68.1) were fixed across complexity. Then, to assess the trial-by-trial relationship between scene complexity and drift rate (v) and boundary separation (a), we fitted eighteen alternative regression models. Both linear models (SC/CE centered around zero), and second-order polynomial models (quadratic) were fitted to examine whether the relationship was curvilinear (e.g. followed an inverted U-shape). We never included both scene statistics simultaneously, as their high correlation leads to multicollinearity and unstable coefficient estimates. To take into account the effect of task instruction on the response boundary a, we estimated two intercepts for this parameter (speed and accuracy) using the `depends_on` key argument. For each model, we ran four separate chains with 5,000 samples. The first 200 samples were discarded (burn), resulting in a trace of 19200 samples. Models were tested for convergence using visual inspection of the group level chains and the Gelman-Rubin statistic, which compares the intra-chain variance of the model to the intra-chain vari-ance of the different runs. It was checked that all group-level parameters had an Rhat between 0.98-1.02. For the best fitting model (lowest DIC), we ran posterior predictive checks by averaging 500 simulations generated from the model’s posterior to confirm it could reliably reproduce patterns in the observed data. Bayesian hypothesis testing was performed on the group-level posterior densities for means of parameters. The probability measure P was obtained by calculating the percentage of the posterior < 0 (see Supplementary section 2).

Experiment 2

A key question is whether the effects found in experiment 1 are driven by the two scene statistics together, as they are generally highly correlated in our natural environment, or whether one of them is the primary cause, as suggested by the SC preference in our optimal HDDM model. To refine our interpretation, we systematically manipulated SC while keeping CE constant (experiment 2a) and vice versa (experiment 2b). Experimental procedure and analyses occurred as in experiment 1, except where otherwise indicated.

Participants

Twenty-four participants (4 males; aged 18-28 years, M = 21.8, SD = 2.7) participated in experiment 2a; Twenty-seven participants (7 males; aged 18-27 years, M = 21.4, SD = 2.5) participated in experiment 2b. All participants gave written informed consent prior to participation and were rewarded with research credits or monetary compensation. The ethics committee of the University of Amsterdam approved the experiment, and all experimental protocols and methods described below were carried out in accordance with the guidelines and regulations.

Stimuli

A new selection of 480 scenes was composed from the same image set as in experiment 1, except that each condition was now defined by either its SC (experiment 2a) or its CE (experiment 2b) values while the other was kept constant at intermediate values (Figure 2.2).

Hierarchical Drift Diffusion Model

In experiment 2a we established that, next to varying response boundary across trial type (ΔDIC = -3426 compared to fixed), varying both parameters across SC was justified to account for the data. This fit produced lower DIC values compared to a fit in which the drift rate (ΔDIC = -70.5), response boundary (ΔDIC = -60.3) or both (ΔDIC = -27.8) were fixed across complexity. Next, we evaluated nine regression models to assess the trial-by-trial relationship between scene complexity (indexed solely by SC), and drift rate and response boundary. For experiment 2b model selection indicated that a model in which, apart from varying response boundary across trial type, the parameters were fixed across CE best explained the observed data. This fit produced lower DIC values compared to a fit in which the drift rate (ΔDIC = -44.7), response boundary (ΔDIC = -76.7) or both (ΔDIC = -48.5) were allowed to vary across complex-ity. Thus, variability in CE alone seems to have no influence on the speed of evidence accumulation or the amount of information required to make a decision. As such, further regression analyses were not justified.

Data and code availability

Data and code to reproduce the analyses are available at the Open Science Framework (https://doi.org/10.17605/OSF.IO/J2AB9) and at https://github.com/noorseijdel/2019_scenestats.

Results

Experiment 1

Data from one participant were excluded for excessive errors (>23%, 2.8 SD > mean). RTs <100 ms were considered "fast guesses" and removed. The repeated-measures ANOVA on RT (on correct trials) revealed main effects of both instruction (speed, accurate) and scene complexity (low, medium, high), but no interaction effect, F(36) = 0.261, $p > 0.77$. Similarly, the repeated-measures ANOVA on error rates revealed main effects but no interaction effect, F(36) = 0.177, $p > 0.83$. As expected, responses were faster and less accurate when given a "speed" instruction, in comparison to "accurate". Because there was no interaction, RTs and error rates were collapsed over speed and accurate trials to further understand how scene complexity modulates decision-making. Bonferroni correction was used for all comparisons.

A repeated-measures ANOVA, with factor scene complexity differentiated RTs across the three conditions, F(2,36) = 19.81, $p < .001$, η^{2par} = .524 (Figure 2.3. Participants responded slower for high (complex) scenes than for medium-, t(18) = -7.293, $p < .001$, and low scenes, t(18) = -3.914, $p = .001$. There was also a main effect on error rates, F(2, 36) = 14.26, $p < .001$, η^{2par} = .442. Participants made more errors for high scenes than for medium, t(18) = -4.493, $p < .001$, and low scenes, t(18) = -2.752, $p = .013$. Remarkably, participants made fewer mistakes on medium scenes than on low SC/CE scenes, t(18) = 3.405, $p = .003$ (Figure 2.3)

Thus, based on the reaction times and error rates, we were able to observe a decrease in performance for low and high complexity scenes. To understand this decrease in performance, we modeled the decision variables drift rate (speed of evidence accu-mulation) and response boundary (evidence requirements). Relative to the null model, the model in which only drift rate was affected by both SC and SC^2 provided the best fit (ΔDIC = -71.0, Figure 2.3), compared to models only including the centered or squared SC values and/or including a varying response boundary (see Supplemen-tary section 2). That is, low and high SC were associated with a decreased drift rate (inverted U-shape; $P < 0.001$), as indicated by a negative shift in the posterior distri-bution. In other words, scene complexity influences the speed of information accu-mulation, resulting in higher reaction times and more errors for low and high complex scenes.

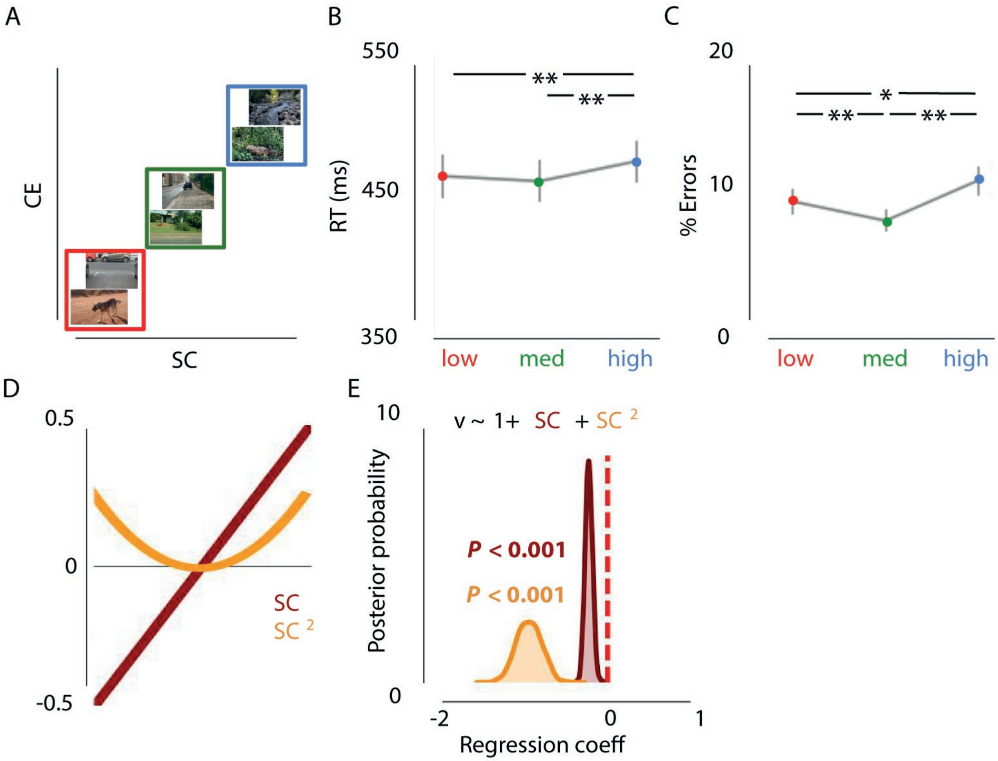

Figure 2.3: **Effects of Spatial Coherence and Contrast Energy on animal vs. non-animal categorization. A)** Examples of the stimuli used in experiment 1. Images varied both in SC and CE (red = low, green = medium, blue = high). **B/C)** Results of experiment 1 indicate worse performance for images with high SC/CE, as indicated by higher RTs and lowered accuracy. Error bars represent 1 SEM. * = $p < .05$, ** = $p < .01$, Bonferroni corrected. Task performance was best for medium SC/CE images. **D)** Schematic representation of the linear and quadratic terms included in the regression model. **E)** Low or high complexity (SC, strongly correlated to CE) was associated with a lower rate of evidence accumulation.

Experiment 2

Results experiment 2a

One participant did not complete the experiment and was excluded from analyses. In contrast to experiment 1, the repeated-measures ANOVA on error rates and RTs showed, apart from the main effects of instruction and scene complexity, an interaction effect, F(42) = 4.351, p = 0.0189. Therefore, results were analyzed separately for "speed" and "accurate" trials to further understand how SC differentially impacts fast or accurate decision strategies.

The repeated-measures ANOVA revealed no main effect of SC on RTs for speeded or accurate trials (all p > 0.104). For error rates, there was a main effect of SC on speed trials, F(1, 44) = 9.189, p < .001, η^{2par} = .295. Participants made fewer er-rors for medium SC scenes compared to both low, t(22) = 3.294, p = 0.003 or high, t(22) = -4.346, p < .001 (Figure 2.4). Notably, SC had no effect on choice errors when participants were motivated to be accurate (p > .103)

Relative to the null model, the model in which drift rate and response boundary were affected by $SC + SC^2$ provided the best fit (ΔDIC = -21; Figure 2.4). As in experiment 1, low and high SC were associated with a decreased drift rate (inverted U-shape), as indicated by negative shifts in the posterior distribution (*P* < .001). Additionally, those scenes were associated with a decreased response boundary (*P* < .001; Figure 2.4), potentially to still allow for a timely response. Thus, SC influences the speed of information accumulation and the evidence requirements, resulting in more errors for low and high complex scenes when pressed for time.

Results experiment 2b

Two participants were excluded because of excessive errors (>25%) or excessive omissions (>40%). A repeated-measures ANOVA with factors scene complexity (low, medium, high) and instruction (speed, accurate) revealed no interaction effects for RTs, F(48) = 0.093, p > 0.9, or errors, F(48) = 1.216, p > 0.3. Consistently, no main effect of CE was observed on RTs or errors when speeded and accurate trials were collapsed (Figure 2.5; all p > 0.306).

Discussion

This study systematically investigated the interaction between low-level statistics in natural scenes and perceptual decision-making processes. Results indicate that scene complexity, as indexed by two parameters (SC, CE), modulates perceptual decisions through the speed of information processing. Experiment 2a/b refined these observations by showing how the isolated manipulation of SC alone results in weaker yet comparable effects, whereas the manipulation of CE has no effect. By using natural stimuli, we show that task performance was best on medium complex images. Overall, these results show that very basic properties of our natural environment influence perceptual decision-making.

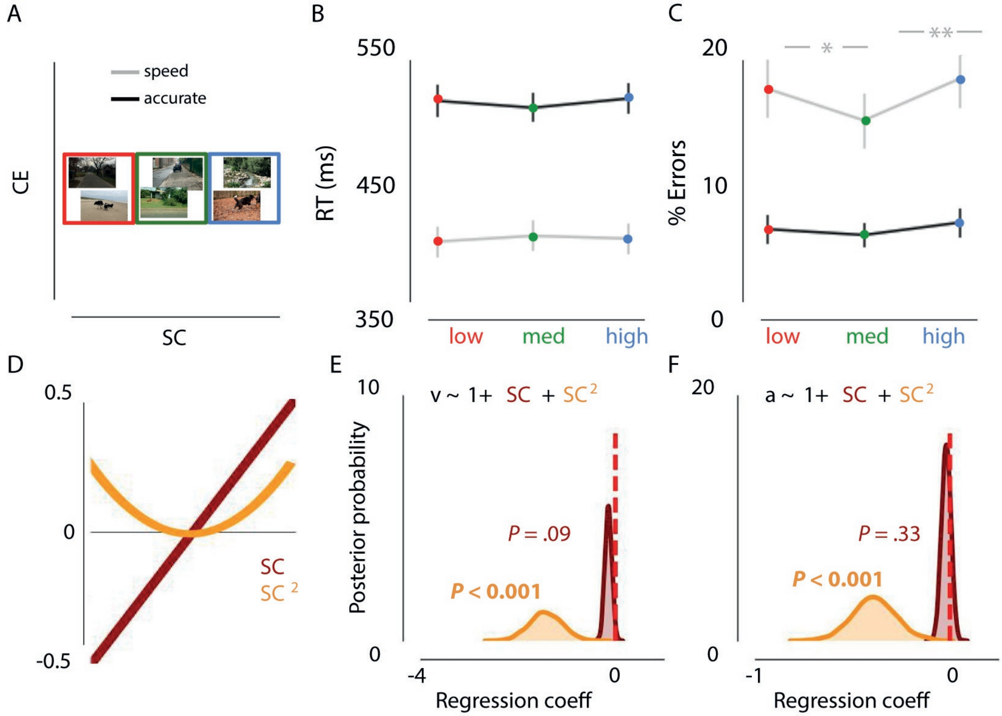

Figure 2.4: **Effects of SC (controlling for Contrast Energy) on decision-making. A)** Examples of the stimuli used in experiment 2a. Images only varied in SC, while CE was kept constant (red = low SC, green = medium SC, blue = high). **B)** Results showed no influence of SC on RT. **C)** Performance was most optimal for images with medium SC complexity in the speed condition, as indicated by a higher accuracy. Error bars represent 1 SEM. * = $p < .05$, ** = $p < .01$, Bonferroni corrected. **D)** Schematic representation of the linear and quadratic term included in the regression model. **E/F)** Negative shifts in the posterior distributions indicated that low or high complexity (SC) was associated with a lower rate of evidence accumulation and required less evidence to reach a decision (inverted U-shape).

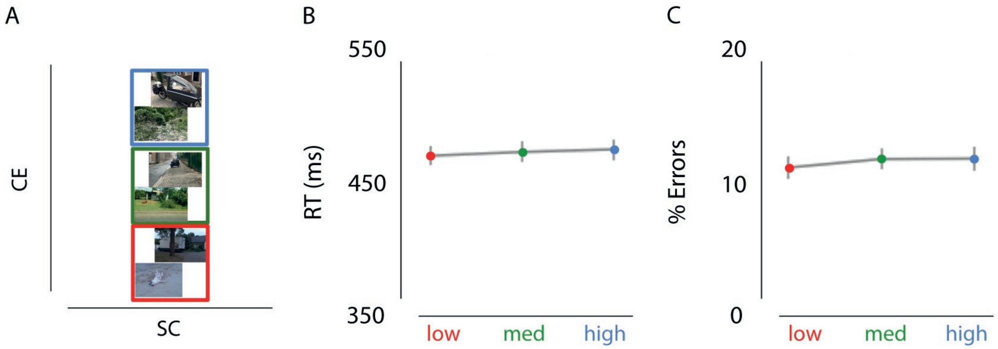

Figure 2.5: **Effects of CE (controlling for Spatial Coherence) on decision-making. A)** Examples of the stimuli used in experiment 2b. Images only varied in CE, while SC was kept constant (red = low CE, green = medium CE, blue = high CE). **B/C)** Results of experiment 2b showed no influence of CE on RT or percentage or errors.

SC and CE together provide a compressed representation of scene complexity. While CE captures information about the amount and strength of edges in a scene, SC indexes higher-order correlations between them, giving an indication of the amount of clutter. In earlier work by Scholte et al. (2009) and Groen et al. (2013); Groen et al. (2016), differences in CE were shown to mainly influence the early part of the ERP, while SC effects arose later (up to 300 ms). In the current study, influences on perceptual decision-making seem to be mainly driven by SC. In experiment 1, when SC and CE were both manipulated, model selection indicated a better fit when changes in drift rate were related to SC (as compared to CE), and in experiment 2 only effects of SC were found. Still, there seems to be an additional influence of CE. The finding that there is no interaction between trial type and complexity condition in Experiment 1 indicates that even for trials in which there is ample time to process the image, scene complexity influences this process. Thus, while participants were faster and more susceptible to making errors when emphasizing speed (compared to accuracy), emphasis on speed or on accuracy did not change the magnitude of the scene complexity effect on both reaction times and errors. We interpret this as showing that the simultaneous manipulation of SC and CE leads to the strongest effects (as compared to experiment 2). In experiment 2a, in which CE was not manipulated, accuracy was decreased for low and high complexity trials only when participants were pressed for time. This suggests that for low and high complexity scenes, visual information processing might be too slow to produce correct responses, especially when participants are motivated to respond quickly and have lower evidence requirements in comparison to accurate instruction trials. In experiment 2a, low and high complexity scenes were, apart from drift rate, also associated with a lowered response boundary. Overall these results suggest that SC is weighed differently when manipulated in isolation. One explanation for the differences between experiment 1 and 2a could be the inherent correlation between the parameters in the real world, as isolating the influence of both parameters separately could have led to an 'unnatural' sub-selection. For this reason, we cannot attribute our results from experiment 2 exclusively to the scene statistics. Whether this is a robust effect should emerge from future research.

From previous studies, using artificial manipulation of stimulus quality, one would expect performance decreases for more complex scenes. For instance, the search slope of reaction times increases with the number of distractors in conjunction search (Wolfe, 1994) and degrading stimulus quality (via spatial filtering) reduces the rate of evidence accumulation (Jahfari et al., 2013). Intuitively 'low' SC scenes are easiest: those scenes are sparser and typically contain the most distinct figure-ground segmentation. Surprisingly, our results suggest a more complex pattern. In experiment 1 and 2a, performance was better on 'medium' than on 'low' scenes. Responses to natural scenes are often hard to predict from studies using artificial stimuli (Felsen & Dan, 2005) because the scenes do not contain simple isolated patterns. But why would scenes with medium SC/CE be processed more efficiently? We outline a number of possible reasons below.

First, it could be that scenes with medium complexity are most commonly encountered in daily life, and that the visual system has become tuned to the statistical regularities of medium scenes (Geisler & Diehl, 2003; Olshausen & Field, 1996), resulting in optimized visual processing.Secondly, it could be the degree to which object context facilitates the recognition process. In natural scenes, objects appearing in a familiar background are detected more accurately and quickly than objects in an unexpected environment (Davenport & Potter, 2004; Greene et al., 2015; Neider & Zelinsky, 2006). Here, most of the 'low' scenes contained little context because the backgrounds were, generally, homogeneous, providing no 'cues' about animal presence or absence. For 'high' images, on the other hand, there may have been too much distraction by spatially unorganized clutter, which does not offer useful cues for animal detection. Third, SC and CE could be related to certain object properties, such as animal size or centrality (the location of the animal in the scene). Additional HDDM analyses however indicated that SC contributed to perceptual decision-making independent of object size, whereas object centrality had no effects (Supplementary Figures S4-S6). Finally, SC/CE could be used as diagnostic information, serving as a building block towards estimating other relevant properties in a scene (e.g. scene clutter, naturalness). Since SC correlates with naturalness ratings (Groen et al., 2013) and, animals are potentially more strongly associated with natural environments, SC could be a diagnostic feature for the animal/non-animal discrimination task. Indeed, post-hoc evaluation of the responses in experiment 1 and 2a indicated a change in bias towards one of the response options (animal or non-animal), depending on the SC value of the scene. However, the pattern of errors, evaluated for animal and non-animal trials separately, was only partly consistent with a naturalness bias (Supple-mentary Figures S7 and S8). In the DDM, effects of a response bias can be explained either by changes in starting point (Δz) or by changes in drift rate (Δv; (Mulder et al., 2012)) or the starting point of the drift rate. Additional modelling suggests that a potential response bias was not reflected in a change in the starting point, and the RT patterns for correct and incorrect trials in our dataset were more in line with a drift bias account (see Supplementary section 5). Crouzet & Serre (2011) have shown that low-level image properties such as SC and CE can relate to human performance in an animal detection task. When they trained a classifier to distinguish between animal/non-animal images based on the Weibull parameters (β and γ), classification performance was above chance, but relatively poor compared to alternative models which included more complex visual features, including oriented contrast (V1-like features) and combinations of oriented linear filter responses (mid-level and higher level features). Moreover, the least animal-like stimuli corresponded to more complex backgrounds, while our analyses of response bias (see Supplementary section 4) suggest the opposite pattern. This suggests that the relation between SC, naturalness and animal detection is not trivial and can vary with stimulus set or image database. Here, we carefully selected images to capture a broad range of CE and SC values, and ensured that animal presence was balanced within each condition. Therefore we believe that the current study is a more sensitive test of effects of low-level contrast statistics on perceptual discrimination than previous post-hoc assessments.

In conclusion, the current study provides clear evidence that SC and CE influence perceptual decision-making in an animal detection task. We propose that, because SC and CE could be plausibly computed in early stages of visual processing, they could indicate the need for more cautious or elaborate processing by providing the system with a global measure of scene complexity (Groen, Jahfari, et al., 2018). Future studies should pinpoint whether this effect is based on the computation of SC and CE directly, as a general measure of complexity, or indirectly, as diagnostic information to estimate other task-relevant scene properties. Given that the rate of evidence accumulation depends on the complexity of the scene, this complexity-dependent adaptation could be reflected in the amount of evidence that is considered sufficient for generating a response. This adaptation, or flexible processing, can help to calibrate the decision process to maximize the goal at hand (e.g. to be accurate or quick).

Supplement to Chapter 2

1. Computation of SC and CE

The following section describes the main computational steps. The code to run the model on an arbitrary input image is available on Github[1].

Natural image statistics: local contrast distribution regularities.

Natural images exhibit much statistical regularity, one of which is present in the distribution of local contrast values. It has been observed (Simoncelli, 1999; Geusebroek and Smeulders, 2002, 2005) that properties that are inherent to natural images, such as spatial fragmentation (generated by the edges between the objects in the scene) and local correlations (due to edges belonging to objects in the image) results in contrast distributions that range between power law and Gaussian shapes, and therefore conform to a Weibull distribution. This regularity (systematic variation in the contrast distribution of natural images) can therefore be adequately captured by fitting a Weibull function of the following form:

$$p(f) = ce^{\frac{(f-\mu)}{\beta}\gamma}$$

Where c is a normalization constant that transforms the frequency distribution into a probability distribution. The parameter μ, denoting the origin of the contrast distribution, is generally close to zero for natural images. We normalize out this parameter by subtracting the smallest contrast value from the contrast data, leaving two free parameters per image, β (beta) and γ (gamma), that represent the scale (beta) and shape (gamma) of the distribution (Geusebroek & Smeulders, 2002, 2005). Beta varies with the range of contrast strengths present in the image, whereas gamma varies with the degree of correlation between contrasts.

LGN model of local contrast statistics: contrast energy and spatial coherence.

In previous work, we found that the beta and gamma parameters of the Weibull distribution can be approximated in a physiologically plausible way by summarizing responses of receptive field models to local contrast (Scholte et al., 2009). Specifically, summing simulated receptive field responses from a model of the parvocellular and magnocellular pathways in the LGN led to accurate approximations of beta and gamma, respectively. In subsequent papers (Groen et al., 2013, 2017) an improved version of this model was presented in which contrast was computed at multiple spatial scales and the LGN approximations were estimated not via summation but by averaging the local parvocellular responses (for beta) and by averaging and divisively normalizing the magnocellular responses for gamma (mean divided by standard deviation). To distinguish the Weibull fitted parameters from the LGN approximations, the LGN-approximated beta value was defined as Contrast Energy (CE) and the LGN-approximated value of gamma as spatial coherence (SC). These modifications, as well

[1] https://github.com/irisgroen/LGNstatistics

as specific parameter settings in the model, were determined based on comparisons between the Weibull fitted values and the CE/SC values, as well as model fits to EEG responses, in separate, previously published image sets (Ghebreab et al., 2009, Scholte et al., 2009). We outline the main computational steps of the model below:

Main computational steps of the model

Step 1: RGB to color opponent space. For each image, the input RGB values were converted to opponent color space using the Gaussian color model described in (Geusebroek, Van den Boomgaard, Smeulders & Geerts, 2001), yielding 3 opponent color values per pixel (grayscale, blue-yellow, red-green; Koenderink, Van De Grind & Bouman, 1972).

Step 2: Multi-scale local contrast detection. Each color opponent layer was convolved with isotropic exponential filters (Zhu and Mumford, 1997) at five octave scales (Croner and Kaplan, 1995). Two separate filter sets were used: smaller filter sizes (0.12, 0.24, 0.48, 0.96, and 1.92 degrees) for CE and larger filter sizes (0.16, 0.32, 0.64, 1.28, and 2.56 degrees) for SC (Ghebreab et al., 2009). Following the LGN suppressive field approach (Bonin et al., 2005), all filter responses were rectified and divisively normalized.

Step 3: Scale selection. Per parameter (CE or SC) and color-opponent layer, one filter response was selected for each image location from their respective filter set using minimum reliable scale selection (Elder and Zucker, 1998). In this MIN approach, the smallest filter size that yields an above-threshold response is preferred over other filter sizes. Filter-specific noise thresholds were determined from a separate image set (Corel database) (Ghebreab et al., 2009).

Step 4: Spatial pooling. Applying the selected filters for each image location results in two contrast magnitude maps: one highlighting detailed edges (from the set of smaller filter sizes, for CE) and the other more coarse edges (from the set of larger filter sizes, for SC) per color opponent-layer. To simulate the different visual field coverage of parvo- and magnocellular pathways, a different amount of visual space was taken into account for each parameter in the spatial pooling step. For CE, the central 1.5 degrees of the visual field was used, whereas for SC, 5 degrees of visual field was used. Finally, the estimated parameter values were averaged across color-opponent layers resulting in a single CE and SC value per image.

2. HDDM model comparison and convergence

First, we evaluated five models in which drift rate (v) and boundary (a) were either fixed or varied across trial type (speed, accurate) and/or scene complexity (low, medium, high).

Table 2.1: **HDDM model [illegible] to determine whether varying across scene complexity was [illegible] to account for the data.** We evaluated five models in which drift rate (v) and boundary (a) were either fixed or varied across trial type (speed, accurate) and/or scene complexity (low, medium, high).

	drift rate (v)	boundary (a)	DIC - exp 1	DIC - exp 2a	DIC - exp 2b
model 0	-	-	-11453.46	-8545.74	-7240.21
model 1	-	instruction	-14857.92	-11971.86	**-10109.50**
model 2	complexity	instruction	-14885.64	-11939.36	-10064.76
model 3	-	instruction, complexity	-14792.76	-11929.10	-10032.79
model 4	complexity	instruction, complexity	**-14926.04**	**-11999.62**	-10060.99

Then, to assess the trial-by-trial relationship between scene complexity and drift rate (v) and boundary separation (a), we fitted eighteen alternative regression models.

Table 2.2: **HDDM Regression models.** Drift rate v and boundary a were either allowed to vary across scene complexity (indexed by SC or CE) or fixed. Both linear models (SC/CE centered around zero), and second-order polynomial models (quadratic) were fitted to examine whether the relationship was curvilinear (e.g. followed an inverted U-shape).

	drift rate (v)	boundary (a)	both (v, a)
SC	model 1	model 2	model 3
SC^2	model 4	model 5	model 6
$SC + SC^2$	model 7	model 8	model 9
CE	model 10	model 11	model 12
CE^2	model 13	model 14	model 15
$CE + CE^2$	model 16	model 17	model 18

Table 2.3: **Means of the posterior distributions.**

parameter	Experiment 1	Experiment 2a
t	0.18	0.23
z	0.27	0.29
$v\_Intercept$	3.69	3.16
v_SC	0.23	-0.14
v_SC^2	-0.96	-1.43
$a\_Intercept(Ac)$	1.93	1.85
$a\_Intercept(Sp)$	1.46	1.47
a_SC	-	-0.02
a_SC^2	-	-0.39

3. HDDM analyses incorporating contextual factors

The following section describes the methods for the additional analyses to evaluate potential contextual factors that could correlate with SC and limit the detection

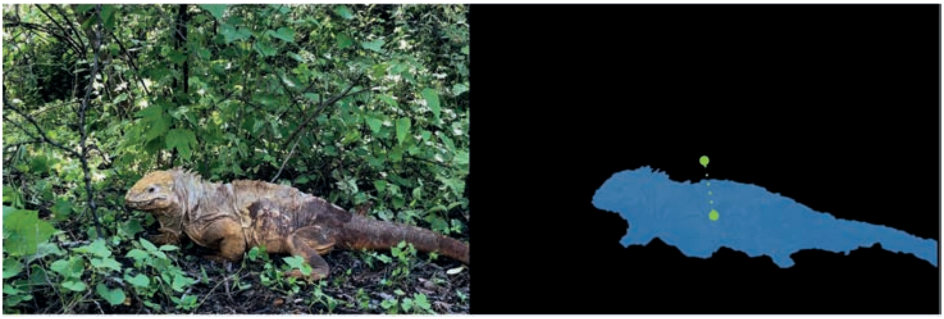

Figure 2.6: **Example of computing object (animal) coverage and centrality.** Object size was computed by taking the percentage of the image that was covered by the animal (manually segmented). Object centrality was computed by taking the distance from the center of mass (CoM) of the animal to the center of the screen (length of green dotted line, in pixels).

task. Specifically, we parameterized two characteristics, object size and centrality. We have focused on these two factors, because just like CE and SC, they were image-computable, i.e. they could be derived by performing calculations on the pixels in the image.

Computing contextual factors

Object size was computed by taking the percentage of the image that was covered by the animal (manually segmented). Object centrality was computed by taking the distance in pixels from the center of mass (CoM) of the animal (computed by interpreting the image as a 2D probability distribution) to the center of the screen (see Supplementary Figure 2.6).

Evaluating the relationship with SC and CE

There was no correlation between SC or CE and object coverage (experiment 1; SC: r = 0.018; CE: r = 0.025) or centrality (SC: r = -0.13; CE: r = -0.09). To evaluate whether SC explains unique variance after accounting for these properties, we included both variables in our HDDM regression analysis, alongside SC.

For experiment 1, results showed an influence of object size (coverage) on the drift rate, with a higher drift rate for images with larger animals as indicated by a positive shift in the posterior distribution (Supplementary Figure S2; $P < .001$). For object centrality, however, we found no effect, and inspection of this variable indicated a low variability: most animals were located quite central. In experiment 2a, as in experiment 1, larger animals were associated with a higher drift rate (Supplementary Figure S3; $P < .001$). Most importantly, for both experiments, the effect of SC remained. This indicates that, even though object size has an influence on the rate of evidence

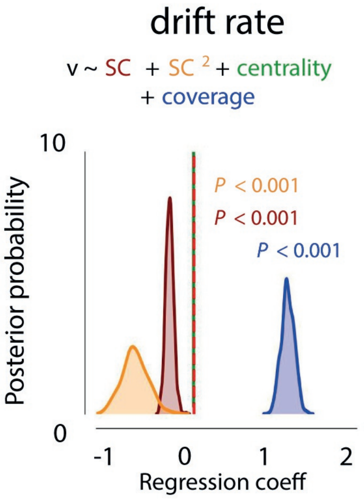

Figure 2.7: **Effects of SC/CE (experiment 1) on drift rate, accounting for object size and centrality.** Bigger animals were associated with a higher rate of evidence accumulation. The effects of SC+SC2 remained, indicating that, even though object size has an influence on the rate of evidence accumulation, SC continues to explain unique variance in the speed of information processing.

accumulation, SC continues to explain unique variance in the speed of information processing. In other words, SC contributes to perceptual decision-making independent of object size, whereas object centrality has no effects.

Full description and code definitions can be found on Github[2]

4. Behavioral analysis evaluating animal/non-animal bias

To investigate whether participants' response bias (towards animal or non-animal) differed with scene complexity, we computed the % animal choices for each participant. Differences between the three conditions (low, med, high) were statistically evaluated using a repeated-measures ANOVA.

For experiment 1, results indicated, apart from a general bias towards the non-animal option (animal choice < 50% for all conditions), that the % animal-responses increases with scene complexity, F(36) = 9.76, $p < 0.001$, η^{2par} = .351 (Supplementary Figure S4). Participants chose 'animal' more often in the high and medium complexity scenes as compared to low, t(18) = -5.104, $p < .001$; t(18) = -2.698, $p = .044$ (Bonferroni corrected). Similar effects were found for experiment 2a (Supplementary Figure S5). There, the percentage of animal responses increased with SC, F(44) = 6.63, $p = 0.003$, η^{2par} = .232. Participants chose 'animal' more often in the high scenes

[2] https://github.com/noorseijdel/2019_scenestats/blob/master/notebooks/Notebook_SceneStats_Context.ipynb

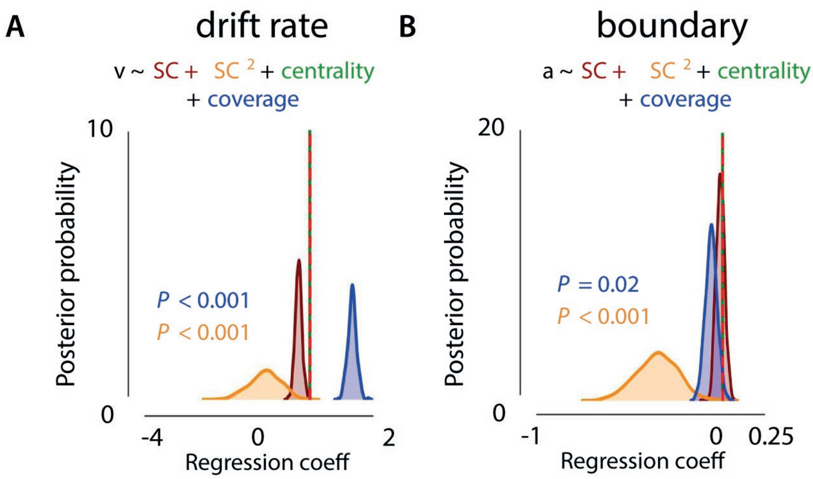

Figure 2.8: **Effects of SC (experiment 2a) on drift rate and response boundary, accounting for object size and centrality.** Bigger animals were associated with a higher rate of evidence accumulation. Again, the effect of SC2 remained, indicating that even though object size has an influence on the rate of evidence accumulation, SC continues to explain unique variance in the speed of information processing.

as compared to low, t(22) = -3.365, *p* = .008 (Bonferroni corrected). In the current experiment, half of the trials in each condition contained an animal. Therefore, this response bias towards animal or non-animal trials can result in an increase in errors in the low and high condition. Analysis of the error rates separately for animal and non-animal trials, indicated for both experiment 1 and experiment 2a that most errors in the low condition were made for animal-trials. In those trials, participants thus seem to 'miss' the animal more often. Errors in high scenes, however, were seemingly not caused by the response bias: while participants reported more animals on non-animal trials (compared to low and medium), they made as many errors on animal trials.

5. HDDM Regression analyses evaluating response bias effects

Following Supplementary section 4, to assess whether SC biases the response (towards animal or non-animal) reflected in changes in the starting point, we fitted several HDDMRegressor models: 1. one model in which we estimate only the response bias z for every complexity condition (low, med, high), such that the bias for animal stimuli is z and the bias for non-animal stimuli is 1-z (z = 0.5 for unbiased decisions in HDDM. 2. one model in which we estimate both v and z. This way, we could measure response-bias (in favor of animal or non-animal) and drift rate for the three conditions (low, med, high) while assuming the same drift rate for both stimuli. 3. one model in which we estimate v, z and a for every complexity condition. 4. one

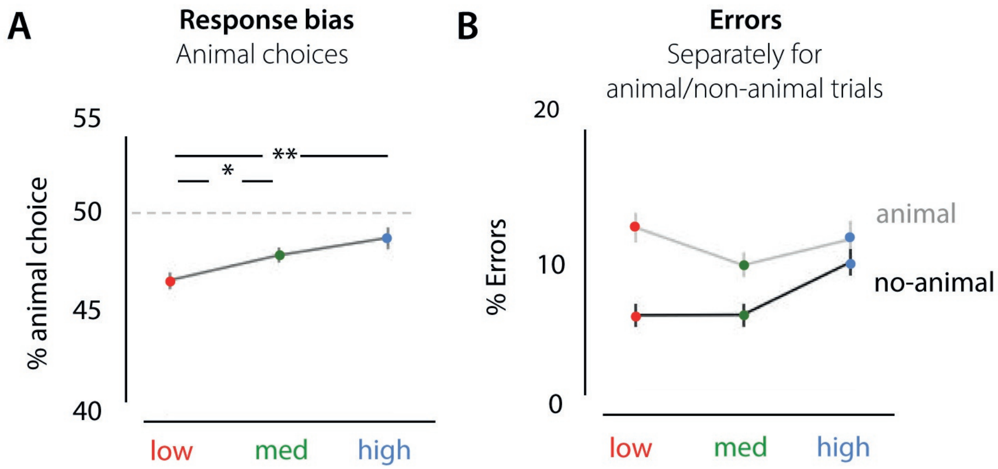

Figure 2.9: **Response bias effects in experiment 1. A)** apart from a general bias towards the non-animal option (animal choice < 50% for all conditions), the % animal-responses increased with scene complexity. **B)** percentage of errors from experiment 1, separately for animal and non-animal trials.

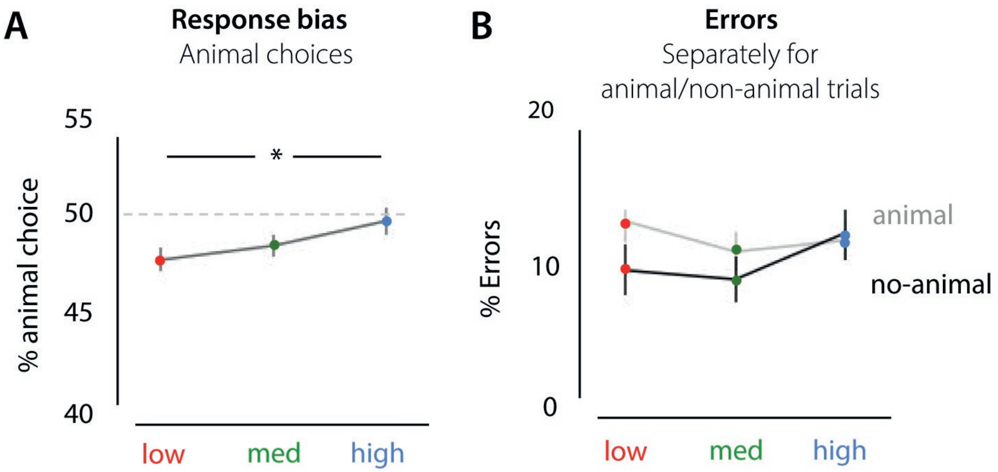

Figure 2.10: **Effects of SC on animal/non-animal responses in experiment 2a. A)** Similar to experiment 1, the % animal-responses increased with SC. **B)** Percentage of errors from experiment 2a, plotted separately for animal and non-animal trials.

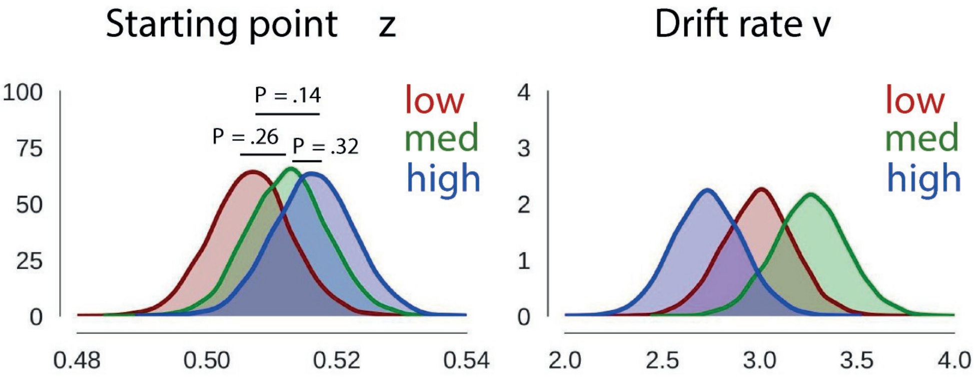

Figure 2.11: **Using HDDMStimCoding to evaluate potential biases towards animal/non-animal across the different conditions for the data obtained in experiment 1.**

model in which we estimate v, z and a for every complexity condition and, using the depends_on key argument, two intercepts for a (speed, accurate) 5. same model as the previous model (4), now using 'medium' as the intercept for z

However, with the properties of our observations and design, models defined in this way do not converge, which makes the interpretation of the parameters uninformative. The traces are non-stationary, and the autocorrelation is high. The histograms look serrated. Full description and code definitions can be found on Github[3]

Then, we fit one model using HDDMStimCoding, in which we estimate v, z for every complexity condition, and a for every complexity condition + speed/accuracy instruction. This model converges. As shown in the figure (Supplementary figure S9) below, the obtained posteriors for z do not differ across our low, med, or high condition. Hence, this evaluation shows no effect of condition (low, med, high) on z when it is allowed to vary.

In the DDM, effects of a response bias can be explained either by changes in starting point (Δz) or by changes in drift rate (Δv; Mulder, Wagenmakers, Ratcliff, Boekel & Forstmann, 2012)) or the starting point of the drift rate. Additional modeling suggests that a potential response bias was not reflected in a change in the starting point and the RT patterns for correct and incorrect trials in our dataset were more in line with a drift bias account:

[3] https://github.com/noorseijdel/2019_scenestats/blob/master/notebooks/Notebook_SceneStats_ResponseBias.ipynb

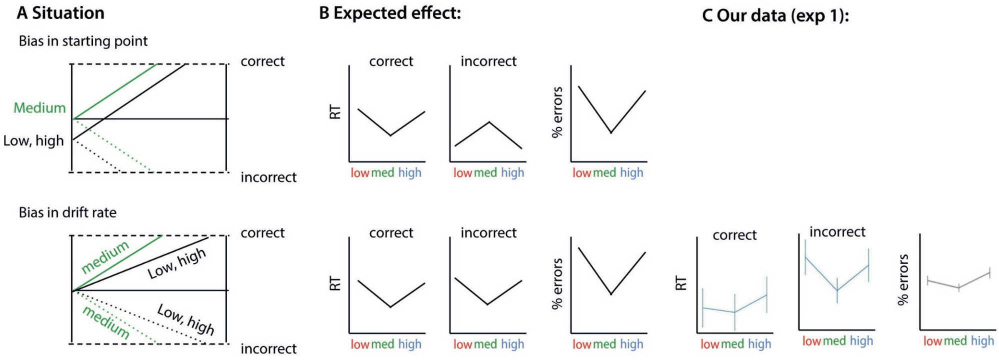

Figure 2.12: **Possible effects of bias on choice behavior (following figure 2 from Mulder et al. (2012)). A)** Effects of bias explained by the drift-diffusion model. When prior information is invalid ('low', 'high') for the choice at hand, subjects will have slower and less correct choices compared with choices where no information is provided (neutral, 'medium'). These effects can be explained by changes in the starting point or the drift rate of the accumulation process. **B)** Both accounts have different effects on RT and accuracy data. **C)** The data from our current experiment is more in line with a drift rate account of response bias.

Chapter 3

Recurrent processing during object recognition: it depends on the need for scene segmentation

Abstract While feed-forward activity may suffice for recognizing objects in sparse scenes, additional visual operations that aid object recognition might be needed for more complex scenes. One such additional operation is figure-ground segmentation; extracting the relevant features and locations of the target object while ignoring irrelevant features. In this study of 60 participants, we show objects on backgrounds of increasing complexity to investigate whether recurrent computations are increasingly important for segmenting objects from more complex backgrounds. Three lines of evidence show that recurrent processing is critical for recognition of objects embedded in complex scenes. First, behavioral results indicated a greater reduction in performance after masking objects presented on more complex backgrounds; with the degree of impairment increasing with increasing background complexity. Second, electroencephalography (EEG) measurements showed clear differences in the evoked response potentials (ERPs) between conditions around 200ms - a time point beyond feed-forward activity and object decoding based on the EEG signal indicated later decoding onsets for objects embedded in more complex backgrounds. Third, Deep Convolutional Neural Network performance confirmed this interpretation; feed-forward and less deep networks showed a higher degree of impairment in recognition for objects in complex backgrounds compared to recurrent and deeper networks. Together, these results support the notion that recurrent computations drive figure-ground segmentation of objects in complex scenes.

This chapter is under review as: Seijdel, N.*, Loke, J.*, van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., de Haan, E.H.F. & Scholte, H.S. (n.d.). Recurrent processing during object recognition: it depends on the need for scene segmentation

Significance statement

The incredible speed of object recognition suggests that it relies purely on the fast feed-forward buildup of perceptual activity. However, this view is contradicted by studies showing that disruption of recurrent processing leads to decreased object recognition performance. Here we resolve this issue by showing that how object recognition is resolved depends on the context in which the object is presented. For objects presented in isolation or in 'simple' environments, feed-forward activity seems sufficient for successful object recognition. However, when the environment is more complex, recurrent processing is necessary to group the elements that belong to the object and segregate them from the background.

Introduction

The efficiency and speed of the human visual system during object categorization suggests that a feed-forward sweep of visual information processing is sufficient for successful recognition (VanRullen & Thorpe, 2002). For example, when presented with objects in a rapid serial visual presentation task (RSVP; Potter & Levy (1969)), or during rapid visual categorization (Thorpe et al., 1996), human subjects could still successfully recognize these objects, with EEG measurements showing robust object-selective activity within 150 ms after object presentation (VanRullen & Thorpe, 2001). Given that there is substantial evidence for the involvement of recurrent processing in figure–ground segmentation (Lamme & Roelfsema, 2000; Wokke et al., 2012), this seems inconsistent with recognition processes that rely on explicit encoding of spatial relationships between parts and suggest instead that rapid recognition may rely on the detection of an 'unbound' collection of image features (Crouzet & Serre, 2011).

Recently, a multitude of studies have reconciled these seemingly inconsistent findings by indicating that recurrent processes might be employed adaptively, depending on the visual input: while feed-forward activity might suffice for simple scenes with isolated objects, more complex scenes or more challenging conditions (e.g. objects that are occluded or degraded), may need additional visual operations ('routines') requiring recurrent computations (Groen, Jahfari, et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Tang et al., 2018). For simple scenes, rapid recognition may thus rely on a coarse and unsegmented feed-forward representation (Crouzet & Serre, 2011), while for more cluttered images recognition may require explicit encoding of spatial relationships between parts. In other words, for those images, extra visual operations to group parts of the object, and to segment this object ('figure') from its background might be needed. Several studies have already shown that the 'segmentability' of a natural scene might influence the degree of recurrent processing. For example, Koivisto et al. (2014) reported that masking, a technique shown to affect mainly recurrent but not feed-forward processing (Fahrenfort et al., 2007), was more effective for objects that were rated as being 'difficult to segregate'. Also in a more recent study we showed that natural scene complexity, providing information about the 'segmentability' of a scene, modulates the degree of feedback activity in the brain (Groen, Jahfari, et al., 2018). However, both studies did not test for effects of segmentation explicitly and used natural scenes that were uncontrolled. Therefore, we here systematically investigated whether scene complexity influenced the extent of recurrent processing during object recognition. To this end, participants performed an object recognition task with objects embedded in backgrounds of different complexity (Figure 3.1), indexed by two biologically plausible measures: the spatial coherence (SC) and contrast energy (CE) (Ghebreab et al., 2009; Groen et al., 2013; Scholte et al., 2009). In half the trials, we impaired feedback activity with visual-masking. In addition to behavioral measures, we measured EEG responses to examine the time-course of visually evoked activity. Besides human participants, we also investigated recognition performance in Deep Convolutional Neural Networks (DCNNs), which received identical visual stimuli as our human participants, and performed a five-choice recognition task.

A convergence of results indicated that recurrent computations were critical for recognition of objects in complex environments, i.e. objects that were more difficult to segment from their background. First of all, behavioral results indicated poorer recognition performance for objects with more complex backgrounds, but only when feedback activity was disrupted by masking. Second, EEG measurements showed clear differences between complexity conditions in the ERPs around 200ms - a time point beyond the first feed-forward visual sweep of activity. Additionally, object category decoding based on the multivariate EEG patterns showed later decoding onsets for objects embedded in more complex backgrounds. This indicated that object representations for more complex backgrounds emerge later, compared to objects in more simple backgrounds. Finally, DCNN performance confirmed this interpretation; feed-forward networks showed a higher degree of impairment in recognition for objects in complex backgrounds compared to recurrent networks. Together, these results support the notion that recurrent computations drive figure-ground segmentation of objects in complex scenes.

Materials and methods

Subjects main experiment

Forty-two participants (32 females, 18-35 years old) took part in a first EEG experiment. Data from two participants were excluded from further analysis due to technical problems. We used this first dataset to perform exploratory analyses and optimize our analysis pipeline (Figure 3.2). To confirm our results on an independent dataset, another twenty participants (13 females, 18-35 years old) were measured. Data from one participant were excluded from ERP analyses, due to wrong placement of electrodes I1 and I2.

Stimuli

Images of real-world scenes containing birds, cats, fire hydrants, frisbees or suitcases were selected from several online databases, including MS COCO (Lin et al., 2014), the SUN database (Xiao et al., 2010), Caltech-256 (Griffin et al., 2007), Open Images V4 (Kuznetsova et al., 2020) and LabelMe (Russell et al., 2008). These five categories were

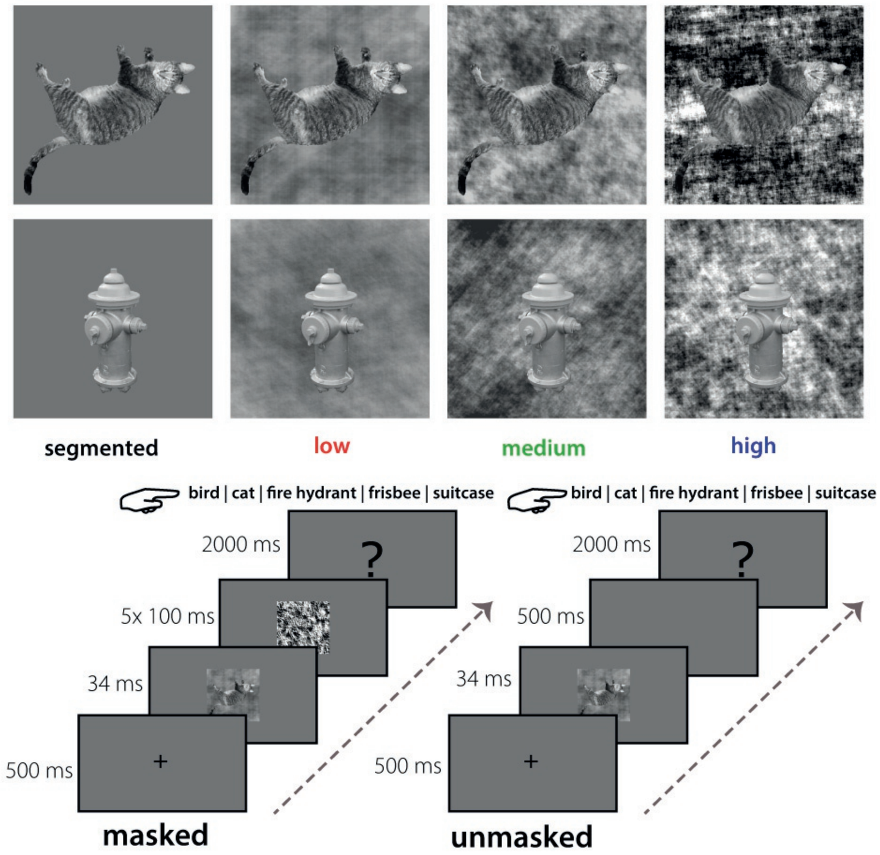
segmented
low
medium
high
bird | cat | fire hydrant | frisbee | suitcase
bird | cat | fire hydrant | frisbee | suitcase
2000 ms
?
5x 100 ms
34 ms
500 ms
+
masked
2000 ms
?
500 ms
34 ms
500 ms
+
unmasked

Figure 3.1: **Stimuli and experimental paradigm.** A) Exemplars of two categories (cat, fire hydrant) from each stimulus complexity condition. Backgrounds were either uniform (segmented; black), or had low (red), medium (green) or high (blue) CE and SC values. B) Experimental design. On masked trials, the stimulus was followed by a dynamic mask (5x100 ms); on unmasked trials this was replaced by a blank screen (500 ms). Participants were asked to categorize the target object by pressing the corresponding button on the keyboard.

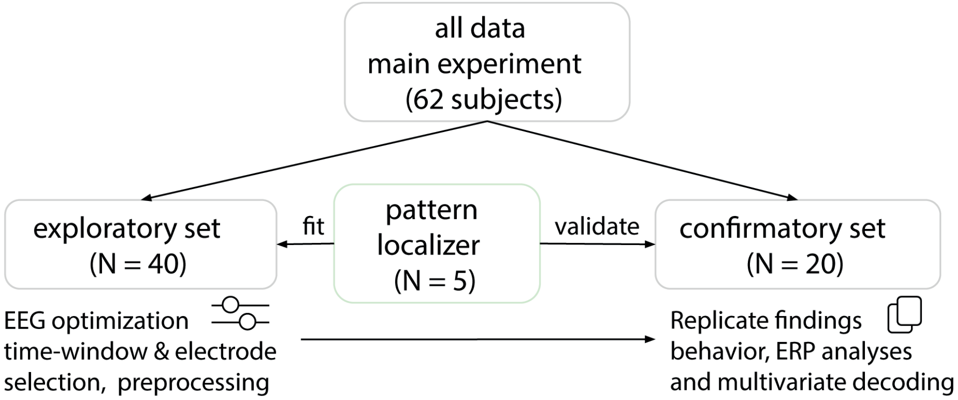

Figure 3.2: **Experimental procedure.** Sixty-two participants took part in the EEG experiment. Data from forty participants were used to perform exploratory analyses. The resulting data (twenty participants) were used to confirm our results. For the decoding analyses, five new participants took part in a separate experiment to characterize multivariate EEG activity patterns for the different object categories.

selected because a large selection of images was available in which the target object was clearly visible and not occluded. For each image, one CE and one SC value was computed using the model described in Ghebreab et al. (2009), Scholte et al. (2009) and Groen et al. (2013). Computing these statistics for a large set of scenes results in a two-dimensional space in which sparse scenes with just a few scene elements separate from complex scenes with a lot of clutter and a high degree of fragmentation. Together, CE and SC appear to provide information about the 'segmentability' of a scene (Groen et al., 2013; Groen, Jahfari, et al., 2018). High CE/SC values correspond with images that contain many edges that are distributed in an uncorrelated manner, resulting in an inherently low figure-ground segmentation. Relatively low CE/SC values on the other hand correspond with a homogenous image containing few edges, resulting in an inherently high figure-ground segmentation (Figure 3.1). Each object was segmented from their real-world scene background and superimposed on three categories of phase scrambled versions of the real-world scenes. This corresponded with low, medium and high complexity scenes. Additionally, the segmented object was also presented on a uniform gray background as the segmented condition (Figure 3.1). For each object category eight low CE/SC, eight medium CE/SC and eight high CE/SC images were selected, using the cut-off values from Groen, Jahfari, et al. (2018), resulting in 24 images for each object category and 120 images in total. Importantly, each object was presented in all conditions, allowing us to attribute the effect to the complexity (i.e. segmentability) of each trial, and rule out any object-specific effects.

Experimental design

Participants performed a 5-choice categorization task (Figure 3.1), differentiating images containing cats, birds, fire hydrants, frisbees and suitcases as accurately as possible. Participants indicated their response using five keyboard buttons corresponding to the different categories. Images were presented in a randomized sequence, for a duration of 34 ms. Stimuli were presented at eye-level, in the center of a 23-inch ASUS TFT-LCD display, with a spatial resolution of 1920*1080 pixels, at a refresh rate of 60 Hz. Participants were seated approximately 70 cm from the screen, such that stimuli subtended a 6.9° visual angle. The object recognition task was programmed in- and performed using Presentation (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). The experiment consisted of 960 trials in total, of which 480 were backward masked trials and 480 were unmasked trials, randomly divided into eight blocks of 120 trials for each participant. After each block, participants took a short break. The beginning of each trial consisted of a 500 ms fixation period where participants focused their gaze on a fixation cross at the centre of the screen. In the unmasked trials, stimuli were followed by a blank screen for 500 ms and then a response screen for 2000 ms. In order to disrupt recurrent processes (Breitmeyer & Ogmen, 2000; Fahrenfort et al., 2007; Lamme et al., 2002), in the masked trials, five randomly chosen phase-scrambled masks were presented sequentially for 500 ms. The first mask was presented immediately after stimulus presentation, each mask was presented for 100 ms; Figure 3.1). The ambient illumination in the room was kept constant across different participants.

Subjects pattern localizer

Five new participants took part in a separate experiment to characterize multivariate EEG activity patterns for the different object categories. For this experiment, we measured EEG activity while participants viewed the original experimental stimuli followed by a word (noun). Participants were asked to only press the button when the image and the noun did not match to ensure attention (responses were not analyzed). A classifier was trained on the EEG data from this experiment, and subsequently tested on the data from the main experiment using a cross-decoding approach. All participants had normal or corrected-to-normal vision, provided written informed consent and received monetary compensation or research credits for their participation. The ethics committee of the University of Amsterdam approved the experiment.

Deep Convolutional Neural Networks (DCNNS)

First, to investigate the effect of recurrent connections, we tested different architectures from the CORnet model family (Kubilius et al., 2018); CORnet-Z (feedforward), CORnet-R (recurrent) and CORnet-S (recurrent with skip connections). Then, to further evaluate the influence of network depth on scene segmentation, tests were conducted on three deep residual networks (He et al., 2016) with increasing number of layers; ResNet-10, ResNet-18 and Resnet-34. “Ultra-deep” residual networks are mathematically equivalent to a recurrent neural network unfolding over time, when the weights between their hidden layers are clamped (Liao & Poggio, 2016). This has led to the hypothesis that the additional layers function in a way that is similar to recurrent processing in the human visual system (Kar et al., 2019). Pre-trained networks were finetuned on images from the MSCoco database (Lin et al., 2014), using PyTorch (Paszke et al., 2019). After initialization of the pretrained network, the model's weights were fine tuned for our task, generating 5 probability outputs (for our 5 object categories). To obtain statistical results, we finetuned the networks ten times for each architecture.

EEG data acquisition and preprocessing

EEG was recorded using a 64-channel Active Two EEG system (Biosemi Instrumentation, Amsterdam, The Netherlands, www.biosemi.com) at a 1024 Hz sample rate. As in previous studies investigating early visual processing (Groen et al., 2013; Groen, Jahfari, et al., 2018), we used caps with an extended 10–10 layout modified with 2 additional occipital electrodes (I1 and I2, which replaced F5 and F6). Eye movements were recorded with additional electro-oculograms (vEOG and hEOG). Preprocessing was done using MNE software in Python (Gramfort et al., 2014) and included the following steps for the ERP analyses: 1) After importing, data were re-referenced to the average of two external electrodes placed on the mastoids. 2) A high-pass (0.1Hz, 0.1Hz transition band) and low-pass (30Hz, 7.5 Hz transition band) basic FIR filters were sequentially applied. 3) an Independent Component Analysis (ICA; Vigario et al. (2000)) was run in order to identify and remove eye-blink and eye-movement related noise components (mean = 1.73 of first 25 components removed per participant). 4) epochs were extracted from -200 ms to 500 ms from stimulus onset. 5) trials were normalized by their 200 ms pre-stimulus baseline. 6) 5% of trials with the most extreme values within each condition were removed, keeping the number of trials within each condition equal. 7) data were transformed to current source density responses (Perrin et al., 1989).

Statistical analysis: behavioral data

Choice accuracy was computed for each condition in the masked and unmasked trials (Figure 3.3). Differences between the conditions were tested using two-factor (Scene complexity: segmented, low, med, high; Masking: masked, unmasked) repeated-measures ANOVAs. Significant main effects were followed up by post-hoc pairwise comparisons between conditions using Sidák multiple comparisons correction at α = 0.05. Behavioral data were analyzed in Python using the following packages: Statsmodels, SciPy, NumPy, Pandas, Matplotlib and Seaborn (Jones et al., 2001; McKinney & Others, 2010; Oliphant, 2006; Seabold & Perktold, 2010).

Statistical analysis: EEG - event related potentials

EEG analyses were carried out in Python, using the MNE software. For each participant, the difference in event-related potential (ERP) to scene complexity was computed within masked and unmasked conditions, pooled across occipital and peri-occipital electrodes (Oz, POz, O1, O2, PO3, PO4, PO7, PO8). This was done by subtracting the signal of each complexity condition (i.e. low, medium or high) from the segmented condition. Doing so enabled us to investigate differences between low, medium and high complex scenes regardless of masking effects. Based on the exploratory dataset, we established five time windows by performing t-tests on every time point for each condition and selecting windows in which the amplitude differed from zero for all complexity conditions (low, med, high). Then, a repeated measures ANOVA with factor Complexity (low, medium, high) and Masking (masked, unmasked) was performed on the average activity in these established time windows.

Statistical analysis: EEG - multivariate classification

The same preprocessing pipeline was used as for the ERP analyses. To evaluate how object category information in our EEG signal evolves over time, cross-decoding analyses were performed by training a Support Vector Machine (SVM) classifier on all trials from the pattern localizer experiment (performed by five different subjects) and testing it on each of the main experiment conditions. Object category classification was performed on a vector of EEG amplitudes across 22 electrodes, including occipital (I1, Iz, I2, O1, Oz, O2), peri-occipital (PO3, PO7, POz, PO4, PO8), and parietal (Pz, P1-P10) electrodes. EEG activity was standardized and averaged across the five time windows derived from the ERP analyses. Statistical significance was determined using a Wilcoxon signed-rank test, and corrected for multiple comparisons using a false discovery rate (FDR) of 0.05.

Data and code availability

Data and code to reproduce the analyses are available at the Open Science Framework (#ru26k) and at https://github.com/noorseijdel/2020_EEG_figureground

Results

Behavior

During the task, participants viewed images of objects placed on top of a gray (segmented), low, medium or high complexity background. On each trial, they indicated which object category the scene contained, using the corresponding keyboard buttons. In half of the trials, the target image was followed by a dynamic backward mask (5x100 ms); the other half of the trials was unmasked (Figure 3.1). Accuracy (percentage correct trials) was computed for each participant. A repeated measures ANOVA on the exploratory dataset (N = 40), with factors background (segmented, low, medium, high) and masking (masked, unmasked) indicated, apart from main effects, an interaction effect. Results indicated that masking impaired performance for objects presented on more complex backgrounds stronger than for less complex backgrounds (F(3,117) = 185.6748, p < .001). Post-hoc comparisons showed that for masked trials, accuracy decreased for both medium (t(39) = 2.88, p(Sidák-corrected) = 0.038) and high (t(39)

= 3.84, p(Sidák-corrected) = 0.003) complexity condition compared to the low condition (all other p > .203). For unmasked trials, all conditions differed from each other, with an incremental decrease in accuracy for objects presented on more complex backgrounds. Analysis of the confirmatory dataset (N = 20) indicated similarly, apart from the main effects, an interaction between masking and background complexity. For masked trials, there was a larger decrease in performance with an increase in background complexity, (F(3, 57) = 101.3338, p < .001). Post-hoc comparisons showed that for masked trials, accuracy decreased for both medium and high complexity conditions compared to the segmented (t(19) = 3.47, p(Sidák-corrected) = 0.003, (t(19) = 3.47, p(Sidák-corrected) = 0.003) and low conditions (t(19) = 4.23, p(Sidák-corrected) < .001, (t(19) = 4.31, p(Sidák-corrected) < .001). For unmasked trials, all conditions differed from each other with the exception of medium - high, with an incremental decrease in accuracy for objects presented on more complex backgrounds.

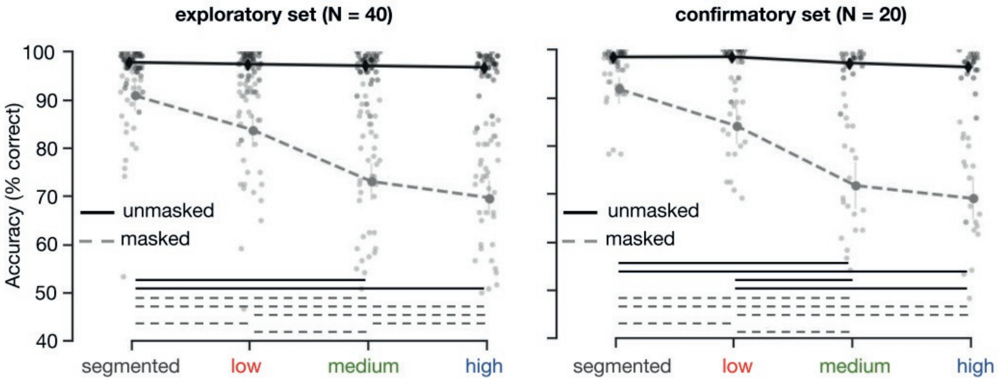

Figure 3.3: **Human performance on the object recognition task.** Performance (percentage correct) on the 5-option object recognition task. For masked trials, performance decreased with an increase in background complexity. The left panel shows results from the exploratory set, on the right results from the confirmatory set are plotted. Error bars represent the bootstrap 95% confidence interval, dots indicate the average performance of individual participants. Significant differences are indicated with a solid (unmasked) or dashed (masked) line.

Network performance

Next, we presented the same images to Deep Convolutional Neural Networks with different architectures. For the CORnets (Figure 3.4, left panel), a non-parametric Friedman test differentiated accuracy across the different conditions (segmented, low, medium, high) for all architectures, Friedman's Q(3) = 27.8400; 24.7576; 26.4687 for CORnet-Z, -RT -S respectively, all p < .001. A Mann-Whitney U test on the difference in performance between segmented and high complexity trials indicated a smaller decrease in performance for CORnet-S compared to CORnet-Z (Mann–Whitney U = 100.0, n1 = n2 = 10, p < .001, two-tailed). For the ResNets (Figure 3.4, right panel), a non-parametric Friedman test differentiated accuracy across the different conditions for ResNet-10 and ResNet-18, Friedman's Q(3) = 23.9053; 22.9468, for ResNet-10 and ResNet-18 respectively, both $p < .001$. A Mann-Whitney U test on the difference in performance between segmented and high complexity trials indicated a smaller decrease in performance for ResNet-34 compared to ResNet-10 (Mann–Whitney U = 100.0, n1 = n2 = 10, $p < .001$, two-tailed). Overall, in line with human performance, results indicated a higher degree of impairment in recognition for objects in complex backgrounds for feed-forward or more shallow networks, compared to recurrent or deeper networks.

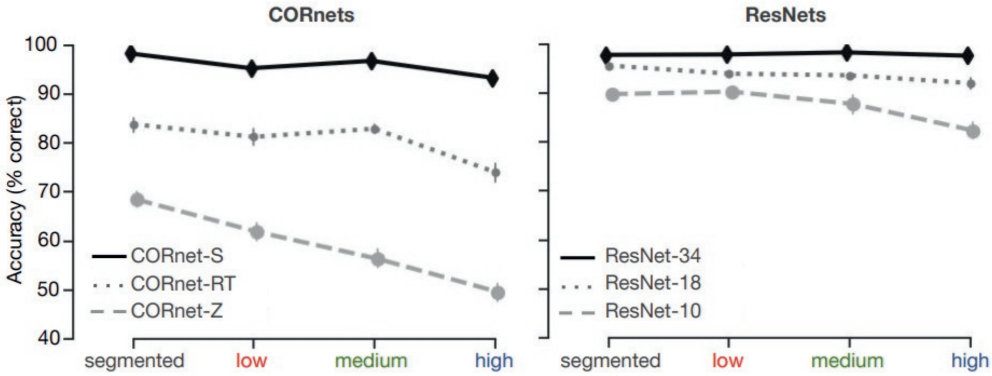

Figure 3.4: **Deep Convolutional Neural Network performance on the object recognition task.** Performance (percentage correct) on the 5-option object recognition task. Networks were finetuned on the 5 target categories, top-1 accuracy was computed. For the CORnets (left panel) performance of the feedforward architecture decreased with an increase in background complexity. For recurrent architectures, this decrease was less prominent. For CORnet-s, there was no difference between conditions. Error bars represent the bootstrap 95% confidence interval.

EEG - event related potentials

To investigate the time-course of figure-ground segmentation in visual cortex, evoked responses to the masked and unmasked scenes were pooled across occipital and peri-occipital electrodes (Oz, POz, O1, O2, PO3, PO4, PO7, PO8), for each condition. Difference waves were generated by subtracting the signal of each condition from the segmented condition (Figure 3.5B/E). Doing so enabled us to eliminate the effect of masking on the EEG signal, and to investigate differences between low, medium and high complex scenes. For each participant, data was averaged across five time windows (based on analyses on the exploratory dataset; see Materials and methods).

For every time window, a Repeated Measures ANOVA was performed on the average EEG amplitude of the difference waves, with Complexity (low, med, high) and Masking (masked, unmasked) as within subject factors. Results on the confirmatory dataset (Figure 3.5D/E/F) showed no main- or interaction effects in the first time win-

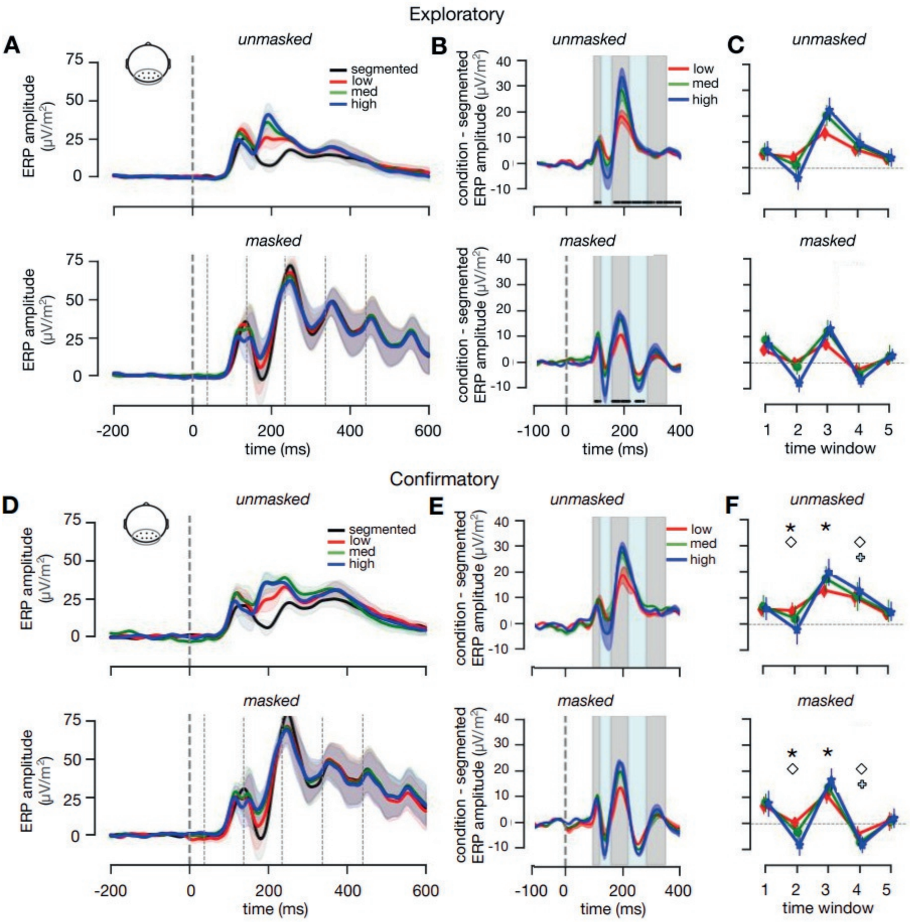

Figure 3.5: **ERP results.** **A)** Average ERP amplitude for segmented, low, medium and high complexity scenes for an occipital-peri-occipital pooling of EEG channels (Oz, POz, O1, O2, PO3, PO4, PO7, PO8) for masked and unmasked trials. Shaded regions indicate SEM across participants. Mask onsets are indicated with thin dashed lines (bottom panel only) **B)** Difference waves were generated by subtracting the signal of each condition from the segmented condition. **C)** Based on significant timepoints in the exploratory dataset, five time windows were defined: 92-115 ms; 120-150 ms; 155-217 ms; 221-275 ms; 279-245 ms) Symbol markers indicate main or interaction effects, asterisk: main effect of condition; diamond: main effect of masking, plus: interaction effect. **D/E/F)** Analyses repeated for the confirmatory dataset.

dow (92-115 ms; Figure 3.5F). Critically, differences between Complexity conditions only emerged in time window 2 and 3 (120-150 ms: F(36) = 22.87, η^{2par} = .56, *p* < .001; 155-217 ms: F(36) = 24.21, η^{2par} = .57, *p* < .001), suggesting a differential con-tribution of recurrent processing to object recognition in varying complexity scenes. In time window 2, there was a main effect of Masking (F(18) = 5.38, η^{2par} = .576, *p* = .03. Only in time window 4 (221-275 ms), an interaction effect of Masking and Complexity, F(18) = 59.60, η^{2par} = .07, *p* < .001 started to emerge.

EEG multivariate classification

To further investigate the representational dynamics of object recognition under different complexity conditions, multivariate decoding analyses were performed on the averaged activity in the five time windows (Figure 3.6. To control for response-related activity (keyboard buttons were fixed across the task), a cross-decoding analysis was performed, by training the classifier on all trials from an independent pattern localizer experiment, and testing it on each of the main experiment conditions (see Methods for details). For unmasked trials, a Wilcoxon signed-rank test on the exploratory dataset indicated successful decoding for segmented trials in all five time windows (Z = 100, *p* < 0.001; Z = 89, *p* < 0.001; Z = 30, *p* < 0.001; Z = 131, *p* < 0.001; Z = 141, *p* < 0.001) and low trials in the first three time windows (92-115 ms; 120-150 ms; 155-217 ms; Z = 198, *p* = 0.007; Z = 82, *p* < 0.001; Z = 61, *p* < 0.001). For objects on medium complex background, successful above-chance decoding emerged slightly later, in time windows 2 and 3 (Z = 200, *p* = 0.012; Z = 110, *p* < 0.001). For objects on high complex background, there was successful decoding in time window 3, Z=216, *p* = 0.045. For masked trials, there was successful decoding for the segmented objects in time windows 1, 3 and 4 , Z = 113, *p* < 0.001; Z = 183, *p* = 0.004; Z = 186, *p* = 0.004, followed by later additional decoding of low (155-217 ms), Z = 138, *p* = 0.001, and high (221-275 ms) complexity trials, Z = 157, *p* = 0.003. There were no significant time windows for medium complexity trials. All p-values reported were corrected by FDR = 0.05. Finally, we aimed to replicate these findings in the confirmatory dataset (N = 20). Overall, results indicated fewer instances of successful object decoding, and if present, slightly delayed compared to the exploratory set. For unmasked trials, results from the Wilcoxon Signed-Ranks test indicated successful decoding for segmented trials in all time windows except the second (92-115 ms; 155-217 ms; 221-275 ms; 279-245 ms), Z = 27, *p* = 0.006; Z = 18, *p* = 0.003; Z = 0, *p* < 0.001; Z = 35, *p* = 0.011. There were no other significant time windows from other unmasked conditions. For masked trials, there was significant decoding for segmented trials in time window 3 and 4 (155-217 ms; 221-275 ms), Z = 36, *p* = 0.031; Z = 38, *p* = 0.031, and for low trials in time window 2, Z = 36 , *p* = 0.050. Overall, these findings showed that dif-ferent objects evoked reliably different sensor patterns when presented in isolation or in 'simple' environments, within the first feed-forward sweep of visual information processing. Additionally, results indicated decreased and later decoding for objects embedded in more complex backgrounds, suggesting that object representations for objects on complex backgrounds emerge later. Finally, these findings also indicate that the object category representations generalized across tasks and participants.

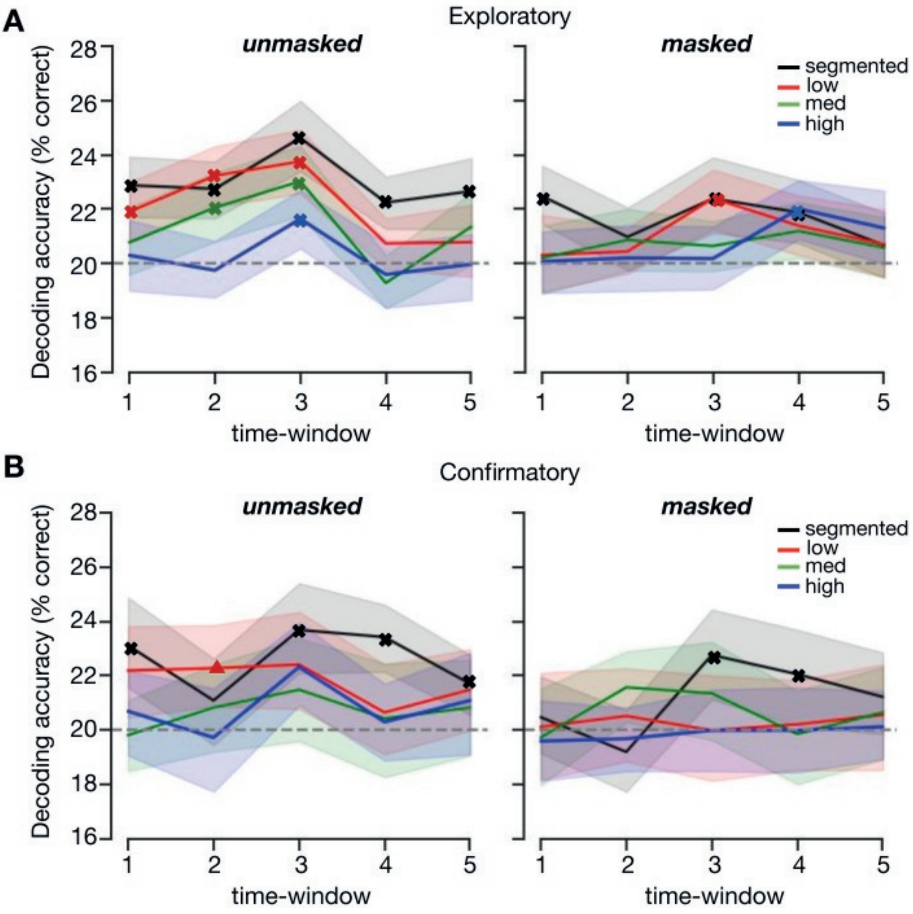

Figure 3.6: **Cross-decoding results using the pattern localizer.** Decoding object category in EEG signal for masked and unmasked trials with varying complexity in the five time windows. Shaded error bars represent the bootstrap 95% confidence interval. Significant results from the Wilcoxon signed-rank test are indicated with a bold x (corrected for multiple comparisons using a false discovery rate of 0.05), a triangle indicates $p = 0.0496$.

Discussion

This study systematically investigated whether recurrent processing is required for figure-ground segmentation during object recognition. A converging set of behavioral, EEG and computational modelling results indicate that recurrent computations are required for figure-ground segmentation of objects in complex scenes.

These findings are consistent with previous findings showing enhanced feedback for complex scenes (Groen, Jahfari, et al., 2018), and visual backward masking being more effective for images that were 'more difficult to segment' (Koivisto et al., 2014). We interpret these results as showing that figure-ground segmentation, driven by recurrent processing, is not necessary for object recognition in simple scenes but it is for more complex scenes.

Effects of scene complexity using artificial backgrounds

In an earlier study, using natural scenes, we already showed that feedback was selectively enhanced for high complexity scenes, during an animal detection task. While there are numerous reasons for using naturalistic scenes (Felsen et al., 2005; Felsen & Dan, 2005; Talebi & Baker, 2012), it is difficult to do controlled experiments with them because they vary in many (unknown) dimensions. Additionally, SC and CE (measures of scene complexity) could correlate with other contextual factors in the scene (e.g. SC correlates with perception of naturalness of a scene (Groen et al., 2013), and could be used as diagnostic information for the detection of an animal. Additionally, previous research has shown that natural scenes and scene structure can facilitate object recognition (Davenport & Potter, 2004; Kaiser & Cichy, 2018; Neider & Zelinsky, 2006). Results from the current study, using artificial backgrounds of varying complexity, replicate earlier findings while allowing us to attribute the effects to SC and CE, and the subsequent effect on segmentability. A limitation of any experiment with artificially generated (or artificially embedded) images is that it's not clear whether our findings will generalize to 'real images' that have not been manipulated in any way. Together with the previous findings, however, our results corroborate the idea that more extensive processing (possibly in the form of recurrent computations) is required for object recognition in more complex, natural environments (Groen, Jahfari, et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Tang et al., 2018).

Time course of object recognition

Based on the data from the exploratory dataset (N = 40), we selected five time windows in the ERPs to test our hypotheses on the confirmatory dataset. For our occipital-peri-occipital pooling, we expected the first feedforward sweep to be unaffected by scene complexity. Indeed, amplitudes of the difference waves, averaged across the selected time windows, indicated no influence of masking or scene complexity early in time (94-110 ms). The observation that all three difference waves deviated from zero, however, indicates that there was an effect of segmentation. In this early time window, background presence thus seems to be more important than the complexity of the background. This difference could be attributed to the detection of additional low-level features in the low, medium and high complexity condition, activating a larger set of neurons that participate in the first feedforward sweep (Lamme & Roelfsema, 2000). In the second and third time window (120-217 ms), differences between the complexity conditions emerge. We interpret these differences as reflecting the increasing need for recurrent processes. Our results are generally consistent with prior work investigating the time course of visual processing of objects under more or less challenging condi-

tions (Cichy et al., 2014; Contini et al., 2017; DiCarlo & Cox, 2007; Rajaei et al., 2019; Tang et al., 2018). In line with multiple earlier studies, masking left the early evoked neural activity (<120 ms) relatively intact, whereas the neural activity after ~150 ms was decreased (Boehler et al., 2008; Del Cul et al., 2007; Fahrenfort et al., 2007; Koivisto & Revonsuo, 2010; Lamme et al., 2002; Lamme & Roelfsema, 2000).
Decoding results corroborated these findings, showing decreased or delayed decoding onsets for objects embedded in more complex backgrounds, suggesting that object representations for those images emerge later. Additionally, when recurrent processing was impaired using backward masking, only objects presented in isolation or in 'simple' environments evoked reliably different sensor patterns that our classifiers were able to pick up (Figure 3.6.

Influence of backward masking

Based on the strong interaction effect on behavior, it is tempting to conclude that complexity significantly increases the effect of masking on recognition accuracy. However, performance on all unmasked trials was virtually perfect (96-97%) raising concerns about ceiling effects obscuring the actual variation between these conditions (Uttl, 2005). Therefore, although masked stimuli show a decrease in performance along increases in complexity; base on the current results we cannot conclude that this is because of masking (i.e. reducing recurrent processes). We do not claim that unmasked segmented, low, med, or high images are equally difficult or processed in the same way (we actually argue for the opposite), but apparently the brain is capable of arriving at the correct answer with enough time. It is hard to come up with an alternative (more difficult) task without affecting our experimental design and subsequent visual processing (e.g. stimulus degradation generally affects low-level complexity; reducing object size or varying object location creates a visual search task that could benefit from spatial layout properties). Combined fMRI and EEG results from an earlier study already showed that for complex scenes only, early visual areas were selectively engaged by means of a feedback signal (Groen, Jahfari, et al., 2018). Here, using controlled stimuli and backward masking, we replicate and expand on these findings. Importantly, results from both EEG and deep convolutional neural networks support the notion that recurrent computations drive figure-ground segmentation of objects in complex scenes.

Consistency of object decoding results

In the exploratory set, results from the multivariate decoding analyses indicated early above chance decoding for 'simple' scenes (segmented and low) in both unmasked and masked trials. For more complex scenes decoding emerged later (medium) or was absent (high) for unmasked trials. In the confirmatory set, however, there were fewer instances of successful object decoding, and if present, successful decoding was delayed. A potential explanation for this finding could be that the sample size in the confirmatory dataset was inadequate for the chosen multivariate decoding analyses, resulting in reduced statistical power. A simulation analysis on the exploratory set, in which we randomly selected 20 participants (repeated 1000 times) indicated reduced decoding accuracy, similar to our confirmatory results. Our choice for the number of participants in the confirmatory dataset thus does not seem to be sufficient (Supplementary Figure 3.7).

Probing cognition with Deep Convolutional Neural Networks

One way to understand how the human visual system processes visual information involves building computational models that account for human-level performance under different conditions. Here we used Deep Convolutional Neural Networks, because they show remarkable performance on both object and scene recognition (e.g. Russakovsky et al. (2015); He et al. (2016)). While we do not aim to claim that DCNNs are identical to the human brain, we argue that studying how performance of different architectures compares to human behavior could be informative about the type of computations that are underlying this behavior. In the current study, it provides an additional test for the involvement of recurrent connections. Comparing the (behavioral) results of DCNNs with findings in humans, our study adds to a growing realization that more extensive processing, in the form of recurrent computations, is required for object recognition in more complex, natural environments (Groen, Jahfari, et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Tang et al., 2018).

Conclusion

Results from the current study show that how object recognition is resolved depends on the context in which the target object appears: for objects presented in isolation or in 'simple' environments, object recognition appears to be dependent on the object itself, resulting in a problem that can likely be solved within the first feedforward sweep of visual information processing. When the environment is more complex, recurrent processing is necessary to group the elements that belong to the object and segregate them from the background.

Supplement to Chapter 3

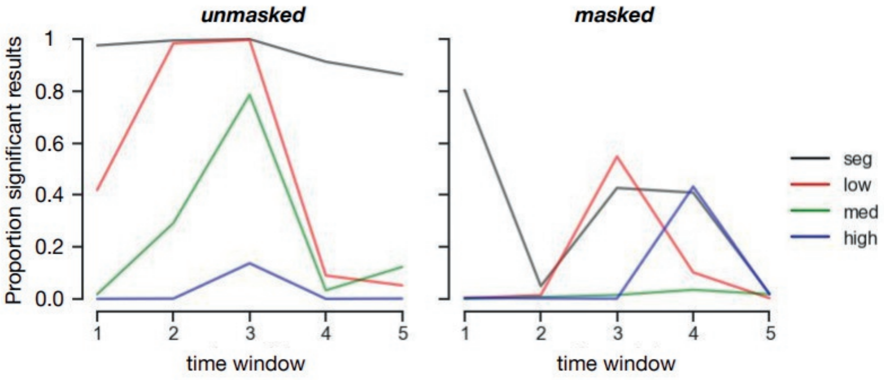

Figure 3.7: **Simulation analysis on the exploratory set.** Random selection of 20 participants (repeated 1000 times) indicated reduced chances of finding significant decoding results. Plotted are the proportion (number of instances divided by 1000) in which the results remained significant.

Chapter 4

Depth in convolutional neural networks solves scene segmentation

Abstract Feedforward deep convolutional neural networks (DCNNs) are, under specific conditions, matching and even surpassing human performance in object recognition in natural scenes. This performance suggests that the analysis of a loose collection of image features could support the recognition of natural object categories, without dedicated systems to solve specific visual subtasks. Research in humans however suggests that while feedforward activity may suffice for sparse scenes with isolated objects, additional visual operations ('routines') that aid the recognition process (e.g. segmentation or grouping) are needed for more complex scenes. Linking human visual processing to performance of DCNNs with increasing depth, we here explored if, how, and when object information is differentiated from the backgrounds they appear on. To this end, we controlled the information in both objects and backgrounds, as well as the relationship between them by adding noise, manipulating background congruence and systematically occluding parts of the image. Results indicate that with an increase in network depth, there is an increase in the distinction between object- and background information. For more shallow networks, results indicated a benefit of training on segmented objects. Overall, these results indicate that, de facto, scene segmentation can be performed by a network of sufficient depth. We conclude that the human brain could perform scene segmentation in the context of object identification without an explicit mechanism, by selecting or "binding" features that belong to the object and ignoring other features, in a manner similar to a very deep convolutional neural network.

This chapter is published as: Seijdel, N., Tsakmakidis, N., de Haan, E.H.F., Bohte, S.M., & Scholte, H.S. (2020). Depth in convolutional neural networks solves scene segmentation. *PLoS Computational Biology, 16*(7), e1008022. doi:10.1371/journal.pcbi.1008022.

Author summary

To what extent do Deep Convolutional Neural Networks exhibit sensitivity to scene properties (e.g. object context) during object recognition, and how is this related to network depth? Linking human visual processing to performance of feed-forward DCNNs with increasing depth, our study explored if and how object information is differentiated from the backgrounds they appear on. We show that with an increase in network depth, there is a stronger selection of parts of the image that belong to the target object, compared to the rest of the image. In other words, network depth facilitates scene segmentation. Given that the operations of a very deep network can be performed by a recurrent network, we speculate that the human brain could perform scene segmentation, in the context of object identification, without an explicit mechanism using recurrent processing.

Introduction

Visual object recognition is so swift and efficient that it has been suggested that a fast feed-forward sweep of perceptual activity is sufficient to perform the task (DiCarlo & Cox, 2007; Serre et al., 2007; VanRullen & Thorpe, 2002). Disruption of visual processing beyond feed-forward stages (e.g. >150 ms after stimulus onset, or after activation of higher order areas) can however lead to decreased object recognition performance (Camprodon et al., 2013; Koivisto et al., 2011), and a multitude of recent findings suggest that while feed-forward activity may suffice to recognize isolated objects that are easy to discern, the brain employs increasing feedback or recurrent processing for object recognition under more ‘challenging’ natural conditions (Groen, Jahfari, et al., 2018; Herzog & Clarke, 2014; Kar et al., 2019; Rajaei et al., 2019). When performing a visual object recognition task, the visual input (stimulus) elicits a feed-forward drive that rapidly extracts basic image features through feedforward connections (Lamme & Roelfsema, 2000). For sparse scenes with isolated objects, this set of features appears to be enough for successful recognition. For more complex scenes, however, the jumble of visual information (‘clutter’) may be so great that object recognition cannot rely on having access to a conclusive set of features. For those images, extra visual operations (‘visual routines’), such as scene segmentation and perceptual grouping, requiring several iterations of modulations and refinement of the feedforward activity in the same and higher visual areas, might be necessary (Hochstein & Ahissar, 2002; Howe, 2017; Lamme et al., 2002; Wyatte et al., 2014). While this view emphasizes that object recognition relies on the integration of features that belong to the object, many studies have shown that features from the background can also influence the recognition process (Bar & Ullman, 1996; Davenport, 2007; Davenport & Potter, 2004; Greene et al., 2015; Joubert et al., 2008; Rémy et al., 2013; Sun et al., 2011). For example, objects appearing in a familiar context are detected more accurately and quickly than objects in an unfamiliar environment, and many computational models of object recognition (in both human and computer vision), use features both from within the object and from the background (Fink & Perona, 2004; Riesenhuber & Poggio, 1999; Torralba et al., 2006). This shows that when subjects recognise an object, figure-ground segmentation has not always occurred completely. One way to understand how the human visual system processes information involves building computational models that account for human-level performance under different conditions. Here we investigate Deep Convolutional Neural Networks (DCNNs). DCNNs are being studied often because they show remarkable performance on both object and scene recognition, rivaling human performance. Recent evidence shows that the depth of DCNNs is of crucial importance for this recognition performance (Russakovsky et al., 2015). In addition to better performance, deeper networks have also been shown to be more human-like (making errors similar to human subjects; Kheradpisheh et al. (2016)). More layers seem especially important when scenes are more difficult or challenging, e.g. because of occlusion, variation, or blurring, where elaborate processing is required (Kar et al., 2019; Rajaei et al., 2019). The very deep residual networks used in current object recognition tasks are nearly equivalent to a recurrent neural network unfolding over time, when the weights between their hidden layers are clamped (Liao & Poggio, 2016). This has led to the hypothesis that the additional layers function in a way that is similar to recurrent processing in the human visual system, and that these additional layers are solving the challenges that are resolved by recurrent computations in the brain. In the current study, we explore how the number of layers (depth) in a DCNN relates to human vision and how depth influences to what degree object segmentation occurs. While we certainly do not aim to claim that DCNNs are identical to the human brain, we argue that they can be studied in a way similar to the way in which we use animal models (DNimals; Scholte (2018)]). First, we focused on the question to what extent DCNNs exhibit the same sensitivity to scene properties (object context) as human participants. To this end, we presented seven Residual Networks (ResNets; He et al. (2016)) with an increasing number of layers and 40 human participants with images of objects that were either presented on a uniform background (segmented), or on top of congruent or incongruent scenes and evaluated their performance. Additionally, for the DCNNs, we controlled the amount of information in the objects and backgrounds, as well as the relationship between them by adding noise or systematically occluding parts of the image. Next, we investigated the role of segmentation on learning ('training'), by training the DCNNs on either segmented or unsegmented objects. A convergence of results indicated a lower degree of segregation between object- and background features in more shallow networks, compared to deeper networks. This was confirmed by the observation that more shallow networks benefit more from training on pre-segmented objects than deeper networks. Overall, deeper networks seem to perform implicit 'segmentation' of the objects from their background, by improved selection of relevant features.

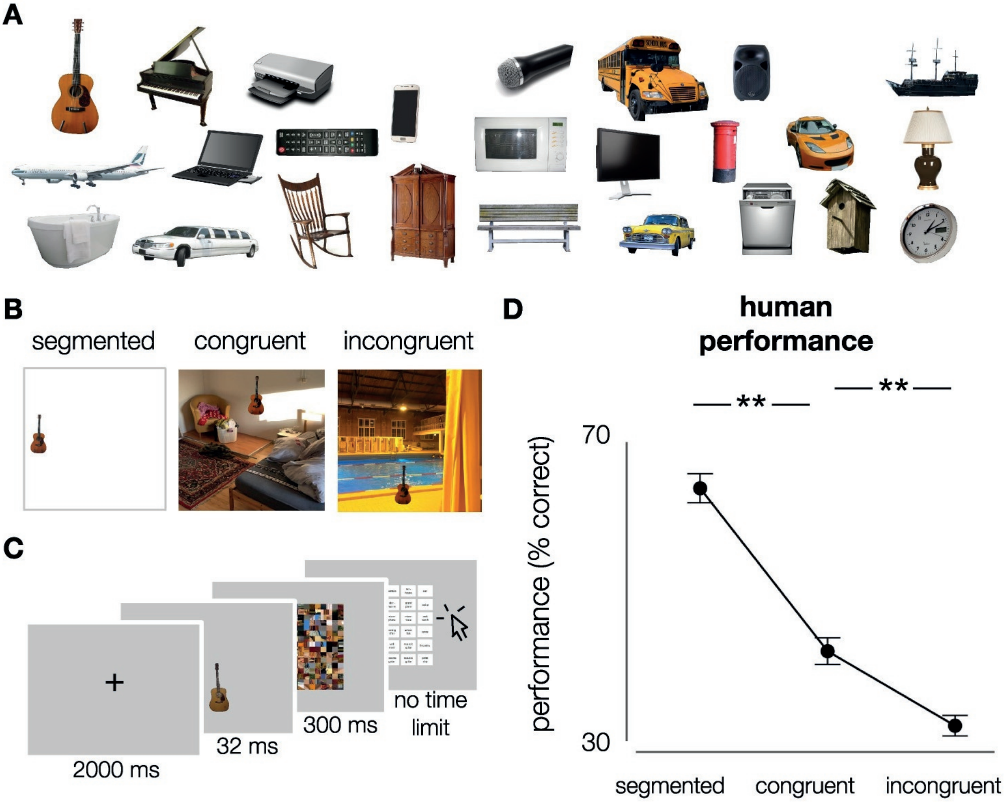

Figure 4.1: **Stimuli and experimental design. A)** Exemplars of the different object categories (cut-out objects from ImageNet validation set). 27 object categories were used in this experiment (subordinate level, based on ImageNet categories). In total, each category contained 10 exemplars. **B)** Stimuli were generated by placing the objects onto white, congruent and incongruent backgrounds (512*512 pixels, full-color). Backgrounds were sampled from the SUN2012 database (Xiao et al., 2010). For human participants, objects were downsized and placed in one of nine possible locations (3*3 grid). For DCNNs, objects were bigger and placed centrally. **C)** Participants performed on an object recognition task. At the beginning of each trial, a fixation-cross was presented in the center of the screen for 2000 ms, followed by an image. Images were presented in randomized sequence, for a duration of 32 ms, followed by a mask, presented for 300 ms. After the mask, participants had to indicate which object they saw, by clicking on one of 27 options on screen using the mouse. After 81 $(1/3)$ and 162 $(2/3)$ trials, there was a short break. **D)** Human performance (% correct) on the object recognition task. Participants performed best for segmented objects, followed by congruent and incongruent respectively. Error bars represent bootstrap 95% confidence intervals.

Results

Experiment 1

Human performance

In experiment 1, participants viewed images of real-world objects placed onto white (segmented), congruent and incongruent backgrounds (Figure 4.1A/B). Images were presented in randomized sequence, for a duration of 32 ms, followed by a mask, presented for 300 ms. After the mask, participants indicated which target object was presented, by clicking on one of 27 options on screen using the mouse (Figure 4.1C; see Materials and methods). Accuracy (percentage correct) was computed for each participant. A non-parametric Friedman test differentiated accuracy across the three conditions (segmented, congruent, incongruent), Friedman's $Q(2) = 74.053$, $p < .001$. Post-hoc analyses with Wilcoxon signed-rank tests indicated that participants made fewer errors for segmented objects, than the congruent, $W = 741$, $p <.001$, and incongruent condition, $W = 741$, $p < .001$ (Figure 4.1D). Additionally, participants made fewer errors for congruent than incongruent trials, $W = 729$, $p < .001$. Overall, results indicate that when a scene is glanced briefly (32 ms, followed by a mask), the objects are not completely segregated from their background and semantic consistency information influences object perception.

Model performance

For human participants, results indicated that (at a first glance) features from the background influenced object perception. Do DCNNs show a similar pattern and how is this influenced by network depth? To investigate the effect of network depth on scene segmentation, tests were conducted on seven deep residual networks (ResNets) with increasing number of layers (6, 10, 18, 34, 50, 101, 152). This approach allowed us to investigate the effect of network depth (adding layers) while keeping other model properties as similar as possible.

We presented 38 different subsets of 243 stimuli to the DNNs, each subset consisting of the same number of images per category and condition that human observers were exposed to (81 per condition, 3 per category). Following the procedure for comparing human performance, a non-parametric Friedman test differentiated accuracy across the three conditions (segmented, congruent, incongruent) for all networks. Using Post Hoc Wilcoxon signed-rank tests with Benjamini/Hochberg FDR correction, differences between the conditions were evaluated for all networks (Figure 4.2; significant differences indicated with a solid line). Results indicated both a substantial overlap and difference in performance between human participants and DCNNs (Figure 4.2). Both were better in recognizing an object on a congruent versus an incongruent background. However, whereas human participants performed best in the segmented condition, DCNNs performed equally well (or better) for the congruent condition. Performance for the incongruent condition was lowest. This effect was particularly strong for more shallow networks (ResNet-6, ResNet-10), and got smaller as the networks got deeper. A Mann-Whitney U test on the difference in performance between congruent and incongruent trials indicated a smaller decrease in performance for incongruent trials for ResNet152 compared to ResNet6 (Mann–Whitney U = 1420.0, $n1 = n2 = 38$, p < .001, two-tailed) For 'ultra-deep' networks it mattered less if the background was congruent, incongruent or even present, behavior that humans also exhibit when these images are shown unmasked. Remarkably, performance of the most shallow network (ResNet-6) was better for the congruent condition compared to the segmented condition. These results suggest that parts of co-varying backgrounds or surroundings influence the categorization of the objects. In other words, there is 'leakage' of the natural (congruent) background in the features for classification, predominantly for the more shallow networks. For object recognition in a congruent scene this is not necessarily a problem, and can even increase or facilitate performance (as compared to the segmented condition). For objects on an incongruent background, however, this impairs classification performance. These results suggest that one of the ways in which network depth improves object classification, is by learning how to select the features that belong to the object, and thereby implicitly segregating the object features from the other parts of the scene.

Then, to determine whether the experimental observations above can be approximated by recurrent computations, we additionally tested three different architectures from the CORnet model family (Kubilius et al., 2018); CORnet-Z (feedforward), CORnet-R (recurrent) and CORnet-S (recurrent with skip connections). The shift in performance from CORnet-Z to CORnet-S showed the same pattern as the shift from ResNet-6 to ResNet-18. This overlap suggests that the pattern of results for deeper ResNets can be approximated by recurrent computations. Because the different CORnet models did not only differ with respect to 'recurrence', but also contained other architectural differences (CORnet-Z not only is feedforward, but it is also shallower than CORnet-S), the differences between the networks could stem from the difference in information flow (feedforward vs. recurrent), or from the different amount of parameters in each network. Taking the results from the ResNets and CORnets together, these findings suggest that one of the ways in which network depth improves object classification, is by learning how to select the features that belong to the object, and thereby implicitly segregating the object features from the other parts of the scene. To confirm this hypothesis, and to gain more insight into the importance of the features in the object vs. the background, Gaussian noise was added to either the object, the background, or both (Figure 4.2B). When noise was added to the complete image (object included), performance decreased for all conditions and all networks. When noise was added to the object only, classification performance also decreased for all conditions Crucially, this decrease was modest for the congruent and particularly severe for the incongruent condition. This indicates that for the congruent condition, also in the no-noise manipulation, performance is heavily dependent on the background for classification. The other side of this conclusion, that in the incongruent condition the features in the background interfere with object classification, is confirmed by the observation that this condition improves when noise is added to the background.

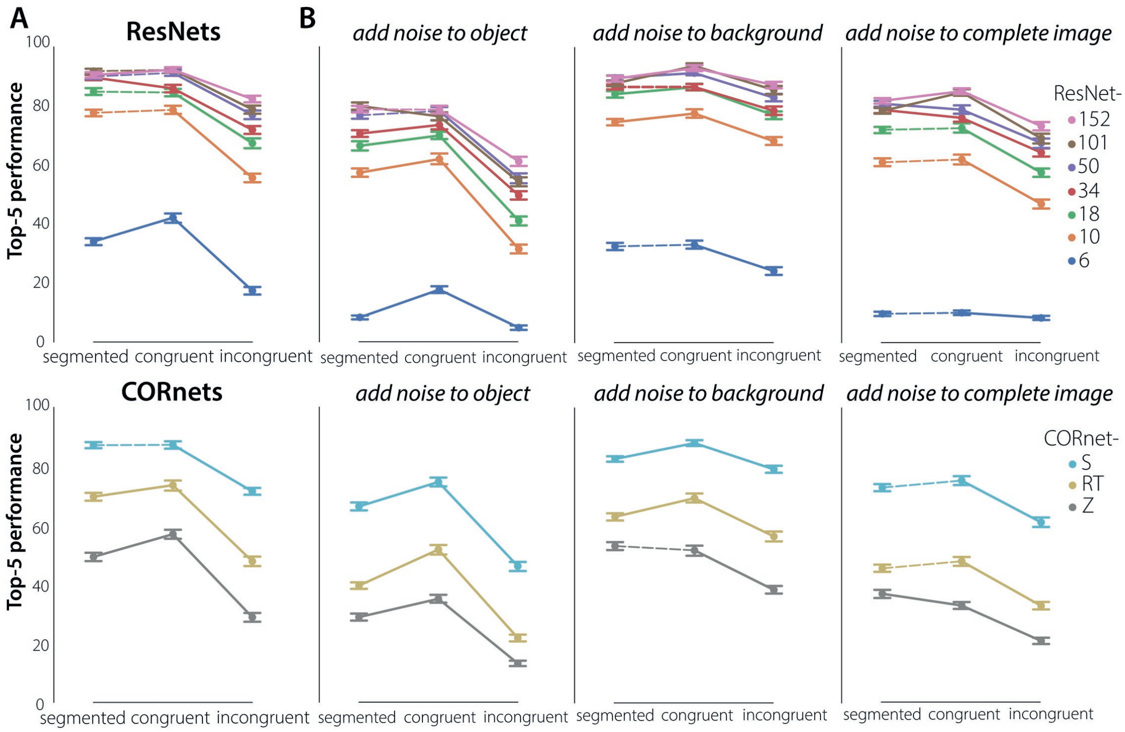

Figure 4.2: **DCNN performance on the object recognition task. A)** DCNN performance on the object recognition task. 38 different subsets of 243 stimuli were presented, each subset consisting of the same number of images per target category and condition (segmented, congruent, incongruent) that human observers were exposed to (81 per condition, 3 per category). For all models, performance was better for the congruent than for the incongruent condition. For the ResNets, this decrease was most prominent for ResNet-6, and got smaller as the models get deeper. For 'ultra-deep' networks it mattered less if the background was congruent, incongruent or even present. For the CORnets, this decrease was most prominent for the feedforward architecture (CORnet-Z). For CORnet-S (recurrent + skip connection) performance was similar to an 'ultra-deep' network. Using Post Hoc Wilcoxon signed-rank tests with Benjamini/Hochberg FDR correction, differences between the conditions were evaluated for all networks. Significant differences are indicated with a solid line, error bars represent bootstrap 95% confidence interval. **B)** DCNN performance on the object recognition task after adding noise to the object, the background, or both.

To further investigate the degree to which the networks are using features from the object and/or background for classification, we systematically occluded different parts of the input image by sliding a gray patch (of either 64*64, 128*128 or 256*256 pixels) across the image in 32 pixel steps. We evaluated the changes in activation of the correct class after occlusion of the different image parts, before the softmax activation function (compared to activation for the 'original' unoccluded image). We reasoned that, if the activity in the feature map changed after occluding a patch of the image, that those pixels were important for classification. For this analysis, positive values indicate that pixels are helping classification, with higher values indicating a higher importance. This reveals the features to be far from random, uninterpretable patterns. For example, in Figure 4.3, results clearly show that the network is localizing the object within the scene, as the activity in the feature map drops significantly when the object (china cabinet in this example) is occluded. To evaluate whether deeper networks are better at localizing the objects in the scene, while ignoring irrelevant background information, we quantified the importance of features in the object vs. background by averaging the change in the feature map across pixels belonging to either the object or the background ('importance'). For each image, importance values of the objects and backgrounds were normalized by dividing them by the activation for the original image. Because performance of ResNet-6 for the 'original' unoccluded images was already exceptionally low, the averaged interference was hard to interpret and remained low, due to many near-zero values in the data. Therefore, we took into account only images that were classified correctly (correct class within Top 5 predictions), resulting in an unequal number of images for each network. Mann-Whitney U tests with Benjamin/Hochberg FDR correction indicated a smaller influence (importance) of background pixels on classification for deeper networks. For those models, pixels from the object had a smaller impact as well, for the segmented and congruent condition.

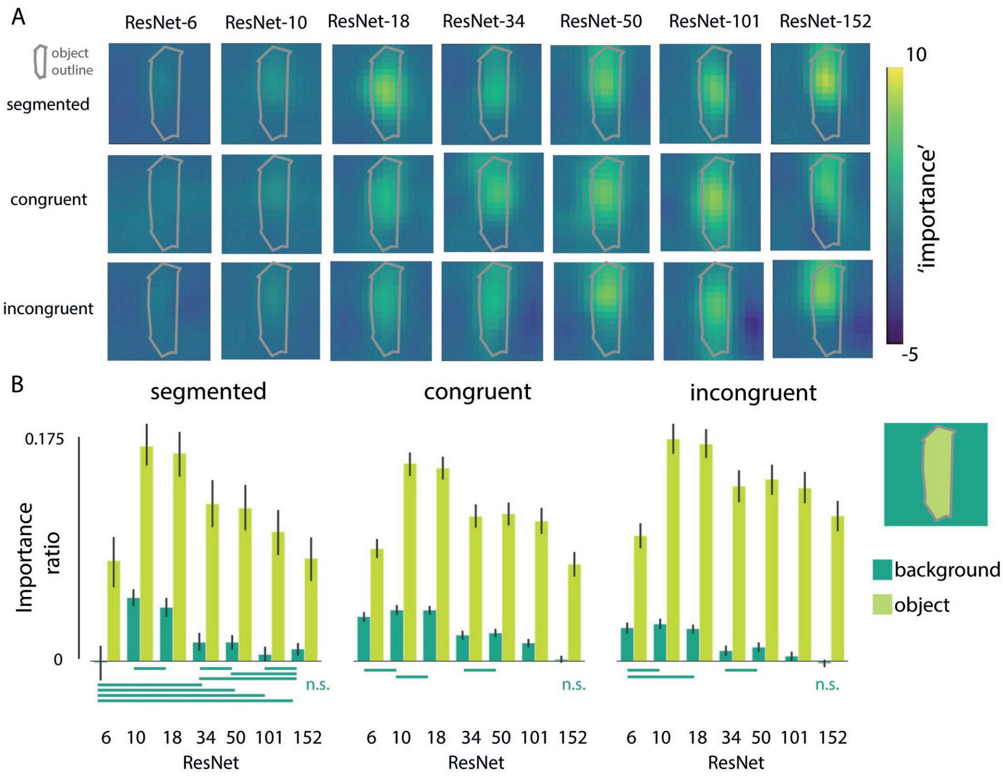

Figure 4.3: **Systematic occlusion of parts of the image. A)** Examples where we occluded different portions of the scene, and visualized how the classifier output for the correct class changed (before the softmax activation function). Images were occluded by a gray patch of 128x128 pixels, sliding across the image in 32 pixel steps. Importance is defined as the relative change in activation after occluding that part of the image (compared to the activation of the 'original' unoccluded image) and is computed as follows: original activation - activation after occlusion / original activation. This example is for illustrative purposes only; maps vary across exemplars. **B)** The relative change in activation (compared to the original image), after occluding pixels of either the object or the background, for the different conditions (segmented, congruent, incongruent). For each image, importance values of the objects and backgrounds were normalized by dividing them by the activation for the original image, resulting in the importance ratio. Error bars represent the bootstrap 95% confidence interval. Non-significant differences are indicated with a solid line below the graph.

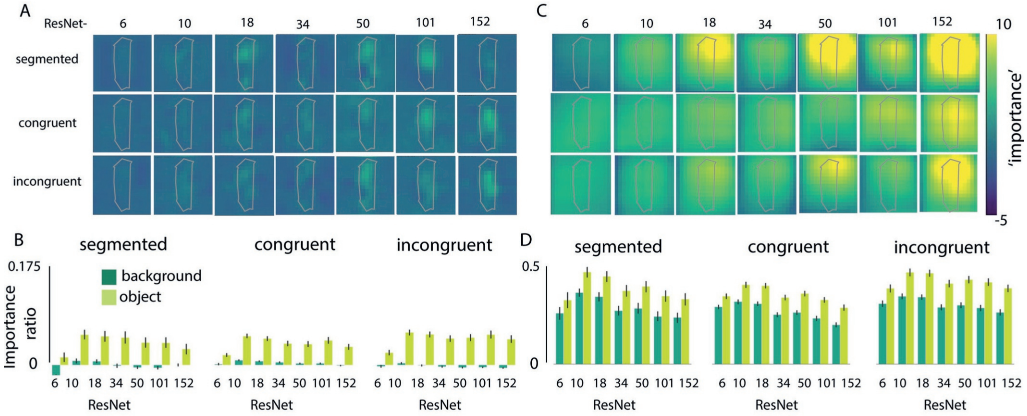

Figure 4.4: **Analysis repeated with a smaller (64x64) and larger (256x256) patch. A)** visualization of the change in classifier output for the correct class, before the softmax activation function after occlusion by a 64x64 patch, sliding across the image in 32 pixel steps. **B)** The relative change in activation (compared to the original image), after occluding pixels of either the object or the background, for the different conditions (segmented, congruent, incongruent). For each image, importance values of the objects and backgrounds were normalized by dividing them by the activation for the original image, resulting in the importance ratio. Error bars represent the 95% confidence interval. **C/D)** Repeated for a large patch (256x256 pixels).

Next, we tested how training was influenced by network depth. If deeper networks indeed implicitly learn to segment object from background, we expect them to show a smaller difference in learning speed, when trained with segmented vs. unsegmented stimuli (as compared to shallow networks).

Experiment 2

Experiment 1 indicated that, when trained on ImageNet, the networks are influenced by visual information from both the object and the background region. In experiment 2, we investigated the influence of background on classification performance when the networks are trained on visual information from the object region only. To do so, we trained four networks (ResNet-6, ResNet-10, ResNet-18, ResNet-34) on a dataset with objects that were already segmented, and on a dataset in which they were unsegmented (i.e. objects embedded in the scene). All images were resized to 128x128 pixels. We used more shallow networks and fewer object classes to reduce computation time. To obtain statistical results, we reinitialized the networks with different seeds and repeated the process for 10 different seeds.

Accuracy of the ResNets was evaluated after each epoch (100 in total) on the validation sets. Results indicated a higher classification accuracy in the early stages of training for the networks trained on segmented objects compared to the networks trained on unsegmented objects (Figure 4.5). Statistical analyses comparing the average accuracies of the first 10 epochs for networks trained on segmented vs. unsegmented objects indicated significant differences for all models (Mann-Whitney U-statistic: U=5.0, p < .001., for ResNet 6, 10, 18 and 34 respectively). In the later stages, accuracy between the two types of models (trained on unsegmented vs. segmented) was similar. Results also indicated a difference between the more shallow networks (ResNet-6), where there is a difference in accuracy between segmented and unsegmented objects for all training epochs, and the deeper networks. For the deeper networks, the difference in accuracy quickly diminishes and finally disappears. Shallow networks trained on segmented stimuli also converged (stabilized) earlier than when they were trained on unsegmented images. Statistical analyses comparing the 'speed of convergence' indicated significant effects of visual training diet (segmented vs. unsegmented) across multiple initialization conditions of the networks, for the more shallow networks (Mann Whitney-U statistic U = 0, p < .001; U = 20.0, p = .012 for ResNet-6 and ResNet-10, respectively). For this analysis, the speed of convergence was defined as the first epoch at which 95% of the maximum accuracy was reached. Deeper networks thus seem to learn to 'segment' the objects from their background during training.

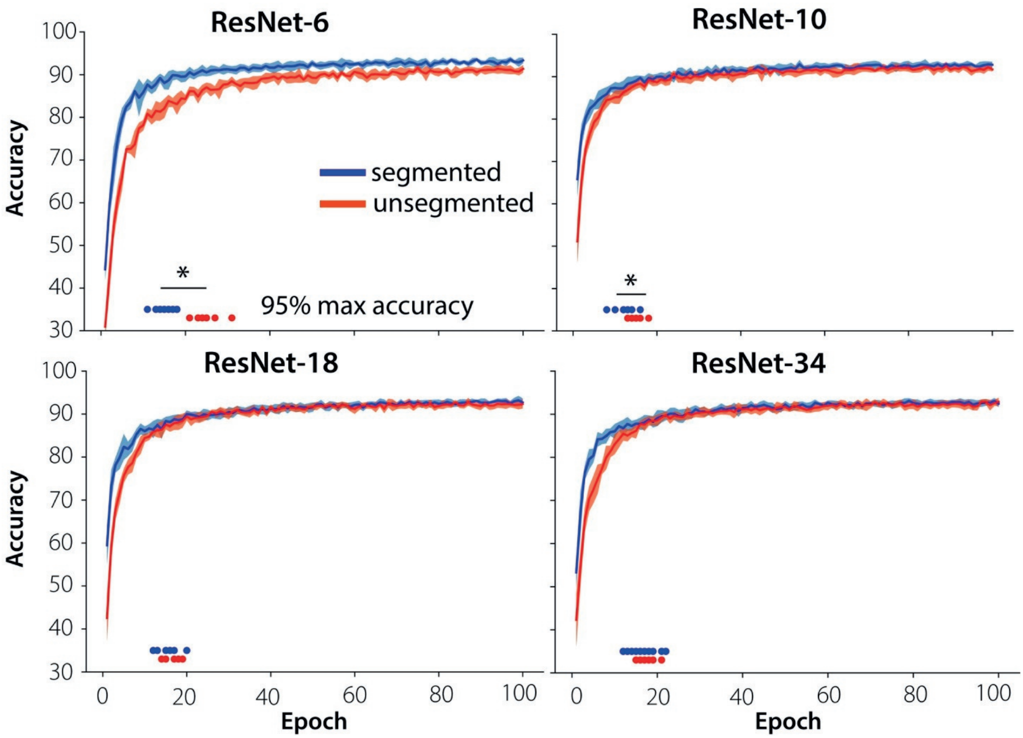

Figure 4.5: **Accuracy during training on segmented vs. unsegmented stimuli.** Networks trained on segmented objects achieve better classification accuracy in the early stages of training than the networks trained on unsegmented objects for shallow networks (ResNet6, ResNet10), and they converge in less epochs. Individual data points indicate the moment of convergence, defined as the first epoch to reach 95% of the maximum accuracy across all epochs.

To better understand the inner workings of our models, we visualized the filter activations of each convolution layer. Visualizing the filter activations of each convolution layer of the networks provides us with heatmaps that show features of a given image, that a corresponding filter is tuned to. This gives an idea of which parts of the image contained the most important features for classification. To obtain these heatmaps, we extracted all the filter activations from the different layers (one 2D-array per filter) for a specific image. Then, for each layer, we summed the absolute value of those arrays together.

Looking at the heatmaps of networks trained on segmented vs. unsegmented data (see Figure 4.6), we see that the heatmaps of the networks trained on segmented ob-jects contain no background activations. For networks trained on unsegmented objects (full images), however, we see that the backgrounds are gradually suppressed inside the network. This indicates that the networks learn to attend to important features (i.e. the objects) and almost eliminate completely the influence of the background, when the depth or capacity of the network is sufficient. This suggests that the network learns to segment the objects before classifying.

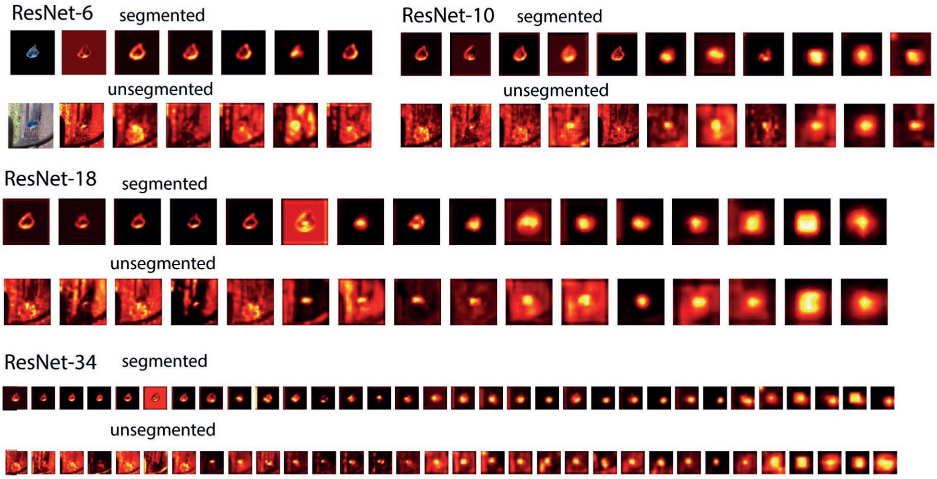

Figure 4.6: **Visualization of the filter activations of each convolution layer for the different networks.** All the filter activations from the different layers (one 2D-array per filter) for a specific image were extracted. heatmaps were generated by summing the absolute value of those arrays together. The lightest part of these heatmaps contain the most important features for classification. Maps for ResNet-34 were resized for visualization purposes.

Discussion

We investigated the extent to which object and context information is represented and used for object recognition in trained Deep Convolutional Neural Networks (DCNNs). Experiment 1 showed both a substantial overlap, and a difference in performance between human participants and DCNNs. Both humans and DCNNs are better in recognizing an object on a congruent versus an incongruent background. However, whereas human participants performed best in the segmented condition, DCNNs performed equally well (or better) for the congruent condition. Performance for the incongruent condition was lowest. This effect was particularly strong for more shallow networks. Further analyses, investigating which parts of the image were most important for recognition, showed that the influence of the background features on the response outcome was relatively strong for shallow networks and almost absent for deeper networks. For shallow networks, the results of experiment 2 indicated a benefit of training on segmented objects (as compared to unsegmented objects). For deeper networks, this benefit was much less prominent. Training on segmented images thus reduced the difference in performance between shallow and deeper networks.

The current results suggest that there is no discrete 'moment' at which segmentation is successful or 'done'. We interpret these findings as indicating that with an increase in network depth there is better selection of the features that belong to the output category (vs. the background), resulting in higher performance during recognition. Thus, more layers are associated with 'more' or better segmentation, by virtue of increasing selectivity for relevant constellations of features. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and might be one of the ways in which scene segmentation is performed in the brain using recurrent computations.

Explicit vs. implicit models of grouping and segmentation

Classic models focussing on grouping and segmentation presume an explicit process in which certain elements of an image are grouped, whilst others are segregated from each other, by a labelling process (Neisser & Becklen, 1975; Treisman, 1999). Several studies have established the involvement of such explicit grouping mechanisms during specific visual tasks. For example, different curve tracing paradigms require grouping of spatially separate contour segments (Roelfsema et al., 1999), and recent findings by Doerig et al. (2019), comparing a wide range of computational models, indicate that an explicit grouping step is crucial to explain different (un)crowding phenomena. Adding explicit segmentation mechanisms to DCNNs is promising to explain human behavior in tasks that require integrating and grouping of global features, or shape-level representations. Our results from behavioral experiments with segmented and unsegmented objects show that when the task is object recognition an explicit segmentation step is typically not necessary. We show that with an increase in network depth, there is a stronger influence of the features that belong to the object on recognition performance, showing that 'implicit' segmentation occurs. When this process becomes more efficient (with a deeper network, or recurrent processing) the result is a situation in which, just as in 'explicit' segmentation, the network (or visual system) knows which features belong together, and which ones do not. Previous studies have already looked into DCNN performance on unsegmented images (Cadieu et al., 2014; Cichy et al., 2017), or have even shown a decrease in classification accuracy for unsegmented, compared to segmented objects (Kheradpisheh et al., 2016)). In those images, however, objects were placed on a random background, thereby often incongruent (or coincidentally, congruent). In the current study, by manipulating the relevance and usefulness of the background information, we could disentangle whether this decrease was due to a segmentation problem, or the presence of incongruent, misleading information.

Contextual effects in object recognition

Different accounts of object recognition in scenes propose different loci for contextual effects (Oliva & Torralba, 2007; Võ et al., 2019). It has been argued that a bottom-up visual analysis is sufficient to discriminate between basic level object categories, after which context may influence this process in a top-down manner by priming relevant semantic representations, or by constraining the search space of most likely objects (e.g. Bar (2003)). Recent studies have also indicated that low-level features of a scene (versus high-level semantic components) can modulate object processing (Võ et al., 2019) by showing that seemingly meaningless textures with preserved summary statistics contribute to the effective processing of objects in scenes. Comparably, in the current study the DCNNs were agnostic to the meaning of the backgrounds, as they were not trained to recognize, for example, kitchens or bedrooms. The current results show that visual context features may impact object recognition in a bottom-up fashion, even for objects in a spatially incongruent location. Previous studies have indicated that explicitly augmenting DCNNs with human-derived contextual expectations (likelihood, scale and location of a target object) was able to improve detection performance, potentially indicating a difference in contextual representations in the networks and the humans (Katti et al., 2019). In the current study, findings show that only training DCNNs on a large dataset (ImageNet), enables them to learn human-like contextual expectations as well.

Feed-forward vs. recurrent processing

Instead of being an ultra-deep feedforward network, the brain likely uses recurrent connections for object recognition in complex natural environments. There are a multitude of findings that have firmly established the involvement of feedback connections during figure-ground segmentation. For example, behavior and neural activity in V1 evoked by figure-ground stimuli are affected by backward masking (Lamme et al., 2002), region-filling processes that are mediated by feedback connections lead to an enhanced neural representation for figure regions compared to backgrounds in early visual areas (Self & Roelfsema, 2014), responses by neurons showing selectivity to border ownership are modulated depending on the location of a 'figure' relative to other edges in their receptive field (Heydt, 2015), and the accuracy of scene segmentation seems to depend on recurrent connections to sharpen the local elements within early visual areas (Self et al., 2019) (and there are many more). The current results do not speak to those findings, but merely indicate that a very deep feedforward architecture is capable of obtaining a 'segmented' representation of an object, without recurrent projections. The interpretation that deeper networks are better at object recognition, because they are capable of limiting their analysis to (mostly) the object –when necessary– is consistent with the idea that deeper networks are solving the challenges that are resolved by recurrent computations in the brain (Liao & Poggio, 2016). Previous findings comparing human behavior or the representational geometry of neural responses to DCNNs (e.g. (Doerig et al., 2019; Khaligh-Razavi & Kriegeskorte, 2014)) often use images that contain (mostly) frontal views of objects on uniform backgrounds. For segmented objects, on a white or uniform background, all incoming information is relevant and segmentation is not needed. For those scenes, feed-forward activity in the brain may suffice to recognize the objects (Groen, Jahfari, et al., 2018). In line with those findings, we also see that even very shallow networks are able to perform well on those scenes. For more complex scenes, on the other hand, the first feed-forward sweep might not be not sufficiently informative, and correctly classifying or recognizing the object might require additional processing. For those scenes, we see a decrease in classification performance, mainly for the more shallow networks. These findings are in line with the global-to-local (or coarse-to-fine) processing framework, in which a coarse visual representation is acquired by the initial feedforward sweep. If this coarse representation is not informative enough to solve the task at hand, additional, more sophisticated visual processes ('routines') can be recruited to refine this representation (Crouzet & Serre, 2011; Epshtein et al., 2008; Groen, Jahfari, et al., 2018; Hochstein & Ahissar, 2002; Lamme & Roelfsema, 2000; Petro et al., 2014; Zheng et al., 2010).

Background congruency

In human natural vision, extraction of gist can lead to a set of expectations regarding the scene's composition, indicating the probability of the presence of a certain object in a scene, but also its most probable locations (Greene et al., 2015; Rémy et al., 2013). In the current study, in incongruent scenes, objects did not only violate the overall meaning of the scene category (semantic violation), but were also placed in a position that was not predicted by the local structure of the scene (syntactic violation). On top of that, objects in the human categorization task were placed in a semi-random location across trials to make the task more difficult. This spatial uncertainty, however, has the additional benefit that it makes the task more comparable to the task we ask DCNNs to perform, as DCNNs have no knowledge about the spatial location. A pilot study using stimuli with centered 3D-rendered objects indicated no difference in performance between congruent and incongruent images. While this is contrary to published literature (Munneke et al., 2013), there are several factors that might explain this difference. First of all, we used 3D-rendered, computer generated objects, placed on natural scenes (real-world pictures, Supporting Figure 4.7). The difference in visual quality and 'style' between the object and the background might have influenced perception, by making it easier to distinguish them from each other. A second reason might be the size of the objects. Compared to the stimuli used by Davenport & Potter (2004) or Munneke et al. (2013), our objects were quite large, in order to obtain good network performance.

Conclusion

With an increase in network depth there is better selection of the features that belong to the output category. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and might be one of the ways in which scene segmentation is performed in the brain.

Materials and methods

Experiment 1

Ethics statement

All participants provided written informed consent and were rewarded with research credits or received a monetary compensation. The experiment was approved by the ethics committee of the University of Amsterdam.

Participants

40 participants (9 males) aged between 18 and 30 years (M = 22.03, SD = 3.02) with normal or corrected-to-normal vision, took part in the experiment. Data from the first two participants were excluded from further data analyses due to technical problems.

Networks

We used Residual Networks (ResNets; He et al. (2016)) as a method to systematically manipulate network depth because this type of network consists of a limited number of fixed components that can be up-scaled without altering the architecture in another way. To evaluate whether the performance of ultra-deep ResNets can be approximated by recurrent computations, we al;so tested three different architectures from the CORnet model family (Kubilius et al., 2018); CORnet-Z (feedforward), CORnet-R (recurrent) and CORnet-S (recurrent with skip connections). Our implementation uses the PyTorch deep learning framework (Paszke et al., 2019) and the torchvision package. ResNet-6 was trained on ImageNet (Russakovsky et al., 2015) with 1 GPU. The other ResNets were downloaded (pretrained).

Stimuli

Images of 27 different object categories were generated by placing cut-out objects from the ImageNet validation set onto white (segmented), congruent and incongruent backgrounds. The categories were defined at a (sub)ordinate level, based on ImageNet categories: acoustic guitar, airliner, bathtub, birdhouse, cab, canoe, cellular telephone, china cabinet, dishwasher, grand piano, laptop, limousine, loudspeaker, mailbox, microphone, microwave, park bench, pirate ship, printer, remote, rocking chair, schoolbus, screen, speedboat, sports car, table lamp, wall clock (Figure 1A). There were ten exemplars for every object category. Backgrounds were sampled from a large database of images obtained from the SUN2012 database (Xiao et al., 2010) (512*512 pixels, full-color). For each category, three typical backgrounds were selected using the five most common places where this object was found within the database (sorted by number of instances inside each scene type). Three atypical backgrounds were manually chosen (Figure 1B). In total, the stimulus set contained 810 images with a congruent background, 810 with an incongruent background and 270 images with segmented objects. To familiarize human participants with the categories, one of the ten exemplars for each category was randomly selected and used in a practice-run. Using the remaining nine exemplars - three for each condition (segmented, congruent, incongruent) - 243 images were generated for the actual experiment. Each exemplar was only presented once for each participant. To ensure participants processed the complete image, exemplars were downsized and placed in one of 9 possible locations (3x3 grid). Importantly, to rule out any effect of 'exemplar-complexity' (e.g. one guitar being easier to recognize than another) or an interaction between the object, location and the background, all possible exemplar-background-location combinations were balanced over participants. For DCNNs, to make the comparison with human participants more valid and to estimate the reliability of the effects in our experiment, we showed different subsets of 243 stimuli to the DCNNs, each subset consisting of the same number of images per category and condition that human observers were exposed to (81 per condition, 3 per category).

Experimental procedure

Participants performed on an object recognition task (Figure 1C). At the beginning of each trial, a fixation-cross was presented for 2000 ms, followed by an image. Images were presented in randomized sequence, for a duration of 32 ms, followed by a mask. The masks consisted of scrambled patches of the images and was presented for 300 ms. After the mask, participants had to indicate which object they had seen, by clicking on one of 27 options on screen using the mouse. After 81 (1/3) and 162 (2/3) trials, there was a short break. Using this paradigm, our human object recognition task was closely analogous to the large-scale ImageNet 1000-way object categorization for which the DCNNs were optimized and thus expected to perform well.

Statistical analysis: human performance

Accuracy (percentage correct) was computed for each participant. Differences in accuracy between the three conditions (segmented, congruent, incongruent) were statistically evaluated using a non-parametric Friedman test. A significant main effect was followed up by Wilcoxon signed-rank tests using a Bonferroni correction at = 0.05, p-values reported in the main text are the adjusted p-values. Data were analyzed in Python.

Statistical analysis: DCNNs

For each of the images, the DCNNs (ResNet-6, ResNet-10, ResNet-18, ResNet-34, Resnet-50, ResNet-101, ResNet-152) assigned a probability value to each of the 1000 object categories it had been trained to classify. For each condition (segmented, congruent, incongruent) the Top-5 Error (%) was computed (classification is correct if the object is among the objects categories that received the five highest probability assignments). Then, to gain more insight in the importance of the features in the object vs the background for classification, we added Gaussian noise to either the object, background, or to both (the complete image) and evaluated performance.

Experiment 2

Results from experiment 1 suggested that information from the background is present in the representation of the object, predominantly for more shallow networks. What happens if we train the networks on segmented objects, when all features are related to the object? To further explore the role of segmentation on learning, we trained ResNets differing in depth on a dataset with objects that were already segmented, and a dataset in which they were intact (i.e. embedded in a scene).

Networks

As in experiment 1, we used deep residual network architectures (ResNets) with increasing number of layers (6, 10, 18, 34). Networks were implemented using the Keras and Theano code libraries (Chollet & Others, 2015; The Theano Development Team et al., 2016). In this implementation, input images were 128x128 randomly cropped from a resized image. We did not use ResNets with more than 34 layers, as the simplicity of the task leads to overfitting problems for the 'ultra-deep' networks.

Stimuli

To train the networks, a subset of images from 10 different categories were selected from ImageNet. The categories were: bird 1 t/m 7, elephant, zebra, horse. Using multiple different types of birds helped us to increase task difficulty, enforcing the networks to learn specific features for each class. The remaining (bigger) animals were added for diversity. From this subselection, we generated two image sets: one in which the objects were segmented, and one with the original images (objects embedded in scenes). Because many images are needed to train the models, objects were segmented using a DCNN pretrained on the MS COCO dataset (Lin et al., 2014), using the Mask R-CNN method (He et al., 2018) (instead of manually). Images with object probability scores lower than 0.98 were discarded, to minimize the risk of selecting images with low quality or containing the wrong object. All images were resized to 128x128 pixels. In total, the image set contained ~9000 images. 80% of these images was used for training, 20% was used for validation.

Experimental procedure

First, we trained the different ResNets for 100 epochs and monitored their accuracy after each epoch on the validation sets. Then, we reinitialized the networks with different seeds and repeated the process for 10 different seeds to obtain statistical results.

Data and code availability

Data and code to reproduce the analyses are available at the Open Science Framework (#gb89u) and at https://github.com/noorseijdel/2019_scenecontext

Supplement to Chapter 4

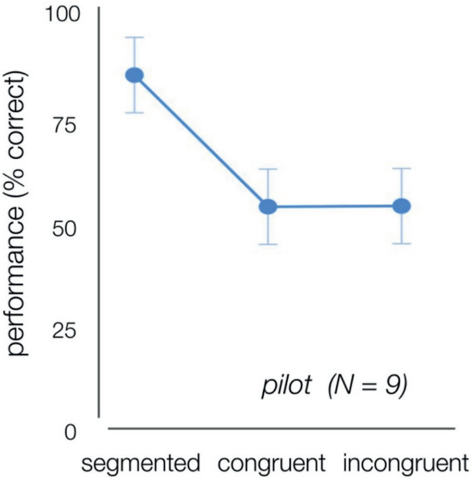

Figure 4.7: **Human performance (% correct) on the object recognition task, using centered 3D-rendered objects on white, congruent or incongruent backgrounds.** Performance was higher for the segmented condition, compared to congruent and incongruent.

Chapter 5

Visual features drive the category-specific impairments of categorization tasks in a patient with object agnosia

Abstract Object and scene recognition both require mapping of incoming sensory information to existing conceptual knowledge about the world. A notable finding in brain-damaged patients is that they may show differentially impaired performance for specific categories, such as for "living exemplars". While numerous patients with category-specific impairments have been reported, the explanations for these deficits remain controversial. In the current study, we investigate the ability of a brain-injured patient with a well-established category-specific impairment of semantic memory to perform two categorization experiments: 'natural' vs. 'manmade' scenes (experiment 1) and 'living' vs. 'non-living' objects (experiment 2). Our findings show that the pattern of categorical impairment does not respect the natural/living versus manmade/non-living distinction. This suggests that the impairments may be better explained by differences in visual features, rather than by category membership. Using Deep Convolutional Neural Networks (DCNNs) as 'artificial animal models' we further explored this idea. Results indicated that DCNNs with 'lesions' in higher order layers showed similar response patterns, with decreased relative performance for manmade (experiment 1) and living (experiment 2) items, even though they have no semantic category knowledge, beyond the pure visual domain. Collectively, these results suggest that the direction of category-effects to a large extent depends, at least in MS' case, on the degree of perceptual differentiation called for.

This chapter is in preparation as: Seijdel, N., Scholte, H.S., & de Haan, E.H.F. (n.d.). Visual features drive the category-specific impairments on categorization tasks in a patient with object agnosia

Introduction

Object or scene recognition requires mapping of incoming sensory information to existing conceptual knowledge about the world. A notable finding in brain-damaged patients is that they may show differentially impaired knowledge of, most prevalently, living things compared to non-living things (Gainotti, 2000). For many years, researchers have been investigating these category-specific semantic deficits. While they generally have been taken as strong evidence for a disturbance of semantic memory, recent findings have highlighted the importance of controlled experimental tasks and perceptual differences. To date, the debate remains unsettled on how this distinction in breakdown of semantic knowledge along the natural/living versus manmade/non-living axis arises (Capitani et al., 2003; Gainotti, 2000; Young et al., 1989).

Some studies have suggested that evolutionary pressures have led to a specialized, distinct neural mechanism for different categories of knowledge (e.g. animals, plants and artefacts), and that category-specific deficits arise from damage to one of these distinct neural substrates (Caramazza & Shelton, 1998; Nielsen, 1946). However, the most widespread views currently hold that they emerge because living and non-living things have different processing demands (i.e. they rely on different types of information). The first (most dominant) of those theories assumes that the storage of semantic information is divided into parts dominated by different knowledge aspects (e.g. perceptual, functional) and proposes that the dissociation arises from a selective breakdown of perceptual compared to functional associative knowledge. While man-made objects have 'clearly defined functions' and are mostly differentiated by their functional qualities, animals have less defining functions and are mostly distinguishable in terms of their visual appearance (Warrington & Shallice, 1984). This 'differential weighting' of perceptual and associative attributes might underlie the dissociation between living and non-living things. Later on this theory was revised to also include other modality-specific knowledge channels, such as a 'motor-related' channel, to support findings indicating greater impairments for certain more 'motor-related' or 'manipulable' items (such as tools or kitchen utensils) compared to larger manmade objects (such as vehicles) (Warrington & McCarthy, 1987).

A number of studies have emphasized the importance of intercorrelations amongst individual semantic features. This intercorrelation theory states that concepts are represented as patterns of activation over multiple semantic properties within a unitary distributed system. This intercorrelation theory is appealing in that it does not rely on damage to specific subtypes of attribute (visual, associative, motor) to produce category-specific deficits (Caramazza et al., 1990; Caramazza & Shelton, 1998; Tyler & Moss, 2001).

Another account holds that living items contain a larger number of structurally similar exemplars (e.g. many different types of trees), requiring a more fine-grained visual analysis for successful recognition (Sartori et al., 1993). In order words, it could be inherently more difficult to visually recognize living things compared to non-living things. This view of the structural description system, and their account for category-specific impairments is consistent with work on normal subjects and animal studies (Gaffan & Heywood, 1993). In line with these findings, a more recent study by (Panis et al., 2017) suggested that category-specific impairments may be explained by a deficit in recurrent processing between different levels of visual processing in the inferotemporal cortex. According to them, category-specificity has a perceptual nature, and the direction can shift, depending on perceptual demand. High structural similarity between stored exemplars might be beneficial for integrating local elements and parts into whole representations because the global and local features of these exemplars are more stable and more highly correlated in the real-world than the features from categories with low structural similarity. At the same time however, high structural similarity may be harmful for matching or precise recognition operations, because there may be more competition between the activated representations (Gerlach, 2009).

Here it's also important to note that different tasks have been used to evaluate patients' ability to recognize objects from different categories. Category-specific impairments have been established both using semantic memory experiments or visual recognition tasks at different levels (picture naming, picture-word matching, categorization). The differences in perceptual demand for these tasks (i.e. on which perceptual information they depend) might underlie the differences in category-specificity that have previously been found.

In the current study, we investigate the ability of a brain-injured patient with a category-specific impairment of semantic memory to perform scene- and object-categorization tasks (Figure 5.1). This patient, MS, has played a crucial role in the development of theories on category-specificity, showing a very clear category-specific deficit on semantic category fluency tests in previous studies. He has shown to perform better than control participants on non-living categories and significantly worse on living items (Young et al., 1989). A recent study showed that his impairments have remained unchanged for more than 40 years (De Haan et al., 2020). MS' problems with living items relative to non-living ones is apparent across a variety of tasks, including mental imagery, retrieval of information and visual recognition (Mehta et al., 1992). However, there is a striking dissociation between MS' preserved ability to access information about category membership in an implicit test (by priming identification of living and non-living items with related category labels), where there is no difference between the categories, and his severe problems in accessing such information in an explicit test (Young et al., 1989). These findings suggest that it's an "access" rather than a "storage" problem. Thus, the question remains as to whether MS can access stored representations of visual stimuli and, if so, what the relationships are between perceptual demand, recognition and semantic memory.

Here, two types of questions were addressed. The first - in order to investigate whether or not his category-specific impairment is dependent on perceptual factors - concerned MS' ability to either categorise visual images as depicting 'natural' vs. 'man-made' scenes (experiment 1) or to categorise visual objects are 'living' vs. 'non-living' things (experiment 2). Our findings show a dissociation between the two tasks, with better performance for inanimate objects, compared to animate objects (as is usually the case), and better performance for naturalistic scenes compared to manmade scenes.

The second question concerned the type of computations that might underlie the

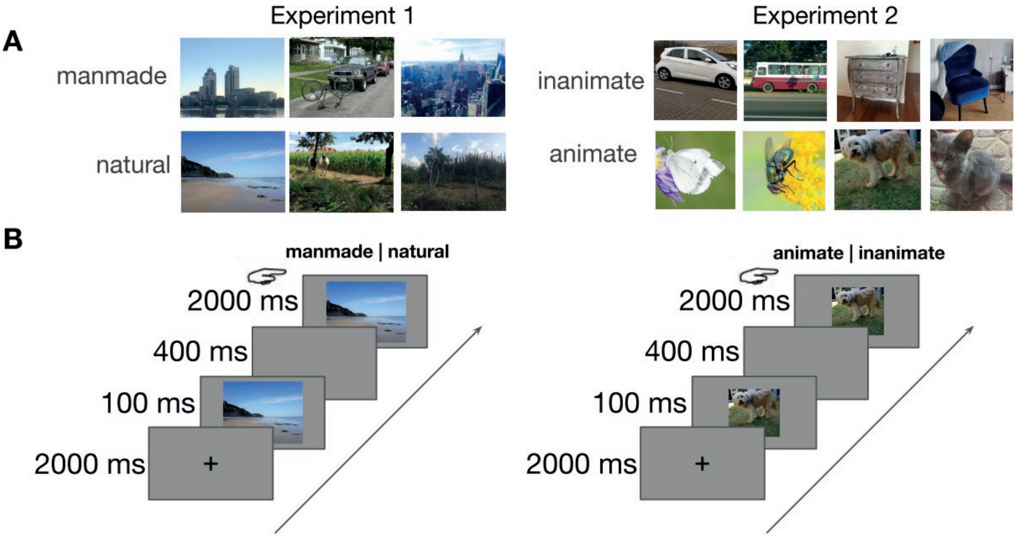

Figure 5.1: **Stimuli and experimental paradigm. A)** Examples (not part of actual stimulus set) of the two categories in experiment 1 (manmade vs. natural) and experiment 2 (inanimate vs. animate). **B)** Experimental design. After a 2000 ms blank screen, the stimulus was shown for 100 ms, followed by a 400 ms blank screen. Then, the image reappeared, and MS was asked to categorize the stimulus by pressing the corresponding button.

observed behavior. Recently, a class of computational models, termed deep convolutional neural networks (DCNNs), inspired by the hierarchical architectures of ventral visual streams demonstrated striking similarities with the cascade of processing stages in the human visual system (Cichy et al., 2016; Güçlü & Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). In particular, it has been shown that internal representations of these models are hierarchically similar to neural representations in early visual cortex (V1-V3), mid-level (area V4), and high-level (area IT) cortical regions along the ventral stream. Therefore, we evaluated performance of different DCNN architectures and compared it to MS' behavior. Results indicated that 'adding lesions' to higher-order layers of a DCNN resulted in response patterns similar to those of MS, with decreased performance for manmade (experiment 1) and living (experiment 2) things. Altogether, results from the current study indicate that, at least in specific cases such as MS, category-specific impairments can be explained by perceptual aspects of exemplars within different categories, rather than semantic category-membership.

Materials and methods

Case history

MS is a former police cadet who contracted herpes encephalitis in 1970 (for a full case description see also Ratcliff (1982)). Most of the ventral temporal cortex of both hemispheres was destroyed, extending to occipital cortex on the right, leaving him with a complete left homonymous hemianopia. He suffers from achromatopsia (Chadwick et al., 2019; Mollon et al., 1980), has severe object agnosia and prosopagnosia (e.g. Newcombe et al. (1989)), but is able to read accurately. His comprehension of what he reads is affected by an impairment of semantic memory. His semantic memory impairment is more marked for living than for non-living things (De Haan et al., 2020; Young et al., 1989).

Anatomical scans (Smits et al., 2019) revealed an, at least partially, intact primary visual cortex (V1) in both hemispheres. Further inspection of the anatomical scan suggests that this part of cortex in the right hemisphere, that could consist of parts of V1 to V4, is disconnected from subsequent cortical areas.

Stimuli

Scenes

240 images (640*480 pixels, full-color) of real-world scenes were obtained from a previous unpublished study by Chow-Wing-Bom et al. (2019). Of these 240 images, 80 images were labeled natural (>90% naturalness rating in an independent experiment), 80 images were man-made (<10% naturalness rating) and 80 images were ambiguous (between 10-90% naturalness rating). Ambiguous trials were collected for a different purpose and are not analyzed in the current study. The stimulus set contained a wide variety of different outdoor scenes including beaches, mountains, forests, streets, buildings and parking lots.

Objects

80 images (512*512 pixels) of animals (dogs, cats, butterflies and flies) and inanimate objects (cars, busses, cabinets and chairs) were selected from several online databases, including MS COCO (Lin et al., 2014), the SUN database (Xiao et al., 2010), Caltech-256 (Griffin et al., 2007), Open Images V4 (Kuznetsova et al., 2018) and LabelMe (Russell et al., 2008).

Experimental design

During the experiments, stimuli were presented for 100 ms, followed by a 500 ms blank screen. Then, the stimulus reappeared for 2000 ms and MS was asked to categorize the image as accurately as possible using one of two corresponding response buttons. Stimuli were presented in a randomized sequence, at eye-level, in the center of a 23-inch ASUS TFT-LCD display (1920*1080 pixels, at a refresh rate of 60 Hz), while MS was seated approximately 70 cm from the screen. The task was programmed in- and performed using Presentation (Version 18.0, Neurobehavioral Systems Inc., Berkeley, CA, www.neurobs.com). After every 40 trials there was a short break. During the task, EEG was recorded.

Statistical analysis: behavioral data

Choice accuracies were computed for each condition in both experiments (Figure 5.2). Differences between the conditions were tested using two-tailed permutation testing with 5000 permutations. Behavioral data were analyzed and visualized in Python using the following packages: Statsmodels, SciPy, NumPy, Pandas and Seaborn (Jones et al., 2001; McKinney & Others, 2010; Oliphant, 2006; Seabold & Perktold, 2010).

Deep Convolutional Neural Networks (DCNNS)

First, to evaluate how many layers were sufficient to accurately perform the categorization tasks, tests were conducted on four deep residual networks (ResNets; He et al. (2016)) with increasing number of layers; ResNet-6, ResNet-10, ResNet-18 and Resnet-34. Pre-trained networks were fine-tuned to perform either the manmade vs. natural categorization task, or the animate vs. inanimate categorization task, using PyTorch (Paszke et al., 2019). Training data was obtained using the SUN2012 database (Xiao et al., 2010) for manmade an natural scenes, and ImageNet (Russakovsky et al., 2015) for animate and inanimate objects. All sets contained a representative variety of different categories, similar to the stimuli used in the experimental task. Each model was initial-ized five times with different seeds to perform statistical analyses. For ResNet-10, the most shallow network that was able to successfully perform the task (>95% accuracy on all conditions), we evaluated categorization performance after 'lesioning' higher-order layers. To this end, we removed one of the 'building blocks', while keeping the skip connection intact.

Data and code availability

Data and code to reproduce the analyses are available at the Open Science Framework (#9h7mf) and at https://github.com/noorseijdel/2020_Object_agnosia.

Results

First, categorization performance (proportion correct) of MS was computed for both categorization tasks. Results from two-sample permutation tests with 5000 permutations indicated higher performance for natural (experiment 1) and inanimate (experiment 2) images (p = 0.007, p = 0.016, respectively). Thus, in the scene categorization task, MS was significantly better at classifying visually the natural compared to manmade environments. In contrast, on the object categorisation task, he was significantly better at assigning the inanimate than the animate exemplars to the correct category.

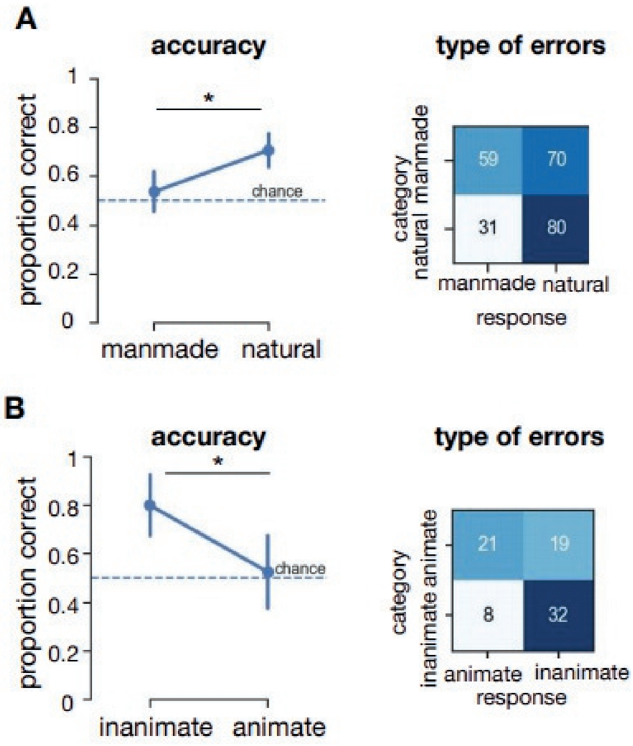

Figure 5.2: **Behavioral results of patient MS in experiment 1 and 2. A)** Results from experiment 1 (manmade vs. natural) Accuracy (proportion correct) per condition. Horizontal black lines indicate the results of two sample permutation tests, two-tailed using 5000 permutations. Error bars represent the bootstrap 95% confidence interval. * = p < 0.05. **B)** Results from experiment 2 (animate vs. inanimate).

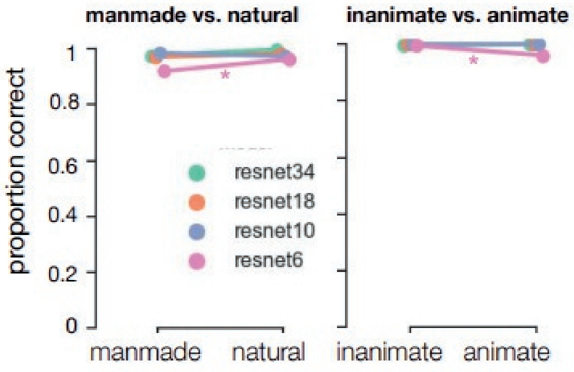

Figure 5.3: **Performance of ResNets with different depth (number of layers) on the images from experiment 1 and 2.** The ResNets were pretrained on ImageNet, and fine-tuned on an independent set of manmade and natural scenes and images containing inanimate and animate objects.

ResNet-10, -18 and -34 all showed virtually perfect performance for both tasks, for all categories (Figure 5.3). For the most shallow network, ResNet-6, there was a slight decrease in performance, specifically for manmade (experiment 1) and animate objects (p = 0.02, p = 0.03, respectively). Overall these results indicate that performance of a shallow ResNet-6 may decrease in a similar fashion as MS. This supports the idea that performance is decreased for specific categories because those stimuli (in our dataset) are more difficult. Still, even for a shallow ResNet-6, the two-option categorization tasks seems too easy.

Finally, we evaluated the performance of ResNet-10 after 'lesioning' higher-order layers (Figure 5.4A). In order to mimic lesions to higher-order areas in the visual processing stream, we removed connections to the final building block of the network (Block 4). Permutation tests with 5000 permutations between ResNet-10 without and with lesion, indicated a decrease in performance after elimination of higher-order layers, specifically for manmade (experiment 1) and animate (experiment 2) images (both p < .001). For natural scenes, there was a slight increase in performance after the removal of higher order layers (p = 0.023). Lesions in earlier layers of the network (blocks 1-3) resulted in a strongly biased response, in which the network generally classified all images as belonging to the same category (Supplementary Figure 5.5). The direction of this bias was variable across different initializations, suggesting that the earlier layers are crucial to obtain a useful representation, and the bias was not caused by the current stimulus set.

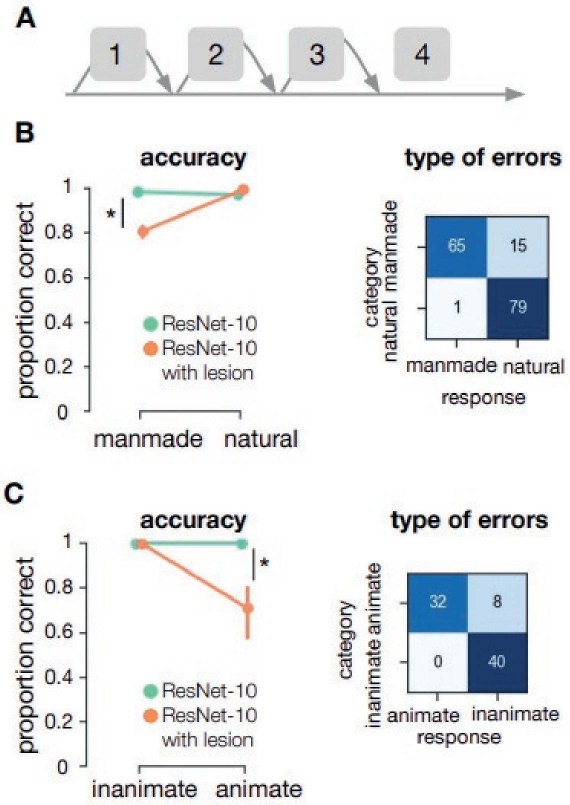

Figure 5.4: **ResNet-10 performance on the images from experiment 1 and 2. A)** Schematic representation of ResNet-10. ResNet are built by stacking blocks (containing the convolution, batch normalization and pooling operations). Bypassing the different blocks, skip connections add the input directly to the next block. Here, we added a lesion to ResNet-10 by removing block 4. **B)** Performance of ResNet-10, with and without lesion, on the categorization tasks.

Discussion

We evaluated the extent to which MS' ability to recognize visual information shows selective impairments for semantic categories. Our findings show a dissociation between two associated tasks (categorization of manmade vs. natural scenes and animate vs. inanimate objects), with better performance for inanimate objects, compared to animate objects (as is usually the case), and better performance for naturalistic scenes compared to manmade scenes. Overall, these results indicate that the category-specific effects, at least for patient MS, are better explained as a visual impairments, invalidating the idea that this is a purely semantic disorder (i.e. by category membership only). This is in line with earlier findings from Young et al. (1989), and suggests that, similar to findings in earlier studies by Gerlach (2001) and Låg (2005), the direction of category-effects to a large extent depends on the degree of perceptual differentiation called for. Using Deep Convolutional Neural Networks as 'artificial animal models' (Scholte, 2018) we further explored the type of computations that might underlie such behavior. Overall, DCNNs with 'lesions' in higher order areas showed similar response patterns, with decreased performance for manmade (experiment 1) and living (experiment 2) things.

Category selectivity in the visual ventral stream

There is an ongoing debate on the emergence of category selectivity in the visual ventral stream of healthy subjects. A popular view is that observed category effects indicate a high-level representation in which neurons are organised around either object category or correlated semantic and conceptual features (Konkle & Oliva, 2012; Kriegeskorte et al., 2008; Mahon et al., 2009). An alternative view is that categorical responses in the ventral stream are driven by combinations of more basic visual properties that covary with different categories (Andrews et al., 2015; Long et al., 2018). The conflation of visual and semantic properties in object images means that category-selective responses could be expected under both accounts. Results from the current study do not speak to these findings, nor include/exclude the possibility for object category-selective responses driven by categorical or semantic properties. However, these findings do indicate that in object recognition impairments (following brain damage to certain regions), apparent category-selectivity can emerge based on basic visual properties.

Object representations in IT

A question that remains unresolved in this study is which visual features might be involved in classification of the different categories, i.e. which dimensions in stimulus or object space are utilized by MS. Recent work by Bao et al. (2020) shows that specialization of different categories in certain regions in IT can be explained by two dimensions, progressing from animate to inanimate (dimension 1), and from more stubby to spiky (dimension 2). Following these dimensions, lesions to different parts of IT should lead to agnosias in specific sectors of object space. For example, the observation that MS' specifically does not recognize insects (which are generally considered more 'spiky' than mammals, Supplementary Figure 5.6) as being animate might be explained by a disturbed 'spiky animate corner' in object space.

Effect of typicality on category-membership decisions

The typicality of a target object is known to influence category-membership decisions (Shoben, 1982). For a given semantic category, the more typical members can be accepted as belonging to that category more quickly than less typical members. In earlier studies, MS also showed faster reactions to more typical exemplars (Young et al., 1989). However, on top of this 'typicality effect', MS showed faster responses to non-living things than living things. In the current study, performance on experiment 2 was merely decreased for insects (Supplementary Figure 5.6). One explanation could be that insects are less typical for the 'animate' condition than mammals, and therefore performance was decreased for these images.

Objects vs. scene categorization

Perceiving a scene involves different information than recognition of objects. Object and scene recognition both require mapping of low-level incoming sensory information to high-level representations and semantic knowledge. Following the reverse hierarchy theory (Hochstein & Ahissar, 2002), coarse and global information is extracted before detailed information becomes available. In particular, this theory suggests that the rapid categorization of real-world scenes with minimal effort (Greene & Oliva, 2009b; Potter, 1975) may be mediated by a global percept of the conceptual 'gist' of a scene. Thus, low- and mid-level properties may be particularly diagnostic for the behavioral goals specific to scene perception, while object recognition might depend on more extensive processing of high-level properties (Groen, Jahfari, et al., 2018; Groen et al., 2017). In the current study, the natural/man-made distinction may be made before basic-level object distinctions. MS could have relied on this global percept, or 'gist' for experiment 1, while this information would not suffice or be informative for experiment 2.

Overall, these results suggest that semantic impairments for certain categories can, at least in MS' case, be explained by differences in perceptual demand and early visual features, rather than by category membership. Additionally, these findings show that utilizing different DCNN architectures (with and without virtual lesions) offers a promising framework when studying human visual cognition.

Supplement to Chapter 5

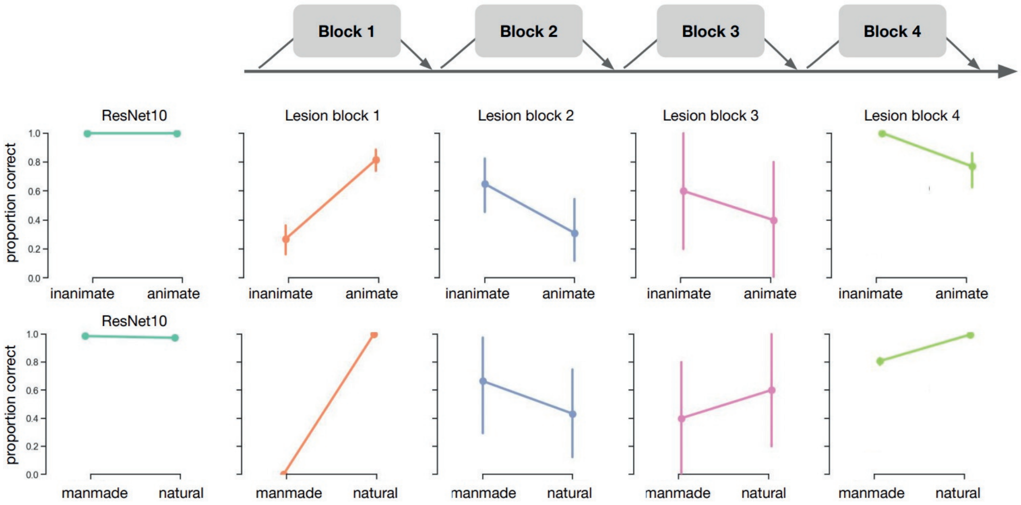

Figure 5.5: **ResNet-10 performance on the images from experiment 1 and 2 after removing block 1, 2, 3 and 4.**

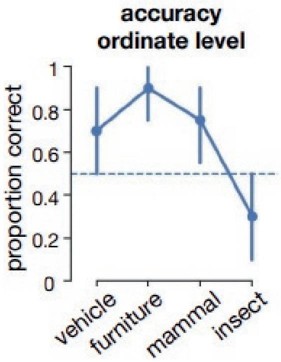

Figure 5.6: **Results from experiment 2 (animate vs. inanimate) on the ordinate level.**

Chapter 6

Summary and Discussion

We see the world in scenes, where objects are embedded and often partially occluded in rich and complex surroundings containing additional objects. How does the brain extract and transform diagnostic low-level visual features into richer representations that facilitate recognition, whilst there are so many factors that affect the appearance of natural object categories? In this thesis, I examined to what extent object and background information is represented and used for object recognition in human subjects and in deep convolutional neural networks. More specifically, I evaluated how different functional architectures or differences in information flow (feed-forward vs. recurrent) exhibit sensitivity to natural scene properties. My experiments focused on the role of natural scene complexity, as indexed by two biologically plausible image statistics, and the manipulation of 'informative' (congruent) information in visual scenes. Overall, results show that recognizing objects in simple scenes can occur in a feed-forward manner, on the basis of a first, coarse representation. For more complex scenes or more challenging situations, additional extensive processing (in the form of recurrent computations) are required. Additionally, results indicate that object recognition can be performed based on feature constellations, without any determination of boundary or segmentation. Finally, it showcases a potential role for DCNNs as artificial animal models of human visual processing. In the following sections I will discuss the obtained results in more detail. Throughout, I will discuss their implications for our understanding of object recognition in natural scenes. Finally, I will go into the broader context of our research and discuss some outstanding questions.

Motivation and summary of the results

The initial motivation came from the findings in Groen, Jahfari, et al. (2018), where we found that object detection was more difficult for scenes with low spatial coherence (SC) and high contrast energy (CE), i.e., high SC/CE values. CE and SC are computed using a simple visual model that simulates neuronal responses in one of the earliest stages of visual processing. Specifically, they are derived by averaging the simulated population response of LGN-like contrast filters across the visual scene (Ghebreab et al., 2009; Scholte et al., 2009). In turn, they could serve as a complexity index that affects subsequent computations towards a task-relevant visual representation. Combined fMRI and EEG results from that study showed that for complex scenes only, early visual areas were selectively engaged by means of a feedback signal. These findings suggested that when the initial global scene impression signals the presence of a high SC/CE scene (indicating that it contains clutter), the visual system has to perform more effortful detailed analysis of the scene, which involves recruitment of information from early visual areas.

In **chapter 2**, we wondered whether these low-level, task-irrelevant properties would also influence perceptual decision-making. In addition, we attempted to dissociate the contributions of the two different axes describing the image complexity 'space' (CE and SC). We used regression analyses in which we included both linear terms as well as second-order polynomials to examine whether the relationship between SC/CE and two parameters from the Drift Diffusion Model (DDM; Ratcliff & McKoon (2008); Wiecki et al. (2013)) was linear or curvilinear (e.g. followed an inverted U-shape). Results indicated that scene complexity, as indexed by our two parameters (SC, CE), modulated perceptual decisions through the speed of evidence accumulation. Our results indicated that the speed of evidence accumulation was related to differences in both SC (linear) and SC^2 (inverted U-shape). That is, low and high SC were associated with a decreased drift rate, as indicated by a negative shift in the posterior distribution. A second experiment refined these observations by showing how the isolated manipulation of SC alone resulted in weaker yet comparable effects, whereas the manipulation of CE had no effect. Overall, these results showed that very basic properties of our natural environment influence perceptual decision-making. Because SC and CE could be plausibly computed in early stages of visual processing, they could indicate the need for more cautious or elaborate processing by providing the system with a global measure of scene complexity.

A question that arose from these findings, was whether the effects were driven by SC and CE 'itself' (low-level regularities) or because they covary with other sources of information in the scene. SC and CE clearly covary with interesting properties of natural scenes, but exactly because of this covariance it is difficult to isolate their impact on visual processing. Therefore, in **chapter 3**, we explored whether these effects could be based on the computation of SC and CE more directly, as a 'general measure' of complexity, or indirectly, as diagnostic information to estimate other task-relevant scene properties (e.g. naturalness). To this end, we manually segmented the objects from their real-world scene backgrounds and superimposed them on phase scrambled versions of the real-world scenes. For each complexity condition, backgrounds were selected using the same cut-off values from Groen, Jahfari, et al. (2018), and each object was presented in all conditions. This allowed us to evaluate the influence of SC and CE and the subsequent effect on segmentability, while removing any (ir)relevant object and context information. Additionally, in half the trials, we hindered recurrent processing with visual backward-masking (Fahrenfort et al., 2007). A convergence of results indicated that recurrent computations were increasingly important for recognition of objects in more complex environments (i.e. objects that were more difficult to segment from their background). First of all, behavioral results indicated poorer recognition performance for objects with more complex backgrounds, but only when feedback activity was disrupted by masking. Second, EEG measurements showed clear differences between complexity conditions in the ERPs around 200 ms - a time window beyond the first feed-forward visual sweep of activity (Lamme & Roelfsema, 2000). Additionally, object category decoding based on the multivariate EEG patterns showed later decoding onsets for objects embedded in more complex backgrounds. This indicated that object representations for more complex backgrounds emerge later, compared to objects in more simple backgrounds. Finally, Deep Convolutional Neural Network (DCNN) performance confirmed this interpretation; feed-forward network architectures showed a higher reduction in recognition performance for objects in more complex backgrounds compared to networks equipped with recurrent connections (Kubilius et al., 2018).

A limitation of any experiment with artificially generated (or artificially embedded) images is that it is unclear whether the findings generalize to 'real images' that have not been manipulated in any way. Together with the previous findings, however, our results corroborate the idea that more extensive processing (in the form of recurrent computations) is required for object recognition in more complex, natural environments (Groen, Jahfari, et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Tang et al., 2018). Nonetheless, the manipulation of SC/CE using artificial vs. naturalistic backgrounds led to slightly different patterns of results. Using artificial backgrounds, scene complexity showed to have a linear effect (seen in chapter 3). When using natural scenes, results showed enhanced performance for medium complex trials (Chapter 2 and Groen, Jahfari, et al. (2018); inverted U-shape). While there are several other factors that could explain this discrepancy (described in the next section), we wondered to what degree real-world scene context influenced recognition performance. Furthermore we wanted to compare the process of scene segmentation for object recognition between human and artificial neural networks in the hope that this could give insight into the question how scene segmentation might be implemented computationally.

Therefore, **in chapter 4**, we evaluated how object and context information is represented and used for object recognition in different DCNNs. More specifically, we investigated how the number of layers (depth) in a DCNN influences scene segmentation and how this compares to human behavior.

Experiment 1 showed both substantial overlap, and differences in performance between human participants and DCNNs. Both humans and DCNNs were better in recognizing an object when it was placed on a congruent versus an incongruent background. However, whereas human participants performed best in the segmented condition (object on homogenous background), DCNNs performed equally well (or better) for the congruent condition. Performance for the incongruent condition was

lowest. This effect was particularly strong for more shallow networks. Notably, the shift in performance from the most shallow network to deeper networks (ResNets; He et al. (2016)) showed the same pattern as the shift from a shallow feedforward architecture to a recurrent architecture (CORnets; Kubilius et al. (2018)], suggesting that there is a functional equivalence between additional nonlinear transformations and recurrence.
Further analyses, investigating which parts of the image were most important for recognition (Zeiler & Fergus, 2014), showed that the influence of background features on the response outcome was relatively strong for less deep networks and almost absent for deeper networks. These findings suggest that one of the ways in which network depth improves object classification, is by learning how to select the features that belong to the object, and thereby implicitly segregating the object features from the other parts of the scene. To complement these findings, we performed an additional experiment in which we tested how *training* was influenced by network depth. If shallow networks fail to correctly recognize objects, merely because they do not learn to implicitly segment the object from the background (while deeper networks do), we expected them to show a larger increase in performance when trained with segmented vs. unsegmented stimuli (as compared to deeper networks). Indeed, results indicated a benefit of training on segmented objects (as compared to unsegmented objects) for more shallow networks. For deeper networks, this benefit was much less prominent. Training on segmented images thus reduced the difference in performance between shallow and deeper networks.

Deep convolutional neural networks thus seem to learn high-level concepts such as objects based on low-level visual input, without existing conceptual knowledge of these concepts. **In chapter 5**, we examined visual processing in a situation in which visual information can no longer be reliably mapped onto existing conceptual knowledge. To this end, we evaluated object and scene categorization in a brain-injured patient MS, with severe object agnosia and category-specific impairments. Our findings show a dissociation between two semantically associated tasks (categorization of manmade vs. natural scenes and animate vs. inanimate objects), with better performance for inanimate objects, compared to animate objects (as is usually the case), and better performance for naturalistic scenes compared to manmade scenes. Using Deep Convolutional Neural Networks as 'artificial animal models' (Scholte, 2018) we further explored the type of computations that might produce such behavior. Overall, DCNNs with 'lesions' in higher order areas showed similar response patterns, with decreased performance for manmade (experiment 1) and living (experiment 2) things. This indicates that behavioral category representations (and subsequent impairments) might be explained by a difference in low-level image statistics or physical properties of the stimuli, and thus by a difference in visual input that they provide to the visual system. Altogether, results from this study indicated that, at least in specific cases such as MS, category-specific impairments can be explained by perceptual aspects of exemplars within different categories, rather than semantic category-membership.

Taken together, our results suggest that recognizing objects in simples scenes, or categorizing very dissimilar target options (e.g. in terms of global properties) could occur on the basis of a first, coarse representation, often described as visual gist (Greene & Oliva, 2009b; Oliva & Torralba, 2006; Torralba & Oliva, 2003). Overall, our findings are line with theories of visual processing proposing that a global impression of the scene accompanies (Rousselet et al., 2005; Wolfe et al., 2011) or precedes (Hochstein & Ahissar, 2002) detailed feature extraction ("coarse-to-fine" processing Hegdé (2008)]). The current results add to this view by showing how this complexity could arise, and what type of functional architecture might produce this behavior. Our results suggest that 'core object recognition' can occur in a feed-forward manner when the visual input is 'simple', but that recurrent computations aid object recognition performance in more challenging conditions. Additionally, our results show that for object recognition, an explicit segmentation step is potentially not necessary. This is in line with recent findings from Tang et al. (2018) and Rajaei et al. (2019), where they showed that backward masking led to a large reduction in human object recognition performance for partially visible or occluded objects. Similarly, both studies found that (more shallow) feed-forward architectures were not robust to partial visibility or occlusion of objects, and that adding recurrent computations led to improvements. From a perspective of vision as subservient to action, this makes sense: if certain visual elements form an object in the first sweep of information, the aim of the brain is often to use this information to characterize or interact with the object, not to go back or zoom in on all possible details about its constituting elements. In order words: to recognize a cat, we do not necessarily need to know where its' legs are (or whether it still has all four). If the first sweep of information is insufficient, it might pay off to wait a little longer and implement recurrent computations to gather more evidence (chapter 2) and obtain a sufficiently detailed representation.

Manipulations of visual processing

The central aim of this thesis was built around two different ways of evaluating visual processing: 1) increasing task difficulty, thereby enhancing the need for recurrent computations, and 2) the effect of decreased quality of visual processing by interfering with recurrent processing or investigating a patient with bilateral temporo-occipital damage. There are a myriad of ways to interfere with visual processing and our design choices have undoubtedly affected our results. Here I will describe the most important varieties in our experimental paradigms, and their implications for our interpretations.

To increase the need for recurrent computations, we mostly focused on the low-level complexity of the visual input (chapter 2-3) and the manipulation of congruent vs. incongruent context information (chapter 4). However, there were several other varying factors in our experimental paradigms.

For example, the amount of response options varied between the different studies, potentially influencing the level of categorization required to accurately perform the task. Objects can be categorized at different levels of abstraction, from superordinate (e.g. animal vs. no-animal in chapter 2 and chapter 5), ordinate (or 'basic', e.g. dog or cat; chapter 3), to more subordinate (e.g. school bus or sports car; Chapter 4). At the perceptual level, features to account for distinct object categories may have differed between the tasks and decreasing the amount of response options may have influenced the amount and/or type of information necessary to analyze the scene (Macé et al., 2009; Rosch et al., 1976). For example, to distinguish between an animate or inanimate object one might rely on more global features, whereas to identify a certain object out of twenty-seven options (chapter 4) more detailed (local) information is needed to accurately distinguish between them.

What is clear from all studies reported in this thesis is that object recognition can (almost) always be solved, given enough time, a-priori knowledge or visual processing capacity. Therefore, in most experiments we additionally manipulated the quality or opportunity for recurrent processing to take place. In the experiments with human participants, we shortened presentation times (ranging from 34 - 100 ms) or applied visual backward-masking. In chapter 5, visual processing of patient MS was severely impaired by lesions to most of the ventral temporal cortex of both hemispheres. For the DCNNs we manipulated network depth, the presence or absence of recurrent connections and the removal of certain connections to 'mimic' lesions.

Taken together, these (sometimes subtle) differences in experimental paradigms and procedures can explain some of the discrepancies in our current findings. For example, in chapter 2, using naturalistic scenes and a superordinate type of task, medium scene complexity was associated with an increased speed of evidence accumulation and enhanced behavioral performance. In that chapter, we discuss several explanations for why scenes with medium CE/SC values could be processed more efficiently, including higher daily frequency (as in, occurring more often in the 'real-world') or the amount of contextual information. In chapter 3, using five 'basic-level' objects embedded in artificially generated backgrounds, higher scene complexity led to an incremental decrease in performance (with visual backward masking). This suggests that low complex naturalistic scenes might be processed differently than artificial scenes. Whether this is because of task demand, expectations or context is unclear from the current results, and should emerge from future research.

Probing cognition with DCNNs

Classic models of object recognition focusing on grouping and segmentation presume an explicit process in which certain elements of an image are grouped, whilst others are segregated from each other, by a labelling process. Our results from behavioral experiments in DCNNs show that, when the task is object recognition, an explicit segmentation step might not be necessary. We interpret these findings as indicating that with an increase in network depth there is better selection of the features that belong to the output category (vs. the background), resulting in higher performance during recognition. Thus, more layers are associated with 'more' or better segmentation, by virtue of increased selectivity for relevant constellations of features. There is thus no discrete 'moment' at which segmentation is successful or 'done'. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and we speculate that it might be one of the ways in which scene segmentation is performed in the brain (using recurrent computations). What these results additionally show is that certain psychological concepts or classifications (e.g. 'object' and 'background') make distinctions that are not recognized by deep convolutional neural networks, and potentially fail to capture the computations of visual processing. For example, while edges and borders of objects were traditionally seen as very important for successful recognition, the current results suggest that we do not necessarily need to detect those in many everyday behaviors. For these networks, objects are not 'things' that 'exist' in a certain location with a clear boundary. Visual properties from all regions in the image are processed, and together result in a robust representation that the network can utilize for classification. Thus, on the basis of low-level input features that map reliably enough onto high-level feature constellations. The 'invention' of deep convolutional neural networks as computational models of the human visual system in this sense allows addressing questions that previously could not be answered (or had not been asked). Without the constraints of experimental set-ups, using DCNNs enables us to ask questions about the underlying mechanisms producing behavior. Instead of building an experiment on the building blocks of psychological concepts, we can explore the borders of experimental manipulations and start asking 'when' and 'how' questions. Crucially, this serves as hypothesis generation, and any obtained new insight from DCNNs will need to be verified and confirmed in human data. Of course, counterarguments can be made to this approach. First, of all, DCNNs are far, far away from being an ultimate model explaining all biological visual processing (Cichy & Kaiser, 2019; Kriegeskorte, 2015; Lindsay, 2020). They generally lack many types of biological properties that are known to be involved in neural processing, they makes different types of errors compared to humans, they generalize poorly beyond the datasets on which they are trained, etc. Clearly, DCNNs are very different from a biological visual system. A second, and perhaps more important, counterargument is that the search space might be too large to solve. It is probably impossible or unfeasible to explore the immense zoo of different architectures, combined with an infinite number of possibilities to investigate different visual diets, training regimes and tasks. One fruitful approach to help navigate or constrain the search space is by combining knowledge from biological vision with existing models. Over the last years there has been an increase in research aiming to augment or equip DCNNs with additional biologically-inspired features and mechanisms. For example, by implementing biological attention mechanisms (Lindsay & Miller, 2018), artificial spiking neural networks (Tavanaei et al., 2019); biological learning rules (Pozzi et al., 2018), or recurrent computations to capture the representational dynamics of the human visual system (Güçlü & Gerven, 2017; Kietzmann, McClure, et al., 2019). Overall, the combination of research in both human and artificial vision offers a promising framework for the investigation of both human visual processing and the development of computational models.

Bibliography

Andrews, T. J., Watson, D. M., Rice, G. E., & Hartley, T. (2015). Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *Journal of Vision*, *15*(7), 3–3.

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychon. Bull. Rev.*, *14*(2), 332–337.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.*, *15*(4), 600–609.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629.

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception*, *25*(3), 343–352.

Biederman, I. (1972). Perceiving Real-World scenes. *Science*, *177*(4043), 77–80.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cogn. Psychol.*, *14*(2), 143–177.

Bodamer, J. (1947). Die prosop-agnosie. *Archiv Für Psychiatrie Und Nervenkrankheiten*, *179*(1-2), 6–53.

Boehler, C. N., Schoenfeld, M. a, Heinze, H.-J., & Hopf, J.-M. (2008). Rapid recurrent processing gates awareness in primary visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, *105*(25), 8742–8747.

Brady, N., & Field, D. J. (2000). Local contrast in natural images: Normalisation and coding efficiency. *Perception*, *29*(9), 1041–1055.

Breitmeyer, B. G., & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review, and update. *Percept. Psychophys.*, *62*(8), 1572–1595.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.*, *10*(12), e1003963.

Camprodon, J. A., Zohary, E., Brodbeck, V., & Pascual-Leone, A. (2010). Two phases of V1 activity for visual recognition of natural images. *J. Cogn. Neurosci.*, *22*(6), 1262–1269.

Camprodon, J. A., Zohary, E., Brodbeck, V., & Pascual-leone, A. (2013). *Two phases of V1 activity for visual recognition of natural images*. *18*(9), 1199–1216.

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cogn. Neuropsychol.*, *20*(3), 213–261.

Caramazza, A., Hillis, A. E., Rapp, B. C., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cogn. Neuropsychol.*, *7*(3), 161–189.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain the animate-inanimate distinction. *J. Cogn. Neurosci.*, *10*(1), 1–34.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Atten. Percept. Psychophys.*, *72*(5), 1283–1297.

Chadwick, A., Heywood, C., Smithson, H., & Kentridge, R. (2019). Translucence perception is not dependent on cortical areas critical for processing colour or texture. *Neuropsychologia*, *128*, 209–214.

Chollet, F., & Others. (2015). *Keras*.

Chow-Wing-Bom, H. T., Scholte, S., De Klerk, C., Mareschal, D., Groen, I. I. A., & Dekker, T. (2019). Development of rapid extraction of scene gist. *PERCEPTION*, *48*, 40–41.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage*, *153*, 346–358.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.*, *17*(3), 1–10.

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based fusion of MEG and fMRI reveals Spatio-Temporal dynamics in human cortex during visual object recognition. *Cereb. Cortex*, *26*(8), 3563–3579.

Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*, *105*, 165–176.

Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition? *Front. Psychol.*, *2*, 326.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Mem. Cognit.*, *35*(3), 393–401.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychol. Sci.*, *15*(8), 559–564.

De Haan, E. H. F., & Cowey, A. (2011). On the usefulness of “what” and “where” pathways in vision. *Trends Cogn. Sci.*, *15*(10), 460–466.

De Haan, E. H., Seijdel, N., Kentridge, R. W., & Heywood, C. A. (2020). Plasticity versus chronicity: Stable performance on category fluency 40 years post-onset. *Journal of Neuropsychology*, *14*(1), 20–27.

Dejerine, J. (1892). Contribution à l'étude anatomopathologique et clinique des différents variétés de cécité verbale. *Mémoires de La Société de Biologie*, *4*, 61–90.

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the non-linear threshold for access to consciousness. *PLoS Biol.*, *5*(10), e260.

Deng, J. D. J., Dong, W. D. W., Socher, R., Li, L.-J. L. L.-J., Li, K. L. K., & Fei-Fei, L. F.-F. L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2–9.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.*, *11*(8), 333–341.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., & Herzog, M. H. (2019). Beyond bouma's window: How to explain global aspects of crowding? *PLoS Comput. Biol.*, *15*(5), e1006580.

Downing, P. E., Chan, A.-Y., Peelen, M., Dodds, C., & Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral Cortex*, *16*(10), 1453–1461.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, *152*, 184–194.

Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proc. Natl. Acad. Sci. U. S. A.*, *105*(38), 14298–14303.

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

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *J. Cogn. Neurosci.*, *19*(9), 1488–1497.

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

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nat. Neurosci.*, *8*(12), 1643–1646.

Felsen, G., Touryan, J., Han, F., & Dan, Y. (2005). Cortical sensitivity to visual features in natural scenes. *PLoS Biol.*, *3*(10), 1819–1828.

Fink, M., & Perona, P. (2004). Mutual boosting for contextual inference. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16* (pp. 1515–1522). MIT Press.

Gaffan, D., & Heywood, C. A. (1993). A spurious category-specific visual agnosia for living things in normal human and nonhuman primates. *J. Cogn. Neurosci.*, *5*(1), 118–128.

Gainotti, G. (2000). What the locus of brain lesion tells us about the nature of the cognitive defect underlying category-specific disorders: A review. *Cortex*, *36*(4), 539–559.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *arXiv Preprint arXiv:1706.06969*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv Preprint arXiv:1811.12231*.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 7538–7550.

Geisler, W. S., & Diehl, R. L. (2003). A bayesian approach to the evolution of perceptual and cognitive systems. *Cogn. Sci.*, *27*(3), 379–402.

Gerlach, C. (2001). Structural similarity causes different category-effects depending on task characteristics. *Neuropsychologia*, *39*(9), 895–900.

Gerlach, C. (2009). Category-specificity in visual object recognition. *Cognition*, *111*(3), 281–301.

Ghebreab, S., Scholte, S., Lamme, V., & Smeulders, A. (2009). A biologically plausible model for rapid natural scene identification. *Adv. Neural Inf. Process. Syst.*, 629–637.

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Front. Comput. Neurosci.*, *8*, 74.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

Goodale, M. A., Milner, A. D., & others. (1992). *Separate visual pathways for perception and action*.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Atten. Percept. Psychophys.*, 1239–1251.

Greene, M. R., & Oliva, A. (2009a). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cogn. Psychol.*, *58*(2), 137–176.

Greene, M. R., & Oliva, A. (2009b). The briefest of glances: The time course of natural scene understanding. *Psychol. Sci.*, *20*(4), 464–472.

Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset*. 20.

Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2010). The role of weibull image statistics in rapid object detection in natural scenes. *J. Vis.*, *10*(7), 992–992.

Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. *J. Neurophysiol.*, *115*(2), 931–946.

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., & Scholte, H. S. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from Low-Level natural image structure to scene category. *Journal of Neuroscience*, *33*(48), 18814–18824.

Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, *7*, e32962.

Groen, I. I. A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V. A., & Scholte, H. S. (2018). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS Computational Biology*, *14*(12), e1006690.

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *372*(1714).

Güçlü, U., & Gerven, M. A. J. van. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlü, U., & Gerven, M. A. J. van. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. *Front. Comput. Neurosci.*, *11*, 7.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2018). Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nat. Rev. Neurosci.*, *9*(6), 467–479.

Hegdé, J. (2008). Time course of visual perception: Coarse-to-fine processing and beyond. *Progress in Neurobiology*, *84*(4), 405–439.

Herzog, M. H., & Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Front. Comput. Neurosci.*, *8*, 135.

Heydt, R. von der. (2015). Figure–ground organization and the emergence of proto-objects in the visual cortex. *Front. Psychol.*, *6*, 10391.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791–804.

Howe, P. D. L. (2017). Natural scenes can be identified as rapidly as individual features. *Atten. Percept. Psychophys.*, *79*(6), 1674–1681.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Jackson, J. H. (1876). *Clinical and physiological researches on the nervous system. I. On the localisation of movements in the brain*.

Jahfari, S., Ridderinkhof, K. R., & Scholte, H. S. (2013). Spatial frequency information modulates response inhibition and decision-making processes. *PLoS One*, *8*(10), e76467.

Jahfari, S., Waldorp, L., Ridderinkhof, K. R., & Scholte, H. S. (2015). Visual information shapes the dynamics of corticobasal ganglia pathways during response selection and inhibition. *J. Cogn. Neurosci.*, 1344–1359.

Jegou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. *European Conference on*

Computer Vision, *5302 LNCS*(PART 1), 304–317.

Jones, E., Oliphant, T., Peterson, P., & Others. (2001). *SciPy: Open source scientific tools for python*.

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *J. Vis.*, *8*(13), 11.1–18.

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations facilitate access to awareness for everyday objects. *Cognition*, *180*, 118–122.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends Cogn. Sci.*

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.*, *22*(6), 974–983.

Katti, H., Peelen, M. V., & Arun, S. P. (2019). Machine vision benefits from human contextual expectations. *Sci. Rep.*, *9*(1), 2112.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, *10*(11), e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci. Rep.*, *6*, 32672.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In *Oxford research encyclopedia of neuroscience*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U. S. A.*, *116*(43), 21854–21863.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Res.*, *46*(11), 1762–1776.

Kleist, K. (1934). *Gehirnpathologie.*

Koivisto, M., Kastrati, G., & Revonsuo, A. (2014). Recurrent processing enhances visual awareness but is not necessary for fast categorization of natural scenes. *J. Cogn. Neurosci.*, *26*(2), 223–231.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *J. Neurosci.*, *31*(7), 2488–2492.

Koivisto, M., & Revonsuo, A. (2010). Event-related brain potential correlates of visual awareness. In *Neuroscience & Biobehavioral Reviews* (Nos. 6; Vol. 34, pp. 922–934).

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, *74*(6), 1114–1124.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modelling biological vision and brain information processing. In *bioRxiv* (p. 029876).

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.*, *12*(4), e1004896.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & others. (2018). CORnet: Modeling the neural mechanisms of core object recognition. *BioRxiv*.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., & Ferrari, V. (2018). *The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale*. http://arxiv.org/abs/1811.00982

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., & others. (2020). The open images dataset v4. *International Journal of Computer Vision*, 1–26.

Lamme, V. a F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, *23*(11), 571–579.

Lamme, V. A. F., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *J. Cogn. Neurosci.*, *14*(7), 1044–1053.

Lauer, T., Cornelissen, T. H., Draschkow, D., Willenbockel, V., & Võ, M. L.-H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, *8*(1), 1–12.

Låg, T. (2005). Category-specific effects in object identification: What is "normal"? *Cortex*, *41*(6), 833–841.

Lewandowsky, M. (1908). Ueber abspaltung des farbensinnes. *European Neurology*, *23*(6), 488–510.

Liao, Q., & Poggio, T. (2016). *Bridging the gaps between residual learning, recurrent neural networks and visual cortex*. http://arxiv.org/abs/1604.03640

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Computer Vision – ECCV 2014*, 740–755.

Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cogn. Neurosci.*, 1–15.

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, *7*, e38105.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024.

Macé, M. J.-M., Joubert, O. R., Nespoulous, J.-L., & Fabre-Thorpe, M. (2009). The time-course of visual categorizations: You spot the animal faster than the bird. *PloS One*, *4*(6), e5927.

Mack, A., Tuma, R., Kahn, S., & Rock, I. (1990). Perceptual grouping and attention. *Bulletin of the Psychonomic Society*, *28*, 500–500.

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, *63*(3), 397–405.

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends Cogn. Sci.*, *20*(11), 843–856.

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: Strategic and incremental information accumulation for scene categorization. *Psychol. Sci.*, *25*(5), 1087–1097.

McKinney, W., & Others. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. In *bioRxiv* (p. 2020.01.08.898288).

Mehta, Z., Newcombe, F., & De Haan, E. (1992). Selective loss of imagery in a case of visual agnosia. *Neuropsychologia*, *30*(7), 645–655.

Mollon, J., Newcombe, F., Polden, P., & Ratcliff, G. (1980). On the presence of three cone mechanisms in a case of total achromatopsia. *Colour Vision Deficiencies*, *5*, 130–135.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *J. Neurosci.*, *32*(7), 2335–2343.

Munneke, J., Brentari, V., & Peelen, M. V. (2013). The influence of scene context on object recognition is independent of attentional focus. *Front. Psychol.*, *4*, 552.

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Res.*, *46*(5), 614–621.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cogn. Psychol.*, *7*(4), 480–494.

Newcombe, F. (1969). *Missile wounds of the brain: A study of psychological deficits*.

Newcombe, F., Young, A. W., & De Haan, E. H. (1989). Prosopagnosia and object agnosia without covert recognition. *Neuropsychologia*, *27*(2), 179–191.

Nielsen, J. M. (1946). *Agnosia, apraxia, aphasia: Their value in cerebral localization*.

Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

Oliva, A. (2005). Gist of the scene. In *Neurobiology of attention* (pp. 251–256). Elsevier.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cogn. Psychol.*, *34*(1), 72–107.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.*, *11*(12), 520–527.

Olmos, A., & Kingdom, F. A. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, *33*(12), 1463–1473.

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, *7*(2), 333–339.

Opelt, A., Pinz, A., Fussenegger, M., & Auer, P. (2006). Generic object recognition with boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, *28*(3), 416–431.

Panis, S., Torfs, K., Gillebert, C. R., Wagemans, J., & Humphreys, G. W. (2017). Neuropsychological evidence for the temporal dynamics of category-specific naming. *Vis. Cogn.*, *25*(1-3), 79–99.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., & others. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 8024–8035.

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol.*, *72*(2), 184–187.

Petro, L. S., Vizioli, L., & Muckli, L. (2014). Contributions of cortical feedback to sensory processing in primary visual cortex. *Front. Psychol.*, *5*, 1223.

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *J. Exp. Psychol.*, *81*(1), 10–15.

Pozzi, I., Bohté, S., & Roelfsema, P. (2018). A biologically plausible learning rule for deep learning in the brain. *arXiv Preprint arXiv:1811.01768*.

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Comput. Biol.*, *15*(5), e1007001.

Ramakrishnan, K., Scholte, H. S., Groen, I. I. A., Smeulders, a W. M., & Ghebreab, S. (2016). *Summary statistics of deep neural network predict temporal dynamics of object recognition*.

Ratcliff, G. (1982). Object recognition: Some deductions from the clinical evidence. *Normality and Pathology in Cognitive Functions*.

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *J. Exp. Psychol. Hum. Percept. Perform.*, *40*(2), 870.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the Two-Choice diffusion model of decision making. *Decision (Wash D C), 2015*.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for Two-Choice decision tasks. *Neural Comput.*, *29*(6), 997–1003.

Rémy, F., Saint-Aubert, L., Bacon-Macé, N., Vayssière, N., Barbeau, E., & Fabre-Thorpe, M. (2013). Object recognition in congruent and incongruent natural scenes: A life-span study. *Vision Res.*, *91*, 36–44.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, *2*(11), 1019–1025.

Roelfsema, P. R., Scholte, H. S., & Spekreijse, H. (1999). Temporal constraints on the grouping of contour segments into spatially extended objects. *Vision Res.*, *39*(8), 1509–1529.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *J. Vis.*, *12*(4).

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and Web-Based tool for image annotation. *Int. J. Comput. Vis.*, *77*(1-3), 157–173.

Sartori, G., Job, R., Miozzo, M., Zago, S., & Marchiori, G. (1993). Category-specific form-knowledge deficit in a patient with herpes simplex virus encephalitis. *J. Clin. Exp. Neuropsychol.*, *15*(2), 280–299.

Scholte, H. S. (2018). Fantastic DNimals and where to find them. In *NeuroImage* (Vol. 180, pp. 112–113).

Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., & Lamme, V. A. F. (2009). Brain responses strongly correlate with weibull image statistics when processing natural images. *J. Vis.*, *9*(4), 29–29.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which artificial neural network for object recognition is most Brain-Like? In *bioRxiv* (p. 407007).

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, *57*, 61.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., & Van Gerven, M. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.

Seibert, D., Yamins, D., Ardila, D., Hong, H., DiCarlo, J. J., & Gardner, J. L. (2016). A performance-optimized model of neural responses across the ventral visual stream. *bioRxiv*, 036475.

Self, M. W., Jeurissen, D., Ham, A. F. van, Vugt, B. van, Poort, J., & Roelfsema, P. R. (2019). The segmentation of Proto-Objects in the monkey primary visual cortex. *Curr. Biol.*, *29*(6), 1019–1029.e4.

Self, M. W., & Roelfsema, P. R. (2014). The neural mechanisms of figure-ground segregation. In *Oxford Handbooks Online*.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*, 399–426.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., & Poggio, T. (2005). A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *Artif. Intell.*, *December*, 1–130.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429.

Shoben, E. J. (1982). Semantic and lexical decisions. In *Handbook of Research Methods in Human Memory and Cognition* (pp. 287–314).

Smits, A. R., Seijdel, N., Scholte, H. S., Heywood, C. A., Kentridge, R. W., & De Haan, E. H. F. (2019). Action blindsight and antipointing in a hemianopic patient. *Neuropsychologia*, *128*, 270–275.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.*, *64*(4), 583–639.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human it well, after training and fitting. *bioRxiv*.

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841.

Sun, H.-M., Simon-Dack, S. L., Gordon, R. D., & Teder, W. A. (2011). Contextual influences on rapid object categorization in natural scenes. *Brain Res.*, *1398*, 40–54.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv Preprint arXiv:1312.6199*.

Tadmor, Y., & Tolhurst, D. J. (2000). Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Res.*, *40*(22), 3145–3157.

Talebi, V., & Baker, C. L., Jr. (2012). Natural versus synthetic stimuli for estimating receptive field models: A comparison of predictive robustness. *J. Neurosci.*, *32*(5), 1560–1576.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., & Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(35), 8835–8840.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2019). Deep learning in spiking neural networks. *Neural Networks*, *111*, 47–63.

The Theano Development Team, Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Snyder, J. B., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., … Zhang, Y. (2016). *Theano: A python framework for fast computation of mathematical expressions*. http://arxiv.org/abs/1605.02688

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412.

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. *Psychol. Rev.*, *113*(4), 766.

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron*, *24*(1), 105–110, 111–125.

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends Cogn. Sci.*, *5*(6), 244–252.

Uttl, B. (2005). Measurement of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science*, *16*(6), 460–467.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *J. Cogn. Neurosci.*, *13*(4), 454–461.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Res.*, *42*(23), 2593–2615.

Vigario, R., Sarela, J., Jousmiki, V., Hamalainen, M., & Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings. In *IEEE Transactions on Biomedical Engineering* (Nos. 5; Vol. 47, pp. 589–593).

Võ, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Curr Opin Psychol*, *29*, 205–210.

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge. Further fractionations and an attempted integration. *Brain*, *110 (Pt 5)*, 1273–1296.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107 (Pt 3)*, 829–854.

Wen, H., Shi, J., Chen, W., & Liu, Z. (2018). Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific Reports*, *8*(1), 1–17.

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, *2017*(14), 36–45.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. *Front. Neuroinform.*, *7*, 14.

Wilbrand, H. (1892). Ein fall von seelenblindheit und hemianopsie mit sectionsbefund. *Deutsche Zeitschrift Für Nervenheilkunde*, *2*(5-6), 361–387.

Wokke, M. E., Sligte, I. G., Steven Scholte, H., & Lamme, V. A. F. (2012). Two critical periods in early visual cortex during figure-ground segregation. *Brain Behav.*, *2*(6), 763–777.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychon. Bull. Rev.*, *1*(2), 202–238.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive*

Sciences, *15*(2), 77–84.

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.*, *5*, 674.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492.

Xu, T., Garrod, O., Scholte, S. H., Ince, R., & Schyns, P. G. (2018). Using psychophysical methods to understand mechanisms of face identification in a deep neural network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(23), 8619–8624.

Young, A. W., Newcombe, F., Hellawell, D., & De Haan, E. (1989). Implicit access to semantic information. *Brain Cogn.*, *11*(2), 186–209.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833.

Zeki, S. (1993). *A vision of the brain*. Blackwell scientific publications.

Zeki, S., McKeefry, D. J., Bartels, A., & Frackowiak, R. S. (1998). Has a new color area been discovered? *Nat. Neurosci.*, *1*(5), 335–336.

Zheng, S., Yuille, A., & Tu, Z. (2010). Detecting object boundaries using low-, mid-, and high-level information. *Comput. Vis. Image Underst.*, *114*(10), 1055–1067.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 487–495). Curran Associates, Inc.

Zihl, J., Cramon, D. von, & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain*, *106 (Pt 2)*, 313–340.

Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Res.*, *50*(20), 2062–2068.

Contributions to the chapters

Chapter 2, published as:

Seijdel, N., Jahfari, S., Groen, I.I.A. & Scholte, H.S. (2020). Low-level image statistics in natural scenes influence perceptual decision-making. *Scientific Reports 10*, 10573. https://doi.org/10.1038/s41598-020-67661-8

- N.S., conceptualization, investigation, formal analysis, software, visualization, writing—original draft, writing—review and editing;
- S.J., conceptualization, formal analysis, software, writing—review and editing;
- I.I.A.G., conceptualization, software, writing—review and editing;
- H.S.S., conceptualization, resources, software, supervision, writing—review and editing.

  All authors contributed to manuscript revision, read and approved the submitted version.

Chapter 3, under review as:

Seijdel, N.*, Loke, J.*, van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., de Haan, E.H.F. & Scholte, H.S. (n.d.). Recurrent processing during object recognition: it depends on the need for scene segmentation

- N.S., conceptualization, investigation, formal analysis, data curation, project administration, software, validation, methodology, visualization, writing—original draft, writing—review and editing;
- J.L., conceptualization, investigation, formal analysis, data curation, project administration, software, validation, methodology, visualization, writing—review and editing;
- R.v.d.K., conceptualization, investigation, data curation, writing—review and editing;
- M.v.d.M., conceptualization, investigation, data curation, writing—review and editing;
- E.Q., conceptualization, investigation, data curation, writing—review and editing;

- S.v.G., conceptualization, writing—review and editing;
- E.H.F.d.H., resources, funding acquisition, supervision, Writing—review and editing.
- H.S.S., conceptualization, resources, software, funding acquisition, supervision, writing—review and editing.

  All authors contributed to manuscript revision, read and approved the submitted version.

Chapter 4, published as:

Seijdel, N., Tsakmakidis, N., de Haan, E.H.F., Bohte, S.M., & Scholte, H.S. (2020). Depth in convolutional neural networks solves scene segmentation. *PLoS Computational Biology, 16*(7), e1008022. https://doi:10.1371/journal.pcbi. 1008022.

- N.S., conceptualization, investigation, formal analysis, data curation, project administration, software, validation, methodology, visualization, writing—original draft, writing—review and editing;
- N.T., conceptualization, investigation, formal analysis, data curation, software, validation, methodology, visualization, writing—original draft, writing—review and editing;
- S.M.B., conceptualization, investigation, project administration, software, resources, methodology, writing—review and editing;
- E.H.F.d.H., resources, investigation, funding acquisition, resoureces, supervision, Writing—review and editing.
- H.S.S., conceptualization, resources, software, methodology, supervision, Writing—review and editing.

  All authors contributed to manuscript revision, read and approved the submitted version.

Chapter 5, under review as:

Seijdel, N., Scholte, H.S., & de Haan, E.H.F. Visual features drive the category-specific impairments on categorization tasks in a patient with object agnosia

- N.S., conceptualization, investigation, formal analysis, data curation, project administration, software, validation, methodology, visualization, writing—original draft, writing—review and editing;
- H.S.S., conceptualization, resources, investigation, software, supervision, Writing—review and editing.

- E.H.F.d.H., conceptualization, resources, investigation, funding acquisition, resources, supervision, writing—review and editing.

  All authors contributed to manuscript revision, read and approved the submitted version.

List of Publications

Described:

Chapter 2 Seijdel, N., Jahfari, S., Groen, I.I.A. & Scholte, H.S. Low-level image statistics in natural scenes influence perceptual decision-making. (2020). *Scientific Reports 10*, 10573. https://doi.org/10.1038/s41598-020-67661-8

Chapter 3 Seijdel, N.*, Loke, J.*, van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., de Haan, E.H.F., & Scholte, H.S. Recurrent processing during object recognition: it depends on the need for scene segmentation. preprint on *BioRxiv*. https://doi.org/10.1101/2020.11.11.377655 [*shared first author; under review*]

Chapter 4 Seijdel, N., Tsakmakidis, N., de Haan, E.H.F., Bohte, S.M., & Scholte, H.S. (2020). Implicit scene segmentation in deeper convolutional neural networks. *PLoS Computational Biology* *16*(7),e1008022 https://doi:10.1371/journal.pcbi.1008022.

Chapter 5 Seijdel, N., Scholte, H.S., & de Haan, E.H.F. Visual features drive the category-specific impairments on categorization tasks in a patient with object agnosia. [*under review*]

Other research

Groen, I.I.A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V.A.F., & Scholte, H.S. (2018). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS computational biology, 14*(12), e1006690.

Smits, A.R., Seijdel, N., Scholte, H.S., Heywood, C.A., Kentridge, R.W., & de Haan, E.H.F. (2018). Action blindsight and anti-pointing in a hemianopic patient. *Neuropsychologia 128*, 270-275.

de Haan, E.H.F., Seijdel, N., Kentridge, R.W., & Heywood, C.A. (2020). Plasticity versus chronicity: stable performance on category fluency 40 years post-onset.

Journal of Neuropsychology 14 (1), 20-27.

Dekkers, L.M.S., Seijdel, N., Weeda, W.D., Jansen, B.R.J., & Huizenga, H.M. Multi-Attribute risky decision making in a sequential context: four Drift Diffusion Model studies [in preparation]

Oosterholt, P., Seijdel, N., & Scholte, H.S. The influence of visual diet on object recognition in computer generated scenes [in preparation]

Short CV

Noortje Seydel was born in Amsterdam on April 20th, 1992. Between 2010 and 2013 she obtained a Bachelor's degree in Psychobiology at the University of Amsterdam. After that, she obtained a Research Master's degree in Psychology, with a specialization in Brain and Cognition and Psychological Methods. She wrote her master thesis under supervision of Dr. Sara Jahfari and dr H. Steven Scholte, on the influence of low-level scene complexity on perceptual decision-making and neural responses. At the end of 2015, she started her PhD project on computational modeling of visual processing during object recognition under supervision of Dr. H. Steven Scholte and Prof. Dr. Edward H.F. de Haan. On April 16 2021, Noor will defend her dissertation in the Agnietenkapel of the University of Amsterdam.

Nederlandse samenvatting (Summary in Dutch)

Schijnbaar zonder enige inspanning interpreteren onze hersenen het licht dat op ons netvlies wordt geprojecteerd, en in een oogwenk herkennen we de voorwerpen om ons heen. In tegenstelling tot een camera, die slechts ruwe visuele informatie opslaat die vanuit de fysieke wereld wordt geprojecteerd, *begrijpen* we onze omgeving in termen van levendige visuele kenmerken, structuren en objecten. Deze prestatie is vooral indrukwekkend omdat objectherkenning een rekenintensief proces is. Een enkel voorwerp, bijvoorbeeld een banaan, kan voor een vrijwel oneindig aantal verschillende projecties op ons netvlies zorgen, afhankelijk van vele factoren, zoals het gezichtspunt, de belichting of zelfs de rijpheid. Bovendien kunnen objecten uit dezelfde categorie verschillen in kleur, grootte, textuur en andere kenmerken. Om het nog ingewikkelder te maken zien we maar zelden een object op zichzelf. We zien de wereld in scènes, waarin objecten zijn ingebed en vaak gedeeltelijk verborgen zijn in een rijke en complexe omgeving. Hoe verwerken de hersenen deze visuele informatie en transformeren ze die tot robuuste visuele representaties van objecten en structuren?

We weten dat het visuele systeem hiërarchisch is opgebouwd. Dat betekent dat de visuele informatie van 'vroege visuele gebieden', zoals de primaire visuele cortex, naar hogere of latere visuele gebieden gaat. Hoe verder de informatie in de hiërarchie komt, hoe complexer de wijze is waarop die informatie verwerkt wordt. Door achtereenvolgens steeds complexere kenmerkcombinaties op te bouwen, zou een enkele golf van activiteit door deze gebieden kunnen volstaan om een object te herkennen. Deze golf van activiteit van vroege naar latere visuele gebieden wordt ook wel de *feedforward sweep* genoemd (zie ook Figuur 1 uit de introductie). Na deze feedforward sweep, is het verwerken echter nog niet altijd afgelopen. Terugkerende signalen vanuit de hogere gebieden kunnen de lagere hersengebieden heractiveren via *feedback* verbindingen. Dit wordt ook wel *recurrent processing* genoemd.

Het hoofddoel in dit proefschrift is het onderzoeken in hoeverre het menselijk brein beïnvloed wordt door eigenschappen van onze natuurlijke omgeving tijdens het herkennen van objecten. We hebben met name onderzocht hoe verschillende functionele architecturen of verschillen in informatieverwerking (feedforward of recurrent processing) informatie onttrekken aan objecten en hun achtergronden.

Om dit psychobiologische proces goed te begrijpen hebben we verschillende experimenten uitgevoerd en verschillende technieken gebruikt. Om de complexiteit van een afbeelding te kwantificeren hebben we gebruik gemaakt van modellen die de statistische eigenschappen van plaatjes berekenen (op een manier die het brein ook zou kunnen uitvoeren). Deze plaatjes hebben we vervolgens laten zien in psychologische experimenten, gecombi-neerd met EEG metingen en beslismodellen om de hersenactiviteit en de processen die optreden tijdens visuele verwerking in kaart te brengen. Bovendien hebben we computermodellen gebruikt als 'kunstmatige diermodellen' van het visuele systeem in mensen, om op die manier meer te leren over de berekeningen die ten grondslag liggen aan succesvolle objectherkenning. Lange tijd waren computers niet in staat om objecten net zo goed te herkennen als mensen. Tegenwoordig, gestimuleerd door grotere datasets en toenemende rekenkracht, hebben de vorderingen in kunstmatige neurale netwerken geleid tot 'visuele systemen' die beginnen te concurreren met men-sen. Aangezien het hier om computermodellen gaat, kunnen we de architectuur ge-makkelijk wijzigen, bepaalde gebieden in het netwerk 'beschadigen', of verschillende architecturen met elkaar vergelijken om de mechanismen of berekeningen die tot ob-jectherkenning leiden te evalueren. Naast het manipuleren van de architectuur, kunnen we ook visuele input manipuleren en evalueren hoe verschillende modellen omgaan met variaties in zintuiglijke input, net als in experimenten met menselijke deelnemers.

In *Hoofdstuk 2* onderzochten we of de complexiteit van een plaatje beïnvloedt hoe de hersenen het plaatje verwerken tijdens het nemen van een beslissing. Om dit te onderzoeken vroegen we deelnemers om wisselend zo snel of zo accuraat mogelijk aan te geven of ze een dier herkenden in verschillende scènes met een lage, gemiddelde of hoge complexiteit (gekwantificeerd door twee statistieken). Analyses met behulp van beslismodellen toonden aan dat de snelheid van informatieverwerking werd beïnvloed door de complexiteit van de scène. Afzonderlijke manipulatie van de twee statistieken verfijnde deze waarnemingen door aan te tonen dat de effecten met name te wijden waren aan de mate van coherentie in de scene.

In *Hoofdstuk 3*, evalueerden we of deze gedragseffecten direct gebaseerd waren op de berekening van SC en CE, als een soort algemene maat voor beeldcomplexiteit, of meer indirect, als diagnostische informatie om andere taak-relevante eigenschappen in te schatten. Onze resultaten suggereren het eerste, omdat we laten zien dat hoe objectherkenning wordt opgelost afhangt van de complexiteit van de context, ook als die context geen taak-relevante eigenschappen bevat: voor objecten die geïsoleerd of in 'eenvoudige' omgevingen worden aangeboden lijkt objectherkenning vooral afhankelijk te zijn van het object zelf, wat resulteert in een situatie die waarschijnlijk kan worden opgelost binnen de eerste feed-forward sweep van visuele informatieverwerking. Wanneer de omgeving complexer of chaotischer is, lijkt recurrent processing nodig om de elementen die bij het object horen te groeperen en het object 'uit te lichten' van de achtergrond.

In *Hoofdstuk 4*, onderzochten we de mate waarin object- en contextinformatie wordt gerepresenteerd en gebruikt voor objectherkenning in verschillende kunstmatige neurale netwerken. We laten zien dat architecturen met meerdere lagen van verwerking (d.w.z. een dieper netwerk) of architecturen met feedback connecties beter in staat zijn om een object te scheiden ten opzichte van de achtergrond, op grond van een toenemende selectiviteit voor de relevante kenmerken.

Tenslotte hebben we in *Hoofdstuk 5* onderzocht wat er gebeurt als visuele informatie niet langer betrouwbaar vertaald kan worden naar bestaande conceptuele kennis. In deze studie testten we object- en scèneherkenning in een patiënt met hersenbe-schadigingen. Een opmerkelijke bevinding bij deze patiënt was dat hij een specifieke prestatievermindering vertoonde voor bepaalde categorieën, zoals voor 'levende dingen' ten opzichte van 'niet-levende dingen'. De resultaten in hoofdstuk 5 laten zien dat de categorie-specifieke effecten, althans voor deze patiënt, niet verklaard kunnen worden door een semantische stoornis alleen. Met behulp van de kunstmatige neurale netwerken probeerden we ook hier weer te onderzoeken welk type bere-keningen dergelijk gedrag zou kunnen produceren. Over het algemeen vertoonden de netwerken met 'beschadigingen' in hogere gebieden (en niet vroege gebieden) vergelijkbare reactiepatronen, met verminderde prestaties voor kunstmatige scenes en levende dingen.

Samenvattend toont het onderzoek in dit proefschrift aan dat hoe objectherkenning wordt opgelost in het brein afhangt van de context waarin het object verschijnt: voor objecten gepresenteerd in een eenvoudige omgeving (bijvoorbeeld een vogel in een strakblauwe lucht), kan herkenning waarschijnlijk worden opgelost binnen de eerste *feed-forward sweep* van visuele informatieverwerking, gebaseerd op een ongebonden verzameling van beeldkenmerken. Voor meer complexe scènes of in meer uitdagende situaties, is aanvullende activiteit (in de vorm van in de vorm van recurrente berekeningen) nodig, om de elementen die bij het object horen te groeperen en te scheiden van de drukke achtergrond.

Acknowledgements (dankwoord)

Eindelijk mag ik mijn dankwoord schrijven! Dit proefschrift is niet zomaar tot stand gekomen, en ik wil daarom heel graag aan een heleboel (!) mensen mijn dank betuigen. Allereerst mijn zeer enthousiaste supervisors, Steven en Edward. Steven, tijdens de reis van elfje naar kuiken (en nu?) heb ik ontzettend veel van je geleerd, met name door je optimisme en vertrouwen –of het nou om een nieuwe analyse, presentatie of DJ performance gaat– goede raad (al heb ik "hou je leren jas aan tijdens het lesgeven" in de wind geslagen) en de ruimte om fouten te maken (*you live you learn*). Je combineert wetenschappelijke visie, methodologische kennis en daadkracht met humor, vrijgevigheid en oprechte interesse in de mensen die met je samenwerken. Ik heb ontzettend veel bewondering voor hoe je in het leven staat en hoop dat we elkaar nog blijven zien voor wetenschap, gesprekken, drank en lekker eten (nu al heimwee naar #scholtelife). Edward, jouw oneindige kennis over neuropsychologische experimenten, oog voor eerdere studies en interessante details in resultaten zorgen ervoor dat afspraken met jou altijd voelen als 'echt onderzoek doen'. Vol enthousiasme, puur gedreven door de onderzoeksvraag en zonder praktische bezwaren (die en die hebben nog wel een interessante stimulus set liggen en een experiment is zo geprogrammeerd) blijf je gefocust op de puzzel van het menselijk visueel systeem.

I would also like to thank Iris Groen, Radek Cichy, Cees Snoek, Marcel van Gerven, Victor Lamme and Pieter Roelfsema for their time and commitment to read my thesis.

Mijn proefpersonen, die zich voor de wetenschap hebben laten scannen en meten en (soms extreem saaie) taakjes hebben uitgevoerd: zonder jullie was dit proefschrift er niet geweest! (al is er maar een oorkonde uitgereikt). I would especially like to express my gratitude to MS for his great efforts and diligent cooperation. Your influence on this thesis and my thinking (about the visual system and life) goes well beyond the presented results.

Bij een deel van de projecten heb ik studenten begeleid die ik graag wil bedanken: Lotte, Yannick, Tim, Jill, Nicole, Evert, Laurens, Lucas, Ron, Matthew en Eva. Ontzettend bedankt voor jullie tijd en enthousiasme!

All (former) members of the Scholtelab: Kandan, Max, Jessica, Lynn, Lukas (over wie

later meer) and our extra lab members Jurriaan and Daniel, #Scholtelife is the best life, I will miss working with all of you.
Zero zero five: Anne, Camile, Jan Willem, Thomas, Lynn, Stijn en Lola, ik had me geen betere kamergenoten kunnen wensen! Het was ontzettend fijn om de serieuzere kanten van het (PhD-) leven maar ook zeker de minder serieuze kanten met jullie te delen. Na een jaar thuiswerken kan ik me eigenlijk niet meer voorstellen dat werk ook zoveel meer kan zijn dan alleen achter een computer zitten, van de science salads soups & savoury pies, de koffietrein en bistable disco tot de komst en viering van de leukste office babies. Ik ga onze kamer ontzettend missen.
Martine en Jolien, veel dank voor jullie onmisbare mentoradviezen (zelfs nu nog!). Iris en Sara, jullie hebben me laten zien dat met wie je werkt het allerbelangrijkst is voor het plezier en de motivatie (zonder jullie had ik ons project echt niet zo lang volgehouden). Lieve collega's van B&C en daarbuiten, heel veel dank voor de fijne tijd! Ik heb genoten van de discussies, lunches en alle borrels in Kriterion, CREA en alle andere veldjes of kroegen (ik zal met name de noord-zuidlijn kroegentocht nooit vergeten).
Anouk, Selma, Nils en Nikki, bedankt dat ik af en toe toch een beetje mee mocht doen met jullie gezellige team FAB4V. Leon, Josipa, Valeria, Esperanza, Anderson, Esmee, Tim, Timo, Peter, Ilja, Heleen, Thelma, Romke, Hilde, Annabeth, Joost, Tulsi, Nicolas, Samuel, Ruben, Simon, Johannes, Marte, Yair, Ien, Manon, Hubert, Ingrid, Lilian, Nikos, Sander, Nicole, Jasper en Marcus - dank voor jullie hulp en gezelligheid!

Mijn eerste dagen als onderzoeker begonnen natuurlijk op Spinoza (Ilja, eeuwig dank voor de aanbeveling) met Tinka, Pia, Jennifer, Michelle en mijn paranimfen Lukas en Diane. Tinka, dit traject was onmogelijk (maar vooral een stuk minder leuk) geweest zonder onze wekelijkse wandelingen, koffiepauzes en geheime momenten om stoom af te blazen of successen te vieren. Lukas en Diane, samen met jullie ben ik dit avontuur gestart en ik ben enorm blij en heel dankbaar dat jullie nu naast me staan. Bij jullie kan ik terecht voor vrijwel alles, van serieus advies over wetenschap, analyses en het leven tot lange avonden dansen, drinken, darten en andere rebelactiviteiten (sowieso alles het liefst met een kleine bi). Ik hoop dat we dat nog heel lang blijven doen!

Naast iedereen op en rondom werk wil ik ook graag mijn lieve vrienden en familie bedanken In meerdere groepen die tegenwoordig allemaal overlappen: Nicole (die zelfs naar colleges komt kijken en met wie tien jaar samenwonen tien jaar feest is), Anna, Laura, Annabel, Benthe, Janina, Juul, Josephine, Senna, Annick, Ceciel, Vera, Jacomien, Vie, Nicky, Wiet, Sanne, Lois, Monika (bedankt voor de geweldige zomer in Berlijn), Hester, Emma (jammer dat we niet langer samen op Roeterseiland hebben kunnen rellen), Omi (waf ya), alle Seijdels, Donders, en iedereen die daarbij hoort: bedankt voor alle vakanties, etentjes, wandelingen, feestjes, concerten en jullie blijvende interesse in mijn onderzoek.

Lieve Gijs, jij hebt de afgelopen jaren vooral gezorgd voor een goede balans tussen werk en ontspannen. Ik word iedere dag gelukkig wakker in Piratenango (I) en heb ontzettend veel zin in de toekomst samen. Met jou kan ik alles aan!

Heel veel dank ook voor mijn stoere lieve zusje Merel. Ik kan me geen leven zonder jou voorstellen, en ben zo blij dat onze band alleen maar hechter wordt naarmate we ouder worden. Tenslotte papa en mama! Ik kan me geen betere (en vooral leukere) ouders wensen en ben jullie ontzettend dankbaar voor jullie onvoorwaardelijke steun en liefde. Voor altijd werken werken werken, leren leren leren en.. plezier.