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

Groen, I.I.A.

Publication date
2014

Citation for published version (APA):
Chapter 4
From image statistics to scene gist: evoked activity reveals neural sensitivity to local contrast statistics during global property categorization

The visual system processes natural scenes in a split second. Part of this process is the extraction of "gist", a global first impression. It is unclear, however, how the human visual system computes this information. Here, we show that when human observers categorize global information in real-world scenes, the brain exhibits strong sensitivity to image statistics. Subjects rated a specific instance of a global scene property, naturalness, for a large set of natural scenes while EEG was recorded. For each individual scene, we derived two physiologically plausible summary statistics by pooling local contrast filter outputs: contrast energy (CE), indexing contrast strength, and spatial coherence (SC), indexing scene fragmentation. We show that naturalness rating is directly related to these statistics, being influenced in particular by SC. At the neural level, both statistics parametrically modulated single-trial event-related potential amplitudes during an early, transient window (100-150 ms), but SC continued to influence activity levels later in time (up to 250 ms). In addition, the magnitude of neural activity that discriminated between man-made versus natural ratings of individual trials was related to SC, but not CE. These results suggest that global scene information may be computed by spatial pooling of responses from early visual areas (e.g., LGN or V1). The increased sensitivity over time to SC in particular, which reflects scene fragmentation, suggests that this statistic is actively exploited to estimate naturalness.

Published as:
Introduction

The remarkable speed with which humans can perceive natural scenes has been studied for decades (Potter, 1975; Intraub, 1981; Thorpe et al., 1996; Fei-Fei et al., 2007b). Many theories of visual processing propose that a global impression of the scene accompanies (Rousselet et al., 2005b; Wolfe et al., 2011) or precedes (Hochstein and Ahissar, 2002) detailed feature extraction ('coarse-to-fine' processing) (Hegdé, 2008). This global percept is often described as visual gist (Torralba and Oliva, 2003; Oliva, 2005). Behavioral experiments confirm that global scene properties are indeed perceived very rapidly: observers can extract global information such as spatial layout, navigability and naturalness in less than 100 ms of exposure time (Greene and Oliva, 2009b).

Scene naturalness may be a particularly interesting example of a global scene property, because it can be judged with less exposure time compared to several other global properties (e.g. depth; Oliva and Torralba, 2001; Greene and Oliva, 2009a) or basic-level categories (e.g. mountain vs. forest; Rousselet et al., 2005b; Joubert et al., 2007; Loschky and Larson, 2010; Kadar and Ben-Shahar, 2012). In computer vision, it was shown that the man-made/natural distinction coincides with low-level regularities of natural scenes, such as the distribution of spatial frequencies (Baddeley, 1997; Oliva et al., 1999), leading to the suggestion that humans might use such 'image statistics' in rapid scene recognition (Schyns and Oliva, 1994; McCotter et al., 2005; Kaping et al., 2007).

It is unclear, however, to which image statistics the brain is sensitive, and how it uses them to extract scene naturalness: besides spatial frequency, colour (Oliva and Schyns, 2000; Goffaux et al., 2005) and local edge alignment (Wichmann et al., 2006; Loschky et al., 2007) are examples of scene features that may play a role.

We propose that to elucidate the role of image statistics in scene perception and their neural representations, we need to address two challenges. First, a plausible computational mechanism by which image statistics are extracted by the brain must be specified. For example, it has since long been recognized that estimating spatial frequency distributions directly from an image by using a global Fourier transformation is a biologically implausible mechanism, given that the visual system receives localized information from the retina (Graham, 1979; Field, 1987). Second, if image statistics are indeed used to extract global properties, they should predict not only general differences between scene categories but visual representations of individual images. This means that single-trial differences, in both neural activity and behavior, should correlate with differences in image statistics.

Here, we addressed these problems by combining computational modeling of physiologically plausible image statistics with single-trial EEG measurements. We approximated image statistics by summarizing the outputs of model receptive fields (Torralba, 2003; Renninger and Malik, 2004; Ghebreab et al., 2009) with two parameters, contrast energy (CE), and spatial coherence (SC) that carry information
about contrast strength and scene fragmentation (Scholte et al., 2009; Groen et al., 2012a, 2012b).

**Figure 4.1.** Stimuli, model and methods. **A)** Examples of images used in the experiment. Images varied considerably in the degree to which they contained exclusively man-made (left) or natural (right) elements, and the set included images for which the distinction may be unclear (middle). **B)** A feed-forward filtering model was used to derive two contrast statistics: contrast energy (CE) and spatial coherence (SC). Three opponent (grayscale, blue-yellow, red-blue) contrast magnitude maps were computed by convolving the image with multi-scale filters (black circles). For CE, a range of smaller filter sizes was used (σ is diameter in degrees); for SC, a range or larger filter sizes. For each image location, a single filter response was selected (red solid circles) from each range using minimum reliable scale selection (see Materials and Methods). These responses were then pooled across a selection of the visual field (red dotted circles): for CE, the resulting responses were averaged; for SC, the coefficient of variation (COV) was computed. These values were averaged across the three color-opponent maps resulting in one CE and SC value per image. **C)** A subset of 160 images (10% of the whole stimulus set, randomly selected) plotted against their CE and SC values. CE (the approximation of the ß-parameter of the Weibull function) describes the scale of the contrast distribution: it varies with the average local contrasts strength. SC (the approximation of the γ-parameter of the Weibull function) describes the shape of the contrast distribution: it varies with the average local contrasts strength. SC (the approximation of the γ-parameter of the Weibull function) describes the shape of the contrast distribution: it varies with the average local contrasts strength. CE (the approximation of the ß-parameter of the Weibull function) describes the scale of the contrast distribution: it varies with the average local contrasts strength. **D)** Single-trial evoked responses (sERPs) to images presented at the center of the screen were computed for each subject. The resulting estimates of sERP amplitude were regressed on CE and SC at each time sample and electrode separately. The design matrix for the regression contained 5 columns; a constant term (c) for the intercept, two columns for CE and two for SC, containing the same parameter values for the first and second image presentation (p1 and p2). These were modeled as separate predictors to examine reliability of the obtained effects across repetitions. The outcome of the analysis is a measure of model fit (explained variance/R²) separately over subjects, time (samples) and space (electrodes). An example is shown for one subject at electrode (Oz).
We tested whether CE and SC were related to behavioral ratings of scene naturalness and examined how they affected the time-course of single-trial event-related potentials (sERPs) during naturalness categorization. Additionally, we examined how they affected discriminability of man-made versus natural ratings based on EEG activity.

Materials and Methods

Subjects
Sixteen subjects (mean age = 26, SD = 6, 3 males) participated in the experiment, which was approved by the Ethical Committee of the Psychology Department at the University of Amsterdam. All participants gave written informed consent prior to participation and were rewarded with study credits or financial compensation (7 euro/hour). The data of two subjects were excluded from analysis because their recordings were incomplete.

Visual stimuli
A stimulus set of 1600 color images (bit depth 24, JPG format, 640x480 pixels) was composed from several existing online databases. The images were selected such that one half of the set contained mostly man-made, and the other mostly natural elements and included images from a previous fMRI study on scene categorization (Walther et al., 2009), and from various datasets used in computer vision: the INRIA holiday database (Jegou et al., 2008), the GRAZ dataset (Opelt et al., 2004), ImageNet (Deng et al., 2009) and the McGill Calibrated Colour Image Database (Olmos and Kingdom, 2004). These different sources assured maximal variability of the stimulus set (Figure 4.1A): it contained a wide variety of indoor and outdoor scenes, landscapes, forests, cities, villages, roads, images with and without animals, objects and people.

Computational modeling

General approach
Natural images exhibit much statistical regularity. One instance of such regularity is present in the distribution of contrast strengths, which ranges between power law and Gaussian and therefore conforms to a Weibull distribution (Simoncelli, 1999; Geusebroek and Smeulders, 2002). This regularity can thus be described by fitting a Weibull function to the contrast distribution, which yields two ‘summary’ parameters that represent the scale (β) and shape (γ) of the distribution (Scholte et al., 2009). The β-parameter varies with the range of contrast strengths present in the image, whereas γ varies with the degree of correlation between contrasts. Previously, we found that these parameters can be approximated in a physiologically plausible way by summarizing responses of receptive field models to local contrast (Scholte et al.,
2009). Specifically, summing the responses from a model of the two main magno-
and parvo-cellular pathways in the LGN led to accurate approximations of $\beta$ and $\gamma$
values, respectively.

Here, we used an improved version of this previous model (Figure 4.1B) where we refer to the approximated $\beta$-value as contrast energy (CE) and to the approximated $\gamma$-value as spatial coherence (SC). In this new model, two major changes have been introduced: 1) contrast is computed at multiple spatial scales and 2) CE is estimated by averaging local contrast values while SC is taken to be the coefficient of variation (mean divided by standard deviation). Importantly, these changes have led to improved results in separate EEG datasets from the one reported here, with different images and different EEG recordings. Specifically, using CE and SC led to an increase in effect size in EEG (see below) by ~20% (Scholte et al., 2009 vs. Ghebreab et al., 2009); the present approximations of CE and SC correlated most strongly with fitted Weibull parameters (Scholte et al., in preparation).

In the present dataset, the approximations (CE and SC) correlated with the fitted parameters ($\beta$ and $\gamma$) with $r = 0.95$ and $r = 0.73$, respectively. The computations involved in filtering and pooling of local contrast are described below.

**Step 1: Local contrast detection**
Contrast filtering was done separately for the three image color layers, which were first converted to opponent color space (grayscale; blue-yellow; red-green; Koenderink et al., 1972). We computed image contrast by convolving each image layer with exponential filters (Zhu and Mumford, 1997) at five octave scales (Croner and Kaplan, 1995). Two separate filter sets were used: one with slightly smaller filter sizes (0.12, 0.24, 0.48, 0.96 and 1.92 degrees) for CE and a range of larger filter sizes (0.16, 0.32, 0.64, 1.28 and 2.56 degrees) for SC (see Ghebreab et al., 2009). Following the LGN suppressive field approach (Bonin et al., 2005), all filter responses were rectified and divisively normalized to account for non-linear neuronal properties.

**Step 2: Scale selection**
Per parameter (CE or SC), one filter response for each image location was selected from their specific filter set using minimum reliable scale selection (Elder and Zucker, 1998). In this MIN approach, the smallest filter size that yields an above-threshold response is preferred over other filter sizes. Noise thresholds were determined in a separate image set (Ghebreab et al., 2009). This step was again performed separately for the three color-opponent layers.

**Step 3: Pooling responses**
Applying the selected filters per image location results in two contrast magnitude maps, one highlighting detailed (from the set of smaller filter sizes) and the other more coarse edges (from the set of larger filter sizes). For the pooling step, a different amount of visual space was taken into account for each parameter. For CE,
the central 1.5 degrees of the visual field was used, whereas for SC, 5 degrees of visual angle were used. Again, these settings were chosen because they yielded the best model fits for regression analyses in separate, independent EEG datasets (Ghebreab et al., 2009; Scholte et al., 2009, and unpublished observations). Finally, parameter values were averaged across color-opponent layers resulting in a single CE and SC value per image (Figure 4.1C).

**Other image statistics**

For comparison of the behavioral results with previous findings of sensitivity of human observers to spatial frequency distributions (e.g. Kaping et al., 2007), Fourier amplitude statistics were computed using the procedure described in Oliva and Torralba (2001), i.e. fitting a line to the rotationally averaged power spectrum. This procedure yields one 'intercept' and 'slope' value per image (see also Groen et al., 2012a).

**Experimental design**

**Experimental procedure**

Subjects completed two EEG recording sessions on two consecutive days. In each session, 1600 images (640x480 pixels) were shown. Every image was repeated once. Images 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 were instructed to indicate, as quickly as possible, whether the image was man-made (‘made by humans’) or natural (‘not made by humans’) using two button boxes, one for each index finger. Response mappings were counterbalanced across participants. To familiarize subjects with stimulus presentation, 30 practice images (never used in the main experiment) were presented without feedback before the first session. Each session was divided in 4 runs, which were subdivided in 5 self-paced mini-blocks: subjects were encouraged to take breaks between blocks. Stimuli were presented using the Presentation software package (www.neurobs.com).

**Behavioral data analysis**

Trials at which the subject failed to respond within 200-1200 ms after stimulus-onset were discarded from analysis (median rejection rate = 0.1%, min = 0%, max = 4%). For each of the 1600 images, we computed the following indices: 1) Naturalness rating, i.e. the average behavioral responses across subjects and repetitions, with 0 indicating that none of the participants rated the scene as natural and 1 indicating that all participants rated the scene as natural; 2) Subject-specific naturalness rating, i.e. the average response across the two repetitions, with 0 indicating a man-made
response, and 1 indicating a natural response; 3) Average reaction time (RT) across subjects and repetitions; 4) Subject-specific RT. These indices were correlated (Spearman’s ρ) with CE and SC, as well as the Fourier slope and intercept values. The resulting p-values were FDR-corrected for the total number of correlations computed. Finally, we also computed for each image, a ‘reliability index’ by comparing the subject-specific naturalness ratings across the two presentations of the image. If the rating was the same across repetitions, reliability was coded as 1, whereas it was coded as 0 if the rating was different. Averaging these numbers across subjects yielded an estimate of reliability for each image.

**EEG data acquisition and preprocessing**

EEG Recordings were made with a Biosemi 64-channel Active Two EEG system (Biosemi Instrumentation BV, Amsterdam, NL, www.biosemi.com). Recording set-up and preprocessing were identical to the procedures described in Groen et al., (2012a, 2012b). We used caps with an extended 10-20 layout, modified with two additional occipital electrodes (I1 and I2, while removing electrodes F5 and F6). Eye movements were monitored with electro-oculograms (EOG). Recording was followed by offline referencing to external electrodes placed on the earlobes. Pre-processing occurred in Brain Vision Analyzer and included a high-pass filter at 0.1 Hz (12 dB/octave); a low-pass filter at 30 Hz (24 dB/octave); two notch filters at 50 and 60 Hz; automatic removal of deflections > 300 mV; epoch segmentation in -100 ms to 500 ms from stimulus onset; ocular correction using the EOG electrodes (Gratton and Coles, 1983); baseline correction between -100 ms and 0 ms; automated artifact rejection (maximal voltage steps 50 µV, minimal/maximal amplitudes -75/75 µV, lowest activity 0.50 µV) and finally, conversion to Current Source Density responses (Perrin, 1989). Median rejection rate was 203 out of 3200 trials (mean 7%; min 1%, max 20%). We also removed, per subject, trials that were excluded based on behavior (see above), which increased the median rejection rate to 8%; overall, 9% of the total amount of data was removed. No trial or electrode averaging was performed: preprocessing thus resulted in a single-trial event-related potential (sERP) specific to each subject, electrode and individual image presentation.

**Regression on single-trial ERPs**

To test whether differences between sERPs were modulated by differences in CE and SC, the sERPs were read into Matlab (Mathworks, Natick, MA, USA), where we conducted linear regression of sERP amplitude on image parameters (Figure 4.1D) using the Statistics Toolbox. For each subject, each electrode and each time-point, the two parameters were entered as ‘predictor variables’, and the z-scored sERP amplitudes as ‘observations’ in the regression model. The two presentations of each image were modeled as two separate predictor columns, in order to examine effects of repetition; the values in each column were z-scored independently (Rousselet et al., 2011). This analysis thus results in a measure of model fit (explained variance or
\( r^2 \) for each subject, electrode and time-point separately. In addition, we read out from the regression results the \( \beta \)-coefficients assigned to each predictor variable, which reflect the regression weights between the image parameters and sERP amplitude (again separately for each subject, electrode and time-point). To assess the significance of these \( \beta \)-coefficients, we tested the subject-averaged \( \beta \)-coefficient against zero using t-tests (separate for each time-point and electrode): a significant test result indicates that the association between a predictor and evoked neural activity is reliably larger than zero at a specific time-point and electrode. To correct for multiple comparisons (time-points, electrodes and parameters), p-values were FDR-corrected at \( \alpha = 0.05 \). The rationale of this approach is similar to conventional fMRI analysis (GLM); we compute the \( \beta \)-coefficients for each predictor in our model and threshold these using multiple-comparison corrected t-values. By doing this for every electrode (‘whole-scalp analysis’), we created spatial maps that indicate which electrodes contain significant \( \beta \)-coefficients.

**Regression results for other image statistics**

For comparison of the regression results for CE and SC with parameters obtained from the Fourier transform, we ran the analyses described above while replacing the predictor columns with Fourier intercept (Fi) and slope (Fs). In addition, we examined the dependence between the two models by running a third regression analysis in which we entered all four parameters (CE, SC, Fi, Fs) together as predictors, yielding a ‘full’ model. By subtracting the \( r^2 \)-values for the first two models from the full model, we obtained a measure of ‘unique explained variance’ by each model (Groen et al., 2012b). These values were again obtained separately per subject, electrode and time-point.

**Linear discriminant component analysis**

To identify decision-related activity in the EEG signal, we used linear discrimination analysis (LDA) as developed by Philiastides and colleagues (Parra et al., 2002; Philiastides and Sajda, 2006; Philiastides et al., 2006). LDA performs logistic regression of binary data on multivariate EEG data to identify spatial weighting vectors (\( w \)) across electrodes that maximally discriminate between conditions of interest (e.g., a face or car stimulus). Here, we used as conditions of interest the ratings made by the subjects during naturalness rating, by entering the trial-specific behavioral responses (man-made or natural) as the binary data. The analysis outcome is a ‘discriminating component’ \( y \), which is specific to activity correlated with one condition while minimizing activity correlated with both task conditions (Philiastides et al., 2006). In our case, we thus isolated activity specific to the decision that a scene was rated as natural.

We used an online available regularized logistic regression algorithm (Conroy and Sajda, 2012) that allows for fast estimation of the discriminating components. We tested various regularization terms (\( \lambda = 10^x \) with \( x = -6:1:2 \)), which yielded similar
results to those reported here for $\lambda = 1e^{-1}$. We computed $y$ for a number of temporal windows (window size $\delta = 20$ ms), to estimate the temporal evolution of discriminant activity over the course of the ERP. Per time window, discriminator performance (Az) across all trials was quantified using the area under the receiver-operator characteristic (ROC) curve, with a leave-one-out cross-validation approach (Duda et al., 2001). Following Blank et al., (2013), significance of performance was evaluated by bootstrapping the Az-values. We used 100 bootstraps per time window (epoch of 600 ms/$\delta = 40$ windows) and subject ($n = 14$) resulting in 42,000 Az-values in total. The overall distribution of these values was used to determine the Az-value leading to a significance level of $\alpha = 0.05$. Finally, the components were projected back on the scalp using a forward model that multiplies the $w$-vector with average ERP amplitude in a specific time window (Parra et al., 2002).

After obtaining an estimate of discriminant component activity at each trial and time-window, the $y$-values were correlated (Spearman's $\rho$) with the CE and SC values of the image presented at that trial, for each subject separately. To assess the significance of these correlations, we again tested the average correlation across subjects against zero using separate $t$-tests for each time-window; the resulting $p$-values were corrected for multiple comparisons using FDR-correction at $\alpha = 0.05$.

**Results**

**Behavior**

Fourteen subjects rated 1600 scene images as either man-made or natural while EEG was recorded. We first examined how behavioral ratings were distributed across the entire set of images. Next, we tested how these results were related to differences in contrast energy (CE) and spatial coherence (SC) estimated from modeled responses to local contrast.

**Naturalness rating and RT for individual images**

On average, subjects rated the trials as man-made and 50.3% as natural (SD = 3.3% for both man-made and natural). Reaction times for man-made versus natural responses did not differ significantly (mean $RT_{\text{man-made}} = 523$ ms, SD = 67 ms; mean $RT_{\text{natural}} = 527$ ms, SD = 79 ms; median $RT_{\text{man-made}} = 497$ ms, SD = 64 ms, median $RT_{\text{natural}} = 504$ ms, SD = 76 ms; all $t(13) < 1$, all $p > 0.56$), showing that subjects did not have a bias towards one particular response. There was however considerable variability in ratings across trials, both within subjects and across subjects.

Within subjects, on average 10% of the trials were rated differently on the second presentation than the first (min = 3%, max = 18%, SD = 4%). For these trials, reaction times were longer than for trials that were rated the same (mean $RT_{\text{same}} = 520$ ms, SD = 70 ms, mean $RT_{\text{different}} = 571$ ms, SD = 90 ms; paired $t$-test $t(13) = 6.1$,}
Across subjects, the scenes could be subdivided evenly in three bins based on differences in naturalness rating (Figure 4.2A): scenes that had a rating below 0.1 (indicating they were rated as natural by less than 10% of the subjects), scenes with a rating above 0.9, and scenes for which the rating was intermediate (0.1-0.9 ratings). Average RT for the intermediately rated images was higher compared to consistently rated images (repeated-measures ANOVA, F(2,13) = 42.6, p = 0.0001); this bin also more often contained scenes for which ratings differed between the repetitions (repeated-measures ANOVA on within-subject consistency, F(2,13) = 138.6, p = 0.0001; Figure 4.2B).

Although subject’s behavior was on average similar for the two scene categories, the increase in variability in rating both within and across subjects for a specific subset of the images shows that some scenes were experienced as more ‘ambiguous’ than others. This is not surprising as the stimulus set was purposely composed to span a wide range of natural images (see Materials and Methods). Next, we examined whether this pattern of results could be explained based on differences in image statistics between the scenes.

**Figure 4.2 Behavioral results.** A) Distribution of naturalness ratings for the entire stimulus set. Roughly 2/3 of the scenes had a high man-made (average < 0.1, blue bin) or high natural rating (> 0.9, red bin), while the remaining images had intermediate ratings (green bin, 0.1-0.9). B) Reaction times (RT; top graph) and within-subject repeat reliability (REL; bottom graph) for the three different bins of trials. Asterisks indicate significant main effects of bin (p < 0.0001). C) Contrast energy (CE) and spatial coherence (SC) values plotted for all 1600 images, color-coded by naturalness rating. The histograms display 10% bins based on either CE or SC, containing the average rating obtained after sorting on SC (top histogram, x-axis), or the average RT after sorting on CE (side histogram, y-axis). D) Intermediately rated images (green bin in A) are found at intermediate SC values, whereas images with high man-made (blue) or natural (red) ratings have on average more extreme SC values. Squares: medians, lines: standard deviations. E) Example images from each bin. Whereas the highly man-made or natural rated images (left) contain either exclusively man-made or exclusively natural elements, one intermediate image (top right) contains a building and a bicycle (man-made) as well as snowy objects and a bush (natural); another (bottom right) has shrubbery (natural) and a fence and wall (man-made).
Contrast energy and spatial coherence predict behavioral performance

The CE and SC values for each image in the image set are shown in Figure 4.2C, with color-coding indicating the average naturalness index per image. Binning images on either CE or SC (side histograms in Figure 4.2C) shows that naturalness rating is related to differences in SC: the higher the SC-value of the scene, the higher the number of subjects that rated the scene as natural. A higher CE-value is associated with shorter RTs. Figure 4.2C also shows that the SC values span a continuous space, rather than two discrete categories. Interestingly, the variability in image ratings is related to SC value: images with intermediate ratings are found at intermediate SC-values (Figure 4.2D; independent-samples Kruskal-Wallis test, $\chi^2(2) = 131.3$, $p = 3e-29$) whereas there is no such effect for CE ($\chi^2(2) = 2.4$, $p = 0.29$). As can be seen in Figure 4.2E, images with intermediate ratings/SC values typically contain patches with both man-made and natural elements. This suggests that SC in particular is diagnostic of naturalness.

We confirmed these observations by correlating the CE and SC values directly with the behavioral measures (Table 4.1). SC correlated with naturalness rating, but not with RT, whereas CE correlated with RT, and to a lesser extent with rating. For comparison we also computed correlations with image statistics derived from the power spectra of the scenes (Oliva and Torralba, 2001), Fourier intercept and slope (see Materials and Methods). These correlations are in a similar range as those with CE and SC (Table 4.1). For this model, however, the two parameters are less well dissociable, as they both correlate significantly with accuracy as well as RT.

<table>
<thead>
<tr>
<th>image parameter</th>
<th>naturalness rating</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho$</td>
<td>$p$</td>
</tr>
<tr>
<td>CE</td>
<td>-0.08</td>
<td>0.0007</td>
</tr>
<tr>
<td>SC</td>
<td>0.39</td>
<td>2.3e-58</td>
</tr>
<tr>
<td>Fs</td>
<td>-0.19</td>
<td>1.3e-15</td>
</tr>
<tr>
<td>Fi</td>
<td>-0.36</td>
<td>2.5e-50</td>
</tr>
</tbody>
</table>

Table 4.1 Correlations of image parameters with naturalness rating and reaction times (RT). CE = contrast energy, SC = spatial coherence, Fs = Fourier slope, Fi = Fourier intercept. Reported are Spearman correlation values ($\rho$), corresponding significance values ($p$), and the percentage of subjects whose individual rating or RT correlated significantly with the parameters (% subjects; FDR-corrected).

The behavioral results support our proposal that CE and SC, computed by pooling of local contrast, capture global scene information. They predict human categorization of scene naturalness to a similar degree as statistics obtained from a global Fourier transformation. In particular, there is a relation between SC and perceived naturalness, whereas CE affects the speed of visual processing.
**EEG regression**

Single-trial ERPs (sERPs) were extracted from the continuous EEG data after which we performed linear regression of sERP amplitude on image statistics. We first examined how CE and SC affected evoked activity to individual images, again comparing them to the Fourier parameters. Next, we tested how the EEG activity itself was related to the behavioral ratings, and whether this relation was dependent on CE and/or SC.

*Explained variance by contrast energy and spatial coherence*

Regression analysis of sERP amplitude on CE and SC values revealed a strong relation between these parameters and evoked neural activity. For all subjects, explained variance was maximal at either Oz or Iz (the occipital mid-line electrodes), ranging between $r^2 = 0.16-0.46$ at 109-137 ms (for all subjects, $p < 1e-10$, FDR-corrected), whereas mean explained variance over subjects was highest at Oz (maximal $r^2 = 0.27$, 117 ms after stimulus-onset; **Figure 4.3A**). Explained variance at this time-point also extended to other electrodes (**Figure 4.3B**); maximal $r^2$ values were reliable in all fourteen subjects at several occipital (I1, O1, O2) and parietal electrodes (P1, P3, P4, P6, P8, P10). These results replicate our earlier findings (Scholte et al., 2009) of ERP sensitivity to statistics derived from local contrast in natural images from ~100 ms after stimulus-onset. They show that differences between individual scenes in evoked neural activity are reliably correlated with differences in CE and SC of the scenes.

*Regression weights for individual parameters*

Given the dissociable effects of CE and SC on behavior, we next tested whether these parameters also differentially influenced sERP amplitude by examining their individual $\beta$-coefficients (see Materials and Methods). At electrode Oz (**Figure 4.3C**), the $\beta$-coefficients for CE were largest during an extended early interval (94-156 ms). In contrast, the largest $\beta$-coefficient for SC was found at a non-overlapping, later interval (160-195 ms). Regression weights associated with each predictor were highly reliable across repetitions.

The difference between the CE and SC effects is illustrated in **Figure 4.3D**, where sERP amplitude for each image relative to its SC and CE value is shown for the two time-points with largest $\beta$-coefficients (113 and 184 ms, respectively). At the first maximum, differences in amplitude are mostly aligned with the CE-axis of the space. At the second maximum, the amplitudes align with the SC-axis instead. Significant but smaller coefficients for SC were also found at two shorter and earlier intervals (78-85 ms and 117-136 ms) and one later interval (226-242 ms). Neural sensitivity to the different image parameters thus differs substantially over the time-course of visual processing. Interestingly, these results again demonstrate dissociable effects of CE and SC, but now on evoked neural activity.
Finally, for both CE and SC, regression coefficients were also significant towards the end of the ERP epoch, most likely reflecting spill-over of effects at other electrodes. These might be electrodes near motor cortex, as this interval is in the RT range (minimal average RT was 436 ms, maximal average RT was 647 ms), and the \( \beta \)-coefficients at Oz are very small at this time-point, indicating low sERP amplitudes. In the next section, we tested to what extent the two image parameters affected responses at other electrodes.

**Figure 4.3** EEG regression results at occipital electrode Oz. **A)** Explained variance for single subjects (colored lines) and mean across subjects (thick black line) over time. **B)** Distribution of mean \( r^2 \) values across the scalp at the maximal time-point (117 ms). Dots indicate electrodes; for red electrodes, \( r^2 \) values at this time points were significant in all participants (FDR-corrected). **C)** Regression weights \( (\beta \text{-coefficients}) \) for each predictor, revealing several significant time windows (lines in bottom). CE-coefficients are maximal early in time (95-156 ms), SC-coefficients later in time (160-195 ms). Shading indicates 95% confidence intervals as obtained from a t-test of the mean \( \beta \)-coefficient against zero. **D)** Average CSD-transformed ERP amplitude (color-coded) per image plotted against its CE and SC values at the time-points of maximal regression weights.

**Whole-scalp results**

The distribution of regression coefficients across the whole scalp mirrors the effects observed at electrode Oz. Generally, the \( \beta \)-coefficients of CE were largest early in time, at occipital electrodes overlying early visual cortex. For SC, on the other hand, the \( \beta \)-coefficients were largest later in time and more sustained at peri-occipital
electrodes that are near mid- or higher-level visual areas (Figure 4.4A). Coefficients of CE were virtually absent after 150 ms, whereas the coefficients of SC remained significant for an extended period across multiple electrodes.

This difference is illustrated in Figure 4.4B, where average sERP amplitude is shown for bins of trials that are sorted based on either CE or SC. At occipital electrodes O1 and O2, the modulation by CE is clearly visible in the early interval, while a sustained modulation by SC is visible at peri-occipital visual electrodes PO7 and PO8. The distinct effects of CE and SC are thus even more evident at the whole-scalp level, with differences in contrast energy giving rise to a transient, early modulation, whereas spatial coherence differences lead to more widespread effects.

Figure 4.4 Whole scalp regression results. A) Topographical representations of t-values (FDR-corrected and thresholded; corrected t = 3.22 for α = 0.05) for β-coefficients at each electrode, reflecting reliable effects of CE and SC on sERP amplitude. The two parameters have different effects: CE influences ERP amplitude early in time, mostly at occipital electrodes, whereas SC gives rise to a sustained correlation later in time (> 150 ms) associated with activity surrounding mid-level visual, peri-occipital electrodes (e.g., PO7 and PO8). B) Subject-averaged ERP amplitude of trials that were sorted and binned in 20% bins based on either CE or SC values, displayed for two different electrode poolings: occipital (O1 and O2) and peri-occipital (PO7 and PO8).

Comparison with Fourier parameters
The regression results demonstrate strong effects of CE and SC on evoked neural activity during categorization of naturalness. However, behavioral naturalness ratings did not only correlate with these image statistics estimated from local contrast, but also with Fourier parameters obtained from a global Fourier transform. Do these
latter parameters also affect evoked activity in a similar way as CE and SC? To test this, we repeated the regression analyses with the Fourier parameters.

Explained variance of the Fourier parameters had a similar time course as for CE and SC, but with on average lower values (Figure 4.5A): the average maximal value was 17% (ranging between 11-29% for single subjects) at electrode Oz around 117 ms, the same electrode and time-point as for CE and SC. The spatial extent of the explained variance was also similar to CE and SC (Figure 4.5B). To compare the two models directly, we estimated the unique explained variance by each pair of parameters by subtracting the $r^2$-values obtained for each model from a full model containing all parameters (see Materials and Methods). Unique explained variance for CE and SC reached a maximum of 11% at 113 ms, whereas the values were much lower for the Fourier parameters (2% at 184 ms; Figure 4.5C). For both models, maxima of unique explained variance were again at Oz, and the whole-scalp results (Figure 4.5D) show that the unique explained variance for Fourier was in fact minimal across all electrodes. This rules out the possibility that Fourier parameters influence neural activity at other brain sites than CE and SC do.

Figure 4.5 Comparison of results with regression analysis on Fourier statistics. A) Explained variance averaged across single subjects for regression of single-trial ERP amplitude on Fourier intercept (Fi) and slope (Fs) (green). For comparison, average $r^2$ for CE and SC is plotted as well (red). Shading indicates standard deviation across subjects. B) Distribution of average $r^2$-values across the scalp at the time-point of maximal $r^2$ (117 ms). Reliable electrodes across subjects (red) were I1, I2, O1, O2, Oz, P3, P4, P5, P6, P7, P8 and POz. C) Unique explained variance for CE and SC (red) compared to Fi and Fs (green), averaged over subjects. Shading indicates standard deviation across subjects. D) Scalp plots of unique $r^2$-values for each model at their respective maximal time-points: for CE and SC, this was at 117 ms (top topographical plot) and at 184 ms for the Fourier parameters (bottom).
As we learned in Chapter 2, in previous findings with naturalistic image categories (Groen et al., 2012a) the Fourier parameters also explained about ~10% variance less than parameters derived from a Weibull fit to the contrast distribution (of which CE and SC are approximations, see Materials and Methods), as well as very little unique variance.

These results show that modulations of single-trial evoked activity during naturalness categorization are also correlated with differences in spatial frequency content (Fourier intercept and slope). However, the effects were weaker compared to those observed for CE and SC, and the Fourier parameters did not explain substantial variance above and beyond the variance explained by CE and SC. This suggests that information that is contained in the Fourier parameters is also captured by CE and SC, and that the latter may provide a more plausible description of evoked neural activity during naturalness categorization.

Role for image statistics in perceptual decision-making?

The regression results suggest that contrast energy and spatial coherence play different roles in visual processing of natural images, with CE giving rise to early, transient effects, while SC leads to sustained effects on evoked activity. Interestingly, the modulations by SC extend into time intervals associated with mid-level stages of visual processing (beyond 200 ms) that are more sensitive to top-down influences (Luck et al., 2000; Scholte et al., 2006) and possibly involved in perceptual decision-making (Philiasstides et al., 2006). Task-relevance of the scene parameters may thus play a role in the differential effects of CE and SC that were observed here. Specifically, the late modulation by SC could reflect an influence of SC on the naturalness decision.

To test how CE and SC affected decision-related neural activity, we used linear discriminant analysis (LDA) to identify discriminating components in the EEG (Parra et al., 2002; Philiasstides and Sajda, 2006; Philiasstides et al., 2006; Conroy and Sajda, 2012; Blank et al., 2013) This is essentially again a single-trial regression, but of behavior onto the EEG: ‘observations’ correspond to the (subject-specific) man-made or natural ratings, and the ‘predictor variables’ consist of the sERP amplitudes at each electrode. This analysis yields two measures: overall discrimination accuracy, summarized in parameter Az, and trial-specific discriminant components y, reflecting evidence toward the natural versus the man-made rating at a given trial (see Materials and Methods). Both measures were determined for consecutive time windows to examine the development of discriminant activity over time (Philiasstides and Sajda, 2006; Philiasstides et al., 2006).

With this analysis, we aimed to address two questions: 1) from what point in time can we reliably predict whether a given trial will be rated as man-made or natural? and 2) To what extent is the strength of evidence towards the man-made or the natural rating modulated by CE and SC?
Discriminating man-made versus natural ratings based on EEG

First, response discrimination accuracy (Az) based on EEG was significantly above chance from a time-window between 80-100 ms (Figure 4.6A). It reached a local maximum between 180 and 200 ms (but remained significant) and started to rise again from 260 ms onwards. Projections of the discriminant activity back on the scalp (insets in Figure 4.6A) show that for the first two of these windows, activity was located at occipital/peri-occipital sites, whereas the activity in the third window was more lateral. Importantly, the early maximum in discrimination accuracy was found at the same moment in time at which SC most strongly affected sERP amplitude (see Figure 4.3C).

Figure 4.6 Discriminant analysis results. A) Discriminator performance (AZ) determined using a leave-one-out procedure (LOO). Colored lines reflect single subject results; the black line is the average across subjects. Grayscale shading indicates standard deviations. The dashed gray line reflects the Az value leading to a significance level of p = 0.05 (obtained from a bootstrap test). Insets represent the scalp distributions of discriminating component activity for three moments in time: the first time window of significant discriminator performance (100-120 ms), the local maximum (180-200 ms), and the sustained effects from 250 ms onwards (distribution shown for the 320-340 ms window). B) Correlation of discriminant component value (y) with contrast energy (red) and spatial coherence (blue). Confidence intervals and p-values (FDR-corrected for the number of time-points and parameters) were obtained by testing the average correlation across subjects against zero at each time window. Time windows with significant correlations are marked with asterisks. C) Discriminating component maps displaying the value of discriminating component amplitude y for each image and time window, averaged over the two single-trial presentations and sorted from low to high on either CE (top) or SC (bottom). Higher component activity (red) indicates evidence for the response ‘natural’.
Second, discriminant components \((y)\) were significantly correlated with SC, from 120 ms onwards (maximal mean \(\rho = 0.18\), 180-200 ms, \(t(13) = 9.1\), \(p = 5.1e^{-7}\); Figure 4.6B). For CE, correlations were present for just one early window (80-100 ms, mean \(\rho = -0.06\), \(t(13) = -3.7\), \(p = 0.002\)), and again much later in time, from 360 ms onwards. Note, however, that for this early window, discrimination accuracy is not yet significant (Figure 4.6A). Figure 4.6C shows discriminant activity for each image, while the images are sorted based on either CE or SC. These component maps reveal that images with high SC-values have strong evidence towards natural responses (red), whereas images with low SC-values are mapped towards man-made responses (blue). No such effects are visible for CE.

These results show that activity discriminating between man-made versus natural responses is present in the EEG from as early as 100 ms. Importantly, the strength of this activity at the single-trial level is related to the SC, but not CE, value of the image.

Overall, the EEG results reveal strong neural sensitivity to image statistics derived from receptive field model responses to local contrast. Neural sensitivity was stronger for CE and SC than for spatial frequency parameters derived by means of a global Fourier transformation. We found dissociable effects of CE and SC on neural activity, across both time (early and late intervals) and space (occipital versus peri-occipital electrodes). This suggests that contrast energy mainly affects activity in early visual areas involved in encoding of the stimulus, whereas spatial coherence also modulates subsequent decision-related activity, which likely involves more visual areas and processing time. The finding that SC, but not CE, is correlated with EEG components that maximally discriminate between the subject-specific behavioral ratings further supports this dissociation.

**Discussion**

*A computational substrate for scene gist perception*

How does the visual system estimate scene gist? We find that for at least one global scene property (naturalness), single-trial differences in behavior and neural activity are related to differences in image statistics derived by integrating local contrast responses. The modulation of neural activity by these statistics as soon as 100 ms after stimulus onset confirms that this information is available early in visual processing. However, the sustained modulations at later time-points reveal a shift in neural sensitivity from contrast energy (CE) to spatial coherence (SC), suggesting a transformation towards coding of more relevant information for estimating naturalness, which appears to be carried by SC.

These results verify our earlier observations of extensive ERP sensitivity to image statistics. We showed that differences in statistics explain a large amount of variance in ERP amplitude (Scholte et al., 2009), to the extent that they can be used to classify which natural image was viewed (Ghebreab et al., 2009). Perceived
similarity of textures and naturalistic images are also related to CE and SC (Groen et al., 2012a, 2012b). Here, we extend these results to scene gist, by linking single-trial differences in scene naturalness and neural activity to differences in image statistics.

Model comparison with spatial frequency distributions

The role of spatial frequency in visual processing has been studied at multiple levels, ranging from local receptive field tuning (De Valois and De Valois, 1990) to the entire image. For example the 'spectral signature' of a scene, reflecting the decay of the 2D power spectrum, can be used to computationally discriminate global scene properties (Torralba and Oliva, 2003). Whereas the power spectrum reflects the distribution of energy across spatial scales ('amplitude'), another type of information is local alignment of spatial frequencies ('phase coherence'), which can for example be quantified from the shapes of contrast distributions (Tadmor and Tolhurst, 2000). It is currently debated which of these two sources (amplitude vs. phase) is more important for scene discrimination (McCotter et al., 2005; Einhäuser et al., 2006; Loschky et al., 2007; Loschky and Larson, 2008).

Our model essentially extracts phase information based on filters modeled after receptive fields (Scholte et al., 2009). However, this information is inferred from contrast computed at, and selected locally from, multiple spatial scales (Elder and Zucker, 1998; Ghebreab et al., 2009) and the information is then spatially integrated across the entire scene. This procedure thus likely captures both phase and some amplitude information, and we have shown previously that our model outperforms separate descriptions of amplitude and phase (Groen et al., 2012a). Here, we again observed that CE and SC explained the neural data better than power spectra alone. This is consistent with the observation that coarse localization of spectral signatures improves computational discrimination of global properties (Torralba and Oliva, 2003). Our results thus support the notion that both amplitude and phase contribute to scene perception (Gaspar and Rousselet, 2009; Joubert et al., 2009).

It is important to emphasize, however, that CE and SC are derived in a very different way compared to traditional measures of phase and amplitude. The latter are obtained by performing a global Fourier transformation, whereas our model integrates local information using averaging and division, which can easily be implemented in a spiking neural network (e.g. using ‘integrate-and-fire’ rules). We thus also attribute the fact that CE and SC provided a better description of neural activity to the physiological plausibility of our model. We propose that our model provides a biologically realistic computational substrate from which image statistics can be derived: local contrast, which could be represented by the population response in visual areas such as LGN or V1.

The role of image statistics in visual perception

The early, transient correlations with CE are consistent with previous reports of early ERP sensitivity to image statistics of natural scenes, demonstrating modulation of
early visual ERP components (e.g. C1 and P1) by energy at different spatial frequencies (Hansen et al., 2011, 2012). During object recognition, early ERP sensitivity to luminance differences (Martinovic et al., 2011), power spectra (Johnson and Olshausen, 2003) and phase scrambling (Rousselet et al., 2008a; Bieniek et al., 2012) have also been reported. In face categorization, ERP sensitivity to phase scrambling has also been found to extend later in time (up to 300 ms; Rousselet et al., 2008b), thus overlapping with our present SC effects. In addition, ERP sensitivity to ‘geometric similarity’ between faces has also been found in this time range (Kahn et al., 2010). Here, we show that modulations in this time range were related to the behavioral stimulus encoding, but also neural processing leading up to the decision outcome.

Supporting this idea, the maximal timing of the late SC effects is close to a discriminating component reported in a set of papers that used a face/car categorization task (Philiastides and Sajda, 2006; Philiastides et al., 2006; Ratcliff et al., 2009). The authors proposed that the D200 discriminating component reflects an intermediate stage between early sensory processing and accumulation of decision-related information, signaling the ‘availability of diagnostic information’. Another study however argued that the modulation of this component was purely related to the addition of phase noise to the stimuli (Bankó et al., 2011). Both claims could be in accordance with our findings, as SC is also sensitive to the addition of phase noise, which leads to a more Gaussian contrast distribution and thus to higher fragmentation of the scene. For the purpose of our task, however, this information could be useful: the brain does not need to discard this information as noise, but may use it as input for the decision. In that case, the similarity in timing with the D200 component supports the notion that spatial coherence contains available information for global property categorization.

Indeed, modulation of neural activity by seemingly simple, low-level image properties such as contrast does not necessarily imply that these properties are irrelevant for image recognition. If they are relatively consistent across categories, they may have become part of a ‘template’ used by the visual system to classify incoming information (Johnson and Olshausen, 2003). The CE and SC parameters reflect variations in the distribution of local contrast: they are thus derived from information that is generally considered to be ‘low-level’. However, since their computation requires integration of this information across the scene, they pick up on global, ‘high-level’ scene information (global energy and scene fragmentation). It is not unlikely that over the course of evolution and/or development, the visual system has adapted and developed templates that are sensitive to such variations in low-level information if they are diagnostic of relevant global properties.

Naturalness as a visual primitive?

Previous work has suggested that the man-made/natural distinction is fundamental in scene perception. Categorization of man-made versus natural scenes occurs faster
than basic-level categories (sea, mountain, city) (Joubert et al., 2007; Greene and Oliva, 2009a; Loschky and Larson, 2010). Basic-level categories from the same superordinate (e.g. sea vs. mountain) level are also more easily confused than those from different levels (e.g. sea vs. city) (Rousselet et al., 2005b). Within global properties, categorization may occur hierarchically, starting with man-made vs. natural (Kadar and Ben-Shahar, 2012). The relation between SC and scene naturalness, as well as the influence of SC on evoked neural activity during naturalness categorization, suggest that the spatial coherence of the scene may drive this early primary distinction.

The fact that humans can quickly decide about naturalness however does not imply that the brain computes it ‘automatically’. Would the brain be interested in determining the naturalness of visual input in everyday viewing? SC varies with scene fragmentation, signaling the relative presence of ‘chaos’ versus ‘order’, rather than an absolute distinction between man-made and natural. However, there is a relation because natural scenes are more likely to be chaotic due to the presence of foliage or other structure that has not been organized by humans (as urban environments are). We thus suggest that the primacy of man-made versus natural reflects early sensitivity to the fragmentation level of visual input. Estimating the fragmentation of incoming visual information may be a useful step in rapid scene processing, for example in order to allocate attention (when presented with a chaotic scene, in which searching for an object will be harder and require more resources; or alternatively, with a highly coherent scene containing a large single object, which needs to be attended in order to assess potential danger) or cognitive control mechanisms. Future experiments should establish to what extent image properties such as CE and SC predict recruitment of attention or control networks.

**Conclusion**

Together, these results suggest that natural image statistics, derived in a physiologically plausible manner, affect the perception of at least one global property: scene naturalness. The results revealed strong neural sensitivity to image statistics when subjects categorized this global property, with decision-related activity specifically being modulated by spatial coherence (SC). We propose that during scene categorization, the brain extracts diagnostic image statistics from pooled responses in early visual areas.

**Acknowledgements**

This work is part of the Research Priority Program ‘Brain and Cognition’ at the University of Amsterdam and was supported by an Advanced Investigator Grant from the European Research Council to VAFL and the Dutch national public-private research program COMMIT to SG.