Running head: IMPORTANCE OF LOCALIZATION IN SCENE GIST

The Importance of Information Localization in Scene Gist Recognition

Lester C. Loschky, Kansas State University

Amit Sethi, University of Illinois at Urbana-Champaign

Daniel J. Simons, University of Illinois at Urbana-Champaign

Tejaswi N. Pydimarri, Kansas State University

Daniel Ochs, Kansas State University

Jeremy L. Corbeille, Kansas State University

Address correspondence to:
Lester Loschky
Assistant Professor
Department of Psychology
471 Bluemont Hall
Kansas State University
Manhattan, KS 66056-5302
Phone: 785-532-6882
E-mail: loschky@ksu.edu

This article may not exactly replicate the final version published in the Journal of Experimental Psychology: Human Perception and Performance. It is not the copy of record.
 DOI: 10.1037/0096-1523.33.6.1431

Abstract

People can recognize the meaning or gist of a scene from a single glance, and a few recent studies have begun to examine the sorts of information that contribute to scene gist recognition. We used visual masking coupled with image manipulations (randomizing phase while maintaining the Fourier amplitude spectrum (RISE: Sadr & Sinha, 2004)) to explore whether and when unlocalized Fourier amplitude information contributes to gist perception. In four experiments, we found that differences between scene categories in the Fourier amplitude spectrum are insufficient for gist recognition or gist masking. While the global 1/f spatial frequency amplitude spectra of scenes plays a role in gist masking, local phase information is necessary for gist recognition, and for the strongest gist masking. Moreover, the ability to recognize the gist of a target image was influenced by mask recognizability, suggesting that conceptual masking occurs even at the earliest stages of scene processing.

The Importance of Information Localization in Scene Gist Recognition

RECOGNIZING THE GIST OF A SCENE

Within a single glance, people can recognize the meaning or "gist" of a scene (Biederman, Rabinowitz, Glass, & Stacy, 1974; Potter, 1976; Rousselet, Joubert, & Fabre-Thorpe, 2005). The term, "gist" is not always clearly defined (though see Oliva, 2005) but is most frequently operationalized as the scene's basic level category, for example "beach" or "street" (Tversky & Hemenway, 1983), and we follow that convention here. Gist information appears to guide viewers' inspection of the scene (Loftus & Mackworth, 1978; Oliva, Torralba, Castelhano, & Henderson, 2003), may aid object recognition in the scene (Boyce & Pollatsek, 1992; Davenport & Potter, 2004; De Graef, De Troy, & D'Ydewalle, 1992; Hollingworth & Henderson, 1998; Palmer, 1975), and affects later memory of the scene (Brewer & Treyens, 1981; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989). Given the speed of gist perception, the information underlying gist recognition may be based on holistic, low-level scene properties (Oliva & Torralba, 2001; Renninger & Malik, 2004; Vailaya, Jain, & Zhang, 1998), rather than based on detecting or recognizing individual objects (c.f., Davenport & Potter, 2004).

If low-level scene information underlies gist recognition, what types of information are used? A recent provocative proposal is that viewers recognize gist based on the unlocalized Fourier amplitude spectrum information in scenes, which consists of the spatial frequencies in the image, from low frequency blobs to high frequency details, at or near the cardinal orientations—horizontal, vertical, and oblique (Gorkani & Picard, 1994; Guyader, Chauvin, Peyrin, Hérault, & Marendaz, 2004; Oliva, Torralba, Guerin-Dugue, & Herault, 1999). For example, most beach scenes have a horizon, conveyed by low frequency horizontal information, while most city scenes do not, but instead have tall buildings, conveyed by more vertical

information across a wider range of spatial frequencies. Importantly, while the Fourier amplitude spectrum can tell us that a beach scene is dominated by low spatial frequency horizontal information, it cannot tell us *where* in the image that information is located (the middle, vs. the top right corner, etc.), which is encoded in the Fourier *phase* spectrum. Thus, if unlocalized amplitude spectrum information is sufficient to recognize gist, then scene layout information (Sanocki, 2003) is not necessary—a counter-intuitive and important claim to test. Recent computational modeling studies have supported this claim (Gorkani & Picard, 1994; Oliva & Torralba, 2001; Oliva, Torralba, Guerin-Dugue, & Herault, 1999). For example, Oliva and Torralba (2001) compared the scene categorization performance of two versions of their Spatial Envelope model, one in which there was coarse spatial localization (the "windowed discrete spectral template" or WDST), and one in which there was none (the "discrete spectral template" or DST). They found that:

> on average among natural and urban landscapes, 92% of the scenes were accurately classified with the WDST [using coarsely localized information] and 86% when using the DST [i.e., unlocalized information]. These results highlight the important role played by the unlocalized spectral components (DST) for representing the spatial envelope properties. The addition of spatial layout information clearly increases performance, but most of this performance level may be attributable to the global distribution of the relevant spectral features. (Oliva and Torralba, 2001, pp. 166-167)

Furthermore, a more recent study by Guyader, et al. (2004), entitled "Image phase or amplitude? Rapid scene categorization is an amplitude-based process," has extended this argument from the domain of computational modeling to that of human gist perception. That

study found that the unlocalized amplitude information contained in phase-randomized scenes provided equivalent scene gist priming to that of normal scene images.

An opposing view argues for the importance of localized information, as evidenced by the response properties of cells in visual cortex. These not only respond best to specific spatial frequencies, at specific orientations, but also to particular locations, and may thus be characterized to a first approximation by wavelets (Field, 1994, 1999; Simoncelli & Olshausen, 2001; Thomson & Foster, 1997). In support of this theory, well-known demonstrations have shown that phase information seems more important than amplitude information for recognizing objects (Oppenheim & Lim, 1981; Piotrowski & Campbell, 1982) and recent carefully controlled studies have shown that phase-randomized objects with normal amplitude spectra are unrecognizable (Sadr & Sinha, 2004; Wichmann, Braun, & Gegenfurtner, 2006). Nevertheless, scenes can be recognized without recognizing their constituent objects (Schyns & Oliva, 1993). Thus, an important question is whether, unlike object recognition, scene gist recognition is possible without configural information, strictly on the basis of unlocalized amplitude spectrum information, or whether phase information is necessary for scene gist recognition as well.

If unlocalized amplitude information is useful for recognizing scene gist, we would like to know when in scene processing is it used. Thus far, only a few facts about the time course of information use in scene gist recognition are known. Near perfect scene gist recognition is possible with masked image presentations as short as 100 ms (Biederman, Rabinowitz, Glass, & Stacy, 1974; Potter, 1976), and scenes can be categorized significantly above chance with only 20 ms masked durations (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Loschky & Simons, 2004). Low spatial frequency information and color seem to dominate early stages of scene gist acquisition (Loschky & Simons, 2004; Oliva & Schyns, 2000; Schyns & Oliva, 1993),

with higher spatial frequency information becoming increasingly important with durations of 50 ms or longer. Given the simplicity of unlocalized amplitude information, one hypothesis is that, if it is useful at all, it should be primarily at very early processing stages, for example, at masked durations at or below 50 ms.

USING VISUAL MASKING TO STUDY SCENE GIST RECOGNITION

Assessing the time course of scene gist recognition generally involves backward masking the scene with another stimulus. Without a mask, sensory persistence cancels out any effects of varying stimulus duration on the information extracted from a stimulus (Loftus & Mclean, 1999; Sperling, 1963). Conversely, by varying the timing of a target and its mask, one can investigate the microgenesis of perception (Breitmeyer & Ogmen, 2006). Masking can also be used to understand the information contributing to a visual task (Delord, 1998). In tasks requiring perception of orientation or spatial frequency content, masking is most efficient when the target and mask are most similar on those dimensions (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). In a task for which low spatial frequency information is important, a low frequency mask is more effective than a high frequency mask, and vice versa for a task in which high frequency information is important (Delord, 1998). For scene gist, we would therefore predict that the information most useful for recognizing gist is also most efficient at masking gist. Thus, by systematically varying both the spatial characteristics of the mask relative to the target, and the timing of the mask onset relative to the target, it should be possible to determine *what* information contributes to scene gist recognition and *when* it contributes.

The Nature of the Mask and Scene Gist Recognition

For most real world scenes, the average Fourier amplitude of spatial frequencies drops off approximately as the reciprocal of spatial frequency, 1/f^α, with α often equal to 1 (Field, 1987)(see Figure 7). In contrast, white noise masks have a flat spatial frequency amplitude distribution, which differs from that of real world scenes. Different types of noise masks (e.g., white noise, 1/f noise) vary in their masking effectiveness depending on the target. For example, Losada and Mullen (1995) found that for Gabor targets masking by 1/f noise was equal for all spatial frequencies, whereas masking by white noise was less effective than 1/f noise at low frequencies and more effective than 1/f noise at higher frequencies. If natural scenes carry important information for recognizing gist in the low spatial frequency range (Schyns & Oliva, 1993), then 1/f noise masks should be more efficient than white noise masks at disrupting scene gist. Consistent with this idea, natural scenes are masked more effectively by low-frequency than high-frequency white noise (Harvey, Roberts, & Gervais, 1983), and conversely, natural scenes are particularly effective at masking low spatial frequency information (Chandler & Hemami, 2003). To our knowledge, however, no studies have directly contrasted the efficiency of 1/f noise versus white noise in masking scene gist.

In addition to the spatial similarity of the target and mask, the conceptual identifiability of the mask also contributes to masking. This has been shown for immediate recognition of letters (Michaels & Turvey, 1979) and faces (Bachmann, Luiga, & Poder, 2005), and for delayed recognition memory for scenes (Intraub, 1984; Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Potter, 1976). Such *conceptual masking* is assumed to operate at a higher cognitive level than spatial masking, and to critically involve switching attention from the target to the mask (Bachmann, Luiga, & Poder, 2005; Intraub, 1984; Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Michaels & Turvey, 1979; Potter, 1976). In the case of scene recognition memory,

evidence for a distinct conceptual masking process comes from the finding that recognizable scenes used as masks interrupt memory consolidation more effectively than "noise" masks composed of random configurations of color and form (Intraub, 1984; Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Potter, 1976). Also, conceptual masking does not occur until the target scene has been processed long enough to identify it (Loftus & Ginn, 1984); recognizable scenes used as masks and noise masks were equally effective with a 50 ms SOA, but recognizable scene masks were more effective with a 350 ms SOA.

Unanswered Questions about Masking Scene Gist Recognition

Despite increasing interest in scene perception research, many different types of scene masks are commonly used, with little justification for the particular mask used and almost no evidence for the relative efficiency of one type of mask over another. The dearth of comparative studies of scene masking is remarkable given the usefulness of masking as a tool to understanding the information underlying scene gist perception and its time course of use (though see Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Rieger, Braun, Bulthoff, & Gegenfurtner, 2005). In fact, as noted above, by varying the spatial and temporal parameters of the mask vis-à-vis the target, we should be able to infer the information used to perceive scene gist and its time course of use. As part of this enterprise, we can address whether noise masking, structural masking, and conceptual masking are distinct masking mechanisms or whether they rely on similar processes.

Consider the evidence for conceptual masking. To demonstrate that conceptual masking results from higher-level semantic processing, a study must first demonstrate that the greater masking by meaningful masks (the conceptual masking effect) is not simply due to greater amplitude similarity between targets and masks when using other scenes as masks. However, to

date, such studies have either insufficiently controlled for amplitude spectrum differences between the meaningful and meaningless masks, or have produced inconclusive results. Several early studies (Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Potter, 1976) used noise masks with little amplitude similarity to the meaningful masks. Another study (Intraub, 1984, Exp 3, Nonsense Condition) carefully equated meaningful and meaningless masks in terms of their shapes, but likely had important differences in terms of their amplitude spectra. A more recent study (Bachmann, Luiga, & Poder, 2005) used noise masks matched to meaningful (face) masks in terms their spatial frequencies, but did not match their orientations. Finally, one study (Intraub, 1984, Exp 3, Inverted Condition) compared upright versus inverted scene masks, thus equating amplitude spectra while varying recognizability. However, unlike the above-mentioned studies, this one did not produce a significant conceptual masking effect, perhaps because the inverted masks were somewhat recognizable. Thus, to more powerfully test the conceptual masking hypothesis in the domain of scene gist recognition, one needs a study that varies the recognizability of masks while controlling for both their spatial frequencies and orientations, given that both are hypothesized to contribute to gist recognition and to spatial masking.

THE CURRENT STUDY

The primary goal of this study is to understand the nature of the information used for scene gist recognition, and in particular to determine whether unlocalized amplitude information is useful for that purpose. The approach uses sophisticated image processing algorithms to randomize localization of amplitude information in scenes, while maintaining other low-level image characteristics, and compares viewers' ability to recognize the gist of such images relative to normal scenes. Then, the same manipulated images are used as visual masks for briefly flashed normal scenes, and compared with normal scenes and white noise as masks, in order to

determine both the information used to recognize gist and its time course. An important point of this integrative approach to visual masking is that it takes into account both its spatial and temporal dimensions, thus laying the foundation for a more principled and systematic use of masking in scene perception research.

To preview the results, Experiment 1 tests whether amplitude spectrum information is sufficient for scene gist recognition, or whether phase spectrum information is necessary, and shows that gist information is increasingly impaired with increased phase randomization, suggesting that amplitude information is insufficient. The remaining three experiments buttress this conclusion using our masking methodology, while also providing strong evidence for both conceptual and spatial masking of gist. Experiment 2 shows that gist masking varies with the degree of mask phase randomization. Experiment 3 replicates this effect, while also showing that masks having 1/f amplitude spectra are more effective at masking scene gist than noise masks having flat amplitude spectra, but that unlocalized amplitude spectrum differences across scene categories make no difference in gist masking. Experiment 4 shows that the apparent conceptual masking effect for scene gist occurs even at the earliest stages of target processing.

EXPERIMENT 1

This experiment explores whether the unlocalized amplitude spectrum and mean luminance information of a scene are useful for gist recognition. To the extent that unlocalized amplitude spectrum information contributes to gist recognition, viewers should be able to recognize scene gist well above chance even when the phase of an image is completely randomized, as long as it retains its amplitude spectrum. Conversely, if phase information is necessary to recognize scene gist, as it is for object recognition (Oppenheim & Lim, 1981; Piotrowski & Campbell, 1982; Sadr & Sinha, 2004; Tadmor & Tolhurst, 1993; Wichmann,

Braun, & Gegenfurtner, 2006), then gist recognition for completely phase-randomized scenes should be at chance levels. Thus, this experiment systematically manipulates the extent of phase randomization and display duration to examine the contribution of unlocalized amplitude information to recognizing the gist of unmasked scenes. Little effect of duration is expected because the images are unmasked.

In addition, this experiment will serve two further functions within the series of experiments reported in this study. First, the results of the current experiment can validate later experiments' use of masking to measure the utility of unlocalized amplitude information for scene gist recognition. To the extent that the direct method used in the current experiment and the masking methods in later experiments produce consistent results, it will validate the masking methods. Second, this experiment will constitute the first step in a rigorous test of the conceptual masking hypothesis. By measuring scene gist recognizability as a function of phase randomization level, later experiments can then determine the effects of mask recognizability on masking when amplitude spectrum information is held constant.

Method

Participants. 96 Kansas State University undergraduate students (60 female, mean age = 19.5 years, age range = 18-44) participated for course credit. All had normal or corrected near vision of at least 20/30 based on a Sloan near acuity letter chart.

Materials. 300 gray scale photographs (1024x674) from the Corel Image Database and other sources were drawn from 10 scene categories, five natural: Beach, Desert, Forest, Mountain, and River, and five man-made: Farm, Home Interior, Market, Pool, and Street, with 30 images in each category. Images were displayed on a 17" Gateway EV910 monitor (85 Hz refresh rate), and viewing distance was fixed to 53.3 cm using a chin rest. Each image subtended

a visual angle of 34.39º x 27.11º. Participants responded using a keyboard, and wore headphones to reduce environmental noise.

Images were modified by parametrically varying their degree of phase randomization, while maintaining their spatial frequency amplitude spectra and mean luminance, similarly to the RISE algorithm (Sadr & Sinha, 2001; Sadr & Sinha, 2004; see also Wichmann, Braun, & Gegenfurtner, 2006 for a similar approach). This process, maintains the energy distribution at each spatial frequency while changing the localization of information within the image (for details, see Appendix A, and Sadr & Sinha, 2004). Figure 1 illustrates the images produced by complete phase randomization for scenes from three categories and their accompanying spatial frequency amplitude spectra based on a Fast Fourier Transformation. The figure shows that important spatial frequency, orientation, and luminance information, which can be used to discriminate between the three images, is maintained in the fully randomized phase versions. Figure 2 shows the degrees of phase randomization used in the experiments. These levels were chosen based on pilot testing to span a wide range of scene identifiability.

[[Insert Figures 1 and 2 about here.]]

Design & procedure. Phase randomization level was a between-subjects factor, with random assignment of levels to subjects. Each participant viewed all 300 images with one of the six levels of phase randomization (0, 0.1, 0.25, 0.4, 0.6, and 1.0 where 0 is a normal image and 1.0 is a fully phase-randomized one). Each participant viewed 60 images (6 from each of the 10 scene categories) at each of 5 different display durations (12, 24, 59, 106, and 200 ms). The durations were chosen to span a wide range, including durations near 10, 50, 100, and 200 ms (though all durations were multiples of the 85 Hz refresh cycle).

Figure 3 (left panel) depicts the sequence of events in an experimental trial. On each trial of the experiment, participants looked at a fixation cross that prompted them to push a key to display the scene image. The image was followed by a 750 ms blank interval, and then by a post-cue scene category label. Subjects pressed 'yes' if the target image matched the post-cue or 'no' if it did not. Each of the 10 cue categories was used equally often, for both valid and invalid trials, and each of the scene categories was cued validly and invalidly equally often. Participants were encouraged to respond with their first impression, whether they were sure or not, and to respond as quickly and accurately as possible. Before beginning the experiment, participants completed a category-learning task with 90 images, 9 from each scene category, in order to acquaint them with the scene category labels. None of these scenes appeared in the experimental trials. They then completed 32 scene gist recognition practice trials, to familiarize them with the experimental task. Images in the practice trials were in the same phase randomization condition as the actual experiment for that subject. Trials were self-paced, and participants were allowed to take breaks at anytime, with the 300 trials generally taking 15 minutes to finish.

[[Insert Figure 3 about here.]]

Results

[[Insert Figure 4 about here.]]

As can be seen in Figure 4, with complete phase randomization, viewers were unable to get any useful information about the scene gist using only the amplitude spectrum information and mean luminance of images. Phase randomization had a robust, monotonic effect on scene gist recognition ($F(5, 90) = 481.16$, $p < .001$), with nearly perfect accuracy (0.95) for unaltered images (0 phase randomization) and chance performance (0.50) with a phase randomization factor of 0.6 or greater.

The influence of stimulus duration on scene gist recognition was more complex (see Figure 4). An overall effect of duration (Pillai's Trace = 0.455, $F(4, 87) = 18.18$, $p < .001$) was qualified by an interaction with the level of phase randomization (Pillai's Trace = 0.67, $F(20, 360) = 3.62$, $p < .001$). Duration had a small but significant effect with no phase randomization (randomization level, hereafter called "RAND" = 0; Pillai's Trace = 0.652, $F(4, 12) = 5.64$, $p = .009$) increasing mean accuracy from .92 to .97, but had essentially no effect with greater than 50% phase randomization (RAND = 0.6; Pillai's Trace = 0.089, $F(4, 12) = 0.291$, $p = .878$; RAND = 1.0; Pillai's Trace = 0.293, $F(4, 12) = 1.24$, $p = 0.345$), with accuracy ranging from .50 to .53. The relatively weak effect of stimulus duration when scenes were normal is due to a ceiling effect and is as expected given that there was no mask. Without a mask, viewers can rely on sensory persistence to process the target after stimulus offset (Breitmeyer & Ogmen, 2006; Loftus & Mclean, 1999; Sperling, 1963). The null effect of duration with high levels of randomization reflects the fact that viewers could get no useful information for gist from such images, regardless of their duration. Thus, duration primarily affected recognition with intermediate levels of phase randomization (levels 0.1-0.4, all Fs $(4, 12) > 12$, all ps $<.001$).

Discussion

Experiment 1 established the relationship between level of phase randomization and scene recognizability for unmasked images of varying durations. Consistent with evidence for the effects of phase randomization on the appearance of scenes and objects (Oppenheim & Lim, 1981; Piotrowski & Campbell, 1982; Sadr & Sinha, 2004; Tadmor & Tolhurst, 1993; Wichmann, Braun, & Gegenfurtner, 2006), viewers obtained no useful gist information from scenes' randomly localized amplitude spectra and mean luminance alone; this strongly suggests that the unlocalized amplitude spectrum of a scene is not sufficient to identify its basic level category.

This finding appears to contradict evidence that unlocalized amplitude information can contribute to scene gist recognition (Guyader, Chauvin, Peyrin, Hérault, & Marendaz, 2004; Oliva & Torralba, 2001; Oliva, Torralba, Guerin-Dugue, & Herault, 1999). However, the current experiment could only provide useful information regarding the spatial dimension of scene gist recognition. By using masking, we can also measure the time course of the use of amplitude information in scene gist recognition.

The current results also provide a tool for exploring the information underlying masking effects when scenes are used as masks. First, we can determine the extent to which phase-randomization produces similar effects on both scene gist recognition, in the current experiment, and scene gist masking, in later experiments. If the effects are similar, this will serve to validate the use of masking to explore the information underlying scene gist recognition. Second, we can determine whether scene recognizability, or amplitude spectra, or both, determine masking effectiveness. Specifically, we can rigorously test the conceptual masking hypothesis by using masks that have equal amplitude spectra but that vary in recognizability by varying their level of phase-randomization.

EXPERIMENT 2

Experiment 2 investigated whether a conceptual masking effect occurs for scene gist recognition, or if masking of scene gist recognition by recognizable scene image masks can be explained simply in terms of the spatial frequency amplitude spectra of the masks. Viewers tried to identify briefly presented target scene images that were followed immediately by scene image masks. Individual mask images were yoked across conditions varying in phase randomization (and, hence, degree of recognizability) but, across these conditions, the masks had identical spatial frequency amplitude spectra and mean luminance. Together with the results of

Experiment 1, the current experiment tested a novel prediction based on the conceptual masking hypothesis, namely that masking will vary monotonically with the degree of mask recognizability. The alternative hypothesis is that amplitude spectrum similarity between target and mask determines masking. This predicts no difference in masking between normal and phase-randomized scene masks sharing identical amplitude information. In sum, if increasing mask recognizability increases scene gist masking, it would be consistent with the conceptual masking hypothesis. Conversely, if scene gist masking is unaffected by mask recognizability, it would be consistent with the amplitude similarity hypothesis.

Method

Participants. 72 Kansas State University undergraduate students (41 female, mean age = 19.3 years, range = 18-23) participated for course credit. All had normal or corrected near vision of at least 20/30, scored using a Sloan near acuity letter chart.

Stimuli. The same set of images was used in this experiment, with the same 6 possible levels of phase randomization. All scenes were used twice, once as a target and once as a mask. The shortest duration (12ms) was replaced by a longer duration (306 ms) to better equate performance across Experiments 1 and 2 (durations used: 24, 59, 106, 200, and 306 ms). As in the original conceptual masking studies that used an RSVP paradigm (Intraub, 1981, 1984; Potter, 1976; Potter & Levy, 1969), we equated the target and mask durations on each trial so that the mask:target duration ratio was constant (1:1) across all target durations. The inter-stimulus interval between target and mask was 0 ms. Thus, target duration equaled the SOA. This approach is well-suited to studying the effect of varying target duration on recognition of masked stimuli.

Design & procedure. The procedure is schematically represented in Figure 3 (middle panel). On each trial, participants viewed a scene followed by a mask from a different scene category (e.g., in Figure 3, a river masked by a market). The target scene was unaltered, but the mask varied in its extent of phase randomization (the extent of randomization was a between-subjects variable). The original pairing of targets and masks was random, but was yoked across phase randomization conditions in order to allow comparisons of the effects of phase randomization with the same image pairs. In the normal image mask condition (RAND = 0), each image was seen twice, once as a target and once as a mask. In the other masking conditions (RAND = 0.1-1.0), each original image was seen once as a target and its phase-randomized version was seen once as a mask. Trial order was randomized for each participant. Following the scene and mask, a label (or "cue") appeared and participants reported whether or not it named the target scene category. The cue was correct on 50% of trials, and when it was incorrect, it never matched the scene category of the mask, and participants were explicitly told this. All categories of cues were used equally often, and were correct equally often.

Results

[[Insert Figures 5 and 6 about here.]]

Masking was strongest with no phase randomization, and decreased with increasing phase randomization ($F(5, 66) = 15.83$, $p < .001$; see Figure 5), suggesting that previous conceptual masking results cannot be explained entirely by greater similarity in the amplitude spectra of scene masks and targets. Masking was also affected by stimulus duration (Pillai's Trace = 0.885, $F(4, 63) = 120.77$, $p < .001$), and the effect of stimulus duration interacted with the level of phase randomization (Pillai's Trace = 0.971, $F(20, 264) = 4.23$, $p < .001$). In essence, the effect of duration was larger for masks with little phase randomization. Figure 6

depicts these relationships as masking effectiveness relative to the unmasked normal image gist recognition from Experiment 1. Masking decreased monotonically with increasing stimulus duration, but only at lower levels of phase randomization (RAND = 0 - 0.25). At the highest levels of phase randomization (RAND = 0.6-1.0), masking is minimal.

Discussion

Masking strength varied monotonically with mask recognizability, consistent with the construct of conceptual masking based on studies of scene recognition memory (Intraub, 1981, 1984; Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Potter, 1976; Potter & Levy, 1969) and immediate recognition of letters (Michaels & Turvey, 1979) and faces (Bachmann, Luiga, & Poder, 2005). The current experiment is the first to show that masking varies monotonically with degree of mask phase randomization and identifiability. These results, together with the fact that mask amplitude spectra were held constant across mask recognizability conditions, weakens the argument that better masking by scenes than by noise masks in previous conceptual masking studies was an artifact of the degree of target/mask amplitude spectrum similarity.

Interestingly, our results were inconsistent with the results of Loftus and Ginn (1984), in that we found a conceptual masking effect even at extremely short SOAs (e.g., 24 ms), when, according to Loftus and Ginn, only perceptual, not conceptual, masking occurs. Our results suggest that conceptual masking can affect even early encoding processes (see also Bachmann, Luiga, & Poder, 2005).

Masking in our experiment varied as a function of mask identifiability even though the amplitude spectra of target and mask were relatively similar. These results were therefore inconsistent with our spatial masking hypothesis, that target/mask amplitude spectrum similarity should produce strong scene masking, based on the spatial masking literature (Carter & Henning,

1971; De Valois & Switkes, 1983; Henning, Hertz, & Hinton, 1981; Legge & Foley, 1980; Losada & Mullen, 1995; Solomon, 2000; Stromeyer & Julesz, 1972; Wilson, McFarlane, & Phillips, 1983; Yang & Stevenson, 1998). A previous study (Harvey, Roberts, & Gervais, 1983) showed that variations in the spatial frequency contents of noise masks have their strongest effects on scene recognition at SOAs < 40 ms. Thus, we might expect to have found the strongest effect of spatial masking by phase-randomized images at our shortest SOA (24 ms). Yet even at that short SOA, the masking by phase-randomized images was minimal. Perhaps this was because Experiment 2 used the same duration for masks and targets (ratio = 1:1) whereas Harvey, et al. (1983) used a stronger mask:target duration ratio of 3:1. Thus, our phase-randomized masks may simply have been too weak to show the effects of amplitude spectrum similarity. Furthermore, because our ISI was fixed at 0 ms, the effects of duration and SOA were confounded. If SOA is the critical temporal variable in masking (Di Lollo, von Muhlenen, Enns, & Bridgeman, 2004; Kahneman, 1967; Turvey, 1973), then by varying ISI while holding target duration constant (thus varying SOA), we should be able to detect the time course of masking effects more readily.

Another possible explanation for the inconsistency of the results with the spatial masking hypothesis is that phase randomization affects not only mask recognizability, but also some other critical variable, such as the mask:target contrast ratio (Breitmeyer & Ogmen, 2006), which in turn affects masking strength. Previous research has shown that phase randomization reduces contrast (Bex & Makous, 2002; Wichmann, Braun, & Gegenfurtner, 2006), which might explain the decreased masking we found as a function of phase randomization. To explore this possibility, we carried out a control experiment in which we reduced the contrast of each normal image mask to match the perceived contrast of its fully phase-randomized version (using the

lower of two raters' perceived contrast ratings), and compared the gist masking produced by each (reduced contrast normal image (i.e., RAND = 0) masks and fully phase-randomized image (RAND = 1.0) masks). As expected, reducing the contrast of the normal image masks increased gist accuracy (i.e., weaker masking)(Normal image mask: $M = .77$, $SD = .08$; Reduced contrast normal image mask: $M = .82$, $SD = .05$; $t(20) = 1.94$, $p < .05$ (one-tailed)). However, the reduced contrast normal image masks still produced lower gist accuracy (i.e., stronger masking) than the fully phase-randomized masks (Phase-randomized mask: $M = .92$, $SD = .07$; ($t(21) = 3.98$, $p < .001$ (one-tailed)). Thus, the reduced masking produced by phase-randomization in this experiment cannot be solely attributed to contrast reduction. Nevertheless, one way to control for contrast effects would be to equalize contrast across all target and mask images.

Experiment 2 also lacked a baseline against which to judge the effects of target/mask amplitude spectrum similarity on gist masking. White noise would suit this purpose well because it has a radically different amplitude spectrum from that of natural scenes. If the 1/f amplitude spectrum of natural scenes is important in the masking of one scene by another, then fully phase-randomized scenes should produce stronger masking than white noise.

A final question is what gist masking effects, if any, are caused by unlocalized amplitude spectrum differences between scene categories. On average, scenes tend to have a 1/f spatial frequency amplitude spectrum, however the amplitude spectra of individual scenes can differ substantially (Langer, 2000). Such differences between scene categories are illustrated in Figure 1, and form the basis for arguing that scenes' unlocalized amplitude spectra are used to recognize their gist (Oliva & Torralba, 2001). One approach to determining the effect of between-category unlocalized amplitude spectrum differences on gist masking is to compare masking caused by a) phase-randomized masks from different scene categories than the target, versus b) phase-

randomized mask versions of the target images themselves. In the latter case, in which target and mask have identical amplitude spectra, one might predict greater masking due to greater amplitude similarity; conversely, one could also predict less masking based on the redundant amplitude spectrum information from the mask helping to categorize the target scene. If, however, masking is equivalent for phase-randomized masks generated from a) scenes from a different category than the target, and b) the target itself, then scenes' unlocalized amplitude spectra likely play little role in scene gist masking, or by extension, in scene gist recognition.

EXPERIMENT 3

This experiment followed up on Experiment 2, which showed conceptual masking of gist by recognizable scene masks, by examining potential spatial masking effects on scene gist caused by information in scenes' unlocalized amplitude spectra. The experiment examined such effects by including two new random phase masking conditions: one having a very different amplitude spectrum from that of scenes, namely white noise, and the other having identical amplitude spectra to that of the targets, namely phase-randomized versions of the targets themselves. If 1/f spatial frequency amplitude spectrum of scenes carries important information for scene gist, then phase-randomized scene image masks should cause greater masking than white noise. Furthermore, if the unlocalized amplitude differences between scene categories carry important information for scene gist, then there should be differences in the scene gist masking produced by the following two types of masks: 1) the phase randomized version of the target, versus 2) the phase randomized version of a scene from a different category.

The experiment controlled for target/mask contrast differences by equalizing the mean luminance and contrast of all targets and masks. It more thoroughly sampled the early SOA range (from 10-50 ms) in which spatial masking is more likely to occur, and used a stronger

mask:target (4:1) duration ratio more likely to produce spatial masking effects. In order to more carefully examine the time course of masking effects, the experiment decoupled target duration and SOA by holding target duration constant and varying SOA.

Method

Participants. 96 Kansas State University undergraduate students (58 female, mean age = 19.0 years, age range = 18 to 29) participated for course credit. Two subjects were excluded for failure to follow instructions. All participants had normal or corrected near vision of at least 20/30, scored using a Sloan near acuity letter chart.

Stimuli. The entire set of images used in the current experiment, including all target and masking images, were equalized for both mean luminance and RMS contrast, with the latter having been shown to be highly correlated with perceived natural image contrast (Bex & Makous, 2002)(see Appendix B for details). This equalization resulted in a contrast reduction for most images; nevertheless, pilot testing indicated that the unmasked normal images were still highly recognizable. The resultant image processing procedure, including details of the phase randomization procedure, and control of mean luminance and RMS contrast, was equivalent to the RISE algorithm (Javid Sadr, 5/8/2006).

Masks in the current experiment included normal images (RAND = 0) and fully phase-randomized images (RAND = 1.0), both of which have, on average, 1/f spatial frequency amplitude spectra, and a set of 300 white noise images, which have flat amplitude spectra (see Figure 7). As shown in Figure 7, the spatial frequency amplitude spectra of the normal and fully phase-randomized masks are identical, and compared to the white noise masks, have more power in the lower frequency range, and less in the high frequency range.

[[Insert Figure 7 about here.]]

Design & procedure. The procedure is schematically represented by Figure 3 (right panel) and except as noted, was identical to Experiment 2. As illustrated in Figure 8, there were four types of masks: 1) a normal image (RAND = 0) from a different scene category than the target, 2) a fully phase-randomized image (RAND = 1.0) from a different scene category than the target, 3) a fully phase-randomized version (RAND = 1.0) of the target image, and 4) a white noise image. Mask type was a randomly assigned between-subjects variable (23-24 subjects each). As in Experiment 2, the target-to-mask pairings were yoked across mask conditions, except for condition 3, in which the mask was the fully phase-randomized version of the target (e.g., Mountain 23 masked by the fully phase-randomized version of Mountain 23).

[[Insert Figure 8 about here.]]

Target duration was fixed at 12 ms, which is near the minimum necessary for above-chance gist recognition performance as shown in Experiment 2 (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005), and the mask duration was fixed at 48 ms, creating a strong (4:1) mask:target duration ratio more likely to show spatial masking effects. As shown in Figure 3 (right panel), a blank inter-stimulus interval (ISI) was presented between the target and mask images for 0-84 ms, creating SOAs (= target duration + ISI) between target and mask of 12, 24, 36, 48, or 96 ms (manipulated within subjects). These SOAs were chosen to focus primarily on the first 50 ms of processing, when one generally finds both the strongest spatial frequency masking effects on scene recognition (Harvey, Roberts, & Gervais, 1983) and many important early scene gist processes (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Renninger & Malik, 2004; Schyns & Oliva, 1993). We also included an SOA near 100 ms, by which time scene gist recognition generally reaches asymptote as shown in Experiment 2 (Biederman, Rabinowitz, Glass, & Stacy, 1974; Potter, 1976). Furthermore, while Experiment 2 showed that

the time course of conceptual masking of immediate gist recognition differs from that previously reported for memory (Loftus & Ginn, 1984), we might still expect that perceptual processes would dominate masking at SOAs < 50 ms and that conceptual processes might be stronger at an SOA of roughly 100 ms (Bachmann, Luiga, & Poder, 2005).

Results

[[Insert Figure 9 about here.]]

Figure 9 shows both conceptual and spatial masking effects on scene gist recognition. The recognizable normal image (RAND = 0) mask condition produced significantly lower gist accuracy (i.e., more masking) than each of the other three unrecognizable mask conditions (Sidak, all *p*s < .001), replicating the conceptual masking effect on immediate gist recognition of Experiment 2. Importantly, Figure 9 also shows that both fully phase-randomized scene mask conditions produced significantly lower gist accuracy than the white noise mask condition (Sidak, both *p*s < .001), suggesting that the unlocalized 1/f amplitude spectrum of scenes is somewhat effective at masking scene gist. On the other hand, Figure 9 shows that the two fully phase-randomized masking conditions produced virtually identical masking (*M* difference = 0.019, *SE difference*= .018, *t(238)* = 1.51, *p* = .877, n.s.), suggesting that unlocalized amplitude spectrum differences between scene categories do not affect gist masking. This is inconsistent with the idea that differences between scene categories in unlocalized amplitude information are useful for recognizing gist.

Figure 9 also shows strong time course effects on scene gist masking, with longer SOAs producing greater scene gist accuracy (Pillai's Trace = 0.684, F(4, 88) = 47.53, p < .001). Furthermore, the time courses of spatial and conceptual masking effects differed (SOA x Masking Condition interaction: Pillai's Trace = 0.546, F(12, 270) = 5.00, p < .001). Figure 9

shows strong spatial masking effects based on scenes' amplitude spectra at the earliest stages of processing (SOA = 12 ms). At that early stage, the three masking conditions with natural amplitude spectra (the normal (RAND = 0) and both fully phase-randomized (RAND = 1.0) conditions) did not differ significantly from each other, but produced significantly lower accuracy (stronger masking) than the white noise mask condition. In contrast, conceptual masking effects were strongest at later stages of scene gist processing (SOA = 94 ms). At that later stage, all three unrecognizable masking conditions (white noise and both fully phase-randomized (RAND = 1.0) conditions) produced greater accuracy (less masking) than the recognizable normal image (RAND = 0) masks.

Discussion

Experiment 3 confirms the importance of phase information in scene gist masking and the likely existence of conceptual masking of immediate gist recognition. This experiment also shows that randomly localized amplitude spectrum differences *between* scene categories are insufficient to produce differences in scene gist masking. Such a result is inconsistent with the idea that inter-category unlocalized amplitude spectrum differences are useful for recognizing scene gist. The theoretical implication is that, although scene categories may well differ in their amplitude spectra, and such differences have been hypothesized to allow gist recognition, this finding suggests that such differences do not contribute to basic level scene gist recognition. This masking result is consistent with the results of Experiment 1, which showed that unlocalized amplitude differences between scene categories were insufficient for scene gist recognition. Nevertheless, the current experiment shows that masks having a 1/f spatial frequency amplitude spectrum are more efficient at masking scene gist than masks having an unnaturally flat amplitude spectrum. This spatial masking effect of the 1/f spatial frequency

amplitude spectrum primarily occurs during early perceptual processes (i.e., SOAs ≤ 50 ms). At later stages of processing (i.e., SOA ≈ 100 ms), scene gist masking is affected more by structured phase information and/or the recognizability of masks than by their unlocalized amplitude spectra. These results are consistent with the idea that masking at short SOAs involves more peripheral processes, while masking at longer SOAs involves more central processes, such as attention (Bachmann, Luiga, & Poder, 2005; Loftus & Ginn, 1984; Michaels & Turvey, 1979).

The above interpretations of the time course of gist masking may need to be tempered, however, because there is an open question about what occurred at the earliest stages of processing in Experiment 3. Specifically, as noted above, at the shortest SOA (12 ms), masking by fully phase-randomized (RAND = 1.0) and normal (RAND = 0) image masks was essentially equal. This may indicate that information encoded by the phase spectrum is of little use for very early peripheral processes. However, Figure 9 suggests that this lack of difference may simply reflect a floor effect in the normal image (RAND = 0) masking condition. If so, then raising performance, for example by using a smaller mask:target duration ratio, should produce differential masking between fully phase-randomized and normal image masks, even at a 12 ms SOA, as found in Experiment 2.

EXPERIMENT 4

This experiment resolves the question of whether unlocalized scene information conveyed by the amplitude spectrum is sufficient for gist masking at the earliest levels of processing, or whether localized information conveyed by the phase spectrum is necessary even then. Experiment 3 suggested that amplitude information may be sufficient at the earliest point in gist processing (SOA = 12 ms), because there was no difference between the completely phase randomized (RAND = 1) and normal (RAND = 0) image masking conditions, but this may have

been due to a floor effect in the latter condition. Experiment 4 resolves this issue by replicating two such masking conditions from Experiment 3, while varying masking strength by means of the mask:target duration ratio. If conceptual masking occurs even at the earliest stages of scene processing, normal image masks should cause greater scene gist masking than completely phase-randomized masks even at 12 ms SOA. Additionally, a no-mask control condition is included to assess the impact of the mean luminance and RMS contrast equalization on baseline unmasked scene gist accuracy.

Method

Participants. 104 Kansas State University undergraduate students (60 female, mean age = 19.4 years, age range = 18 to 30) participated for course credit. All participants had normal or corrected near vision of at least 20/30, scored using a Sloan near acuity letter chart.

Stimuli. The stimuli were a subset of those used in Experiment 3 (described below in terms of mask types).

Design & procedure. The design and procedure was identical to that of Experiment 3, except as follows. First, we used only two of the mask types used in Experiment 3: 1) a normal image (RAND = 0) from another scene category than the target, and 2) a fully phase-randomized image (RAND = 1.0) from another scene category than the target, with mask type a between-subjects variable (46 subjects for each mask type, and 12 subjects in a no-mask control condition, with random subject-to-condition assignment). As in Experiments 2 and 3, the target-to-mask pairings were yoked across mask conditions.

The most important difference from Experiment 3 was that the mask:target duration ratio was varied from 1:1 to 4:1, by fixing target duration at 12 ms and varying mask duration from 12-48 ms (manipulated between subjects, 14-17 subjects randomly assigned per condition). This

was the key manipulation of the current Experiment. All ISIs and SOAs were identical to those in Experiment 3.

Results

[[Insert Figure 10 about here.]]

The current experiment explains the apparently equivalent gist masking caused by localized and unlocalized scene amplitude information at the earliest stages of gist processing in Experiment 3. As shown in Figure 10, we replicated the Experiment 3 interaction between mask type and SOA, such that the SOA effect was greater in the fully phase-randomized (RAND = 1) than the normal image (RAND = 0) masking condition (Pillai's Trace = 0.450, $F(4, 83) = 16.96$, $p < .001$), which is consistent with a possible floor effect in the normal image masking condition. To manipulate this possible floor effect, we varied the mask:target duration ratio, and the three panels of Figure 10 show that this strongly affected accuracy ($F(2, 86) = 19.83$, $p < .001$), irrespective of mask type ($F(2, 86) = 1.04$, $p = .359$, n.s.). The key tests of the floor effect were in terms of several *a priori* planned comparisons. First consider the masking condition that replicates the key condition in Experiment 3, the normal (RAND = 0) image mask, at 12 ms SOA, with the strongest masking ratio (Mask: Target = 4:1, Figure 10, top panel). Accuracy in this condition ($M = 0.51$, $SD = .04$) did not differ significantly from chance (0.5) ($t(13) = 1.07$, $p = .303$ (two-tailed), n.s.), and there was a flat, 0 slope from 12-24 ms SOA (both means = 0.51), which together strongly suggest a floor effect. No such floor effect is found in the weakest masking ratio condition (1:1 mask:target ratio, Figure 10, bottom panel), with accuracy at the shortest SOA (12 ms) in the normal image condition (RAND = 0) ($M = 0.56$, $SD = .05$) significantly above chance (0.5), $t(16) = 5.26$, $p < .001$ (two-tailed), and a positive slope from 12-24 ms SOA (0.2% accuracy increase/ms SOA). (The intermediate masking ratio condition

(2:1 mask:target ratio, Figure 10, middle panel), produced results closer to the 4:1 masking ratio condition, with accuracy (M = 0.53, SD = .03) only slightly, but significantly, greater than chance (0.5), $t(14) = 3.81$, $p < .002$ (two-tailed), but a flat slope between the 12 and 24 ms SOAs (both means = 0.53)). We therefore conclude that there was a floor effect in Experiment 3 at the shortest SOA (12 ms) in the normal image (RAND = 0) masking condition, which was largely due to the strong (4:1) mask:target duration ratio.

This floor effect created a false equivalency in the gist masking caused by localized and unlocalized amplitude spectrum scene information at the earliest stages of processing in Experiment 3. The top panel of Figure 10, where the floor effect is found, replicates the equivalent masking by localized and unlocalized amplitude information found in Experiment 3 (RAND = 0: M = 0.51, SD = .04; RAND = 1: M = 0.53, SD = .06), $t(27) = -1.17$, $p = .251$ (two-tailed), n.s.). In contrast, amplitude localization strongly affects scene gist masking even at the shortest SOAs in the bottom panel of Figure 10, where there is no floor effect. There we see significantly lower accuracy in the normal image masking condition (RAND = 0: M = 0.56, SD = .05) than in the fully phase-randomized masking condition (RAND = 1: M = 0.63, SD = 0.08), $t(31) = -2.67$, p =.012 (two-tailed)). (In the intermediate masking ratio (2:1) condition, the results are similar to those in the 4:1 masking ratio condition, with no significant difference in accuracy between the normal (RAND = 0: M = 0.53, SD = .03) and fully phase-randomized image masking conditions (RAND = 1: M = 0.55, SD = 0.05), $t(28) = -1.54$, p =.135 (two-tailed).) We therefore conclude that localized amplitude scene information is important for gist masking even at the earliest stages of processing.

Finally, the inclusion of the no-mask control condition allowed us to gauge the overall effect of the mean luminance and RMS contrast equalization on perception of the targets, relative

to that of Experiment 1, which were not equalized. A comparison of the 12 ms duration normal image (RAND = 0) condition in Experiment 1 (Figure 4)(M = 0.92) with the no-mask condition in Experiment 4 (Figure 10)(M = 0.86) shows that the equalization did somewhat reduce accuracy, though accuracy was still quite high. This set an upper bound for accuracy in the masking conditions, and one can see in the bottom panel of Figure 10 that the fully phase-randomized (RAND = 1) masking condition was approaching this level of accuracy at the longest SOA (95 ms), whereas this was not the case in the normal image (RAND = 0) masking condition.

Discussion

The results of Experiment 4 show that even at the earliest stages of processing (i.e., 12 ms SOA), scene gist perception depends on localized information. Specifically, gist recognition was less disrupted by fully phase-randomized scene masks than by normal scene masks, though both had identical amplitude spectra. Thus, even at very earliest stages of scene gist processing, unlocalized amplitude spectrum information is insufficient for scene gist recognition. Also, consistent with Experiment 2, conceptual masking of immediate gist occurs at even the earliest stages of processing, in contrast to what has been shown for scene memory (Loftus & Ginn, 1984).

GENERAL DISCUSSION

The current study examined the role of unlocalized amplitude spectrum information in recognizing scene gist (Gorkani & Picard, 1994; Guyader, Chauvin, Peyrin, Hérault, & Marendaz, 2004; Oliva & Torralba, 2001; Oliva, Torralba, Guerin-Dugue, & Herault, 1999; Oppenheim & Lim, 1981; Wichmann, Braun, & Gegenfurtner, 2006), and showed that it is insufficient to recognize a scene's basic level category, based on converging evidence from the

effects of phase randomization on both unmasked gist recognition and the masking of gist recognition. The current study also provides the strongest test to date of the conceptual masking hypothesis, by ruling out the hypothesis that conceptual masking can be explained simply in terms of target:mask amplitude spectrum similarity. Furthermore, this study lays the foundation for more theoretically-based and systematic uses of masking to study scene perception, by applying knowledge from the extensive literature on masking in spatial vision to understanding scene gist recognition.

The Role of Unlocalized Amplitude Spectrum Information in Scene Gist Recognition

The current study contributes to our understanding of the processes involved in scene gist recognition. The fact that people recognize scene gist so incredibly quickly suggests that it may be based on very early processing of low-level stimulus dimensions. One such candidate dimension is the unlocalized amplitude spectra of scenes. To test this hypothesis, we randomized the phase spectra of scenes, while maintaining their amplitude spectra, luminance, and contrast using the RISE algorithm (Sadr & Sinha, 2004), and measured the effects this had on both unmasked scene gist recognition and scene gist masking. Experiment 1 showed that unmasked scenes with 60% or greater phase randomization could not be identified above chance, though their unique amplitude spectra remained unchanged. Then, using masking, Experiments 2-4 showed that between-scene category differences in the amplitude spectra of masking images make no difference in scene gist masking, though the general 1/f amplitude distribution of scenes does. Together, these results suggest that scenes' amplitude spectra provide only limited information for recognizing scene gist.

The current results are therefore inconsistent with the success of the Spatial Envelope model in classifying scenes (at 86% accuracy) using only stationary, globally distributed,

unlocalized information from scenes' amplitude spectra (Oliva & Torralba, 2001), suggesting that human observers may not be sensitive to such information, even though it is potentially useful for scene classification. However, such an argument cannot explain the current study's inconsistency with Guyader, et al. (2004), who found that scenes' randomly localized amplitude spectra significantly primed human observers' scene gist recognition. Importantly, that study used a very simple two-category discrimination task ("beach" vs. "city") and only found relatively small priming effects on reaction times (15-18 ms). Thus, it may be that such effects are only detectable with a more constrained categorization task and response time measures. Alternatively, unlocalized amplitude spectrum information may be useful for identifying scenes, but only at the level of the perceptually primitive "natural" versus "man-made" scene distinction, which Oliva and Torralba (2001) argue is the most fundamental. Such a hypothesis is entirely consistent with the results of Guyader and colleagues (2004) who argued that their "beach" versus "city" distinction, a simplified case of the "natural" versus "man-made" distinction, was based entirely on clear orientation differences (i.e., horizontal = beach, vs. vertical = city). We are currently testing this hypothesis in a series of studies.

Masking studies have played a crucial role in developing theories of spatial vision (Carter & Henning, 1971; De Valois & Switkes, 1983; Henning, Hertz, & Hinton, 1981; Legge & Foley, 1980; Losada & Mullen, 1995; Solomon, 2000; Stromeyer & Julesz, 1972; Wilson, McFarlane, & Phillips, 1983; Yang & Stevenson, 1998), and the current study suggests that masking can be similarly helpful for understanding the information used to recognize gist, particularly given that the results from our masking experiments (Exp 2-4) were consistent with the results of a direct measure of scene gist recognition (Exp 1). Using logic similar to DeLord (1998), we have argued that the information that most efficiently masks scene gist is also the most useful for

recognizing scene gist. Our results are entirely consistent with arguments that second order image statistics based on the unlocalized amplitude spectrum provide insufficient information to recognize scenes. Instead, higher order image statistics that include spatial localization, such as wavelets, are necessary to capture the critical information for recognizing scenes (Field, 1987, 1999; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001; Thomson & Foster, 1997). Such a claim is consistent with more recent versions of the Spatial Envelope Model, which emphasize the importance of spatially localized coding, and specifically layout, in recognizing scenes (Oliva, 2005; Oliva & Schyns, 2000; Oliva & Torralba, in press; Sanocki, 2003; Sanocki & Epstein, 1997; Schyns & Oliva, 1993). Indeed, the results of the current study indirectly support claims for the importance of layout in gist recognition because decreased gist recognition and gist masking accompany the loss of layout information produced by phase randomization. This suggests a prediction worth testing in further research, that increasing layout information in masks will increase scene gist masking.

Our results are also consistent with the idea that low spatial frequency information is important for recognizing scene gist (Loschky & Simons, 2004; McCotter, Gosselin, Sowden, & Schyns, 2005; Oliva & Schyns, 2000; Schyns & Oliva, 1993). Experiment 3 showed that phase-randomized scene images, which have 1/f spatial frequency amplitude spectra, are more efficient at masking scene gist than are white noise images, which have relatively more high frequency information but less low frequency information. Our results are also consistent with Harvey, et al. (1983), who found that lower frequency noise masks were more efficient than higher frequency masks at disrupting scene recognition, but inconsistent with the recent results of Bacon-Mace and colleagues (2005), who found that higher frequency noise masks were more effective at masking animal detection in scenes. One explanation for the latter discrepancy is

that animal detection, a subset of object recognition at the superordinate level, depends more on higher frequency information, whereas scene gist recognition depends more on lower frequency information (though see Oliva & Schyns, 1997).

Conceptual Masking of Immediate Scene Gist Recognition

The current study provides a rigorous test of the existence of conceptual masking as distinct from noise and structural masking, using a novel approach in which we systematically varied masks' recognizability while holding their amplitude spectra, mean luminance, and contrast constant. The study was thus able to show that scene gist masking varies monotonically with mask identifiability, while largely ruling out an alternative explanation in terms of spatial masking based on target/mask similarity in the unlocalized Fourier amplitude domain (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). Such a rigorous test of the conceptual masking hypothesis is important, given that the noise masks in previous conceptual masking studies (Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988; Potter, 1976; Potter & Levy, 1969) shared few if any spatial characteristics with the scene target images (though see Bachmann, Luiga, & Poder, 2005; Intraub, 1984, Exp 3, Inverted Condition). Consistent with predictions based on previous spatial masking research, the results of Experiment 3 showed greater masking by phase-randomized scene images having 1/f amplitude spectra than by white noise, which has a flat amplitude spectrum. However, the 1/f amplitude spectrum only affected gist masking at early stages of perceptual processing (SOAs $\leq$ 50 ms), as would be predicted by conceptual masking theory (Loftus & Ginn, 1984; Loftus, Hanna, & Lester, 1988). On the other hand, inconsistent with previous work on the time course of conceptual masking (Loftus & Ginn, 1984), when we

carefully controlled masking strength, recognizable normal scene images produced greater masking than unrecognizable images, which is the hallmark of conceptual masking, from even the earliest stages of target processing (i.e., the shortest SOA of 12 ms).

It is important to point out that an alternative version of the spatial masking hypothesis may still explain why normal scene images are more efficient than phase-randomized images at masking gist. Normal images may be better scene gist masks because they contain spatially localized higher order structure critical for scene gist recognition. In fact, further studies in our laboratory using noise that has been coerced to share the wavelet-based texture statistics of scenes indicates that such noise more efficiently masks scene gist than the fully phase-randomized scene masks of the current study (Loschky et al., 2006). In order to make a claim for conceptual masking, it is critical to eliminate such alternative, simpler, low-level masking explanations.

Finally, the current study is an important first step towards providing a principled basis for choosing spatial and temporal mask parameters for use in studies of scene perception. We have shown that, after controlling for the spatial parameters of target and mask amplitude spectra, mean luminance, and RMS contrast, and the temporal parameters of mask:target duration ratio, and SOA, a normal scene is a more efficient at masking gist than a fully phase-randomized version of that scene. However, we have also shown that, given the same controls, noise having a 1/f amplitude spectrum is more efficient than white noise at masking gist. More generally, we have shown that masking can be used to elucidate the types of information used to recognize scene gist. These findings provide important information for vision scientists studying both the information underlying scene gist and the time course of scene perception, because masking is necessary for studying the effects of stimulus duration on scene processing.

References

Bachmann, T., Luiga, I., & Poder, 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*(2), 139-153.

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.

Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of the Optical Society of America, A, Optics, Image Science and Vision., 19*(6), 1096-1106.

Biederman, I., Rabinowitz, J., Glass, A., & Stacy, E. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*, 597-600.

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-Verlag.

Breitmeyer, B. G., & Ogmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision*. Oxford: Clarendon Press.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), Apr 1981, 1207-1230.

Carter, B. E., & Henning, G. B. (1971). The detection of gratings in narrow-band visual noise. *Journal of Physiology, 219*(2), 355–365.

Chandler, D. M., & Hemami, S. S. (2003). Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions. *Journal Of The Optical Society Of America A-Optics Image Science And Vision, 20*(7), 1164-1180.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 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*(1), 11-18.

Delord, S. (1998). Which mask is the most efficient: A pattern or a noise? It depends on the task. *Visual Cognition, 5*(3), 313-338.

Di Lollo, V., von Muhlenen, A., Enns, J. T., & Bridgeman, B. (2004). Decoupling stimulus duration from brightness in metacontrast masking: Data and models. *Journal Of Experimental Psychology-Human Perception And Performance, 30*(4), 733-745.

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*(12), 2379-2394.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation, 6*(4), 559-601.

Field, D. J. (1999). Wavelets, vision and the statistics of natural scenes. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 357*(1760), 2527-2542.

Gorkani, M. M., & Picard, R. W. (1994). *Texture orientation for sorting photos "at a glance"*.

Guyader, N., Chauvin, A., Peyrin, C., Hérault, J., & Marendaz, C. (2004). Image phase or amplitude? *Comptes Rendus Biologies, 327*, 313-318.

Harvey, L. O., Roberts, J. O., & Gervais, M. J. (1983). The spatial frequency basis of internal representation. In H. G. Geissler, H. F. J. M. Buffart, E. L. J. Leeuvenberg & V. Sarris (Eds.), *Modern issues in perception* (pp. 217–226). Berlin: VEB Deutscher Verlag der Wissenschaften.

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 A, 71*, 574-581.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Joumal of Experimental Psychology: General, 127*(4), 398-415.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7*(3), 604-610.

Intraub, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 115-125.

Kahneman, D. (1967). An onset-onset law for one case of apparent motion and metacontrast. *Perception and Psychophysics, 2*(12-A), 577-584.

Langer, M. S. (2000). Large-scale failures of f(-alpha) scaling in natural image spectra. *Journal of the Optical Society of America A-Optics & Image Science, 17*(1), 28-33.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America, 70*(12), 1458-1471.

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

Loftus, G. R., & Ginn, M. (1984). Perceptual and conceptual masking of pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(3), 435-441.

Loftus, G. R., Hanna, A. M., & Lester, L. (1988). Conceptual masking: How one picture captures attention from another picture. *Cognitive Psychology, 20*(2), 237-282.

Loftus, G. R., & Mclean, J. E. (1999). A front end to a theory of picture recognition. *Psychonomic Bulletin and Review, 6*(3), 394-411.

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*(2), 250-260.

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Forristal, N., Corbeille, J., et al. (2006). The roles of amplitude and phase information in scene gist recognition and masking [Abstract]. *Journal of Vision, 6*(6), 802a.

Loschky, L. C., & Simons, D. J. (2004). The effects of spatial frequency content and color on scene gist perception [Abstract]. *Journal of Vision, 4*(8), 881a.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. (2005). The use of visual information in natural scenes. *Visual Cognition, 12*(6), 938-953.

Michaels, C. F., & Turvey, M. T. (1979). Central sources of visual masking: Indexing structures supporting seeing at a single, brief glance. *Psychological Research-Psychologische Forschung, 41*(1), 1-61.

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 Academic Press.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72-107.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 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*(3), 145-175.

Oliva, A., & Torralba, A. (in press). Building the gist of a scene: The role of global image features in recognition. In (Vol. 155).

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.

Oliva, A., Torralba, A., Guerin-Dugue, A., & Herault, J. (1999). Global semantic classification using power spectrum templates. *In Proceedings of The Challenge of Image Retrieval, Electronic Workshops in Computing series, Springer-Verlag: Newcastle.*

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607-609.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. *IEEE, Proceedings, 69*, 529-541.

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

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 And Cognition, 15*(4), 587-595.

Piotrowski, L. N., & Campbell, F. W. (1982). A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception, 11*(3), 337-346.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning & Memory, 2*(5), 509-522.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*(1), 10-15.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44*, 2301–2311.

Rieger, J. W., Braun, C., Bulthoff, H. H., & Gegenfurtner, K. R. (2005). The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *Journal of Vision, 5*(3), 275-286.

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*(6), 852-877.

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

Sadr, J., & Sinha, P. (2004). Object recognition and Random Image Structure Evolution. *Cognitive Science, 28*(2), 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*(5), 374-378.

Schyns, P., & Oliva, A. (1993). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*, 195-200.

Sekuler, R. W. (1965). Spatial and temporal determinants of visual backward masking. *Journal of Experimental Psychology, 70*(4), 401-406.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience, 24*(1), 1193-1216.

Solomon, J. A. (2000). Channel selection with non-white-noise masks. *Journal of the Optical Society of America A, 17*(6), 986-993.

Sperling, G. (1963). A model for visual memory tasks. *Human Factors, 5*, 19-31.

Stromeyer, C. F., & Julesz, B. (1972). Spatial-frequency masking in vision: Critical bands and spread of masking. *Journal of the Optical Society of America A, 62*(10), 1221-1232.

Tadmor, Y., & Tolhurst, D. J. (1993). Both the phase and the amplitude spectrum may determine the appearance of natural images. *Vision Research, 33*(1), 141-145.

Thomson, M. G. A., & Foster, D. H. (1997). Role of second- and third-order statistics in the discriminability of natural images. *Journal of the Optical Society of America, 14*(9), 2081-2090.

Turvey, M. T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review, 80*(1), 1-52.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15*(1), 121-149.

Vailaya, A., Jain, A., & Zhang, H. J. (1998). On image classification: City images vs. landscapes. *Pattern Recognition, 31*(12), 1921-1935.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. (2006). Phase noise and the classification of natural images. *Vision Research, 46*, 1520-1529.

Wilson, H. R., McFarlane, D. K., & Phillips, G. C. (1983). Spatial frequency tuning of orientation selective units estimated by oblique masking. *Vision Research, 23*(9), 873-882.

Yang, J., & Stevenson, S. B. (1998). Effect of background components on spatial-frequency masking. *Journal of the Optical Society of America, 15*(5), 1027-1035.

Author Note

Lester Loschky, Department of Psychology, Kansas State University; Amit Sethi, Department of Computer and Electrical Engineering, University of Illinois at Urbana-Champaign; Daniel J. Simons, Department of Psychology, University of Illinois at Urbana-Champaign; Tejaswi N. Pydimarri, Department of Computer and Information Science, Kansas State University; Daniel Ochs, Department of Psychology, Kansas State University; Jeremy L. Corbeille, Department of Psychology, Kansas State University.

This study contains some information that has been presented at the Annual Meeting of the Psychonomic Society (2005) and the Annual Meeting of the Vision Sciences Society (2006), with the abstract of the latter 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 wish to 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 experiments. The authors also wish to thank David Field, Javid Sadr, and Jian Yang for their helpful discussions of the paper.

Correspondence concerning this article should be addressed to Lester Loschky, Department of Psychology, Kansas State University, 471 Bluemont Hall, Manhattan, Kansas 66506-5302. Electronic mail may be sent via Internet to loschky@ksu.edu.

Appendix A

Basic Concepts of the Fourier Transform Applied to Digital Images

A 2-D Fourier transform *F(u,v)* expresses a 2-D function *f(x,y)* as a weighted linear combination of spatially shifted 2-D sinusoidal basis functions $e^{2i\pi(ux+vy)}$ as shown in equation 1.

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{2i\pi(ux+vy)}dudv \tag{1}$$

The Fourier transform is calculated using equation 2.

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-2i\pi(ux+vy)}dxdy \tag{2}$$

A digital image of size *MxN* pixels is a discrete signal that can be expressed as a 2-D array *f(x,y)*, where *x* is an integer from 0 to *M-1*, and *y* is an integer from 0 to *N-1*. It can be expressed as a linear combination of *MN* Fourier basis functions $e^{2i\pi(ux/M+vy/N)}$ as shown by equation 3.

$$f(x,y) = \sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)e^{2i\pi(ux/M+vy/N)} \tag{3}$$

The 2-D array of multiplicative weights *F(u,v)* is called the Fourier coefficients or the discrete Fourier transform (DFT) of the image *f(x,y)*. In general, the image pixel values *f(x,y)* are real non-negative numbers, whereas the Fourier coefficients *F(u,v)* are complex numbers. The DFT array has the same size as the image array. The DFT is calculated according to the equation 4.

$$F(u,v) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-2i\pi(ux/M+vy/N)} \tag{4}$$

The Fast Fourier Transform (FFT) is an algorithm to efficiently compute the DFT, and often times DFT and FFT are used synonymously. Depending on the context, the term "FFT" is used to describe the FFT algorithm or the output of the algorithm (the FFT array).

A Fourier coefficient, which is a complex number, can be represented as a sum of a real and an imaginary number. Alternatively, it can be represented in the polar form as a product of a real and non-negative number (called the magnitude) and a complex number with unit magnitude. The angle of the complex number with unit magnitude to the positive real axis on the unit circle in the complex number plane represents the phase of the complex number. These relations are represented in the equation 5.

$$F(u,v) = a_{u,v} + ib_{u,v} = r_{u,v}e^{i\theta_{u,v}} \quad (5)$$

In equation 5, $r_{u,v}$ represents the magnitude (also known as the amplitude) corresponding to the frequency *(u,v)* cycles per spatial dimension, and $\theta_{u,v}$ represents the phase of the basis function of that frequency $e^{2i\pi(ux/M+vy/N)}$. The phase determines the shift (spatial offset) of the sinusoidal pattern represented by the basis function $e^{2i\pi(ux/M+vy/N)}$. Thus, the important information about the image structure and location of patterns is embedded in the phase of the FFT. Magnitude, on the other hand, stores the energy or gradient information of patterns of various spatial frequencies. The original image can be recovered from the FFT array by taking the inverse FFT (IFFT) of the FFT array. The IFFT algorithm is quite similar to the FFT algorithm with some minor differences as represented by the equations 3 and 4.

Implementing the RISE Algorithm

To control the energy distribution in the spatial frequencies of an image, the magnitude of the Fourier transform should be preserved. The image appearance (spatial structure) can still be changed by altering the phase information. This is the basic idea behind RISE, which stands for Random Image Structure Evolution (Sadr & Sinha, 2001; Sadr & Sinha, 2004). To progressively degrade the image structure, the phase at every spatial frequency location *(u,v)* is linearly and progressively interpolated between its original value and a target value. The extent of

interpolation towards the target value is controlled by a parameter α (which is a number between *0* and *1*) common for all the locations. The parameter α determines the extent of damage to the original image structure (*0* representing the unaltered image). For each location *(u,v)*, a target phase $\Phi_{u,v}$ is chosen as a random value between $-\pi$ and π. Half of the locations *(u,v)* are chosen at random for further alteration to their target values. The target phase $\Phi_{u,v}$ of such a chosen location is compared to the original phase $\theta_{u,v}$ (which is also (reduced to) a number between $-\pi$ and π). If the two have the same sign, then no further action is taken. Otherwise, 2π is added to $\Phi_{u,v}$ if $\theta_{u,v}$ is positive and -2π is added to $\Phi_{u,v}$ if $\theta_{u,v}$ is negative. This is done to ensure that at least half of the interpolated phases will not cross zero during interpolation. Having too many phases close to zero tends to produce white corners in the IFFT image (which will be the RISE image in this case). The altered Fourier coefficients $F_R(u,v)$ are expressed in terms of the original magnitude $r_{u,v}$, original phase $\theta_{u,v}$, target phase $\Phi_{u,v}$ and the interpolation factor α by equation 6.

$$F_R(u,v) = r_{u,v} e^{i((1-\alpha)\theta_{u,v} + \alpha\phi_{u,v})} \tag{6}$$

When we start with an original image of real numbers *f(x,y)*, we are guaranteed that its FFT will satisfy the property that any given Fourier coefficient will be the complex conjugate of the coefficient at the diametrically opposite location in the 2-D FFT array (considering the zero/DC frequency as the center, and wrap around at the boundary of the array). This property is expressed in equation 7, where the complex conjugate of a complex number z is expressed by z^*.

$$F(u,v) = F^*(M-u, N-v) \tag{7}$$

Taking the complex conjugate is same as preserving the magnitude and taking the negative of the phase. This means that the equations 8 and 9 must hold.

$$r_{u,v} = r_{M-u,N-v} \tag{8}$$

$$\theta_{u,v} = -\theta_{M-u,N-v} \tag{9}$$

However, the inverse also holds true. This means that we need to guarantee that the RISE image $f_R(x,y)$ is also an array of real numbers by enforcing a constraint similar to equation 7 on the modified FFT $F_R(u,v)$. Since the $F(u,v)$ and $F_R(u,v)$ share the same magnitude (see equations 5 and 6) this constraint can be satisfied by ensuring that equation 10 holds for all (u,v).

$$((1-\alpha)\theta_{u,v} + \alpha\phi_{u,v}) = -((1-\alpha)\theta_{M-u,N-v} + \alpha\phi_{M-u,N-v}) \tag{10}$$

Using equation 9, equation 10 can be satisfied if equation 11 holds for target phase $\Phi_{u,v}$.

$$\phi_{u,v} = -\phi_{M-u,N-v} \tag{11}$$

Equation 11 represents the mathematical constraints needed to ensure that the RISE image $f_R(x,y)$ (which is the IFFT of $F_R(u,v)$) will be an array of real numbers. These constraints are enforced in our algorithm. Equations 9 and 11 also constrain the phase at half the maximum frequencies to be zero, when (u,v) is $(0,0)$, $(0,N/2)$, $(M/2,0)$, or $(M/2, N/2)$, when M and N are multiples of two, and they usually are.

Finally, after computing the RISE image, some of the pixel values can have very small imaginary parts due to the limit of numerical precision associated with the computing setup. This residual imaginary part is discarded, and only the real part is kept. In addition, some of the real parts of the pixel values may be negative or may be outside the display range of the image system. All image pixel values are linearly scaled and shifted by a common amount to fit the display limits. Common display limits are *0* to *1* and *0* to *255*, and are often quantized. For example, pixel values can be integers between *0* and *255* on most systems. Such linear scales and shifts can also be tailored to match the average pixel intensity or the RMS contrast (but not necessarily both) of the original and the RISE images. However, if the original image itself is also linearly scaled and shifted then both the average pixel intensity and RMS contrast can be

matched. Such an algorithm is described in Appendix B.

Appendix B

Equalizing Mean Luminance and RMS Contrast of All Images in a Set

Let the coordinates of a pixel be represented by the ordered pair (x, y), where *x* and *y* are integers ranging from 1 to the number of columns (M) and rows (N) respectively in the image. Let the pixel intensity at a location (x, y) be represented by $I(x, y)$. Let the mean intensity be represented by $\hat{I}$, and RMS contrast be represented by $\dot{I}$. The mean intensity and RMS contrast can be calculated as shown in equations 12 and 13.

$$\hat{I} = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N} I(x, y)}{MN} \tag{12}$$

$$\dot{I} = \sqrt{\frac{\sum_{x=1}^{M}\sum_{y=1}^{N} (I(x, y) - \hat{I})^2}{MN}} \tag{13}$$

Let us suppose that we have a set of *n* images I_i, where *i* ranges from 1 through *n*. We want to find a linear operator (scale and shift) for each image so that when these operators are applied to the pixel intensities of their associated images, all the resultant images have the same mean and RMS contrast after applying their respective operators. A simple two-step method for this is to make these images zero mean and unit contrast by applying an appropriate linear operator to each image, followed by applying a common linear operator to all the resultant zero mean and unit contrast images such that the resultant images of this second step occupy the entire range of displayable image intensities (or a sub-range thereof). The first step can be done as follows. Let I_i' represent the resultant images of the first step that have zero mean and unit

RMS contrast. The linear operator for the image intensities of I_i, that results in I_i' is represented as shown in equation 14.

$$I_i' = \frac{I_i - \hat{I}_i}{\dot{I}_i} \tag{14}$$

Now, the maximum and the minimum pixel values in all the images, $I_{\max}$ and $I_{\min}$, can be represented as shown in equations 15 and 16.

$$I_{\max} = \max_{i,x,y} I_i'(x, y) \tag{15}$$

$$I_{\min} = \min_{i,x,y} I_i'(x, y) \tag{16}$$

Note that $I_{\min}$ will be negative because the images I_i's are zero mean. Now, we want to scale and shift these images so that $I_{\max}$ and $I_{\min}$ respectively map to 0 and 255 (assuming that to be the displayable range) in the final images represented by I_i''. This can easily be done as shown in equation 17

$$I_i'' = (I_i' - I_{\min}) \times \frac{255}{I_{\max} - I_{\min}} \tag{17}$$

This can suitably be quantized for the display system (such as an integer in most cases). As is obvious from the equation 17, the new mean intensity $\hat{I}''$ and RMS contrast $\dot{I}''$ for all the images I_i'' is the same since the same linear operator is applied to all the zero mean and unit contrast images resulting from step 1. This new mean and contrast can be represented as shown in equations 18 and 19.

$$\hat{I}'' = \frac{-255 I_{\min}}{I_{\max} - I_{\min}} \tag{18}$$

$$\dot{I}'' = \frac{255}{I_{\max} - I_{\min}} \tag{19}$$

And, since the algorithm consists of applying two linear operators to every image, the entire transformation from I_i to I_i'' is a linear operation.

This completes the outline of our algorithm for equalizing the mean luminance and RMS contrast of all images in a set.

Figure Caption

Figure 1. Top row: Example images in the normal condition (RAND = 0). The images differ in terms of their dominant spatial frequencies, orientations, and mean luminance levels. Middle row: Fully randomized phase versions (RAND = 1.0) of the example images. Note that the images have maintained their differences in terms of spatial frequencies, orientations, and mean luminance levels. For example, Beach 15 is dominated by low frequencies at an oblique orientation, Street 4 has more high frequencies with a dominant vertical orientation, and Mountain 18 has low to medium spatial frequencies at all orientations. Bottom row: Fast Fourier Transform (FFT) spatial frequency amplitude images for each of the scenes. (The FFT spatial frequency amplitude images for the original (top row) and fully randomized (middle row) versions of each scene are identical, thus only one is shown for each.) FFT spatial frequency amplitude images represent energy (contrast) by brightness, spatial frequency by distance from the center of the graph, and orientation by radial orientation on the unit circle, with a 0 at the 12 o'clock position (i.e., scene-based orientation coordinates are shifted 90º clockwise). Energy has been multiplied by 20 to enhance visibility of higher frequencies. The above-noted differences between the three scenes in frequencies and orientations are evident.

Figure 2. Example image with the six levels of phase randomization (ranging from 0-1.0) used in the study. "RAND = 0" represents a phase randomization factor of 0 (a normal image); "RAND = 1.0" represents a phase randomization factor of 1.0 (completely randomized).

Figure 3. Schematics of the events in a trial in Experiments 1 (left), 2 (middle), and 3 (right). Note that in Experiments 2 and 3, the mask type varied in terms of level of phase randomization,

and in Experiment 3 included white noise. ISI = interstimulus interval; SOA = stimulus onset asynchrony.

Figure 4. Scene identification accuracy as a function of phase randomization factor (RAND = 0-1.0) and stimulus duration of mask and target (ms). "RAND = 0" represents a phase randomization factor of 0 (a normal image); "RAND = 1.0" represents a phase randomization factor of 1.0 (completely randomized).

Figure 5. Scene identification accuracy as a function of masking scene phase randomization factor (RAND = 0-1.0) and stimulus duration of mask and target (ms). "RAND = 0" represents a phase randomization factor of 0 (a normal image); "RAND = 1.0" represents a phase randomization factor of 1.0 (completely randomized).

Figure 6. Masking effect on scene identification accuracy as a function of phase randomization factor (RAND = 0-1.0) and stimulus duration of mask and target (ms). "RAND = 0" represents a phase randomization factor of 0 (a normal image); "RAND = 1.0" represents a phase randomization factor of 1.0 (completely randomized).

Figure 7. Spatial frequency amplitude spectra of the masking conditions used in the study: normal image (RAND = 0), completely phase-randomized (RAND = 1.0), and white noise. Spatial frequency amplitude values are averaged across all orientations and across all 300 images in each condition.

Figure 8. Relationship between target and mask images in Experiment 3. Examples of the four types of masks are shown in the right column. RISE phase randomized images were fully randomized (RAND = 1.0).

Figure 9. Scene identification accuracy as a function of mask type and stimulus onset asynchrony (SOA)(ms). "RAND = 0" represents a phase randomization factor of 0 (a normal image); "RAND = 1.0" represents a phase randomization factor of 1.0 (completely randomized); "Different Category" represents a masking scene from a different scene category than the target scene; "Same Image" represents a masking scene which is a fully randomized phase version of the target scene. Error bars represent 95% confidence intervals for each mean.

Figure 10. Scene identification accuracy as a function of mask type, mask:target duration ratio, and stimulus onset asynchrony (SOA)(ms). "RAND = 0" represents a phase randomization factor of 0 (a normal image); "RAND = 1.0" represents a phase randomization factor of 1.0 (completely randomized); "No-Mask" represents an unmasked control condition; "Mask:Target = 4:1" represents a mask to target duration ratio of 4 to 1 (48:12 ms). Error bars represent 95% confidence intervals for each mean.

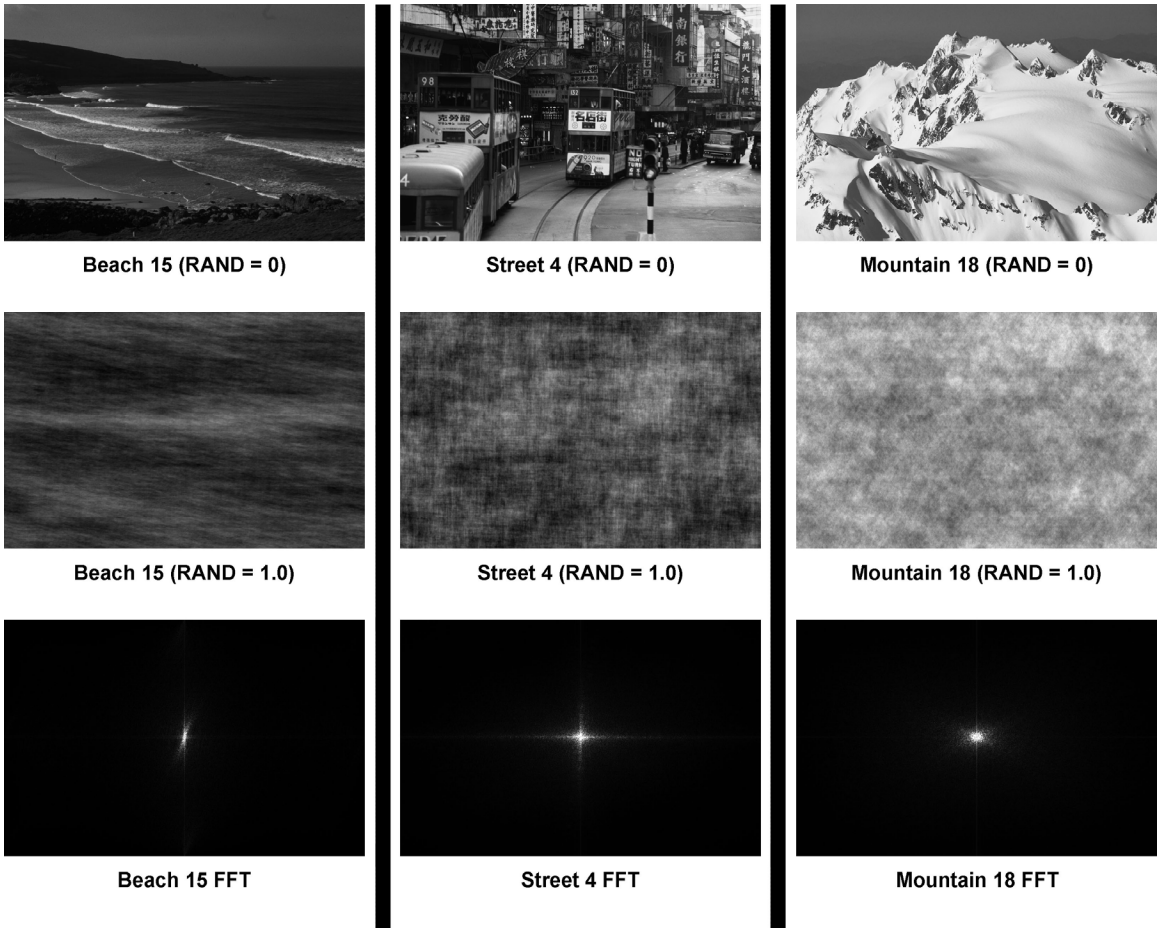
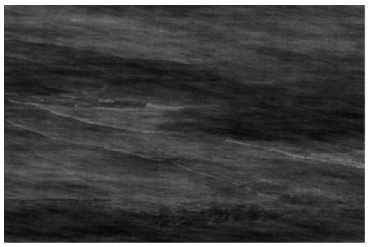

Beach 15 (RAND = 0)
Street 4 (RAND = 0)
Mountain 18 (RAND = 0)
Beach 15 (RAND = 1.0)
Street 4 (RAND = 1.0)
Mountain 18 (RAND = 1.0)
Beach 15 FFT
Street 4 FFT
Mountain 18 FFT

Figure 1

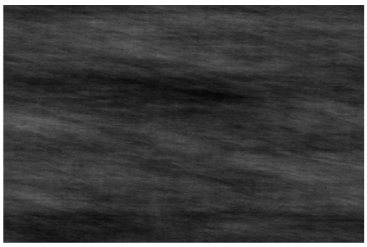
Beach 15 (RAND = 0)

Beach 15 (RAND = 0.1)

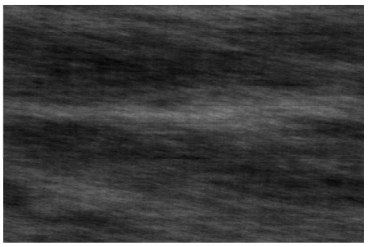
Beach 15 (RAND = 0.25)



Beach 15 (RAND = 0.4)



Beach 15 (RAND = 0.6)



Beach 15 (RAND = 1.0)

Figure 2

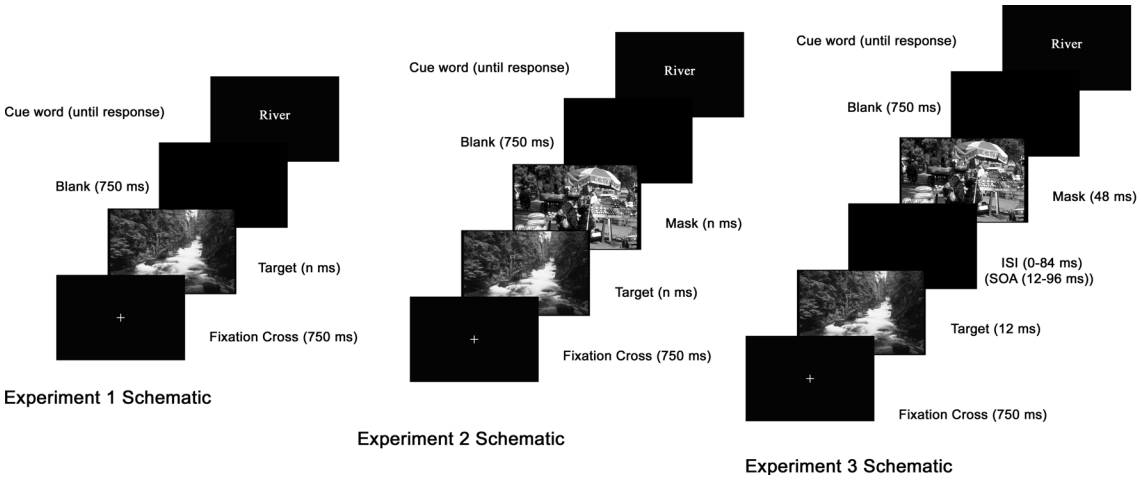

Cue word (until response)
River
Blank (750 ms)
Target (n ms)
Fixation Cross (750 ms)
Experiment 1 Schematic
Cue word (until response)
River
Blank (750 ms)
Mask (n ms)
Target (n ms)
Fixation Cross (750 ms)
Experiment 2 Schematic
Cue word (until response)
River
Blank (750 ms)
Mask (48 ms)
ISI (0-84 ms)
(SOA (12-96 ms))
Target (12 ms)
Fixation Cross (750 ms)
Experiment 3 Schematic

Figure 3

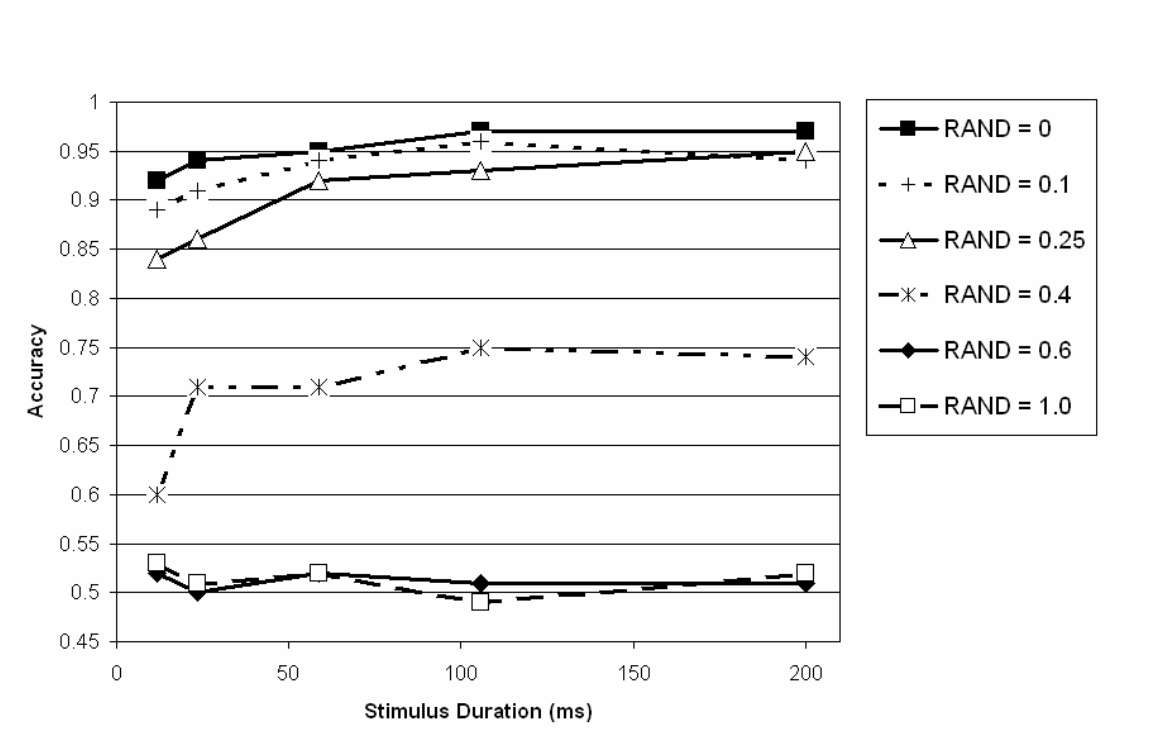

Accuracy
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0
50
100
150
200
Stimulus Duration (ms)
RAND = 0
RAND = 0.1
RAND = 0.25
RAND = 0.4
RAND = 0.6
RAND = 1.0

Figure 4

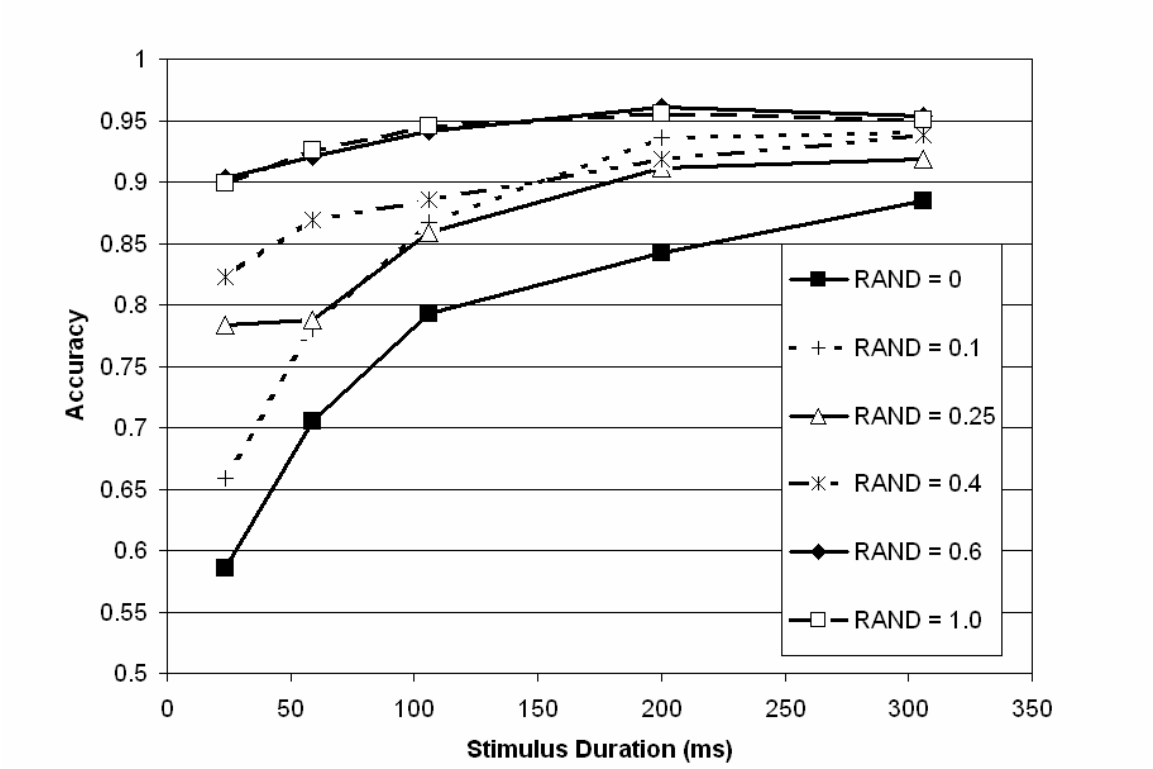

1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
50
100
150
200
250
300
350
Stimulus Duration (ms)
RAND = 0
RAND = 0.1
RAND = 0.25
RAND = 0.4
RAND = 0.6
RAND = 1.0

Figure 5

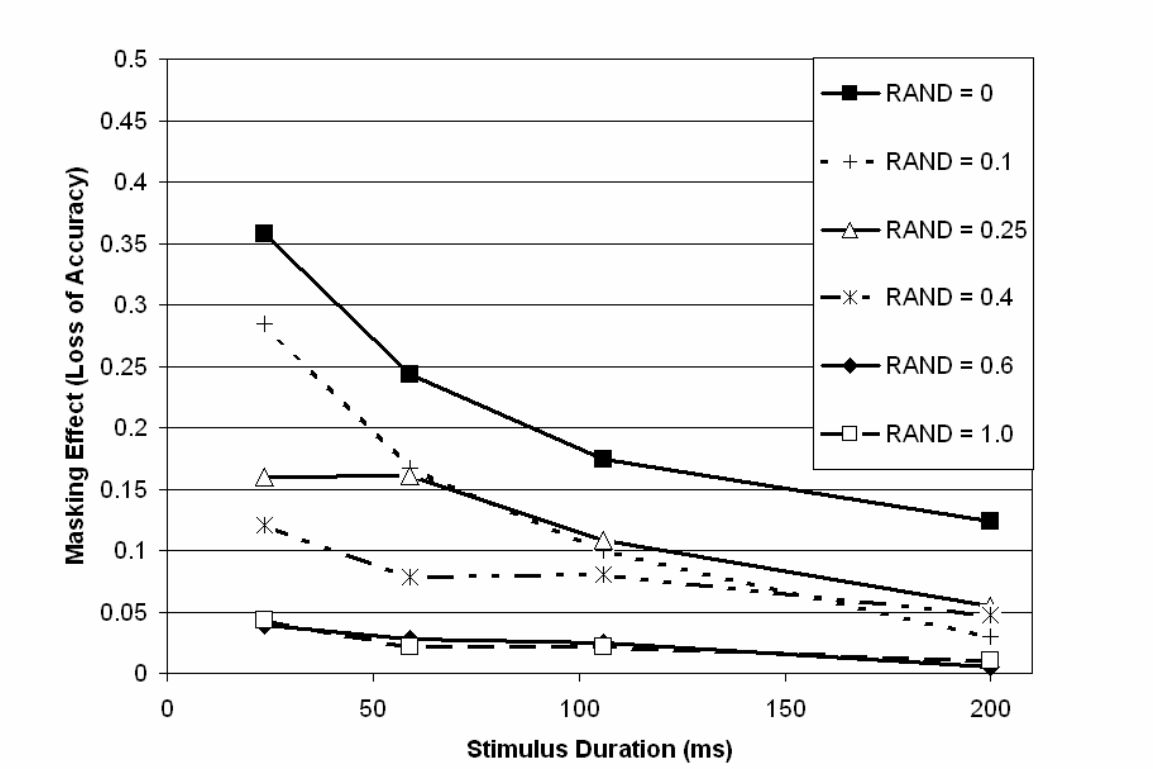

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Masking Effect (Loss of Accuracy)
0
50
100
150
200
Stimulus Duration (ms)
RAND = 0
RAND = 0.1
RAND = 0.25
RAND = 0.4
RAND = 0.6
RAND = 1.0

Figure 6

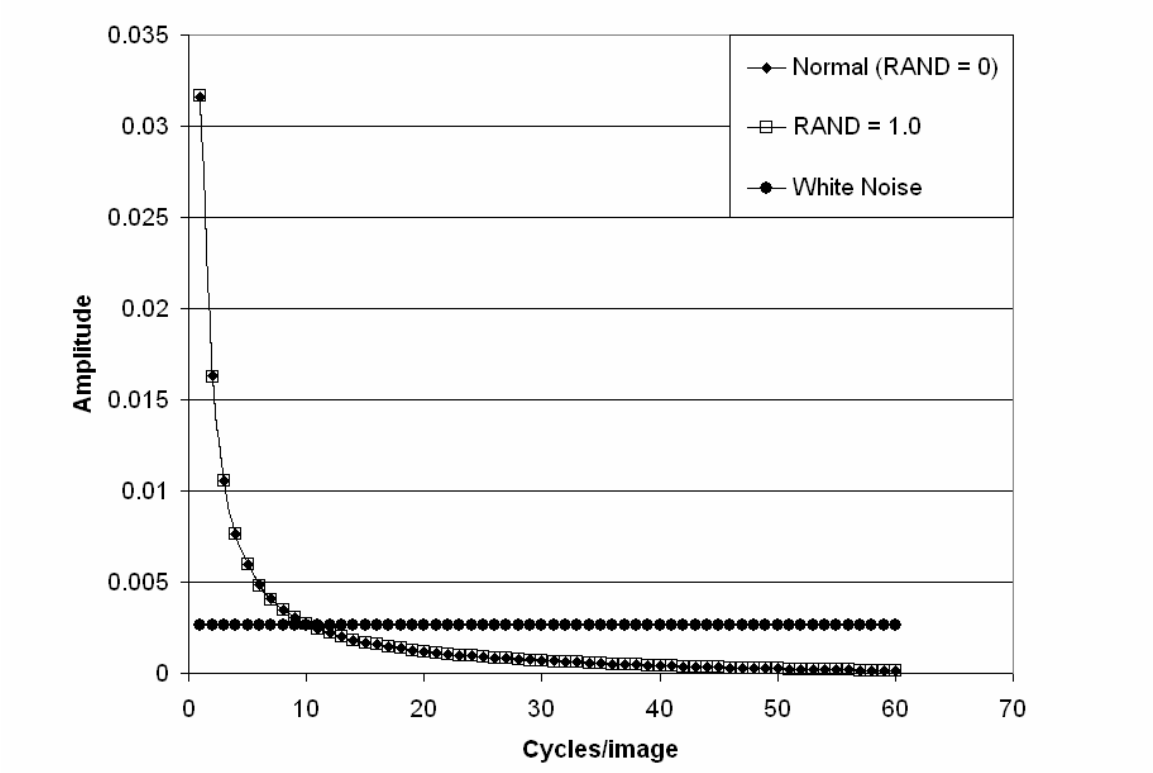

0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Amplitude
0
10
20
30
40
50
60
70
Cycles/image
Normal (RAND = 0)
RAND = 1.0
White Noise

Figure 7

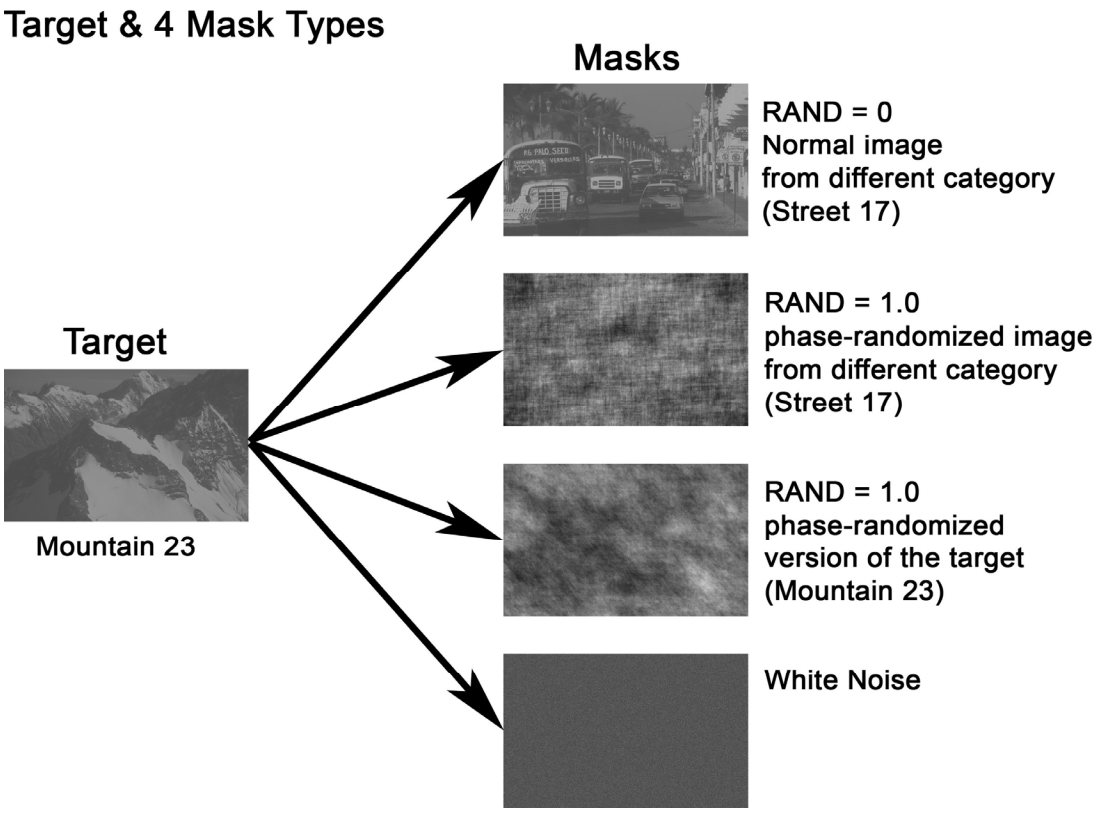

Target & 4 Mask Types
Masks
Target
Mountain 23
RAND = 0
Normal image
from different category
(Street 17)
RAND = 1.0
phase-randomized image
from different category
(Street 17)
RAND = 1.0
phase-randomized
version of the target
(Mountain 23)
White Noise

Figure 8

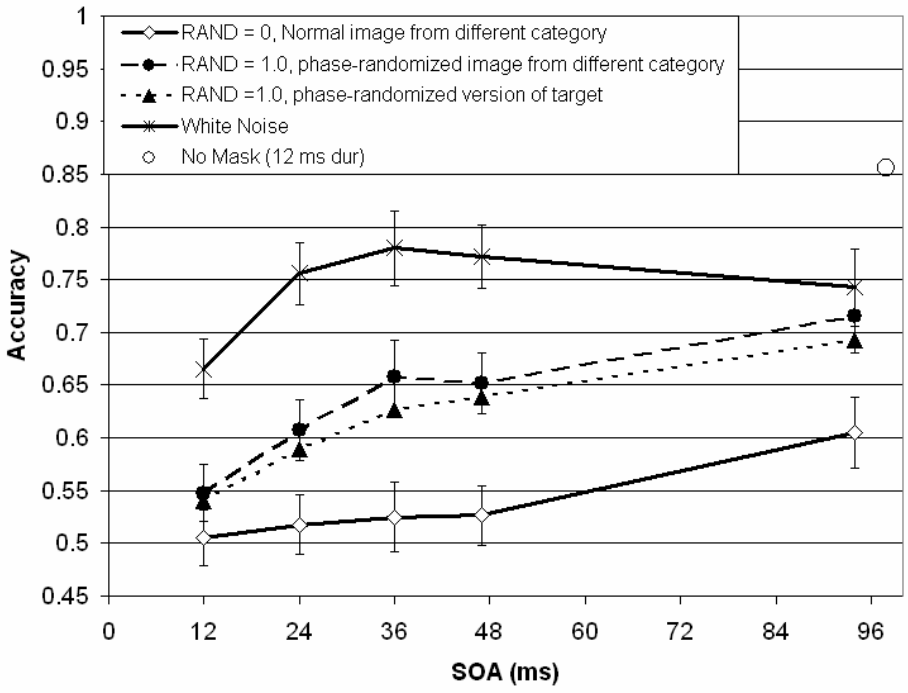

RAND = 0, Normal image from different category
RAND = 1.0, phase-randomized image from different category
RAND =1.0, phase-randomized version of target
White Noise
No Mask (12 ms dur)
Accuracy
SOA (ms)
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0
12
24
36
48
60
72
84
96

Figure 9

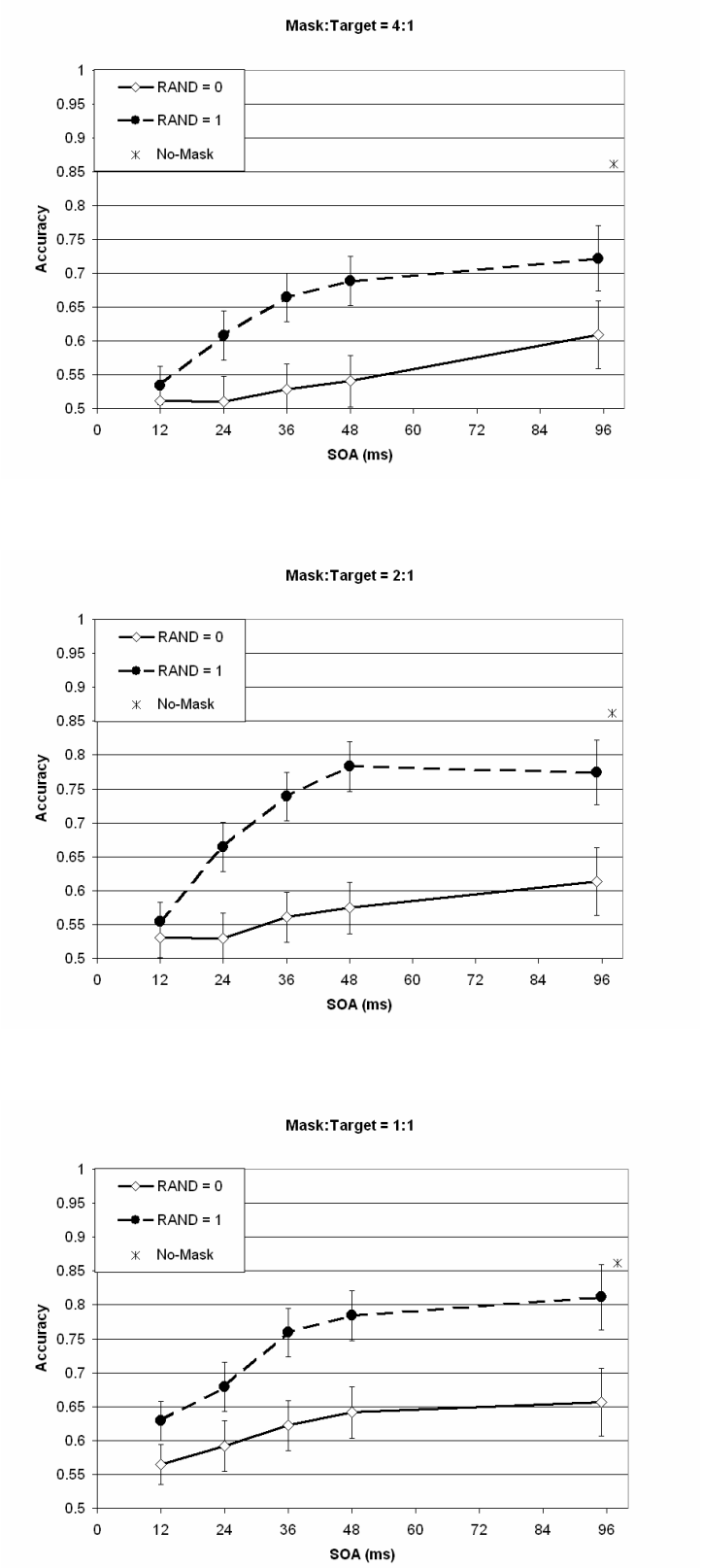

Mask:Target = 4:1
RAND = 0
RAND = 1
No-Mask
Accuracy
SOA (ms)
Mask:Target = 2:1
RAND = 0
RAND = 1
No-Mask
Accuracy
SOA (ms)
Mask:Target = 1:1
RAND = 0
RAND = 1
No-Mask
Accuracy
SOA (ms)

Figure 10