Global ensemble texture representations are critical to rapid scene perception

Timothy F. Brady [1] Anna Shafer-Skelton [1] George A. Alvarez [2]

[1] Department of Psychology
Univ. of California, San Diego
9500 Gilman Dr #0109
La Jolla, CA, 92093

[2] Department of Psychology
Harvard University
33 Kirkland St
Cambridge, MA 02138

Please address correspondence to:

Timothy F. Brady
Department of Psychology
Univ. of California, San Diego
9500 Gilman Dr #0109
La Jolla, CA, 92093
Email : timbrady@ucsd.edu

Abstract

Traditionally, recognizing the objects within a scene has been treated as a prerequisite to recognizing the scene itself. However, research now suggests that the ability to rapidly recognize visual scenes could be supported by global properties of the scene itself rather than the objects within the scene. Here, we argue for a particular instantiation of this view: that scenes are recognized by treating them as a global texture and processing the pattern of orientations and spatial frequencies across different areas of the scene without recognizing any objects. To test this model, we asked whether there is a link between how proficient individuals are at rapid scene perception and how proficiently they represent simple spatial patterns of orientation information (global ensemble texture). We find a significant and selective correlation between these tasks, suggesting a link between scene perception and spatial ensemble tasks but not non-spatial summary statistics In a second and third experiment, we additionally show that global ensemble texture information is not only associated with scene recognition, but that preserving only global ensemble texture information from scenes is sufficient to support rapid scene perception; however, preserving the same information is not sufficient for object recognition. Thus, global ensemble texture alone is sufficient to allow activation of scene representations but not object representations. Together, these results provide evidence for a view of scene recognition based on global ensemble texture rather than a view based purely on objects or on non-spatially localized global properties.

Keywords: ensemble perception; statistical summary perception; scene recognition; navigation; visual texture.

Statement of Public Significance

People can recognize visual scenes rapidly and accurately, determining the meaning of complex visual scenes in less than 1/10th of a second. Intuitively, we might expect such rapid scene recognition to proceed from the bottom up: first we recognize objects, then the configuration of these objects and then the entire scene. However, object recognition is not necessary for accurate scene recognition, and people can rapidly recognize scenes even when they cannot recognize any individual objects. Here, we provide evidence that one way the visual system performs this rapid non-object-based scene recognition is by treating scenes as “textures” and looking at the distribution of orientations and spatial frequencies across the entire scene at once.

People can recognize visual scenes rapidly and accurately, determining the meaning of a complex scene in less than 100ms (Intraub, 1981; Potter & Faulconer, 1975; Thorpe, Fize, & Marlot, 1996). Intuitively, we might expect such rapid scene recognition to proceed from the bottom up: first we recognize edges, then object parts, then entire objects, and then we eventually recognize the configuration of these objects and then the entire scene. Indeed, classic models of vision have generally predicted such a structure for visual recognition and treated objects and their relations as the basic unit of visual scene recognition (e.g., Biederman, Mezzanotte, & Rabinowitz, 1982; Marr, 1982).

However, object recognition is not necessary for accurate scene recognition: people can rapidly recognize scenes even when they cannot recognize any individual objects (Oliva & Torralba, 2006; Schyns & Oliva, 1994). Furthermore, people can recognize global features of a scene before they can identify the image category (Greene & Oliva, 2009b), and these global properties, rather than the objects present in a scene, seem to drive the confusions people make between rapidly presented scenes (Greene & Oliva, 2009a). This suggests that a representation of scene layout, independent of objects, may play a major role in rapid scene recognition (Sanocki & Epstein, 1997; Sanocki, 2003). An important role for spatial layout, rather than objects, is also consistent with the neural evidence from regions of the brain that preferentially respond to scenes over individual objects, like the parahippocampal place area (PPA; Epstein & Kanwisher, 1998). These regions are sensitive to scene layout but considerably less sensitive to objects and other scene content (Epstein, 2005; Park, Brady, Greene, & Oliva, 2011).

How could people recognize the meaning and spatial layout of a scene rapidly without using objects? One possibility is that initial scene perception occurs by rapidly encoding patterns of orientation and spatial frequency across an image – effectively treating the scene as a holistic entity and examining spatial variations in its texture. Consistent with this proposal, computational models have shown that the information present in the pattern of orientation and spatial frequencies across an image is sufficient to categorize a scene and to determine some global properties of the scene, including its spatial layout (Oliva & Torralba, 2001, 2006; Renninger & Malik, 2004), and can explain the relative difficulty of different scene categorization tasks (Sofer, Crouzet, & Serre, 2015). For example, Oliva and Torralba (2006) show that preserving the spatial frequency and orientation distribution of an image, but pooling it across each quadrant of an image (e.g., in a 2x2 grid), is nevertheless sufficient to determine the natural or man-made-ness of an environment, as well as any 3D perspective in the image (Ross & Oliva, 2010). Preserving more spatial information (e.g., pooling separately in each cell of an 6x6 or 8x8 grid) additionally preserves the average depth of the scene as well as the degree of openness (e.g., the extent to which a horizon line is visible; Ross & Oliva, 2010). Thus, even a very simple texture representation of a scene – a grid of spatial frequency and orientation information – is computationally sufficient to recognize significant information about the spatial layout and 3D structure of a scene, even when little or no information about individuated objects is preserved. Even very limited information – e.g., only the amplitude spectrum of a scene, with no spatial information at all – can provide some information about the scene (e.g., the amount of vertical orientation can cue whether a scene is a city or a beach; Guyader, Chauvin, & Peyrin, 2004; see also Honey et al., 2008; Kaping, Tzvetanov, & Treue, 2007), although without spatial information, this seems to be limited and not sufficient to recognize the scene gist (Loschky et al., 2007). In addition, the amplitude spectrum alone cannot account for even the human ability to perform basic distinctions like natural vs. man-made, which can be performed rapidly and accurately even with image sets where the amplitude spectrum has been equated (Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009). Thus, spatial information being preserved is critical to recognizing scenes based on texture properties.

----- Figure 1 about here -----

Are people actually sensitive to patterns of orientation and spatial frequency information across an image? The literature on ‘spatial ensemble perception’ argues that people are able to compute spatial distributions of low-level features very quickly and efficiently, at least in simple displays. For example, people can efficiently compute the distribution of orientations in the top and bottom of a grid of gabor elements (Alvarez & Oliva, 2009), or the spatial distribution of simple color squares (Brady & Tenenbaum, 2013) and seem to store and use this information (e.g., Brady & Alvarez, 2015). People can also compute these spatial ensemble statistics when attention is diffusely spread (Alvarez & Oliva, 2009) and in their periphery (Balas, Nakano, & Rosenholtz, 2009), consistent with a role in scene recognition. These spatial ensemble patterns, while made up of simple elements like gabors, nevertheless closely mimic the patterns of orientated elements used in computer vision algorithms to holistically recognize scenes (e.g., Oliva & Torralba, 2006), raising the question of whether human sensitivity to these patterns in simple displays, like grids of gabors, arises because of their role in allowing for rapid recognition of the spatial structure of scenes (e.g., Figure 9 in Brady, Konkle, & Alvarez, 2011).

In addition to this spatial ensemble information, people are also sensitive to even simpler, non-spatial ensemble information, like the mean and variance of basic feature dimensions (often referred to as summary statistics). For example, participants can rapidly extract the mean size of a set of circles (Ariely, 2001; Chong & Treisman, 2003) or the average emotion of a set of faces (Haberman & Whitney, 2007). Whereas the representations required to perform spatial ensemble tasks must preserve spatial information (e.g., the top is mostly horizontal; bottom is mostly vertical; Alvarez & Oliva, 2009), summary statistics do not. While computation of summary statistics requires pooling information across space, it does not involve the recognition of spatial patterns, since all information must be pooled into a single representation of the average. Although it has been proposed that scene recognition relies on such summary statistic processing (e.g., Wolfe, Võ, Evans, & Greene, 2011, pg. 81), representing properties such as spatial layout requires the preservation of how information is distributed across space. Thus, it remains to be determined how related these non-spatial summary statistic representations are to scene recognition.

Here we examine the role of such summary statistics, spatial ensemble statistics, and similar global ensemble texture representations in visual scene recognition. In a first experiment, we use an individual differences design to show that the same participants who perform best on a spatial ensemble task also show the most activation of scene representations in brief displays. This suggests a link between spatial ensemble processing and rapid scene recognition. However, we find no relationship between non-spatial summary statistics and scene recognition. In a second experiment, we show that preserving only global-ensemble-texture information (in particular, a spatial distribution of orientations and spatial frequencies) in scenes is sufficient to allow participants to activate scene representations. In a third experiment, we show that this link between spatial ensembles and scenes is selective: preserving the same information in images of objects is insufficient to allow activation of object representations. Overall, our data provide evidence for the role of rapid global ensemble texture processing in rapid scene recognition, as well as suggesting the spatial ensemble tasks may tap into these same global ensemble texture processing mechanisms.

Experiment 1: Individual differences

In Experiment 1, we examine the relationship between rapid scene recognition, spatial ensemble perception, and summary statistics in simple displays using an individual differences approach. Specifically, we ask whether skill at spatial ensemble processing predicts individual participants' rapid scene recognition ability above and beyond general factors, like motivation, working memory capacity and non-spatial summary perception.

As a measure of spatial ensemble processing, we use a modified version of a task developed by Alvarez and Oliva (2009). Participants have to detect changes to a grid of high spatial frequency gabor elements while their attention is diffusely spread (so they cannot focus on the individual gabor elements). Sometimes nothing changes; sometimes all the individual elements rotate, but these changes do not change the global structure of the display; and sometimes all the individual elements rotate and these changes also affect the global structure/ensemble of the display (see Figure 2A). We ask whether participants who are particularly sensitive to the ensemble structure changes are the same participants who are best at rapid scene recognition.

As a measure of rapid scene recognition, we use the *object recognition* task of Davenport and Potter (2004). We ask participants to recognize objects, and these objects can appear on top of informative scene backgrounds (e.g., a priest in a church), or on top of uninformative scene backgrounds (e.g., a priest on a football field). The only difference between conditions is the scene backgrounds, and thus any benefit to object recognition from informative scenes must be driven by participant's rapid scene recognition ability. We chose this task rather than a direct measure of scene recognition because a task where naming scenes was directly relevant would need to use extremely brief presentations with strong dynamic masks (e.g., a single frame; Greene & Oliva, 2009b), and we found in pilot experiments that individual differences in scene recognition were swamped in such tasks by the vigilance and motivational factors that are prevalent in such tasks. Furthermore, the object recognition literature has shown robust effects of background scenes on object recognition (e.g., Biederman et al., 1982; Boyce, Pollatsek, & Rayner, 1989; Boyce & Pollatsek, 1992; Davenport & Potter, 2004; although see Hollingworth & Henderson, 1998, 1999), and scenes are known to rapidly influence objects in such object recognition tasks (e.g., Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008). Thus, the facilitation of object recognition by scenes can be usefully used as a measure of rapid scene processing.

Finally, we also measured participants' ability to compute non-spatial summary statistics: in particular, the average orientation of a set of Gabor elements. People are able to quickly and accurately report summary statistics across sets of objects: e.g., the average orientation of a set (Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001) or the average size of a set (Ariely, 2001; Chong & Treisman, 2003; see Alvarez, 2011 for a review). These tasks do not require the preservation of spatial information, and thus are distinct from spatial ensemble tasks as well as from the texture representations that have been used in computational models of scene perception (e.g., Oliva & Torralba, 2001, 2006). Because the task is, however, dependent on the global spread of attention and the processing of multiple gabor elements, it serves as a control condition for the spatial ensemble task—it allows us to disambiguate the role of spatial information and global ensemble texture patterns, which are present in the spatial ensemble task but not present in the summary statistic task, from the role of processing multiple gabor elements and spreading attention globally, which are present in both tasks. It also allows us to examine whether even such summary statistic tasks might be related to scene recognition, as has been claimed (e.g., Wolfe et al., 2011).

----- Figure 2 about here -----

Method

Participants

50 individuals (age range 18- 35) from the Cambridge, MA, and Harvard University community participated. All participants gave informed consent and had normal or corrected-to-normal vision. All individuals completed each of our 3 conditions to allow us to examine how performance on different tasks correlates across individuals. This enables us to ask whether these tasks could be supported by the same underlying mechanism or whether they must be supported by independently operating mechanisms (e.g., Vogel & Awh, 2008; Wilmer, 2008).

Spatial ensemble processing measure

Participants performed 200 trials of a change-detection task in which an 8x8 grid of gabor patches (50% contrast; ~2 cycles/deg; each subtending 1°by 1°) was briefly flashed (250ms) and then reappeared (300ms later). The patches were aligned so that the top of the screen consisted of nearly vertical items (±22.5° from vertical) and the bottom consisted of nearly horizontal items (±22.5° from horizontal), or the opposite pattern (vertical bottom, horizontal top); see Figure 2A. When the grid reappeared, 50% of the time all of the patches' orientations were identical. The other 50% of the time, they had all rotated by 45°. On 50% of change trials, these 45° rotations altered the global pattern of orientations in the display (local+ensemble changes; e.g., the top went from roughly vertical to horizontal and bottom from horizontal to vertical). The other half of the time, the global pattern remained the same despite each element rotating by 45° (local-only changes; e.g., the top remained roughly vertical and bottom remained roughly horizontal). The amount of local change to each gabor was identical in the local+ensemble change condition and the local-only change condition—the only difference between these two conditions is the presence of an ensemble change. To the extent that participants are sensitive to the ensemble structure, it should be easier to notice changes on local+ensemble trials than local-only trials (see Alvarez & Oliva, 2009 for a similar task and logic).

Each trial started with a distractor task that encouraged participants to spread their attention globally rather than focusing on particular elements: Every 150ms a character appeared at a random location on the screen (6-11 characters), and participants had to count how many of these characters were digits (vs. letters). After an unpredictable number of characters, rather than a digit or letter appearing, the grid of gabors appeared. Participants responded to the gabor task first (change/no change), but they were instructed to focus primarily on the digit task to ensure that they kept their attention globally spread.

To assess performance on the change detection task, we calculated d' to quantify participants' sensitivity to the changes in the local-only and local+ensemble conditions. We then calculated an *ensemble benefit score* by using regression to remove performance with local-only changes from performance with local+ensemble changes.

We used regression, not subtraction, because this results in an ensemble benefit score that has no correlation with performance in the local-only condition and a positive correlation with performance in the local+ensemble condition. In our task, where the presence of ensemble changes is likely to be helpful to performance but their absence is not actively negative for performance, this is the more valid analysis technique (e.g., DeGutis, Wilmer, Mercado, & Cohan, 2013; D. Ross, Richler, & Gauthier, 2014)[1]. Note that by regressing out performance in local-only from performance in local+ensemble, we also eliminate effects of motivation, change detection ability and other general factors from our ensemble benefit score. This is because these factors are present in the local-only condition as well as the local+ensemble condition. We performed the regression on z-scored values of d' so that the resulting coefficients are comparable across our tasks.

Rapid scene recognition measure

Our rapid scene recognition measure was based on the task employed by Davenport and Potter (2004). We presented participants with quickly flashed images of objects on top of scenes, and they had to report the identity of the object in a free response format. To the extent that participants are quicker and more accurate at rapid scene recognition, they should have higher accuracy for informative scene backgrounds (e.g., a priest in a church) than uninformative scene backgrounds (e.g., a priest on a football field). The objects are identical in the two conditions and only the usefulness of the scene differs, so this comparison, despite participants being asked about objects and not scenes, provides our index of rapid scene recognition.

We used 27 images of objects and 27 images of backgrounds combined into 27 informative and 27 uninformative object-background pairs (from Davenport & Potter, 2004); see Figure 2B. Each participant completed 54 trials, with each trial consisting of one object-background pair. Trials began with a fixation cross, and then the image (~28° X 17°) was presented for 84ms, followed by a mask for 200ms. Then participants had to type the name of the object they had seen. The masks consisted of checkerboard-scrambled versions of scenes. The same objects appeared twice for each participant, once

[1] In general, whether to use regression or subtraction depends on the task: if one condition is a true baseline, and the other condition only adds a factor on top, then regression is preferred (as in the current experiment). If one condition has a factor and the other has a negative version of that factor (e.g., if our local-only condition instead had actively misleading ensemble information), then subtraction is the more valid technique

on an informative background and once on an uninformative background. We counterbalanced the stimuli so that half of the objects appeared first in an informative background and half in an uninformative background.

Participants' responses were scored as correct only if they named the exact object (e.g., 'priest' or 'pope' or 'religious figure', not just 'man'). This scoring was done by two independent coders without knowledge of the condition represented by each response. The two coders' scores were in strong agreement, as they agreed on the correct/incorrect judgment of 96.8% of trials.

We calculated a scene benefit score by using regression to remove participants' performance on trials with the uninformative scenes from their performance on trials with the informative scenes. This regression also eliminates effects of motivation, object processing ability and other general factors from our scene benefit score. We performed this regression on z-scored values of percent correct. Using regression in this case is justified if the informative backgrounds are helpful for recognizing the objects, whereas uninformative backgrounds are unhelpful (rather than actively misleading). If uninformative backgrounds were actively misleading, then subtraction would be the preferred analysis technique (e.g., we should derive the scene benefit score from subtracting performance in the uninformative condition from performance in the informative condition). To disambiguate these, we would need a "neutral" condition. However, no neutral condition is feasible—there is no such thing as a scene that is exactly like other scenes, but makes no predictions at all about what objects are most likely to be present. Previous work has presented the objects without backgrounds (e.g., Davenport & Potter, 2004), but no-background conditions (or 1/f noise) make segmenting the object from the background much easier than it is in normal scenes. Consequently, these conditions are not truly neutral, but instead are significantly easier than conditions with true scenes. Since most objects can appear in most situations (e.g., none of the scenes is a physically impossible place for any of our objects), it seems most consistent to use regression, and we use that as our main measure. However, we report the effects using both regression and subtraction to show that the choice of analysis method is not critical to the conclusions.

Object-based summary statistics measure

This task was designed to measure participants' skill at computing summary statistics and was based on the task employed by Haberman, Brady, and Alvarez (2015). Participants completed 60 trials of a task where they had to report the average orientation of a grid of 4 gabor patches (see Figure 2C).

Each display consisted of 4 oriented gabors (~1 cycles/deg) varying in orientation. The 4 items were always ±5° and ±15° from the mean orientation, which was chosen randomly on each trial. Each gabor was located approximately 3° from fixation and subtended approximately 3.5°. Participants saw the display of gabors for 1 second and then after a 1 second delay, a test item appeared at the center of the screen. They had to adjust this item to reflect the average orientation of the set using their mouse. On each trial, we can compute an error measure as the angle, in degrees, between the correct response and participants' response, resulting in a distribution of errors across trials. We then fit a mixture model of a von Mises distribution and a uniform distribution to these error distributions using the MemToolbox (Suchow, Brady, Fougnie, & Alvarez, 2013), as is common in visual working memory experiments (e.g., Zhang & Luck, 2008). The standard deviation of this von Mises distribution (z-scored) was our measure of fidelity. This mixture model approach allowed us to assess the fidelity of participants' summary statistic computation independent of any lapse trials, which helps make our measure independent of participants' motivation level. While this model-based approach provides a more realistic measure of participant's ability to compute summary statistics, all of the same qualitative conclusions hold if we analyze mean absolute error without removing lapse trials.

Results

Main effects. Participants performed well in the spatial ensemble task, with 90.3% correct in the distractor digit counting task (S.E.M.: ±0.7%), and, looking at only trials with a correct digit response, a mean d’ of 1.1 (S.E.M.: ±0.1) in the local-only change detection condition and of 2.6 (±0.2) in the local+ensemble change condition. The difference between these two conditions was reliable, suggesting participants did, on average, take advantage of the ensemble structure (*t*(49)=12.7, *p*<0.0001, *Cohen’s d*=1.8; see Figure 3A).

----- Figure 3 about here -----

In the rapid scene recognition task, participants accurately recognized 72.1% (±1.8%) of the objects on the uninformative backgrounds but recognized 79.4% (±1.7%) on the informative backgrounds, a reliable effect of the scene background (*t*(49)=8.5, *p*<0.0001; see Figure 3B). Despite being a relatively small effect, this difference was highly consistent across participants, with a Cohen’s *d* of 1.2 and with only 3/50 participants showing better performance with uninformative than informative backgrounds.

In the summary statistic task, participants had an average fidelity of 13.7° (±0.62°, measured as the standard deviation of the von Mises distribution; see Figure 3B), with a lapse rate of 8.3% (±2.5%). Looking at all trials, rather than using the mixture model, and computing average absolute error rather than fitting a distribution, gives an average error of 14.6° (±1.3°).

----- Figure 4 about here -----

Reliability. Our primary interest is in the degree to which our different measures correlate with one another. However, the correlation observed between two variables is limited by the reliability with which those variables are measured. Thus, we first assessed the reliability of all of our measures using Spearman-Brown corrected split-half reliability (Brown, 1910; Spearman, 1910). All of our measures were highly reliable: participants’ performance at object-recognition on informative and uninformative backgrounds (*r*=0.95, *r*=0.93, respectively), d’ for local-only and local+ensemble change detection (*r*=0.95, *r*=0.86), and fidelity and lapse rate in the summary statistic task (*r*=0.85, *r*=0.86) all had reliability estimates greater than 0.85. Thus the maximum observable correlations between our tasks range from 0.85 to 0.92 (Nunnally Jr, 1970).

Correlations between tasks. Our main question of interest is the extent to which summary statistic processing and spatial ensemble processing are related to rapid scene recognition. To measure this, we used our scene benefit score, calculated by regressing performance with uninformative scenes out of performance with informative scenes (see Methods), our ensemble benefit score, calculated by regressing local+only performance out of the local+ensemble performance, and our measure of fidelity in the summary statistic task, calculated by removing lapse trials and calculating the standard deviation of participants' remaining reports.

We find that participants' ensemble benefit score is a significant predictor of their scene benefit score (r=0.46, r^2=0.21, p=0.001; see Figure 4A). In other words, the same participants who are good at detecting changes to the spatial ensemble structure are the participants who benefit most from informative scenes in an object recognition task. Since we regressed out performance at closely matched control conditions (e.g., uninformative scenes and local-only changes), this relationship cannot reflect motivation, general skill at object recognition or other general factors. Thus, 21% of the variance in our measure of rapid scene recognition can be explained by participants' sensitivity to the spatial structure of oriented gabors, consistent with the hypothesis that rapid scene recognition is supported by global ensemble texture processing of a scene.

By contrast, we find no significant relationship between performance at our summary statistic task and participants' scene benefit score (r=-0.14, r^2=0.02, p=0.35; Figure 4B) or ensemble benefit score (r=-0.15, r^2=0.02, p=0.30; Figure 4C). Thus, despite making use of the same local elements (gabors) and requiring both integration over multiple elements and a diffuse spread of attention, a simple summary statistic computation does not appear to be tied to rapid scene recognition or to the more texture-based spatial ensemble task (as it explains less than 2.3% of the variance in each). This result requires some revision to the assumption that all ‘ensemble’ tasks tap a similar ability, and suggests that spatial ensemble tasks may be more directly related to scene recognition.

If we use subtraction rather than regression to calculate the scene congruency effect (see Methods), we still find no relationship between the summary statistic task and participants’ scene benefit score (*r*=-0.00, r^2=0.00, *p*=0.99), and a significant relationship between the ensemble benefit score and the scene benefit score (*r*=0.28, r^2=0.08, *p*=0.048).

Discussion

Participants who were most sensitive to changes in spatial ensemble structure were also the participants most influenced by scene backgrounds in an object recognition task. This provides support for the hypothesis that spatial ensemble representations, or global ensemble texture more broadly, partly underlies rapid scene recognition. By contrast, computation of object-based summary statistics (i.e., average orientation) did not relate to scene recognition, as measured by our tasks, despite the similarity in the gabor stimuli used in the spatial ensemble task and the summary statistic task and the need for selection of all of the items in both tasks.

Broadly speaking, this provides evidence for a global view of rapid scene recognition, where information about a scene’s spatial layout is computed primarily based on the rapid encoding of patterns of orientation and spatial frequency across an image (e.g., Oliva & Torralba, 2006). These findings also highlight the strength of individual differences research for linking computational theories with cognitive models, and open the door to using individual differences to further examine the relationship between cognitive and neural models of scene perception. Our data also argue for a particular instantiation of a global scene recognition: a representation based on the spatial distribution of orientation and spatial frequency across a scene; as opposed to a global scene representation based on low-frequency information (e.g., Schyns & Oliva, 1994) or non-spatially localized global properties

(Greene & Oliva, 2009a). The layout information in these displays is carried by high spatial frequencies, not low spatial frequencies (e.g., if you blur these displays, you get a uniform gray field), suggesting the distribution of high spatial frequency information is critical, not low spatial frequency information. In addition, because they are not semantically meaningful, these spatial ensemble displays do not have properties like temperature or navigability (Greene & Oliva, 2009a). Thus, the connection we find between the spatial ensemble task and scene processing provides evidence that the spatial distribution of orientation at relatively high spatial frequencies – as used in computer vision models of spatial layout properties (Ross & Oliva, 2010) -- is related to scene recognition.

We controlled for general factors like motivation, working memory capacity, and object recognition by using a design with paired conditions. We also showed that not all global attention tasks correlate with rapid scene recognition, even ones dependent on very similar sets of gabor elements, like our summary statistic task. This suggests that the relationship we observe with scene recognition is selective to the processing of spatial patterns. By contrast, summary statistic tasks like the average orientation of gabors seem to have the majority of their individual differences explained by participant's precision at processing the individual gabors themselves (e.g. Haberman, Brady & Alvarez, 2015).

Nevertheless, individual differences are relatively indirect; a more direct measure would provide stronger evidence of a link between patterns of orientation and spatial frequency in an image and rapid scene recognition. Thus, in Experiment 2 and 3, we directly manipulate images in order to preserve only global ensemble texture information and ask whether this is sufficient to drive scene recognition (but not object recognition).

Experiment 2: Sufficiency of global ensemble texture for scenes

In a second experiment, we ask whether preserving only global ensemble texture information but eliminating the semantic meaning of scenes is still sufficient to activate scene representations.

Our primary manipulation is to "texturize" the scenes; that is, to eliminate all semantic information in the scenes and render them unrecognizable, and preserve only a small part of the spatial distribution of orientation and spatial frequency (see Figure 5 for example stimuli). In particular, we preserve only the power at 4 spatial frequencies and 6 orientations in a 6 by 6 spatial grid. This discards approximately 99.5% of the information from the original scenes[2], but preserves the limited set of spatial information about orientation and spatial frequency that we have proposed is critical for some aspects of scene recognition.

To measure scene recognition with texturized scenes, we once again use a task based on Davenport and Potter (2004). In particular, we ask whether participants are better at recognizing objects that follow textures derived from informative scenes (e.g., those that fit with the objects) as opposed to textures derived from uninformative scenes. This would be expected only if this texture information preserves sufficient information to drive the scene processing pathway and activate relevant scene representations to a sufficient extent to allow for the priming of relevant objects (perhaps based on the spatial layout of the scene, which is known to be available in such texture information; e.g., Ross & Oliva, 2010).

We modified the paradigm used in Experiment 1, in this case presenting the scenes before the objects-to-be-recognized, and thus having the scenes serve as primes for the objects (as in Palmer, 1975), rather than having the objects embedded in the scenes (as in Davenport & Potter, 2004). We did this because (1) the objects are easier to segment from texturized backgrounds (making the task too easy in some cases), and (2) because inserting objects into scenes changes the global scene statistics of the images, and the presence of consistent vs. inconsistent objects tends to change the global image features differently (e.g., Banno & Saiki, 2015; Gaspar & Rousselet, 2009; Mack & Palmeri, 2010). By

[2] While pixels are a poor measure of information, this reduces the simplest representation of the stimuli from 150,000-300,000 numbers (px) to 864 numbers (6 orientation/4 spatial frequencies in a 6x6 grid); and, under any encoding model, is a significant compression of the stimuli.

keeping the scenes intact without occluding them with objects, we allow participants to process the scene statistics without interference from overlapping objects.

To gauge the level of performance using texturized-scenes, we first ran a version of the experiment using non-texturized scenes. In Experiment 2A, we asked participants to recognize objects following intact grayscale scenes that were either informative or uninformative about the identity of the objects. In Experiment 2B, we asked participants to recognize the exact same objects, but now following texturized versions of the same scenes, which preserved only the distribution of orientation and spatial frequency information, but which were unrecognizable at the basic-level (e.g., oven, tennis court).

Method

Participants

50 participants were recruited on Amazon's Mechanical Turk for Experiment 2A (with non-texturized scenes). We expected a smaller effect size in Experiment 2B (with texturized scenes), so 100 participants were recruited for Experiment 2B. All participants were from the United States, were over 18, and gave informed consent in accordance with the procedures and protocols approved by the Harvard Committee on the Use of Human Subjects. Turk users form a representative subset of adults in the United States (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011), and data from Turk are known to closely match data from the lab on visual cognition tasks (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013). All participants indicated they had normal or corrected-to-normal color vision. All participants were paid 1 dollar for several minutes of their time and none of the participants participated in multiple experiments (all participants are identified by a unique ID by Amazon).

Stimuli

Stimuli consisted of the 27 object-scene pairs from Experiment 1 (taken from the set created by Davenport & Potter, 2004), augmented by 23 additional pairs to create 50 informative object-scene pairs. Each object and scene was also paired with a different object and scene to create uninformative object-scene pairs, as in Experiment 1 and Davenport and Potter (2004). In this experiment, the scenes did not contain the objects, but instead were separate images. The objects and scenes were both presented in grayscale to remove color as a cue. In addition, the objects were presented on 1/f noise backgrounds to make it more difficult to see and categorize the objects (see Figure 5). The scenes were a mixture of indoor, outdoor, and urban places, and were paired with objects of various kinds [animals, people, things], at different sizes (from close views of an oven or desktop to large-scale views of a mountain). In particular, the stimuli consisted of: airport [pilot]; barn [tractor]; basketball court [basketball player]; a bathroom [tub]; bathroom counter [a comb]; battle ground [soldier]; baseball field [mitt]; beach [surfer]; bowling alley [bowler]; cemetery [gravestone]; church [priest]; desert [cactus]; football field [football player]; field [buffalo]; fire station [fireman]; forest [deer]; grass [butterfly]; hallway [table]; helipad [helicopter]; hospital [doctor]; ice rink [figure skater]; kitchen [knife]; library [student]; living room [couch]; mountain trail [donkey with rider]; mud pit [pig]; NASCAR racer track [racecar]; NFL football game [referee]; ocean [fish]; inside of oven [pie]; parade [trumpet player]; parking lot [car]; path in a park [jogger]; ping pong table [paddle]; resort [a boat]; restaurant kitchen [chef]; rocks/stones [penguin]; sand [sandcastle]; savannah/field [zebra]; a ship's deck [life preserver]; the sky [hot air balloon]; snowy hill [sled]; space (earth/stars) [space shuttle]; a street (flat view) [a biker]; a supermarket [shopping basket]; a tennis court [racket]; a theater [ballerina]; a racehorse track [race horse with jockey]; a street (perspective view) [a truck]; and underwater [turtle].

In Experiment 2A, the unmanipulated scenes were presented. In Experiment 2B, they were first "texturized", using the algorithm of Oliva & Torralba (2006). In particular, the images were divided up into a 6x6 grid, and in each grid cell the power was estimated at 4 spatial frequencies x 6 orientations. This reduces the hundreds of thousands of pixels of information in an image to just 864 numbers, discarding approximately 99.5% of the information in each image when the image is (naively) coded in pixels. Under any coding algorithm, the image ends up highly compressed and most information is discarded. Then, a random white noise image was generated, and this image was iteratively coerced to have the same distribution of orientations and spatial frequencies in each cell as the original image did. At each iteration, the noise is decomposed using a bank of multiscale-oriented filters and the magnitude output of the filters is averaged over each grid cell, then these features are modified to more closely match the 4 x 6 spatial frequency/orientation features of the target image in each cell. Through an iterative process, the noise image more and more closely matches the statistics of average orientation/spatial frequency of the original image in each of the 6x6 cells. Before applying the iterative adjustment to the white noise image, the adjustment factor for each of the 6x6 cells is scaled up to the original size of the image with bicubic interpolation, resulting in some smoothing, which is why the images do not display grid artifacts.

This texturized version of the scenes preserves most of the orientation and spatial frequency information from the original image, but their spatial organization is only loosely preserved. This destroys the majority of the recognizable features of the image but preserves some information about the spatial layout of the scene (e.g., Oliva & Torralba, 2006; see Figure 5B).

We ensured that the images were no longer recognizable as a basic-level (e.g., kitchen, forest, etc.) by running a control experiment in which 30 naïve participants were shown these images and asked via free response to guess what kind of image they were generated from or most closely resembled. Participants could not succeed at this task. Even with very liberal grading criteria, only 3.4% of the images were recognized, and this was largely due to participants' tendency to guess the same answer for many images (e.g., people called many of the images beaches, even when this was incorrect). To demonstrate this, we shuffled the labels and images relative to each other so the labels were graded with different scenes than the participants saw; 2.9% - 4.8% of the labels were still judged as correct across each of 3 random shuffles. Thus, it is unlikely any of the responses reflected true recognition of the scenes, as a similar percent correct was found with the correct labeling or with shuffled labels. Thus, the texturized images were generally unrecognizable at the basic-level.

----- Figure 5 about here -----

Procedure

We presented participants with images of scenes (2A) or texturized scenes (2B) for 500ms, followed by briefly flashed object images for 100ms, and then a mask (the same masks used in Davenport & Potter, 2004 and in Experiment 1). Participants then had to report the identity of the object in a free response format. Each participant saw all 50 objects, with half paired with an informative scene and half paired with an uninformative scene (2A) or a texturized version of those same scenes (2B). To the extent that the prime scenes/textures drive participant's scene recognition system and thus prime the relevant objects, participants should have higher accuracy when preceding scenes or textures contain informative vs. uninformative information. The objects are identical in the two conditions, and only the usefulness of the prime scene/texture differs—so this comparison, despite participants being asked about objects and not scenes, provides our index of whether the prime scenes/textures successfully drive the scene recognition system. By using texturized-scenes, Experiment 2B allows us to ask if the same informative scene benefit is present even when only a simple distribution of low-level information is preserved: e.g., enough to provide information, at least in theory, about the spatial layout of the scene (e.g., Ross & Oliva, 2010), but without any basic-level recognition.

As in Experiment 1, participants' responses were scored as correct only if they named the exact object (e.g., 'priest' or 'pope' or 'religious figure', not just 'man'). This scoring was once again done without knowledge of the condition represented by each response (e.g., blind to condition).

Results

In Experiment 2A, with meaningful scenes as primes, participants accurately recognized 72.7% (±2.2%) of the objects primed by uninformative backgrounds but recognized 82.4% (±2.2%) primed by the informative backgrounds, a reliable effect of the scene's informativeness ($t(49)=7.91$, $p<0.0001$; see Figure 6A). Thus, the benefit of informative scenes on object recognition (e.g., Davenport & Potter,

2004) replicates even with grayscale scenes (see Munneke, Brentari, & Peelen, 2013) and even with the scene as a prime rather than with participants having to segment the object from the scene (e.g., Palmer, 1975).

Is preserving only a distribution of spatial frequencies and orientations in the texturized-scene condition sufficient to drive an object recognition benefit (Experiment 2B)? We found that participants accurately recognized 76.5% (±1.0%) of the objects primed by texturized versions of uninformative backgrounds but recognized 79.4% (±0.9%) of the objects primed by texturized-informative backgrounds, a reliable effect of the informativeness of the texturized scene (*t*(99)=3.11, *p*=0.002; see Figure 6B). Thus, the texturized scenes, which are not recognizable at the basic-level, nevertheless prime the identity of objects that are consistent with the original scenes. This suggests that preserving only the spatial distribution of orientation and spatial frequency is sufficient to drive the scene pathway and allow the activation of scene representations and the associated object representations.

The effect of informativeness was reliable not only across participants, but also across items (object-scene pairs; *t(*49)=3.10, *p*=0.003). This suggests that the effect is generalizable across the scenes we showed, rather than driven by just a few pairs of scenes and objects. Given the diversity of our stimulus set (indoor; outdoor; urban; natural, with far views, close views; and animals, people and things), this shows significant generalization of the effect. The effect was also not driven by the small chance of participant's recognizing a texturized-scene. If we calculate a priming effect using only the scenes that not a single participant guessed the identity of in the control experiment, we find a priming effect of 3.4% (which is significantly greater than zero; *t*(19)=3.11, *p*=0.006); with scenes that at least one person guessed the identity of, the priming effect was only 1.8%, a numerical smaller effect (the opposite of what would be predicted). This difference for ever-recognized vs. never-recognized scenes was not significant (*t*(48)=1.01, *p*=0.32).

----- Figure 6 about here -----

Discussion

We found that even texturized versions of informative scenes were sufficient to drive an object recognition advantage, although this advantage was less than that provided by the full scenes (which convey a lot of other information, including semantics). This provides further support for the idea that global pattern information, like the spatial distribution of orientations and spatial frequencies is sufficient to activate some aspects of scene representations. This may be because these global ensemble textures preserve information about scene layout (e.g., Oliva & Torralba, 2006; Ross & Oliva, 2010), and spatial layout information alone is sufficient to generate predictions about which objects are commonly present in the activated scene, thereby facilitating object detection and recognition (Bar, 2004; Bar et al., 2006). This texture information may also be sufficient to activate other aspects of scene representations (e.g., affordances; Greene & Oliva, 2010).

We used facilitation of object recognition as our measure of whether scenes were sufficiently processed to activate scene representations. Our results suggest that global ensemble texture representations are sufficient to activate representations of related objects, suggesting that object-scene consistency effects may be in part driven by global scene structure rather than solely by the semantic information in recognizable scenes. This claim is consistent with some previous work which has also pointed to the fact that object-scene consistency effects can be driven by spatially global representations of scenes. For example Munneke, Brentari and Peelen (2013) showed that the spatial location of attention had little effect on the scene benefit for objects, suggesting a more global, gist-based representation might be responsible.

Overall, the current results suggest that sensitivity to the distribution of orientations and spatial frequencies – what we call global ensemble texture – can activate scene representations, perhaps because this information is critical to the representation of scenes' spatial layout. Combined with

Experiment 1, these results reinforce the proposed link between such global ensemble texture and scene recognition.

Experiment 3: Is global ensemble texture particularly informative for scenes?

In the first two experiments, we showed that (1) the same participants who are the best at recognizing global pattern in simple grids of gabor elements are also the best at rapid scene recognition, and (2) preserving only a grid of orientation and spatial frequency information is sufficient to drive the scene pathway, at least enough to activate and prime relevant objects. In both cases, we suggested this is because of a link between *scenes* in particular and global ensemble texture patterns. Indeed, computational work has shown that such texture representations are particularly informative for scenes, since such texture patterns preserve information about 3D scene structure (e.g., Ross & Oliva, 2010).

In a third experiment, we asked whether global ensemble texture information provided information that was particularly relevant for scene representations, as we have hypothesized, or whether global ensemble texture was instead equally useful for driving object recognition systems. In particular, we designed a stimulus set and experiment that mirrored that of Experiment 2A and 2B, but rather than using scenes and texturized-scenes as primes, we used objects (3A) and texturized-objects (3B). We reasoned that if the preservation of global ensemble texture information is informative only for scenes and not for objects, as would be expected if it is driven primarily by sensitivity to spatial layout, then, despite the presence of a strong priming effect from texturized-scenes (in Experiment 2B), we should abolish all priming effects by using texturized objects (in Experiment 3B).

Experiment 3 was thus identical to Experiment 2, except using objects rather than scenes as primes: an informative object prime (e.g., a basketball hoop) or uninformative object prime (e.g., a cooking pot) was shown, followed by an object to be recognized (e.g., a basketball player), after which the object was masked and then participants had to type the name of the object they saw. The objects that needed to be recognized were identical to those in Experiment 2.

Existing work has shown that object-to-object consistency gives rise to object recognition benefits, just as scene-to-object consistency give rise to object recognition benefits. For example, Davenport (2007) showed in a paradigm very similar to that of Davenport and Potter (2004) that informative objects facilitated free responses for naming other objects (see also Auckland, Cave, & Donnelly, 2007). Thus, we reasoned that objects should serve as primes exactly as well as scenes (Experiment 3A). This allows us to investigate whether texturizing those objects preserves the priming effect as it did for scenes (Experiment 3B). We used pilot experiments to choose the prime objects, which allowed us to match performance with the informative-object primes (Exp. 3A) to the performance of informative-scene primes (Exp. 3B), thus providing an equal starting point for asking about how texturizing the primes affects performance in scenes and objects.

Method

Participants

50 participants were recruited on Amazon's Mechanical Turk for Experiment 3A, which we expected to have a similar effect size to Experiment 2A. To choose a sample size for Experiment 3B, we did a power calculation based on the data from Experiment 2B. Because we hypothesized that texturized-objects might not lead to a priming effect, we made sure we had 95% power to detect the same size priming effect we observed with texturized-scenes (Cohen's d=0.31). Achieving this power requires 136 participants. Thus, in Experiment 3B, we recruited 150 participants, giving ample power to detect a priming effect if one is present with texturized-objects. All participants were from the United States, were over 18, and gave informed consent in accordance with the procedures and protocols approved by the Harvard Committee on the Use of Human Subjects. All participants indicated they had normal or corrected-to-normal color vision. All participants were paid 1 dollar for several minutes of their time and

none of the participants participated in multiple experiments (all participants are identified by a unique ID by Amazon).

----- Figure 7 about here -----

Stimuli

Stimuli consisted of the same 50 objects as in Experiment 2, but rather than scenes serving as primes, related objects instead served as primes (e.g., a cooking pot for a chef; a basketball hoop for a basketball player; a checkered flag for a race car; see Figure 7). In Experiment 3A, the prime-objects were presented normally. In Experiment 3B, they were first "texturized", using the same algorithm as described in Experiment 2B.

As in Experiment 2, we ensured that the texturized object-prime images were difficult or impossible to recognize at a basic-level (e.g., pot, bunny, etc.) by running a control experiment in which 30 naïve participants were shown the texturized-object images and asked via free response to guess what kind of image they were generated from or most closely resembled. Participants were generally unsuccessful at this task (6.3% correct), although there were 4 images that were recognized a significant portion of the time (a snake, a rabbit, a giraffe and a fork) – all cases where the "outline" of the image was sufficient to drive recognition in cases where participants were explicitly asked to recognize the object. It remains unlikely that participants would recognize these objects in the context of the experiment, but, to ensure the possibility of recognition did not effect our results, we look at performance with these images separately as well as analyzing all images together.

Procedure

The procedure was identical to that of Experiment 2, except with prime objects (3A)/prime texturized-objects (3B) rather than prime scenes/texturized-scenes.

Results

In Experiment 3A, with recognizable objects as primes, participants accurately recognized 73.7% (±1.9%) of the objects primed by uninformative objects but recognized 82.9% (±1.5%) primed by the informative objects, a reliable effect of the prime object (*t*(49)=5.89, *p*<0.0001; see Figure 6C). Thus, the basic benefit of informative objects on object recognition was very similar to the effect of informative scenes on object recognition (benefit of informative scenes: 9.7%, benefit of informative objects: 9.2%).

Is preserving only a distribution of spatial frequencies and orientations in the texturized-object condition sufficient to drive an object recognition benefit (Experiment 3B) as it was with scenes? Participants accurately recognized 77.2% (±1.0%) of the objects primed by texturized versions of uninformative objects and recognized 77.2% (±1.0%) of the objects primed by texturized versions of informative objects. Thus, there was no reliable effect of the informativeness of the texturized object prime (*t*(149)=0.12, *p*=0.90; see Figure 6D). Moreover, comparing Experiments 2B and 3B shows that the benefit for texturized-objects (-0.08%) was significantly smaller than the benefit for texturized-scenes (2.9%; *t*(248)=2.63, *p*=0.009), showing an interaction between experiments. Thus, while the texturized scenes nevertheless prime the identity of objects that are consistent with the scenes, the texturized objects do not. This is despite the fact that fully recognizable objects and scenes result in the same priming effect. This suggests that preserving only the spatial distribution of orientation and spatial frequency is sufficient to drive the scene pathway but not the object pathway.

As with the texturized-scenes, we can break down the effect by whether the prime object was recognized or not. If we calculate a priming effect using only the prime-objects that not a single participant guessed the identity of in the control experiment, we find a priming effect of 0.5%; with prime-objects that at least one person guessed the identity of, the priming effect was -0.4%. This difference is not significant (*t*(48)=0.52, *p*=0.61). Thus, the small chance of a texturized-scene or texturized-object being recognized by a participant does not seem to modulate the priming effect.

Discussion

We found that texturized versions of prime objects were insufficient to drive an object recognition advantage. Thus, while the texturized scenes prime the identity of objects that are consistent with the scenes, the texturized objects do not. This is despite the fact that fully recognizable objects and scenes result in similar size priming effects. This suggests that preserving only the spatial distribution of orientation and spatial frequency – the global ensemble texture -- is sufficient to drive the scene pathway but not the object pathway.

General Discussion

In Experiment 1, we found that participants who were most sensitive to changes in spatial ensemble structure were also the participants most influenced by scene backgrounds in an object recognition task. This suggests a link between spatial ensemble processing and rapid scene recognition. In a second experiment, we showed that preserving only global ensemble texture information in scenes is sufficient to allow participants to activate scene representations. In a third experiment, we show that this link between global ensemble texture and scenes is selective to scenes: preserving the same information in images of objects is insufficient to allow activation of related object representations. Overall, our data support the hypothesis that global ensemble texture representations can drive activation of scene information during rapid scene recognition. This is consistent with computer vision models showing the sufficiency of global patterns of orientation and spatial frequency for recognizing scene information (Oliva & Torralba, 2001, 2006; Renninger & Malik, 2004; Sofer et al., 2015) and in particular, information about spatial layout (e.g., Ross & Oliva, 2010).

Our data argue against a purely object-based view of scene recognition in favor of a more global account. Our data also point to a particular instantiation of global scene recognition: a representation based on the spatial distribution of orientation and spatial frequency across a scene; as opposed to a global scene representation based on low-frequency information (e.g., Schyns & Oliva, 1994) or non-spatially localized global properties (Greene & Oliva, 2009a). For example, because the displays from the spatial ensemble task in Experiment 1 and the tasks of Experiments 2 and 3 are not semantically meaningful, these spatial ensemble displays do not have properties like temperature or navigability (Greene & Oliva, 2009a). Thus, the connection we find between the spatial ensemble tasks and scene processing and the preservation of priming from texturized-scenes provides evidence for a global scene recognition system based at least in part on the spatial distribution of orientation at relatively high spatial frequencies rather than solely based on affordances and other semantic global properties (Greene & Oliva, 2009a).

It remains an open question at what level such ensemble texture effects operate. For example, the priming effects of Experiment 2 could be relatively high-level or low-level. At a high level, participants might directly perceive spatial layout in our texturized-scenes, allowing them to activate the relevant object representations. Alternatively, the effects could arise at a lower-level; for example, participants might be primed by large homogenous regions in the scene to expect large objects vs. small ones. One important note here is that any account needs to explain why priming is preserved for texturized-scenes but eliminated for texturized-objects. Thus, some scene-specific information must be posited, even in low-level accounts.

While our results suggest some role for global ensemble texture in scene recognition, global ensemble texture information is certainly not the only thing relevant to scene recognition. People accumulate a great deal of information about scenes over multiple saccades and integrate this information into a rich scene representation (e.g., Hollingworth & Henderson, 2002; Hollingworth, 2004, 2006; Malcolm, Nuthmann, & Schyns, 2014). In addition, more fine-grained information, like junctions between contours, are also relevant to how participants rapidly recognize scenes (e.g., Walther & Shen, 2014). However, our results do point to the possibility that scene processing may be partially reliant on distributions of orientation and spatial frequency that are not totally localized.

Separate object and scene processing pathways

Bar (2004), among others, has argued that low spatial frequencies might be processed quickly to arrive at a perceptual hypothesis about the identity of an object. Our proposal is related but different, in that the global ensemble texture information we propose helps underlie scene recognition is primarily reflected in a spatial distribution of high spatial frequency information rather than the low spatial frequency information. For example, if blurred with a low-pass filter, the stimuli from our spatial ensemble task (Figure 1A) become a uniform gray field. While low frequency information may be particularly informative for objects, as it preserves overall shape contours (e.g., Bar, 2004), the distribution of relatively high-spatial frequency information has previously been shown to be particularly informative for scene layout (e.g., Ross & Oliva, 2010).

This suggests a possible dissociation between the processing of scenes and the processing of objects, which may be related to the known dissociation between how these stimuli are processed in the ventral visual pathway (e.g., Kanwisher, 2010). In general, our data are consistent with a two-pathway view of the brain's processing of visual scenes, in which one focal attention-bound pathway (e.g., LOC, pFS) processes object information while a second non-attentional (or distributed attention) pathway processes scene information via global ensemble texture and spatial layout (e.g., OPA/TOS, PPA) (Park et al., 2011; Wolfe et al., 2011). In particular, neuroimaging studies of scene-selective brain regions suggest that, of all the ways scenes differ from objects, the dimensions most relevant for these brain regions are the spatial layout of the scenes and their visual texture rather than the number of objects present or how complicated the relations between objects are (Cant & Xu, 2012; Dilks, Julian, Paunov, & Kanwisher, 2013; Epstein & Kanwisher, 1998; Epstein, 2005). This is consistent with the idea that global ensemble texture information may be particularly relevant for scenes, rather than objects, and that this may be related to such texture information's utility for determining the spatial layout of a scene.

One possibility is that these two pathways – an object pathway and a scene pathway -- process all scenes simultaneously (e.g., Wolfe et al. 2011). For example, when viewing a single scene, areas like LOC may process information about the objects and content while simultaneously areas like PPA process information about spatial layout (e.g., Park, Brady, Greene & Oliva, 2011).

Effect of image statistics on object and scene recognition

We argue that rapid scene recognition may rely on global ensemble texture processing, but that object recognition requires more information than just the global ensemble texture of the object. In particular, we find that our texturized objects (Exp. 3B) are insufficient to prime related objects, whereas the same texturization process preserves enough information about scenes to prime related objects (Exp. 2B). However, there do seem to be some circumstances where participants can make very basic distinctions about the objects an image contains based on global image statistics. In particular, there is a significant literature on rapid detection of whether an animal is present in a scene or not (Kirchner & Thorpe, 2006; and some related work on tasks like vehicle detection; VanRullen & Thorpe, 2001). These tasks show that participants can very rapidly detect whether an image contains an animal. However, some have argued that rather than doing object recognition per se, participants may succeed at these tasks in part by analyzing the images holistically and asking whether their global image statistics (like their amplitude spectra) are consistent with what would be expected of an image with an animal in it (e.g., Torralba & Oliva, 2003). However, the extent to which this is true remains unclear (Crouzet, Joubert, Thorpe, & Fabre-Thorpe, 2012; Fabre-Thorpe, 2011; Gaspar & Rousselet, 2009) and for the most part, this strategy appears to be useful only for making superordinate-level categorizations about large central objects, rather than a more general property of object recognition (Fabre-Thorpe, 2011).

While the global ensemble texture of the object alone does not support basic-level object recognition, it is possible that large objects may affect the global ensemble texture of scenes, making the scenes more or less recognizable. For example, recent work on rapid scene recognition has shown that even with very rapid scene categorization, participants are faster to recognize a scene in the presence of congruent objects (compared to incongruent objects) (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). However, this effect of objects on scene recognition can actually be modeled by considering the ways in which adding large objects to a scene affects the global image statistics of a scene (Mack & Palmeri, 2010). In particular, differences in the global ensemble texture of the congruent vs. incongruent images are sufficient to explain this effect without any appeal to object recognition per se. Thus, these data are consistent with our claim that rapid scene recognition may be particularly related to global ensemble texture processing, and, at least in some cases, object congruency effects may be caused not by object recognition processes per se but by the way objects affect global ensemble texture and thus scene recognition (Mack & Palmeri, 2010). Note that in the current experiments, we are interested in the opposite effect (the extent to which scenes prime object recognition), so our use of global ensemble texture does not conflict with the results of Mack and Palmeri (2010) but instead provides additional support for a global ensemble texture view of scene recognition. In addition, in Experiments 2 and 3, we presented the prime scenes/objects and test objects sequentially to avoid any interactions in how the objects modified the scene statistics or object statistics in a simultaneous display.

Throughout the current set of experiments, we used a task where scene recognition was measured only indirectly, through its facilitation of object recognition. We based this decision on the robust literature suggesting scenes influence objects in an interactive manner during early recognition (e.g., Joubert et al. 2008). In Experiment 2, we show this facilitation of object recognition can occur even with limited scene information (only the global ensemble texture). In many ways, this very limited scene context is similar to the paradigm used in contextual cueing experiments (Chun & Jiang, 1998). In these paradigms, contextual information is often just the location of relevant distractor objects in a display of simple discrete objects (like T's and L's). Having a consistent and recurring background context can help make decisions about target objects – like which direction a sideways T is facing – easier (Brady & Chun, 2007; Kunar, Flusberg, & Wolfe, 2006). One interesting implication of this is to ask whether global ensemble texture information might be particularly useful for guiding visual search during contextual cueing and other memory-based tasks where limited "scene" information is used to guide object-based tasks.

Choice of texture representation

Many studies have relied on the Portilla and Simoncelli algorithm (Portilla & Simoncelli, 2000) to preserve low-level information while discarding high-level information in natural images. In the current experiments, we instead make use of a model based on V1-like features (the GIST model of Oliva & Torralba, 2001; 2006). We made use of this texture algorithm because we are most interested in how people represent spatial structure – e.g., the top of the image being largely made-up of vertical elements and the bottom horizontal elements, an important clue to spatial layout -- which is not the kind of structure the Portilla and Simoncelli texture model represents. In fact, the Portilla and Simoncelli algorithm assumes stationarity (homogeneity) across the image (Portilla & Simoncelli, 2000). Thus, while this algorithm preserves important texture information, it does not preserve the kind of spatial layout information we are interested in the current experiments (see Figure 8 for examples).

----- Figure 8 about here -----

Of course, non-stationary texture models could be employed that are considerably more sophisticated than our simple grid of orientations and spatial frequencies model. However, one benefit of the simpler texture algorithm we use is that the analogy between the representation of global ensemble texture we use here and the spatial ensemble Gabor-task we use in Experiment 1 is extremely direct: Both are limited to a set of orientations at fixed spatial frequencies and grid locations. The success of even this simple texture algorithm at preserving spatial layout information but discarding semantic information and object-based information provides a motivation for why participants might be good at the spatial ensemble tasks we employ in Experiment 1, and why performance in such tasks might be related to scene recognition.

Distinctions between summary statistic tasks and spatial ensemble tasks

In Experiment 1, we found that computation of non-spatial summary statistics (i.e., average orientation) did not relate to scene recognition, despite the similarity between the gabor elements and global attention required in the spatial ensemble task and the summary statistic task. In the context of our task, this suggests that the correlation we find between spatial ensembles and scene recognition is not driven purely by the ability to globally attend to multiple gabor elements. However, this data also suggests that spatial ensembles and non-spatial summary statistics may be distinct. In particular, the major constraint on computing summary statistics like the mean may be how precisely the individual elements are represented, as this places a limit on the possible precision of such statistical summaries (e.g., Alvarez, 2011; Haberman et al. 2015). In other words, non-spatial summary statistics like the mean orientation of a set may be more related to the precision of individual object representations, while, ensemble representations that require the preservation of distributions of spatial information may be particularly related to scene recognition.

Alternatively, there may be aspects of our task that results in the summary statistic task being performed differently than the spatial ensemble task. For example, consistent with existing studies of summary statistics, we used relatively long 1 second exposures (e.g., Haberman et al., 2015; Sweeny & Whitney, 2014). Thus, participants in this task may have performed it with serial attention, weakening the link to scene recognition.

Conclusion

The present series of studies argues for an important link between global ensemble texture information and scene recognition. We first used an individual differences approach to establish a relationship between rapid scene perception and spatial ensemble processing (but not non-spatial statistical summary perception), a kind of global ensemble texture representation. We then showed the sufficiency of global ensemble texture information for activating scene representations, but not object representations, using a priming paradigm. Together, these studies provide support for the hypothesis that global ensemble texture representations partly underlie rapid scene recognition.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661311000040

Alvarez, G. A., & Oliva, A. (2009). Spatial Ensemble Statistics: Efficient Codes that Can be Represented with Reduced Attention. *Proceedings of the National Academy of Sciences*, *106*, 7345.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162. Retrieved from http://pss.sagepub.com/content/12/2/157.short

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review*, *14*(2), 332–337. http://doi.org/10.3758/BF03194073

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13. Retrieved from http://www.journalofvision.org/content/9/12/13.short

Banno, H., & Saiki, J. (2015). The use of higher-order statistics in rapid object categorization in natural scenes. *Journal of Vision*, *15*(2), 1–20. http://doi.org/10.1167/15.2.4.doi

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Schmidt, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 449–54. http://doi.org/10.1073/pnas.0507062103

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. http://doi.org/10.1093/pan/mpr057

Biederman, I., Mezzanotte, R., & Rabinowitz, J. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. Retrieved from http://www.sciencedirect.com/science/article/pii/001002858290007X

Boyce, S. J., & Pollatsek, a. (1992). Identification of objects in scenes: the role of scene background in object naming. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *18*(3), 531–543. http://doi.org/10.1037/0278-7393.18.3.531

Boyce, S. J., Pollatsek, a, & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology. Human Perception and Performance*, *15*(3), 556–566. http://doi.org/10.1037/0096-1523.15.3.556

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science*, *22*(3), 384. Retrieved from http://pss.sagepub.com/content/22/3/384.short

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Brady, T. F., & Chun, M. M. (2007). Spatial constraints on learning in visual search: Modeling contextual cuing. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(4), 798–815. Retrieved from http://camplab.psych.yale.edu/articles/PDFs for Website/Brady2007_HumanPerception.pdf

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4. Retrieved from http://www.journalofvision.orgwww.journalofvision.org/content/11/5/4.short

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8295.1910.tb00207.x/full

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://doi.org/10.1177/1745691610393980

Cant, J., & Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *The Journal of Neuroscience*, *32*(22), 7685–7700. Retrieved from http://www.jneurosci.org/content/32/22/7685.short

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12535996

Crouzet, S. M., Joubert, O. R., Thorpe, S. J., & Fabre-Thorpe, M. (2012). Animal Detection Precedes Access to Scene Category. *PLoS ONE*, *7*(12), 1–9. http://doi.org/10.1371/journal.pone.0051471

Dakin, S., & Watt, R. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181–3192. Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698997001338

Davenport, L. J., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15(8)*, 559.

DeGutis, J., Wilmer, J., Mercado, R., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87–100. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027712002041

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *The Journal of Neuroscience*, *33*(4), 1331–1336. Retrieved from http://www.jneurosci.org/content/33/4/1331.short

Epstein, R. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13506280444000607

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9560155

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology*, *2*(OCT), 1–12. http://doi.org/10.3389/fpsyg.2011.00243

Gaspar, C. M., & Rousselet, G. a. (2009). How do amplitude spectra influence rapid animal detection? *Vision Research*, *49*(24), 3001–3012. http://doi.org/10.1016/j.visres.2009.09.021

Greene, M., & Oliva, A. (2009a). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010028508000455

Greene, M., & Oliva, A. (2009b). The briefest of glances The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472. Retrieved from http://pss.sagepub.com/content/20/4/464.short

Greene, M., & Oliva, A. (2010). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human …*. Retrieved from http://psycnet.apa.org/journals/xhp/36/6/1430/

Guyader, N., Chauvin, A., & Peyrin, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process. *Comptes Rendus …*. Retrieved from http://www.sciencedirect.com/science/article/pii/S163106910400071X

Haberman, J., Brady, T., & Alvarez, G. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental …*. Retrieved from http://psycnet.apa.org/journals/xge/144/2/432/

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17(17)*, R751.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: the roles of short- and long-term visual memory. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(3), 519–37. http://doi.org/10.1037/0096-1523.30.3.519

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition*, *14*(4-8), 781–807. http://doi.org/10.1080/13506280500193818

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology …*. Retrieved from http://www.psy.ed.ac.uk/research/vc/pdfs/33-1998_Hollingworth_Henderson_ consistent scene.pdf

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychologica*, *102*(2-3), 319–343. http://doi.org/10.1016/S0001-6918(98)00053-5

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended object in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 113.

Honey, C., Kirchner, H., VanRullen, R., M., E., R., N., R., V., … V., F. A. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision*, *8*(12), 9–9. http://doi.org/10.1167/8.12.9

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

Joubert, O. R., Fize, D., Rousselet, G. a, & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, *8*(13), 11.1–18. http://doi.org/10.1167/8.13.11

Joubert, O. R., Rousselet, G. a, Fabre-Thorpe, M., & Fize, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision*, *9*(1), 2.1–16. http://doi.org/10.1167/9.1.2

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*(26), 3286–3297. http://doi.org/10.1016/j.visres.2007.09.013

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. http://doi.org/10.1073/pnas.1005062107

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition*, *15*(1), 12–19. http://doi.org/10.1080/13506280600856660

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776. http://doi.org/10.1016/j.visres.2005.10.002

Kunar, M. a, Flusberg, S. J., & Wolfe, J. M. (2006). Contextual cuing by global features. *Perception & Psychophysics*, *68*(7), 1204–16. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2678916&tool=pmcentrez&rendertype=abstract

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(6), 1431–1450. http://doi.org/10.1037/0096-1523.33.6.1431

Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision*, *10*(3), 11.1–11. http://doi.org/10.1167/10.3.11

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: strategic and incremental information accumulation for scene categorization. *Psychological Science*, *25*(5), 1087–97. http://doi.org/10.1177/0956797614522816

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, Henry Holt and Co. *Inc., New York, NY*. Retrieved from http://scholar.google.com/scholar?q=Marr%2C+1982&btnG=&hl=en&as_sdt=0%2C22#4

Munneke, J., Brentari, V., & Peelen, M. V. (2013). The influence of scene context on object recognition is independent ofattentional focus. *Frontiers in Psychology*, *4*(AUG), 1–10. http://doi.org/10.3389/fpsyg.2013.00552

Nunnally Jr, J. (1970). *Introduction to psychological measurement.* New York, NY: McGraw-Hill. Retrieved from http://psycnet.apa.org/psycinfo/1970-19724-000

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. Retrieved from http://www.springerlink.com/index/K62TG81W8352G71H.pdf

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0079612306550022

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. http://doi.org/10.3758/BF03197524

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *The Journal of Neuroscience*, *31*(4), 1333–1340. http://doi.org/10.1523/JNEUROSCI.3885-10.2011

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*, 739.

Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, *40*(1), 49–70. http://doi.org/10.1023/A:1026553619983

Potter, M., & Faulconer, B. (1975). Time to understand pictures and words. *Nature*, *253*, 437–438. Retrieved from http://www.nature.com/nature/journal/v253/n5491/abs/253437a0.html

Renninger, L., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311. Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698904001919

Ross, D., Richler, J., & Gauthier, I. (2014). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*. Retrieved from http://link.springer.com/article/10.3758/s13428-014-0497-4

Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1), 2.1–25. http://doi.org/10.1167/10.1.2

Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010028503000021

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*. Retrieved from http://www.jstor.org/stable/40063215

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195. Retrieved from http://pss.sagepub.com/content/5/4/195.abstract

Sofer, I., Crouzet, S., & Serre, T. (2015). Explaining the timing of natural scene understanding with a computational model of perceptual categorization. *PLoS Comput Biol*. Retrieved from http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004456

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8295.1910.tb00206.x/full

Suchow, J., Brady, T., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10), 9. Retrieved from http://www.journalofvision.org/content/13/10/9.short

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903–13. http://doi.org/10.1177/0956797614544510

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. Retrieved from http://fias.uni-frankfurt.de/~triesch/courses/260object/papers/SpeedOfProcessing.pdf

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412. Retrieved from http://informahealthcare.com/doi/abs/10.1088/0954-898X_14_3_302

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668. http://doi.org/10.1068/p3029

Vogel, E. K., & Awh, E. (2008). How to Exploit Diversity for Scientific Gain: Using Individual Differences to Constrain Cognitive Theory. *Current Directions in Psychological Science*, *17*(2), 171–176. http://doi.org/10.1111/j.1467-8721.2008.00569.x

Walther, D. B., & Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychological Science*, *25*(4), 851–60. http://doi.org/10.1177/0956797613512662

Wilmer, J. (2008). How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. *Spatial Vision*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2586597/

Wolfe, J., Võ, M., Evans, K., & Greene, M. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364661310002536

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. Retrieved from http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature06860.html

Figure Legends

Figure 1. One way for participants to recognize a scene would be to make use of global ensemble texture information, like the distribution of orientations and spatial frequencies, which has been shown to be computationally sufficient to recognize the spatial layout and category of a scene (e.g., Ross & Oliva, 2010); e.g., features like perspective, depth of view, and other spatial layout characteristics. For example, a scene can be transformed into only loosely localized information about its spatial frequency and orientation distribution, which can then be transformed into information about the 3D layout and category of the scene.

Figure 2. Methods for the 3 parts of Experiment 1. (A) In the spatial ensemble task, participants had to detect changes to a grid of gabor elements that appeared at an unexpected time for a brief duration while they performed another task (counting digits). The grid of gabors appeared briefly, then disappeared. When the display reappeared after a brief blank, it could sometimes be identical to before the blank (no change); or all of the individual gabor elements could have rotated by 45° (change trials). On every change trial, all of the individual gabor patches rotated by 45°, but on local-only trials (left), the way the elements rotated kept the ensemble structure the same (vertical on top, horizontal on bottom), whereas on local+ensemble trials (right), the 45° rotations changed the ensemble structure; for example, in the example in the figure, the top is now horizontal and the bottom vertical. The gabors in these example displays are larger and have higher contrast than the gabors used in the actual experiment. (B) In the rapid scene recognition task, participants saw a briefly flashed object on top of an irrelevant scene background (84ms) followed by a mask for 200ms. They then had to type the name of the object. On some trials, the scene background was informative because it was consistent with the object (left), whereas on other trials the scene background was uninformative (right). The difference between these conditions provides a selective measure of scene processing, as only the scenes differ between the conditions. (C) In the summary statistic task, participants saw a grid of 4 gabor elements for 1s and had to remember the average orientation of the set during a 1s delay and then report it by adjusting a gabor to match this average orientation using the mouse.

Figure 3. Main effects across all 50 participants for the (A) spatial ensemble task (d’ at detecting changes), (B) scene task (percent correct in recognizing objects), and (C) summary statistic/mean orientation task (standard deviation of participant’s reports, as estimated from the mixture model). Error bars represent within-participant standard errors of the mean.

Figure 4. Results. (A) Participants' performance for local+ensemble after controlling for their performance on local-only changes (ensemble benefit) was a strong predictor of their performance recognizing objects in informative scenes after controlling for their performance with uninformative scenes (scene benefit). The same participants who benefited most from ensemble changes in the spatial ensemble task with gabors were also the ones who benefited most from informative scenes. (B) The orientation summary statistic task, by contrast, did not significantly correlate with either the scene benefit or the ensemble benefit.

Figure 5. Methods of (A) Experiment 2A, (B) Experiment 2B. In both experiments, a grayscale prime scene or texture was presented, followed by a brief presentation of a grayscale object on a noise background, followed by a mask. Then participants had to type the name of the object they saw. In Experiment 2B, the prime was a texturized scene, designed to be unrecognizable but containing the same spatial distribution of orientations and spatial frequencies.

Figure 6. Results of (A) Experiment 2A, (B) Experiment 2B, (C) Experiment 3A, and (D) Experiment 3B. In Experiment 2A, there was a significant effect of the prime scene; participants performed better when the scene was informative. The same was true in Experiment 3A, where a prime object generated better performance when informative than uninformative. However, with texturized images, there was a major distinction between scenes and objects: In Experiment 2B, there was a significant effect of the prime texturized-scene, where people did better when the texture was generated from informative scenes than when it was generated from uninformative scenes. However, there was no benefit in Experiment 3B from informative texturized-object primes.

Figure 7. Methods of (A) Experiment 3A, (B) Experiment 3B. In both experiments, a grayscale prime object or texture was presented, followed by a brief presentation of an object on a noise background, followed by a mask. Then participants had to type the name of the object they saw on the texture background (the second object). In Experiment 3B, the prime was a texturized object, designed to be unrecognizable but containing the same spatial distribution of orientations and spatial frequencies.

Figure 8. (A) Scene images used in Experiment 2. (B) Texturized-versions of these scenes using our grid of orientation and spatial frequencies algorithm (based on Oliva & Torralba, 2006). (C) Texturized-versions of these scenes using a popular algorithm that assumes stationarity (homogeneity), by Portilla and SImoncelli (2000). You can see that the algorithm we use, which is considerably simpler and retains fewer image features than the Portilla and SImoncelli algorithm, nevertheless preserves spatial layout information better than the Portilla and Simoncelli algorithm because it does not assume spatial homogeneity across the image and is designed as a model of scene structure rather than explicitly as a model of visual texture.

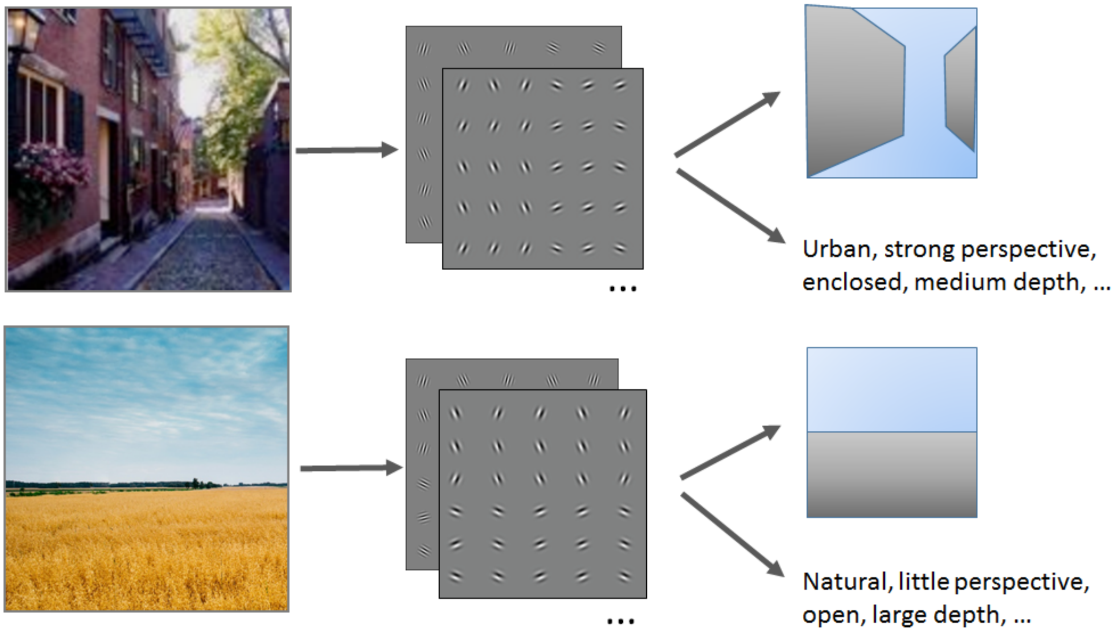
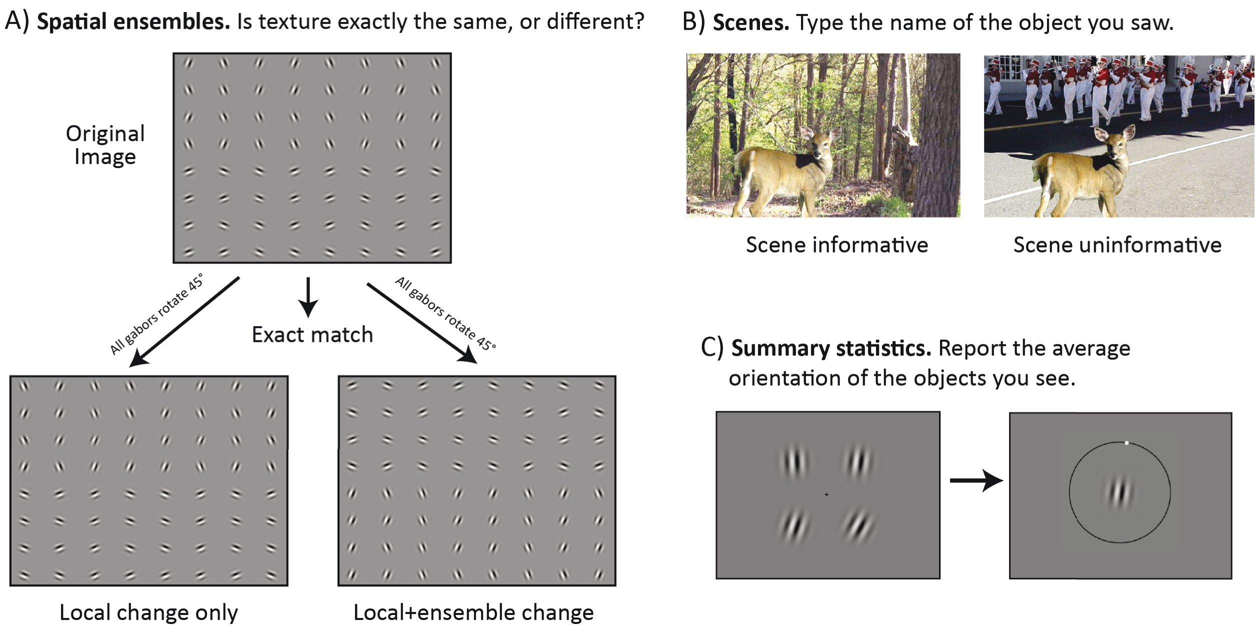

Urban, strong perspective,
enclosed, medium depth, ...
Natural, little perspective,
open, large depth, ...

Figure 1. One way for participants to recognize a scene would be to make use of global ensemble texture information, like the distribution of orientations and spatial frequencies, which has been shown to be computationally sufficient to recognize the spatial layout and category of a scene (e.g., Ross & Oliva, 2010); e.g., features like perspective, depth of view, and other spatial layout characteristics. For example, a scene can be transformed into only loosely localized information about its spatial frequency and orientation distribution, which can then be transformed into information about the 3D layout and category of the scene.


A) **Spatial ensembles.** Is texture exactly the same, or different?

B) **Scenes.** Type the name of the object you saw.

C) **Summary statistics.** Report the average orientation of the objects you see.

Figure 2. Methods for the 3 parts of Experiment 1. (A) In the spatial ensemble task, participants had to detect changes to a grid of gabor elements that appeared at an unexpected time for a brief duration while they performed another task (counting digits). The grid of gabors appeared briefly, then disappeared. When the display reappeared after a brief blank, it could sometimes be identical to before the blank (no change); or all of the individual gabor elements could have rotated by 45° (change trials). On every change trial, all of the individual gabor patches rotated by 45°, but on local-only trials (left), the way the elements rotated kept the ensemble structure the same (vertical on top, horizontal on bottom), whereas on local+ensemble trials (right), the 45° rotations changed the ensemble structure; for example, in the example in the figure, the top is now horizontal and the bottom vertical. The gabors in these example displays are larger and have higher contrast than the gabors used in the actual experiment. (B) In the rapid scene recognition task, participants saw a briefly flashed object on top of an irrelevant scene background (84ms) followed by a mask for 200ms. They then had to type the name of the object. On some trials, the scene background was informative because it was consistent with the object (left), whereas on other trials the scene background was uninformative (right). The difference between these conditions provides a selective measure of scene processing, as only the scenes differ between the conditions. (C) In the summary statistic task, participants saw a grid of 4 gabor elements for 1s and had to remember the average orientation of the set during a 1s delay and then report it by adjusting a gabor to match this average orientation using the mouse.

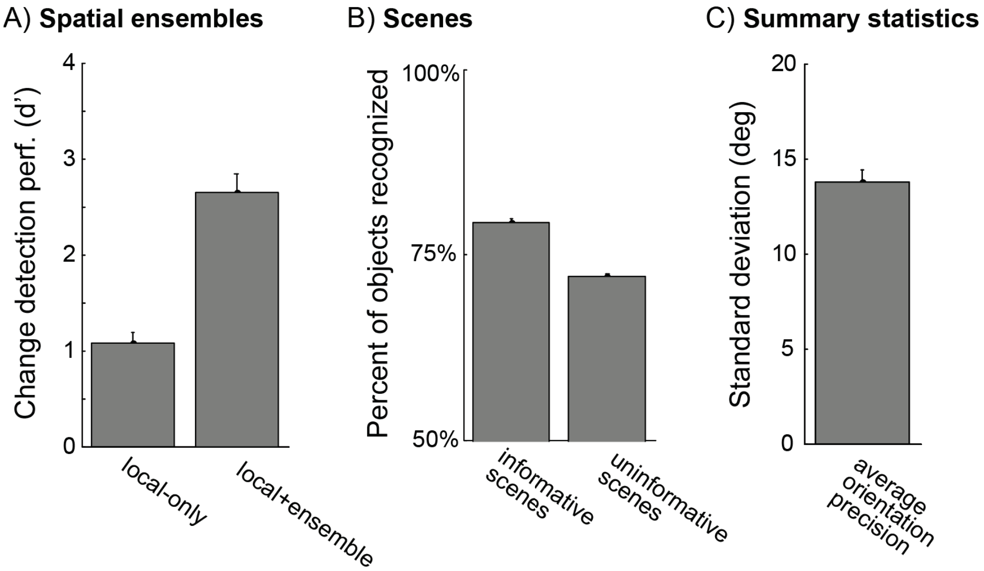


Figure 3. Main effects across all 50 participants for the (A) spatial ensemble task (d' at detecting changes), (B) scene task (percent correct in recognizing objects), and (C) summary statistic/mean orientation task (standard deviation of participant's reports, as estimated from the mixture model). Error bars represent within-participant standard errors of the mean.

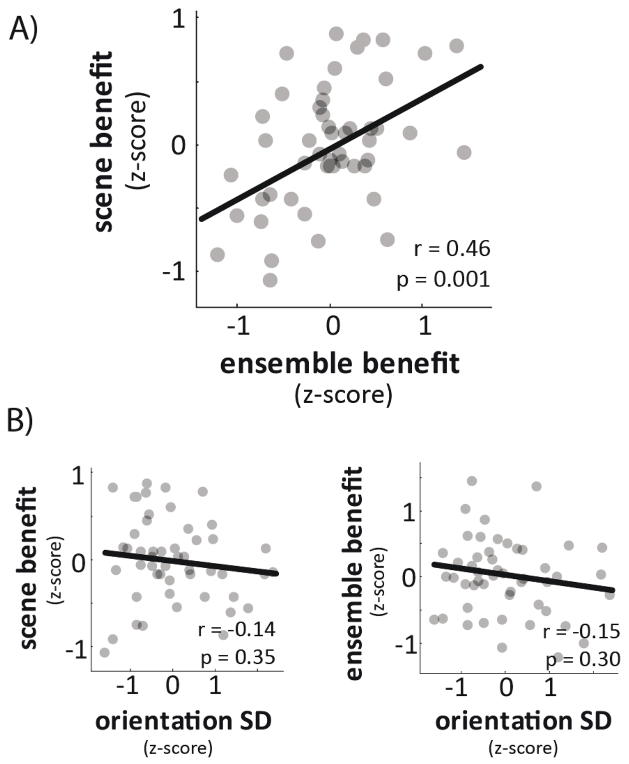


Figure 4. Results. (A) Participants' performance for local+ensemble after controlling for their performance on local-only changes (ensemble benefit) was a strong predictor of their performance recognizing objects in informative scenes after controlling for their performance with uninformative scenes (scene benefit). The same participants who benefited most from ensemble changes in the spatial ensemble task with gabors were also the ones who benefited most from informative scenes. (B) The orientation summary statistic task, by contrast, did not significantly correlate with either the scene benefit or the ensemble benefit.

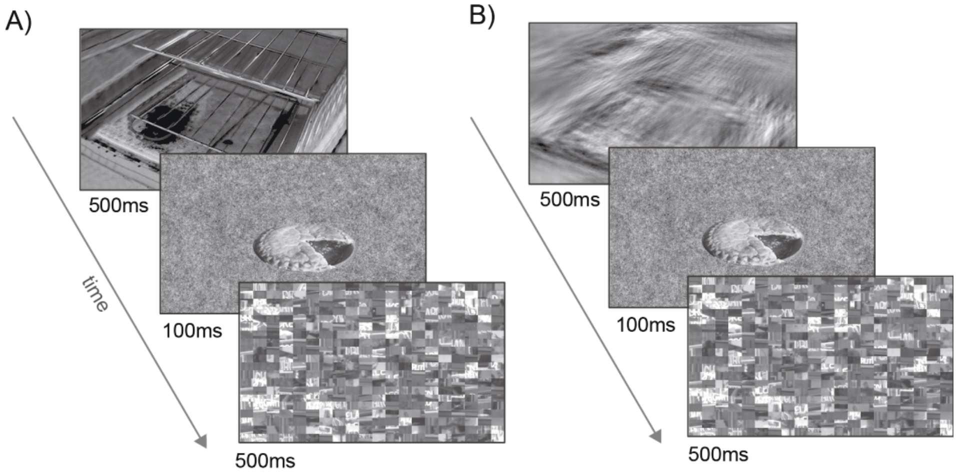


Figure 5. Methods of (A) Experiment 2A, (B) Experiment 2B. In both experiments, a grayscale prime scene or texture was presented, followed by a brief presentation of a grayscale object on a noise background, followed by a mask. Then participants had to type the name of the object they saw. In Experiment 2B, the prime was a texturized scene, designed to be unrecognizable but containing the same spatial distribution of orientations and spatial frequencies.

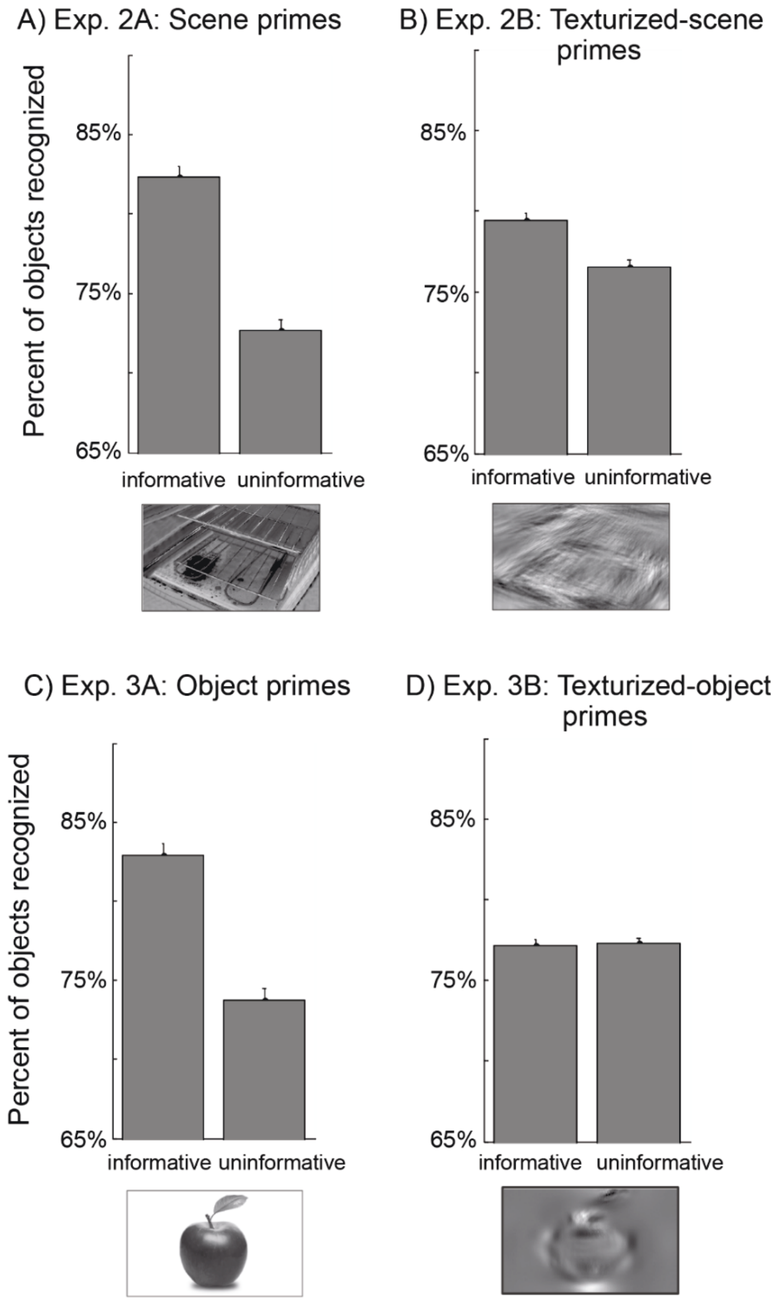


Figure 6. Results of (A) Experiment 2A, (B) Experiment 2B, (C) Experiment 3A, and (D) Experiment 3B. In Experiment 2A, there was a significant effect of the prime scene; participants performed better when the scene was informative. The same was true in Experiment 3A, where a prime object generated better performance when informative than uninformative. However, with texturized images, there was a major distinction between scenes and objects: In Experiment 2B, there was a significant effect of the prime texturized-scene, where people did better when the texture was generated from informative scenes than when it was generated from uninformative scenes. However, there was no benefit in Experiment 3B from informative texturized-object primes.

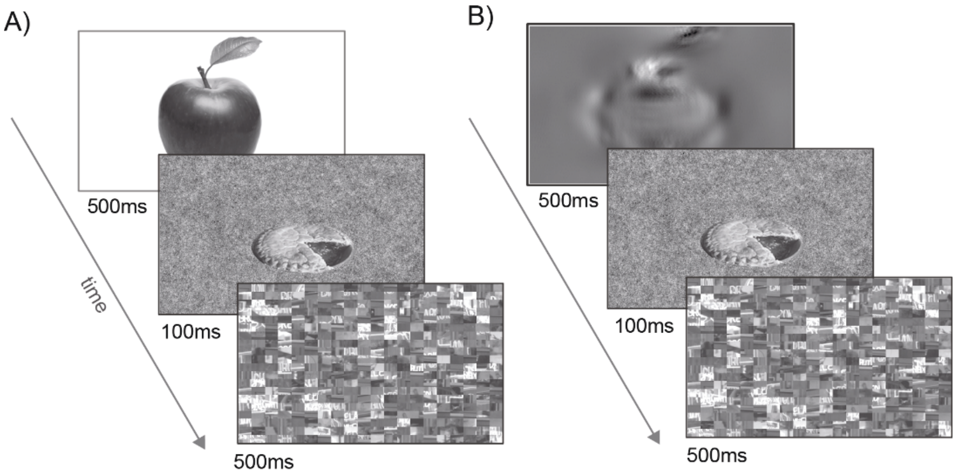


Figure 7. Methods of (A) Experiment 3A, (B) Experiment 3B. In both experiments, a grayscale prime object or texture was presented, followed by a brief presentation of an object on a noise background, followed by a mask. Then participants had to type the name of the object they saw on the texture background (the second object). In Experiment 3B, the prime was a texturized object, designed to be unrecognizable but containing the same spatial distribution of orientations and spatial frequencies.

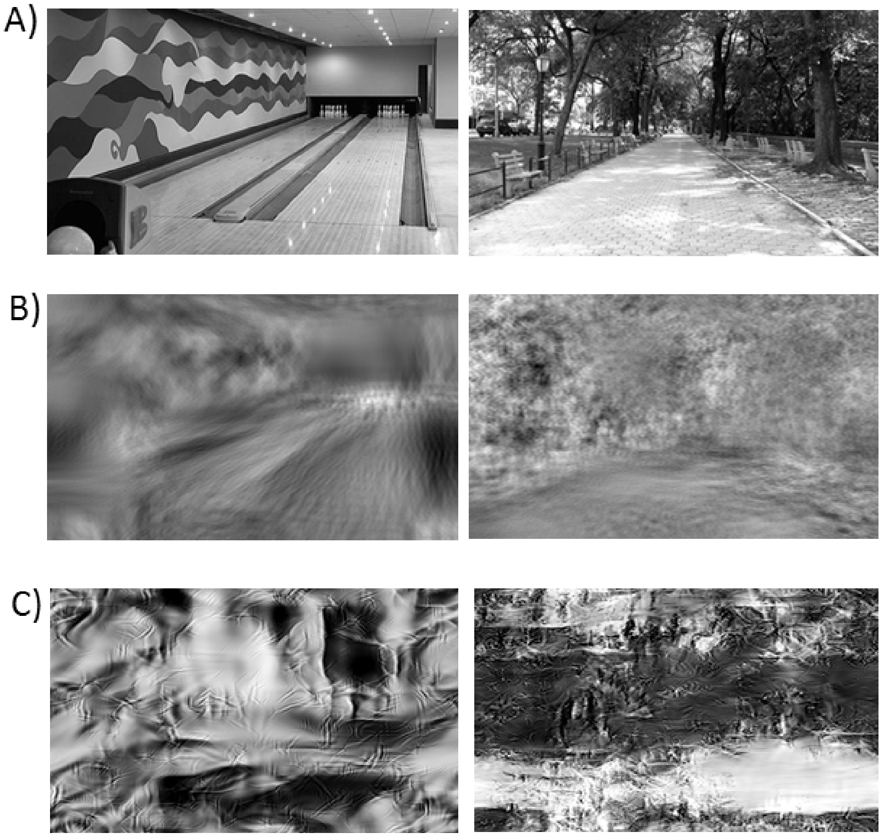


Figure 8. (A) Scene images used in Experiment 2. (B) Texturized-versions of these scenes using our grid of orientation and spatial frequencies algorithm (based on Oliva & Torralba, 2006). (C) Texturized-versions of these scenes using a popular algorithm that assumes stationarity (homogeneity), by Portilla and SImoncelli (2000). You can see that the algorithm we use, which is considerably simpler and retains fewer image features than the Portilla and SImoncelli algorithm, nevertheless preserves spatial layout information better than the Portilla and Simoncelli algorithm because it does not assume spatial homogeneity across the image and is designed as a model of scene structure rather than explicitly as a model of visual texture.