Attention, Perception, & Psychophysics
2010, 72 (2), 427-444
doi:10.3758/APP.72.2.427

The role of higher order image statistics in masking scene gist recognition

Lester C. Loschky
Kansas State University, Manhattan, Kansas

Bruce C. Hansen
Colgate University, Hamilton, New York

Amit Sethi
University of Illinois at Urbana-Champaign, Urbana, Illinois

and

Tejaswi N. Pydimarri
Kansas State University, Manhattan, Kansas

In the present article, we investigated whether higher order image statistics, which are known to be carried by the Fourier phase spectrum, are sufficient to affect scene gist recognition. In Experiment 1, we compared the scene gist masking strength of four masking image types that varied in their degrees of second- and higher order relationships: normal scene images, scene textures, phase-randomized scene images, and white noise. Masking effects were the largest for masking images that possessed significant higher order image statistics (scene images and scene textures) as compared with masking images that did not (phase-randomized scenes and white noise), with scene image masks yielding the largest masking effects. In a control study, we eliminated all differences in the second-order statistics of the masks, while maintaining differences in their higher order statistics by comparing masking by scene textures rather than by their phase-randomized versions, and showed that the former produced significantly stronger gist masking. Experiments 2 and 3 were designed to test whether conceptual masking could account for the differences in the strength of the scene texture and phase-randomized masks used in Experiment 1, and revealed that the recognizability of scene texture masks explained just 1% of their masking variance. Together, the results suggest that (1) masks containing the higher order statistical structure of scenes are more effective at masking scene gist processing than are masks lacking such structure, and (2) much of the disruption of scene gist recognition that one might be tempted to attribute to conceptual masking is due to spatial masking.

Scene gist recognition—typically operationally defined as scene categorization following a brief presentation (Kaping, Tzvetanov, & Treue, 2007; Loschky et al., 2007; McCotter, Gosselin, Sowden, & Schyns, 2005; Oliva & Schyns, 2000; Rousselet, Joubert, & Fabre-Thorpe, 2005)—may be the first meaningful stage of scene perception, possibly occurring even before the recognition of constituent objects in a scene (Oliva, 2005; Oliva & Torralba, 2001). Scene gist activates prior knowledge, which influences scene processing, including attention (Eckstein, Drescher, & Shimozaki, 2006; Gordon, 2004; Loftus & Mackworth, 1978; Torralba, Oliva, Castelhano, & Henderson, 2006), possibly object recognition (Boyce & Pollatsek, 1992; Davenport & Potter, 2004; De Graef, De Troy, & d'Ydewalle, 1992; Hollingworth & Henderson, 1998; Palmer, 1975), and memory (Brewer & Treyens, 1981; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989). Importantly, scene gist recognition is rapid, with performance improving over the first 40-msec stimulus onset asynchrony (SOA) of masked presentation, and near perfect performance after a 100-msec SOA (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Fei-Fei, Iyer, Koch, & Perona, 2007; Loschky et al., 2007). The incredible speed of scene gist processing suggests that it utilizes very low-level image features that are perhaps processed in parallel (Rousselet, Fabre-Thorpe, & Thorpe, 2002). This raises a key question: To what degree are second- versus higher order image statistics used to rapidly recognize scene gist? Below, we will briefly review the nature of second- and higher order image statistics, followed by the hypotheses and design of the present study.

The Amplitude Spectra of Scenes and Gist Recognition

One approach to understanding the low-level structure of scene images comes from the global 2-D discrete Fou-

L. C. Loschky, loschky@ksu.edu

rier transform (DFT). The 2-D DFT treats an image as a complex 2-D luminance waveform, represented as the sum of sinusoidal waveforms of different amplitudes, frequencies, orientations, and phases. The amplitude that is plotted as a function of spatial frequency and orientation is often referred to as the *amplitude spectrum*. The phase angles, plotted as a function of spatial frequency and orientation, are represented by the *phase spectrum* (Shapley & Lennie, 1985; Smith, 2007). Both the amplitude and phase spectra are argued to be important in representing image structure (Hansen, Essock, Zheng, & DeFord, 2003; Tadmor & Tolhurst, 1993). Specifically, the global amplitude spectrum of a scene only contains information about contrast as a function of spatial frequency and orientation, without regard to their image locations (i.e., the spatial domain). Thus, without *any* contrast information, no image structure would be visible. On the other hand, the phase spectrum of a scene determines where different image frequencies are aligned (Kovesi, 1999; Marr, 1982; Morrone & Burr, 1988; Morrone & Owens, 1987; Wang & Simoncelli, 2004), thereby determining the formation of local image structures in an image. Thus, an image possessing systematic phase coherence possesses localized image structure. Conversely, an image lacking phase coherence (e.g., a phase-randomized image) would have its contrast as a function of spatial frequency and orientation randomly distributed across the image; that is, it would be an image with nonlocalized structure.

Several recent computational models of scene categorization have proposed using spatially unstructured scene information (i.e., their global amplitude spectra, hereafter referred to as second-order image statistics[1]), in order to categorize them (e.g., Gorkani & Picard, 1994; Guerin-Dugue & Oliva, 2000; Guyader, Chauvin, Peyrin, Hérault, & Marendaz, 2004; Herault, Oliva, & Guerin-Dugue, 1997; Oliva & Torralba, 2001). The idea that amplitude spectrum statistics are important for gist recognition was supported by Oliva and Torralba's (2001) spatial envelope computational model, which achieved 86% recognition accuracy using only second-order scene statistics, as compared with a 12.5% chance level. Although the spatial envelope model is a computational model and there is thus no reason to expect that the human brain would necessarily use similar mechanisms, some have gone as far as to argue that "the amplitude spectrum alone is sufficient for natural scene categorization" (Guyader et al., 2005, p. 5642). For example, Guyader et al. (2004) compared the priming of scene gist by normal scenes and by phase-randomized scenes, and found that both were equivalent. This was important because phase-randomized scenes contain the same second-order statistics as do normal scenes, but none of the higher order statistics carried in the phase spectra. Additionally, Kaping et al. (2007) reported results that were interpreted as supporting the hypothesis that second-order image statistics are sufficient for scene categorization.

In contrast, other research has questioned the usefulness of second-order image statistics for gist recognition. Several studies have shown that incrementally destroying the higher order image statistics of scenes by incrementally randomizing their phase spectra, while preserving their second-order image statistics, reduces gist recognition from near perfect to chance levels (Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009; Loschky & Larson, 2008; Loschky et al., 2007). Furthermore, Loschky et al. (2007) provided converging evidence by building on the long history of using masking to study the structure of spatial vision (Carter & Henning, 1971; de Valois & Switkes, 1983; Henning, Hertz, & Hinton, 1981; Legge & Foley, 1980; Losada & Mullen, 1995; Sekuler, 1965; Solomon, 2000; Stromeyer & Julesz, 1972; Wilson, McFarlane, & Phillips, 1983). They used a scene gist masking paradigm that systematically disrupted the higher order image statistics (by incremental phase randomization) of scene image masks, which incrementally reduced gist masking strength. Going a step further, Loschky et al. (2007) found that masking a target scene by a fully phase-randomized version of itself (which thus had identical second-order statistics to the target) caused no more scene gist masking than when the target scene was masked by a fully phase-randomized version of a scene from a different category (thus having different second-order statistics). These results suggest that (1) higher order image statistics carried in the phase spectrum are important for categorizing scenes, and (2) the differences in the second-order image statistics of scenes from different scene categories are not particularly useful for recognizing scene gist.

The Present Study

The present study tested the hypothesis that the higher order statistics of scenes are necessary for scene gist recognition. Fully phase-randomized scenes lack higher order scene statistics (Thomson, 1999, 2001a) and are unrecognizable (Joubert et al., 2009; Loschky & Larson, 2008; Loschky et al., 2007). Thus, testing our hypothesis required comparison stimuli that, in contrast with phase-randomized images, did contain higher order statistical relationships such as those in natural scenes (Thomson, 2001a, 2001b), but, similar to phase-randomized images, were not recognizable. Texture images, generated from scenes using the texture synthesis algorithm of Portilla and Simoncelli (2000), seemed to fit both requirements.[2] Their algorithm analyzes an input image using wavelets at different spatial frequency bands, orientations, and locations, to derive a higher order statistical model of texture in that image. It then takes random noise and iteratively coerces it to match the statistical parameters of its model of the input image. Because it is a wavelet model, it uses higher order statistical information, encoding the edges, contours, and spatial patterns of textures. Furthermore, we can globally quantify the higher order statistics of such textures using the recently developed phase-only second spectrum measure (Seidler & Solin, 1996; Thomson, 2001a, 2001b). We therefore created texture images to use as masks, using real-world scenes from different categories as inputs to the texture synthesis algorithm.

Although the Portilla and Simoncelli (2000) algorithm was not designed to construct textures from scenes, when it is applied to scenes or other "images that are structured and highly inhomogeneous," it will "capture the local structure of the original images, albeit in a globally disorganized

fashion" (pp. 62–63). However, if most scene images are relatively spatially inhomogeneous—that is, have an irregular layout—then texture images that are generated from scenes might be expected to create unrecognizable textures. Figure 1 shows three example scenes and texture images generated from them, and seems to bear out this intuition. This is important because an alternative explanation for the reduced masking produced by phase-randomized scenes as masks, as compared with normal scenes, is that phase-randomized scenes are less recognizable; that is, more recognizable masks produce more masking—a phenomenon known as *conceptual masking* (Bachmann, Luiga, & Põder, 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Loschky et al., 2007; Potter, 1976). However, masking produced by unrecognizable texture masks could not be explained that way. Conversely, the texture algorithm

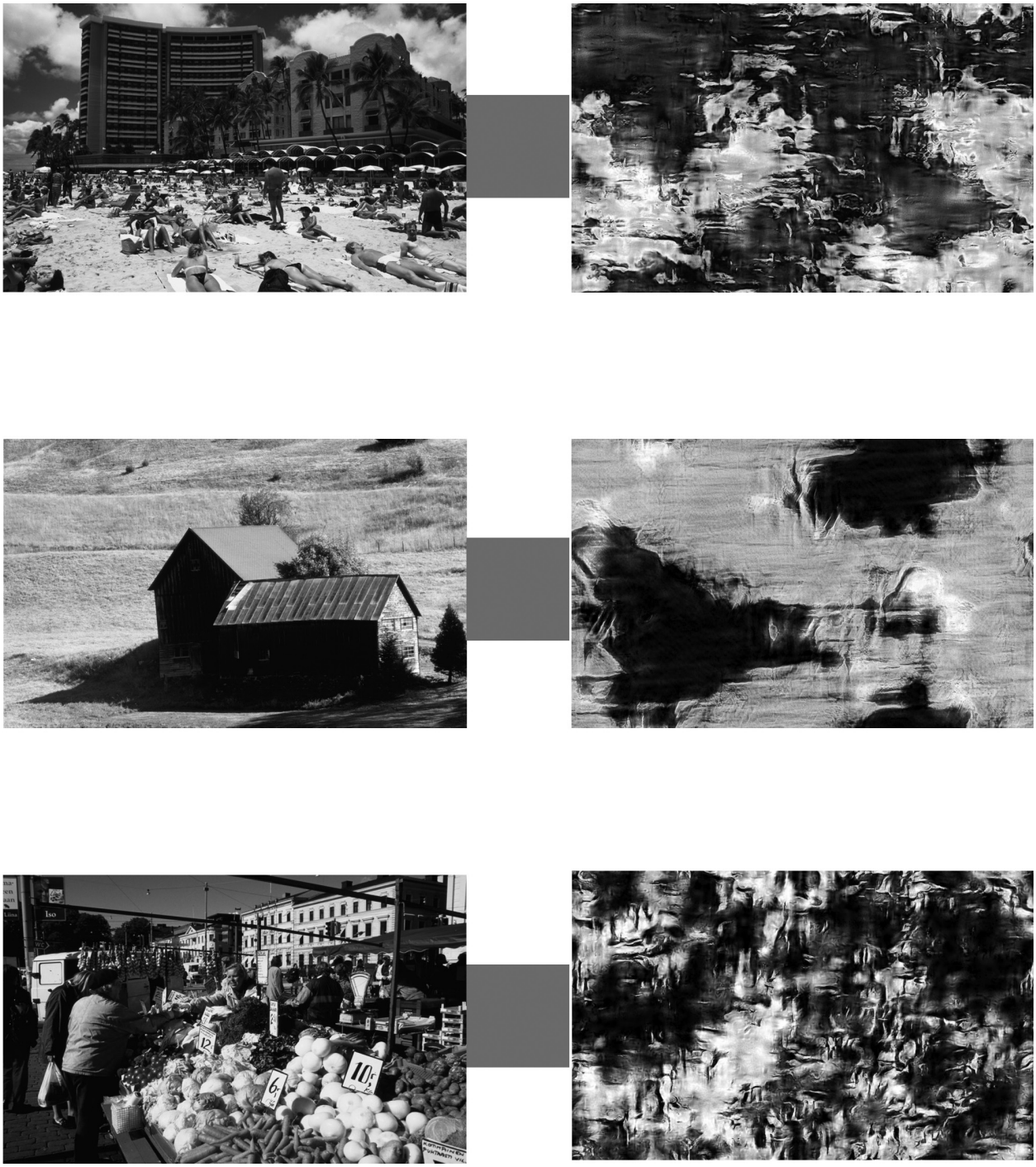

Figure 1. Three example scenes and the scene textures derived from them using the texture synthesis algorithm (Portilla & Simoncelli, 2000).

will generate images with higher order statistics, such as those in natural scenes (see Figure 2). Thus, such texture images might be very powerful masks in a scene categorization task. That is, if (1) scene texture masks contain more higher order statistics than do phase-randomized scenes, and (2) both scene texture masks and phase-randomized scene masks are equally unrecognizable, then (3) we can compare masking by these two types of images to determine the roles of higher order statistical structure versus conceptual masking in explaining the effects of phase randomization on scene gist masking.

The present study used visual masking to test whether a particular form of spatial information—higher order scene statistics—was useful for a given visual task, scene gist recognition, which is similar to the use of masking in spatial vision masking studies to test the multichannel model of vision (Carter & Henning, 1971; de Valois & Switkes, 1983; Henning et al., 1981; Legge & Foley, 1980; Losada & Mullen, 1995; Sekuler, 1965; Solomon, 2000; Stromeyer & Julesz, 1972; Wilson et al., 1983). The effects of higher order image statistics on scene categorization were then contrasted with the effects of conceptual masking (Bachmann et al., 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976).

In Experiment 1, we compared the scene gist masking produced by masks varying in their second- and higher order statistics: normal scenes, scene textures, phase-randomized scenes, and white noise. We also varied

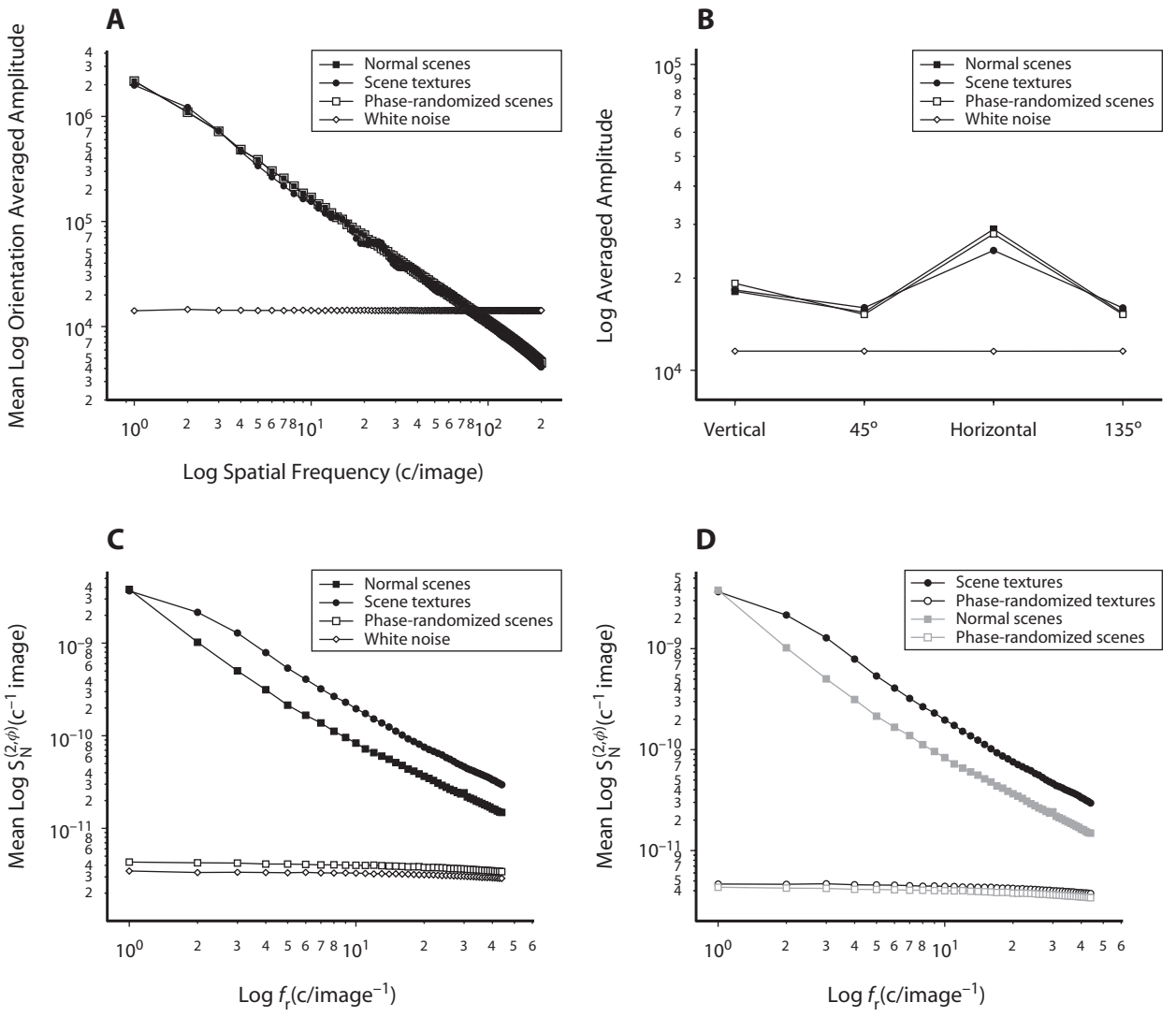

Figure 2. Second-order and higher order statistics of masks used in Experiment 1. (A) Amplitude spectra (averaged over orientations) for normal scene, scene texture, phase-randomized scene, and white noise masking images in Experiment 1. (B) Average orientation (averaged over spatial frequencies) in four bands (0º, 45º oblique, 90º, and 135º oblique) for the normal scene, scene texture, phase-randomized scene, and white noise masking images in Experiment 1. (C) Second spectra (averaged over orientations) for the normal scene, scene texture, phase-randomized scene, and white noise masking images in Experiment 1. (D) Second spectra (averaged over orientations) for the scene texture and phase-randomized scene texture masking images used in the control experiment for Experiment 1. For purposes of comparison, the normal scene and phase-randomized scene masking conditions from Experiment 1 are also shown in gray.

processing time (i.e., target-to-mask SOA). The results showed a masking strength hierarchy of scenes > scene textures > phase-randomized scenes > white noise. In a control study, we compared masking by scene textures and their phase-randomized versions and showed that the former caused stronger masking.

In Experiment 2, we tested the hypothesis that conceptual masking caused the stronger masking by scene textures rather than by phase-randomized scenes in Experiment 1 because the scene textures were significantly more recognizable than were the phase-randomized scenes. However, the results failed to support this hypothesis.

In Experiment 3, we tested the hypothesis that more recognizable scene textures caused more gist masking, and we showed that scene texture recognizability explained only 1% of the masking variance. This again failed to support the hypothesis that the greater masking by texture images rather than by phase-randomized images in Experiment 1 was caused by conceptual masking. Thus, the remaining alternative explanation of Experiment 1's results was in terms of differences in the masks' higher order scene statistics.

EXPERIMENT 1

In Experiment 1, we tested the hypothesis that higher order statistical structures in scenes—such as edges, lines, and combinations of both in unique spatial configurations—are necessary for recognizing scene gist. We examined this via the scene gist masking paradigm of Loschky et al. (2007, Experiment 3), with four mask types having different degrees of second- and/or higher order image statistics: normal scenes, scene textures, fully phase-randomized scenes, and white noise. We assumed that the second-order statistics of all mask types but the white noise would be quite similar, whereas the higher order statistics of the normal scenes and scene textures would differ from the phase-randomized scenes and white noise. On the basis of these assumptions, we hypothesized that (1) if higher order statistical structure is necessary for scene gist recognition, scene texture masks should disrupt scene gist more than should phase-randomized scene masks, and (2) if such higher order statistical structure is sufficient for scene gist recognition, scene texture masks should disrupt scene gist as much as or more than normal scenes as masks.

In Experiment 1, we also investigated the time course of scene gist processing associated with each of the above types of masks. We varied the SOA between the target and mask images to determine when each form of information is more or less useful for disrupting scene gist. Thus, the more important a given form of information is at a given point in processing, the stronger the masking caused by it should be.

Method

Participants. A total of 120 Kansas State University students (64 female; mean age = 18.93 years) participated for course credit. All had an uncorrected or corrected near acuity of 20/30 or better.

Stimuli. We started with 300 monochrome scene images (1,024 × 674 pixels)[3] that were divided equally among 10 basic level categories: 5 natural (beach, desert, forest, mountain, and river) and 5 man-made (farm, home, market, pool, and street). From these images, we created 300 synthesized texture versions, using the Portilla and Simoncelli (2000) texture synthesis algorithm. We also created a set of fully phase-randomized versions of the original 300 scene images using the RISE algorithm with a randomization factor of 1 (for more details, see Loschky et al., 2007, Appendix A; Sadr & Sinha, 2001, 2004). We then created 300 white noise images. We then equalized the mean luminance and RMS contrast of the complete set of 1,200 normal scenes, scene textures, fully phase-randomized scenes, and white noise images (luminance $M = 107.83$, RMS contrast = 19.33) (see Loschky et al., 2007, Appendix B).

Figure 2A shows a plot of the mean orientation-averaged amplitude spectra of the four image categories (normal scenes, scene textures, fully phase-randomized scenes, and white noise). This shows that the orientation-averaged amplitude spectra were essentially the same for all mask types except white noise. We quantified this by calculating the slope of the orientation-averaged amplitude spectra (on logarithmic axes) for each of the four mask sets on the central 674 × 674 pixel regions of the stimuli.[4] The average slopes for the four sets were: normal scenes, $M = -1.29$, $SD = 0.15$; phase-randomized scenes, $M = -1.30$, $SD = 0.14$; scene textures, $M = -1.29$, $SD = 0.13$; and white noise images, $M = -0.001$, $SD = 0.007$. There were no significant differences (independent t tests, $df = 598$) between normal scenes, phase-randomized scenes, or scene textures ($p > .05$), and all three were significantly different from white noise masks ($p < .001$). To test for differences in amplitude as a function of orientation, we analyzed orientation bias by averaging the amplitude coefficients within a 45º band of orientations (across all spatial frequencies) centered on four orientations (vertical, 45º oblique, horizontal, and 135º oblique) for each of the mask sets used in Experiment 1 (refer to Hansen et al., 2003, for further details). The results are shown in Figure 2B. There were significant main effects of orientation for normal scenes [$F(3,897) = 593.8$, $p < .001$, $\eta_p^2 = .67$], scene textures [$F(3,897) = 303.5$, $p < .001$, $\eta_p^2 = .51$], and phase-randomized scenes [$F(3,897) = 535.4$, $p < .001$, $\eta_p^2 = .64$] (see note 4), but not for white noise [$F(3,897) = 1.4$, $p > .05$, $\eta_p^2 = .005$]. All other mask types were significantly different from white noise ($p < .001$), with $\eta_p^2 > .80$ for all comparisons. A 2 (normal scenes vs. scene textures) × 4 (orientation) repeated measures ANOVA showed a significant main effect of image type (i.e., normal scenes vs. scene textures) [$F(1,598) = 15.2$, $p < .001$], but with a very small effect size ($\eta_p^2 = .025$), and a significant interaction [$F(3,1794) = 44.5$, $p < .001$] that also had a very small effect size ($\eta_p^2 = .069$). Thus, although the large number of images contributed to the observed significance, the size of the effects was quite small and, as is shown in Figure 2B, the orientation bias between normal scenes and scene textures is quite similar, with a slightly smaller horizontal bias in the scene textures. Therefore, given the aforementioned similarities, any differences in masking between the normal scenes, fully phase-randomized scenes, or scene textures cannot be easily explained in terms of the second-order statistics of their amplitude spectra.

We compared the higher order statistics of the different mask types by measuring their phase-only second spectra (Seidler & Solin, 1996; Thomson, 2001a, 2001b; see the Appendix of the present study for details). Figure 2C shows the mean orientation-averaged phase-only second spectra for each of the four mask types. The phase-only second spectrum is a global measure of the phase-alignment of pairs of spatial frequencies as a function of their frequency differences, or offsets. Thus, a phase-only second spectrum slope (on logarithmic axes) that is significantly greater than 0 indicates phase alignments (Hansen & Hess, 2007; Morrone & Burr, 1988; Morrone & Owens, 1987) that create edges, lines, contours, and unique spatial configurations in an image (Thomson, 2001b). Furthermore, the area under the phase-only second spectrum curve is a measure of sparseness. Thus, the more area under the phase-only second spectrum curve, the more efficient image coding should be (Thomson, 2001a, 2001b). Figure 2C shows very similar slopes for the normal scenes and scene textures, both of which are very different from the fully phase-randomized scenes and white noise, which are basically 0 since they have no phase alignments. The scene-texture images,

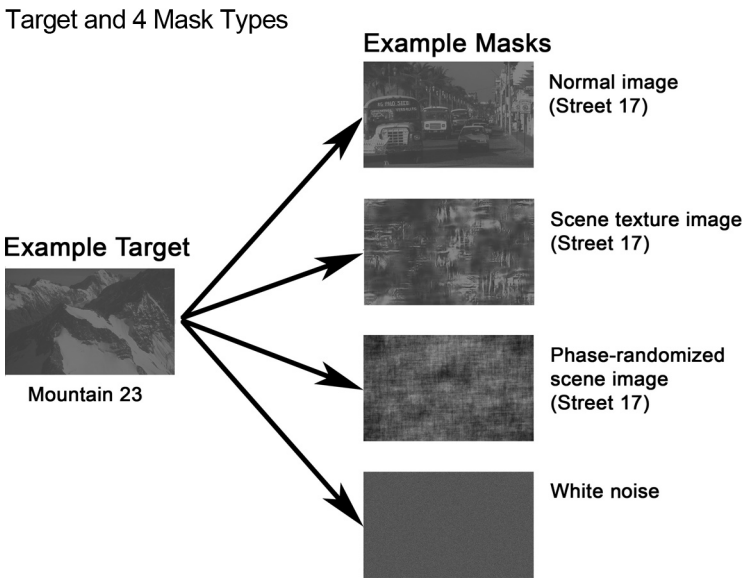

Figure 3. Experiment 1: Example target and four mask types—normal scenes, scene textures, phase-randomized scenes, and white noise. Note that the white noise condition is difficult to resolve at the small spatial scale of the figure, but that it was quite visible in the actual experiment.

derived from wavelet-filtered scenes, have somewhat higher phase-only second spectra than do their original scenes, which is consistent with the finding that wavelet filtering increases scene image sparseness (Thomson, 2001a). We calculated the slope of the orientation-averaged phase-only second spectra for each of the four mask sets: normal scenes, $M = -1.22$, $SD = 0.16$; phase-randomized scenes, $M = -0.08$, $SD = 0.02$; scene textures, $M = -1.27$, $SD = 0.12$; and white noise images, $M = -0.06$, $SD = 0.02$. The slopes of scene texture and phase-randomized scene masks were significantly different [$t(598) = 158$, $p < .001$], with a large effect size of $d = 17.0$. The slopes of the scene textures and the normal scenes were also significantly different [$t(598) = 4.7$, $p < .001$], with an average difference of 0.058, producing a small to moderate effect size of $d = 0.44$. Their area under the curve was significantly different as well [$t(598) = 7.58$, $p < .001$]. Thus, if the phase-only second spectrum contains a useful structure for scene gist recognition, scene textures should produce stronger gist masking than should phase-randomized scenes, and different masking than should normal scenes.

Images were shown on a 17-in. Gateway EV910 monitor (85 Hz refresh rate) at a screen resolution of 1,024 × 768, with the viewing distance fixed at 53.3 cm by a chinrest. Each image subtended a visual angle of 34.39º × 27.11º.

Design and Procedures. As is shown in Figure 3, there were four types of masks: (1) a normal image from a different scene category than that of the target, (2) a scene texture from a different category than that of the target, (3) a fully phase-randomized image from a different category than that of the target, and (4) a white noise image. We also included a no-mask condition as a baseline. Mask type (including no-mask) was a randomly assigned between-subjects variable (24–25 participants each). The target-to-mask pairings were yoked across mask conditions.

Figure 4, left panel, shows a schematic of a trial in this experiment. Participants first saw a fixation cross, prompting them to press a button to start the trial. Then, 750 msec later, they saw a target image that was flashed for 12 msec. After a variable interstimulus interval (ISI; 0–82 msec; SOA of 12–94 msec), a mask appeared for 35 msec, producing a relatively strong 3:1 mask:target duration ratio. After a 750-msec blank interval, a cue word (from the 10 target categories) was presented, either validly or invalidly describing the preceding target image. Using a keyboard, participants pressed the "YES" key if they thought the cue validly described the target, or the "NO" key if it did not. A random half of the trials had valid cues, with cue validity equal across all target and cue categories. Participants were encouraged to respond as quickly and as accurately as possible, and to go with their first impression if unsure. There were 300 self-paced trials, with each image presented once in a random order. Before the experiment, participants were familiarized with the 10 scene categories and the experimental task by seeing a separate set of 90 labeled scene images for 1 sec each, and then doing 32 practice trials, without feedback.

For the repeated measures analyses throughout the entire study in which the assumption of sphericity was violated, we used the reasonably conservative Huynh–Feldt epsilon correction to adjust all degrees of freedom.

Results

Figure 5 shows scene gist recognition accuracy as a function of the masking conditions and processing time (SOA). As is shown in the figure, the mask types greatly differed in their gist masking strength [$F(3,93) = 83.15$, $p < .001$].

We carried out planned comparisons in order to test the differences between adjacent pairs of masking conditions (in the order: normal scene, scene texture, phase-randomized scene, white noise, no-mask). Scene texture masks produced stronger scene gist disruption than did fully phase-randomized scene masks ($\Delta = 0.047$, $SE = 0.017$, $p = .007$), consistent with the hypothesis that higher order statistical scene structure, as measured by the phase-only second spectrum, is necessary for scene gist recognition. Specifically, gist masking strength decreases when higher order structure is eliminated by phase randomization and increases when higher order structure is present.

Nevertheless, scene texture masks produced weaker scene gist disruption than did normal scene masks ($\Delta = 0.057$, $SE = 0.017$, $p = .001$). Although this is consistent with the fact that the two conditions differed in their

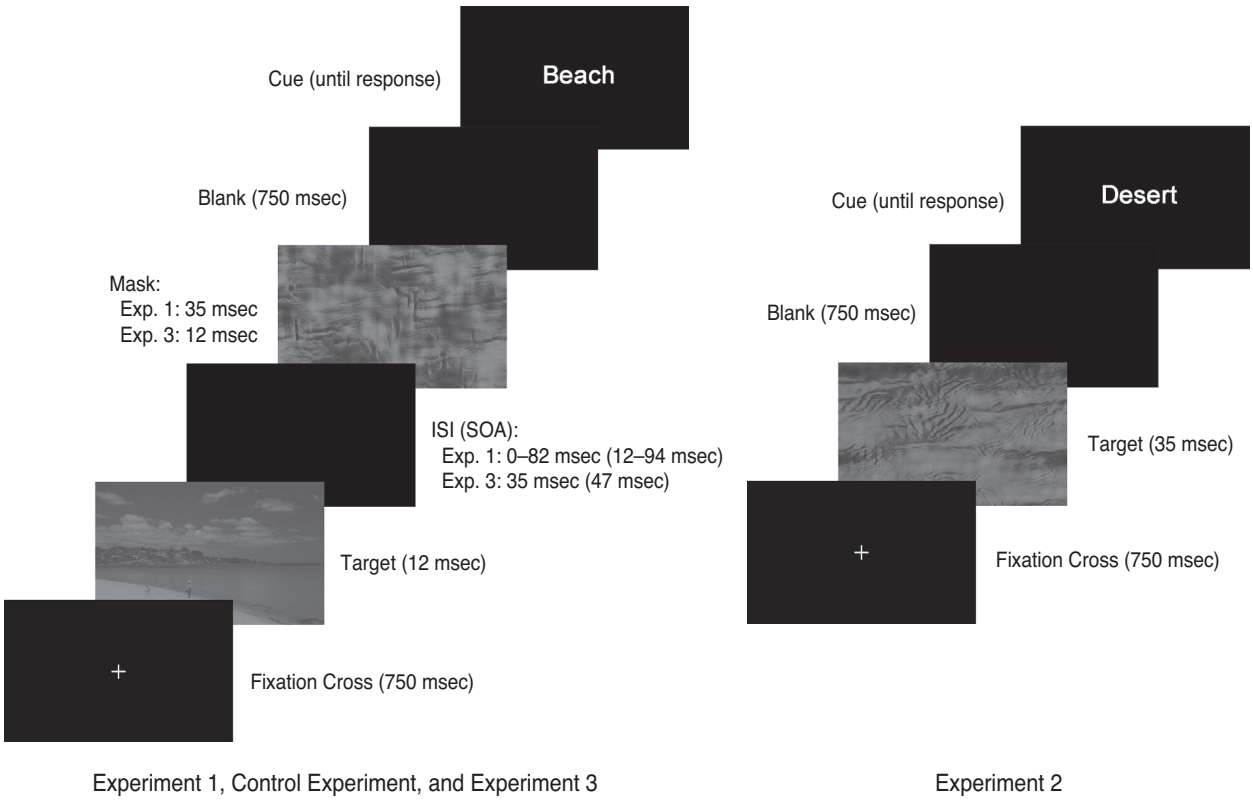

Figure 4. Trial schematics for Experiments 1–3. Left: Schematic for Experiment 1, control experiment, and Experiment 3, which involved masking. Right: Schematic for Experiment 2, in which targets were unmasked.

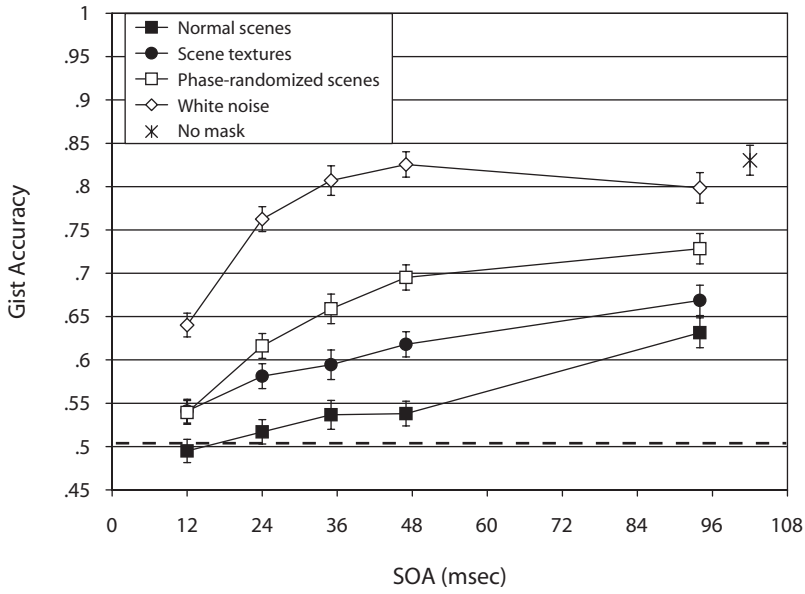

Figure 5. Experiment 1 data: Scene gist recognition accuracy as a function of the four different mask types (normal scenes, scene textures, phase-randomized scenes, and white noise) and processing time (stimulus onset asynchrony, SOA), and the no-mask condition (target = 12 msec; mask = 36 msec). Error bars represent standard errors of the means.

phase-only second spectra, it also suggests a possible nonlinear relationship between phase-only second spectrum magnitude and masking. It is also consistent with the hypothesis that the higher order statistical structure that is measured by the phase-only second spectrum is not sufficient for full scene gist recognition—that is, some information that is not measured by the phase-only second spectrum, which is missing from the scene texture masks but which is contained in normal scene masks, is necessary for normal scene gist recognition.

Furthermore, phase-randomized scene masks produced stronger gist disruption than did white noise masks ($\Delta = 0.119$, $SE = 0.017$, $p < .001$) (Loschky et al., 2007), consistent with the hypothesis that the second-order statistics of scenes carry some information that is useful for scene gist, perhaps because they are dominated by lower spatial frequencies. Since the normal scene masks and scene texture masks (both of which have similar amplitude spectra to the phase-randomized scene masks) also share the same masking strength advantage over white noise, this suggests that part of their strength may derive from their second-order statistics as well.

Finally, the white noise masks disrupted scene gist recognition significantly more than did the no-mask condition ($\Delta = 0.064$, $SE = 0.017$, $p < .001$), indicating that even masks sharing no statistical structure with scenes caused some degree of scene gist disruption.

Figure 5 also shows an expected effect of processing time (i.e., SOA) [$F(3.85, 358.36) = 100.96$, $p < .001$]. Interestingly, there was also a clear interaction between mask type and processing time [$F(11.56, 358.36) = 6.71$, $p < .001$]. Part of this may be due to a floor effect at the shortest SOA (12 msec). Thus, the apparently equal masking strength of scene texture masks and phase-randomized scene masks at a 12-msec SOA should be treated with caution. Using similar methods, Loschky et al. (2007, Experiment 4) found a similar floor effect at a 12-msec SOA, but they showed that differences in masking strength appeared there when a weaker mask:target duration ratio (e.g., 1:1) was employed. However, a more informative aspect of this interaction is the small effect of processing time in the normal scene mask condition, which reflects its very strong masking—except at the longest SOA (96 msec)—in contrast with the small effect of processing time in the white noise masking condition, reflecting its weak masking, except at the shortest SOA (12 msec).

Discussion

The results of Experiment 1 are consistent with our predictions based on differences between several mask types in terms of their second- and higher order image statistics. Scene textures have steeper phase-only second spectra slopes than do the phase-randomized scenes, because the texture synthesis algorithm (Portilla & Simoncelli, 2000) uses oriented band-pass wavelet filters that capture spatially localized information (Field, 1999). This enables the algorithm to encode structures such as edges, lines, contours, and unique spatial configurations. These are decimated by phase randomization and thus cannot be encoded by second-order image statistics from the amplitude spectrum (see Figure 3). Experiment 1's results suggest that such differences affect scene gist masking and thus provide some of the first evidence for the necessity of measurable higher order statistical scene information for scene gist recognition. This conclusion is consistent with a body of research and theory on scene image statistics, which argues for the importance of higher order localized information in terms of neural coding strength (see, e.g., Field, 1987, 1994, 1999; Olshausen & Field, 1996; Simoncelli, 2003; Simoncelli & Olshausen, 2001; Thomson, 2001a, 2001b). Conversely, the present results are inconsistent with arguments that much or most of the important information for recognizing scenes is contained in the second-order statistics of the amplitude spectrum (Gorkani & Picard, 1994; Guerin-Dugue & Oliva, 2000; Guyader et al., 2005; Guyader et al., 2004; Herault et al., 1997; Kaping et al., 2007; Oliva & Torralba, 2001).

Nevertheless, there is a problem in attributing the difference in masking produced by scene textures versus phase-randomized scene masks solely to differences in their higher order statistical properties, because—although both mask types had very similar amplitude spectra—they were not identical, which could have affected their masking. We therefore carried out a control experiment that was identical in all respects to Experiment 1, with the single exception that there were only two mask type conditions: (1) the scene textures used in Experiment 1 ($n = 14$) versus (2) fully phase-randomized versions of the same scene textures ($n = 15$), with both mask types equalized for RMS contrast and mean luminance. Most importantly, the amplitude spectra of both masking conditions were, by definition, identical, whereas their higher order phase-only second spectra were markedly different (Figure 2D). The average phase-only second spectrum slope for the phase-randomized scene textures was $M = -.07$, $SD = .023$, and it was meaningfully and significantly different from the average phase-only second spectrum slope of the nonphase randomized scene textures [$t(598) = 161.27$, $p < .001$, $d = 16.0$]. Figure 6 shows the masking results. Importantly, scene textures caused significantly greater masking than did their phase-randomized versions [$F(1,27) = 6.39$, $p = .018$]. There was also the expected main effect of processing time (i.e., SOA) [$F(3.48, 94.16) = 29.38$, $p < .001$], but no significant interaction between mask type and processing time [$F(3.48, 94.16) < 1$, $p = .618$]. Critically, the stronger masking that was caused by scene textures rather than by phase-randomized scene textures cannot be attributed to differences in their amplitude spectra, since they were nonexistent, but can be attributed to the large differences in their phase-only second spectra.

We also ran another masking condition that was not mentioned previously, involving masking by the texture version of the target image. This produced less masking than did a texture of a different category (and almost identical masking to that of the phase-randomized conditions). This reduction in masking may have been caused by integrating information from the target with texture information from the same category. However, fully understanding this result requires a different experimental design in which

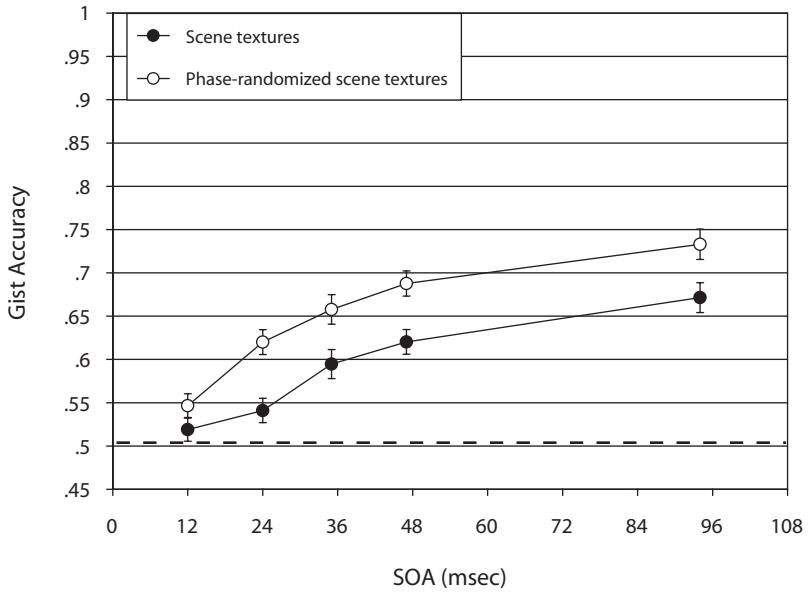

Figure 6. Control experiment for Experiment 1 data: Scene gist recognition accuracy as a function of the two different mask types (scene textures and phase-randomized scene textures) and processing time (stimulus onset asynchrony, SOA) (target = 12 msec; mask = 36 msec). Error bars represent standard errors of the means.

the target/mask categorical similarity factor (i.e., the similarity between the target and mask categories) is crossed with the phase (randomized vs. normal) factor. Data from such a study (Loschky et al., 2009) show a reduction in masking due to matching the category of the target with the category of the mask, but this occurs only for masks containing higher order statistics, not for those containing only second-order statistics. The latter result is consistent with the previously mentioned results of Loschky et al. (2007): that there was no difference between masking caused by a phase-randomized version of the target versus a phase-randomized version of an image from a different category than the target. That result suggested that second-order statistics do not cause category-specific masking, but instead cause a more generalized masking, perhaps due to their possessing 1/*f* amplitude spectra.

An interesting question is, can the higher order statistical structure captured by the phase-only second spectrum explain the difference in masking between scene texture masks versus normal scene masks? Specifically, scene textures possessed phase-only second spectra with steeper slopes and higher spectral magnitudes, but produced less masking than did normal scene images. Yet, masks containing higher order image statistics did produce stronger masking than did masks lacking such statistical structure. Thus, the effect of scene structure measured by the phase-only second spectrum on scene gist masking may likely follow a nonlinear function. For example, scene gist masking effects for higher order statistical properties may occur only above some threshold difference in the phase-only second spectrum, which in the present study was found when comparing the texture versus phase-randomized mask conditions, but not when comparing the texture versus normal scene mask conditions. If so, the stronger masking caused by the normal scene masks rather than by texture masks may not be due to their relatively small difference in the phase-only second spectrum, which would leave open the question of what caused their observed masking differences. One possible explanation is in terms of heterogeneous configuration information (e.g., layout), which, by the very nature of texture, is not captured from scenes by the texture algorithm (Portilla & Simoncelli, 2000), but is contained in normal scene images. Perhaps such global configuration information in the normal scene masks resulted in their stronger gist masking, in comparison with that in the scene texture masks.

A very different explanation for the difference in masking strength between the scene texture masks and normal scene masks is that the latter were more recognizable and thus produced more conceptual masking (Bachmann et al., 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976). By the same logic, the difference in masking strength between the scene texture masks and phase-randomized scene masks might also be explained in terms of differential recognizability. Thus far, we have pointed to the apparent unrecognizability of the scene textures in Figure 1 to rule out such an argument. However, completely ruling out this explanation requires that we empirically measure the recognizability of the scene texture masks versus the phase-randomized scene masks. In Experiment 2, we addressed this issue.

EXPERIMENT 2

We interpreted the results of Experiment 1 in terms of the differential importance of second-order versus higher order statistics for recognizing scene gist. However, an alternative explanation for the stronger masking caused by scene texture masks rather than by phase-randomized scene

masks is in terms of conceptual masking (Bachmann et al., 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976). Specifically, the scene texture masks might be more recognizable than the phase-randomized scene masks, and might thus produce stronger masking.

Recognizable masks generally produce greater masking than do those that are unrecognizable—the conceptual masking effect (Bachmann et al., 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976). Thus, conceptual masking theory would predict that, as a mask becomes less recognizable, it would cause less masking, and vice versa (Loschky et al., 2007). For example, Loschky et al. (2007) found that incremental phase randomization resulted in both incrementally worse scene gist recognition and incrementally weaker scene gist masking, which, taken together, are consistent with a conceptual masking explanation. Thus, to the degree that the scene texture masks of Experiment 1 were more recognizable than the phase-randomized scene masks, it would be consistent with a conceptual masking explanation of the Experiment 1 results. Conversely, if the texture and phase-randomized masks are equally unrecognizable, then we would instead explain the texture masking advantage in terms of their higher order statistical properties measured by the phase-only second spectrum.

Although the scene textures are only approximations of actual textures in scenes, it is possible that they might be recognizable as those scenes, on the basis of recent proposals that texture information may be important for scene classification (Fei-Fei & Perona, 2005; Renninger & Malik, 2004). Others have argued that at least some scene categories (e.g., forests) may be recognized using texture information (Oliva & Torralba, 2001). Thus, a secondary aim of Experiment 2 was to directly test such claims by measuring the recognizability of texture images (from Experiment 1) that were generated from scenes using the texture synthesis algorithm (Portilla & Simoncelli, 2000).

Method

Participants. A total of 75 Kansas State University students (42 female; mean age = 19.04 years) participated in the study for course credit. None of them had participated in any related experiments. All of the participants had an uncorrected or corrected near acuity of 20/30 or better.

Stimuli. The stimuli in Experiment 2 were identical to those used in Experiment 1 (minus the white noise images, which are inherently unrecognizable).

Design and Procedure. Image condition was the independent variable, with participants randomly assigned to see either normal scene images ($n = 21$) or texture images ($n = 24$). Later, a fully phase-randomized scenes condition was added ($n = 30$). Figure 4, right panel, shows a schematic of the events in a trial. Participants first saw a fixation cross, prompting them to press a button to start the trial. Then, 750 msec later, they saw an image that was flashed for 35 msec (the duration of masks in Experiment 1). After a 750-msec blank interval, a cue word (from the 10 target categories) was presented, either validly or invalidly describing the preceding image. Using a keyboard, participants pressed the "YES" key if they thought the cue validly described the preceding image, or the "NO" key if they thought that it did not. Half of the trials had valid cues (randomly selected), with cue validity equal across all target and cue categories. Participants were encouraged to respond as quickly and as accurately as possible, and to go with their first impression if unsure. There were 300 self-paced trials in the experiment, with each image presented once in a random order. Prior to the experiment, participants were familiarized with the 10 scene categories and the experimental task by seeing a separate set of 90 labeled scene images for 1 sec each, and then by doing 32 practice trials without feedback.

Results

As is shown in Figure 7, although gist recognition accuracy for unmasked normal scenes flashed for 35 msec was nearly perfect, accuracy for scene texture images that were generated from the same scenes was just barely—though reliably—above chance [$t(23) = 2.804$, $p = .01$]. More importantly, Figure 7 shows a tiny difference in the recognizability of scene texture images and of fully phase-randomized scenes [textures, $M = 0.52$, $SD = 0.03$; phase-randomized scenes, $M = 0.50$, $SD = 0.02$; $t(33.81) = 1.96$, $p = .058$, n.s.]. Thus, the hypothesis that the masking advantage of scene texture masks over phase-randomized scene masks was due to greater recognizability of the scene texture images—that is, conceptual masking—seems highly unlikely.

On the other hand, Figure 8 shows that gist recognition for texture images did vary across scene categories, from somewhat below chance for most man-made categories (because of a high miss rate) to clearly above chance for several natural categories [$F(7.36, 530.14) = 10.69$, $p < .001$]. Furthermore, this main effect of category interacted with the type of image [category × image type interaction: $F(14.73, 530.14) = 8.67$, $p < .001$]. As is shown in Figure 8, this interaction was primarily due to the fact that the most recognizable categories of normal scenes are not the most recognizable categories based on either scene textures or phase-randomized scene amplitude spectra. The clear difference in the recognizability of natural versus man-made categories of scene textures provides a clue to what the useful information from scene textures is. Note that the Portilla and Simoncelli (2000) texture synthesis algorithm picks up patterns in an input image and generates texture images in which those patterns are distributed homogeneously. Thus, those scene categories that are the most recognizable as texture images may be those that are the most recognizable based on homogeneous patterns—namely, forests, deserts, and mountains. Conversely, the categories that were the least recognizable based on homogeneous patterns were farms, streets, and markets. Thus, recognizing these latter categories may depend more on inhomogeneous configuration information (i.e., layout—Sanocki, 2003; Sanocki & Epstein, 1997; Schyns & Oliva, 1994). However, this hypothesis must be tested in further research because, to our knowledge, there is currently no standard metric of image homogeneity.

Discussion

Overall, the scene textures were unrecognizable. The 52% gist recognition rate for scene textures was just barely above chance, and only slightly different from the 50% gist recognition rate for fully phase-randomized scenes. These results fail to support the hypothesis that the masking advantage of texture images over fully phase-randomized

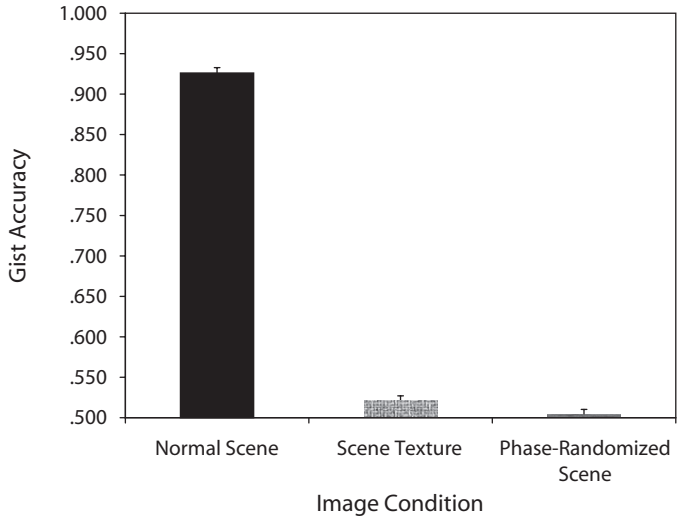

Figure 7. Unmasked scene gist recognition accuracy in Experiment 2 as a function of image type (normal scenes vs. scene textures vs. phase-randomized scenes) (duration = 35 msec). Error bars represent standard errors of the means.

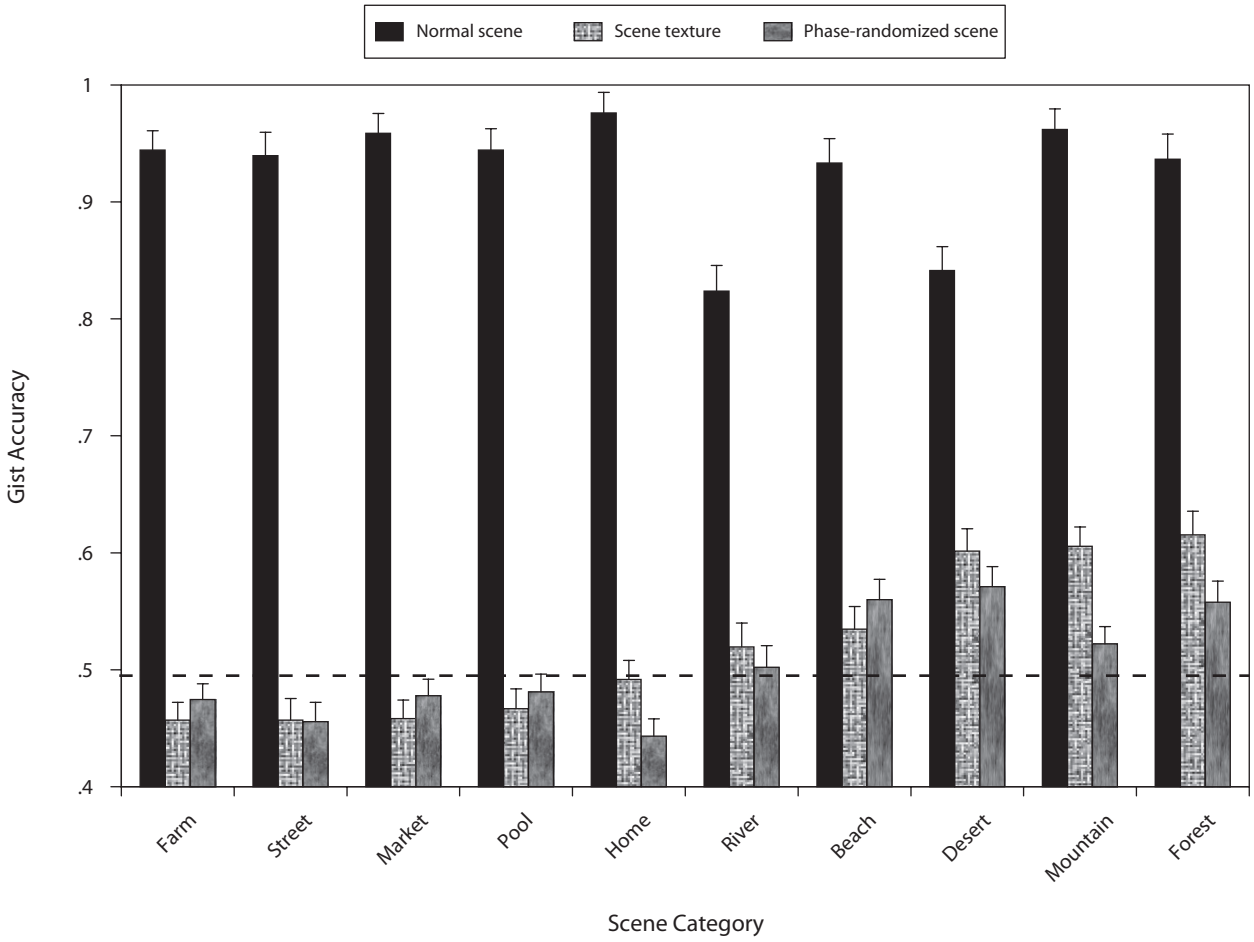

Figure 8. Unmasked scene gist recognition accuracy in Experiment 2 as a function of image type (normal scenes vs. scene textures vs. phase-randomized scenes) and scene category (duration = 35 msec). Error bars represent standard errors of the means.

images was due to conceptual masking. Thus, the results strengthen the opposing hypothesis that the masking advantage for scene texture masks was due to their having higher order statistical scene properties.

Nevertheless, although we failed in Experiment 2 to support a conceptual masking explanation for the strong scene gist masking produced by scene textures, its results nevertheless left the possibility open for such an explanation. Specifically, the variability in the gist recognizability of individual scene texture images, which ranged from 13% to 92%, was consistent with the hypothesis that some more recognizable scene texture images might have caused conceptual masking. If so, this might have been enough to produce the difference in masking that was found between scene texture masks and fully phase-randomized scene masks.

EXPERIMENT 3

In Experiment 3, we investigated whether the masking produced by a scene texture mask varies as a function of its recognizability. In Experiment 2, we showed that certain scene texture images were recognized at a fairly high level of accuracy. Those recognizable scene texture masks may have conceptually masked scene gist to a moderate degree (Bachmann et al., 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Loschky et al., 2007; Potter, 1976), thus explaining the mean masking strength advantage for scene texture masks over fully phase-randomized scene masks in Experiment 1. In Experiment 3, we rigorously tested this hypothesis by factorially combining the most and least recognizable scene texture masks in each scene category with an equal number of target images, and performing a regression analysis of masked scene gist accuracy on scene texture mask recognizability.

Method

Participants. A total of 68 Kansas State University students (37 female; mean age = 19.4 years) participated in the study for course credit. None had participated in any related experiments. All of the participants had an uncorrected or corrected near acuity of 20/30 or better.

Stimuli. The stimuli were a subset of those used in Experiments 1 and 2, and we will describe the selection criteria below.

Design and Procedures. The following experimental design elements and procedures were specific to Experiment 3.

High versus low recognizability masks. We chose the two least recognizable and two most recognizable scene texture images from each of the 10 scene categories in Experiment 2. Doing this maximized the chances of finding a conceptual masking effect that was caused by variation in scene texture mask recognizability. The average texture recognizability greatly differed between the two sets (most recognizable, $M = .72$, $SD = .09$; least recognizable, $M = .32$, $SD = .08$). The fact that the average accuracy for the 20 least recognizable scene textures was well below chance (.50) suggested a nonrandom bias. An analysis of these texture images showed a bias to reject them as members of their respective scene categories (76.6% misses on validly cued trials). However, when we further analyzed individual texture images, we found fairly uniform distributions of false alarms across invalid cue categories, suggesting that these unrecognizable texture images were not biased toward any particular interpretations.

Targets selected for average recognizability. In order to produce representative levels of target accuracy for each of 10 categories in Experiment 3, we selected four target images that were close to each scene category's mean in Experiment 1. We selected targets on the basis of their mean accuracy in the scene texture mask condition with a 47-msec SOA, which was the mask type and SOA planned for the experiment (explained further below).

Factorial combinations of targets and masks. Experiment 3 was designed to uniquely identify the masking effect of each individual mask. We factorially combined each mask with a set of 20 target images; thus, the recognizability of target images was held constant across masks, and doing this allowed us to determine the average target accuracy produced by each mask.

Between-subjects division of combinations into two sets. Because a factorial combination of 40 masks × 40 targets (1,600 trials) would fatigue our introductory psychology subject pool participants, we divided the masks and targets into two sets of 20 masks × 20 targets each (400 trials per set). Across the two sets, the natural and man-made target scene categories were approximately evenly divided (Set 1, beach, forest, farm, home, pool; Set 2, desert, mountain, river, market, street). Experiment 2's results suggested that both sets would produce equal average masked target accuracy (previous average accuracy rates: Set 1, $M = .644$; Set 2, $M = .646$). We then randomly assigned participants to the two sets (Set 1, $n = 36$; Set 2, $n = 32$).

The events in a trial were nearly identical to those in Experiment 1, with the following exceptions (see Figure 4, left panel). We chose to use a 1:1 mask:target duration ratio (both target and mask = 12 msec) to avoid a floor effect, and we chose a constant 47-msec SOA because it produced the strongest advantage for scene texture masks over fully phase-randomized scene masks in Experiment 1, which may have been the result of conceptual masking. All other procedures were identical to those in Experiment 1.

Results

The two stimulus sets and participant groups produced very similar mean accuracy (Group 1, $M = .72$, $SE = .013$; Group 2, $M = .69$, $SE = .014$); thus, they were combined for all further analyses. Contrary to the hypothesis that conceptual masking caused the texture masking advantage in Experiment 1, there was no effect of texture mask recognizability on scene gist accuracy. Mean gist accuracy when masked by the most recognizable scene texture masks ($M = .70$, $SE = .009$) was virtually identical to that when masked by the least recognizable scene texture masks ($M = .71$, $SD = .010$).

We carried out a stepwise logistic regression for accuracy as a function of mask recognizability, as well as two potentially strong variables of little theoretical interest: target ID (the specific target image) and mask ID (the specific mask image). Mask ID was entered first and was significant ($p < .001$), and target ID was entered second and was also significant ($p = .019$). Thus, the experiment was very sensitive to differences between masks (and to a lesser extent, it was also sensitive to differences between targets, although they had been selected to be of average gist recognizability). However, contrary to the hypothesis that the recognizability of individual texture masks caused their masking advantage in Experiment 1, mask recognizability did not reach significance ($p = .225$) and so was not entered into the equation. Figure 9 is a scatterplot of target gist accuracy as a function of scene texture mask recognizability, and shows no clearly discernable relationship

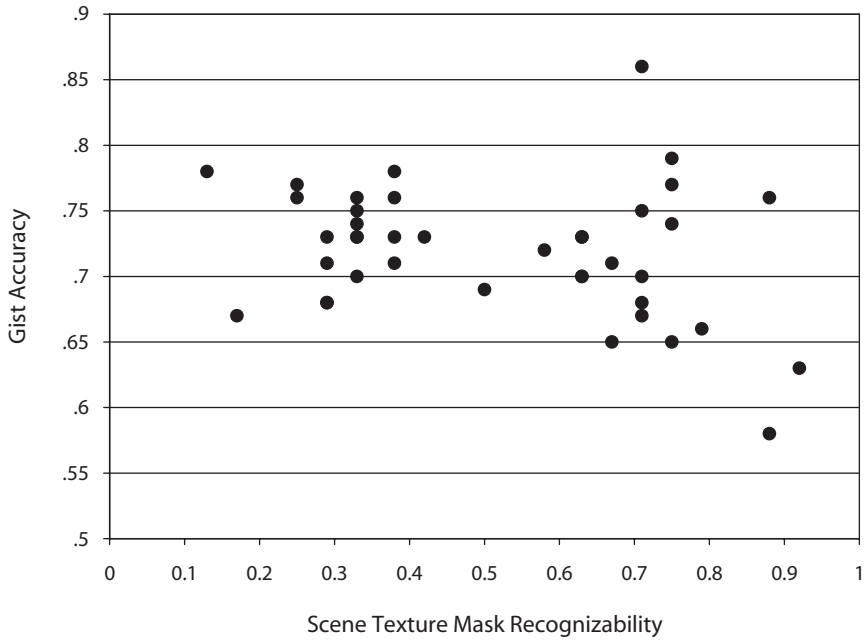

Figure 9. Scatterplot of target scene gist accuracy in Experiment 3 as a function of scene texture mask recognizability (as determined in Experiment 2).

between them. This was quantified by a linear regression of mask recognizability on accuracy, which, contrary to the conceptual masking hypothesis, produced a nonsignificant $R^2 = .013$ ($F < 1$), showing that only 1% of the gist masking variance was explained by texture mask recognizability. Similarly, there was a nonsignificant, small, negative Pearson correlation between mask recognizability and accuracy ($r = -.01$, $p = .07$, n.s.), whose magnitude clearly showed a lack of any meaningful relationship between texture mask recognizability and gist masking.

Discussion

The results of Experiment 3 provided no support for the hypothesis that scene texture masks produce conceptual masking. Thus, the masking advantage for scene texture masks over phase-randomized scene masks in both Experiment 1 and the control experiment was not due to conceptual masking. This failed hypothesis was in opposition to an alternative hypothesis, that masks containing higher order statistical scene structure more effectively disrupt scene gist recognition than do masks containing only second-order statistical scene information, which was supported by the results of both Experiment 1 and the control experiment.

CONCLUSIONS

Our present study addressed the issue of whether measurable higher order scene statistics are important for rapidly recognizing the category of a scene, or its "gist." We did this by controlling second-order scene statistics and by manipulating higher order scene information available in images and measuring their strength in masking scene gist. This follows the well-established practice of using masking to explore perceptual mechanisms in spatial vision (Carter & Henning, 1971; de Valois & Switkes, 1983; Henning et al., 1981; Legge & Foley, 1980; Losada & Mullen, 1995; Sekuler, 1965; Solomon, 2000; Stromeyer & Julesz, 1972; Wilson et al., 1983). A key issue was the relative roles of second- versus higher order image statistics in processing scene gist, which is related to a long-standing debate about the relationship between the statistical properties of the visual environment and the structure of mammalian visual systems. Specifically, the tuning of simple cell bandwidths to the $1/f$ amplitude spectrum (Field, 1987; Tolhurst, Tadmor, & Chao, 1992) has led to arguments emphasizing the importance of the second-order scene statistics in real-world perception (Párraga, Troscianko, & Tolhurst, 2000, 2005). This was mirrored by a number of computational models of scene classification emphasizing the importance of second-order statistics of scenes (Gorkani & Picard, 1994; Guerin-Dugue & Oliva, 2000; Guyader et al., 2004; Herault et al., 1997; Oliva & Torralba, 2001), and supporting human scene gist recognition studies (Guyader et al., 2005; Guyader et al., 2004; Kaping et al., 2007). Conversely, the higher order statistics of scenes have been argued to be more important for distinguishing different types of images, from both a computational and an evolutionary perspective (Field, 1994, 1999; Olshausen & Field, 1996; Simoncelli, 2003; Simoncelli & Olshausen, 2001; Thomson, 2001a, 2001b). Importantly, human scene gist recognition studies supporting this latter point of view have only recently begun to appear and have been limited to showing the negative impact on scene gist recognition of removing higher order statistical relationships (Joubert et al., 2009; Loschky & Larson, 2008; Loschky et al., 2007). The contribution of the present study, which showed a positive impact of measured higher order statistical in-

formation for human gist recognition, can be understood within the context of this larger debate.

In Experiment 1, we showed that masks containing only second-order amplitude spectrum scene statistics are less effective in masking scene gist recognition than are masks also possessing significant higher order scene statistics. We hypothesized that image structure encoded by higher order scene statistics was necessary for full scene gist masking. Consistent with this hypothesis, masking conditions with lower phase-only second spectrum magnitude (or flat orientation-averaged phase-only second spectrum slopes) produced weaker scene gist masking, whereas those with higher phase-only second spectrum magnitude (or steeper second-spectrum slopes) produced stronger gist masking. Nevertheless, we also found that the fact that a mask contains greater overall phase-only second spectrum magnitude is insufficient for full scene gist masking. This was shown by the fact that scene texture masks, which had a larger phase-only second spectrum magnitude (and a relatively steeper slope) than did normal scene masks, nevertheless produced a weaker disruption of scene gist than did normal scene masks. Nevertheless, the critical finding of the present study is that images possessing significant higher order image structure alone (i.e., unaccompanied by semantically meaningful content) are more effective at masking scene gist recognition than are images lacking such statistical relationships.

The present study also focused on explaining scene gist masking in terms of two opposing explanatory frameworks: conceptual masking versus spatial masking, with the latter being explained by second-order statistics carried by the amplitude spectrum versus higher order statistics, as measured by the phase-only second spectrum. Importantly, however, in most previous studies of conceptual masking, researchers have measured its effects on conceptual short-term memory (Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976), rather than on scene gist recognition (Loschky et al., 2007). In conceptual masking theory (Potter, 1976), it has been argued that perceptual masking processes occur up to about a 100-msec SOA, leading up to target identification, after which conceptual masking processes occur until roughly a 300-msec SOA, during conceptual memory consolidation. According to this theory, processing the meaning of the mask switches attention from the target to the mask, stopping target memory consolidation (Loftus et al., 1988). Interestingly, Loschky et al.'s (2007) results suggested that mask recognizability, as manipulated by phase-randomization, affected even early perceptual processes in scene gist recognition. However, consistent with the claim that early masking effects are perceptual rather than conceptual (Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976), the present study proposed that such early masking effects are largely explained in terms of spatial masking.

The relative contributions of conceptual versus spatial masking in the present study can be understood with the help of Tables 1 and 2. Table 1 shows the mean masked accuracy for each of the masking conditions in Experiment 1 and in the control experiment (across SOA). Conceptual masking effects for conceptual short-term memory are generally inferred from the finding of greater masking by meaningful scene masks rather than by meaningless noise masks (Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976). We can approximate the conceptual masking effect for scene gist recognition in the present study by comparing gist accuracy with white noise masks versus normal scene masks, as is shown in Table 2. It is tempting for one to explain this difference, which shows stronger masking by normal scenes than by white noise, in terms of conceptual masking. However, both fully phase-randomized scenes and scene textures cause greater gist masking than does white noise, and Experiments 2 and 3 have ruled out conceptual masking as an explanation for this masking. Indeed, Table 2 shows that the difference in masking produced by white noise versus phase-randomized scene masks is just over half (0.53) of the difference in masking produced by white noise versus normal scene masks. Thus, second-order scene statistics can account for half of what we might otherwise have attributed to conceptual masking. Similarly, scene texture masks produce nearly three fourths (0.74) of the effect that we might have attributed to conceptual masking. This additional 21% of the scene gist masking effect is instead explained by the higher order scene statistics measured by the phase-only second spectrum. Thus, three fourths of the scene gist masking effect that we were tempted to attribute to conceptual masking is instead explained by spatial masking due to second- and higher order scene statistics, thus leaving only one fourth (0.26) of the effect unexplained by spatial masking.

We can apply the same logic to answer the question that began our present study: How much of the decrement in masking caused by phase randomization can we attribute to the loss of higher order statistics? Table 3 shows the

Table 1
Mean Accuracy for Four Mask Types Used in Both Experiment 1 and the Control Experiment

Mask Type	Masked Accuracy
Normal scene	.54
Scene texture	.60
Phase-randomized scene or texture	.65
White noise	.77

Table 2
Proportion of the Conceptual Masking Effect (White Noise − Normal) Accounted for by Spatial Masking in Experiment 1

	Masked Accuracy Difference	Proportion of White Noise − Normal
White noise − normal	.22	1.0
White noise − phase-rand	.12	.53
White noise − texture	.17	.74

Note—Conceptual masking effect, the difference in masked accuracy between the white noise masking condition and the normal scene as mask condition; white noise, white noise used as a mask; normal, normal scene used as a mask; phase-rand, phase-randomized scene used as a mask; texture, scene texture used as a mask.

proportion of the phase-randomized scene (or texture) masking effect (i.e., phase-randomized masks—normal masks) accounted for by masks having higher order scene statistics. This shows that almost half (0.45) of the decrement in scene gist masking caused by phase randomization can be attributed to the loss of the higher order scene statistics measured by the phase-only second spectrum. The remainder of the masking effect may indeed be due to conceptual masking (Bachmann et al., 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus et al., 1988; Potter, 1976), although there are still questions about their time course; specifically, such effects are occurring when only perceptual masking, not conceptual masking, assumedly occurs. Alternatively, some part of that remaining masking effect may be accounted for by scene information not captured by the texture synthesis algorithm (Portilla & Simoncelli, 2000) and thus not in scene texture masks—scene layout (Sanocki, 2003; Sanocki & Epstein, 1997; Schyns & Oliva, 1994). Consistent with this hypothesis, a recent study showed that much of what might have been attributed to conceptual masking was explained by the mask's spatial layout (Michod & Intraub, 2008). Further research should disentangle these alternative explanations of scene gist masking. Nevertheless, almost 75% of the scene gist masking effect that one might have been tempted to attribute to conceptual masking was explainable in terms of spatial masking, and 45% of the decrement in masking that was due to phase randomization was attributable to the loss of higher order statistical structure. Both are important findings for understanding conceptual masking as it affects scene gist.

The results of the present study also challenge popular models that successfully classify scenes using second-order image statistics. An important element of the spatial envelope model (Oliva & Torralba, 2001, 2006; Oliva, Torralba, Castelhano, & Henderson, 2003; Torralba, 2003; Torralba et al., 2006) is the idea that much of the useful information in scenes is contained in their second-order statistics from the amplitude spectrum, on the basis of their scene classification modeling results (Oliva & Torralba, 2001). Similarly, authors of behavioral studies (Guyader et al., 2004; Kaping et al., 2007) have argued that amplitude spectrum information is critical or even sufficient for recognizing scene gist. However, the results of Experiment 1 suggest that scenes' higher order statistics contain critical information for scene gist recognition. In fact, the spatial envelope model has evolved over time to include an increasingly important role for higher order statistics through windowing or wavelet filters (Oliva & Torralba, 2001, 2006; Torralba & Oliva, 2002). Our results show the importance of such higher order scene statistics encoded by wavelet filters—as measured by the phase-only second spectrum—for scene gist masking, and thus for human scene gist recognition performance and scene classification models, such as the spatial envelope model. There is also an interesting agreement between Oliva and Torralba's (2001) scene classification modeling results and our masking results regarding the relative roles of second-order versus higher order statistics in scene gist recognition. Specifically, Oliva and Torralba's (2001) model achieved 86% accurate scene categorization relative to 12.5% chance, using only second-order amplitude spectrum statistics. Adding higher order statistics through spatially localized filters only increased performance to 92%. Analogously, we found that just over 50% of the masking that we wanted to explain (i.e., the masking difference between white noise, which has no information useful for scene gist, and normal scenes, which contain all of the information useful for scene gist) was produced by fully phase-randomized scenes, which contained only second-order amplitude spectrum statistics. Another 21% of the masking that we wished to explain was produced by the masks' having higher order statistics, measured by the phase-only second spectrum.

Table 3
Proportion of the Phase-Randomized Scene Masking Effect (Phase-Randomized − Normal) Accounted for by Masks Having Higher Order Statistical Scene Structure in Experiment 1

	Masked Accuracy Difference	Proportion of Phase-Rand − Normal
Phase-rand − normal	.11	1.0
Phase-rand − texture	.05	.45

Note—Phase-randomized scene masking effect, the difference in masked accuracy between the phase-randomized scene as mask condition and the normal scene as mask condition; phase-rand, phase-randomized scene used as a mask; normal, normal scene used as a mask; texture, scene texture used as a mask.

Our present study also speaks to the value of texture information for recognizing scene gist. Generally, textures that were synthesized from real-world scenes were unrecognizable. Yet, certain scene categories containing repeated homogeneous patterns were more recognizable, such as the "natural" categories of desert, forest, and mountain. Conversely, the scene categories that were the least recognizable from textures appeared to be those for which a heterogeneous global spatial configuration (i.e., layout) was most important. Since gist recognition based on scene-texture information was very poor (52%, with 50% being chance), layout may be very important for recognizing scene gist (Sanocki, 2003; Sanocki & Epstein, 1997; Schyns & Oliva, 1994). Future studies should investigate the complementary roles of local texture patterns versus global spatial configurations (i.e., layout) in recognizing scene gist.

Finally, our present study is useful for scene perception researchers who use visual masking. Currently, most scene perception studies using visual masks lack a strong justification for the type of masks they use. Our study shows the relative strengths of several types of masks (varying in their second- and higher order image statistics) over time, thus providing a theoretically driven empirical basis for understanding scene masking effects. Finally, by ruling out conceptual masking by two types of highly effective masks (phase-randomized scenes and scene textures), our study clarifies the locus of their effects and outlines a practical method for further investigating the spatial information used to recognize scene gist.

AUTHOR NOTE

The present study contained some information that was presented at the Annual Meeting of the Vision Sciences Society (2006), with the abstract of the presentation having been published in the *Journal of Vision*. This work was supported by funds from the Kansas State University Office of Research and Sponsored Programs, and by the NASA Kansas Space Grant Consortium. The authors acknowledge the work of Bernardo de la Garza, Katie Gibb, Jeff Burns, John Caton, Stephen Dukich, Ryan Eshelman, Nicholas Forristal, Kaci Haskett, Hannah Hess, Zach Maier, Rebecca Millar, Kwang Park, and Merideth Smythe, who helped carry out the present experiments. The authors also thank Eero Simoncelli for sharing the Texture Synthesis software, Mitchell Thomson for discussions of the phase-only second spectrum, and David Field, Eero Simoncelli, Daniel J. Simons, and Tyler Freeman for their helpful discussions of the article. Correspondence concerning this article should be addressed to L. C. Loschky, Department of Psychology, Kansas State University, 471 Bluemont Hall, Manhattan, KS 66506-5302 (e-mail: loschky@ksu.edu).

REFERENCES

BACHMANN, T., LUIGA, I., & PÕDER, E. (2005). Variations in backward masking with different masking stimuli: II. The effects of spatially quantised masks in the light of local contour interaction, interchannel inhibition, perceptual retouch, and substitution theories. *Perception*, **34**, 139-154.

BACON-MACE, N., MACE, M. J., FABRE-THORPE, M., & THORPE, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, **45**, 1459-1469.

BOYCE, S., & POLLATSEK, A. (1992). An exploration of the effects of scene context on object identification. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp. 227-242). New York: Springer.

BREWER, W. F., & TREYENS, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, **13**, 1207-1230.

BURTON, G. J., & MOOREHEAD, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, **26**, 157-170.

CARTER, B. E., & HENNING, G. B. (1971). The detection of gratings in narrow-band visual noise. *Journal of Physiology*, **219**, 355-365.

DAVENPORT, J. L., & POTTER, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, **15**, 559-564.

DE GRAEF, P., DE TROY, A., & D'YDEWALLE, G. (1992). Local and global contextual constraints on the identification of objects in scenes. *Canadian Journal of Psychology*, **46**, 489-508.

DE VALOIS, K. K., & SWITKES, E. (1983). Simultaneous masking interactions between chromatic and luminance gratings. *Journal of the Optical Society of America*, **73**, 11-18.

DONG, D. W., & ATICK, J. J. (1995). Statistics of natural time-varying images. *Network*, **6**, 345-358.

ECKSTEIN, M. P., DRESCHER, B. A., & SHIMOZAKI, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, **17**, 973-980.

FEI-FEI, L., IYER, A., KOCH, C., & PERONA, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, **7**(1, Art. 10), 1-29.

FEI-FEI, L., & PERONA, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In C. Schmid, S. Soalto, & C. Tomasi (Eds.), *Computer vision and pattern recognition, 2005* (Vol. 2, pp. 524-531). Los Alamitos, CA: IEEE Computer Society Press.

FIELD, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, **4**, 2379-2394.

FIELD, D. J. (1993). Scale-invariance and self-similar "wavelet" transforms: An analysis of natural scenes and mammalian visual systems. In M. Farge, J. C. R. Hunt, & J. C. Vassilicos (Eds.), *Wavelets, fractals and Fourier transforms: New developments and new applications* (pp. 151-193). Oxford: Oxford University Press, Clarendon Press.

FIELD, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, **6**, 559-601.

FIELD, D. J. (1999). Wavelets, vision and the statistics of natural scenes. *Philosophical Transactions of the Royal Society A*, **357**, 2527-2542.

GORDON, R. D. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 760-777.

GORKANI, M. M., & PICARD, R. W. (1994). Texture orientation for sorting photos "at a glance." In S. Peleg & S. Ullman (Eds.), *Proceedings of the 12th IAPR International Conference on Pattern Recognition* (pp. 459-464). Los Alamitos, CA: IEEE Computer Society Press.

GUERIN-DUGUE, A., & OLIVA, A. (2000). Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, **21**, 1135-1140.

GUYADER, N., CHAUVIN, A., BERT, L., MERMILLOD, M., HÉRAULT, J., & MARENDAZ, C. (2005). Rapid visual scene categorization relies mainly on amplitude spectrum. *Investigative Ophthalmology & Vision Science*, **46**, E-Abstract 5642.

GUYADER, N., CHAUVIN, A., PEYRIN, C., HÉRAULT, J., & MARENDAZ, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process. *Comptes Rendus Biologies*, **327**, 313-318.

HANSEN, B. C., & ESSOCK, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision*, **4**, 1044-1060.

HANSEN, B. C., & ESSOCK, E. A. (2005). Influence of scale and orientation on the visual perception of natural scenes. *Visual Cognition*, **12**, 1199-1234.

HANSEN, B. C., ESSOCK, E. A., ZHENG, Y., & DEFORD, J. K. (2003). Perceptual anisotropies in visual processing and their relation to natural image statistics. *Network: Computation in Neural Systems*, **14**, 501-526.

HANSEN, B. C., & HESS, R. F. (2007). Structural sparseness and spatial phase alignment in natural scenes. *Journal of the Optical Society of America A*, **24**, 1873-1885.

HENNING, G. B., HERTZ, B. G., & HINTON, J. L. (1981). Effects of different hypothetical detection mechanisms on the shape of spatial-frequency filters inferred from masking experiments: I. Noise masks. *Journal of the Optical Society of America*, **71**, 574-581.

HERAULT, J., OLIVA, A., & GUERIN-DUGUE, A. (1997). Scene categorisation by curvilinear component analysis of low frequency spectra. In M. Verleysen (Ed.), *Proceedings of the 5th European Symposium on Artificial Neural Networks* (pp. 91-96). Bruges, Belgium: D Facto.

HOLLINGWORTH, A., & HENDERSON, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, **127**, 398-415.

INTRAUB, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 115-125.

JOUBERT, O. R., ROUSSELET, G. A., FABRE-THORPE, M., & FIZE, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision*, **9**(1, Art. 2), 1-16. doi:10.1167/9.1.2

KAPING, D., TZVETANOV, T., & TREUE, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition*, **15**, 12-19.

KEIL, M. S., & CRISTOBAL, G. (2000). Separating the chaff from the wheat: Possible origins of the oblique effect. *Journal of the Optical Society of America B*, **17**, 697-710.

KOVESI, P. (1999). Image features from phase congruency. *Videre*, **1**, 1-26.

LEGGE, G. E., & FOLEY, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America*, **70**, 1458-1471.

LOFTUS, G. R., & GINN, M. (1984). Perceptual and conceptual masking of pictures. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 435-441.

LOFTUS, G. R., HANNA, A. M., & LESTER, L. (1988). Conceptual masking: How one picture captures attention from another picture. *Cognitive Psychology*, **20**, 237-282.

LOFTUS, G. R., & MACKWORTH, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 565-572.

LOSADA, M. A., & MULLEN, K. T. (1995). Color and luminance spatial tuning estimated by noise masking in the absence of off-frequency looking. *Journal of the Optical Society of America A*, **12**, 250-260.

LOSCHKY, L. C., HANSEN, B. C., FINTZI, A., BJERG, A., ELLIS, K., FREEMAN, T., ET AL. (2009, May). *Basic level scene categorization is affected by unrecognizable category-specific image features*. Poster presented at the 8th Annual Meeting of the Vision Sciences Society, Naples, FL.

LOSCHKY, L. C., & LARSON, A. M. (2008). Localized information is necessary for scene categorization, including the natural/man-made distinction. *Journal of Vision*, **8**(1, Art. 4), 1-9.

LOSCHKY, L. C., SETHI, A., SIMONS, D. J., PYDIMARI, T., OCHS, D., & CORBEILLE, J. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **33**, 1431-1450.

MARR, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

MCCOTTER, M., GOSSELIN, F., SOWDEN, P., & SCHYNS, P. (2005). The use of visual information in natural scenes. *Visual Cognition*, **12**, 938-953.

MICHOD, K. O., & INTRAUB, H. (2008). Conceptual masking: Is concept the key, or does layout play a role? *Visual Cognition*, **16**, 120-123.

MORRONE, M. C., & BURR, D. C. (1988). Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society B*, **235**, 221-245.

MORRONE, M. C., & OWENS, R. A. (1987). Feature detection from local energy. *Pattern Recognition Letters*, **6**, 303-313.

OLIVA, A. (2005). Gist of a scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251-256). Burlington, MA: Elsevier.

OLIVA, A., & SCHYNS, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, **41**, 176-210.

OLIVA, A., & TORRALBA, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, **42**, 145-175.

OLIVA, A., & TORRALBA, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, **155**, 23-36.

OLIVA, A., TORRALBA, A., CASTELHANO, M. S., & HENDERSON, J. M. (2003). Top down control of visual attention in object detection. *IEEE Proceedings of the International Conference on Image Processing*, **1**, 253-256.

OLSHAUSEN, B. A., & FIELD, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607-609.

PALMER, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, **3**, 519-526.

PÁRRAGA, C. A., TROSCIANKO, T., & TOLHURST, D. J. (2000). The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology*, **10**, 35-38.

PÁRRAGA, C. A., TROSCIANKO, T., & TOLHURST, D. J. (2005). The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model. *Vision Research*, **45**, 3145-3168.

PEZDEK, K., WHETSTONE, T., REYNOLDS, K., ASKARI, N., & DOUGHERTY, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 587-595.

PORTILLA, J., & SIMONCELLI, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, **40**, 49-71.

POTTER, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 509-522.

RENNINGER, L. W., & MALIK, J. (2004). When is scene identification just texture recognition? *Vision Research*, **44**, 2301-2311.

ROUSSELET, G. A., FABRE-THORPE, M., & THORPE, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, **5**, 629-630.

ROUSSELET, G. A., JOUBERT, O. R., & FABRE-THORPE, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, **12**, 852-877.

RUDERMAN, D. L., & BIALEK, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, **73**, 814-818.

SADR, J., & SINHA, P. (2001). *Exploring object perception with random image structure evolution* (No. Memo #2001-06). Cambridge, MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

SADR, J., & SINHA, P. (2004). Object recognition and random image structure evolution. *Cognitive Science*, **28**, 259-287.

SANOCKI, T. (2003). Representation and perception of spatial layout. *Cognitive Psychology*, **47**, 43-86.

SANOCKI, T., & EPSTEIN, W. (1997). Priming spatial layout of scenes. *Psychological Science*, **8**, 374-378.

SCHYNS, P., & OLIVA, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, **5**, 195-200.

SEIDLER, G. T., & SOLIN, S. A. (1996). Non-Gaussian 1/f noise: Experimental optimization and separation of high-order amplitude and phase correlations. *Physical Review B*, **53**, 9753-9759.

SEKULER, R. W. (1965). Spatial and temporal determinants of visual backward masking. *Journal of Experimental Psychology*, **70**, 401-406.

SHAPLEY, R., & LENNIE, P. (1985). Spatial frequency analysis in the visual system. *Annual Review of Neuroscience*, **8**, 547-583.

SIMONCELLI, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, **13**, 144-149.

SIMONCELLI, E. P., & OLSHAUSEN, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, **24**, 1193-1216.

SMITH, J. (2007). *Mathematics of the discrete Fourier transform (dft) with audio applications* (2nd ed.). Available at http://books.w3k.org/.

SOLOMON, J. A. (2000). Channel selection with non-white-noise masks. *Journal of the Optical Society of America A*, **17**, 986-993.

STROMEYER, C. F., III, & JULESZ, B. (1972). Spatial-frequency masking in vision: Critical bands and spread of masking. *Journal of the Optical Society of America*, **62**, 1221-1232.

SWITKES, E., MAYER, M. J., & SLOAN, J. A. (1978). Spatial frequency analysis of the visual environment: Anisotropy and the carpentered environment hypothesis. *Vision Research*, **18**, 1393-1399.

TADMOR, Y., & TOLHURST, D. J. (1993). Both the phase and the amplitude spectrum may determine the appearance of natural images. *Vision Research*, **33**, 141-145.

THOMSON, M. G. A. (1999). Higher-order structure in natural scenes. *Journal of the Optical Society of America B*, **16**, 1549-1553.

THOMSON, M. G. A. (2001a). Beats, kurtosis and visual coding. *Network: Computation in Neural Systems*, **12**, 271-287.

THOMSON, M. G. A. (2001b). Sensory coding and the second spectra of natural signals. *Physical Review Letters*, **86**, 2901-2904.

TOLHURST, D. J., TADMOR, Y., & CHAO, T. (1992). Amplitude spectra of natural images. *Ophthalmic & Physiological Optics*, **12**, 229-232.

TORRALBA, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A*, **20**, 1407-1418.

TORRALBA, A., & OLIVA, A. (2002). Depth estimation from image structure. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **24**, 1226-1238.

TORRALBA, A., & OLIVA, A. (2003). Statistics of natural image categories. *Network*, **14**, 391-412.

TORRALBA, A., OLIVA, A., CASTELHANO, M. S., & HENDERSON, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, **113**, 766-786.

VAN DER SCHAAF, A., & VAN HATEREN, J. H. (1996). Modelling the power spectra of natural images: Statistics and information. *Vision Research*, **36**, 2759-2770.

WANG, Z., & SIMONCELLI, E. P. (2004). Local phase coherence and the perception of blur. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems* (pp. 786-792). Cambridge, MA: MIT Press.

WILSON, H. R., MCFARLANE, D. K., & PHILLIPS, G. C. (1983). Spatial frequency tuning of orientation selective units estimated by oblique masking. *Vision Research*, **23**, 873-882.

NOTES

1. Here, we briefly discuss the relationship between second-order and higher order statistics of scene images. The distribution of amplitude as a function of spatial frequency and orientation in the frequency domain is based on the degree of correlation between the luminances of all possible pixel pairs in the spatial (i.e., image) domain, and these are referred to as the second-order image statistics of a scene. That is, the Fourier amplitude spectrum (or power spectrum, or autocorrelation function) directly assesses the second-order statistical relationships of the pixel luminances across a scene image. Although much information about scene content can be obtained by assessing the second-order statistics of scenes (Burton & Moorehead, 1987; Dong & Atick, 1995; Field, 1987; Hansen & Essock, 2004, 2005; Keil & Cristobal, 2000; Oliva & Torralba, 2001; Ruderman & Bialek, 1994; Switkes, Mayer, & Sloan, 1978; Tadmor & Tolhurst, 1993; Tolhurst, Tadmor, & Chao, 1992; Torralba & Oliva, 2003; van der

Schaaf & van Hateren, 1996), other work has shown that second-order image statistics by themselves cannot explain other important attributes of real-world images (Thomson, 1999, 2001a, 2001b). This is because the Fourier phase spectrum consists of image statistics higher than the second-order (i.e., higher order) statistics, involving relationships among more than two pixel luminances (Thomson, 2001a).

2. We thank David Field for first suggesting this idea.

3. The size of the original images from the Corel image database was 1,024 × 768 pixels. These were trimmed to remove black margins (from the photographic negatives) and were resized to 1,024 × 674 pixels without changing the cropped images' aspect ratios. From these images, texture images were then generated at a size of 1,024 × 640 pixels, due to constraints of the algorithm. However, the experimental software (Experiment Builder) displayed all images so that they filled the 1,024 × 768 pixel screen resolution.

4. In order to conduct a proper Fourier analysis that would produce an equal number of spatial frequencies for each orientation, all image analyses were conducted on the central 674 × 674 pixel region of the stimuli. Thus, any subtle variation between the amplitude measurements of intact scenes and phase-scrambled scenes can be explained by the fact that the analyses were run on the central regions of the stimuli (which would otherwise be identical).

APPENDIX
Calculating the Phase-Only Second Spectrum of an Image

The phase-only second spectrum offers a global assessment of the strength of sinusoidal fluctuations (i.e., signal variance) as a function of different spatial frequency offsets. For example, a large value in the phase-only second spectrum shows the presence of a significant interaction among a number of sinusoidal modulations that either sum up to or differ by that particular offset. Thus, the phase-only second spectrum assesses the degree of phase alignment across all spatial frequencies in an image. Natural scene images exhibit a linear falloff (on logarithmic axes) in the phase-only second spectrum magnitude as a function of frequency offset (Thomson, 2001a, 2001b), which indicates a strong degree of alignment across spatial frequencies (i.e., higher second spectrum magnitude at smaller frequency offsets relative to larger offsets). The edges, lines, and contours that make up natural images arise from frequency alignment across a large range of spatial frequencies (Field, 1993; Morrone & Burr, 1988; Morrone & Owens, 1987), to which the phase-only second spectrum is quite sensitive.

The following phase-only second spectrum calculation is based on the steps described in Thomson (2001b) and Seidler and Solin (1996). In the following, the pixel intensity at location coordinates (x,y) in an image I is denoted by the corresponding lowercase letter $i_{x,y}$.

For a given image I, the following steps will give the phase-only second spectrum $S^{(2,\phi)}$.

1. Filter out the highest spatial frequency octave of the image in order to avoid artifacts introduced by the edges of the image, and so on. Let $I' = f(I)$, where f is the band-pass filter that removes the highest spatial frequency octave and the zero frequency component.

2. Calculate the discrete Fourier transform of I' using a standard FFT algorithm. Let $J = DFT(I')$, where DFT denotes a discrete Fourier transform.

3. Whiten J in order to remove the contribution of the amplitude spectrum. Let $\hat{J}$ be the whitened Fourier transform of the band-pass image I', so that

$$\hat{j}_{u,v} = \frac{j_{u,v}}{abs(j_{u,v})},$$

where *abs* represents the magnitude operation.

4. Synthesize an image from $\hat{J}$ using the inverse discrete Fourier transform, which can be calculated using a standard FFT algorithm. Let $V = IDFT(\hat{J})$, where $IDFT$ is the inverse discrete Fourier transform.

5. Normalize the power spectrum of V^2. First, calculate the point-by-point squared signal of V, which we denote by V^2. Compute M, the power spectrum of V^2. Normalize M to obtain N, so that

$$n_{u,v} = \frac{m_{u,v}}{4P^2(f_H - f_L)^2},$$

where P is the number of pixels and f_H and f_L are the high and low frequencies of the band-pass filter used in Step 1.

6. Subtract the Gaussian background to obtain the phase-only second spectrum $S^{(2,\phi)}$, so that

$$s_{u,v}^{(2,\phi)} = n_{u,v} - \frac{2(f_H - f_L - f_{u,v})^2}{(f_H - f_L)^2},$$

where $f_{u,v}$ is the spatial frequency at the point (u,v).

As a final note, the phase-only second spectrum calculation described previously is highly sensitive to pixel clipping (i.e., if one generates an image containing values that fall outside of the 0–255 range and subsequently saves it as an image file [i.e., .bmp, .tif, etc.], the pixel values falling outside the 0–255 range will be clipped to either 0 or 255). Pixel clipping has the effect of adding unintentional edges around the clipped region(s), which will show up in the phase-only second spectrum as phase alignment. Thus, great care is needed to ensure that stimuli do not exceed the 0–255 pixel luminance range before conducting a phase-only second spectrum analysis. Also, since the phase-only second spectrum is a global measure of phase alignment, stimuli should have square dimensions and ideally possess dimensions that are a power of two.

(Manuscript received August 17, 2008;
revision accepted for publication September 13, 2009.)