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

Groen, I.I.A.

Publication date
2014

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 6

Scene statistics predict neural feedback to low-level visual areas during object categorization in natural scenes

Numerous experiments show that object recognition in natural scenes occurs extremely rapidly, suggesting that only feed-forward visual processing is required for this process. On the other hand, recurrent processing is believed to be involved in many operations relevant for object recognition, such as grouping and figure ground segmentation. Here, we examined whether scene complexity affects the degree to which recurrent activity occurs during object recognition. To this end, we systematically manipulated the complexity of natural scenes by selecting scene stimuli based on two low-level, biologically plausible scene statistics: contrast energy (CE) and spatial coherence (SC). These statistics summarize local edge intensity and higher-order correlations between edges in a scene and are diagnostic of the degree of scene sparseness vs. clutter (e.g. whether it contains a bird against the sky versus a deer in the woods). In the fMRI scanner, subjects categorized animal vs. non-animal scenes that were divided into three conditions: low, medium or high clutter, as defined by CE and SC. Slowed reaction times and increased error rates indicated that categorization was especially difficult for the high clutter scenes. In early visual areas, higher fMRI activity for animal scenes compared to non-animals was found only in the high clutter condition. In contrast, in the lateral occipital complex (LOC) activity was higher for animals than for non-animals regardless of clutter. In the parahippocampal place area (PPA), activity was higher for non-animals, but less so for scenes with high clutter. Separate ERP recordings suggested that the increased early visual activity for high clutter animal scenes is not due to low-level differences between animal and non-animal scenes, but that it reflects visual feedback activity from ~200 ms after stimulus onset. These results suggest that the degree to which feedback is employed during object detection in natural scenes depends on the complexity of the scene as described by CE and SC. For sparse scenes with clear figure-ground segmentation, feed-forward activity may be sufficient, whereas for cluttered scenes, object recognition involves increased feedback.

Introduction

High-level visual areas in the brain are selective for different types of stimuli such as scenes, objects and faces (Grill-Spector and Malach, 2004). These areas exhibit remarkable response properties. For example, neurons in human lateral occipital cortex exhibit selectivity for individual objects as soon as 100 ms after stimulus onset (Liu et al., 2009). Even when objects are embedded in real-world scenes, they are recognized very rapidly, as indicated by ERP markers (Thorpe et al., 1996; VanRullen and Thorpe, 2001). Object recognition thus appears to be a visual task that can be solved by the first wave of visual processing (VanRullen and Thorpe, 2002). This claim is supported by computational modeling results showing that human performance in object categorization in natural scenes can be matched by a feed-forward hierarchical network (Serre et al., 2007).

However, the ventral visual stream (VVS), in which this network is thought to be implemented (Ungerleider and Mishkin, 1982), does not conform to a strict feed-forward hierarchy. The VVS contains long-range connections across hierarchical levels as well as lateral connections at the same level (Kravitz et al., 2013). Moreover, it has long been known that in this network, feedback connections are abundant (Rockland and Pandya, 1979; Salin and Bullier, 1995). Electrophysiological evidence from monkeys indicates that feedback activity is crucial for perception of a figure in a textured background (Lamme, 1995; Zipser et al., 1996). This feedback-mediated figure-ground modulation is thought to group elements of a figure together (Roelfsema, 2006). In particular, feedback may be useful for processing of different levels of detail in visual stimuli (Roelfsema et al., 2000) and integrating these into a conscious, coherent percept (Lamme and Roelfsema, 2000).

Moreover, 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, indicating that recurrent processing between visual areas is in fact necessary for object recognition in natural scenes. What purpose might this processing serve? Possibly, feedback is required when the feed-forward representation is noisy or of low quality. Wyatte et al., (2012) used backward masking and computer simulations to show that interfering with feedback is more disruptive for occluded or degraded objects, and suggested that recurrent activity is necessary for robust recognition if the feed-forward sweep fails to provide this information. Similarly, Koivisto et al., (2013) found that masking (which is thought to selectively interrupt feedback processing; Lamme et al., 2002; Fahrenfort et al., 2007; van Loon et al., 2012) had weaker effects for scenes in which animals were 'easily segregated' compared to scenes with 'more demanding backgrounds'. Thus, in natural scene perception, feedback-mediated figure-ground modulation could be selectively employed depending on the complexity of the bottom-up input.
How can the brain determine the complexity of bottom-up input? In natural scenes, an initial global impression of the scene - its gist (Oliva, 2005) – already contains spatial layout and basic-level category information (Oliva and Torralba, 2006). This information which can be derived from statistical regularities in scenes (Torralba and Oliva, 2003), such as the shape of the distribution of spatial frequencies in an image. We recently showed that statistics derived from local contrast distributions in scenes are informative about scene complexity (Scholte et al., 2009; Ghebreab et al., 2009; Groen et al., 2013). These statistics can be derived in a biologically plausible way from early visual responses by means of two summary parameters that reflect the amount of contrast energy (CE) and spatial coherence (SC) in scenes. Together, CE and SC describe an image space in which simple, sparse scenes with clearly segregated objects are projected on the bottom left, whereas complex, cluttered images containing a large amount of scene fragmentation are on the top right (see Figure 4.1B). Thus, these statistics are informative about whether the scene contains one or a few objects that are clearly distinguishable from the background. As a consequence, they could predict the degree to which feedback is needed for object recognition.

Here, we tested this hypothesis by measuring brain activity to scenes that varied systematically in their CE/SC values. Specifically, we measured fMRI activity while subjects categorized animal vs. non-animal scenes from different parts of the image space. In addition, we obtained event-related potentials (ERPs) from a separate experiment to examine the time-course of differential evoked activity for animal vs. non-animal scenes. If CE and SC affect the degree of recurrent activity, scenes with high CE/SC values should give rise to higher activity in early visual areas compared to scenes from different parts of the scene space. Moreover, this manipulation should be reflected in activity modulations at later time-points in the visual time course, which would indicate a potential contribution of feedback above and beyond feed-forward activity modulations.

**Methods**

**fMRI experiment**

**Subjects**
Twenty-five subjects (7 males, age 19-26 years; mean = 21.6, SD = 1.7) participated in the fMRI experiment. All subjects had normal or corrected-to-normal vision. All participants provided informed consent and were financially compensated for their time. The ethics committee of the University of Amsterdam approved the experiment. One participant, who had a median reaction time of 2 standard deviations above average as well as 7.5% non-responses on GO trials was excluded from further data analysis.
Stimuli
Scenes were selected from a larger set of 4800 scenes used in a separate EEG experiment (Groen et al., 2010; see below under 'EEG experiment'). For each scene, one CE and one SC value was computed by simulating the output of contrast-sensitive receptive fields and integrating these responses across the scene by averaging (CE) and divisive normalization (SC). In natural scenes, CE and SC typically correlate highly with parameters of a Weibull function fitted to the distribution of contrast values. CE is a biologically realistic approximation of the distribution width (the scale parameter of the function), whereas SC is an approximation of its shape (the degree to which the function describes a power law or a Gaussian distribution). These two statistics thus capture information about the overall presence of edges in an image (CE) and higher-order correlations between them (SC). As a result, images with high CE values contain strong figure-ground segmentation, whereas images with high SC values are cluttered or textured. The details of the model that implements this computation are described in Chapters 3 and 4 and in Ghebreab et al., (2009).

Here, we used the CE and SC values to selectively sample scenes with low or high complexity. We created three conditions: LOW, MEDIUM and HIGH (Figure 6.1A), whereby each condition was defined by its CE/SC values. Each condition consisted of 160 images, half of which contained an animal. Importantly, within condition animal and non-animal images were matched in their CE and SC values such that these two categories did not differ from each other in their mean (all t(158) < 0.13, all p > 0.89) or median values (Wilcoxon rank sum test all z < 0.16, all p > 0.87). Images were randomly selected from the larger set of scenes solely based on their image statistics and annotations. Subsequent inspection revealed that the animal images contained a wide variety of animals including pets such as dogs and cats but also wildlife, reptiles and fish. Non-animal images consisted of urban, landscape and indoor scenes and also contained a variety of objects, ranging from vehicles to household items. Several exemplars from each condition are provided in Figure 6.1B. Here, we can see clearly that selecting images on image statistics results in a selection of cluttered or textured scenes in the HIGH condition and to more sparse, easily segmentable scenes in the LOW and MEDIUM condition.

Experimental procedure
In the fMRI experiment, we used an animal vs. non-animal categorization in a stop signal paradigm (Figure 6.1C). Subjects performed 480 trials in total, divided over 2 separate runs. Each trial lasted 2000 ms and started with a fixation cross of variable duration (500-750 ms jittered with 50 ms intervals), after which a scene was presented for 100 ms. Scenes were 640x480 pixels and were back-projected on a 61x36 cm LCD screen using Presentation (Neurobehavioral Systems, Albany, CA, USA) that was viewed through a mirror attached to the head coil at ~120 cm viewing distance.
There were two trial types: GO and STOP trials. On GO trials, subjects had to indicate whether the stimulus was an animal or non-animal scene before the trial ended (i.e. at maximum within 1250 ms). Subjects indicated their response using a hand-held button box with the index or middle finger. If they did not respond in time, a screen displaying the word 'miss' appeared for 2000 ms. On STOP trials, a beep signal was presented over the headphones indicating that subjects had to withhold their response. At the start of the experiment, the time interval between the stimulus and beep (stop signal delay) was initialized at 250 ms and was adjusted in a staircase procedure based on the stopping performance (Jahfari et al., 2011).

Overall, 25% of the scenes were shown in STOP trials, and these trials contained an equal number of animal and non-animal scenes (the same scenes for all subjects). Trials were presented in two randomized sequences that were counterbalanced across subjects. All analyses reported here included only the GO trials. For these trials, animal and non-animal scenes were still matched in their CE and SC values per condition (means: all $t(118) < 1.13$, all $p > 0.26$; medians: all $z < 1.10$, all $p > 0.28$).

**Figure 6.1** Stimuli and design. A) Image statistics of the stimuli: each scatter point represents a scene sampled from the image space described by CE and SC. Scenes had either low (red), medium (green) or high (blue) CE and SC values. Within these conditions, CE and SC values were matched between scenes with (A) and without animals (NA). B) Example scenes from each condition. C) Experimental design of the fMRI experiment. On GO trials, subjects indicated whether the scene contained an animal or not. On STOP trials, an auditory signal was given at a variable inter-trial-interval (ITI) after scene presentation, signaling that subjects had to withhold their response. Only GO trials were analyzed.
**Behavioral data analysis**

Behavioral categorization performance was measured by computing mean RT and sensitivity (d') based on the GO trials for each subject. Correct categorization of an animal scene was counted as a hit and incorrect categorization of a non-animal scene as a false alarm. To obtain a combined measure of speed and accuracy we computed Inverse Efficiency Scores (IES; Townsend and Ashby, 1978) by dividing the average RT for each subject by their overall percentage correct on both animal and non-animal scenes. Differences in average d', RT and IES between the three conditions (LOW, MEDIUM, HIGH) were tested with repeated-mesures ANOVAs and paired t-tests. Data were analyzed in Matlab (Mathworks, Natick, MA, USA) and SPSS 17.0 (IBM, Armonk, USA).

**fMRI recordings**

The BOLD-MRI data was acquired in a single scanning session, over the course of two runs on a 3T scanner (Philips Achieva XT with a 32-channel head-coil). In each run 255 T2*-weighted GE-EPI recordings (TR = 2000 ms, TE = 27.6 ms, FA = 76.1°, SENSE = 2, FOV = 240 mm², matrix size = 80², 37 slices, slice thickness 3 mm, slice gap = 0.3 mm) were made. During acquisition we also recorded breath-rate and the pulse-oxidization signal. In addition, a separate functional localizer scan was recorded, to identify the fusiform face area (FFA), the parahippocampal place area (PPA) and lateral occipital complex (LOC) in each subject. Using the same acquisition settings as for the main experiment, subjects viewed a series of houses, faces, objects as well as phase-scrambled scenes. To sustain attention during functional localization, subjects pressed a button when an image was directly repeated (12.5% likelihood). A 3D-T1 weighted scan (TE = 3.8 ms, TR = 8.2 ms, FA = 8°, FOV = 256², matrix size = 256², 160 sagittal slices, slice thickness = 1mm) was acquired after the functional runs. This scan was used to register the functional volumes of each run to the structural brain, after which they were registered to standard MNI (Montreal Neurological Institute) space.

**fMRI data analysis: main experiment**

Analysis was performed using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl) and custom Matlab code. The functional data were motion- (Jenkinson et al., 2002) and slice-time corrected. Subsequently a median filter was applied to remove low frequencies and whiten the data. Next, the data was spatially smoothed with a Gaussian filter with a 5 mm FWHM kernel. The resulting preprocessed scans were subjected to voxel-wise event-related GLM analysis using FILM (Woolrich et al., 2001) by convolving the onset times of each trials with a double gamma function to model the hemodynamic response function. We generated explanatory variables (EVs) by dividing trials in the following conditions: LOW animal, LOW non-animal, MEDIUM animal, MEDIUM non-animal, HIGH animal, and HIGH non-animal. For these EVs, only correct GO trials
were included; STOP, error and non-misses were modeled as separate EVs. In addition, heartbeat and breathing variables were included as nuisance variables. This GLM results in an estimate of the magnitude of the BOLD signal for each EV in each voxel. Based on these estimates, the following contrasts of interest were computed for each individual subject: LOW (animal > non-animal), MED (animal > non-animal) and HIGH (animal > non-animal). Because the trial design lacked a common baseline condition, all analyses were conducted on these differential activity measures.

*fMRI data analysis: localizer scans*

The localizer scans were preprocessed for the purpose of another study conducted within the same experimental session (Jahfari et al., in preparation). Again, the data were motion- and slicetime corrected and prewhitened. In addition, they were spatially smoothed using a 4mm Gaussian filter and were temporally filtered by means of a high-pass filter (sigma = 50s). A GLM was fitted with following EVs: for FFA, faces > (houses and objects); for PPA, houses > (faces and objects) and for LOC, (intact scenes > scrambled scenes).

The resulting statistical maps were masked with anatomically defined regions from the Harvard-Oxford cortical-structural atlas implemented in FSL. For FFA, these were the temporal occipital and occipital fusiform gyrus; for PPA, the parahippocampal gyrus and lingual gyrus as posterior as MNI coordinate y-74; for LOC, the lateral occipital cortex inferior division). Significant voxels within these masks were thresholded at z = 2.3 (FFA and PPA) or z = 3.0 (LOC) and manually adjusted within each subject to remove individual voxels that were not part of clusters found within the masked regions.

**ROI analysis**

We examined the contrasts of interest explained above within regions-of-interest (ROIs) derived from the functional localizer scans (FFA, PPA and LOC) as well as an anatomical mask of V1 derived from the Jülich histological atlas implemented in FSL (Eickhoff et al., 2005). We first performed the ROI analysis for each hemisphere separately. However, because initial inspection of the results did not indicate any differences between hemispheres, all results are reported for bilateral ROIs. Within the ROIs the normalized activity (i.e., the t-value for the differential contrast of interest) was averaged over voxels. The resulting values were compared across conditions using repeated-measures ANOVAs for each individual ROI.

**Whole brain analysis**

We performed a whole-brain analysis for the same contrasts of interest as the ROI analysis. The resulting statistical maps were first pooled across runs (fixed effects) and then across subjects (mixed effects using FLAME1; Woolrich, 2008). To examine where in the brain the three contrasts differed from one another, we ran the following additional contrasts: HIGH (animal > non-animal) > LOW (animal > non-animal);
HIGH (animal > non-animal) > MEDIUM (animal > non-animal); and MEDIUM (animal > non-animal) > LOW (animal > non-animal). Results were corrected for multiple comparisons by means of cluster correction (z = 2.3, p < 0.05; Worsley, 2001).

**EEG experiment**

**Subjects**

Twenty-one participants (7 males, 22-33 years old, mean 25.6, ± SD = 2.5) took part in the EEG experiment. All participants had normal or normal-to-corrected vision, provided written informed consent and received financial compensation. The ethics committee of the University of Amsterdam approved the experiment. Two subjects were excluded in preprocessing: one subject based on the presence of excessive alpha activity in the EEG, and one because this subject had a history of epilepsy.

**Experimental procedure**

Subjects viewed 7200 different scene stimuli while performing various go-no go object recognition tasks (Groen et al., 2010). Stimuli were presented on a 19-inch Ilyama CRT-monitor (1024x768 pixels, frame rate 60 Hz). Subjects were seated 90 cm from the monitor such that stimuli subtended ~14x10° of visual angle. On each trial, one image was randomly selected and presented in the center of the screen on a grey background for 100 ms, on average every 1500 ms (range 1000 - 2000 ms). Subjects searched for either animals or vehicles in one of four tasks: detection, or superordinate, basic-level or subordinate categorization. For the purpose of this paper, only the trials for scenes used in the fMRI experiment were analyzed. Of these scenes, 130 were presented during an animal vs. background task (object detection); 84 during a vehicle vs. background task (object detection); 110 during an animal vs. vehicle task (super-ordinate categorization); 109 during a vehicle vs. animal task (super-ordinate categorization), and 47 in a cats vs. other animals task (basic-level categorization). These task differences were not considered here; we were only interested in whether evoked activity to the scenes that were used in the fMRI study to constitute the LOW, MEDIUM and HIGH conditions (and that were matched in their CE and SC values) differed early in time between animal and non-animal scenes.

**EEG recording and analysis**

EEG recording and preprocessing was identical to other chapters described in this thesis. We obtained evoked potentials for each scene (n = 480), subject (n = 19), electrode (n = 64) and time point (n = 154). Per condition, we computed the grand average ERP to animal and non-animal scenes, after which difference waves were derived for each individual subject. The ERPs and difference waves were pooled across two different sets of visual electrodes: an early occipital pool (Oz, Iz, O1, O2, I1 and I2) and a medial peri-occipital pool (POz, PO3, PO4, Pz, P1, P2, P3, and P4).

In each condition, the difference wave between animal and non-animal scenes was tested against zero using separate one-sample t-tests for each time
sample and electrode pooling. In analogy with the fMRI analysis, difference waves were compared across conditions using the following paired-sample t-tests: HIGH (animal vs. non-animal) vs. LOW (animal vs. non-animal); HIGH (animal vs. non-animal) vs. MEDIUM (animal vs. non-animal); and MEDIUM (animal vs. non-animal) vs. LOW (animal vs. non-animal). P-values were corrected for multiple comparisons (number of tests, poolings and time-points) using FDR correction (pFDR = 0.013).

Results

Behavior

Average sensitivity (d’) and reaction time for LOW, MEDIUM and HIGH complexity scenes are presented in Figure 6.2. A repeated-measures ANOVA on d’ revealed a significant main effect of condition (F(2,23) = 9.96, p = 0.001). Subsequent paired comparisons indicated that sensitivity in the MEDIUM condition was significantly higher compared to the LOW (paired-samples t(23) = -3.2, p = 0.004) and HIGH condition t(23) = 3.5, p = 0.002) (Figure 6.2A). The LOW and HIGH condition did not differ from each other in average sensitivity (t(23) = 1.8, p = 0.08). A main effect of condition was also found on RT (F(2,23) = 5.84, p = 0.012). Subjects responded significantly slower in the HIGH condition compared to the LOW (t(23) = 2.7, p = 0.01) and the MEDIUM conditions (t(23) = 2.8, p = 0.01) (Figure 6.2B). There was no significant difference in RT between LOW and MEDIUM (t(23) = 0.16, p = 0.87).

These results show that subjects were best able to perform the animal vs. non-animal categorization task for scenes with intermediate CE/SC values. Although performance in the LOW condition was worse than for the intermediate condition, subjects responded equally fast for both types of scenes. For high complexity scenes, however, subjects both performed less well and responded more slowly. This suggests that the HIGH condition was experienced as most difficult.

Figure 6.2 Behavioral results. A) Sensitivity (d’) in the animal categorization task per condition. Colored lines show the data for individual subjects; gray bars the averages. B) Average reaction times (RT) in each condition. C) IES scores per condition. Error bars represent S.E.M.. ** p < 0.01, * = p < 0.05.
To confirm this conclusion, we combined the speed and accuracy scores in a single inverse efficiency score (IES) measure, as proposed by Townsend and Ashby (1978), which weighs the impact of speed and accuracy (Bruyer and Brysbaert, 2011; see Methods and Materials). Statistical comparison of IES across conditions (Figure 6.2C) indicated a main effect of condition ($F(2,23) = 14.4, p < 0.0001$). The highest IES values were found in the HIGH condition, which differed significantly from the MEDIUM ($t(23) = 4.5, p = 0.0001$) and LOW condition ($t(23) = 3.1, p = 0.005$), that also differed from one another ($t(23) = 3.0, p = 0.006$). Thus, when simultaneously taking into account speed and accuracy, the HIGH condition was most difficult.

**ROI analysis**

The difference in BOLD activity for animal vs. non-animal scenes in each condition was computed in four regions of interest: LOC, FFA, PPA and V1. FFA and LOC both responded more strongly to scenes with animals than to scenes without animals (Figure 6.3A-B), but these responses did not differ across conditions (all $F(2,23) < 1.7$, all $p > 0.19$). In PPA, the opposite pattern was observed (Figure 6.3C): it responded more strongly for scenes without animals than scenes with animals. This activity difference however differed significantly across conditions ($F(2,23) = 3.5, p = 0.04$): it was significantly larger in the MEDIUM condition compared to the HIGH ($t(23) = -2.5, p = 0.021$) but not compared to the LOW condition ($t(23) = 1.8, p = 0.08$). Finally, in V1 a main effect of condition was observed ($F(2,23) = 4.1, p = 0.03$); differential activity for animals vs. non-animal scenes was larger in the MEDIUM condition compared to both the LOW ($t(23) = -2.3, p = 0.03$) and the HIGH ($t(23), p = 0.03$) condition.

These results show that the scene complexity manipulations affect activity levels in high-level scene selective area PPA. Of the three conditions, scenes of medium complexity give rise to strongest differential activity between animal and non-animal scenes. These scenes were also best categorized by participants. For low complexity scenes, for which behavioral sensitivity was diminished but which were processed equally fast, we observe a less strong differential signal, but this effect is not significantly different from the intermediate condition. For the most difficult, high complexity scenes, however, we find a significantly less strong differential signal in high-level PPA. Conversely, in low-level visual area V1 differential activity is present for scenes of high complexity only.

These findings suggest that categorization of scenes of intermediate complexity is facilitated because they induce a clear differential signal in activity in PPA. By comparison, scenes of high complexity induce less difference in PPA activity and give rise to additional differential activity in V1. Information encoded in low-level areas may thus be selectively recruited for categorization of high complexity scenes. To examine whether any condition effects were present beyond our ROIs, we next tested to what extent voxels across the whole brain exhibited differential activity between animal and non-animal scenes.
Whole brain analysis

Whole brain statistical maps (Figure 6.4) for the comparison (animal > non-animal) and (non-animal > animal) revealed clusters of activity in lateral visual and medial high-level visual cortex, respectively. Cluster coordinates for all contrasts are reported in Table 6.1. For the animal > non-animal contrast (Figure 6.4A), bilateral clusters overlying lateral occipital areas were present in each condition. Moreover, in the HIGH condition, additional differential activity was present in low-level visual areas. Indeed, contrasting these statistical maps revealed a large cluster in several early visual areas including V1 (Figure 6.4B). For non-animal > animal (Figure 6.4C), bilateral clusters were present in the MEDIUM condition, whereas only right-lateralized clusters in the LOW and HIGH condition survived whole-brain cluster-correction. However, contrasting the difference between non-animal > animal scenes between conditions resulted in no significant clusters.

These results largely concur with the effects observed in the ROI analysis. In all three conditions, reliable differences are found in lateral occipital areas. In contrast, the MEDIUM scenes give rise to more reliable differential activity in medial parahippocampal areas, whereas HIGH scenes selectively induce differential activity in early visual areas. Since the animal and non-animal scenes were carefully matched in low-level image statistics within conditions, it is not likely that the activity
differences in V1 are driven by differences in low-level properties between these scenes. We thus suggest that this low-level visual differential activity results from increased neural feedback elicited by the need for increased detailed scene information in the high complexity condition.

Figure 6.4 Whole brain results. A) Statistical parametric maps for animal (A) > non-animal (NA) contrast for each condition. From left to right, MNI coordinates for each transversal slice are \( z = [10, 2, -6] \). B) Statistical parametric maps for non-animal > animal contrast for each condition. From left to right, MNI coordinates are \( z = [-2, -8, -14] \). C) Differences in the animal vs. non-animal activity between conditions. MNI-coordinates: \( [x = 6, y = -88, z = 6] \). Maps were cluster-corrected and thresholded at \( z = 2.3 \). Color scales range from \( z = 2.5 \) to 5 (A and B) and \( z = 2.3 \) to 3.5 (C).
<table>
<thead>
<tr>
<th>contrast</th>
<th>size</th>
<th>log10(p)</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>areas included</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>animal &gt; non-animal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOW</td>
<td>43656</td>
<td>15.6</td>
<td>45</td>
<td>-66</td>
<td>-9</td>
<td>V4, V5, lat. occ. cortex. inf. division, inf. temp. gyrus, temp. occ. fusiform. gyrus (R)</td>
</tr>
<tr>
<td></td>
<td>43472</td>
<td>15.6</td>
<td>-41</td>
<td>-71</td>
<td>-8</td>
<td>V4, V5, lat. occ. cortex. inf. division, inf. temp. gyrus, temp. occ. fusiform. gyrus (L)</td>
</tr>
<tr>
<td>MED</td>
<td>39296</td>
<td>15.7</td>
<td>45</td>
<td>-70</td>
<td>-6</td>
<td>V4, V3v, lat. occ. cortex. inf. division, inf. temp. gyrus, temp. occ. fusiform. gyrus, occ. fusiform gyrus (R)</td>
</tr>
<tr>
<td></td>
<td>39976</td>
<td>15.9</td>
<td>-42</td>
<td>-74</td>
<td>-6</td>
<td>V4, V5, lat. occ. cortex. inf. division, inf. temp. gyrus, temp. occ. fusiform. gyrus. (L)</td>
</tr>
<tr>
<td>HIGH</td>
<td>142168</td>
<td>36.1</td>
<td>3</td>
<td>-74</td>
<td>-3</td>
<td>V1, V2, V3, V4, V5, lat. occ. cortex. inf. division, inf. temp. gyrus, temp. occ. fusiform. gyrus. (L+R)</td>
</tr>
<tr>
<td><strong>non-animal &gt; animal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOW</td>
<td>4585</td>
<td>1.96</td>
<td>27</td>
<td>-45</td>
<td>-9</td>
<td>parahippocampal gyrus post. div., lingual gyrus, temp. occ fusiform cortex (R)</td>
</tr>
<tr>
<td>MED</td>
<td>29872</td>
<td>5.78</td>
<td>4</td>
<td>49</td>
<td>2</td>
<td>parahippocampal gyrus post. div., lingual gyrus, temp. occ. fusiform cortex, supracalcarine cortex, precuneus (L+R)</td>
</tr>
<tr>
<td>HIGH</td>
<td>4392</td>
<td>1.81</td>
<td>26</td>
<td>-47</td>
<td>-1</td>
<td>parahippocampal gyrus post. div. lingual gyrus, temp. occ. fusiform cortex. supracalcarine cortex, precuneus (R)</td>
</tr>
<tr>
<td><strong>animal &gt; non-animal, between conditions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIGH &gt; MED</td>
<td>14536</td>
<td>6.32</td>
<td>7</td>
<td>-85</td>
<td>-2</td>
<td>V1, V2, V3V, V4, occ. fusiform gyrus, lingual gyrus (R)</td>
</tr>
<tr>
<td></td>
<td>7992</td>
<td>3.61</td>
<td>-15</td>
<td>-95</td>
<td>19</td>
<td>V1, V2, V3V, inf. parietal lobule, occ. fusiform gyrus, lingual gyrus (R)</td>
</tr>
<tr>
<td>HIGH &gt; LOW</td>
<td>21016</td>
<td>8.94</td>
<td>2</td>
<td>-90</td>
<td>4</td>
<td>V1, V2, V3V, inf. parietal lobule, lingual gyrus, (L+R)</td>
</tr>
</tbody>
</table>

Table 6.1 Cluster coordinates for the significant contrasts in the whole-brain analysis. COG = center of gravity, p = p-value, lat. = lateral, occ. = occipital, temp. = temporal, inf. = inferior, post. = posterior, L = left, R = right. Coordinates are reported in MNI space. Areas included in the clusters were determined by determining overlap of local maxima within clusters with the probability maps of the Juelich histological Atlas and the Harvard-Oxford Cortical Structural Atlas implemented in FSL. Note that for the contrasts HIGH (animal - non-animal), MED (non-animal - animal) and [HIGH (animal-non-animal) - MED (animal-non-animal)], the clusters are bilateral, forming one cluster across hemispheres (see also Figure 6.4).
However, due to the poor temporal resolution of the fMRI signal, we cannot exclude the possibility that early visual activity differences were in fact present due to other perceptual differences between the animal and non-animal scenes than those captured by CE and SC. To test our hypothesis that the activity differences in early visual areas were feedback-related, we used ERP recordings for the same scenes obtained in previously conducted experiment (Groen et al., 2010). We used these ERPs to test at what moment in time animal and non-animal evoked responses started to differ on electrodes overlying visual cortex. Since scenes were matched in image statistics, we predicted that there would be no differences between animal and non-animal scenes before 100-150 ms, beyond which feedback starts to affect ERP activity (Lamme and Roelfsema, 2000; Fahrenfort et al., 2007; Scholte et al., 2008).

**ERP results**

Evoked responses to the LOW, MEDIUM and HIGH animal and non-animal scenes are presented in Figure 6.5. Consistent with our previous EEG studies, early (~100-150 ms) ERP amplitudes at occipital sites were increasingly negative for scenes with high CE/SC values (Figure 6.5A). Critically, however, whereas animal and non-animal ERPs in the LOW and MEDIUM conditions started to diverge between 120 and 150 ms, they were completely overlapping up to 200 ms in the HIGH condition. In contrast, at peri-occipital sites (Figure 6.5B), we observe that the ERPs diverged at ~200 ms, with the largest differences being present in the HIGH condition.

![Figure 6.5 ERP results. Average ERP amplitude for animal (solid gray lines) and non-animal scenes (dashed gray lines) and the corresponding difference wave (solid black lines) per condition for A) a posterior occipital (Oz, Iz, O1, O2, I1 and I2) and B) a medial peri-occipital pooling (POz, PO3, PO4, Pz, P1, P2, P3, and P4).](image)
Statistical comparison of the differences waves shows that for the occipital pooling (Figure 6.6A), the difference wave in the HIGH condition is much less reliable than the other two conditions. In fact, for none of the conditions the difference waves reached sustained significance until the end of the epoch (thick significance lines in Figure 6.6A). In contrast, at the peri-occipital pooling (Figure 6.6B) the difference wave in the HIGH condition is extremely reliable from 180 ms onwards. Moreover, it is significantly larger than the difference wave in both the LOW and MEDIUM condition (see blue stars/crosses in Figure 6.6B). The results for the posterior occipital pooling thus show that it is unlikely that the differential activity in early visual areas in the HIGH condition is driven by uncontrolled low-level properties. In fact, in the HIGH condition, a reliable ERP difference between animal and non-animal scenes arises later than in the other two conditions.

Moreover, we observed an enhanced ERP difference beyond 200 ms at peri-occipital sites in this condition. This effect might be the electrophysiological correlate of the early visual area difference observed in the fMRI experiment. However, it is surprising that this effect was only present at the peri-occipital, but not the occipital pooling, given that in the enhanced fMRI activity for animal vs. non-animal scenes extended all the way down to V1. This could be due to the fact that V1 is largely located in the calcarine fissure on the medial surface between the two hemispheres and its activity is thus not clearly reflected in the ERP signal. Instead, other visual areas than V1 such as V2, V3 or V3A, which are located on the cortical surface and which all also had increased differential fMRI activity in the HIGH condition (see Table 6.1) could have contributed to the enhanced difference in the ERP signal.

In any case, in both the occipital and peri-occipital pooling the ERP differences in the HIGH condition appeared beyond 200 ms, which suggests it cannot arise from a feed-forward signal alone, but is likely modulated by recurrent activity. Together, these results thus suggest that differences between animal and non-animal scenes in early visual areas are enhanced by a feedback signal selectively employed for scenes with high complexity.

Figure 6.6 Difference waves for the A) posterior occipital and B) medial peri-occipital pooling for each condition overlaid and statistically compared. Thick lines: significant difference wave deflections from zero; shadings: 95% confidence intervals. Symbol markers: significant differences between conditions.
Discussion

In this chapter, we investigated whether the image statistics contrast energy (CE) and spatial coherence (SC) affect object recognition in natural scenes. We find that for scenes with high CE and SC values, i.e. with low figure-ground segmentation and high clutter, animal categorization performance is decreased compared to scenes with lower CE and SC values. Using fMRI, we find that this effect is characterized by a decrease in differential activity between animal and non-animal scenes in high-level scene-selective areas. In low-level visual areas, the opposite pattern was observed: there was more differential activity for the complex, cluttered scenes. Separate ERP recordings support the interpretation that this difference likely results from neural feedback, rather than feed-forward activity differences between the scenes.

These results show that when scene complexity is varied, different areas become involved in representing the difference between animal and non-animal scenes. For simple scenes with clear figure-ground segmentation, only high-level areas are involved, whereas low-level areas are selectively recruited for complex scenes. Because the enhanced activity in early visual areas appears to result from a feedback signal, we interpret these findings as showing a selective need for increased detailed analysis for the scenes of high complexity. For simple scenes, the initial, coarse representation provided by the feed-forward sweep is enough for high-level areas to obtain information about the presence or absence of a salient (animal) object. For highly cluttered, complex scenes, however, the feed-forward sweep is not sufficiently diagnostic: it merely signals that there is a lot of ‘stuff’ in the scene. To detect and categorize the objects it contains, information about individual line segments and other features encoded in low-level areas need to be integrated with the initial impression by means of recurrent activity.

This interpretation fits with the idea that the brain processes information in a global-to-local manner, first constructing a coarse impression of the scene by means of feed-forward activity, after which detailed analysis takes place via reentrant processing (Hochstein and Ahissar, 2002). This detailed analysis could take the form of different visual routines based on the visual input and task requirements (Roelfsema et al., 2000). In the visual routine framework, the feed-forward sweep provides visual cortex with a base representation. If the visual task can be solved based on this representation alone, no further operations are necessary and e.g. action selection can be initiated. If, however, the base representation is not sufficiently informative, visual routines can be applied existing of elemental operations such as contour grouping, curve tracing and texture segmentation. These processes require feedback activity because they need to recover features belonging to a single object encoded at lower visual areas (Roelfsema et al., 2000). Our results suggest that the degree of detailed analysis that is required depends on the complexity of the scene as reflected in global image statistics.
CE and SC could very well be involved in the formation of the base representation because they can be computed in a feed-forward manner from non-oriented, low-level contrast responses that are for example present in LGN (Scholte et al., 2009). Moreover, as we have seen in Chapters 2-5, they strongly affect the amplitude of early evoked neural activity as measured with EEG. If these parameters are indeed a good characterization of the overall complexity of the feed-forward information in natural scenes, the brain might use them as information 'markers' to determine whether and/or which further operations are necessary.

**A role for image statistics in signaling the need for feedback**

The influence of the bottom-up input on subsequent scene analysis has been studied before by Schyns and Oliva, who proposed that extraction of coarse 'blobs' precedes detailed analysis at other spatial scales depending on the diagnostic value of the blobs (Schyns and Oliva, 1994; Oliva and Schyns, 1997). Similarly, Bar et al., (2006) suggested that a low spatial frequency representation of the input is carried all the way to orbitofrontal cortex in the feed-forward sweep, and then used to direct top-down facilitation on the visual input. That feedback can be employed flexibly depending on the visual input was also suggested by the masking study of Wyatte et al., (2012), who decreased the quality of the visual input using occlusion and contrast reduction. By means of computational modeling, these authors showed that a feedback component was necessary to accurately predict the effects of masking on behavioral categorization performance, but that this was the case for degraded scenes only.

Here, we did not degrade or manipulate the scenes but rather used CE and SC to sample real-world variation in scene complexity. Our results suggest that these scene statistics convey meaningful information that affects the dynamics of visual processing. Sampling scenes based on real-world properties constitutes a more ecologically valid approach and indicates that the observed effects could in fact occur in the brain during real-life perception. Moreover, our results provide new neuro-imaging evidence in support of the suggestion that scenes with more demanding backgrounds lead to enhanced employment of recurrent activity (Koivisto et al., 2013). In that particular study, behavioral effects of masking were found to be stronger for scenes that were 'less easily segmented', which was assessed post-hoc by independent observers that were asked to rate scenes on this property. Here, we provide an a priori, quantitative approach to this question, showing that ease of segmentation can be manipulated beforehand by constraining scene selection based on physiologically plausible image statistics.

Importantly, however, not only the complexity of the visual input, but also top-down factors may determine whether the representation provided by the feed-forward sweep is sufficient or not. Whereas the feed-forward representation for low CE/SC scenes may be sufficient for animal detection, this might not be the case for more complex tasks such as basic-level or subordinate categorization or identification. In
addition, when searching for objects in natural scenes, feedback is likely involved in the application of attentional templates (Peelen and Kastner, 2014), which could differ between different types of tasks (Malcolm and Henderson, 2009), for example in terms of the level of detail that is necessary to solve the task (Oliva and Schyns, 1997). So, although accurate object recognition can likely be achieved based on feed-forward information (Serre et al., 2007) for certain types of scenes and under certain tasks instructions, in real-life scene perception there likely is a bidirectional interplay between bottom-up and top-down requirements (Malcolm et al., 2014) which will affect the amount of feedback activity that is employed.

Differential activity in PPA vs. LOC

An interesting finding in our study is that CE and SC affected differential activity between animal and non-animal scenes in PPA, but not in LOC. The LOC is usually defined as the area that is more sensitive to intact than scrambled objects (Malach et al., 1995) and has been proposed to be crucially involved in object recognition (Grill-Spector, 2003). Animal vs. non-animal object distinctions are prominently present in LOC, both when objects are presented in isolation (Kriegeskorte et al., 2008b) and when they are embedded in natural scenes (Naselaris et al., 2012). Accordingly, we find a strong and reliable difference between the animal and non-animal scenes in our study in this area in all conditions. Apparently, the LOC activity differences between animals and non-animal scenes are thus not strongly modulated by whether the object is embedded in a complex, cluttered scene or not. Importantly, this does not imply that LOC is not at all sensitive to scene statistics. In fact, our group recently showed that fMRI activity in posterior parts of lateral occipital cortex (LO1/LO2) is modulated by the SC of natural scenes (Scholte et al., 2013). However, in the current study the animal and non-animals scenes were matched on SC (and CE). The results show that differential LOC activity between animals and non-animal scenes is of equal magnitude in each part of the CE/SC space: it thus seems that this information is encoded independently of the scene background.

PPA, on the other hand, is defined as the area that responds stronger to scenes than to non-scene stimuli (Epstein, 2005). This selectivity has been attributed to sensitivity of PPA to spatial layout information (Epstein and Kanwisher, 1998; Kravitz et al., 2011; Park et al., 2011) or scene category (Walther et al., 2009). However, PPA is also sensitive to contextual associations (Bar and Aminoff, 2003), spatial qualities of individual objects (Troiani et al., 2014) and to texture and "object ensembles" (Cant and Xu, 2012). In line with these findings, a recent study independently manipulated gradients of object and spatial information in scenes (Harel et al., 2012) and found that PPA was in fact sensitive to variations in both object and scene information, unlike area LOC, which cared only about the objects but not the background information. Consistently, we here find that differential object activity in PPA (but not LOC) is modulated by the broader scene context.
Potential effects of the stop-signal task
An important detail in our study is that subjects were engaged in a stop-signal task, rather than a simple animal-non-animal categorization task. This task is generally thought to affect action selection processes reflected in motor activity buildup (Verbruggen and Logan, 2009), rather than sensory processing, and we excluded all trials in which a stop signal may have interfered with visual processing. However, recent evidence suggests that this task also targets sensory representations (Salinas and Stanford, 2013), in the sense that sensory information can modify ongoing motor plans. Indeed, this is precisely the reason why we used this task in the first place. To better understand potential effects of sensory information on action selection, we will examine in a separate, future report how CE and SC influence strategic adjustments in action control (Jahfari et al., 2013) and how these changes may be reflected in effective connectivity in prefrontal go- and stop brain networks (Jahfari et al., 2011).

Conclusion
In sum, we provide evidence for increased feedback activity during animal categorization in natural scenes depending on the complexity of the scene. These results fit with the suggestion that a coarse impression of scenes based on feed-forward activity not only precedes (Hochstein and Ahissar, 2002) but also affects subsequent detailed analysis of the scenes (Roelfsema et al., 2000). In particular, our results suggest that feedback is more strongly employed during object detection if the scenes are cluttered, as described by contrast energy and spatial coherence. The selective need for “vision with scrutiny” (Hochstein and Ahissar, 2002) may thus be signaled by summary representations in high-level scene-selective areas.