UvA-DARE (Digital Academic Repository)

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):
Groen, I. I. A. (2014). *Scene statistics: neural representation of real-world structure in rapid visual perception*. [Thesis, fully internal, Universiteit van Amsterdam].

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.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:29 Feb 2024

Chapter 7
Summary and Discussion

In this thesis, I examined neural representation of real-world structure as described by natural scene statistics. My experiments focused on the role of two summary parameters derived from local image contrast, which capture statistical regularities in the contrast distribution of real-world scenes. In the first part of the thesis, I examined the role of this information in the neural representations of perceptual similarity *(Chapter 2)*, texture invariance *(Chapter 3)* and the global scene property naturalness *(Chapter 4)*. In the second part, I studied to what extent neural sensitivity to these statistics was modulated by top-down task instructions *(Chapter 5)* and whether scene statistics affected object recognition in scenes *(Chapter 6)*. My results suggest that natural image statistics contribute substantially to the formation of initial visual representations of real-world scenes and that the visual system possibly exploits this information in order to optimize their processing.

Motivation and main findings

My initial motivation to study the role of scene statistics in visual processing came from the findings by Scholte et al., (2009) and Ghebreab et al., (2009), who found that visual activity evoked by natural scenes was well described by scene statistics. This was an intriguing finding because it suggested that the brain is adapted to statistical regularities in real-world scenes, resulting from the physical rules that govern our environment (Geusebroek and Smeulders, 2002), specifically those that influence the degree of fragmentation in scenes (Geusebroek and Smeulders, 2003). We interpreted the finding as showing that the visual system is strongly tuned to this statistical regularity, and we asked whether this tuning served a purpose in visual processing. This thesis is centered on this question: what information do scene statistics carry and how is this information used in visual perception?

In particular, we wondered whether these statistics could perhaps play a part in rapid perception of natural scenes which has puzzled researchers for decades (Potter, 1975; Intraub, 1981; Thorpe et al., 1996; Fei-Fei et al., 2007b). It is largely still unclear how the brain interprets these "tremendously complex stimuli containing different sources of information, including edges, surfaces, textures, local objects and spatial layout" (Mullin and Steeves, 2013). Had we maybe discovered a 'trick' that the brain plays in order to quickly estimate scene content?

In order to test this hypothesis, we addressed a number of issues in this thesis. First of all, we argued that any scene statistic under consideration should be biologically plausible, in the sense that the human brain, not just a computer program, should be capable of computing it. The scene statistics we studied characterize the histogram of local contrast (i.e., local differences in the luminance values between pixels) in scenes. The histogram summarizes the range of contrasts strengths present in the entire scene. In the previous papers (Ghebreab et al., 2009;

Scholte et al., 2009), this histogram was characterized by fitting a Weibull function to it. This gave us two summary parameters: one indicated the width of the histogram (the β-parameter) and the other the shape of the histogram (the degree to which it looks like a Gaussian or a power-law, as reflected in the γ- parameter). However, although local contrast has a clear biological validity (see *Chapter 1*), the characterization of this information by means of a Weibull fit does not, as it is unclear how a neural system might perform such a mathematical fitting operation. In this thesis, we thus improved our model by using physiologically plausible approximations of the β- and γ- parameters. Such an approximation can for example be derived by summing the responses of local contrast detectors found in the X- and Y-type ganglion cells of the retina or the magno- and parvocellular layers of the LGN (Scholte et al., 2009). In the thesis, I referred to these approximations as contrast energy (CE) and spatial coherence (SC).

This nomenclature was based on the findings of *Chapter 2*, which is the only chapter in which we still used the fitted β- and γ-parameters. In this chapter, we found that computer-generated images filled with disks (*dead leaves)* that were manipulated in terms of occlusion, spatial distribution and size, clustered in the Weibull space described by the β- and γ-parameters in a particular way (Figure 2.4A). Images with occluding disks, which thus had high contrast borders, had high β-parameters. This parameter measures the mean strength of the local contrast and reflects the overall energy of the scene elements, and we therefore referred to the approximation of this property as CE. We also noted that images for which the edges belonged to a single, large disk had low γ-values, whereas highly cluttered images with many edges had high γ- values. We thus chose to describe this property as the relative presence of spatially coherent edges, i.e. SC (note that this parameter also has a formal interpretation as spatial coherence because it describes the correlation between local contrasts, which is low for cluttered scenes and high for single-object scenes).

A second issue we addressed was whether scene statistics could predict not only neural sensitivity during visual processing, but also the behavioral outcome of the visual process, i.e. perceptual experience. In *Chapter 2* and *3,* we thus examined how scene statistics affected EEG signals *and* how perceptual image similarity mapped onto differences in scene statistics. Moreover, since previous literature had suggested a particular role for scene statistics in the formation of an initial, global impression of scenes (scene gist; Torralba and Oliva, 2003; Oliva and Torralba, 2006), we examined in *Chapter 4* to what degree the scene properties described by the scene statistics were predictive of global information that subjects can rapidly extract from an image. We again presented both EEG evidence and behavioral measures to study a potential role of scene statistics in the capability of human subjects to judge the global property of scene naturalness, and also integrated these two sources of data in a single-trial decoding analysis of behavior from EEG data.

Third, when testing models of information against the brain and behavior, it is important to compare results to other models in order to establish a baseline measure and to test how different elements of the model contribute to its explanatory power. Thus in *Chapters 2, 3* and *4* we compared our results with other statistical measures in the same scenes such as spatial frequency and luminance distributions, by for example computing unique explained variance values for each type of scene statistic or comparing representational spaces based on different scene statistics.

Finally, in the second part of the thesis, we shifted our focus away from establishing the contribution of scene statistics to perceptual similarity and gist perception towards the interaction of this information with task requirements. The motivation for this was that information encoding in the brain is not a purely bottom-up process: it interacts with top-down factors that originate in higher-level parts of the brain. In *Chapter 5*, we thus again used EEG to examine how neural sensitivity to scene statistics was affected by the requirement to categorize the scenes on global information. In *Chapter 6*, rather than focusing on the global category or overall similarity of the scenes, we considered a fundamental task that we employ on a daily basis: recognition of objects in scenes. In particular, we were interested in how the global scene representation driven by the scene statistics would affect the ease with which the objects could be categorized.

Below, I discuss the results from each chapter in more detail, after which I go into the broader context of the research and questions raised by our findings.

Part 1: Understanding the information contained in scene statistics

Perceptual similarity between images is captured by scene statistics

In the first study, we examined more carefully what type of information the scene statistics carry and whether their previously observed effects on evoked activity indeed reflected a role for scene statistics in visual perception. To this end, in *Chapter 2* we performed an EEG experiment with naturalistic images. As explained above, the appearance of the images was manipulated such that they formed different categories. Using dissimilarity analysis, we showed that the organization of the categories in the Weibull parameter space (Figure 2.4A) predicted categorical ordering in both neural activity and in behavioral perception.

In this chapter, we also directly compared our results to two other types of scene statistics that contained related information to the Weibull statistics and, importantly, that had been proposed before to play a role in human visual perception (McCotter et al., 2005; Oliva and Torralba, 2006; Kaping et al., 2007) or in neural coding (Brady and Field, 2000; Tadmor and Tolhurst, 2000). We found that that the Weibull statistics explained more variance in neural responses than other scene statistics, and importantly, best predicted the errors that subjects made during categorization. This prediction was not perfect, suggesting that scene statistics do not explain the perceived similarity entirely. Nevertheless, the results suggested that scene statistics in fact may have functional role in visual scene perception.

However, the deadleaves images were quite unusual stimuli in the sense that even though their low-level properties are similar to natural scenes (Hsaio and Millane, 2005), they do not really look like scenes. One could thus argue that low-level information is all subjects could 'work with', and that our behavioral effects might not generalize to more meaningful stimuli. In *Chapter 3*, we thus improved the ecological validity by using textures, i.e. photographs of real-world materials such as wool, bread and sand. Since our previous study had shown that the overlap of stimulus categories in scene statistics space was predictive of neural and perceived similarity, we here manipulated the statistics of the textures by selecting stimulus categories based on their variance in this space (now using the approximations CE and SC of the fitted Weibull parameters for the first time). Again we observed that the categorical ordering predicted by the Weibull space very nicely mapped onto evoked activity and behavioral categorization, also for these more natural images.

In *Chapter 3*, we also first started to consider not only the overall dissimilarity as predicted by the two scene parameters combined, but also potential *differences* in the influence of CE and SC on the time course of the evoked response. As in the subsequent chapters, we found that CE mainly affected evoked activity early in processing (100-150 ms). SC, on the other hand, also reliably affected evoked activity at later time points, and its effects even extended more 'post-perceptual stages' of visual processing (Luck et al., 2000) such as perceptual decision-making (Philiastides and Sajda, 2006). This again led us to believe that the information captured by scene statistics - in particular by SC - could play a functional role in the formation of neural representations.

Do scene statistics contribute to rapid gist perception?

In *Chapter 4,* we finally turned to real-world scenes in order to examine whether the scene statistics that we now had studied in more detail were involved in rapid scene perception. This time, we chose not to use distinct categories or to select certain scenes beforehand, but to examine for a large set of natural scenes how the categorical distinction between man-made and natural environments was related to differences in CE and SC. Previous behavioral studies (Joubert et al., 2007; Greene and Oliva, 2009a; Loschky and Larson, 2010; Kadar and Ben-Shahar, 2012) had indicated that naturalness is a global scene property that the brain potentially picks up very early in visual processing, and it was also suggested that statistical regularities might aid in this computation (Oliva and Torralba, 2001; Torralba and Oliva, 2003). Given the early effects of CE and SC on evoked neural activity, it seemed plausible that they would contribute to this man-made/natural distinction.

Our behavioral results indicated that CE and SC were indeed correlated with naturalness, and importantly, that scenes with intermediate SC vales were more often difficult to categorize as either man-made or natural. In the evoked activity, we again observed a similar pattern as in *Chapter 3*: early effects for CE and later effects for SC. Interestingly, whereas the early CE effects were very *transient,* basically disappearing beyond 150 ms, the late SC effects were more sustained, lasting up to 300 ms. The behavioral and neural results came together in the single-trial decoding analysis (see Figure 4.6), in which we tested how well the behavioral naturalness rating for each scene could be predicted based on the ERP data. Our results showed that scenes with higher SC values lead to relatively stronger neural evidence for the behavioral response that the scene was natural.

Does this mean that SC is 'the neural correlate of naturalness'? As explained in *Chapter 4*, we do not propose this to be the case. Rather, we think that the man-made/natural distinction covaries with SC because SC is informative about the degree of organization in scenes. Natural scenes tend to be less organized because there are no humans that bring order into the scene by organizing it into roads, buildings, living rooms etc. The SC axis of our scene statistics space (see Figure 4.1B) describes a gradient from more chaotic to more organized scenes, which thus also contains information about natural vs. man-made environments. However, careful inspection of the behavioral correlations of CE/SC with naturalness rating (Figure 4.2C) shows that there are in fact natural scenes that are reliably rated as natural but yet have low SC values. It turns out that these are mostly scenes of beaches or empty meadows with clear skies behind them: i.e., natural scenes with a low degree of fragmentation because they simply contain very little 'stuff'. Thus, the semantic naturalness distinction does not map onto the scene statistics one on one. This suggests that further processing, beyond the mere extraction of scene statistics, is still required for this distinction, and likely so for other global categories.

It is therefore important to note that we do not claim that our scene statistics are a 'complete' model of early visual processing that can explain global perception in its entirety. Rather, we think that CE and SC quantify potential sources of information in scenes that the brain may use in order to construct these categorical representations, such as scene fragmentation.

Part 2: Top down factors and task requirements

An interesting question that arose based on our findings in Part 1 is why the degree of fragmentation in a scene is relevant to compute in early visual processing. A potential explanation for this is put forward in the remaining chapters, in which we related our effects to existing theoretical frameworks that suggest that visual processing occurs in a coarse-to-fine manner (Lamme and Roelfsema, 2000; Hochstein and Ahissar, 2002; Rousselet et al., 2004). According to such theories, visual processing starts with a phase in which information is automatically and rapidly extracted in parallel, leading to an initial representation that lacks visual detail and is thus global in nature. After this, a phase of more detailed, scrutinized analysis follows, which can be modulated by top-down factors such as attention or task requirements. Based on *Chapters 5-6*, we argue that scene statistics may be involved in the formation of this initial global percept *(Chapter 5)* and that they may even affect the degree to which subsequent scene analysis is required *(Chapter 6)*.

Automaticity and task-dependency of neural sensitivity to scene statistics

We hypothesized that if CE and SC indeed contribute to the automatic formation of an initial global percept, neural sensitivity to this information should not be affected by top-down manipulations to the scenes. Moreover, we wondered whether the late sensitivity to SC was also part of an automatic processing phase, or whether this effect could be related to more voluntary processing of information. In *Chapter 5*, we put this hypothesis to the test by manipulating task requirement while subjects viewed the scenes. We again showed participants man-made and natural scenes (a smaller selection from the stimuli using in *Chapter 4*, but with similar distribution of CE and SC values) but at the same time we also presented focal letter or peripheral outline stimuli. In different task blocks subjects either categorized the scenes or the distracter stimuli. The scene-evoked ERPs painted a clear picture: early visual sensitivity to CE and SC was completely unaffected by the task. Similarly, the man-made/natural distinction in grand-average ERP amplitude was not affected either, and its significance strictly seemed to follow the time course of CE and SC.

By contrast, the later effects did depend on the task manipulations, and the results seemed to suggest that neural sensitivity to SC was only enhanced when subjects were actively categorizing the scenes. At least, Experiment 2 of *Chapter 5* showed that the disappearance of the late SC effect was not due to suppression of peripheral stimulation. Interestingly, there was also a clear difference in the topography of neural sensitivity across the EEG recording sites between the early and late effects, in the sense that late task-dependent ERP modulations by SC were much more widespread across the scalp compared to the early, automatic effects. It thus seems that during the initial, parallel processing phase, computation of scene statistics occurs in an obligatory fashion, but that this information can be flexibly 'broadcasted' across the visual system if scenes are attended, possibly because the information aids categorization of the scene on a global property.

Effects of scene statistics on object recognition

Although humans can apparently efficiently categorize global scene properties such as naturalness, we do not perform such a task on a daily basis (at least not consciously). If scene statistics indeed contribute to the formation of the first, global impression of scenes, could they be predictive about the *quality* of this representation, and its use for everyday visual tasks such as object recognition? Given that CE and SC inform about the strength and coherence of edges in a scene (*Chapter 2-3*) *and* are possibly involved in the formation of the initial coarse percept of a scene (*Chapter 4-5*), we hypothesized that the ease of object recognition in scenes would be modulated by these global scene statistics.

Therefore, in *Chapter 6*, we used the CE and SC values as selection criteria to manipulate the coherence of the initial global scene representation. As can already be guessed from the example pictures provided in Figure 6.1B, object recognition is more difficult for scenes with low coherence and high energy (i.e., high CE/SC values). Moreover, our combined fMRI and EEG results provide evidence for the presence of coarse-to-fine processing in the brain by showing that for these scenes only, early visual areas were selectively engaged by means of a feedback signal. Thus, when the initial global scene impression signals the presence of a high CE/SC scene, meaning that it contains chaos and clutter, the visual system has to perform more effortful detailed analysis of the scene (to find the object and segment it from the background), which involves recruitment of information from early visual areas.

The broader picture

Visual perception: from local to global to local

Interestingly, object recognition in natural scenes constitutes a classic problem in computer vision, in which decades of research have been spent on the development of algorithms that could detect an object by tracing its outline and then segmenting it from the background. Despite these efforts, such algorithms never matched human performance in general object recognition tasks. However, when computer vision scientists started to consider models that did not necessarily attempt to segment the object, but rather encode its *context* statistically, a significant jump in object classification performance was made (e.g., Uijlings et al., 2013). In the successful Bag-of-Words approach (BoW; Fei-Fei et al., 2007a; Jégou et al., 2011), scenes are summarized in histograms of local feature elements, which are then statistically characterized and used for classification. This model thus exploits relations between object features and background elements, using their global covariance as a way to classify scenes (as such, it can become possible that a BoW model that detects boats is in fact classifying whether there is a hole in the water).

Similarly, CE and SC effectively summarize local contrast responses as if they would be listed in a histogram, and thus characterize similarity between scenes based on *globally integrated local information*. Our results show that statistical information derived from this local contrast integration can contribute to a global percept of scene fragmentation, and thereby signal the probability that a coherent object is present in the scene, which can subsequently guide further analysis of the local elements. In particular SC might be relevant in this respect because it seems most predictive of the 'objectness' of the images.

Unfortunately, in *Chapter 6* we did not manipulate CE and SC separately, so we cannot claim that it is only SC that drives the feedback representation. In fact, we suspect that the low CE values may have hindered feed-forward object recognition for low CE/SC scenes, because even though these scenes were highly coherent (or mostly empty, similar to the beaches in *Chapter 4*), subjects had more difficulty categorizing these scenes. If we take into account the entire image space in Figure 4.1B, feedback would probably be least required for scenes in the top left of the space, which have high energy and high coherence; most feedback would be required for the scenes in the bottom right, with low energy and low coherence.

However, those types of scenes rarely occur in the real world; as can be seen in all the figures of the scene statistics space provided in this thesis, CE and SC are correlated such that scenes with high SC values tend to have higher CE values. Those scenes that do have high CE and SC values always contain a clear object that almost jumps out of the background (Figure 4.1B); it would be hard to not detect such an object in a recognition task. Extreme examples of such scenes might be cut-outs of stimuli that are often used in object categorization experiments. Conversely, scenes with low CE and SC values are homogenous and textured: scenes with extreme values probably look rather 'cloud-like', like the transparent dead leaves categories in *Chapter 2*. Perhaps, interpretation of such stimuli is more easily modified by feedback-mediated top-down influences, for example when one is trying to 'see a face in the clouds'.

What we can learn from natural scenes

I believe that there are two main strengths to the approach taken in this thesis. First of all, we used stimuli that came from (or were models of) our real visual environment. Most knowledge of the visual system comes from studies that used simplified, artificial stimuli, such as abstract shapes, or cutouts of objects on uniform backgrounds. However, in daily life, the visual system is not confronted with isolated objects, but with a world that exists as scenes: rich, information-dense stimuli, often consisting of a background and multiple objects. Furthermore, findings obtained with simple stimuli do not necessarily generalize to real-world perception (Felsen et al., 2005). Ultimately, we would like to understand how the brain encodes real-world stimuli, rather than idealized stimuli. This thesis directly contributes to this endeavor.

Second, I believe the current work is first to use a parametric, quantitative neuroimaging (i.e., EEG, fMRI) approach to study the contribution of scene statistics in visual processing. We operationalized global scene information with just two parameters, but supported this reduction with a biologically plausible model, and we developed analysis methods (single-trial regression, dissimilarity analysis on EEG) to relate this information to neural processing. This contrasts with previous work in which mostly Fourier information was used to either measure or manipulate global information in scenes. Although the visual system is arguably sensitive to spatial frequency, the global Fourier transformation is a mathematical procedure that cannot be implemented in the brain as such. As was hinted at in the previous paragraph, realistic models of visual processing need to specify how the *local* information from receptive fields is integrated into a *global* presentation (that then potentially feeds back onto local information). Another difference with these previous approaches is that we did not characterize global information by putting semantic labels on certain global dimensions based on conscious perceptual judgement (e.g., openness, naturalness as in Greene and Oliva, 2009b). Instead, we attempted to 'let the brain speak', by examining how clustering of neural signals was related to statistical properties of scenes, and how this in turn affected behavior.

Of course, counterarguments can be made against both these contributions. First, it has been argued that using natural stimuli in visual research is actually not a good idea at all. In 2005, two contrasting perspective articles on this topic were published in Nature Neuroscience, which represent the two extremes of this debate. Opponents of using natural stimuli argue that the lack of control in such stimuli undermines the conclusions that can be drawn from the data: because there are too many inter-correlated sources of information in the scenes, the appropriate stimulus-response relation cannot be established (Rust and Movshon, 2005; "In praise of artifice"). Proponents of natural stimuli however argue that it provides more ecologically valid approach and that we need to stimulate the brain with this information *because* scenes contain complex regularities (Felsen and Dan, 2005, "A natural approach to studying vision"). The research presented in this thesis is clearly in the second camp. There are indeed often correlations between properties of scenes, but we think that the brain may have adapted to these properties. Precisely because "the function of the system is intimately related to the properties of the visual stimuli commonly found in the natural environment" (Felsen and Dan, 2005, page 1643), using natural scene stimuli can reveal unique response properties that are not present for artificial stimuli. Moreover, experiments with natural scenes require less a priori assumptions about which stimulus parameters are relevant. Moreover, explicit quantitative modeling of information in the scenes, as we did in this thesis, can help in elucidating which scene properties are important for perception and *how* they are interrelated.

Second, although our neural information mapping approach revealed substantial sensitivity of evoked activity to image statistics, this does not prove that the brain actually *uses* this information. Our scene statistics clearly covary with interesting properties of natural scenes, but exactly because of this covariance it is difficult to isolate their influence (see above). I believe that if such covariances exist, the brain likely exploits them to optimize processing, rather than regarding low-level regularities as noise. After all, visual processing starts out with encoding of low-level information (such as contrast), from which the brain actively *constructs* a high-level representation (i.e, there is no representation of 'animal' on the retina). The research in this thesis suggests that statistical summaries such as CE and SC might have a part to play in this constructive process. But to show that the brain really computes this information, we (ironically) may need to complement the research with natural scenes with experiments that use more controlled stimuli after all.

For example, in a recent separate study we found that *dead leaves* images that were manipulated to have a certain range of SC values gives rise to parametric modulations of activity in retinotopic areas in posterior LOC, the brain area associated with object recognition (Scholte et al., 2013). This does not immediately suggest that LOC thus responds to SC and not to objects. However, information encoded in these areas could be a first step towards an object representation (and perhaps signal whether more detailed information is necessary for this representation, see *Chapter 6*). Generally, these results encourage us to change our conception of these high-level areas from Platonic 'object detectors' towards 'neural pattern recognizers'. At least, they challenge us to think about *how* the low-level information needs to be transformed into high-level representations.

Outstanding questions

The dynamics of visual processing following the global percept

Over the course of Chapter 2-4, we discovered the importance of the SC parameter. Whereas CE gave rise to a transient effect (which is consistent with neural reports of contrast sensitivity in artificial stimuli), this information may not be as interesting from a neural coding perspective, other than that it can perhaps predict how easily or rapidly a stimulus is processed. For SC, however, we found sustained effects (*Chapter 3-4*), which (at least partly) appear to be task-dependent (*Chapter 5*). Moreover, the temporal profile of the regression weights seems to suggest an oscillatory component in this sensitivity (see Figure 4.3B). It would be very interesting to study the neural mechanisms underlying this sensitivity by looking at oscillatory dynamics using time-frequency analysis. This might also help us understand better how this information might lead to modulations of the induced feedback activity (which could be reflected in specific frequency bands) reported in *Chapter 6*.

Another interesting question for future research is whether the SC of the scene predicts the effects of causal interference in visual processing. We could use TMS to test whether scenes with higher SC values indeed give rise to more feedback activity, as predicted by our results from *Chapter 6*. As in Koivisto et al., (2011), we could selectively disturb visual processing in early visual areas at either early (feed-forward) or late (feedback) time windows. In addition, we can select the scenes such that their SC values are parametrically increasing. If feedback is indeed selectively employed based on the SC value of the scene, there should be no interference of visual object recognition performance due to TMS in late time windows for scenes with low SC values, and strong interference for scenes with high SC values.

Development and flexibility of neural sensitivity to scene statistics

Another interesting question concerns the flexibility of the observed neural sensitivity. We already showed that it can differ depending on task demands (*Chapter 5*), but how malleable is this sensitivity? Is it shaped by evolutionary factors or might it be related to our visual experience, and if so on what timescale? One experiment that might address this is inspired by the popular television show in the Netherlands called *Farmer Seeks Wife*. In our experiment, it would be scientists seeking farmers, in order to examine whether farmers - who have been exposed to different scene environments compared to people living in urban environments - might have different sensitivity to variance in scene statistics. Another possibility to study the development of sensitivity to scene statistics would be to have subjects grow up in a sensory deprived experiment. This would be ethically dubious for humans, but it might be possible in animal studies. To what extent sensitivity to scene statistics is present in other species is a very interesting question in and of itself. However, a way to influence sensitivity to scene statistics in humans could be to use visual adaptation techniques (e.g, as in Kaping et al., 2007, who used Fourier spectra manipulations to bias perception to certain statistical ranges) and examine how prolonged exposure to stimuli with certain scene statistics affect neural responses and sensitivity to scenes with different statistics.

Local vs. global statistics

Finally, because the main focus of this thesis is on the relation between scene statistics and global similarity or global scene category, we always computed CE and SC based on local contrast across the entire scene. However, this information can also be extracted for local patches within the scenes. This type of local scene statistics could for example be relevant for (object) saliency detection and the prediction of eye movements (Yanulevskaya et al., 2011). Moreover, as we suggested in *Chapter 3*, the statistics of texture, as described by the CE and SC values, might help the visual system detect local physical invariances that could potentially serve as reliable building blocks in the construction of the global percept. This does not necessarily have to occur in a single transformation step: the pooling of local structure by means of summary statistics could occur repeatedly across the different stages of the visual hierarchy in order to create a stable and coherent representation of the relevant information (DiCarlo and Cox, 2007; Freeman and Ziemba, 2011). Thus, local patch-based statistics as computational units within hierarchical models are an interesting direction to explore in future research.

Conclusion

In this thesis, I report that the human visual system exhibits strong sensitivity to scene statistics derived from local contrast. I explored how these statistics shape neural representations, investigated the temporal dynamics of this neural sensitivity, and examined how it affects the visual processing of object information. I demonstrated a potential role for scene statistics in perceived similarity of naturalistic textures, scene gist perception and dynamics of object recognition. Overall, the results suggest that visual cortex might compute scene statistics in order to estimate the degree of coherence, i.e. the amount of organized structure vs. chaos that is present in the visual input, and use this information in the process of rapid scene categorization, or to dynamically adjust its processing depending on the task at hand.