Scene statistics: neural representation of real-world structure in rapid visual perception

Groen, I.I.A.

Publication date
2014

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 1
Introduction

From the moment you wake up and open your eyes, your brain starts to process visual information. Light reflecting from your environment falls onto the eye lens, which projects it onto the retina in the back of the eye. Photoreceptors in the retina translate light intensity into a neural signal, which travels via the optic nerve to the brain (Figure 1.1A). But how does your brain ‘make sense’ of light? More precisely, how does it translate light rays into a visual representation of the outside world?

This fundamental question lies at the heart of visual neuroscience. Decades of research in humans and animals have taught us a great deal about how vision is implemented in the brain. We know that a large chunk of cortex at the back of our brain is dedicated to visual processing. This visual cortex itself exists of many sub-areas that are specialized for different types of visual information. Some areas preferably respond to simple stimuli such as lines or colors; other areas only become active when a specific object is visible, such as a face or a house (Figure 1.1A).

Despite this cortical specialization, we do not experience a fragmented world existing of individual lines and objects. Instead, in our conscious perception, these different types of visual information are integrated into a coherent whole. This suggests that the visual areas do not operate in isolation; in fact, they form an intricate, large-scale network. This dedicated network somehow converts the light rays arriving at the retina into a complete representation of our environment.

Figure 1.1 Research question and approach taken. A) When presented with visual stimulation, the brain translates light reflecting from the image into a neural signal. From the eye, signals travel (red arrows) via sub-cortical structures such as the thalamus (not shown) to the visual cortex. Thalamic nuclei respond to contrast, i.e. light to dark transitions in an image. The first cortical visual area, located in the visual pole at the back of the brain, preferably responds to contrast with a certain orientation. Higher-level areas become responsive to entire objects, and are sometimes even specialized for specific object types such as faces, houses or bodies. It is unclear how all this information is combined in a coherent representation of the entire scene. B) The reported research is situated at the interface of psychology, neuroscience and computer vision (yellow star). We investigated subjective human perception of scenes with behavioral categorization experiments (psychology); we looked into the human brain to examine neural responses to visual scenes using modern neuroimaging techniques (neuroscience); and finally, we quantified information in the scenes by means of computational modeling (computer vision).
An extensive literature exists on the anatomy and functional specialization of this brain network (e.g., Grill-Spector and Malach, 2004; Kravitz et al., 2013). However, understanding vision also requires understanding the computations that are involved in visual processing (Olshausen and Field, 2005; DiCarlo et al., 2012). In other words, we do not only want to know what is represented where in the brain, but also how these visual representations come about.

So how does the brain construct a coherent representation of an entire visual scene? This question has been approached before from multiple scientific disciplines (Figure 1.1B). Experimental psychology has examined our visual experiences during real-world scene perception. Neuroscience has measured how the brain responds to scene-specific information. Finally, computer vision - a branch of informatics that is concerned with the implementation of vision in artificial systems - has developed computational models of visual processing (Poggio and Serre, 2013).

The research presented in this thesis is located at the intersection of these three different fields (Figure 1.1B). We think that to understand how real-world scene representations are formed in the brain, it is necessary to relate experimental neural and behavioral findings to computational models of vision. This is because computational models can quantify the information that drives visual responses and thereby shapes the neural representations that underlie our visual experience.

In particular, this thesis explores the idea that a specific type of visual information known as natural image statistics plays an important role in the formation of visual representations of real-world scenes. Natural image statistics are real-world regularities that might help in representing visual information in an efficient way. One key hypothesis tested in this thesis is that the brain uses these statistical regularities to rapidly recognize and classify real-world scenes. We examined this question by explicitly modeling scene statistics and testing these models against behavioral and brain data.

In this interdisciplinary approach, we thus aimed to relate psychological, neural and computational representations with one another. This is challenging because it involves dealing with different methods. In this thesis, we both developed new, and applied existing methodology that allows this, such as single-trial and dissimilarity analysis on time-resolved EEG data.

Below, I first briefly review previous psychological and neuroscientific findings on real-world scene perception. Secondly, I explain in more detail what natural image statistics are and why they may be important for scene representation in the brain. Thirdly, I explain how computational models can be used to represent scene information, and how we can relate these models to behavioral and neural data. Finally, a brief description of the remaining chapters of this thesis is provided.

Scene perception is fast and effortless
A key feature of scene perception is that it occurs extremely fast. Already in the 1970s - when computers were not yet available - experiments were performed to test
how much information we can extract from a single glance at a scene. Potter (1975) showed series of photos in rapid succession on a projector while observers searched for a particular target picture. They were either shown the target beforehand, or only given a verbal instruction, e.g. ‘search for two men drinking beer’. Even when the photos were only shown for 1/8 of a second, the observers were still able to detect the target pictures with high accuracy. Importantly, they were equally accurate when they were only given the more abstract verbal instruction. This experiment thus showed that even with very brief visual exposure, the brain can process scenes to a level at which it is possible to identify their semantic content or meaning.

This conclusion has now been corroborated by many of these 'rapid serial visual presentation' experiments (Potter, 2012). Nowadays, we can easily replicate this experiment ourselves by simply flipping through the photo album on our laptop or smartphone. Even if you do it very fast, it is surprisingly easy to recognize the content of each picture: for example where it was taken, or objects and people in it. But how does your brain achieve this? In the 1990s, researchers began to examine how much time the brain takes to distinguish different types of scenes. They found that electroencephalography (EEG) measures of visual responses to scenes containing targets - such as animals or vehicles - and no-target scenes already differ at 150 ms after picture presentation (Thorpe et al., 1996; VanRullen and Thorpe, 2001). If subjects only needed to make an eye movement towards the relevant category, this could even be done within 120 ms (Kirchner and Thorpe, 2006).

Our brain thus seems to process scene information in a fast and efficient way: it needs only a single glance at a scene to extract relevant information within a fraction of a second. Another demonstration of this remarkable efficiency is that when observers have to detect objects in two scenes in parallel, EEG responses are just as fast as when they see only one scene at a time (Rousselet et al., 2002). Even more strikingly, when subjects are simultaneously engaged in a different task (with different stimuli), they are still quite good at categorizing animals or vehicles in scenes presented in another part of their visual field (Li et al., 2002).

Importantly, in the last study, good performance was only found when the stimuli in the second task were scenes: it did not occur for – seemingly - more simple discrimination of letters or disks. Together, these results have led researchers to believe that natural scenes are a 'superior' type of stimulus that the brain might process in a specialized manner (Braun, 2003). For example, the scene may serve as a context that facilitates rapid object detection (Bar, 2004; Oliva and Torralba, 2007). But what makes scenes so easy to process? Intuitively, we might think that because it can consist of infinitely many features, a scene is visually complex. On the other hand, because of its complexity, it also has more information that the brain can exploit. In particular, low-level cues could aid in forming a rapid, first impression of a scene, which is often referred to as the gist of the scene (Schyns and Oliva, 1994; Oliva, 2005).
A role for natural image statistics?

How might the brain extract scene gist? By contrasting neural activity to scenes versus single objects, neuroimaging and physiological studies have found scene-selective areas within visual cortex (Epstein, 2005; Dilks et al., 2013). These areas appear to be particularly sensitive to the fact that scenes (unlike single objects) have a spatial layout (Epstein and Kanwisher, 1998; Kravitz et al., 2011; Park et al., 2011). In addition, it is possible to uncover a handful of scene categories based on activity in these regions, e.g. whether the scene depicts a forest or a beach (Walther et al., 2009). This categorical information may in fact be related to differences in spatial layout between scene types. For example, the layout of forest scenes often consists of closed vertical contours, whereas beaches expand horizontally. In their ‘Spatial Envelope’ model, Oliva and Torralba proposed that such properties might be very important for scene gist representation (Oliva and Torralba, 2001).

For these properties to be useful for rapid scene understanding, it should not take very long for the brain to extract them from the visual input. Behavioral experiments have shown that basic-level scene categorization - into beaches, forests, etc. - is equally fast as object categorization (Rousselet et al., 2005b). Interestingly, the extraction of global structural properties of the scenes appears to be even faster: observers require less exposure time to decide whether a scene has an 'open' layout, or whether it is 'natural', than whether it is a city or a forest scene (Greene and Oliva, 2009a), and they also respond faster to such global structural properties in a go-no go task (Joubert et al., 2007). Other behavioral paradigms recently also showed that categorization of global structural properties precedes that of scene category (Loschky and Larson, 2010; Kadar and Ben-Shahar, 2012).

But how can the brain rapidly extract global scene properties without first performing a detailed analysis of its constituent elements - or even its category? This is where natural image statistics come into play. The structural properties of scenes are governed by regularities in low-level properties of the visual environment. For example, forests tend to have higher spatial frequency content (Torralba and Oliva, 2003), whereas beaches or city scenes are dominated by low spatial frequencies (Figure 1.2A). Similarly, the distribution of local contrast values differs between scene categories (Brady and Field, 2000; Tadmor and Tolhurst, 2000; Scholte et al., 2009). Cluttered natural scenes (such as forests) have more Gaussian (bell-shaped) distributions compared to sparse man-made scenes that contain one or a few salient objects (such as a building) (Figure 1.2B). Perhaps the brain makes use of these statistical properties of the visual environment in order to form an initial structural impression of a visual scene.

Surprisingly, very little research has been done on the question of whether and how the visual system of the brain encodes scene statistics. Theoretical work shows that real-world low-level regularities - such as the range of contrast levels - may have shaped the development of response properties of individual visual neurons, e.g., their receptive field shape and their tuning preferences (Olshausen
and Field, 1996a; Simoncelli, 2003; Geisler, 2008). This type of neural adaptation may have occurred because it increases the efficiency of information transmission by individual neurons (Laughlin, 1981; Vinje and Gallant, 2000, 2002). It is less clear, however, how these low-level regularities affect rapid recognition of entire natural scenes in humans. Some behavioral effects of scene statistics on rapid categorization have been demonstrated (Einhäuser et al., 2006; Wichmann et al., 2006; Kaping et al., 2007; Loschky and Larson, 2008; Joubert et al., 2009), but how they affect representations at the neural level is not well known.

**Figure 1.2** Low-level regularities in natural scenes. **A)** Scene categories differ in terms of their spatial frequency content. Compared to the city scene, the forest scene contains more high spatial frequencies. This is reflected in higher power for these frequencies in 2D power spectra derived from a Fourier transformation of the scene (right; average over 6000 images). The $f_x$ and $f_y$ axes represent the frequencies at each possible orientation in the image; c/p is cycles per pixel. The contour lines represent 50% (inner contour) and 80% (outer contour) of the total spectral energy ("spectral signatures"; copied from Torralba and Oliva, 2003, Figure 2; see also Baddeley, 1996). **B)** A similar difference between these two types of scenes is reflected in their contrast distribution, which indicates how often strong and weak contrasts are present across the scene. The distribution of forest-like scenes contains many contrasts of intermediate strength, and therefore has a Gaussian shape. Urban-like scenes have fewer, but stronger contrasts, resulting in a sparse, power-law shaped distribution.

**Modeling visual information**

In order to address the question of whether the human brain is sensitive to scene statistics during rapid scene categorization, it is necessary to consider how the visual system might compute scene statistics. In the Spatial Envelope model, spatial frequency regularities are exploited by organizing scenes based on principal components of their power spectra (Oliva and Torralba, 2001). These components can be mapped onto many of the global properties discussed above, such as 'openness', 'naturalness' and 'depth' (Greene and Oliva, 2009b). However, it is unclear how the computation of these components may be implemented in the brain.

Given that the visual cortex samples information locally, it is in fact unlikely that it can perform a computation akin to a whole-scene Fourier transformation that is necessary to derive the principal components (Graham, 1979; Field, 1987). Biological models of visual processing commonly start with conversion of the light intensities arriving at the retina to black vs. white (i.e., contrast) responses from neurons with
small, local receptive fields (Hubel and Wiesel, 1968). Each neuron thus only 'sees' a small part of the visual scene, effectively responding to local contrast present in that part of the scene. At further stages of processing, these responses are compressed and normalized (Heeger et al., 1996) before they are spatially combined in higher visual areas (Serre et al., 2007). Contrast distributions such as those depicted in Figure 1.2B, which summarize local contrast across the whole scene, could thus potentially be computed more easily in the brain than Fourier spectra.

Further interest in contrast distributions comes from computer vision, which showed that not only the distributions of man-made and natural scenes, but of almost all natural images tend to range between power-law and Gaussian. This family of distributions is well described by a so-called Weibull function (Simoncelli, 1999). The Weibull regularity is a consequence of physical differences in fragmentation between scenes (Geusebroek and Smeulders, 2003). If a scene consists of many independent parts, there is necessarily a wide range of local contrast strengths present (Figure 1.3A). If, on the other hand, the scene consists of a single object or space, contrast is restricted to a particular range (Geusebroek and Smeulders, 2002). Thus, it appears that the Weibull function well characterizes the dynamic range of initial visual responses when viewing natural images.

In previous work, Scholte et al., (2009) showed that estimating these characteristics of the contrast distribution of a scene does not require a whole scene transformation, or even a Weibull fit. Instead, it can be done by simply summatating the

![Image of Figure 1.3](image_url)

**Figure 1.3** Computing and approximating contrast distribution statistics. A) The shape and scale of contrast distributions of natural scenes can be summarized in two parameters derived from a Weibull fit ($\beta$ and $\gamma$). Hypothetical parameter values are shown below the Weibull fits of the distributions for illustration. B) Schematic of a biologically realistic model that approximates the two Weibull parameters by means of spatial summation. CE is the approximation of $\beta$; SC is the approximation of $\gamma$.  

12
contrast responses at the first relay stage in visual processing after the retina (see Figure 1.1A), the lateral geniculate nucleus (LGN). By simulating these local responses to individual scenes, it is possible to derive two summary parameters that strongly correlate with the scale ($\beta$) and shape ($\gamma$) parameters of the Weibull function (Figure 1.3B).

The scale parameter describes the average contrast strength in the scene, whereas the shape parameter describes to what degree the contrast distribution resembles a power law or a Gaussian distribution. Scenes with high $\beta$ values often contain strong figure-ground segregation (e.g. because of the presence of an object with sharp edges), whereas images with high $\gamma$ values are cluttered or textured (Figure 1.4). Throughout this thesis, we therefore refer to the approximation of the $\beta$-parameter as contrast energy (CE), and to the approximation of the $\gamma$-parameter as spatial coherence (SC). Together, these two approximations provide a potential biologically plausible substrate for the computation of one particular scene property, the distribution of local contrast (Ghebreab et al., 2009).

The main focus of this thesis is to study to what extent the human brain makes use of these particular scene statistics in real-world visual processing. However, we also compared this model to other scene statistics to examine whether its biological plausibility in fact leads to better prediction of brain responses during rapid scene categorization.

**Figure 1.4** In the parameter space derived from the model in Figure 1.3B, scenes with a high amount of clutter/texture are located on the right, whereas scenes with single objects are found on the left. Scenes with strong figure-ground segregation/depth are at the top, whereas flat scenes are at the bottom. The images are from various online databases; see Methods and Materials of Chapter 4 on page 82.
Testing information against the brain

Our aim to relate scene statistics to the brain requires us to go beyond commonly applied methods in cognitive neuroscience. These (mass) univariate methods traditionally compare neural activity to a small number of stimulus categories (e.g., scenes with and without animals), to test where or when in the brain they are differently represented. However, when using real-world stimuli, this approach becomes problematic because it is more difficult to fully control the stimuli. For example, animal scenes will contain other things besides animals; perhaps they will have more bushes in the background. These bushes might potentially also contribute to an overall difference between animal and non-animal scenes. When testing real-world stimuli, we should thus not simply examine scene selectivity (where/when is the difference?), but scene information (what is driving the difference?).

By modeling statistical parameters for each individual scene, we have a more specific, information-based hypothesis: if the brain is indeed sensitive to these statistics, their values should predict the magnitude of differences in neural responses between these individual scenes. In other words, scenes that have almost the same scene statistics should be 'similar' in terms of their brain responses, whereas scenes with very different scene statistics should be 'dissimilar'.

One way to test this is by using linear regression of single-image evoked neural activity on scene statistics (Scholte et al., 2009). Instead of delivering conclusions such as 'the difference occurs at 150 ms' or 'this area is more active for animals than non-animals', this analysis gives information about model fits: how much of the variance between scenes is captured by scene statistics? Which model performs best? In addition, obtained regression weights between image properties and neural activity can be inverted to 'decode' which scene was seen by participants on a given trial (Kay et al., 2008; Ghebreab et al., 2009). In this thesis, we used regression analysis on single-stimulus EEG activity (Figure 1.5A) to compare neural sensitivity to different scene statistics. In addition, we examined the time course of the obtained regression weights in order to study when in visual processing scene statistics affected neural activity.

Another elegant way to relate models to neural data is representational similarity analysis (RSA; Kriegeskorte et al., 2008). RSA translates the differences in neural activity between individual stimuli into a representational space. Scenes that give rise to very different brain responses will be far away in the space, whereas scenes with similar responses are near one another (Figure 1.5B). In this thesis, we used representational spaces to examine to what degree brain activity during scene perception was shaped by scene statistics, using RSA for the first time on EEG data (Figure 1.5C). With RSA, differences in neural activity can be directly compared to differences in scene statistics, by matching a model representational space to the neural representational space (Figure 1.5D). Moreover this allows for evaluation of more complicated representational organizations than traditional univariate methods or decoding analyses (Kriegeskorte and Kievit, 2013).
Figure 1.5 Neural information-mapping techniques used in this thesis. A) In single-trial regression analysis, a time course is obtained for each individual scene, which reflects the amplitude of activity evoked by the stimulus (sERP). We then test the effect of scene statistics on sERP amplitude using a regression model that relates the time course to the scene parameters (CE and SC) for each time-point separately. As a result, we obtain a measure of explained variance ($r^2$) over time, as well as regression weights for each parameter (not shown; here, separate weights were obtained for two repetitions of the same image, $p_1$ and $p_2$, which were modeled as separate columns). In this way, we can examine at what moment in visual processing of information contained in the regressors starts to play a role in visual processing. B) Scene statistics can be used to construct a representational space in which each scene is a point in the 2-dimensional space formed by the scene parameters CE and SC. Clustering-by-category in the space indicates that the statistics are diagnostic of category. C) Just as with the scene statistics, differences in evoked activity can be used to construct a representational space, in which each point of measurement (in this case, EEG electrodes) is a dimension of the space. This multi-dimensional space can be summarized in a dissimilarity matrix, which represents the dissimilarity between each scene and every other scene by means of color-coding (red = dissimilar; blue = similar). Again, this measure can be derived for each time point separately, such that emergence of representational structure over time can be observed. D) The neural dissimilarity matrix (left) can be directly compared with a distance matrix (right) in scene statistics built from the space shown in B.

Outline of this thesis

The first three experimental chapters examine how single-trial EEG responses are related to scene statistics. Throughout these three chapters, there are two parallel developments. First, there is a general increase in complexity of the visual stimulation: we start out with naturalistic abstractions of scenes, after which we move to textures and then to real-world photos. Second, our scene statistics become more biologically plausible: the Weibull fit is still used to obtain the statistical parameters in the first chapter, whereas the approximations based on spatial summation (i.e., the CE and SC values) are used in later chapters.
In Chapter 2, we compare different scene statistics to test which best describes perceptual similarity (experienced by human observers) of abstract images called *dead leaves*. The results indicate that our Weibull model performs better than alternative statistics such as Fourier components: it explains more variance in the single-stimulus evoked EEG and correlates more strongly with categorization errors observed in an independent behavioral categorization experiment. In addition, using RSA we show that images with similar scene statistics give rise to similar neural responses and are perceived as more similar when viewed side by side.

In Chapter 3, we extend this conclusion to real-world textures, by showing that texture materials whose scene statistics are highly constant under different viewing conditions give rise to more 'stable' neural representations (as measured with EEG) and are less easily confused with other texture materials in two different behavioral categorization paradigms. We also rule out a potentially confounding contribution of luminance differences in perceived similarity of the textures.

In Chapter 4, we show that neural categorization of one particular global scene property - naturalness - is accompanied by sensitivity to scene statistics. We find extensive modulations of EEG activity by CE and SC when subjects perform a man-made vs. natural scene categorization task. Again we find that our model outperforms Fourier components in terms of predicting brain activity and behavioral categorization of natural scenes. In addition, we applied single-trial decoding to show that the amount of evidence for a 'natural' decision is affected by scene statistics, in particular SC, which correlates with naturalness because natural scenes tend be more chaotic (less spatially coherent) than man-made scenes.

Having established that scene statistics indeed shape neural representations in scene perception, we change our focus to consider potential top-down influences on neural processing of this information in the remaining chapters. In Chapter 5, we examine whether task instruction affects the previously observed neural sensitivity to scene statistics during naturalness categorization. From the spatiotemporal dynamics observed in two different EEG experiments, we conclude that at the initial stages of visual processing, sensitivity to this information is automatic, i.e. it is not affected by task manipulations. In contrast, later in visual processing, this information is flexibly maintained, possibly because it serves a role in categorization of naturalness.

In Chapter 6, we examined how scene statistics affect the detection of objects in the scenes. We show, using fMRI in conjunction with separate EEG recordings, that scene statistics affect the amount of feedback activity that the visual cortex employs to successfully detect an animal in a scene. This suggests that the scene structure - as captured by scene statistics - does not only shape neural scene representations, but also affects other fundamental visual tasks.

Finally, in Chapter 7, we summarize the main findings of this thesis and discuss their implications for our understanding of scene representation in the human brain.