Dissociation of salience-driven and content-driven spatial attention to scene category with predictive decoding of gaze patterns

Thomas P. O'Connell

The Ohio State University, Columbus, OH, USA
Department of Psychology, Yale University, New Haven, CT, USA

Dirk B. Walther

The Ohio State University, Columbus, OH, USA
Department of Psychology, University of Toronto, Toronto, ON, Canada

Scene content is thought to be processed quickly and efficiently to bias subsequent visual exploration. Does scene content bias spatial attention during task-free visual exploration of natural scenes? If so, is this bias driven by patterns of physical salience or content-driven biases formed through previous encounters with similar scenes? We conducted two eye-tracking experiments to address these questions. Using a novel gaze decoding method, we show that fixation patterns predict scene category during free exploration. Additionally, we isolate salience-driven contributions using computational salience maps and content-driven contributions using gaze-restricted fixation data. We find distinct time courses for salience-driven and content-driven effects. The influence of physical salience peaked initially but quickly fell off at 600 ms past stimulus onset. The influence of content effects started at chance and steadily increased over the 2000 ms after stimulus onset. The combination of these two components significantly explains the time course of gaze allocation during free exploration.

Introduction

Humans have access to high-level scene information with very brief presentation times (Fei-Fei, Iyer, Koch, & Perona, 2007; Potter & Levy, 1969), even when attention is engaged with another task (Li, VanRullen, Koch, & Perona, 2002). This fast extraction of high-level scene information, sometimes referred to as "gist," is purported to serve subsequent biasing of visual attention for tasks such as navigation (Greene & Oliva, 2009b), avoidance of threats (Righart & de Gelder, 2008), visual search (Castelhano & Henderson, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006; Wolfe, Võ, Evans, & Greene, 2011), or providing context for more detailed visual exploration (Bar, 2004) and memory encoding (Brockmole, Castelhano, & Henderson, 2006; Chun & Jiang 1998; Hollingworth, Williams, & Henderson, 2001). Torralba et al. (2006) explored the interaction of physical salience, global scene properties, and specific task for visual search, showing that combining the three contributions in a Bayesian framework predicts eye movements within novel scenes. In visual search tasks, both visual (Castelhano & Henderson, 2007) and semantic (Castelhano & Heaven, 2010) scene category cues do not facilitate visual search, but scene category does facilitate transfer of contextual cues between semantically related scenes (Brockmole & Võ, 2010).

Whereas these studies focused on the mediating effect of high-level scene information on task-related behavior, here we explore the direct effect of scene content on overt attention in the absence of a specific task. Free-exploration paradigms have been used to investigate the influence of low-level scene information on gaze patterns (Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005; Tatler, Baddeley, Gilchrist, 2005). Here we aim to extricate the effect of high-level scene content from effects related to physical salience in order to determine if readily available high-level scene content directly affects spatial attention.

Scene content encompasses a variety of partially overlapping aspects of natural scenes, such as global scene information related to the spatial frequency spectrum (Oliva & Torralba, 2001), objects in scenes

Citation: O'Connell, T. P., & Walther, D. B. (2015). Dissociation of salience-driven and content-driven spatial attention to scene category with predictive decoding of gaze patterns. *Journal of Vision, 15*(5):20, 1–13. http://www.journalofvision.org/content/15/5/20, doi: 10.1167/15.5.20.

doi: 10.1167/15.5.20 Received June 12, 2014; published April 29, 2015 ISSN 1534-7362

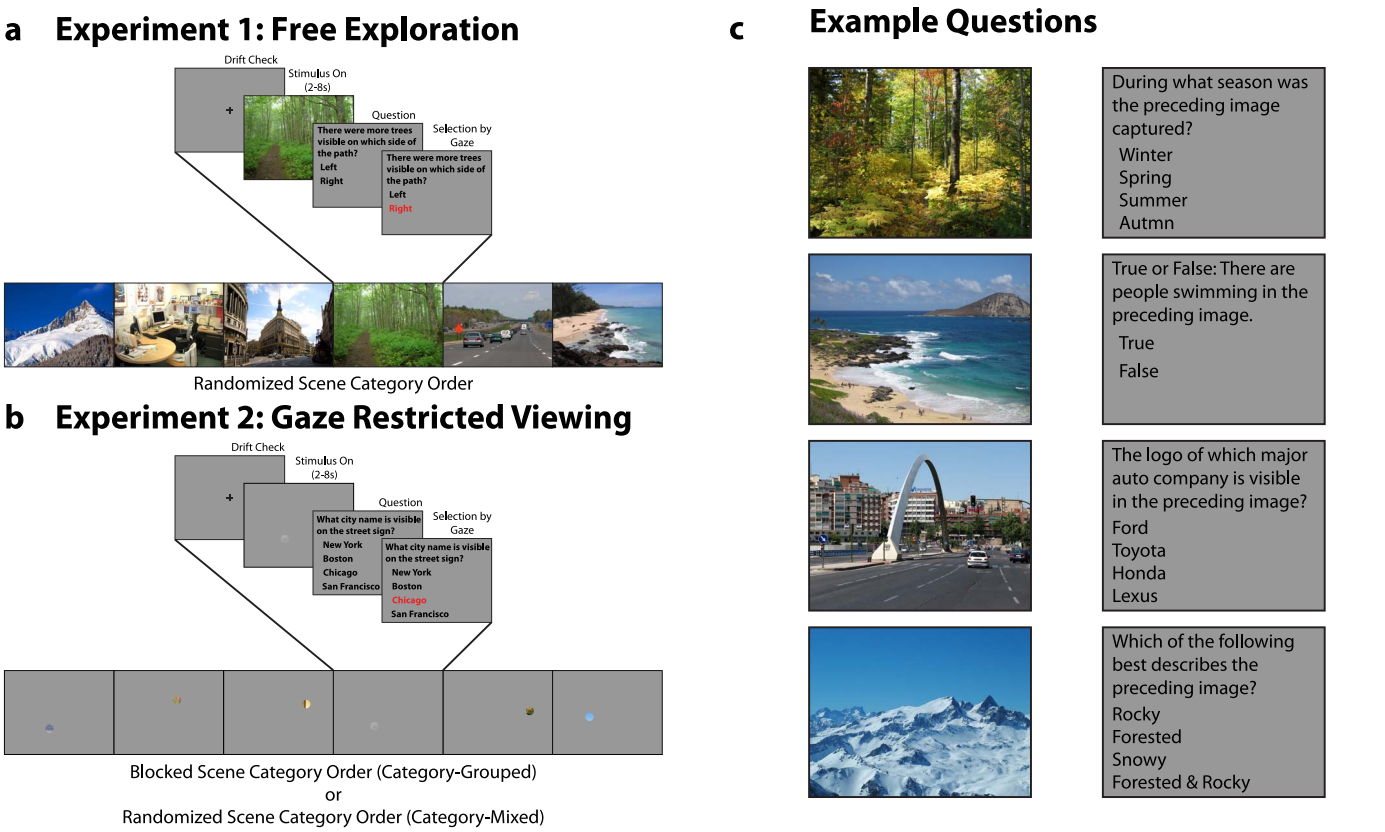

Figure 1. (a) Structure of Experiment 1. Participants viewed natural scene pictures from six scene categories (beaches, city streets, forests, highways, mountains, and offices, presented randomly), and answered questions after 33% of the items. (b) Structure of Experiment 2. Participants viewed natural scene pictures through a gaze-contingent moving window. In one condition, pictures were blocked by scene category to potentiate any category-specific attentional processes (category-grouped), whereas in another condition, pictures were drawn equally from all scene categories (category-mixed). (c) Example questions that appeared in 33% of trials.

(Hollingworth et al., 2001; MacEvoy & Epstein, 2011), scene category (Greene & Oliva, 2009a; Tversky & Hemenway, 1983; Walther, Caddigan, Fei-Fei, & Beck, 2009), or scene layout (Greene & Oliva, 2009b; Kravitz, Peng, & Baker, 2011; Park, Brady, Greene, & Oliva, 2011). In the current paper, we focus on one aspect of high-level scene content, scene category, without claiming comprehensiveness or exclusivity in explaining the relationship between other types of high-level scene content and overt attention. Human observers readily categorize natural scenes (e.g., beaches, forests, offices, city streets) with presentation times as short as 13 ms (Fei-Fei et al., 2007; Greene & Oliva, 2009a; Loschky et al., 2007; Potter & Levy, 1969; Walther et al., 2009). Functional Magnetic Resonance Imaging (fMRI) research has shown that scene category is represented in several scene-selective brain regions (Walther et al., 2009) and that relatively sparse structural information from a scene is sufficient to represent scene category (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). Here we seek to determine how this readily available scene category information affects observers' subsequent viewing behavior and how this information is applied to guide attention in task-neutral conditions.

Many studies have explored the distinguishing characteristics of goal-driven versus salience-driven selection biases (Corbetta & Shulman, 2002; Desimone & Duncan, 1995; Egeth & Yantis, 1997; Folk, Remington, & Johnston, 1992; Itti & Koch, 2001; Jonides, 1981; Kastner & Ungerleider, 2000; Posner & Petersen, 1990; Posner, Snyder, & Davidson, 1980; Wolfe, Cave, & Franzel, 1989) and the effect of physical salience on eye movements within natural scenes (Parkhurst et al., 2002; Peters et al., 2005; Tatler et al., 2005). Here we dissociate task-neutral salience-driven and content-driven biases on spatial attention within natural scenes as a function of time. In the absence of goal-driven task effects on attention, scene category could affect spatial attention through salience-driven mechanisms, caused by systematic regularities in the distribution of physically salient features, or through content-driven mechanisms, caused by previous experience with the same scene category.

To dissociate these two mechanisms, we conducted two eye-tracking experiments. Participants either had unrestricted access to the entire image (Experiment 1: Free exploration; Figure 1a) or had their gaze restricted to a 7° gaze-contingent moving window (Experiment 2: Gaze-restricted exploration; Figure 1b). During free exploration, participants could make use of physical salience and scene category information to guide eye movements. In contrast, gaze-restricted viewing blocked peripheral access to the physical salience of each scene, thus encouraging the use of content-driven biases to guide attention.

We used a novel gaze decoding technique to measure the contributions of salience-driven and content-driven biases towards the allocation of spatial attention. Gaze pattern decoding predicts stimulus category membership by comparing global gaze patterns to behaviorally and computationally derived reference patterns (see General methods for detailed explanation). Varying the type of information used as the reference pattern allowed us to control which attentional contribution is assessed by a given analysis. In each case, the time window of gaze data input to the gaze pattern decoding algorithm was varied to measure the time course of salience-driven and content-driven contributions over the first two seconds of viewing a natural scene. Using our method, we show that scene category biases allocation of spatial attention in task-neutral conditions during free exploration, and we dissociate salience-driven and content-driven contributions to category-specific patterns of eye movements.

General methods

Stimuli

In this study, 216 color photographs (800 × 600 pixels) depicting six natural scene categories (beaches, city streets, forests, highways, mountains, and offices; 33 images per category) were used. The images were downloaded from the Internet and rated as the best exemplars for their respective category from a larger database of images (Torralbo et al., 2013). Example images can be seen in Figure 1.

Apparatus

Eye movements were recorded using an EyeLink1000 eye tracker (SR Research, Ottawa, ON, Canada), which uses infrared pupil detection and corneal reflection to track eye movements at 1000 Hz. The EyeLink1000 was mounted using a tower setup, which stabilized participants' heads with chin and forehead rests. Eye movement data were recorded monocularly from participants' dominant eyes. Subjects were positioned 50 cm from a CRT monitor (21 in. diagonal) with a resolution of 800 × 600 pixels and a refresh rate of 150 Hz. Stimuli were presented on the entire screen and subtended 44° × 33° of visual angle. Stimulus presentation and response recording were controlled with a PC, running the Windows XP operating system and Experiment Builder software (SR Research). Responses were recorded using a game controller connected to the parallel port of the eye-tracking computer.

Calibration procedure

Eye-tracker calibration and calibration validation were completed using a nine-point fixation sequence. Each trial was preceded by a drift check, in which participants focused on a central fixation mark. The eye tracker was recalibrated when drift exceeded 1° of visual angle and at the beginning of each block. The mean spatial accuracy of the eye tracker across all calibrations was 0.49° ($SD = 0.13°$) of visual angle.

Gaze pattern classification

To assess the contribution of salience-driven and content-driven biases towards attentional allocation, we developed a classification technique to predict an image's scene category from evoked gaze patterns. We held out one participant's gaze data and computed category predictions using the other participants' data for training. This procedure was repeated for each participant in a leave-one-subject-out (LOSO) cross validation.

To capture the overall pattern of fixations for each scene category, fixation density maps (FDMs) were calculated from the training data as the sum of impulse functions, centered at the locations of fixations and weighted by fixation duration. The maps were then convolved with a two-dimensional Gaussian kernel, followed by normalization to zero-mean and unit-standard deviation across spatial locations to obtain z-score maps:

$$F'(x,y) = \frac{1}{\sum_{f=1}^{N} d_f} \sum_{f=1}^{N} d_f \cdot \exp\left(\frac{-(x_f - x)^2 - (y_f - y)^2}{\sigma^2}\right) \quad (1)$$

$$F(x,y) = \frac{F'(x,y) - \overline{F'}}{\sigma_{F'}} \quad (2)$$

Here N is the number of fixations in a given trial, (x_f, y_f) is the fixation location, d_f is the duration of fixation

f, $\overline{F'}$ is the mean, and $\sigma_{F'}$ is the standard deviation over all locations in the unnormalized FDMs. The standard deviation of the Gaussian kernel σ was chosen to be 0.55° of visual angle (10 pixels) in approximation of the average spatial accuracy of the eye tracker. Category-specific grand FDMs were computed for each of the six scene categories by averaging over all trials of a category for all subjects in the training set:

$$\overline{F_{\text{cat}}} = \frac{1}{N_{\text{trials}}(\text{cat})} \sum_{t=1}^{N_{\text{trials}}(\text{cat})} F_t \quad (3)$$

In order to control for spatial biases common across all scene categories, most prominently the center bias, we computed marginal FDMs by subtracting the average over all category-specific grand FDMs from each of the category-specific FDMs:

$$M_{\text{cat}} = \overline{F_{\text{cat}}} - \frac{1}{N_{\text{cat}}} \sum_{c=1}^{N_{\text{cat}}} \overline{F_c} \quad (4)$$

In each scene category's marginal FDM, positive values represent regions fixated more than the overall average across categories, and negative values represent regions fixated less than the overall average across categories. Marginal fixation density maps for each scene category can be seen in Figure 2a for Experiment 1 and Figure 2c for Experiment 2.

We assessed the overall pattern of physical salience for each scene category using computational salience maps (SMs), generated with the Saliency Toolbox (Walther & Koch, 2006). This model of bottom-up attention computes salience solely from contrasts of low-level image properties (color contrasts, luminance contrasts, orientation contrasts; Itti, Koch, & Niebur, 1998), without having any notion of scene content. It has been shown to predict observers' fixations when information is limited to these low-level image properties (Peters et al., 2005). Salience maps were generated individually for each image using default parameters at 1/16 the image resolution, up-sampled to the image resolution, and averaged across images belonging to the same scene category to create category-specific grand SMs. Marginal SMs were calculated for each category by subtracting the overall grand SM from category-specific SMs, in analogy to Equation 4. For a given scene category's marginal SM, positive values indicate regions with higher than average salience across categories, and negative values represent regions with lower than average salience across categories. Marginal salience maps for each scene category can be seen in Figure 2b.

To test whether gaze patterns were discriminative of scene category with a given reference pattern, we modified the normalized scanpath salience method (Peters et al., 2005) for assessing the goodness-of-fit between fixation data and SMs. Our gaze pattern decoding analysis predicts scene category using fixation data by generating metrics representing the goodness-of-fit between single-trial gaze patterns and marginal FDMs or SMs for each scene category. Gaze pattern metrics were calculated separately for each category by summing marginal reference values weighted by test data fixation duration across fixated locations. For a given trial from the test data, gaze pattern metrics were generated for each category in the following manner:

$$G_{\text{cat}}(\text{trial}) = \frac{1}{\sum_{f=1}^{N} d_f} \sum_{f=1}^{N} d_f \cdot M_{\text{cat}}(x_f, y_f) \quad (5)$$

Here, N is the number of fixations in the trial, d_f is the fixation duration, M_{cat} is the marginal reference map, and (x_f, y_f) is the fixated location. The category with the largest G_{cat} was selected as the predicted category for that trial:

$$\text{prediction}(\text{trial}) = \arg\max_{\text{cat}} G_{\text{cat}}(\text{trial}) \quad (6)$$

Using LOSO cross validation we obtained category predictions for each subject, using all other subjects to compute FDMs. Prediction accuracy was computed as the fraction of trials with correct category predictions and compared to chance (1/6) using a one-tailed t test over individual subjects.

To assess the time course of category-specific contributions to attentional control, the full gaze decoding analysis was run separately for 17 cumulative intervals of fixation data, ranging from 0–400 ms to 0–2000 ms in steps of 100 ms. This allowed us to assess how category-specific biases in spatial attention (whether driven by physical salience or category-specific selection history) accumulate over the course of a trial. T tests were conducted using Bonferroni adjusted alpha levels of 0.0029 per test (0.05/17).

Experiment 1: Free exploration

Methods

Subjects

Twenty-two young adults (13 male, ages 18–41, mean age 20.2 years) completed Experiment 1. All participants had normal or corrected-to-normal vision and received partial course credit for completion of the experiment. The experiment was approved by The Ohio State University Independent Review Board.

Procedure

Participants viewed images depicting natural scenes in three blocks of 66 trials (198 trials total per

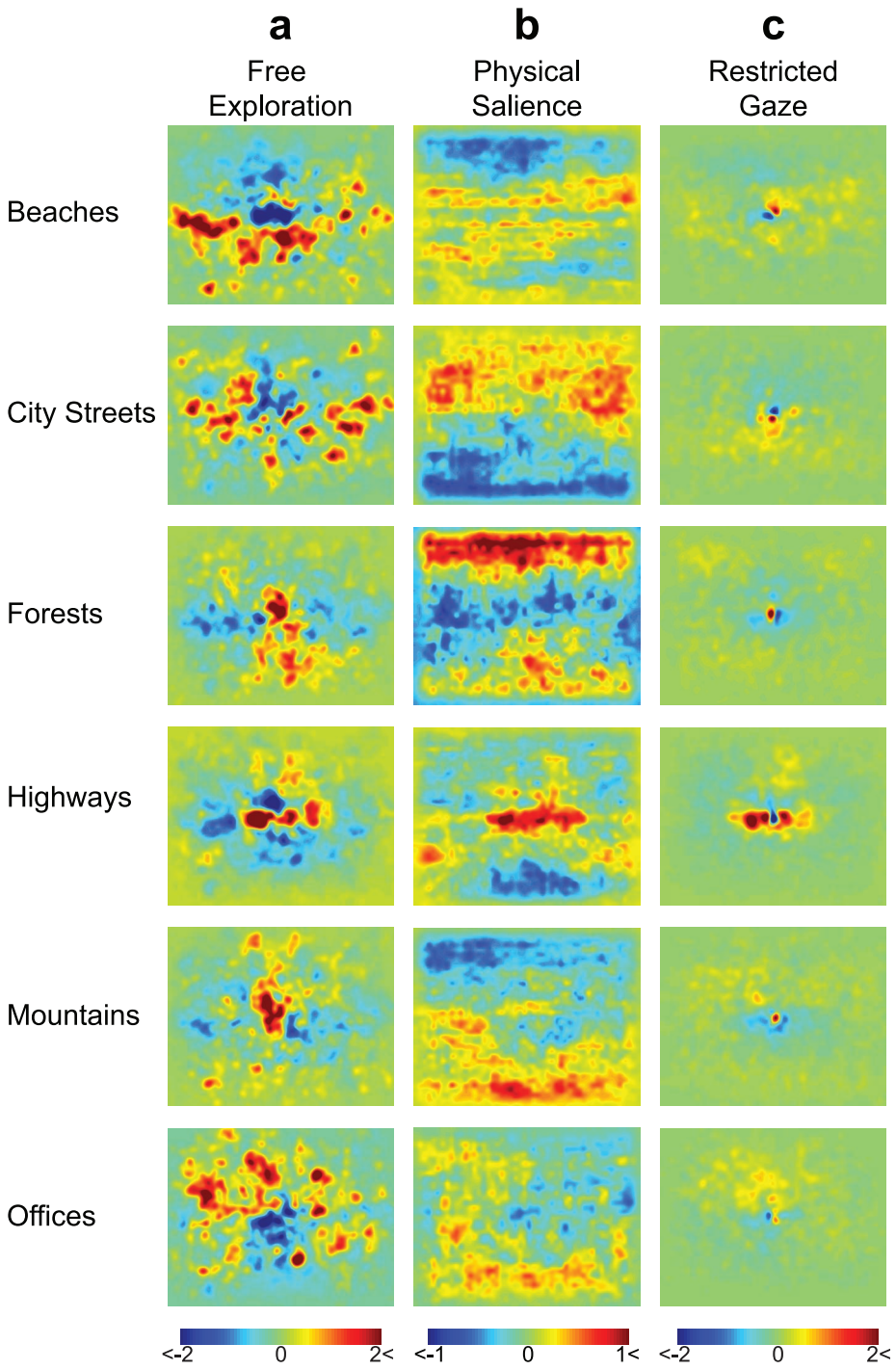

Figure 2. Example marginal maps used as references for gaze pattern decoding. (a) Marginal fixation density maps for free exploration calculated with data from all 22 subjects in Experiment 1. (b) Marginal salience maps calculated using the Saliency Toolbox (Walther & Koch, 2006) for all 216 color photographs. (c) Marginal fixation density maps for gaze-restricted exploration calculated with data from all 40 subjects in Experiment 2. Note that maps for the gaze pattern decoding analysis were computed only from the 21 (Experiment 1) or 39 (Experiment 2) training subjects in each LOSO cross validation fold.

participant), with equal number of scenes from each of the six categories, randomly interleaved. The eye tracker was calibrated before each block and recalibrated during a block if the spatial error of the camera exceeded 1° of visual angle. Stimuli for each participant were randomly sampled from the 216 overall image pool. Each trial started with a drift check, followed by a color photograph of a scene presented for a random time interval between 2 and 8 s (Figure 1a). A random time interval was used to ensure participants' gaze was

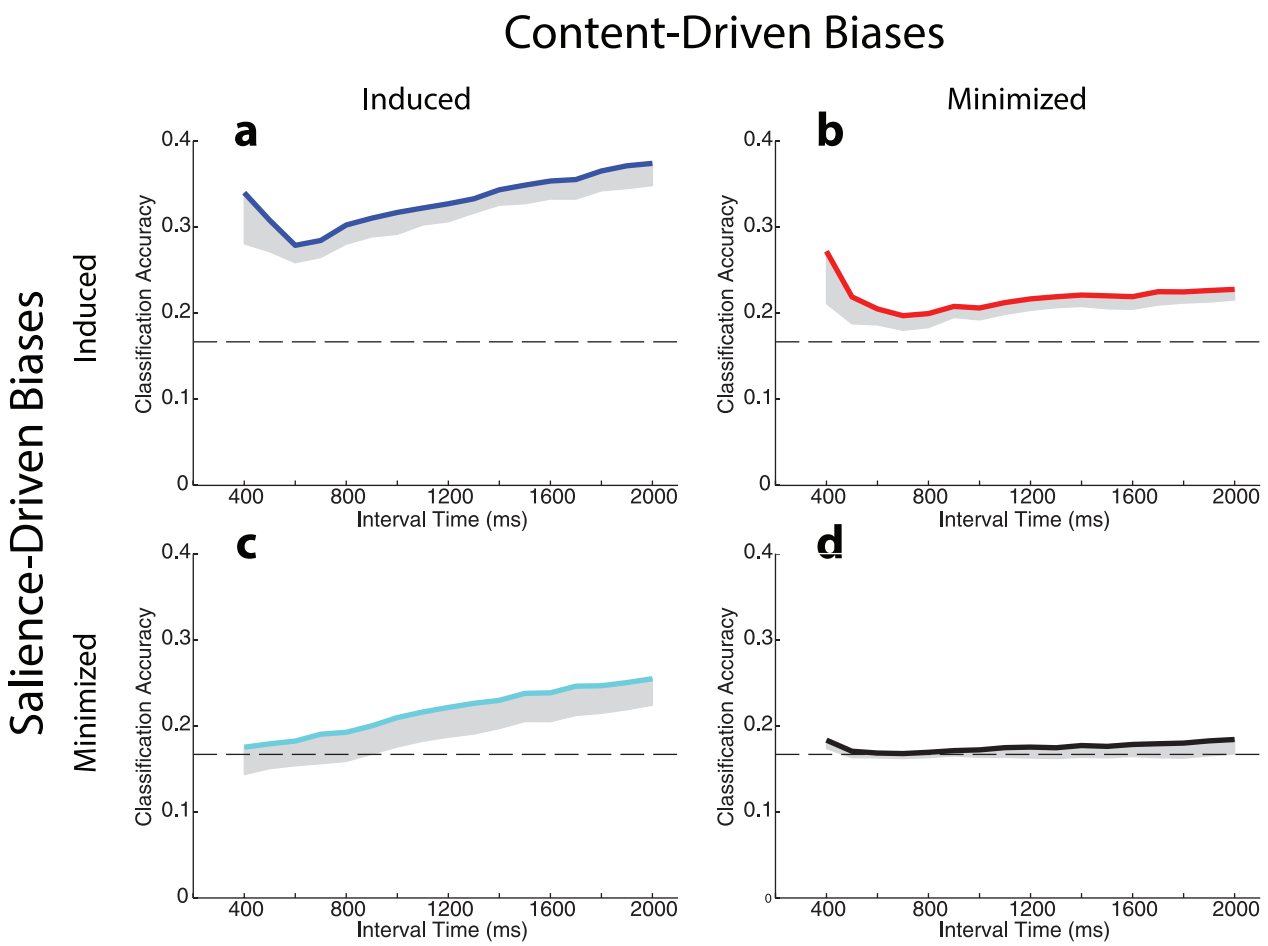

Figure 3. Results for gaze pattern decoding analyses. Note that time intervals are cumulative to assess the accumulation of category-specific information over the course of a trial. (a) Results from the free-exploration analysis. Free-exploration gaze data (Experiment 1) were used to calculate marginal fixation maps and make scene category predictions. (b) Results from the salience-driven component analysis. Free-exploration gaze data were compared to marginal salience maps to predict scene category. (c) Results from content-driven component analysis. Gaze-restricted data (Experiment 2) were used to calculate marginal fixation maps and make scene category predictions. (d) Results from the gaze-restricted paradigm confirmation analysis with salience maps as a reference (Experiment 2). Gaze-restricted data were compared to marginal salience maps to predict scene category. Gray shading shows the Bonferroni-corrected 99.7% confidence interval.

not biased by prediction of the trial offset. After one third of the trials, participants were asked a multiple choice or true/false question about the preceding image to encourage thorough exploration of each image. The questions were varied to ensure no task-related effects (such as object search bias) on subjects' eye movements (see Figure 1c for example questions). Participants responded to each question by fixating on their selected response and pressing a button on a game controller connected to the presentation computer to confirm their selection.

Results

Free exploration

Participants freely explored scenes, with a mean saccade length of 2.71° of visual angle ($SEM = 0.093°$).

We used gaze pattern decoding with 22-fold LOSO cross validation in order to determine if the gaze patterns elicited during free exploration of natural scenes discriminate scene category. In this analysis, we predicted scene category from marginal FDMs of participants exploring scene images in a free-exploration paradigm. Participants had full access to the physical salience of each image, and thus attentional allocation could be biased by both salience-driven and content-driven components. The use of marginal FDMs from the free-exploration paradigm as the reference pattern (Figure 2a) allowed us to assess the contribution of both salience-driven and content-driven biases in conjunction. Classification rates (Figure 3a) at each interval were compared to chance accuracy (1/6) using a one-tailed t test and corrected for multiple comparisons using Bonferroni correction.

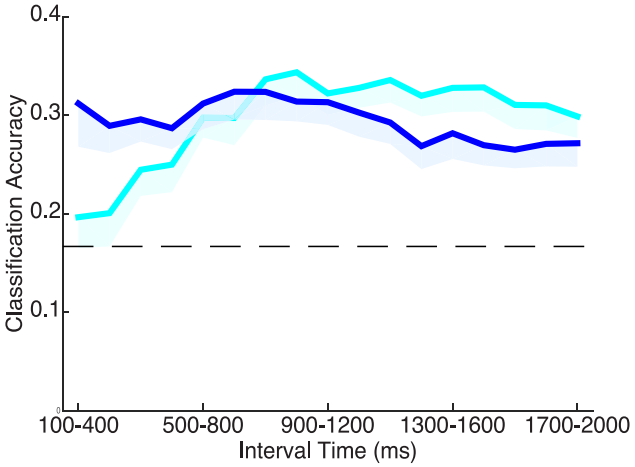

Figure 4. Results for the sliding-window gaze pattern decoding analyses using free-exploration gaze data (blue) and gaze-restricted data (turquoise). The 300 ms sliding window was used to assess the amount of category-specific information present during various time intervals in a trial. Shading shows the Bonferroni-corrected 99.7% confidence intervals.

At the earliest time interval (0–400 ms), scene category was predicted correctly in 34.0% of the trials ($t_{21} = 8.93$, $p = 1.34 \times 10^{-8}$). Accuracy decreased for the next two intervals, down to 27.9% ($t_{21} = 16.97$, $p = 9.69 \times 10^{-14}$) for 0–600 ms, before steadily increasing to 37.4% ($t_{21} = 24.43$, $p = 6.66 \times 10^{-17}$) for 0–2000 ms. Classification accuracy was above chance (1/6) for all time intervals ($t_{21} > 8.93$, $p < 1.34 \times 10^{-8}$), indicating that gaze patterns vary across scene categories. The shape of the time course is best described as a linear decrease from 0–400 ms to 0–600 ms, followed by a linear increase from 0–600 ms to 0–2000 ms (Figure 3a).

Additionally, we conducted a control analysis to assess whether category-specific gaze allocation is present throughout the course of a trial or confined to specific parts of a trial. We repeated the same decoding analysis as above with a 300 ms sliding time window instead of cumulative time intervals. Windows started at 100 to 400 ms and were shifted by 100 ms up to 1700 to 2000 ms. We found that scene category classification rates survived Bonferroni correction at every time window ($t_{21} > 10.11$, $p < 1.61 \times 10^{-9}$), indicating that category-specific fixation patterns were elicited throughout the course of a trial during free exploration of natural scenes (Figure 4).

The result shown in Figure 3a may be caused by two separate mechanisms. First, images of the same category are likely to have similar layout (e.g., one beach scene resembling another beach scene more than a forest scene), and therefore, similar spatial distribution of physically salient locations. Second, content-driven biases may bias attention in category-specific patterns to facilitate extraction of behaviorally relevant information.

Salience-driven biases

We used marginal SMs (Figure 2b) in place of marginal FDMs as a reference for gaze pattern decoding to determine if patterns of physical salience predict evoked gaze patterns. Classification accuracies were calculated for each participant at each time interval and averaged across individuals (Figure 3b).

As before, we found classification accuracy above chance (1/6) for every time interval ($t_{21} > 5.11$, $p < 4.61 \times 10^{-5}$), albeit accuracy was overall lower than in the previous analysis (Figure 3b). We found relatively high prediction accuracy of 27.2% ($t_{21} = 5.30$, $p = 2.94 \times 10^{-5}$) for the earliest time interval (0–400 ms), followed by a decrease down to 19.7% ($t_{21} = 5.51$, $p = 1.81 \times 10^{-5}$) at 0–700 ms. We also see a slight increase again for later time intervals up to 22.8% ($t_{21} = 15.13$, $p = 9.14 \times 10^{-13}$) for 0–2000 ms.

Experiment 2: Gaze-restricted viewing

In order to isolate the role of content-driven biases in guiding attention under task-neutral conditions, we modified our experimental procedure by restricting gaze to the perifoveal region. This modification was designed to suppress access to the physical salience of the scene outside the fixated location and potentiate the influence of content-driven biases. Subjects viewed images in two conditions: (a) images within a block all came from the same scene category with a word cue introducing the category before each block (category-grouped condition) and (b) images within a block were drawn randomly from all six scene categories (category-mixed condition). The former condition was designed to potentiate any category-specific content-driven biases maximally, while the latter condition was designed to assess whether our semantic manipulation potentiated category-specific biases above and beyond biases initiated via rapid categorization of the scene through the moving window (Larson & Loschky, 2009).

Methods

Subjects

Forty undergraduates at The Ohio State University (21 male, ages 18 to 41, mean age 19.95 years) participated in Experiment 2 for partial course credit. All participants had normal or corrected to normal vision, and none of them had participated in Experiment 1. The experiment was approved by The Ohio State University Independent Review Board.

Since Experiment 2 contained two separate experimental conditions, the number of trials per participant per condition was smaller in Experiment 2 than Experiment 1. The number of participants in Experiment 2 was chosen to equate power between the two experiments: 198 trials × 22 participants = 4,356 trials for Experiment 1; 108 trials per condition × 40 participants = 4,320 trials per condition for Experiment 2.

Procedure

In Experiment 2, participants viewed images depicting natural scenes in a gaze-restricted paradigm that limited the availability of physical salience from the image (see Reingold et al., 2003, for details on gaze-contingent viewing paradigms). For the duration of each trial (2–8 s), participants viewed each image through a moving circular window with a diameter of 7° of visual angle that revealed 2.4% of the image area around the fixated location (Figure 1b). The rest of the image was masked by an opaque gray overlay. Each participant viewed all 216 images. Participants viewed 12 blocks of images containing 18 trials apiece. In six blocks, images were grouped by scene category, and participants were informed of the category for the upcoming block by a centrally presented word cue preceding the block in order to maximize content-driven biases (category-grouped condition). In the other six blocks, an equal number of images from all scene categories were randomly interleaved (category-mixed condition). The eye tracker was calibrated before each block and recalibrated during a block if the spatial error of the camera exceeded 1° of visual angle. The same questions used in Experiment 1 were asked after one third of trials to ensure subjects thoroughly explored each image. Data analysis was identical to Experiment 1.

Results

Gaze-restricted paradigm confirmation

How does restricting visibility of the image to the perifoveal region affect participants' exploration of the scenes? Not surprisingly, mean saccade length across conditions in Experiment 2 ($M = 1.75°$, $SEM = 0.073°$) was significantly shorter than in Experiment 1 ($t_{60} = 6.52$, $p = 1.64 \times 10^{-8}$; unpaired t test). However, did the manipulation also suppress salience-driven attentional biases as intended? If this was the case, then gaze patterns from Experiment 2 should no longer predict scene category when marginal SMs are used as reference patterns. This is indeed what we found (Figure 3d) for all time intervals in the category-grouped condition, except for the 0–400 ms interval (18.3%; $t_{39} = 4.91$, $p = 1.65 \times 10^{-5}$) and for the 0–2000 ms interval (18.4%; $t_{39} = 3.00$, $p = 2.40 \times 10^{-3}$). Prediction was not significantly above chance accuracy after Bonferroni correction for any other interval ($t_{39} < 2.63$, $p > 0.0122$). Shorter saccade length alone cannot explain this result. If participants were using salience-driven biases to guide fixations, even at shorter range within the moving window, we would still expect physical salience maps to significantly predict fixation patterns in the gaze-restricted paradigm.

Content-driven biases

Having established that biases on visual attention driven by physical salience are largely suppressed in the gaze-restricted paradigm, we explored the role of content-driven biases on the allocation of spatial attention. We used gaze-restricted marginal FDMs (Figure 2c) as a reference for gaze pattern decoding in a 40-fold LOSO cross validation. Prediction accuracy was averaged over participants at each time interval and compared to chance accuracy (1/6) using a one-tailed t test with Bonferroni correction for multiple comparisons.

We did not find prediction accuracy to be significantly above chance (1/6) for early time intervals (Figure 3c) in the category-grouped condition. Accuracy first became significant at 20.0% ($t_{39} = 2.94$, $p = 2.80 \times 10^{-3}$) for 0–900 ms and then increased steadily up to 25.5% ($t_{39} = 8.39$, $p = 2.89 \times 10^{-10}$) for 0–2000 ms. We found similar results in the sliding window control analysis (Figure 4). Classification was at chance for the earliest windows (100–400 ms and 200–500 ms) and then steadily increased for later time intervals ($t_{21} > 8.98$, $p < 4.89 \times 10^{-11}$).

The category-mixed condition was designed to probe the influence of prior expectations through explicit cuing and blocking of the categories. We found that all classification results for cumulative intervals were the same for the category-mixed as the category-grouped condition, both for the salience map analysis, $t(39) < 1.27$ (paired t tests), and the content-driven bias analysis, $t(39) < 1.49$.

Combined results

Collective contributions of salience-driven and content-driven biases

Do the measured salience-driven (Figure 3b) and content-driven (Figure 3c) components sufficiently describe the pattern of gaze allocation during free exploration (Figure 3a)? In a post hoc analysis we summed the classification rates elicited in the salience-driven analysis and the content-driven analysis and subtracted chance (1/6) to assess the combined

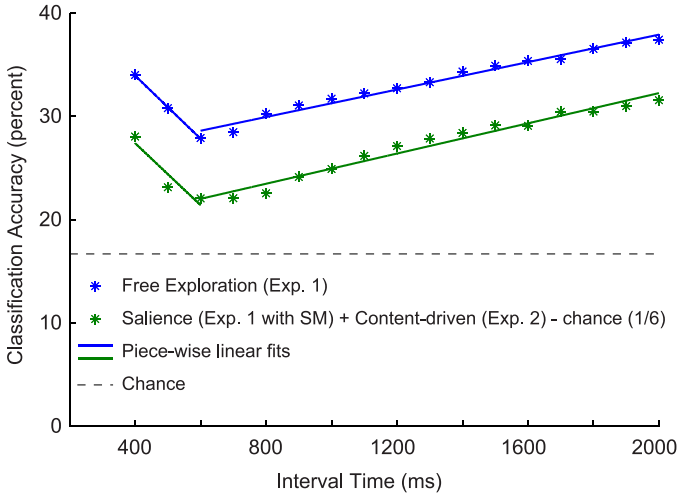

Figure 5. Summing salience-driven contributions (Figure 3b) and content-driven contributions (Figure 3c) results in a curve (green) with the same shape as free exploration (Figure 3a; blue curve). The slopes of the two curves are statistically the same (see text for analysis).

contribution of these two attentional components. The resultant curve subjectively exhibited the same shape as the free-exploration classification analysis (Figure 5). To better quantify their similarity, we fitted both curves with two straight lines, one for the initial drop in classification accuracy from 400–600 ms and one for the subsequent recovery, starting at 600 ms. The slopes for the initial drop are −0.306 (95% confidence interval: [−0.478, −0.133]) for free exploration and −0.283 (95% confidence interval: [−0.483, −0.082]) for the sum of salience-driven and content-driven contributions. The slopes for the later recovery are 0.0664 (95% confidence interval: [0.0537, 0.0792]) for free exploration and 0.0736 (95% confidence interval: [0.0547, 0.0925]) for the sum of salience-driven and content-driven contributions.

In both cases, the mean of the slopes for the sum of contributions are within the confidence interval of free exploration, and vice versa. Two-sample t tests comparing the set of individual subjects' slopes of (free exploration minus physical salience contributions; $N = 22$) and content-based contributions ($N = 40$) show no significant differences ($t < 1$ for both slopes). To explicitly test the equality of these two sets of slopes, we performed a test for common language effect size (CLES; McGraw & Wong, 1992). In this test we considered each pairing of subjects from the first group (free exploration minus physical salience) with subjects from the second group (content-based). The test showed that over all 880 pairings, slopes are about equally likely to be larger for the first as for the second group (CLES = 0.485 for the initial drop; CLES = 0.446 for the later recovery). Based on this analysis, we conclude that the combination of salience-driven and content-driven biases accounts for gaze behavior observed during free exploration in our experiments. However, given that our model was constructed as a post hoc fit of the observed data, it remains possible that other factors, such as individual differences in eye movement patterns, may explain a significant portion of the variance in gaze allocation behavior. Note that we did not make any assumptions about the relative strengths of stimulus- and content-driven contributions. The decoding accuracy for each of the two contributions was determined by way of excluding the respective other contribution by way of analysis and experimental manipulation. In the additive analysis presented above, we simply combined the two isolated components back together and compared the result with the original free-viewing behavior. In fact, we did not completely succeed in explaining free-exploration behavior by these two components, as is demonstrated by the remaining constant offset between the free-exploration accuracy and the summed accuracies for salience-driven and content-driven components. This offset implies that free-exploration behavior contains some other, so far unexplained, contributions.

Discussion

We have shown that scene category significantly affects patterns of gaze allocation during free exploration of natural scenes, providing evidence that high-level scene information directly biases spatial attention in the absence of an explicit task. Additionally, we dissociated salience-driven and content-driven contributions towards category-specific patterns of gaze allocation in our two experiments. We found that the combination of salience-driven and content-driven biases largely accounts for the effects observed in a free-exploration paradigm, suggesting that, taken together, our measured components explain free-exploration gaze behavior in natural scenes. Furthermore, we show proof of concept for a gaze pattern decoding methodology to effectively dissociate these two components.

In our free-exploration paradigm, we investigated the overall effect of scene category on allocation of spatial attention under task-neutral conditions when both salience-driven and content-driven effects were potentially biasing attentional control. Using gaze patterns evoked in the unrestricted viewing paradigm as the input and reference pattern for gaze pattern decoding, we found that gaze patterns significantly predicted an image's scene category. This result highlights that scene category plays a significant role in directly guiding spatial attention.

What category-specific scene information may bias spatial attention? One possibility is that scenes of the same category share spatial distributions of physically salient locations and these salience-driven features drive attention in category-specific patterns. We found evidence for this possibility by showing that physical salience is the primary driver in allocation of spatial attention for the first several hundred milliseconds of viewing a natural scene, and that physical salience continues to play a role in guidance of spatial attention through at least 2000 ms of viewing. Previous studies have found a similar time course for the role of salience in viewing behavior (Parkhurst et al., 2002; Tatler et al., 2005).

Additionally, category-specific gaze allocation may be biased by content-driven factors common to scenes that share category. In support of this possibility, we found that human observers allocated attention in category-specific patterns even when access to physical salience was severely restricted. Since salience-driven biases were suppressed (Figure 3d) and task was held constant, these category-specific fixation patterns most likely reflect content-driven biases that guide attention towards regions of a scene that had been advantageous to attend to in prior experience with scenes of the same category.

What visual components may underlie content-driven attentional allocation in natural scenes? Different scene categories vary in their constituent objects (Greene, 2013) and the locations of objects within scenes show constraints relative to the spatial layout of a scene and other objects in the scene (Bar, 2004; Oliva & Torralba, 2007). For example, cars tend to appear on roads in the lower half of the visual field in city street and highway scenes, pedestrians tend to appear in the lower half of the visual field on sidewalks in city street scenes and on sand beach scenes, computers and keyboards tend to co-occur on desks in office scenes, and rocks tend to appear in the lower half of the visual field in forest and mountain scenes. Given these restrictions across images sharing scene category, identifying an image's scene category should provide information about where behaviorally relevant objects may appear in relation to the spatial layout of the scene and in relation to other objects in the scene.

Wu, Wick, and Pomplun (2014) reviewed several types of scene information that have been shown to guide attention in scenes and may explain our observed content-driven biases. One possibility is that these biases reflect scene–object relations and the allotment of attention towards objects that tend to have behavioral relevance in a given scene category. For example, research has shown that text (Cerf, Frady, & Koch, 2009; Wang & Pomplun, 2012) and human bodies and faces (Buswell, 1935; Yarbus, 1967; Cerf et al., 2009; Fletcher-Watson, Findlay, Leekam, & Benson, 2008) attract attention in free exploration and visual search, even when doing so hurts task performance. Our scene categories contained differential amounts of text (e.g., traffic signs in highway scenes, computer screens in office scenes) and human figures (e.g., pedestrians in city street scenes, vacationers in beach scenes), which could conceivably lead to the category-specific fixation patterns detected in our experiments. In our gaze-restricted paradigm, categorization of the scene should provide information about where to find behaviorally relevant objects in the scene, even without visual access to the entire image.

Another possibility is that object–object relations reflecting co-occurrence and semantic relationships among objects in a scene may drive attention in category-specific patterns. Object co-occurrence relationships have been shown to facilitate visual search in simple arrays (Chun, 2000; Chun & Jiang, 1998, 1999) and natural scenes (Castelhano & Heaven, 2010; Mack & Eckstein, 2011; Wu, Wang, & Pomplun, 2014). Hwang, Wang, and Pomplun (2011) showed that humans are more likely to direct attention toward objects that are semantically related to the currently attended object during search and exploration in natural scenes. Therefore, category-specific fixation patterns could also arise from the tendency to fixate on semantically related or frequently co-occurring objects. For example, if the first fixation in an office scene in our gaze-restricted paradigm fell on a computer screen, the location of co-occurring or semantically related objects (e.g., computer mouse, keyboard) could be inferred without access to the entire image and attention may be biased towards locations where these objects are most likely to be found.

Our failure to find any differences in classification rates between the category-grouped and category-mixed conditions in our gaze-restricted paradigm suggests that our measured content-driven biases are not initiated by prior semantic scene information but driven via rapid visual categorization of the scene. Larson and Loschky (2009) found that scene category could be determined with high accuracy when scenes were presented for just 106 ms and participants only had access to a window spanning 5° of visual angle in the center of the scene. This suggests that even without category grouping or a semantic category cue, participants were likely able to rapidly determine the category of a scene in our gaze-restricted paradigm. Given the rapid rate at which scene category and gist can be extracted (Potter, 1976; Thorpe, Fize, & Marlot, 1996), it is unlikely that the lag in content-driven biases compared to salience-driven biases is due to a lag in processing physical salience compared to scene category.

What might explain the differing time courses in salience-driven and content-driven attentional alloca-

tion? These differences may be explained by use of different attentional allocation strategies over the course of a trial. Physical salience is extracted very rapidly (Nothdurft, 2002) and cues attention in an exogenous manner. Such exogenous cuing has been shown to have a fast time course with fully effective attentional bias only 150–200 ms following cue onset (Posner et al., 1980). Therefore, this information should be available to guide spatial attention by the time the first eye movement is made. While scene information is also extracted very rapidly (Potter, 1976; Thorpe et al., 1996), endogenous cuing based on this high-level information is likely to take longer. Indeed, Posner et al. (1980) showed that endogenous (content-driven) attention takes considerably longer to become fully effective than exogenous (salience-driven) attention. This finding offers an explanation for the slower deployment of content-driven than salience-driven biases in guiding spatial attention in the current experiments. It is important to note that the slower deployment of content-driven attention does not contradict the rapid processing of scene information. The additional time needed for content-driven biases to become fully effective is likely due to the control processes involved in endogenous cuing.

Based on this evidence, we therefore posit that the initial high prediction accuracy followed by a drop-off seen in the free-exploration analysis (Figure 3a) is caused by salience-driven attentional biases, and that the slow recovery of prediction accuracy for later time intervals is due to an interaction of salience-driven and content-driven biases. Initial eye movements, guided mostly by physical salience, may play a role in orienting oneself in a newly encountered environment and initially processing the most salient aspects of the scene. In the absence of a specific task, later fixations, guided in part by physical salience and scene content, may sample locations in the scene that contained useful information or objects in similar previously encountered scenes. Taken together, these two effects significantly explain task-neutral gaze behavior in the free exploration condition.

Finally, we introduced the gaze pattern decoding method for predicting stimulus features using evoked gaze patterns. We demonstrated that this method successfully predicts the scene category of a single image by comparing single-trial gaze patterns to reference patterns of gaze data or computational salience. This method is relatively simple to implement and offers flexibility regarding the classification reference pattern. This allows for the explicit testing of competing hypotheses, via the chosen reference pattern, using the same classification procedure. We expect that such an explicit decoding-based approach to the analysis of gaze data will be applicable to other scenarios to assess the differential contributions of attentional components to gaze behavior.

Conclusion

We have shown that scene category biases spatial attention such that the category of individual scene images can be predicted from evoked gaze patterns. Since no specific task was used in the experiment, this is a direct effect of high-level scene information on the allocation of overt spatial attention. Additionally, we have shown that both salience-driven and content-driven biases guide spatial attention over 2000 ms of viewing a natural scene. These results represent compelling evidence that (a) high-level scene information such as scene category directly influences spatial attention and (b) the contributions of salience-driven and content-driven biases can be dissociated using prediction-based analysis of eye tracking data.

Keywords: scene category, eye movements, eye-tracking, scene perception, bottom-up attention, top-down attention, salience scene content

Acknowledgments

Commercial relationships: none.
Corresponding author: Thomas P. O'Connell.
Email: thomas.oconnell@yale.edu.
Address: Department of Psychology, Yale University, New Haven, CT, USA.

References

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617–629.

Brockmole, J. R., Castelhano, M. S., & Henderson, J. M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 699–706.

Brockmole, J. R., & Võ, M. L. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception, and Psychophysics*, *72*, 1803–1813.

Buswell, G. T. (1935). *How people look at pictures*. Chicago, IL: The University of Chicago Press.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to

visual search in scenes. *Attention, Perception, & Psychophysics, 72*, 1283–1297.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 753–763.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71.

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*, 360–365.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*, 201–215.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18,* 193–222.

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology*, *48,* 269–297.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1):10, 1–29, http://www.journalofvision.org/content/7/1/10, doi:10.1167.7.1.10. [PubMed] [Article]

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37*, 571–583.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 1030–1044.

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4,* 777.

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

Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176.

Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin and Review*, *8*(4), 761–768.

Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*, 1192–1205.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*, 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jonides, J. (1981). Voluntary versus automatic control over the mind's eye's movement. *Attention and Performance IX, 9,* 187–203.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23,* 315–341.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, *31*, 7322–7333.

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, http://www.journalofvision.org/content/9/10/6, doi:10.1167/9.10.6. [PubMed] [Article]

Li, F. 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*, 9596–9601.

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*, 1431–1450.

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience, 14*(10), 1323–1329.

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision, 11*(9):9, 1–16, http://www.journalofvision.org/content/11/9/9, doi:10.1167/11.9.9. [PubMed] [Article]

McGraw, K. O., & Wong, S. P. (1992). A common

language effect size statistic. *Psychological Bulletin, 111,* 361–365.

Nothdurft, H. C. (2002). Attention shifts to salient targets. *Vision Research, 42*, 1287–1306.

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

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*, 520–527.

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*, 1333–1340.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107–123.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*, 2397–2416.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13,* 25–42.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General, 109*, 160–174.

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

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

Reingold, E. M., Loschky, L. C., McConkie, G. W., & Stampe, D. M. (2003). Gaze-contingent multiresolutional displays: An integrative review. *Human Factors, 45*, 307–328.

Righart, R., & de Gelder, B. (2008). Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, Affective, & Behavioral Neuroscience, 8*, 264–272.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.

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

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766–786.

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. *PLoS One, 8*(3), e58594.

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

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience, 29*, 10573–10581.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of National Academy of Sciences, USA, 108*, 9661–9666.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.

Wang, H. C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision, 12*(6):26, 1–17, http://www.journalofvision.org/content/12/6/26, doi:10.1167/12.6.26. [PubMed] [Article]

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 419–433.

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

Wu, C. C., Wang, H. C., & Pomplun, M. (2014). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research, 105,* 10–20.

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5,* 1–13.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.