The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/195556

Please be advised that this information was generated on 2019-01-26 and may be subject to change.
Visual processing of scenes in the first tens of milliseconds relies on global image summary statistics rather than localized processing. Although natural scenes typically involve our entire visual field, scenes are usually presented experimentally at limited eccentricity. Receptive-field size increases with foveal eccentricity while increasingly pooling activity from local receptive fields. Here, we asked to what extent an observer’s performance on a scene-gist perception task depends on the contents of the scene as well as on the eccentricity of the scene. We manipulated the scene content by applying window and scotoma masks. In addition, we changed presentation eccentricity independent of image content by upscaling and downscaling the scenes. We find that discrimination is strongly affected when the scene is presented with a window of 5°, showing only the central part rather than the whole scene. Performance is, however, eccentricity scale independent provided that the same scene content is presented and a comparable area of the surface of primary visual cortex is activated. We furthermore show that this eccentricity scale independence holds for shorter presentation times, down to 17 ms in some scene-discrimination tasks, but not for the naturalness-discrimination task.

Introduction

Humans are able to understand the visual environment with high precision and speed, regardless of visual complexity. Within just one glance, 19–67 ms, one can extract enough meaningful information to perform basic-level semantic classifications (e.g., field, ocean, or forest) and classify spatial and functional properties of one’s surroundings (e.g., concealment or navigability; Greene & Oliva, 2009; Kaplan, 1992; Oliva, 2005; Oliva & Torralba, 2001). This ability is referred to as perceiving the gist of a scene (Oliva, 2005). Gist perception plays a crucial role in survival because it drastically constrains the appropriate sequences of actions and contributes to an almost reflexlike selection (Kaplan, 1992). Simply illustrated, it takes a split second to recognize that one can hide well in a forest, while it will take a longer time to identify the specific types of trees. This raises the question: How is the visual system able to extract such a rich percept from the complex and cluttered visual environment in such a short time?

Crucially, our visual world is not just random uncorrelated information—it has structure, regularity, and redundancy (Cohen, Dennett, & Kanwisher, 2016; Field, 1987; Geisler, 2008; Kersten, 1987). The distinct and typical distribution of low-level features of different types of scenes, also called global summary statistics, provides our visual system with cues that summarize the whole environment. Psychophysical evidence has indeed shown that fast scene categorization can be achieved without the recognition of individual objects using only coarse-scale information (oriented blobs with spatial organization), suggesting that coarse features capture the necessary diagnostic information for scene perception (Guyader, Chauvin, Peyrin, Héralt, & Marendaz, 2004; Musel et al., 2013; Musel, Chauvin, Guyader, Chokron, & Peyrin 2012; Schyns & Oliva, 1994; Torralba & Oliva, 2003).

This robust performance despite spatial-frequency information is believed to result from fast feed-forward mechanisms in the visual system, consisting of differently tuned spatial-frequency channels with different temporal dynamics (Bar, 2003; Bullier, 2001; Hegdè, 2008; Hughes, Nozawa, & Kitterle, 1996; Parker, Lishman, & Hughes, 1992; Peyrin et al., 2005; Schyns &...
One elaboration of this concept is the retina-based model (Bar, 2003). In this model the gist percept of a scene is carried by the magnocellular pathway, which has a very robust and fast response to low spatial frequencies even at very low contrasts, making it an ideal candidate for the extraction of coarse-scale information (Derrington & Lenniet, 1984; Ginsburg, 1986; Hughes et al., 1996). Rods and parasol ganglion cells comprising this magnocellular path have a distinct nonuniform distribution throughout the retina. Their density decreases with foveal eccentricity (Curcio & Allen, 1990; Curcio, Sloan, Kalina, & Hendrickson, 1990), while their receptive-field size increases, pooling the activity from more and larger local receptive fields (Bouma, 1970). This nonuniform distribution would therefore predict that the gist of the scene is processed differently depending on its eccentricity. Indeed, an fMRI experiment by Musel et al. (2013) has shown that low-pass-filtered scene images activate the anterior primary visual cortex (V1), which contains more peripheral receptive fields. In contrast to this, high-pass-filtered scene images are processed more posteriorly, closer to the foveal representation of visual space.

A second model, the flexible-use model, states on the other hand that information in the whole frequency spectrum is needed, but that this information is globally pooled and summarized over large spatial areas (Ehinger & Rosenholtz, 2016; Schyns & Oliva, 1994, 1997). The flexible-use model would therefore predict that performance on a gist-of-the-scene task would be eccentricity independent as long as the stimuli are scaled to match spatial-pooling regions.

In this article, we investigate how scene-gist recognition interacts with eccentricity. The involvement of foveal vision (<1°), parafoveal vision (1°–5°) and peripheral vision (>5°) in scene-gist recognition has been addressed in several studies. For example, Van Diepen, De Graef, Lamote, and Van Wijndaele (1994) have shown that scene recognition was only slightly degraded when the foveal information was masked with noise. Expanding on this earlier work, Larson and Loschky (2009) and Larson, Freeman, Ringer, and Loschky (2014) examined the contribution of combined foveal and parafoveal vision (hereafter referred to as central vision) to scene categorization, using the window-and-scotoma paradigm. In this paradigm, the window reveals the central part of the image and blocks peripheral information, whereas the scotoma blocks the central part and reveals only peripheral information. These studies found that scene-categorization performance was practically unaffected when central vision was blocked, whereas blocking peripheral information caused a serious decrease in performance. However, when the dramatic differences in visual content between the two conditions were accounted for, central vision appeared to outperform peripheral vision on a pixel-by-pixel basis (Larson & Loschky, 2009). Furthermore, human observers performed more accurately with central vision for shorter processing times, whereas peripheral vision provided sufficient information only at longer processing times (Larson et al., 2014). Recent studies have confirmed this central advantage, with identical stimuli presented at different eccentricities (Boucart, Moroni, Thibaut, Szaffarczyk, & Greene, 2013; Thibaut, Tran, Szaffarczyk, & Boucart, 2014). Remarkably, however, scene-categorization performance in those studies remained well above chance up to the furthest eccentricity tested (70°), which is in line with the high performance of individuals with central-vision loss due to age-related macular degeneration (Thibaut et al., 2014; Tran, Rambaud, Despretz, & Boucart, 2010).

A number of concerns motivated our experiments. Firstly, the window and scotoma interfere with the global summary statistics, making it hard to interpret the results. For example, regardless of the model of scene perception, a window and a surface-matched peripheral ring may evoke different performances because of their different cuts of the scene image and therefore of its statistics. Secondly, regarding the more recent studies by Boucart et al. (2013) and Thibaut et al. (2014), presenting the same stimulus at different eccentricities and limited to one side of the fovea clearly degrades the typically immersive quality of natural scenes. This approach curtails the potential of scene-processing networks that use correlations reaching across diametrically opposed sides of the fovea. Finally, visual cortical magnification implies that stimuli at larger eccentricities will stimulate less cortical surface. Thus, the apparent central advantage for scene-gist processing may simply reveal a limitation to the extent of the activated network rather than a constraint from peripheral processing per se.

### Experiment 1

The first experiment was designed to dissociate the contributions of global summary statistics, the amount of the activated cortical surface, and the eccentricity dependence of scene-gist perception. To this end, observers performed several different scene-discrimination tasks to ensure a general scene-perception effect. It is well know that different tasks can be resolved using different kinds of diagnostic information. For example, observers can use global information, such as the long horizontal and vertical lines in urban structures compared to more textured zones and undulating contours in natural scenes, to discriminate different types of scene (Oliva & Torralba, 2001). But at the
same time they may also identify individual objects such as windows or cars compared to trees in this same naturalness task. Furthermore, we combined the traditional window-and-scotoma paradigm with conditions that downscaled and upscaled the scene image to different eccentricities. Using scaled annulus presentations, we maintained the immersive quality of the visual scene by stimulating diametrically opposed regions in relation to the fovea.

Next, we compared images presented at different sizes with unbalanced and balanced amounts of activated cortical V1 surface, taking the cortical magnification factor into account. In this way we were able to investigate whether equivalent stimulation, in terms of both stimulus content and amount of activated cortical surface, can lead to differences in performance as a function of eccentricity.

Methods

Participants

Twenty-one participants (14 women, seven men; average age: 24 ± 4 years) were recruited at the Donders Institute. Nine completed all conditions across two sessions; six were not available for the second session and were replaced in conditions CO1 and P2. Participants had normal or corrected-to-normal vision. All participants gave full written consent prior to their participation and were compensated with so-called participation hours. The local ethics committee of the Faculty of Social Sciences of Radboud University approved the experimental procedures (Protocol No. ECSV2016-2208-41) as noninvasive observational experiments with healthy adult human participants.

Stimuli

We selected 256 gray-level photographs from the SUN database (Oliva & Torralba, 2001). Greene and Oliva (2009) have ranked these images on prototypicality in several different categories. We explored the effects of stimulus content and eccentricity for different scene-discrimination tasks, to establish whether our findings are generic for scene-gist processing. Participants therefore performed two global-property classifications (naturalness and concealment) and one semantic, basic-level task (field vs. ocean; Figure 1, Table 1). These tasks were chosen because of the very different strategies that need to be adopted to perform them. Each category contained 50 images, giving us a total of 300 images. The gamma-corrected images measured 256 × 256 pixels, with constant mean luminance and mean contrast for all.

We varied the inner and outer radii of two central conditions, one combined central–peripheral condition, and two peripheral conditions (for a summary, see Figure 2A). In the first central condition (C1), central tunnel vision was simulated using a window with a radius of 5° of the scene, similar to experiments by Larson and Loschky (2009) and Larson et al. (2014). This forced participants to base their scene discrimination solely on central, local information. In the second central condition (C2), the whole scene image was scaled down into a radius of 5°, preserving the global summary statistics, albeit transferred to a higher spatial-frequency range. These two central conditions were compared with the complementary peripheral condition (P1), in which we used an annulus showing only peripheral information, with an outer radius of 40° and an inner radius of 5° that blocked central vision.

A second peripheral condition (P2) and a combined condition (CO1) were added, in which image content and amount of activated cortical area were equated. P2 has an outer radius of 40° eccentricity, and its downscaled version formed CO1. Note that CO1 is thus both a downscaled version of P2 (i.e., including the scotoma) and P2’s complementary. In order to determine the amount of activated cortical V1 surface, we used the following equation based on the retinotopic measurement by Wu, Yan, Zhang, Jin, and Guo (2012):

\[ y = \frac{\int_{r_{\text{inner}}}^{r_{\text{outer}}} (553.99 - 123.98 \ln(x)) \, dx}{7.5} \]

In this formula, \( r_{\text{inner}} \) and \( r_{\text{outer}} \) are the inner and outer radii in degrees, and \( y \) is the integrated V1 surface area in square millimeters, representing the activated cortical surface. With this equation, we found that an annulus with an inner radius of 3.6° and an outer radius of 12° (CO1) activates approximately the same amount of cortical V1 surface as an annulus with an inner radius of 12° and an outer radius of 40° (P2). Figure 2 and Supplementary Table S1 show the amount of activated cortical V1 surface per condition. Image resolution changes dramatically depending on the radius, because of the fixed number of pixels in all scene images (256 × 256 pixels). Therefore the maximum presentable spatial frequency (based on the pixel dimensions) varied from 25.6 c/° in C2 to 10.7 c/° in CO1 and even 3.2 c/° in P1 and P2.

Noise sensitivity

Scene images and noise-mask images were perturbed by adding pink-noise patterns (1/f). Most natural scene images have a 1/f amplitude drop-off, meaning that they have higher power in the lower frequencies (Field, 1987). Therefore, we believe that pink noise provides a more balanced perturbation across the spatial-frequency range of our scene images than, for example,
white noise. White noise effectively masks the higher frequencies of the natural images more, potentially forcing participants to rely more on lower spatial frequencies because they are less masked rather than due to a true preference of processing. Thus, different signal-to-pink-noise ratios (SNRs) allowed us to quantify the amount of information presented equally across all frequencies. This manipulation therefore allowed us to quantify the amount of information necessary for scene-gist perception (see Figure 1).

**Design and procedure**

Stimuli were generated using custom-made software in MATLAB (MathWorks, Natick, MA) employing Psychtoolbox routines (Brainard, 1997). Stimuli were displayed on a 21-in. CRT monitor (Sony GDM-F520) with a resolution of 1,280 × 1,024 pixels, on a mean-luminance background of 21 cd/m² and at a refresh rate of 85 Hz. Participants were seated 24 cm in front of the CRT monitor in an otherwise dark room. Head movements were minimized with a chin rest. Data collection in the eccentricity-scaling conditions was spread over 2 days: We collected conditions C1, C2 and P1 on the first day and CO1 and P2 on the second day. The scene-discrimination task was block-randomized, and participants were told at the beginning of each block which task to perform. The presentation order of

<table>
<thead>
<tr>
<th>Category</th>
<th>Task</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturalness</td>
<td>Urban</td>
<td>The scene is an urban environment.</td>
</tr>
<tr>
<td></td>
<td>Natural</td>
<td>The scene is a natural environment.</td>
</tr>
<tr>
<td>Concealment</td>
<td>Low</td>
<td>You would be easily seen while standing in this scene, and there are not many spots to hide objects.</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>The scene contains many accessible hiding spots, and there may be hidden objects in the scene.</td>
</tr>
<tr>
<td>Semantic</td>
<td>Field</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ocean</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Instructions of the discrimination tasks given to participants.

Figure 1. Example images for the three discrimination tasks. Images in the upper row illustrate the two global-properties discrimination tasks and the basic-level, semantic discrimination task. The lower rows illustrate the effects of the systematic signal-to-noise ratio (in dB) manipulation using pink noise.
the eccentricity-scaling conditions within one day was randomized and counterbalanced across participants. All SNR levels were randomly interleaved in a single block, and each SNR level was repeated 40 times.

Testing began with a practice block of 10 trials to get used to the stimuli. Each trial was preceded by a fixation circle, to indicate the start of the trial and invite participants to achieve steady fixation. After a rest period that varied in length unpredictably between 1,000 and 1,800 ms, the scene image was presented for 150 ms, followed by a dynamic mask (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005) to limit any additional sensory processing after the stimulus presentation. We used the Greene and Oliva (2009) protocol, in which four noise-mask images were presented for 40 ms each. These noise-mask images were generated with the texture-synthesis algorithm designed by Portilla and Simoncelli (2000). This algorithm creates meaningless noise images that conserve marginal and first-order statistics as well as higher order ones while discarding object and spatial-layout information, for optimal masking (Greene & Oliva, 2009). After the masking, participants were required to perform a two-alternative forced-choice task and identify the scene according to its category while maintaining fixation. Depending on the instruction, they were instructed to make this decision as accurately but also as quickly as possible (see Figure 3).

All participants were experienced viewers in psychophysical experiments. Fixation was monitored with an EyeLink II (SR Research Ltd., Ottawa, Canada). The EyeLink II measured eye positions at a sampling rate of 500 Hz with a stated spatial resolution of <0.5°.

**Analyses**

For this experiment, we discuss two benchmarks: the percentage correct at the highest SNR (25 dB), corresponding to maximum performance, and the analysis of the psychophysical fit that effectively is a measurement of the sensitivity to noise. A 5 (between participants: eccentricity scales) × 3 (within participant: scene discrimination) mixed ANOVA was performed in SPSS 25 (IBM) to compare participants’ maximum performances. We used pairwise deletion to deal with the missing values of the replaced participants, which drops the missing values while keeping the cases. This analysis yields a coarse estimation of the effects of eccentricity, but also shows direct differences in difficulty of the three scene-gist tasks. In addition, we re-examined part of the data with a Bayesian t-test in SPSS 25. From this one derives a Bayes factor, in particular for those findings that depend on a significant null hypothesis, by comparing the fit of the data for the null hypothesis with the alternative hypothesis using Bayesian information criteria. The Bayes factor returns an estimate of the likelihood that the null hypothesis is true (H0:H1) and thus the conditions are actually equal (Jeffreys, 1961; Lee & Wagenmakers, 2013). This means that a Bayes factor
of 10 suggests that the data are 10:1 in favor of the null hypothesis—or, rather, that the null hypothesis is 10 times as likely.

For each condition, accuracy data were pooled across participants and fitted (Wallis, Baker, Meese, & Georgeson, 2013) using a maximum-likelihood method (Wichmann & Hill, 2001), with a logistic function defined as

\[ F(SNR; \tilde{a}, \tilde{b}) = \frac{1}{1 + e^{-(\tilde{a} \times SNR + \tilde{b})}} , \]

where \( \tilde{a} \) estimates the slope and \( \tilde{b} \) estimates the true threshold. The standard error was estimated with a nonparametric bootstrap procedure and was used to calculate the 95% confidence intervals (CI\(_{95\%}\)). Participants were not always able to reach 100% accuracy in all conditions due to the short processing time used in this experiment, so lapses were set to maximum performance. Function fitting was implemented in MATLAB using the Palamedes toolbox (Kingdom & Prins, 2009).

Results

Figure 4 summarizes all parameters used to examine eccentricity scale dependence of scene-gist perception. All plots show the pooled data averaged across all participants. The central-vision conditions (C1 and C2) are reported in different shades of green, the peripheral-vision conditions (P1 and P2) are reported in blue, and the combined condition (CO1) is in blue-green.

Maximum performance

A mixed ANOVA was used to examine whether and how maximum performance was affected by eccentricity. Figure 4B shows a significant interaction among the five different eccentricity scales and the three scene-discrimination tasks on maximum performance, \( F(8, 110) = 2.9, p = 0.006, \eta = 0.18 \). Simple main-effects analyses of the scene-discrimination tasks show that participants, on average, score significantly better on the naturalness task (\( M = 0.88, CI_{95\%} [0.86, 0.89] \)) than the concealment task (\( M = 0.81, CI_{95\%} [0.79, 0.83] \)) and the semantic task (\( M = 0.71, CI_{95\%} [0.69, 0.74] \)), \( ps < 0.001 \); and significantly better on the concealment task than the semantic task (\( p < 0.001 \)). Simple main-effects analyses of the eccentricity conditions show that for C1—the central-vision condition using only local information—performance is significantly worse than for both P1 and P2 (\( p < 0.001 \)) in all scene-discrimination tasks. C2, preserving the global summary statistics, also scores significantly worse than P1 and P2 on the naturalness and field-versus-ocean tasks (\( p < 0.001 \)). However, participants’ performance in the conceal-
ment task is significantly better in C2 than C1 ($p = 0.003$), while it is the same as in P1 and P2 ($p = 0.098$ and $p = 0.051$, respectively). Participants therefore perform better using peripheral vision than central vision, even when the global summary statistics are preserved.

However, when comparing CO1 and P2—which both excite approximately equal surface areas in V1—
we observe no significant difference ($p = 0.24$, $p = 0.52$, and $p = 0.65$, respectively, for the naturalness, concealment and field-versus-ocean tasks). To support this finding, we calculated the Bayes factors to estimate the likelihood that performance in these conditions was indeed equal. The Bayes factors were 1.2 for naturalness, 3.5 for concealment, and 3.6 for field vs. ocean, meaning that the null hypothesis was 1.2, 3.5, and 3.6 times as likely. These values indicate no evidence for the naturalness condition, and moderate evidence for the other conditions, that participants' performance is independent of eccentricity when a comparable area of the cortical V1 surface is activated (Jeffreys, 1961).

Psychophysical fits

Figure 4A illustrates the fits to the pooled accuracy data for the various eccentricity conditions, and therefore the effects of noise on scene-gist perception. Each plot shows data from the three discrimination tasks, respectively. The estimated fit parameters and the 95% confidence intervals are reported in Supplementary Table S2. Values of $R^2$, resulting from the goodness-of-fit test, varied between 0.06 and 0.84. We note that condition C1 in the field-versus-ocean task has an $R^2$ of 0.06, which is very low in comparison to the other conditions. Besides a fairly poor goodness of fit, we also find that the fit in Palamedes did not converge on a solution in this particular condition. Any estimated fit parameters are therefore unreliable and will not be included in the analysis.

Figure 4C shows a comparison between the estimated thresholds of these fitted functions. Simply eyeballing the 95% confidence intervals shows us that C1 has a higher threshold than CO1 and the peripheral conditions in the naturalness task, and additionally a higher threshold than C2 in the concealment task. We observe that C2 also has a significantly higher threshold than CO1 and the peripheral conditions. As Figure 4D shows, eccentricity does not consistently affect the slopes. In the naturalness task, we see flatter slopes for CO1 and C1 than for the two peripheral conditions. In the concealment task, we see a flatter slope for C1 and C2 compared to the two peripheral conditions.

Cortical V1 surface

The eccentricity dependence of performance appears to be explained by the extent to which networks are activated. We analyzed this hypothesis directly with a linear regression of the maximum performance and the threshold and the slope to the amount of activated cortical V1 surface in square millimeters (see Figure 5). We find an increase in maximum performance with amount of activated cortical V1 surface: slope $= 0.0003$, $F(1, 220) = 8.13$, $p < 0.001$ (Figure 5A). The threshold decreased with the activated cortical V1 surface: slope $= -0.02$, $F(1, 220) = -3.06$, $p = 0.0024$ (Figure 5B); and we find no linear dependence between the slope and the activated cortical V1 surface: slope $= -0.0012$, $F(1, 220) = -1.56$, $p = 0.12$ (Figure 5C).

Discussion

Limiting the view to only the first 5° of the scene in central vision was found to be especially disruptive, because this significantly disrupts the use of global summary statistics. Scenes are typically characterized by the spatial arrangement of different structures. For example, street scenes normally consist of a road,
buildings, and the sky, arranged in a predictable way (Oliva & Torralba, 2001). This also means that the scene’s diagnostic local information is not always evenly distributed spatially. In our experiments, we chose to minimize the role of unevenly distributed local information by selecting images that could not be recognized based solely on a single object.

On average, participants’ performance in the different discrimination tasks was significantly less accurate and less robust when only local information was available, consistent with earlier work (Larson & Loschky, 2009; Van Diepen et al., 1994). In contrast, when the global summary statics were available, performance was more robust to noise in the naturalness and concealment task, and even more accurate in the concealment task. Nonetheless, downscaling the complete scene into central vision severely limits performance compared to the peripheral presentation conditions.

Central-vision stimulation (C1), however, activates only 318.9 mm² of cortical V1 surface, compared to 857.7 mm² for its peripheral counterpart (P1). When the amount of activated cortical V1 surface is also equated (CO1 vs. P2), we find no performance difference. This result is inconsistent with Larson and Loschky’s report (2009) on the critical radius required to balance performance viewing the scene through a window (central viewing) or with a scotoma (peripheral viewing). Those researchers showed that observers required a critical radius 2.5 times larger than the cortical magnification factor would predict. In contrast to this, we here present evidence showing that observers can process the gist of the scene very precisely and robustly regardless of which eccentricity it is presented at, provided that the scene stimulus has identical diagnostic information and activates a nearly identical area of V1 cortex. Because CO1 and P2 stimuli differ by a factor of about 3.5 in extent, this finding suggests that the spatial-frequency content may change considerably without affecting performance.

Several earlier studies have suggested that processing time for scenes depends on the spatial-frequency content (Greene & Oliva, 2009; Kaplan, 1992; Oliva & Torralba, 2001). For example, hybrid-stimuli experiments, using competing images with different spatial-frequency filtering, have observed spatial-frequency effects for presentations times as short as 30 ms (Joubert et al., 2007; Rousselet et al., 2005; Oliva & Torralba, 2001; Schyns & Oliva, 1994). In addition, Kauffmann et al. (2014) have shown that reactions to a succession of low-to-high spatial-frequency-filtered copies of the same scene occurred more rapidly than for the opposite ordering, in 150 ms. Remarkably, discrimination performance was very similar despite the reaction-time differences. For this reason, we speculate that our presentation time of 150 ms was simply too long to reveal possible eccentricity scale dependencies, because the processing within parallel spatial filters has completed across the entire spatial-frequency range within this presentation time. More precisely, it may be that 150 ms after stimulus onset all the diagnostic information necessary for scene recognition is available, and therefore similar performance is achieved, although through complementary networks stimulated by CO1 and P2. In Experiment 2 we investigated effects of presentation time while equating scene content and the amount of activated cortical surface for two different eccentricity ranges.

### Experiment 2

#### Methods

#### Participants

Twenty participants (15 female, five male; average age: 22.2 ± 6 years) completed the second experiment. They all had normal or corrected-to-normal vision and gave full written consent prior to their participation. The local ethics committee of the Faculty of Social Sciences of the Radboud University approved the experimental procedures (Protocol No. ECSW2016-2208-41).

#### Stimuli

A subset of the stimuli from Experiment 1 was used. The three scene-discrimination tasks (naturalness, concealment, and field-versus-ocean) were assessed for a second time using the conditions CO1 and P2, which ensured a nearly identical amount of cortical activation and resulted in equal performance as in Experiment 1. The inner and outer radii of the central condition were 3.6° and 12°, and for the peripheral condition they were 12° and 40°.

#### Design and procedure

Participants were presented a similar trial structure as in Experiment 1 (see Figure 3), with the exception that in this experiment we varied presentation times; no noise was applied. The presentation times were 17, 50, 100, and 133 ms. Each participant viewed six experimental blocks (2 eccentricity scales × 3 scene-discrimination tasks). Presentation time was block-randomized, and blocks were counterbalanced across participants. For each trial, participants performed a two-alternative forced-choice task, indicating whether the presented image was part of one category or the other. They were instructed to give their responses as accurately and as quickly as possible. All participants...
were experienced viewers, and fixation was monitored with the EyeLink II.

**Analyses**

We analyzed percentage correct depending on presentation time. The average percentage correct was compared in a $2 \times 3 \times 3 \times 4$ repeated-measures ANOVA with within-group variables of scene-discrimination task, eccentricity scaling, and presentation time. In addition, Bayes factors were estimated to give the likelihood that performance in the two conditions was actually equal.

**Results**

Figure 6 summarizes the results of the second experiment, showing the presentation-time dependence of scene-gist perception in the different eccentricity-scaling conditions with approximately equal amounts of activated V1 area (CO1 vs. P2). A repeated-measures ANOVA was used to examine whether and how accuracy and reaction times were affected.

A significant interaction was found between the discrimination task and the eccentricity scaling, $F(2, 36) = 5.58, p = 0.007$. As expected, the simple main effects show that accuracy increases with longer presentation times in all conditions—naturalness: $F(3, 54) = 45.3, p < 0.0001$; concealment: $F(3, 54) = 21.7, p < 0.001$; field-versus-ocean: $F(3, 54) = 47.8, p < 0.001$. Only in the naturalness task did the presentation time interact significantly with eccentricity scaling, $F(3, 54) = 5.5, p = 0.002$. Pairwise comparisons of presentation times between central and peripheral stimulation show that participants scored significantly better with shorter presentation times using peripheral stimulation than central stimulation—17 ms: $t(18) = -5.05, p < 0.001$; 50 ms: $t(18) = -3.00, p = 0.008$; 100 ms: $t(18) = -2.54, p = 0.021$. In the concealment and field-versus-ocean tasks, we observe no significant difference, $-1.1 < t(18) < 1.5, 0.15 < p < 0.98$. In addition, we calculated the Bayes factors to estimate the likelihood that performance with different presentation times was indeed equal. The Bayes factors of the different presentation times for the concealment and field-versus-ocean tasks were all between 3 and 10, except for 17 ms in the concealment task (Bayes factor of 2.0). Therefore, there is moderate evidence that participants perform equally well even at brief presentation times for these tasks.

**Discussion**

Accuracy for peripheral vision was higher than for central vision at shorter processing times, but only in the naturalness task. This suggests that scene-gist perception can benefit from the peripheral, relatively larger receptive fields in this specific task. We did not find differential effects of central or peripheral vision in the concealment or field-versus-ocean tasks. This is in contrast to the only other study on temporal dynamics, to our knowledge, that has shown a moderate but significant central advantage at shorter processing times (Larson et al., 2014). However, as those researchers concluded, their results are not necessarily caused by low-level benefits of central vision. In their experiment, central advantages were found when the scotoma and window conditions were randomly presented. When the probability of peripheral stimulation was increased, the central advantage disappeared. Larson et al. showed with these results that scenes are temporally asymmetically processed in relation to preferential selective attention to the center—that is, in line with the zoom-
out hypothesis. The zoom-out hypothesis states that selective attention proceeds to zoom out from the fovea over the course of a fixation (Eriksen & St. James, 1986; Eriksen & Yeh, 1985; Larson et al., 2014).

In contrast to this, in our experiment we block-randomized the central and peripheral conditions. Participants reported that when stimulus eccentricity scaling changed for a new block, they actively adapted their attention strategy to perform the task in the new block. By block-randomizing the conditions, we gave participants time to change the focus of attention and thus minimize the possible effect of a mismatch between spatial attention and eccentricity scaling. Altogether, the results of this second experiment indicate that peripheral vision can process the scene at least as rapidly and accurately as central vision, and for some tasks it may even outperform central processing.

General discussion

Previous studies have shown that scene perception can be performed rapidly (Greene & Oliva, 2009; Kaplan, 1992; Rousselet et al., 2005; Oliva, 2005; Oliva & Torralba, 2001) using only coarse feature information (Oliva & Schyns, 1997; Schyns & Oliva, 1994). Spatial resolution is known to decrease dramatically with foveal eccentricity, while the activity of local receptive fields is increasingly pooled (Curcio et al., 1990; Curcio & Allen, 1990). This suggests that scene-gist perception performance might be dependent on eccentricity. Yet our experiments demonstrated that scene-gist perception could be accomplished with similar accuracy independent of eccentricity if the proper diagnostic information is provided and a comparable amount of cortical surface activated.

We found in Experiment 1 that scene-gist perception is significantly degraded when only a small fraction of the image is presented in central vision. This result is consistent with those of Larson and Loschky (2009), who report that scene perception is moderately but significantly disrupted when the central 5° of the scene is presented as opposed to the whole scene. Using a window and a scotoma, as in their study, enables simulation of visual-information acquisition during the course of a single fixation in a very naturalistic way. This method has been used to investigate eye-movement behavior during scene exploration, for example (Van Diepen et al., 1994). We stress, in line with current scene-perception models, the importance of global structure or spatial relationships in scene-gist perception (Bar et al., 2006; Oliva & Torralba, 2001). The window and the scotoma present different cuts of the scene and will therefore inherently change available spatial information, disrupting the previously mentioned parameters. We therefore conclude that for a fair comparison, observers should be provided with an equal amount of diagnostic information to perform the scene-perception task. Indeed, our finding that scene discrimination tolerates the complete downscaling into central vision better than selecting only the central part is in agreement with the importance of providing proper diagnostic information. We suspect that the poor performance for the central window in previous studies is thus a consequence of the method rather than an inherent eccentricity limitation on scene discrimination of the visual system.

We furthermore noticed a task dependence in both our experiments that we interpret as follows. First, we note that the tasks may allow opportunism: Some tasks can be solved using different kinds of diagnostic information, be it the global summary statistics or local information. Naturalness discrimination, for example, has been frequently reported to be an easy task (Boucart et al., 2013; Greene & Oliva, 2009), because it can be performed with both coarse global summary statistics and diagnostic local information. Alternatively, opportunism may apply regarding spatial location rather than type of information; the natural- versus-urban discrimination task shows high spatial stationary behavior of the statistics (Oliva & Torralba, 2001), meaning that all across the image each region is about equally informative in performing the task. Indeed, Oliva and Torralba (2001) have shown that for naturalness discrimination, all spatial regions are about equally predictive.

In contrast, the concealment and field-versus-ocean tasks did not appear to offer such possibilities. Performance in the concealment task was significantly reduced when participants were shown only the central 5°, demonstrating the need for global scene information. Likewise, participants barely performed above chance level in the field-versus-ocean task even when the entire scene was downscaled. It appears that this task was a very difficult one under our conditions, and it might be possible that our choice of grayscale images and images without clear objects removed the diagnostic information that is normally used for this task, such as characteristic objects or color. The task dependence shows us that different tasks can be accomplished using different kinds of information. We thus emphasize the importance of presentation of proper diagnostics in comparing scene performance.

Interestingly, scene discrimination is still less effective when the full image is shown centrally instead of to the periphery. This would appear to be in line with the previously mentioned scene-perception models claiming that scene perception follows a coarse-to-fine sequence (Bar, 2003; Bullier, 2001; Hegdé, 2008; Hughes et al., 1996; Parker et al., 1992; Peyrin et al., 2005; Schyns & Oliva, 1994). Low spatial frequencies...
have temporal precedence over high spatial frequencies, as shown by hybrid images and image-presentation sequences from low-pass to high-pass (Kauffmann et al., 2014; Musel et al., 2012; Schyns & Oliva, 1994). The small receptive fields in central vision are tuned to higher spatial frequencies and finer details than in the periphery, which has a high sensitivity to lower spatial frequencies and coarser visual structures (DeValois, & DeValois, 1988). The periphery processes this coarser information faster than central vision processes the finer details. Thus, for the same presentation time, images presented in the periphery have effectively been presented longer.

Yet subjects perform equally well when the amount of activated cortical surface for presentations at different eccentricities is accounted for. This assertion holds even down to 17-ms presentation times in most tasks and indicates an eccentricity scale independence. This scale independence challenges the low-spatial-frequency focus in the coarse-to-fine sequence model, and highlights the fact that coarse and fine distinctions are not equivalent to low- and high-spatial-frequency distinctions. Oliva and Torralba (2006) emphasize that low spatial frequencies are not necessarily preferred in the early stages of visual processing. In fact, those studies show that the visual system can selectively choose a suitable spatial scale depending on the task (Oliva & Schyns, 1997; Oliva & Torralba, 2006; Schyns & Oliva, 1997). For this reason, our observations could be better explained by this multiscale representation (Oliva & Torralba, 2006).

In summary, we raised concerns about previous studies’ conclusion that gist perception benefits from central vision. We addressed our two main concerns by first presenting the same diagnostic information independent of stimulus eccentricity. Second, we compared conditions with equal activated cortical surface and, therefore, presumably similar amounts of pooled numbers of local receptive fields. We demonstrated that under these circumstances, performance on scene-gist discrimination tasks was similarly high independent of eccentricity.

**Keywords:** scene perception, peripheral vision, central vision, coarse-to-fine, eccentricity independence

## Acknowledgments

This work was supported by the European Union FP7 Marie Curie IDP Grant (FP7-PEOPLE-2013-ITN). The authors wish to acknowledge the work of Rowanne Steiner, who helped carry out the experiments. The authors also wish to thank Sinziana Pop and Milena Kästner for proofreading the final version of the manuscript.

Commercial relationships: none.

Corresponding author: Anna C. Geuzebroek.

Email: A.Geuzebroek@donders.ru.nl.

Address: Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroscience, Radboud University, Nijmegen, the Netherlands.

## References


