A general account of peripheral encoding also predicts scene perception performance

A general account of peripheral encoding also predicts scene perception performance

Krista A. Ehinger

Harvard Medical School and Brigham & Women's Hospital, Cambridge, MA, USA

Ruth Rosenholtz

CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA
Department of Brain & Cognitive Sciences, MIT, Cambridge, MA, USA

People are good at rapidly extracting the "gist" of a scene at a glance, meaning with a single fixation. It is generally presumed that this performance cannot be mediated by the same encoding that underlies tasks such as visual search, for which researchers have suggested that selective attention may be necessary to bind features from multiple preattentively computed feature maps. This has led to the suggestion that scenes might be special, perhaps utilizing an unlimited capacity channel, perhaps due to brain regions dedicated to this processing. Here we test whether a single encoding might instead underlie all of these tasks. In our study, participants performed various navigation-relevant scene perception tasks while fixating photographs of outdoor scenes. Participants answered questions about scene category, spatial layout, geographic location, or the presence of objects. We then asked whether an encoding model previously shown to predict performance in crowded object recognition and visual search might also underlie the performance on those tasks. We show that this model does a reasonably good job of predicting performance on these scene tasks, suggesting that scene tasks may not be so special; they may rely on the same underlying encoding as search and crowded object recognition. We also demonstrate that a number of alternative "models" of the information available in the periphery also do a reasonable job of predicting performance at the scene tasks, suggesting that scene tasks alone may not be ideal for distinguishing between models.

Introduction

Lab experiments have shown that scene recognition is extremely fast: In less than 100 ms, human observers can name the basic-level category of a scene (Oliva & Schyns, 1997; Rousselet, Joubert, & Fabre-Thorpe, 2005), detect if a scene shows an animal (S. Thorpe, Fize, & Marlot, 1996), and recognize scene attributes such as naturalness, openness, and navigability (Greene & Oliva, 2009; Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009). Some scene tasks, such as determining whether or not a scene contains an animal, can be performed in peripheral vision with minimal attention (Li, VanRullen, Koch, & Perona, 2002; S. J. Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001; VanRullen, Reddy, & Koch, 2004). (Note that the term "peripheral" is used inconsistently in the field. Throughout this paper, we use it to mean "outside the rod-free fovea," i.e., "extrafoveal.")

This rapid scene perception is problematic for many models of visual encoding that assume complex representations must be built up by serially attending to various objects and regions in a scene. Models based on visual search have suggested that the information available preattentively consists of only individual feature bands; in the absence of focused attention, one cannot bind these features to either spatial locations or other feature bands (A. M. Treisman & Gelade, 1980). If such models are correct, then the only way scene perception can be rapid and seemingly preattentive is if it relies solely on individual feature bands, such as color or orientation. Is this a possible explanation? For example, the orientation of edges in a scene provides information about its spatial layout: Perspective views of man-made spaces contain diagonal lines, and straight-on views contain mostly vertical and horizontal lines. Recognizing the navigability of a scene might just be a matter of distinguishing scenes with mostly diagonal edges (hallways, streets, etc.) from scenes with mostly horizontal/vertical edges (walls, building fa-

Citation: Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision*, *16*(2):13, 1–19, doi:10.1167/16.2.13.

doi: 10.1167/16.2.13 Received June 5, 2015; published November 18, 2016 ISSN 1534-7362

cades, etc.). However, if this were the case, then we would expect visual search for scenes to be easy: For example, a navigable scene should pop out from non-navigable scenes in a search display just like a diagonal line would pop out in a display of vertical lines. However, this is not the case (Greene & Wolfe, 2011). Other easy scene discrimination tasks, such as animal versus nonanimal, are similarly difficult when presented as search tasks, which suggests that these discriminations are not based on a single, low-level feature contrast (VanRullen et al., 2004). And although there are some simple features that correlate with "animalness" in the commonly used databases, these features alone do not explain performance on these rapid perception tasks (Wichmann, Drewes, Rosas, & Gegenfurtner, 2010).

It is generally agreed that individual feature bands are not sufficient to represent the gist of a scene (Rensink, 2001; A. Treisman, 2006; Wolfe, 2007). This has led to suggestions that scene processing is special and uses a separate pathway from visual search, perhaps subject to less restrictive capacity limitations and/or utilizing a different encoding of the visual input (e.g., Wolfe, Võ, Evans, & Greene, 2011).

But rather than assuming that scene-related tasks operate on a separate pathway with a different encoding than other tasks, such as visual search, it is more parsimonious to assume a common encoding underlies both types of tasks. We have previously argued (Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012) that difficult search may arise not from a need for serial attention to bind features, but rather from limitations of peripheral vision. Perhaps scene tasks are often easy and search tasks often difficult because of the information available in peripheral vision for those tasks. Scene tasks may rely primarily on information readily available in the periphery, and difficult visual search tasks may require information that becomes unreliable with distance from the point of fixation. In other words, scene perception may be special but not in the sense of using a different nonlimited capacity pathway.

Much of everyday scene perception takes place in peripheral vision while attention is engaged in another task. When navigating through the world, you are generally doing many visual tasks simultaneously: looking for a particular turn in the road, avoiding other pedestrians, reading a sign, maybe checking a map on your phone. Many of these tasks, especially those involving reading, require fixating and attending an object of interest. However, many navigational tasks, such as avoiding obstacles, can be done while focal attention is engaged elsewhere (Hyman, Sarb, & Wise-Swanson, 2014; Tractinsky & Shinar, 2008). Because the fovea occupies only a small portion of the visual field, many of these everyday scene perception tasks must take place primarily in the peripheral visual field.

Peripheral vision seems to play a particularly important role in navigation tasks. People are worse at navigating through real-world environments when their peripheral vision is blocked by blinders (Toet, Jansen, & Delleman, 2007, 2008) and worse at navigating virtual environments when given a limited field of view (Van Rheede, Kennard, & Hicks, 2010). Although there is some evidence that scene information can be processed more efficiently in central vision (Larson & Loschky, 2009), peripheral vision may have the advantage of parallel processing, allowing people to quickly process information about a large proportion of the space around themselves although with lower quality than would be obtained in central vision. This is vital for detecting obstacles and other hazards and maintaining a sense of where the observer is in the environment.

We have previously proposed a model of peripheral vision, known as the texture tiling model (TTM). This model is based on suggestions that peripheral vision encodes its inputs with a rich set of summary statistics (Rosenholtz, Huang, & Ehinger, 2012), pooled over local regions of an image. These summary statistics could probably be computed quite efficiently, allowing for rapid scene perception. However, this encoding would not necessarily provide a strong signal for visual search because pooling features over sizeable spatial regions might make it difficult for the visual system to distinguish between a peripheral patch containing the target and a number of other "distractor" items and a patch containing only distractors.

Similar models have previously been proposed explicitly to explain scene perception, utilizing different types of features and pooling schemes. For example, the GIST model (Oliva & Torralba, 2001) pools orientation information over large spatial regions of an image. This encoding predicts human performance on scene layout tasks (Ross & Oliva, 2010) and context-guided eye movements in visual search (Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009). However, more complex models that include more complex features and/or smaller pooling regions are more predictive of human performance on rapid perception tasks (Crouzet & Serre, 2011). In particular, the texture statistics of Portilla and Simoncelli (2000) are fairly good predictors of human performance on simple rapid scene perception tasks, and this is true when the statistics are both pooled over an entire image (Crouzet & Serre, 2011) and pooled within local regions that overlap and tile the image (Rosenholtz, Huang, & Ehinger, 2012).

The TTM explains performance on a number of nonscene tasks, such as crowded letter recognition (Balas, Nakano, & Rosenholtz, 2009; Keshvari &

Rosenholtz, 2016) and symbol recognition (Keshvari & Rosenholtz, 2016; Rosenholtz et al., 2012; Zhang, Huang, Yigit-Elliott, & Rosenholtz, 2015). It also correlates well with visual search performance (Rosenholtz, Huang, Raj, Balas, & Illie, 2012; Zhang et al., 2015). There is some evidence that TTM captures the information lost and maintained in early vision, possibly in area V2 (Freeman & Simoncelli, 2011; Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013). The model is (of course) by no means perfect. Wallis, Bethge, and Wichmann (2016) have demonstrated that matching Portilla and Simoncelli (2000) statistics over small regions in the periphery does not generate metamers indistinguishable from the original image. Alexander, Schmidt, and Zelinsky (2014) have shown that fixation patterns on search displays with model-synthesized targets and distractors are different than fixation patterns on original search displays, suggesting the model lacks some information available in peripheral vision. Finally, although model and peripheral performance are correlated in crowding and search tasks, there clearly remains variance not explained by the model.

In spite of TTM's imperfections, its previous performance predicting crowding and visual search tasks makes it a good candidate model for testing whether a single encoding might underlie all of these tasks (Rosenholtz, 2016). Several points are important to note: First, we are not saying that the same process identifies crowded objects, searches for a target, and gets the gist of a scene. Obviously, at some level, these tasks have their own underlying mechanisms. Rather, we test the possibility that all these tasks are subject to a single bottleneck and have available the information from a single encoding as opposed to one bottleneck/encoding for search and another for scenes. We suggest that the information that gets through that bottleneck governs which tasks are easy and which are hard. Second, clearly the simpler model of vision is one with one encoding rather than a different encoding for scenes. As a result, the bar is low for supporting a unified encoding in vision. We merely need at least one model to perform reasonably well on the tasks in question: visual crowding, visual search, and getting the gist of a scene. Even if the model leaves some of the variance unexplained and clearly does not capture exactly the information available in that encoding (for example, it does not match peripheral appearance), good performance across a range of tasks calls into question the significantly more complicated alternative model that requires different encodings for different tasks.

Here we asked whether TTM can predict performance on scene perception tasks. We started by gathering ground truth: We asked participants to perform a variety of scene perception tasks while fixating in scenes. We used a single fixation in each image in order to control the visual input across the periphery and masked the foveal portion of the image to ensure that participants could only use extrafoveal (peripheral and parafoveal) information to perform the task. We then compared the results to performance when free-viewing the scenes. We aimed to include a wide variety of scene tasks in order to study scene tasks with a range of difficulty. After establishing the range of performance on these tasks, we investigated whether the same model (TTM) that has previously shown promise at predicting search and recognition of crowded peripheral objects could also predict performance on these scene perception tasks.

Experiment 1

We asked people to perform a variety of scene perception tasks while fixating centrally on an image with the foveal portion of the image blocked by a mask. We used outdoor urban scenes as stimuli and looked at four broad types of tasks, shown in Figure 1: detecting an object in the periphery, identifying the general scene category, identifying the specific geographic location (e.g., New York or Paris), and describing the spatial layout.

We picked tasks that were expected to have a wide range of difficulty. Previous work has shown that basic-level scene category and spatial layout information (e.g., openness and navigability) can be identified in a brief glance when foveal information is also available (Greene & Oliva, 2009; Oliva & Torralba, 2001) as well as with only peripheral information (Boucart, Moroni, Szaffarczyk, & Tran, 2013; Boucart, Moroni, Thibaut, Szaffarczyk, & Greene, 2013; Larson, Freeman, Ringer, & Loschky, 2014; Larson & Loschky, 2009; Tran, Rambaud, Despretz, & Boucart, 2010). Object classification (e.g., "animal" or "dog") can be performed in a glance and in the periphery (S. Thorpe et al., 1996; S. J. Thorpe et al., 2001); however, these tasks typically use images in which the object to be recognized is fairly prominent. Detecting small objects in a multiobject display normally requires eye movements (A. M. Treisman & Gelade, 1980; Wolfe, 2007), so we might expect this task to be harder than other scene perception tasks.

It is not known whether people can identify the city depicted in a scene in a single glance. On the one hand, it seems like this task might require a detailed examination of the scene to look for characteristic styles of architecture, types of plants, signs, statues, and other objects in the scene. However, computer vision systems can predict the geographic location of an image using coarse global GIST features (Hays & Efros, 2008)

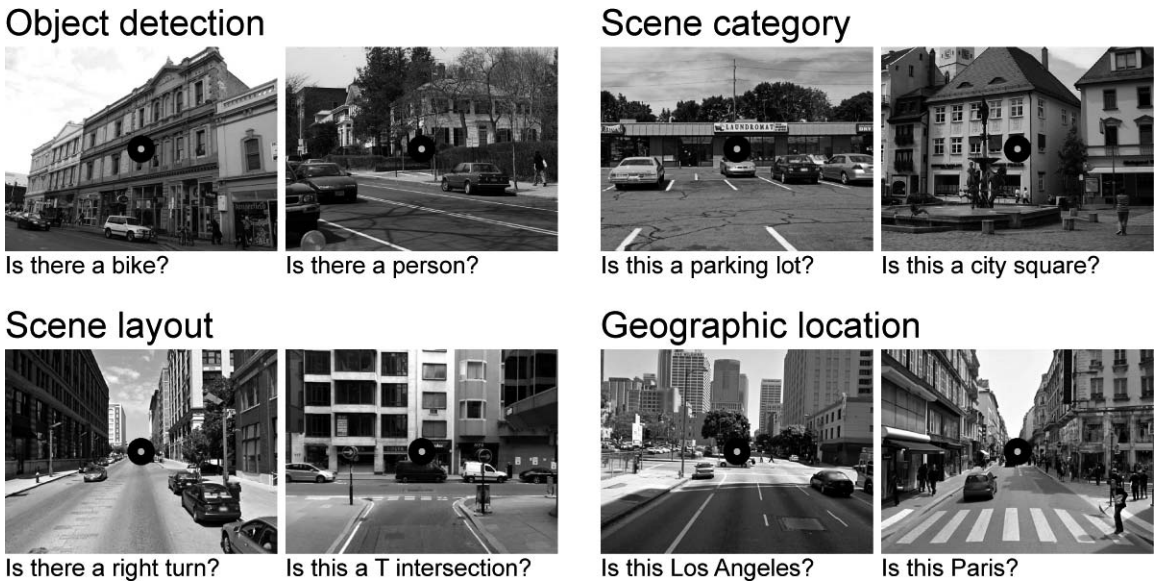

Figure 1. Examples of the types of scene perception tasks included in the experiment. An example of a target ("yes" response) image is shown with each question.

and automatically discover the regions and objects representative of different cities, such as Paris, from image patches (Doersch, Singh, Gupta, Sivic, & Efros, 2012). It has recently been argued that simple features computed over the whole visual field are suitable for recognizing specific locations (Eberhardt & Zetzsche, 2013; Eberhardt, Zetzsche, & Schill, 2016). If this is true, then it may be possible to do localization tasks in a single fixation on an image.

Methods

Participants

Twenty-four participants (16 female) were recruited from the Massachusetts Institute of Technology community. Participant age ranged from 19 to 51 (mean 28, *SD* 10); participants reported normal or corrected-to-normal vision. All participants gave informed consent and were paid to take part in the experiment.

Design

The experiment was a between-subjects design with half of the participants free-viewing the images and half of the participants required to maintain fixation. In each viewing condition, participants were asked the same set of 20 yes/no questions about natural scenes. Five questions were included from each of the four question groups: presence/absence of an object, scene category or gist, road layout, and geographic location. The order of the question blocks and the order of trials within each block were randomized for each participant.

Materials and apparatus

Stimuli consisted of 400 photos of urban environments. Each image appeared twice as a stimulus for two different tasks: The scene category stimuli were also used in the object detection tasks, and the geographic location stimuli were used for the layout tasks. The 200 images used as stimuli for the road layout and geographic location questions were collected from Google Street View, and the 200 images used as stimuli for the object presence and scene category questions were taken from the SUN database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010) or collected from the Internet. The images used in the object presence tasks were selected so that the target object appeared in only one location in the image. Additionally, object presence was counterbalanced with scene category, so that there were an equal number of target-present and target-absent trials in each scene category. Object size ranged from 0.5° to 7.3° on the diagonal (mean 2.6°, *SD* 1.3°) and object eccentricity (distance from the center of the image to the center of the object) ranged from 4.3° to 10.9° (mean 7.9°, *SD* 1.3°). The target scene categories ("downtown," "parking lot," "plaza," "residential neighborhood," and "shop front") were selected from the list of scene categories in the SUN database. Geographic location and road layout were also counterbalanced so that each road layout class appeared equally often in each city. For each of the scene category and layout tasks, foil images were drawn randomly from the other four categories. The foil

images for the "Is this Europe?" geographic location task were taken from Asia and North America, and the foils for the city location tasks were a random selection of other U.S./European cities and cities on other continents. Images were grayscale and 640 pixels wide by 480 pixels high in size. To ensure that participants could not use foveal information to perform the tasks, the center of each image was covered with a black circle 32 pixels (1° visual angle) in radius. Participants were able to use any information outside this central region to perform the task, including parafoveal information. Images were presented at 15° by 20° visual angle on a 34 cm by 60 cm monitor (Acer GD235HZ 23.6-in. LCD) with a resolution of 1920 by 1080 pixels and a refresh rate of 120 Hz. In the fixating viewing condition, participants were seated in a headrest with a viewing distance of 50 cm from the screen in a dim room, and eye position was tracked with an Eyelink 2000 eye-tracking system. In the free-viewing condition, participants did not use a headrest but were seated about 50 cm away from the screen.

Procedure

The 20 scene-perception questions were presented in blocks of 40 images per block. At the start of each block, participants were shown the question for that block (for example, "Is this London?") and two example images to illustrate the difference between the "yes" and "no" categories (e.g., a picture of London and a typical distractor scene). In the fixating condition, each trial was preceded by a central fixation cross, and the image appeared only after the participant was fixating the cross. Participants were required to maintain fixation on the center of the image, and whenever eye position moved more than 1° from the central position, the image was replaced with a uniform gray mask. Gaze position was tracked monocularly (right eye only) at 1000 Hz. Eye-tracker calibration was performed at the start of the experiment by having the participant fixate nine targets with a subsequent validation. Recalibration was done in between trials as needed. In both viewing conditions, image presentation time was unlimited, but participants were asked to respond as soon as they knew the answer to the question by pressing 1 ("yes") or 2 ("no") on a keyboard. Participants were not given feedback about whether or not their response was correct.

Results and discussion

Trials with a response time greater than 3 *SD* above the mean in each viewing condition were dropped from analysis (1.3% of trials in the fixating condition and 1.9% of trials in the free-viewing condition). In addition, one participant in the fixating condition reversed the response keys during one block, so this block was dropped from analysis. A scatterplot of the accuracy on each task when fixating or free-viewing is shown in Figure 2. Accuracy was averaged for each subject across all the images within each task. Accuracy in the free-viewing condition varied across the four task types: people were most accurate on the object detection tasks (average 92% correct) and scene category tasks (87%), less accurate on the spatial layout tasks (81%), and least accurate on the geographic location tasks (70%). A one-way, within-subject ANOVA showed a significant effect of task type, $F(3, 33) = 34.95$, $p < 0.01$, and post hoc Tukey honestly significant difference (HSD) tests showed a significant difference ($p < 0.01$) between each pair of tasks except for the object detection and spatial tasks, which were not significantly different from each other. There was also a significant effect of task type in the central fixation task (one-way, within-subject ANOVA), $F(3, 33) = 34.7$, $p < 0.01$. Post hoc Tukey HSD tests showed significant differences ($p < 0.01$) between each pair of task types except for the object detection and scene layout tasks, which were not significantly different. Participants in the central fixation condition were most accurate on the scene category tasks (average 84% correct), followed by the spatial layout (75%), object detection (72%), and geographic location tasks (65%).

In both viewing conditions, people were more accurate at the basic-level scene categorization tasks and less accurate at the more specific geographic location tasks: Recognizing a general scene category, such as "plaza" or "downtown street," was easier than distinguishing London streets from those in Rome or Tokyo. This parallels findings from rapid object categorization: Recognizing a general category ("animal") is easier and faster than recognizing more specific categories ("dogs" or "birds"; Mace, Joubert, Nespoulous, & Fabre-Thorpe, 2009; VanRullen & Thorpe, 2001). Similarly, superordinate scene category (man-made or natural) can be detected more quickly than a basic-level category, such as "mountain" (L. C. Loschky & Larson, 2010).

The difference between fixating and free-viewing performance on each individual task is also shown in Figure 2. In general, performance on these tasks was slightly higher when participants were allowed to make multiple fixations in the images: accuracy dropped about 20% on average in the object detection tasks and about 5% on average in the other scene perception tasks. A 2 (viewing condition, between-subjects) × 20 (task, within-subject) ANOVA with accuracy as the dependent measure showed a significant main effect of viewing condition, $F(1, 440) = 119.7$, $p < 0.01$; a significant main effect of task, $F(19, 440) = 118.3$, $p <$

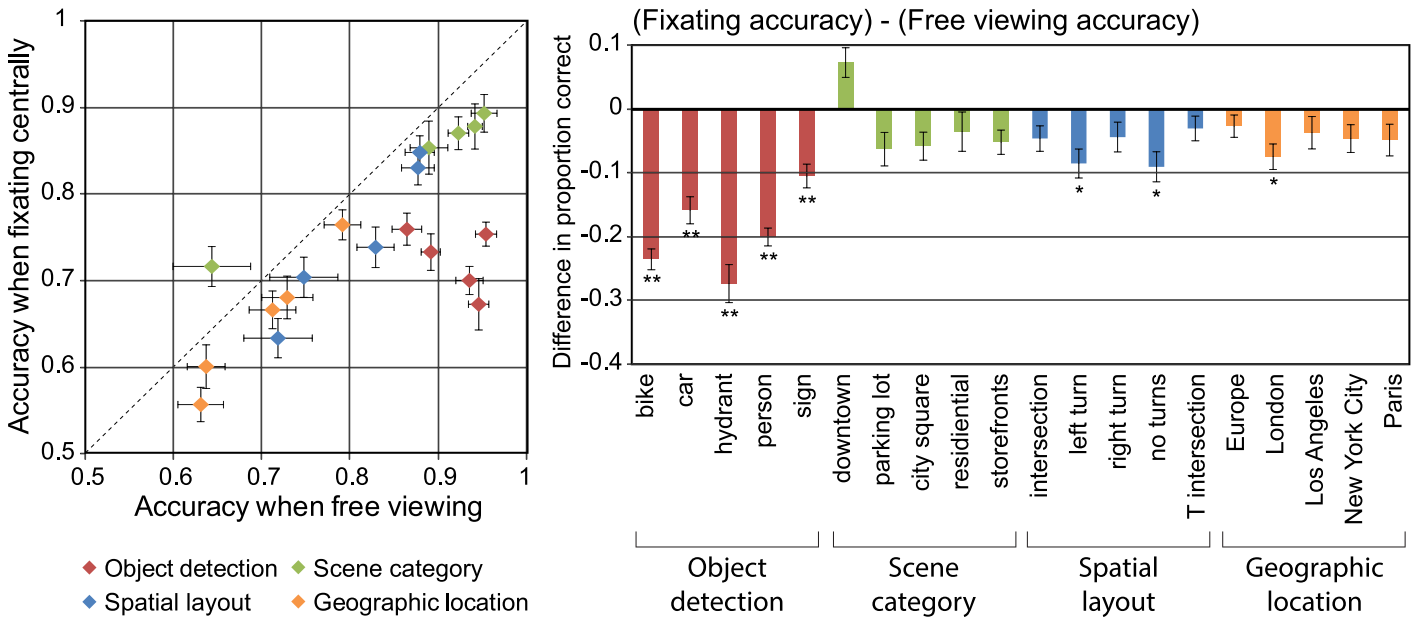

Figure 2. Scatterplot of performance on scene perception tasks when free-viewing versus fixating centrally in the image and a graph of the difference in accuracy for each task.

0.01; and a significant interaction, $F(19, 440) = 25.7$, $p < 0.01$. Bonferroni-corrected, one-sample t tests were used to determine whether the difference in performance was significantly different from zero. There was a significant drop in performance for all of the object detection tasks—people were less accurate at these tasks when they were required to fixate centrally than when free-viewing the images. However, performance on most of the scene perception tasks was not significantly different. The only tasks that showed a significant performance drop when fixating were "Is there a right turn only?," "Are there no turns?," and "Is this London?".

Response times across tasks in the two viewing conditions are shown in Figure 3. In both viewing conditions, responses were fastest in the scene categorization tasks and slowest in the geographic location tasks. In general, response times were faster in the fixating condition, which is somewhat surprising because this condition should have been more difficult, and the recorded response times include periods when the stimuli was masked due to attempted saccades. On

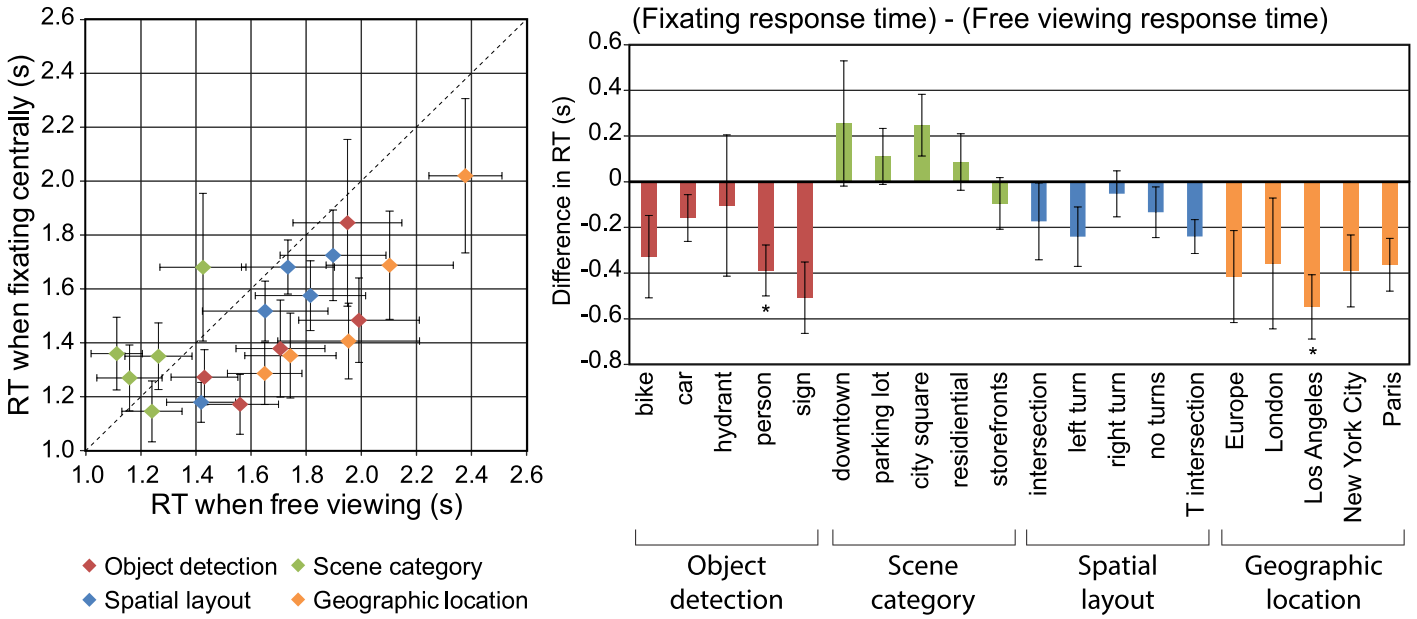

Figure 3. Comparison of response times for the scene tasks when fixating versus free-viewing.

the other hand, participants in the fixating condition may have responded more quickly because they were only allowed to use the information from a single fixation, and the free-viewing participants could explore the images more thoroughly to ensure that their responses were correct. Bonferroni-corrected paired t tests were used to compare fixating and free-viewing response times in the individual tasks; only the "Is there a person?" and "Is this Los Angeles?" tasks were significantly different ($p < 0.05$). Both tasks showed faster response times in the fixating condition than in the free-viewing condition.

It is perhaps not surprising that object detection tasks are more difficult when fixating centrally in an image. The target objects were all fairly small and eccentric, and we designed the tasks so that people could not guess object presence based on scene priors (for every target-present scene, we included a target-absent foil from the same scene category). The scene tasks that were more difficult when fixating may also have required detecting small, eccentric features in the images. As shown in Figure 1, the left- and right-turn images were often less distinctive than the other types of street layout: Sometimes a turn onto a minor side street would only be marked by a subtle gap in the buildings, which might not be easily detected without an eye movement. This may have led to more errors when people were asked to detect left turns only or detect which scenes had no turns. The "Is this London?" task may also have particularly benefited from eye movements that allowed people to read the text of signs or look for evidence of left- or right-drive traffic.

However, most of the scene perception tasks were not significantly more difficult when people were required to perform the tasks in peripheral vision without eye movements. This means that, for the most part, the features needed to recognize general scene categories, distinguish intersections from straight roads, or recognize specific geographic locations are readily available in peripheral vision. This by itself is an interesting result: Although we expected from previous studies that basic-level scene categorization and spatial layout classification would be easy in peripheral vision, it is interesting to note that even the finer-grained city-level classification is not much harder in a single fixation than when free-viewing scenes.

Now we return to the question of whether the same visual encoding might underlie performance at both these scene tasks and at visual search and crowded object recognition. Or does one need to postulate a separate channel for scene processing with different capacity limitations and different available information? In the next experiment, we tested the hypothesis that a single encoding scheme might operate for both sets of tasks by testing whether we could predict performance on the scene tasks using an encoding model that has previously shown promise at explaining a number of visual search results.

a. Original image b. Full-field mongrel

Figure 4. An example image from Experiment 1 and the corresponding mongrel version used in Experiment 2.

Experiment 2

To better understand how these images might be represented by the peripheral visual system, we used a paradigm based on Balas et al. (2009). We created images that captured, for each image, the information we hypothesized to be available in peripheral vision (essentially at a glance) then asked a second group of participants to do the same 20 scene perception tasks with these new, modified stimuli. An example of a modified "mongrel" image is shown in Figure 4. This encoding, which we call the TTM (Rosenholtz, Huang, & Ehinger, 2012), represents images in terms of texture statistics computed in pooling regions across the visual field.

We looked at how accuracy in classifying these modified images across the different tasks compared to the accuracy of fixating participants performing the same tasks in Experiment 1. To the extent that the hypothesized encoding captures the information lost and preserved by peripheral vision, performance free-viewing the synthesized images should predict performance when fixating the original images for a wide range of tasks and stimuli. A good fit between model predictions and performance but with variance unaccounted for would suggest a need for model improvement yet would still support a single encoding model. On the other hand, if scene perception requires additional information not available for other tasks, such as visual search (Rensink, 2001; A. Treisman, 2006; Wolfe, 2007), then the model should incorrectly predict poor performance for fixated scene tasks relative to free-viewing scene tasks. The plot of model prediction (y) versus scene task performance (x) would have a slope near zero rather than near one with little of the variance accounted for by the model.

We have previously demonstrated that our hypothesized encoding predicts difficulty for a number of

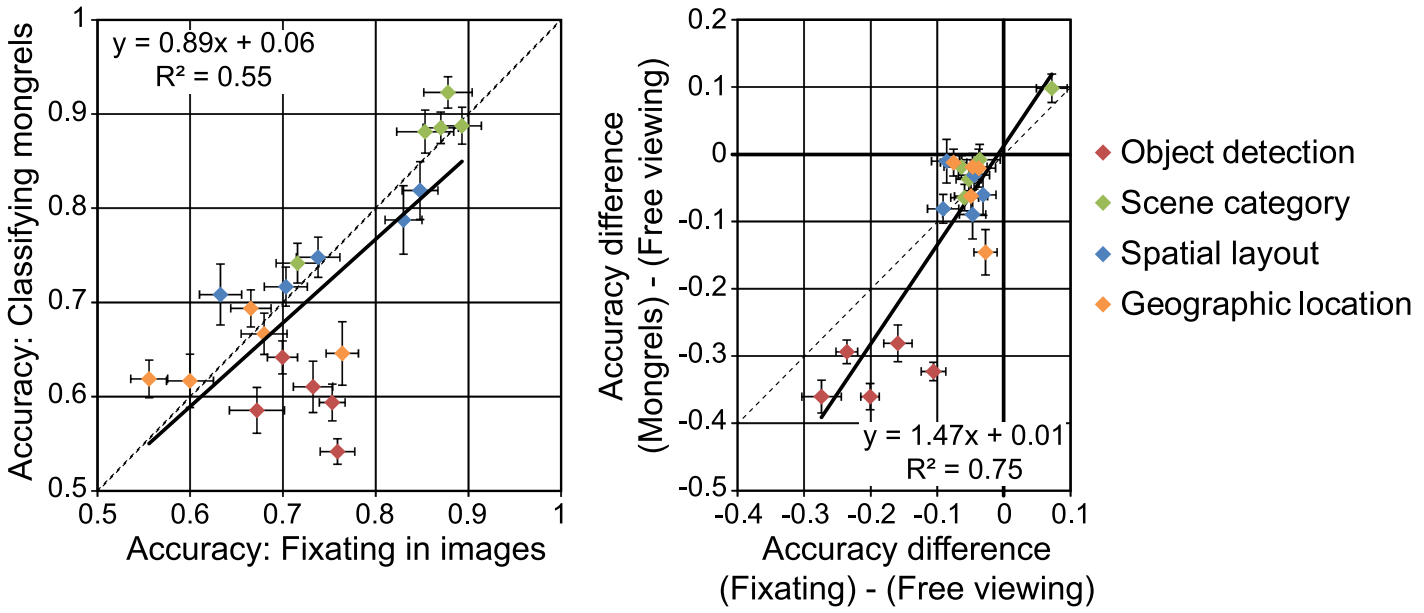

Figure 5. Model predictions (*y*-axis) versus performance. The left graph compares accuracy in the mongrel classification task to accuracy of the fixating participants from Experiment 1. The dotted line indicates $y = x$; if a model perfectly predicted fixating performance, all tasks would lie on this line. The solid line is the best fit from linear regression (slope, intercept, and R^2 are indicated on the graph). The right graph compares the drop in accuracy when performing these tasks with mongrels (vs. free-viewing the unmodified images, Experiment 1) to the drop in performance from fixating versus free-viewing. If the drop in accuracy when using mongrels was the same as the drop when fixating, points would fall on the dashed line, $y = x$. Points below the line indicate the conditions for which mongrel viewing impairs performance relative to the free-viewing condition more than the fixating does. The solid line is the best fit from linear regression (slope, intercept, and R^2 are indicated on the graph).

visual search tasks (Rosenholtz, Huang, Raj, Balas, & Illie, 2012; Zhang et al., 2015) as well as various crowded object recognition tasks (Balas et al., 2009; Keshvari & Rosenholtz, 2016; Zhang et al., 2015). The current experiment investigates whether the same model with the same set of features can also predict performance on these scene tasks.

Methods

Participants

Sixty participants took part in the image classification tasks on Amazon's Mechanical Turk service. No demographic data was collected from these participants. All of the individuals who participated in the Mechanical Turk task were located in the United States and had a good track record with the Mechanical Turk service (at least 100 HITs completed with an acceptance rate of 95% or better). All participants gave informed consent and were paid to take part in the experiment.

Design

Participants were asked the same 20 yes/no questions that had been presented in the central fixation task. Questions were presented in blocks of 40 trials with block and trial order randomized for each participant. Participants completed as many blocks as they wished (up to 20) and were able to quit the experiment at any point between blocks. Each question block was completed by 12 different participants.

Materials

For each of the stimuli images used in Experiment 1, we created a corresponding full-field mongrel by matching the texture statistics of a Gaussian noise image using the method described by Rosenholtz, Huang, and Ehinger (2012). The synthesis algorithm is as follows: Starting at a central fixation point, the algorithm tiles the image with square, overlapping pooling regions whose size increases with distance from fixation according to Bouma's law (Bouma, 1970). Within each pooling region, the model measures feature statistics from the original image and coerces the noise to have the same statistics using Portilla and Simoncelli's (2000) texture synthesis. Synthesis is initiated by assuming that the foveal region, a 1° radius around the fixation point, is reconstructed perfectly. Then, moving outward, each subsequent pooling region is synthesized using the previous partial synthesis result as the seed for the texture synthesis process. The lowest spatial frequency statistics are synthesized first, and

then higher spatial frequency information is added in a coarse-to-fine manner. The process iterates a number of times over the whole image. After each iteration, the foveal region and the border between the image and its background are reimposed on the output.

Procedure

Participants completed the classification tasks on their own computer, using a web interface on the Amazon Mechanical Turk website. Participants were told that the purpose of the study was to determine how well people could recognize images "distorted by digital noise" and were shown examples of images with their corresponding full-field mongrels. Participants were told they would answer yes/no questions about scenes. To discourage self-selection, participants were not shown the specific question they would answer in that block until after they pressed a button to start the block. In each block, participants were given a single question (for example, "Is this London?") and were shown mongrel versions of the 40 images that had been used as stimuli for that question in Experiment 1. Participants were allowed to study each image for as long as they wished and then clicked one of two buttons beneath the image to indicate "yes" or "no." Participants received feedback after each response.

Results and discussion

If a participant quit a block partway through, those trials were recorded but dropped from analysis, and the block was rerun with another participant. There were also a few cases in which a trial was recorded multiple times due to a browser issue; these duplicates were also dropped. A total of 324 trials were dropped, leaving 9,600 trials.

Because participants were not required to complete all 20 questions in the online classification tasks, we were worried that they might opt out of blocks they found particularly difficult, leaving those tasks to be completed by participants with more expertise. This could potentially be an issue for the geographic location tasks: For example, the "Is this New York?" task might have attracted a disproportionate number of New Yorkers if other participants dropped out of this task. However, this behavior was not very common. We marked blocks as "dropout" blocks if a participant completed 25% or more of the block but opted not to submit their results. There were only five dropout blocks in the mongrel classification task. Four of these were object detection blocks (e.g., "Is there a car?"), and one was a geographic location task ("Is this Los Angeles?"). It should be noted that we cannot say why participants abandoned these tasks; they may have quit because they found the tasks particularly difficult, or they may have simply gotten bored or distracted or run into technical problems. Regardless, we do not think this low rate of dropout would have significantly affected the results.

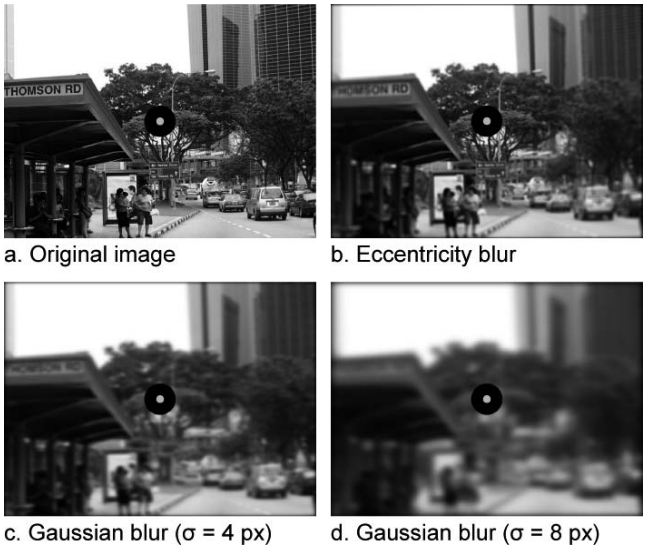

Figure 6. (a) An example stimulus image from Experiment 1. (b–d) Blurred versions of this image used in Experiment 3.

For each classification task, we compared accuracy in the mongrel classification task to the accuracy of participants fixating in the same images (Figure 5, left graph). A linear regression across tasks shows that accuracy with the modified images is generally similar to the accuracy when free-viewing for most of these tasks. The object detection tasks are the exception: These tasks were more difficult for participants in the mongrel classification task than they were for participants fixating in the original images.

With this analysis, it is not clear whether the model is actually predicting fixating performance or just the baseline (free-viewing) difficulty of these tasks. Most of these tasks are not much harder in the periphery, which means any model that can predict free-viewing performance will be a good predictor of the fixating performance. To address this issue, we also compared the average performance with the modified images to the average drop in performance on each task: the difference in accuracy between fixating and free-viewing participants. This is the difference in performance predicted by the type of information loss represented by our model visual crowding. This predicted difference was compared to the actual difference between fixating and free-viewing performance that we observed in Experiment 1. Scatterplots of the actual versus predicted difference by task is shown in Figure 5 (right graph). Linear regression was performed to determine how well the model predicted the difference between fixating and free-viewing across the range of scene perception tasks; this fit explained 75% of the variance across tasks. This

suggests the TTM can explain a good portion of the difficulty of these scene perception tasks although the slope of the linear fit is not quite one. This seems to be due primarily to the object detection tasks: the TTM overestimates how difficult these tasks should be when fixating.

The difference between the mongrel classification and free-viewing performance was generally similar in magnitude to the difference between the fixating and free-viewing conditions in Experiment 1. Tasks that were easy when fixating were about equally easy when free-viewing mongrel images, and tasks that were more difficult to perform when fixating were similarly difficult with the mongrels. The mongrel images simulate the effects of crowding in the peripheral visual field, so the similar drop in accuracy suggests that scene perception when fixating may be partially explained by crowding.

It is impossible to *prove* that an encoding model is correct simply by showing that it can predict behavioral results. One can only gather evidence in support of the model. If one tests a model on only a handful of tasks, one runs the risk that those tasks may not discriminate well between models, a point we return to in the next experiment. The local statistics encoded in our model have been previously shown to predict performance in other peripheral vision tasks, including visual search and crowding (Balas et al., 2009; Rosenholtz, Huang, Raj, Balas, & Illie, 2012). Combined with these previous findings, the current results from these scene perception tasks provide further evidence that the peripheral visual field uses a statistical summary encoding, pooling visual features over local regions of the peripheral visual field. This pooled, summary information is sufficient for many scene perception tasks, such as determining whether a scene is a residential or city street, whether a road turns left or right, or whether an image depicts an American or European city.

Experiment 3

As a control, we also wanted to know how well these scene tasks alone could discriminate between models of peripheral vision. To test this, we degraded the scene images in ways that differed significantly from the information loss modeled in Experiment 2 and asked to what extent those image degradations predicted performance on the scene tasks. There are many different types of noise that could be used to degrade images, but for simplicity, we chose blur. If performance with these other image degradations could also predict performance on the scene tasks, we should be wary of choosing a model of peripheral vision based on these tasks alone.

We tested three image degradations. First, visual acuity falls off with eccentricity in the periphery (Anstis, 1974). We could mimic this acuity loss by applying an eccentricity-based blur to our original images (Figure 6b). We also generated two sets of uniformly blurred stimuli with mild ($\sigma = 4$ pixels $= 2$ c/°, Figure 6c) or moderate ($\sigma = 8$ pixels $= 1$ c/°, Figure 6d) Gaussian blur.

To be clear, this was *not* a test of which model of peripheral vision is the best. We compared performance with the blurry image degradations to our model to get a sense of how well the scene tasks discriminated between these models. However, even if one of these image degradations better predicted performance on the scene tasks than our summary statistic model, we could not conclude that the "model" represented by that image degradation is a better general-purpose peripheral encoding model than the TTM. None of the tested blurs represented a viable model of peripheral encoding. None of the blurs could explain crowding (Lettvin, 1976). The two uniform blurs could not explain peripheral acuity experiments (Anstis, 1974; L. Loschky, McConkie, Yang, & Miller, 2005) because the blur is too high and not eccentricity-dependent.

Methods

Except as noted below, all methods were the same as in Experiment 2.

Participants

A total of 197 participants took part in the image classifications tasks on Amazon's Mechanical Turk service. None of these participants had taken part in the mongrel classification task.

Design

The three image conditions were run separately in a between-subjects design, so each participant only saw images from one of the three conditions.

Materials

We created uniformly blurred versions of our stimuli by filtering the images with a Gaussian filter with $\sigma = 4$ pixels (2 c/° blur) or 8 pixels (1 c/° blur). The blur-with-eccentricity images were created by convolving each pixel in the image with an averaging disk filter of radius $1 + 0.43 \cdot ecc$ pixels (or $0.03 + 0.0134 \cdot ecc°$), where *ecc* is eccentricity of the pixel in degrees. This radius is approximately one third of the threshold letter height reported by Anstis (1974). The same falloff with eccentricity would be obtained using the frequency

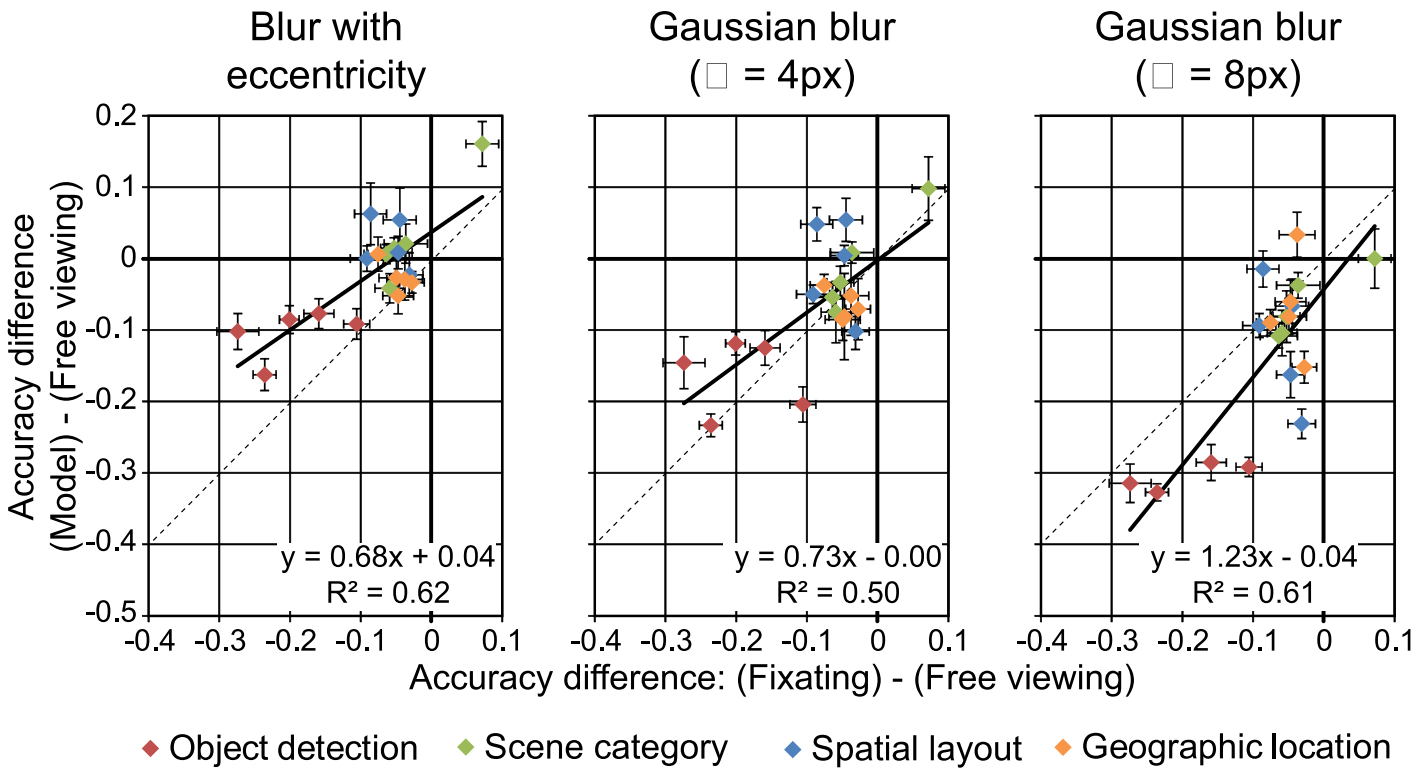

Figure 7. Model predictions (*y*-axis) versus performance. Graphs compare the drop in accuracy when performing these tasks with blurred images versus the drop in performance from fixating versus free-viewing. The dotted line indicates $y = x$; if a model perfectly predicted fixating performance, all tasks would lie on this line. The solid line is the best fit from linear regression (slope, intercept, and R^2 are indicated on the graph).

cutoff formula of L. Loschky et al. (2005):

$$f_{\text{cut}} = 43.1 \frac{E2}{E2 + ecc} \quad (1)$$

E2 is the eccentricity at which the resolution is halved, so higher values denote less blur; we assume E2 = 1.73. This is a lower value of E2 than suggested in Miller et al. (E2 = 3.11), but various papers have found a range of best-matched values for E2: Geisler and Perry (1998) use E2 = 2.3; Abdelnour and Kalloniatis (2001) use E2 = 2.5. Because we thought that eccentricity blur alone would probably underpredict the information loss in the periphery, we made a point of erring on the side of too much blur rather than too little. Using E2 = 1.73 gives a level of blur that should be just barely detectable when fixating centrally in the images. Examples of the stimuli from the three blur conditions are shown in Figure 6.

Results and discussion

As in Experiment 2, we dropped trials from incomplete blocks and trials that were duplicated due to recording issues. There were 250 dropped trials in the 8-pixel blur condition and 173 dropped trials in each of the other blur conditions, leaving 9,600 trials in each condition. As in Experiment 2, we looked at how often participants abandoned blocks that were partially (at least 25%) complete to determine whether participants might be opting out of tasks they found particularly difficult. This was most common in the 8-pixel blur condition (10 blocks abandoned) and less common in the 4-pixel blur and blur-with-eccentricity conditions (four and five blocks abandoned, respectively). A plurality of the abandoned blocks (five) in the 8-pixel blur condition involved the object detection tasks.

For each classification task, we compared the average performance of participants viewing the modified images to the average free-viewing performance on that task from Experiment 1. This is the difference in performance predicted by the type of information loss represented in each of the models (eccentricity blur or one of the two types of uniform blur). This predicted difference was compared to the actual difference between the fixating and free-viewing performance that we observed in Experiment 1. Scatterplots of the actual versus predicted difference by task for the three models is shown in Figure 7. Linear regression was performed for each model to determine how well it predicted the difference between fixating

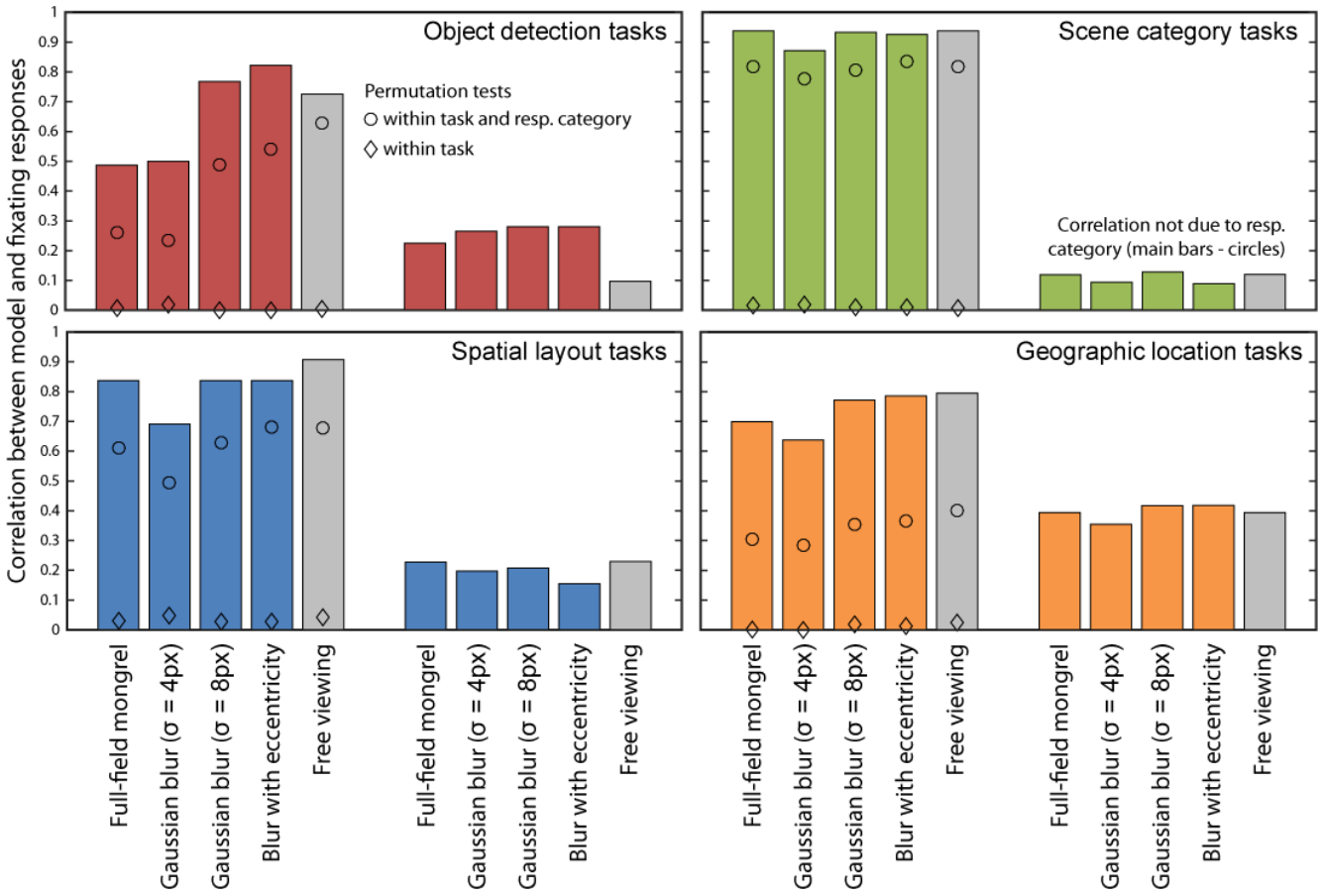

Figure 8. Correlations between fixating participants' responses and models to the individual images used in these tasks. The left-hand set of bars shows overall correlations; the circles and diamonds indicate baseline correlations obtained from permuting responses to the images within each task or within each task and ground-truth response condition ("yes" or "no"). Ninety-five percent confidence intervals were computed but are too small to show in this figure. The right-hand bars show the difference between the overall correlation and the latter control and indicate how well each model captures the variation in responses, which is not simply due to the model getting the correct ground-truth response.

and free-viewing across the range of scene perception tasks.

The images that were blurred to simulate the falloff in visual acuity with eccentricity are the best linear fit to the accuracy difference observed between the fixating and free-viewing conditions in Experiment 1 ($R^2 = 0.62$). However, this is lower than the fit to our model of crowding (TTM) from Experiment 2 ($R^2 = 0.75$). In addition, most of the points lie above the diagonal, which means that people were generally better at performing these scene tasks when free-viewing eccentricity-blurred images than when fixating the original images. This suggests that the falloff in acuity over the peripheral visual field does not completely explain the difference between fixating and free-viewing performance; there is some additional information loss in the periphery, and we can see the effects of that loss in performance of the scene tasks.

The two uniform blur conditions were poorer fits to the accuracy difference observed in Experiment 1. Performance with the moderately blurred (8-pixel blur) images was generally worse than fixating performance, and the linear fit was similar to the eccentricity-blurred images ($R^2 = 0.61$). Performance with the less blurred (4-pixel blur) images was more similar to fixating performance on average, but the linear fit was worse ($R^2 = 0.50$). This suggests that some of the fixating performance on these scene perception tasks may be explained by low-resolution information in the periphery. However, models that included some higher spatial frequency features (the eccentricity blur and full-field mongrels) fit the data better than these low-resolution models.

Finally, it is worth asking how well each model can predict responses to the individual images used in the scene perception tasks. To investigate this, we used the method from Crouzet and Serre (2011) to look at the correlations between model responses and the responses of participants in the fixating condition of Experiment 1. The results are shown in Figure 8. The

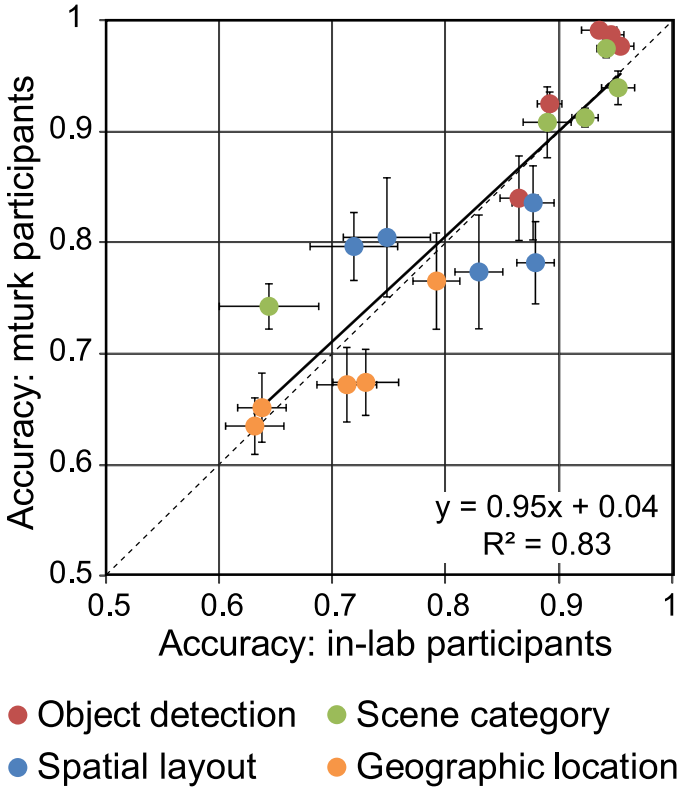

Figure 9. Accuracy of Mechanical Turk participants compared to in-lab participants (Experiment 1). Both groups are performing the tasks while free-viewing.

percentage of fixating participants who said "yes" to a given image was compared to the percentage of workers who said "yes" to the modified version of that image in the classification tasks. We also show the correlation between fixating and free-viewing participants (Experiment 1) in each task. Following Crouzet and Serre (2011), we compared overall correlations to two baselines: a standard permutation test in which the responses were randomly shuffled within each task and a restricted permutation test in which the responses were randomly shuffled within each task and ground-truth response category (e.g., just within the ground-truth "no" images for the "Is there a car?" task). The latter baseline indicates how much of the overall correlation is due to a model getting the ground-truth response correct. The differences between the overall correlation and the within-response baseline (shown separately in Figure 8) show how well each model captures the variation in image difficulty, separate from how well the model could classify the images.

The left sets of bars in each part of Figure 8 show the overall correlations between fixating participants and models. Fixating responses correlated most highly with free-viewing responses and with the two models that had the least image distortion (blur with eccentricity and 8-pixel Gaussian blur). This reflects the fact that most of these tasks can be performed nearly as well in the periphery as when free-viewing: Responses when fixating should be well correlated to responses when free-viewing the original (or very slightly distorted) images. However, most of this correlation is due to the models getting the ground-truth response correct: Fixating participants were generally accurate at most of these tasks as were free-viewing participants and the less-blurred models. More important is the comparison between the overall correlation and the permutation test (right sets of bars), which shows how well each model captures the range of difficulty of the individual images. Most of the models do about equally well when compared in this fashion; the best correlating model varies across tasks. For most tasks, free-viewing performance is well correlated with fixating performance even after controlling for ground-truth correct responses. The object detection tasks are the exception, but this is because the free-viewing responses were nearly at ceiling. This suggests that this type of correlation analysis may not be ideal for distinguishing between models that preserve too much image information because fixating responses are generally highly correlated to free-viewing responses.

Finally, it should be reiterated that blur alone is not a viable model of peripheral vision because it would not explain performance on a range of tasks. For example, difficulty identifying a crowded peripheral target compared to an unflanked target cannot be explained by blur or loss of acuity (Lettvin, 1976). However, the

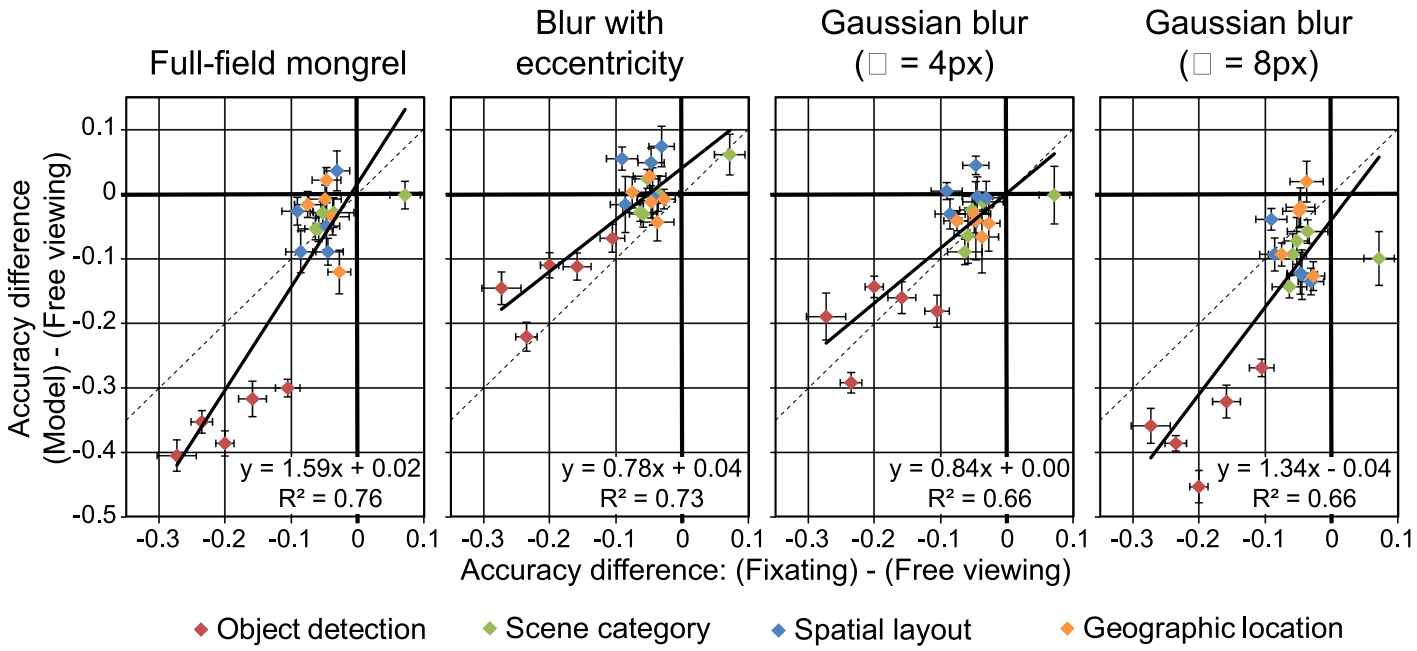

Figure 10. Model comparisons using the free-viewing accuracy of Mechanical Turk participants as the baseline for the models.

fact that in the present work several different levels of blur as well as TTM did reasonably well at predicting scene task performance and the fact that many scene tasks cluster at the easy end of the spectrum—they can be performed very rapidly in a single fixation—means they may not be the best tasks for discriminating between models of peripheral vision. Although our scene tasks covered a range of difficulty, most of them were not much more difficult when fixating than when free-viewing. This makes it difficult to determine exactly what features are preserved in peripheral vision because many of these tasks may only require a very coarse representation of the image. However, this issue is not unique to scene perception tasks. Most visual tasks probably do not require all of the information available to the visual system, which is why it is imperative to test any model of vision on a wide range of tasks.

Experiment 4

One concern with the design of Experiments 2 and 3 is that the image classification tasks were run online using a different participant pool from the one used in the lab-based fixation and free-viewing tasks. Previous work has demonstrated that participants in online experiments on Amazon Mechanical Turk generally give the same performance as participants run in the lab despite the differences in experiment setting and differences in the average demographics of these groups (Crump, McDonnell, & Gureckis, 2013). However, because we are using image classification data from an online task to predict in-lab performance, it is particularly important for our study to know that the online and in-lab participants have similar performance on these scene tasks. If the online participants have a very different baseline performance on these tasks—for example, because they are more likely to guess randomly or more likely to be experts at recognizing specific cities—then their performance on the image classification tasks is less useful for predicting the information loss in peripheral vision. To compare baseline performance, we asked Mechanical Turk participants to perform the free-viewing classification tasks from Experiment 1, using the same online interface as had been used in Experiments 2 and 3.

Methods

The methods were the same as in Experiments 2 and 3, but instead of degraded images, participants performed the yes/no classification tasks on the original, unmodified images used in Experiment 1. As in Experiment 1, the center of each image was covered by a black circle 32 pixels in radius, and participants were not given feedback after each response. A total of 103 participants took part in the tasks on Amazon's Mechanical Turk service. None of these participants had taken part in the other classification tasks, and this experiment was run after we had finished data collection for those tasks in order to ensure that there was no opportunity for participants in the degraded

image conditions to see any of the original, unaltered images.

Results and discussion

As in Experiments 2 and 3, we dropped trials from incomplete blocks and trials that were duplicated due to recording issues. In total, 361 trials were dropped, leaving 9,600 trials. To determine whether participants were opting out of certain tasks, we looked at how often blocks were partially (at least 25%) completed and then abandoned. There were 12 such blocks in Experiment 4: one was an object-detection task ("Is there a fire hydrant?"), three were scene categorization tasks, four were spatial layout tasks, and four were geographic location tasks. The "Is this Paris?" task was the only individual task abandoned by more than one participant; three of the 12 dropped blocks that involved this question.

Overall, the performance of the Mechanical Turk participants was very similar to the performance obtained in the lab. An accuracy comparison with the best-fit linear regression is shown in Figure 9. The correlation between the groups on these scene tasks is 0.91, $R^2 = 0.83$. The intercept of the regression is near zero, which suggests that there is no great difference in the mean performance of the two participant pools; Mechanical Turk participants aren't systematically underperforming compared to the lab-based participant pool. The slope of the regression is near one, which suggests that there is no difference between the groups across harder or easier tasks as might be the case if Mechanical Turk participants were opting out of these tasks or guessing more frequently in the harder tasks.

Because the performance on free-viewing tasks is similar, it seems unlikely that the results in Experiments 2 and 3 are driven by differences between the Mechanical Turk and lab-based participant pools. We can also use the Mechanical Turk free-viewing performance as an alternate "ground-truth" baseline for the image classification tasks in Experiments 2 and 3; rather than subtracting the lab participants' free-viewing accuracy from the mongrel or blurred image classification, we can subtract the Mechanical Turk participants' average free-viewing accuracy. This may give a better estimate of the cost of each image manipulation if the small performance differences between the online and lab-based participants are actually due to systematic differences between these groups and not just random noise. Comparisons using this alternate baseline are shown in Figure 10. In general, the results are similar, but the linear fits are better, particularly for the 4-pixel blur and blur-with-eccentricity images.

This does not mean that uniform Gaussian blur is a correct model for the peripheral visual field; that would be inconsistent with many previous results. However, it does suggest that many scene tasks can be performed with only the low spatial frequency portions of the image (as previously shown by Schyns & Oliva, 1994).

General discussion

We compared fixating and free-viewing performance on a range of scene perception tasks: detecting objects, recognizing scene categories and spatial layout, and identifying specific geographic locations. For many of these tasks, accuracy was not significantly different when fixating centrally or free-viewing the images, consistent with previous work showing that people are able to accurately perform a variety of scene perception tasks in a single fixation on a briefly presented image. Only the object detection tasks were consistently more difficult when fixating relative to free-viewing. Because the fovea occupies only a small percentage of the visual field, much of the visual processing that occurs during a single fixation on a scene must occur in the parafovea and periphery.

It is quite likely that only a limited set of visual features are necessary to perform some of the scene tasks tested. For example, the GIST model uses only a subset of the statistics in the TTM (Balas et al., 2009; Freeman & Simoncelli, 2011; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Illie, 2012), computed over very large pooling regions, but this is sufficient to recognize basic-level scene categories and spatial layout (Oliva & Torralba, 2001; Ross & Oliva, 2010). Because scene perception may be accomplished with only a subset of the features available in the periphery, these tasks alone are not ideal for determining exactly what features are available to extrafoveal vision. However, the goal of these experiments is not to find the bare minimum set of features required for each individual scene perception task, but to test whether a single model of peripheral vision can explain performance on a wide range of tasks, including scene perception, crowding, and visual search. Because each visual task may be accomplished with a different subset of features, no single type of task is ideal for testing models of peripheral vision; a general model must be validated with a range of tasks.

Additionally, we cannot use these results to dismiss the role of attention in scene perception tasks. Previous work seems to rule out a model of scene perception based on serially attending to each object in the scene in favor of a more holistic, global process for tasks such as rapid scene categorization. However, responses in our tasks were relatively slow, and even in the fixating

condition, people may have been able to deploy covert attention to multiple regions of the image. In particular, people may have tried to use covert attention to search for the targets in the object detection task although covert search would still be constrained by the information loss in peripheral vision.

These scene perception results, when combined with previous work on crowding and visual search, provide support for a summary statistic encoding in peripheral vision. According to this account, scene perception and classic search tasks use the same underlying visual encoding, and whether a task is difficult or easy to do in a glance depends on whether the necessary features are readily available in the periphery. The peripheral visual system computes a rich set of summary statistics over some feature space within pooling regions distributed across the visual field. Coarsely pooled features are sufficient for extracting the broad structures and texture surfaces in the scene, so this encoding can support scene perception tasks such as identifying the basic-level scene category and recognizing the spatial layout of the scene. It also conveys some information about objects in the scene and their likely locations, but it is not sufficient to perform tasks that require very fine-grained feature localization or discrimination, such as visual search or crowded letter recognition or peripheral recognition of objects in scene contexts in the current study.

However, the information available may be perfectly sufficient for everyday navigational tasks. People can determine in a glance if a street is mostly commercial or residential buildings and even recognize the kind of building and road details that distinguish different cities, such as Paris or Los Angeles. This kind of scene gist information may be useful for way-finding tasks. People are also able to extract spatial layout information peripherally, and this perception of surface orientation and position may help people avoid and detect obstacles without requiring focal attention. Furthermore, although object detection is notably worse in the periphery, scene category and layout may be used to guide attention and eye movements to the likely location of objects. Although the representation in the periphery is impoverished relative to the fovea, it may be well designed for navigation, allowing people to quickly process a wide field of view in order to build and maintain a representation of the space around themselves.

Assuming that there is a single, general encoding mechanism underlying a range of visual tasks is more parsimonious than assuming different encoding mechanisms for different tasks, e.g., a separate pathway for scene perception (Wolfe, 2007) that is separate from the visual processing of other types of displays or different kinds of attention for scenes (Rensink, 2001; A. Treisman, 2006). We propose that all visual tasks are affected by a single bottleneck: a compressed, summary-statistic representation in peripheral vision that limits the features available across the visual field. Which tasks are difficult or easy depends on how well they can be accomplished with this representation. Some tasks, such as recognizing the gist of a scene, are easily accomplished in the periphery with these limited features or a subset thereof, and other tasks, such as searching for small objects in clutter, are extremely difficult. Determining exactly what features are available across the peripheral visual field is difficult because different tasks may use different subsets of features and rely on their own underlying mechanisms to process those features. However, we do not need a perfect model of the periphery in order to test the basic question of whether a single encoding can explain a range of results. Although the TTM is not a perfect model of the feature representation in the periphery, it is close enough to be able to predict performance on a variety of visual tasks. The fact that this model can also support scene perception suggests that there is no need to assume separate pathways for scene perception versus other visual tasks: A single model of peripheral encoding can explain performance across tasks. This suggests the possibility of a unified account of visual encoding underlying much of visual processing.

Keywords: scene perception, peripheral vision, crowding, parafoveal vision, navigation

Acknowledgments

The authors would like to thank Ali Jahanian for help running the free-viewing condition of Experiment 1. This work was supported by an IIS-1607486 to Dr. Rosenholtz as part of the NSF/NIH/ANR/BMBF/BSF Collaborative Research in Computational Neuroscience Program.

Commercial relationships: none.
Corresponding author: Krista A. Ehinger.
Email: k.a.ehinger@gmail.com.
Address: Visual Attention Lab, Harvard Medical School and Brigham & Women's Hospital, Cambridge, MA, USA.

References

Abdelnour, O., & Kalloniatis, M. (2001). Word acuity threshold as a function of contrast and retinal eccentricity. *Optometry and Vision Science, 78*, 914–919.

Alexander, R. G., Schmidt, J., & Zelinsky, G. J. (2014).

Are summary statistics enough? Evidence for the importance of shape in guiding visual search. *Visual Cognition, 22*(3–4), 595–609.

Anstis, S. M. (1974). Letter: A chart demonstrating variations in acuity with retinal position. *Vision Research, 14*(7), 589–592.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12):13, 1–18, doi:10.1167/9.12.13. [PubMed] [Article]

Boucart, M., Moroni, C., Szaffarczyk, S., & Tran, T. H. C. (2013). Implicit processing of scene context in macular degeneration. *Investigative Ophthalmology & Visual Science, 54*(3), 1950–1957. [PubMed] [Article]

Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., & Greene, M. (2013). Scene categorization at large visual eccentricities. *Vision Research, 86*, 35–42.

Bouma, H. (1970, Apr 11). Interaction effects in parafoveal letter recognition. *Nature, 226*, 177–178.

Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition? *Frontiers in Psychology, 2*(326), 1–15.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*, 3, doi:10.1371/journal.pone.0057410.

Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. A. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics (SIGGRAPH 2012), 31*, 3.

Eberhardt, S., & Zetzsche, C. (2013). Low-level global features for vision-based localization. In M. Ragni, M. Raschke, and R. Stolzenburg (Eds.), *Proceedings of the K1 2013 workshop on visual and spatial cognition* (pp. 5–12).

Eberhardt, S., Zetzsche, C., & Schill, K. (2016). Peripheral pooling is tuned to the localization task. *Journal of Vision,* in press.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*, 945–978.

Freeman, J., & Simoncelli, E. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*, 974–981.

Geisler, W. S., & Perry, J. S. (1998). A real-time foveated multi-resolution system for low-bandwidth video communication. In B. Rogowitz & T. Pappas (Eds.), *SPIE proceedings, vol. 3299: Human vision and electronic imaging III* (pp. 294–305). Bellingham, WA: SPIE.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464–472.

Greene, M. R., & Wolfe, J. M. (2011). Global image properties do not guide visual search. *Journal of Vision, 11*(6):18, 1–9, doi:10.1167/11.6.18. [PubMed] [Article]

Hays, J., & Efros, A. A. (2008). IM2GPS: Estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 3436–3443). New York: IEEE.

Hyman, I. E., Sarb, B. A., & Wise-Swanson, B. M. (2014). Failure to see money on a tree: Inattentional blindness for objects that guided behavior. *Frontiers in Psychology, 5*, 1–7.

Keshvari, S., Rosenholtz, R. (2016). Pooling of continuous features provides a unifying account of crowding. *Journal of Vision, 16*(3):39, 1–15, doi:10.1167/16.3.39. [PubMed] [Article]

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, doi:10.1167/9.1.2. [PubMed] [Article]

Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 471–487.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision, 9*(10):6, 1–16, doi:10.1167/9.10.6. [PubMed] [Article]

Lettvin, J. Y. (1976). On seeing sidelong. *The Sciences, 16*(4), 10–20.

Li, F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, USA, 99*(14), 9596–9601.

Loschky, L., McConkie, G., Yang, J., & Miller, M. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition, 12*(6), 1057–1092.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made prior to basic-level

distinctions in scene gist processing. *Visual Cognition, 18*(4), 513–536.

Mace, M. J. M., Joubert, O. R., Nespoulous, J.-L., & Fabre-Thorpe, M. (2009). The time-course of visual categorizations: You spot the animal faster than the bird. *PLoS ONE, 4*(6), 1–12.

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

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.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–71.

Rensink, R. A. (2001). Change blindness: Implications for the nature of visual attention. In M. Jenkin & L. Harris (Eds.), *Vision and attention* (pp. 169–188). New York: Springer.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*, 437–457.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3*, 1–15.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12*(4):14, 1–17, doi:10.1167/12.4.14. [PubMed] [Article]

Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision, 10*(1):2, 1–25, doi:10.1167/10.1.2. [PubMed] [Article]

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852–877.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Thorpe, S., Fize, D., & Marlot, C. (1996, June 6). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*, 869–876.

Toet, A., Jansen, S. E. M., & Delleman, N. J. (2007). Effects of field-of-view restrictions on speed and accuracy of manoeuvring. *Perceptual and Motor Skills, 105*(3), 1245–1256.

Toet, A., Jansen, S. E. M., & Delleman, N. J. (2008). Effects of field-of-view restriction on manoeuvring in a 3-D environment. *Ergonomics, 51*(3), 385–394.

Tractinsky, N., & Shinar, D. (2008). Do we bump into things more while speaking on a cell phone? In *Proceedings of ACM CHI 2008 conference on human factors in computing systems* (pp. 2433–2442). New York: ACM.

Tran, T. H. C., Rambaud, C., Despretz, P., & Boucart, M. (2010). Scene perception in age-related macular degeneration. *Investigative Ophthalmology & Visual Science, 51*(12), 6868–6874. [PubMed] [Article]

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*, 411–443.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97–136.

van Rheede, J. J., Kennard, C., & Hicks, S. L. (2010). Simulating prosthetic vision: Optimizing the information content of a limited visual display. *Journal of Vision, 10*(14):32, 1–15, doi:10.1167/10.14.32. [PubMed] [Article]

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience, 16*(1), 4–14.

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.

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision, 16*(2):4, 1–30, doi:10.1167/16.2.4. [PubMed] [Article]

Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision, 10*(4):6, 1–27, doi:10.1167/10.4.6. [PubMed] [Article]

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York: Oxford.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Science, 15*(2), 77–84.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN Database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3485–3492). New York: IEEE.

Zhang, X., Huang, J., Yigit-Elliott, S., & Rosenholtz, R. (2015). Cube search, revisited. *Journal of Vision, 15*(3):9, 1–18, doi:10.1167/15.3.9. [PubMed] [Article]