Atten Percept Psychophys (2015) 77:1239–1251
DOI 10.3758/s13414-015-0859-8

What you see is what you expect: rapid scene understanding benefits from prior experience

Michelle R. Greene · Abraham P. Botros · Diane M. Beck · Li Fei-Fei

Published online: 17 March 2015

Abstract Although we are able to rapidly understand novel scene images, little is known about the mechanisms that support this ability. Theories of optimal coding assert that prior visual experience can be used to ease the computational burden of visual processing. A consequence of this idea is that more probable visual inputs should be facilitated relative to more unlikely stimuli. In three experiments, we compared the perceptions of highly improbable real-world scenes (e.g., an underwater press conference) with common images matched for visual and semantic features. Although the two groups of images could not be distinguished by their low-level visual features, we found profound deficits related to the improbable images: Observers wrote poorer descriptions of these images (Exp. 1), had difficulties classifying the images as unusual (Exp. 2), and even had lower sensitivity to detect these images in noise than to detect their more probable counterparts (Exp. 3). Taken together, these results place a limit on our abilities for rapid scene perception and suggest that perception is facilitated by prior visual experience.

Keywords Scene understanding · Prior probability · Free-response

Research in high-level visual perception has shown that human observers have a truly impressive ability to recognize complex real-world scenes in a mere glance. Upon viewing a new scene for less than 250 ms, observers are able to name the scene at a semantic level (Potter, 1976), to categorize the scene (Torralbo et al., 2013; Walther, Caddigan, Fei-Fei, & Beck, 2009), to name a few large objects (Fei-Fei, Iyer, Koch, & Perona, 2007) including animals (Thorpe, Fize, & Marlot, 1996), to understand spatial properties such as depth (Gajewski, Philbeck, Pothier, & Chichka, 2010; Greene & Oliva, 2009) and affordance properties such as navigability (Greene & Oliva, 2009), and even to rate a scene for aesthetics (Kaplan, 1992). However, these studies may have biased participants toward success and overestimated our rapid scene understanding abilities: In addition to using highly typical stimuli, for which there are strong top-down expectations, most of the tasks have promoted or leveraged those expectations. For example, many studies have presented observers with a target class of scenes, such as scenes containing animals (Thorpe et al., 1996) or forest scenes (Greene & Oliva, 2009), and have asked observers to detect target scenes among the nontarget distractor scenes. However, such explicit categorization tasks provide a strong top-down signal biasing visual processing toward features that are diagnostic of the target class (Johnson & Olshausen, 2003; McCotter, Gosselin, Sowden, & Schyns, 2005). In other words, if an observer reports seeing (e.g.) an animal in a scene, we do not know whether this is because she has fully processed the image or because she detected diagnostic animal features (Evans & Treisman, 2005). Rapid scene understanding has also been evaluated by asking observers to write descriptions of briefly viewed images (Fei-Fei et al., 2007). Although this task may reflect a less biased view of what is understood from a brief glance at a scene, the results can still be influenced by expectations. What an observer writes depends not only on what she has perceived, but also on her inferences given the information she has gleaned. These inferences will, in turn, influence what she remembers, what she chooses to mention, and any guesses or assumptions that she makes. Because observers are prone to false recollections based on inference (Brewer & Treyans, 1981), this is a serious problem for the free-report paradigm.

Electronic supplementary material The online version of this article (doi:10.3758/s13414-015-0859-8) contains supplementary material, which is available to authorized users.

M. R. Greene (✉) · A. P. Botros · L. Fei-Fei
Department of Computer Science, Stanford University, 353 Serra Mall, Room 240, Stanford, CA 94305, USA
e-mail: mrgreene@stanford.edu

D. M. Beck
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Although theories of optimal coding, such as predictive-coding models, have posited that prior experience and expectations can be used to disambiguate complex visual input (Rao & Ballard, 1999), our survival depends on being able to rapidly and accurately detect novelty in the environment, and surprising information seems to guide visual attention (Walther & Koch 2006). Given the strong statistical regularity of the natural world (Olshausen & Field, 1996; Torralba & Oliva, 2003), these two coding principles are rarely in conflict. However, by examining how the visual system handles violations of visual expectations, we can understand the extent to which our first visual representations depend on matching the current input to stored representations of typical past experience.

In the present experiments, we presented observers with images of improbable real-world situations (or visually and semantically matched control images) and asked them to write a comprehensive description of everything that they saw in the scene (Fei-Fei et al., 2007). The free-response paradigm allows us to understand a participant's overall understanding of a scene, which includes more than just the scene's category and objects (Zelinsky, 2013). By comparing the descriptions of typical ("probable") and unusual ("improbable") scenes, we can disentangle perception from mere inference in rapid scene perception. Since the probable and improbable image pairs did not differ in terms of low-level visual features, the results could not be driven by bottom-up conspicuity or salience.

Our results indicated that observers strongly rely on prior probabilities in rapid scene perception: They failed to describe many of the unexpected details in the improbable scenes, while simultaneously writing in many false details (Exp. 1). Furthermore, these deficits appear to be perceptual in origin. Participants required a remarkably long image presentation time to reliably report that an improbable scene was unusual (Exp. 2), and they even had difficulties detecting briefly presented improbable images in noise (Exp. 3). Taken together, these results show that it takes observers much longer to understand and even perceive improbable visual images, indicating that our rapid scene categorization abilities depend critically on our prior experience with real-world environments, highlighting the importance of our lifetime of experience with typical environments to our ability to rapidly parse the complex visual world.

Experiment 1: Written descriptions

In order to understand how prior experience influences our ability to rapidly perceive scenes, we asked observers to write detailed descriptions of briefly viewed scenes that depicted either very-low-probability events in the world or visually matched images depicting more typical events.

Method

Materials

Image selection The image database consisted of 100 images, composed of 50 image pairs. Each pair contained an improbable image and a probable image that was hand-chosen to match the style, content, and structure of the improbable image as much as possible. Unusual images were collected from the Web and were chosen to depict low-probability real-world events that were free from overtly emotional content. Example image pairs are shown in Fig. 1. These images were screened from a larger set of images and rated by five observers for oddness as well as emotional content in a pilot experiment (see the Supplementary Materials for details). To the best of our knowledge, these images were real-world photographs and not the product of photo manipulation.

Image-based analysis: saliency and image feature differences In order to determine what (if any) influence

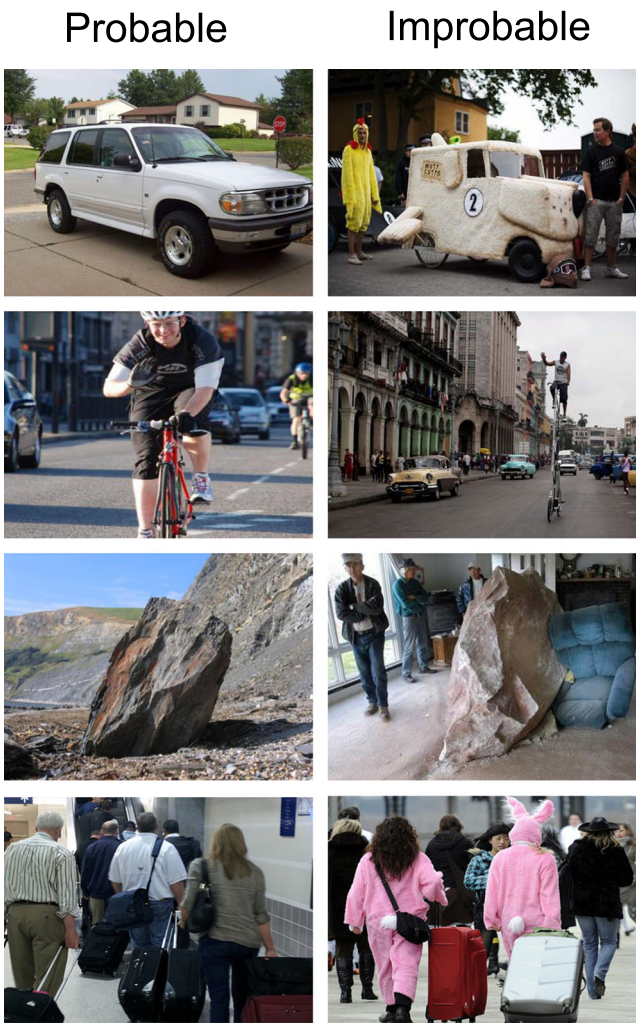

Fig. 1 Examples of matched probable and improbable image pairs

visual salience had on responses, we analyzed each of our images using the Itti and Koch (2000) saliency toolbox for MATLAB (Walther & Koch, 2006). We manually created tight bounding boxes around the central feature or concept most integral to the meaning of each image. We computed the area of each box and found no significant differences between the probable and improbable images [$t(49) < 1$]. We then assessed the mean and max saliency magnitude within the bounding boxes, and found no significant differences between the probable and improbable images in the mean saliency of these regions [$t(49) = 1.22$, $p = .23$], nor in the maximum [$t(49) < 1$]. Therefore, any differences in observers' perceptions of these images cannot be attributed to the salience of the images, nor to the spatial extent of the scenes' meaningful content.

In order to ensure that our probable and improbable images could not be distinguished according to low-level visual features, we computed four types of biologically relevant visual features for each of our images: color histograms, scene gist features, edge density, and multiscale Gabor filter weights.

Color histograms Images were converted from RGB into LAB color space, and two-dimensional histograms were created from the a* and b* channels of each image using 50 bins per channel (Oliva & Schyns, 2000).

Multiscale Gabor wavelets This model expresses an image's dominant orientations and spatial frequencies and is similar to those used to model responses in early visual areas (Kay, Naselaris, Prenger, & Gallant, 2008). Images were down-sampled to 128 × 128 pixels and convolved with a bank of Gabor filters at three spatial scales (3, 6, and 11 cycles per image with a luminance-only wavelet that covered the entire image), four orientations (0, 45, 90, and 135 deg), and two quadrature phases (0 and 90 deg). An isotropic Gaussian mask was used for each wavelet, with its size relative to spatial frequency such that each wavelet had a spatial frequency bandwidth of one octave and an orientation bandwidth of 41 deg. Wavelets were truncated to lie within the borders of the image.

Gist features These features represent summary statistics of scenes and represent a successful baseline for scene classification in computer vision. Images were down-sampled to 350 × 350 pixels and represented with the Gist descriptor of Oliva and Torralba (2001). This descriptor creates a summary representation of a scene by measuring the dominant orientations at multiple spatial scales, coarsely localized throughout the image plane.

Edge density Edge density was measured by summing the edge elements from a Canny edge map of each image. The probable and improbable images did not have significantly different edge densities [$t(49) < 1$]. Since this was a relatively coarse measurement, we also fit Weibull functions to the distribution of the edge contrasts for each image. The two parameters of the Weibull distribution have been shown to be useful for distinguishing among different types of scenes (Scholte, Ghebreab, Waldorp, Smeulders, & Lamme, 2009), and also seem to be driving early neural responses to scenes (Groen, Ghebreab, Prins, Lamme, & Scholte, 2013). However, our image set did not differ significantly in either the beta [$t(49) = 1.62$, $p = .11$] or the gamma [$t(49) = 1.8$, $p = .07$] parameters of the Weibull distribution.

SVM analysis Given the multidimensional natures of the color, Gabor, and gist features, we employed a classifier to test the extent to which these features could be used to distinguish the probable from the improbable images. The logic of this approach is that if a classifier can use a feature to predict whether an image is probable or improbable, the two image groups differ according to this feature, and human observers might make use of this difference in perception. On the other hand, an inability to classify the scenes by a given feature can be taken as evidence that the two image groups do not differ in terms of that feature.

The image features (color histograms, Gabor wavelets, or Gist descriptor) were fed into a support vector machine with a linear kernel. The task of the classifier was to predict whether an image depicted a probable or improbable situation. Each image was used separately for testing, with the remaining images being used for training. Both the wavelet and color histograms yielded 44% correct performance at classifying an image as probable or improbable (not different from chance, $p = .27$ binomial test). Gist features led to 45% correct classifications (not different from chance, $p = .38$). Combining all features yielded 42% correct performance (not different from chance, $p = .13$). Given the low level of performance and the simplicity of these features, we also trained an SVM classifier on the top-level features from a state-of-the-art neural network (Sermanet et al., 2013) to represent the best-case scenario for the contribution of low-level visual features (Razavian, Azizpour, Sullivan, & Carlsson, 2014). This classifier achieved 59% correct classifications (not better than chance, $p = .09$, binomial test). Taken together, these image-based analyses indicated that any observed differences between the improbable and probable image pairs were unlikely to be attributed to differences in the low-level visual features.

Image presentation The stimuli were presented at 15.8 × 10.8 deg of visual angle on a 21-in. CRT monitor (resolution 1,280 × 1,024) with an 85-Hz refresh rate. Pattern masks were created by making a texture of each experimental image using the Portilla and Simoncelli (2000) texture synthesis algorithm.

Participants

Ten participants (ages 19 to 25; seven male, three female; all native English speakers with normal or corrected-to-normal

vision) took part in Experiment 1. They provided informed consent and were compensated for their time.

Design and procedure

Each participant viewed 50 images total. Of these, 25 were improbable and 25 were probable images. Observers saw either the probable or the improbable version of each pair, and the version was counterbalanced across observers. Each image was viewed once for one of five presentation times (24, 47, 82, 153, and 506 ms), and the presentation times were counterbalanced across participants such that the final data set contained one written description of each image at each presentation time across the ten participants. Our sample size allowed us to examine our primary hypotheses concerning differences in image group (probable or improbable), while maintaining a reasonable workload for the participants who rated the image descriptions (see below).

The 50 images were shown to participants in a random order. Each trial commenced with a fixation point for 500 ms, followed by the experimental image, followed by a dynamic pattern mask of four pattern masks, chosen randomly from the set of masks, shown in an RSVP stream of 24 ms each (Greene & Oliva, 2009). Participants were instructed to type a detailed description of the image and to be as thorough and accurate as possible. In order to ensure that the descriptions were not abbreviated due to time pressure, participants were given a full hour to complete the experiment. They were not given any information about the types of images they would be viewing.

Assessing the written descriptions We used crowdsourcing to quantitatively evaluate the written descriptions. Workers on Amazon's Mechanical Turk (AMT) rated and assessed the quality of the text descriptions with respect to the photograph. Assessment was carried out in three different phases with 157 independent workers. Five individuals assessed each image and its associated description. Workers qualified for our task by having a previous approval rating at or equal to 98% for at least 2,000 previous AMT tasks. In addition, the potential workers were required to pass an extensive qualification and training session culminating in a graded exam. In the training, potential workers viewed detailed example trials along with explanations of the correct responses. The images that were used in the training were taken from the pool of nonexperimental images described in the Supplementary Materials. The tests were formulated exactly as the real assignments, and prospective workers were required to respond correctly to all of the test questions in order to gain eligibility to participate in real assignments. In addition to the 157 qualified workers, 73 workers attempted the training but failed.

In the first phase of assessment, workers viewed an image along with the text description given by one of the participants from Experiment 1. These workers were asked to rate the quality of the description from 0 (*very bad*) to 4 (*outstanding*). For the improbable images, workers were also asked to rate the degree to which the description captured the oddness of the scene, on a 0 (*did not understand at all*) to 3 (*understood completely*) scale. In order to assess observers' understanding of the objects and details within the images, workers were asked to click on keywords within the descriptions to indicate which words were object or scene names. In the second phase of assessment, AMT workers were asked to label keywords containing adjectives that described any of the object and scene terms identified in the first phase of the assessment. Descriptors included the number, appearance, emotion, action, and position of an object. Any descriptor that did not fit into these categories could be listed as "other." In the last phase of assessment, the workers indicated which of the previously identified keywords (objects, scenes and descriptors) were actually present in the image. For each stage of the assessment, five workers graded each response. For rankings, the average of the five participants' ratings was used, and for keywords, the ruling of the majority was used in the analysis.

In order to assess the completeness of the responses, one of the authors (A.B...) wrote ground-truth descriptions of each image after viewing for unlimited time. These responses were also graded by AMT workers using the previously described procedure. The probable and improbable ground-truth descriptions did not differ significantly in word length [$t(49) < 1$], number of keywords [$t(49) = 1.5$, $p = .14$], number of objects listed [$t(49) = 1.34$, $p = .19$], or the number of details or descriptors mentioned [$t(49) = 1.45$, $p = .15$].

Results and discussion

Example descriptions for two images can be found in Fig. 2a. Overall, participants wrote an average of 106 words per image (range: 8–743). Of these, 19 words on average were identified as being keywords (object name, scene name, or descriptor). The number of keywords increased significantly with increased presentation time [$F(4, 36) = 2.66$, $p < .0001$, $g\eta^2 = .38$] but did not vary with the probability of the image [$F(1, 9) < 1$], nor did probability interact with presentation time [$F(4, 36 = 1.01$, $p = .42$, $g\eta^2 = .008$]; see Fig. 2b.

Although participants wrote descriptions of similar length for both the probable and improbable scenes, the quality of these descriptions varied considerably. We found a significant main effect of scene type on the quality ratings, with descriptions of improbable images being rated as lower quality than those for the probable images ($M = 1.45$ vs. 1.88) [$F(1, 9) = 28.6$, $p < .001$, $g\eta^2 = .24$]. As expected, the ratings increased with presentation time [$F(4, 36) = 79.8$, $p < .0001$, $g\eta^2 = .71$]. We observed a marginal interaction between scene type and presentation time [$F(4, 36) = 2.58$, $p = .053$, $g\eta^2 = .20$], suggesting that description quality improved more with presentation time for the improbable than for the probable images; see

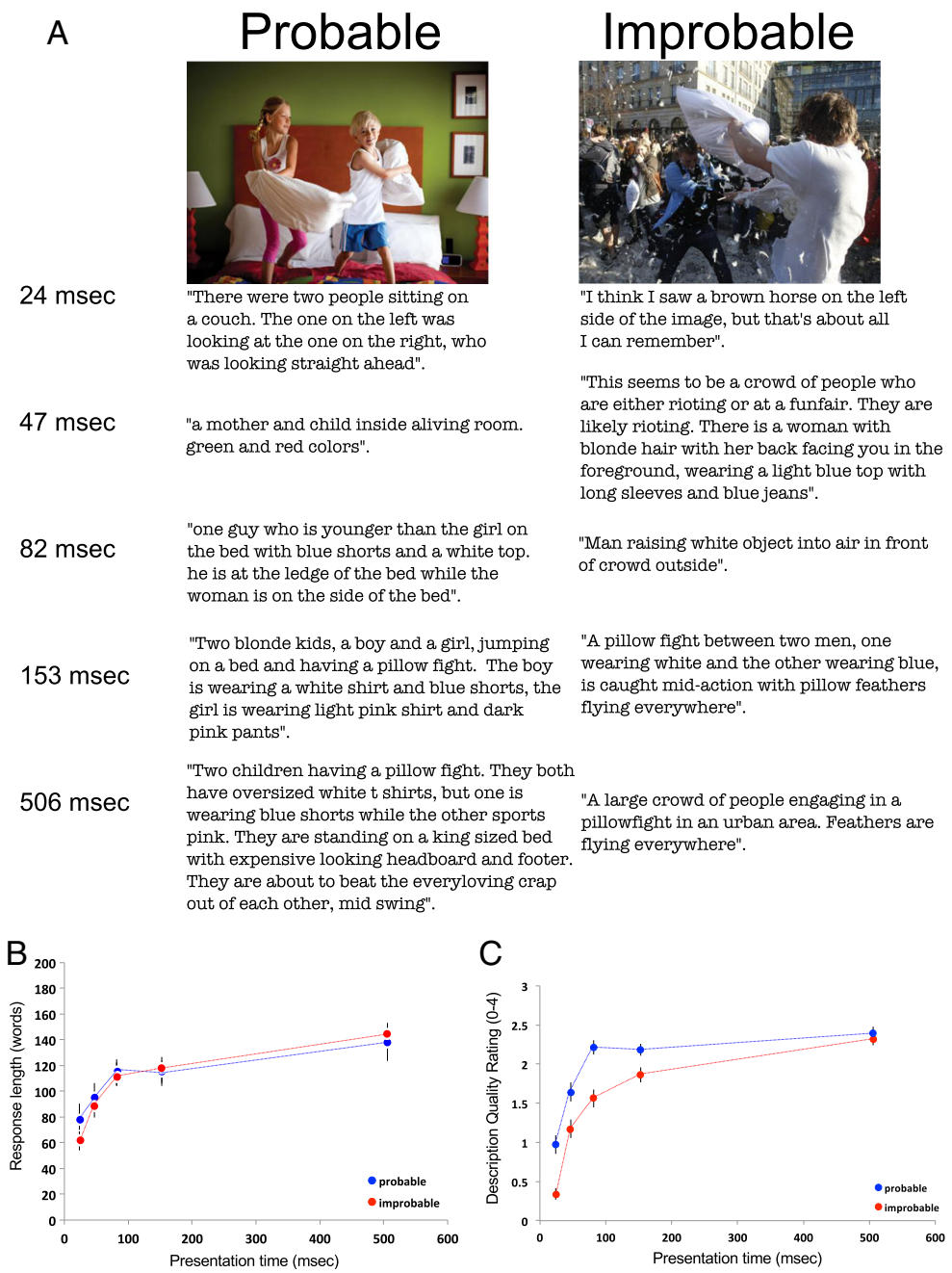

Fig. 2 (**a**) Sample free responses to a pair of images from each presentation time. (**b**) Numbers of keywords used for the probable and improbable images at each presentation time. (**c**) Description quality ratings for the probable and improbable images at each presentation time. Error bars indicate ±1 *SEM*

Fig. 2c. These differences were not found in the ground-truth descriptions, however, since these were rated as having similar quality by the AMT workers (M = 2.9 vs. 3.1, respectively) [$t(49) = 1.85$, $p = .07$]. If anything, the ground-truth descriptions of the improbable images were more highly rated than those of the probable images.

Why did observers write worse descriptions of the improbable images? In order to investigate, we examined (1) the degree to which the descriptions captured the unusual aspects of the image; (2) the number of items present in the images that were not mentioned (misses); and (3) the number of items mentioned by observers that were not actually present in the images (false alarms).

(1) Descriptions of improbable images did not capture oddness

For each improbable image, AMT workers rated how well the writer seemed to understand the oddness of the scene on a 0–3

scale. This assessment increased monotonically with presentation time, from an average of 0.12 for 24-ms presentations to 2.06 for 506-ms presentations. Tellingly, only nine of the 50 improbable images received a top score even at the longest presentation time, suggesting that even when viewing for 506 ms—enough time to have executed one saccade (Rayner & Pollatsek, 1992)—observers had insufficient time to fully understand the odd features of the improbable scenes.

Additional analyses revealed that the oddness ratings were not significantly modulated by the visual saliency of the unusual scene aspects. The overall quality score of an image description was not strongly related to the mean salience of the image ($r = .03$). Similarly, the oddness score of the improbable images was not strongly related to the image's mean salience ($r = .08$). Taken together, these results indicate that salience alone did not drive the quality of the scene descriptions.

(2) Descriptions of improbable images missed many important scene features

In order to examine omissions, we compared the experimental descriptions to the ground-truth descriptions. For each description, a hit rate was calculated as the proportion of keywords (objects, scene categories, and descriptive adjectives) identified in the ground-truth description that were correctly mentioned in the experimental descriptions. Observers had a higher hit rate for objects in probable than in improbable scenes ($M = .31$ vs. $M = .22$) [$F(1, 9) = 32.3$, $p < .0005$, $g\eta^2 = .24$], and the hit rate increased with presentation time [$F(4, 36) = 32.5$, $p < .0001$, $g\eta^2 = .55$], but there was no significant interaction between these two factors [$F(4, 36) < 1$; see Fig. 3a]. Paired t tests indicated that the improbable scenes had a significantly lower hit rate than their probable counterparts at every presentation time before 506 ms, suggesting that observers saw fewer objects in the unusual scenes. Although it is striking that observers reported fewer than half of a scene's objects on average, this result is consistent with the low estimates of working memory capacity for objects in scenes when guessing is controlled for (Liu & Jiang, 2005).

Since most participants did not write scene category terms in their descriptions, we could not compute reliable hit rates for these terms. Nonetheless, the results for scene terms are reported in the Supplemental Materials.

For descriptor terms (descriptions of appearance, actions, quantities, or positions for both objects and scenes), the hit rate also increased with presentation time [$F(4, 36) = 23.1$, $p <$

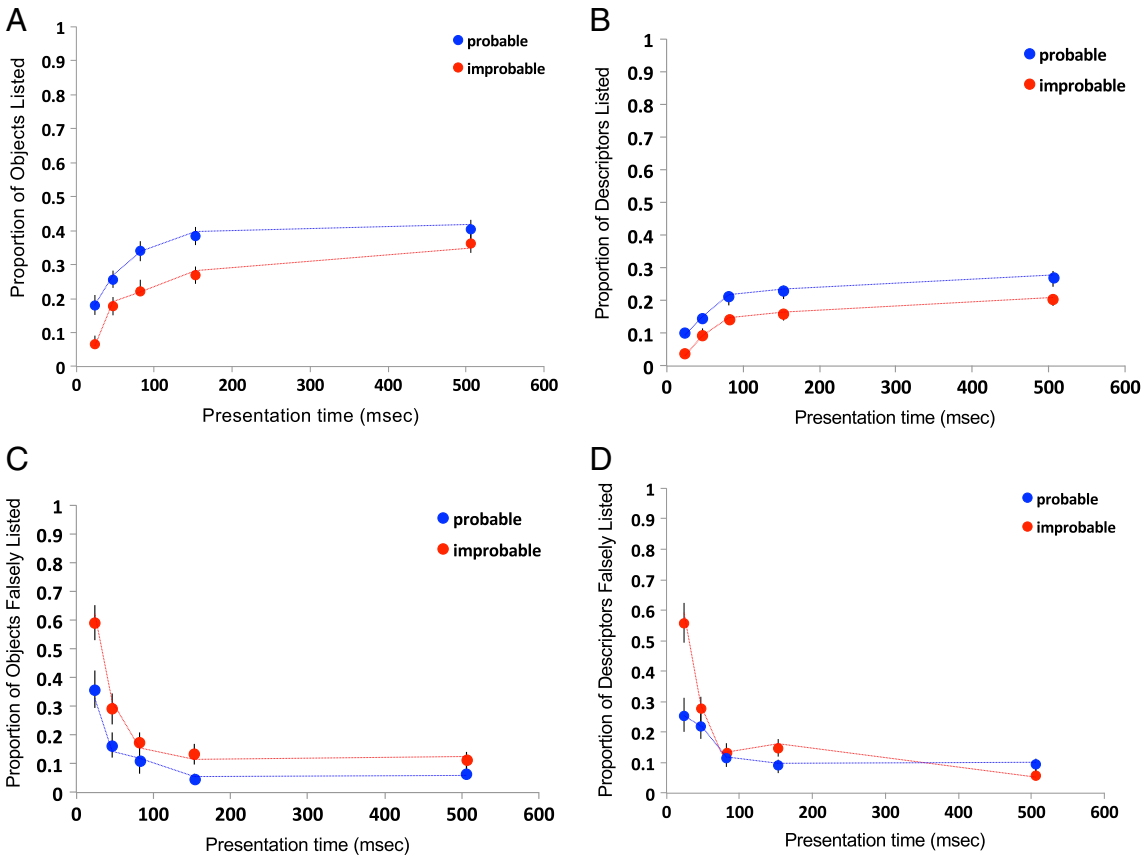

Fig. 3 (**a**) Proportions of objects named for the probable and improbable scenes at each presentation time. (**b**) Proportions of descriptors correctly named at each presentation time. (**c**) Proportions of objects falsely named at each presentation time. (**d**) Proportions of descriptors falsely identified at each presentation time. Error bars indicate ±1 *SEM*

.0001, $g\eta^2 = .49$], and probable images had a higher hit rate ($M = .19$) than improbable images ($M = .13$) [$F(1, 9) = 40.9$, $p < .0001$, $g\eta^2 = .22$]. However, again image typicality did not interact with presentation time [$F(4, 36) < 1$; see Fig. 3b]. Paired t tests indicated that improbable scenes had lower hit rates than did probable scenes at all presentation times.

Next, we examined the hit rates for each type of object descriptor (scene descriptors were omitted because they were rare). Overall, the hit rates were low and ranged from .22 (action) to .07 (position). The hit rate for each type of descriptor (action, appearance, position, and quantity) increased with presentation time (all ps $< .05$; see the Supplemental Materials). Although probable images resulted in higher hit rates for the action and appearance object descriptors (each $p < .005$; see the Supplement), probability did not significantly influence the hit rates of the position and quantity descriptors. No significant interactions were observed between presentation time and scene probability, as is shown in Fig. 3b. That position and quality descriptors were frequently missed is consistent with a number of other studies in scene perception that have demonstrated that we often do not pick up details in a glance (Fei-Fei et al., 2007). The especially low hit rate for position information is in line with the fact that observers are often not able to localize objects that they have detected in briefly presented scenes (Evans & Treisman, 2005), and that scene-selective areas such as the parahippocampal place area (PPA) are insensitive to mirror reversals of a scene that change the locations of objects within it (Dilks, Julian, Kubilius, Spelke, & Kanwisher, 2011).

(3) Descriptors of unusual images included many errors

To what extent did the image descriptions contain incorrect or fabricated details? For each type of keyword (object and descriptor), we computed a false alarm rate from the third phase of the AMT assessment, in which workers indicated whether a keyword was correct or incorrect. For object keywords, the participants had an average false alarm rate of .20 and a lower false alarm rate for probable ($M = .15$) than for improbable ($M = .26$) images [$F(1, 9) = 15.8$, $p < .001$, $g\eta^2 = .11$]. As expected, the false alarm rates decreased with increased presentation time [$F(4, 36) = 19.2$, $p < .0001$, $g\eta^2 = .52$]. We did not observe an interaction between image probability and presentation time [$F(4, 36) = 1.6$, $p = .19$, $g\eta^2 = .11$]; see Fig. 3c. Paired t tests indicated significantly more false alarms to objects in improbable than in probable scenes only at the 24-ms and 153-ms presentation times.

As with hit rates, we pooled all descriptors together to calculate an overall descriptor false alarm rate of .19. As expected, descriptor false alarms decreased with presentation time [$F(4, 36) = 20.7$, $p < .0001$, $g\eta^2 = .54$]. Additionally, the probable images had a lower false alarm rate ($M = .15$) than did the improbable images ($M = .23$) [$F(1, 9) = 14.8$, $p < .001$, $g\eta^2 = .06$], and there was a significant interaction between presentation time and image typicality [$F(4, 36) = 5.4$, $p < .005$, $g\eta^2 = .10$; see Fig. 3d].

Altogether, we found that written descriptions of briefly viewed improbable scenes were systematically poorer than descriptions written about matched typical scenes. To what extent are these deficits perceptual in nature? In Experiments 2 and 3, we examined this question in detail.

Experiment 2: Presentation time threshold

Experiment 1 demonstrated that participants' descriptions of briefly viewed improbable images were significantly worse than their descriptions of more probable images, suggesting that the fidelity of our initial scene representations depends on one's visual experience with the depicted environments and situations. Were these failures a result of having a less accurate perception of the improbable images, or merely of participants being unable to accurately describe the unusual aspects of the image? If the impoverished descriptions were due to writers' inability to express the strange situations, then we would expect to find equal performance for both scene types in a categorization task. However, if the description deficits were due to an impoverished initial representation, then we would expect observers to need more time to even categorize an image as improbable.

Method

Materials The stimuli consisted of the 100 images used in Experiment 1. We presented this experiment on a 21-in. CRT monitor (resolution 1,280 × 960 at 100 Hz). Stimuli subtended 15.8 × 11.6 deg of visual angle.

Participants A total of 21 participants (12 female, nine male; ages 18–36, with normal or corrected-to-normal vision) participated in Experiment 2. None of these individuals had participated in Experiment 1. They provided informed consent and were compensated for their time.

Design and procedure We employed a linear 3-up–1-down psychophysical staircase on image presentation times in order to determine how long participants needed to view an image to accurately classify it as probable or improbable. Observers viewed all 100 images from Experiment 1 in a random order, with the initial presentation time set to 100 ms. We employed separate psychophysical staircases on both groups of images in order to estimate the image durations needed to support the classification of images as probable versus improbable. We predicted longer durations for improbable images.

Participants were instructed that they would see images that depicted either scenes of typical, daily-life events or scenes that contained events and activities that were very improbable

in the world, and that they were to classify each scene as probable or improbable. Each trial commenced with a fixation point for 500 ms. A scene image was then shown for 10–200 ms, as required by the staircase, followed by a dynamic pattern mask that was identical to that of Experiment 1. The participant then indicated with a keypress whether the image was probable or improbable. No performance feedback was given. When a participant had answered three consecutive trials correctly in a given staircase, the presentation time of subsequent trials decreased by 10 ms (to a floor of 10 ms), whereas each incorrect answer resulted in an increase of presentation time by 10 ms.

For each participant and each condition, the final presentation time threshold was calculated as the average of four values: the final presentation time viewed, the minimum presentation time at which the observer could achieve at or above 75% correct, the mean presentation times of the last three correctly answered trials, and the modal presentation time. Since we did not observe any systematic differences in the threshold estimates from the four methods used [one-way analyses of variance: $F(3, 75) < 1$ for probable images; $F(3, 75) = 1.33$, $p = .27$, $g\eta^2 = .02$, for improbable images], averaging over multiple threshold estimates would produce a more robust estimation.

Results and discussion

Two participants were dropped from the analysis due to floor performance (>20% of trials at the maximum presentation time for either of the two staircases).

Critically, participants required significantly less viewing time to classify an image as probable ($M = 47$ ms) than to classify it as improbable ($M = 135$ ms) [$t(18) = 6.53$, $p < .0001$]; see Fig. 4. Interestingly, the presentation time required for the probable scenes was quite similar to the presentation times required to classify the scenes into basic-level categories using the same method (50 ms, range 30–67 ms; Greene & Oliva, 2009), whereas the 135-ms masked presentation time to classify an image as improbable represents a deviation from much of the data in scene understanding, which have almost universally shown outstanding performance in scene-understanding tasks.

We wanted to ensure that the large difference in thresholds for probable and improbable images was not due to a subset of observers being biased toward always classifying images as "probable." If some observers had a strong inclination to classify an image as probable, this could result in an artificially low threshold for probable images and an artificially high threshold for improbable ones. Of the 19 observers included in the analysis, we identified only four who had a significant bias toward responding "probable" using a binomial test. When we reanalyzed the data without these observers, we still found a significantly longer presentation time threshold for improbable ($M = 126$ ms) than for probable ($M = 54$ ms) images [$t(14) = 5.4$, $p < .0005$]. Although this pattern of results cannot solely be attributed to a few biased participants, we could not fully rule out bias in this experiment. Many of the participants had a slight tendency toward answering "probable," possibly driving up the presentation times for improbable images while driving down the presentation times for probable images. However, the magnitude of the duration effect (improbable presentation threshold minus probable duration threshold) showed no correlation with the bias of the observer ($r = .01$), so it is unlikely that the difference between the groups was due solely to bias.

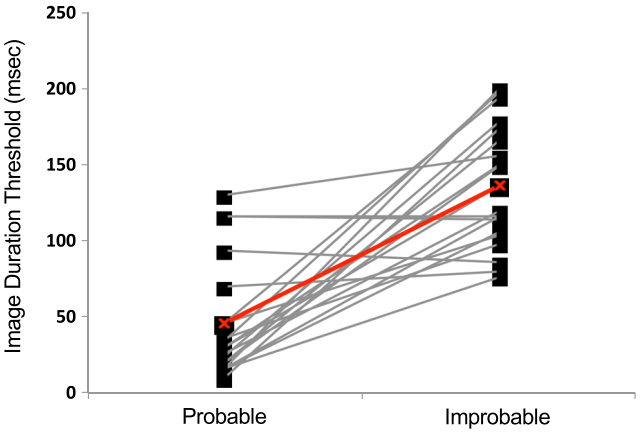

Fig. 4 Distribution of presentation time thresholds for probable and improbable images in Experiment 2. Individual participants are shown in gray, whereas the mean of all participants is shown in the central color line

Overall, Experiment 2 demonstrated that although participants can detect scene improbability in a glance, this classification task requires a great deal more image exposure than do most scene classification tasks. Although some observers had a significant bias toward classifying a scene as probable, observers without this bias still needed to view improbable images for more than twice as much time as the probable images in order to make reliable classifications. These results suggest that the poor descriptions observed in Experiment 1 were not due solely to observers' inabilities to describe the unusual scenes in words, and they give credence to the idea that the deficits are perceptual in nature. At what stage of perceptual processing does image probability begin to exert an effect? Experiment 3 tested the extent to which even image detection can be influenced by the prior probability of the scene.

Experiment 3: Detection

So far, we have shown that human rapid scene perception abilities are modulated by scene probability: Descriptions of briefly viewed improbable scenes were markedly error-prone relative to matched probable scenes (Exp. 1), and observers

needed to view an improbable scene for well over 100 ms before they understood that anything was amiss (Exp. 2), suggesting that the deficits observed in the first experiment were not simply due to observers' inability to verbally express what they had seen, or to the fleeting nature of conceptual short-term memory (Potter, 1976). However, we were still left with the fundamental question of how early in perception image probability has an effect. We reasoned that if the deficits occur early in perceptual processing, then observers should have less sensitivity to detect improbable images in noise, a low-level task. In Experiment 3, we presented participants with trials consisting of either a phase-randomized scene image or one of the experimental images, and tested for observers' detection sensitivities for both probable and improbable images.

Method

Materials The stimuli consisted of the 100 scene images used in Experiments 1–2, as well as full-color, fully phase-randomized versions of these scenes. Additional images for the practice session were taken from the original set of probable images that were not chosen for the final set (see the Supplementary Materials).

Participants Fourteen participants (eight female, six male; ages 19–30, with normal or corrected-to-normal vision) took part in Experiment 3. None of these participants had taken part in Experiment 1 or 2, and all provided informed consent and were compensated for their time.

Design and procedure After a short block of three practice trials to familiarize participants with the task, the experiment commenced with a block of 100 trials designed to obtain the 75% presentation time thresholds for detecting an image as a scene or phase-randomized noise image. Half of the trials were real-world scenes, and half were full-color phase-scrambled versions of these scenes. Each image was followed by the same dynamic pattern mask used in the previous experiments, based on probable scenes that were not used in the main experiment. As with Experiment 2, the initial presentation time was 100 ms, and a 3-up–1-down psychophysical staircase was employed to keep observers at 75% correct. The threshold for each observer was defined as the lowest presentation time with at least 75% correct performance. Observers were instructed to respond with a keypress as to whether an image was "intact" versus "scrambled."

After the threshold time was determined, each observer viewed an experimental block of 200 images, consisting of the 100 experimental images and 100 phase-randomized versions of these scenes; see Fig. 5a. The images were shown in a randomized order and viewed for the presentation time determined in the initial block. Each image was followed by the dynamic pattern mask. Performance feedback was given during practice trials only.

Results and discussion

The average presentation time threshold across participants was 28.9 ms (range: 10–80 ms). The data from one participant were removed from the analysis due to near-ceiling performance in the main block (92% correct), indicating that the presentation time threshold was overestimated during staircasing.

In the experimental block, we found that observers' detection sensitivities (using d') for probable images were higher than those for improbable images ($M = 2.21$ vs. 1.78) [$t(12) = 6.5$, $p < .0001$; see Fig. 5b]. All but one observer had a higher d' for probable images.

This result indicates that it was more difficult for observers to distinguish the improbable images from noise than to distinguish their probable counterparts. Although some researchers have found similar information requirements for detection and categorization (Grill-Spector & Kanwisher, 2005), others have found that categorization tasks can be made more difficult without affecting detection (Mack, Gauthier, Sadr, & Palmeri, 2008). Our results show that both categorization and detection are negatively affected in cases in which we cannot rely on past experience and expectations. Although category typicality has been shown to influence scene categorization in the same task (Torralbo et al., 2013), the probable images in our study were chosen to be as similar as possible to their improbable pairs. Additionally, image-level analyses indicated that our images could not be classified as probable or improbable on the basis of low-level visual features such as color or edge density. Taken together, these results indicate that subtle differences between scenes can influence our abilities to rapidly recognize not only their content, but also that they are in fact scenes. These results also strongly support the view that our rapid scene recognition abilities are aided by our familiarity with the visual inputs.

General discussion

These three experiments demonstrate that human observers cannot accurately understand very improbable visual scenes in a brief glance. Observers wrote poorer descriptions of improbable scenes (Exp. 1), had more difficulty in classifying situations as improbable (Exp. 2), and even had difficulties in determining that briefly presented improbable images were scenes, rather than phase-randomized noise (Exp. 3). Altogether, these results support the view that rapid scene perception is aided by our lifetime of visual experience in typical environments, as has been suggested by theories of optimal visual coding.

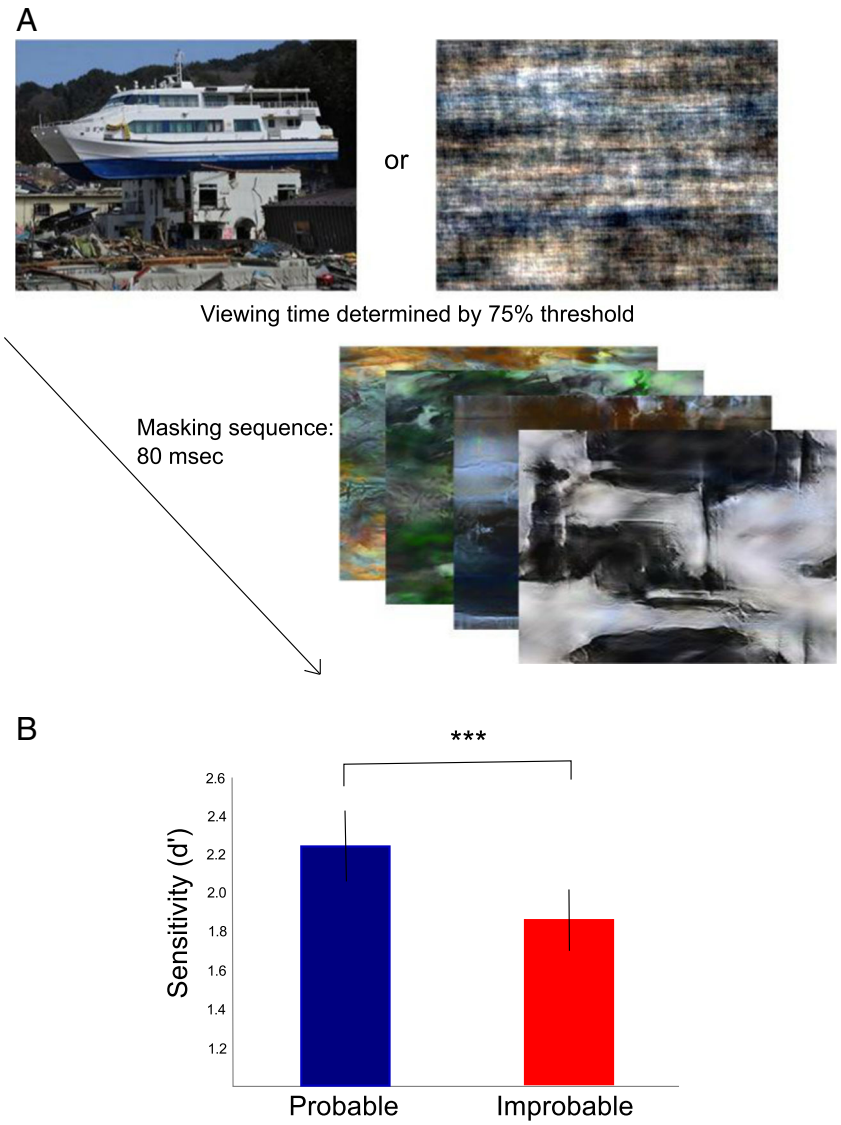

Fig. 5 (**a**) Image presentation sequence for Experiment 3. (**b**) Detection sensitivities (*d'*) for probable and improbable images. Error bars indicate ±1 *SEM*

In this work, we examined scenes that we had classified as "probable" or "improbable" real-world situations, rather that focusing on typicality within a scene category, such as beach scenes. Although we feel that this encompassing definition is more true to the overall meaning of a scene (Zelinsky, 2013), we can draw parallels between our definition of typicality and other scene perception studies that have examined the role of either category typicality (Ehinger, Xiao, Torralba, & Oliva, 2011; Torralbo et al., 2013) or object typicality (Biederman, Mezzanotte, & Rabinowitz, 1982; Davenport & Potter, 2004; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008; Rémy et al., 2013; Võ & Henderson, 2009) on scene understanding. It has been shown that images that are more typical exemplars of a scene category are recognized better by both human observers (Torralbo et al., 2013) and computer vision classifiers (Ehinger et al., 2011), possibly because typical images have lower within-class variability (Torralbo et al., 2013). Similarly, priming studies have shown facilitated visual processing for stimuli that conform to an observer's expectations (Eger, Henson, Driver, & Dolan, 2007; Esterman & Yantis, 2010). However, in these studies, low-level differences between typical and atypical images could have driven the results. Our results show that observers are better at recognizing more-typical situations, even when these images cannot be distinguished from images of improbable situations on the basis of low-level visual features, and thus they support the view that our visual experience allows us to create more-efficient visual codes.

Evidence from the eye movement literature also supports the notion that unusual visual stimuli lead to slower perceptual and cognitive processing. Although visually salient inconsistent objects can draw early saccades (Becker, Pashler, & Lubin, 2007; Loftus & Mackworth, 1978; Underwood,

Templeman, Lamming, & Foulsham, 2008), most studies have shown that first saccades are not systematically drawn to inconsistent objects (Henderson, Weeks, & Hollingworth, 1999; Rayner, Castelhano, & Yang, 2009; Võ & Henderson, 2009), suggesting that the objects might not be fully understood in the first fixation. Furthermore, inconsistent objects are fixated longer than other objects (Becker et al., 2007; Henderson et al., 1999; Loftus & Mackworth, 1978; Rayner et al., 2009; Underwood et al., 2008; Võ & Henderson, 2009), possibly indicating the increased visual or cognitive processing necessary to understand these objects. Taken together, these results support the view that unusual stimuli require processing above and beyond that needed by more typical stimuli. This literature has remained controversial, due in part to the difficulties with controlling experimenter-manipulated images for visual salience and conspicuity. The present results demonstrate that observers show deficits in understanding unusual visual scenes from the real world that are not distinguishable from typical images solely by their low-level visual features.

In contrast to the relative difficulties of initially recognizing atypical input, unusual elements tend to be remembered better once they are recognized (Isola, Xiao, Parikh, Torralba, & Oliva, 2013; Lampinen, Copeland, & Neuschatz, 2001). Indeed, participants' memories for objects that are consistent with a particular scene schema can even be at chance (Lampinen et al., 2001), suggesting that observers may adopt an efficient memory-encoding scheme in which attention is preferentially directed toward the least typical elements of an image and the remainder of the image is encoded as a schema (Brady, Konkle, & Alvarez, 2011). Taken together, these results suggest that the goals of top-down processing for visual perception are different from the goals of top-down processing for memory: Vision uses expectations to disambiguate complex input, whereas memory uses expectations to shorten the representation in memory, sometimes at the expense of representational uniqueness.

Future work will examine the generation of this schema over time. The keywords in Experiment 1 were assessed in a binary way—as either correct or incorrect. However, an examination of the descriptions shows that some incorrect responses were somewhat understandable, whereas others seemed completely random. For example, the participant who viewed the urban pillow fight image for 47 ms described the image as a "riot" (see Fig. 2), and indeed, urban pillow fights and riots share some featural and conceptual similarities. Many of the responses seemed to be conceptually driven. For example, for an image of two women sitting at a "bathroom themed" restaurant with seats made from toilets and a table made from a sink, all participants wrote about the "restaurant," but no participants made mention of the unusual seats or table. Other responses seemed driven by features from the visual masks. For example, a participant viewing the top-right improbable image from Fig. 1 described this image as ". . . another movie picture. There is a fire on the right. The scene is generally blue and orange. There is a house on fire and running people on the left." Still other descriptions defy simple classification. A participant viewing a row of Batmobiles driving down the road described the image as "A dead German Shepard" (sic). Therefore, future work will examine these false responses directly, asking how many of the incorrect responses were driven by assumptions given by scene schema? Did these occur at all presentation times? Understanding the type of inference errors made during rapid or noisy perception will be key to understanding errors made in the eyewitness testimony of events.

The present results challenge the generally accepted view that scene "gist" understanding can take place within a feedforward "sweep" through the ventral visual stream (Fei-Fei et al., 2007; Potter, Wyble, Hagmann, & McCourt, 2014; Thorpe et al., 1996). Why, then, are all scenes not created equally for perception? One possibility is that these deficits are more related to visual working memory than to perception. Perhaps observers can keep fewer objects in the improbable scenes in working memory because the probable scenes afford more opportunities to create compressed memory representations (Brady, Konkle, & Alvarez, 2009). Although this hypothesis could explain the patterns of results obtained in Experiments 1 and 2, this view cannot explain why the observers in Experiment 3 were worse at merely detecting improbable scenes. Instead, the results imply that the initial visual representation activates a template of a frequently viewed scene that acts in a top-down manner to refine the perceptual representation (Bar et al., 2006; Summerfield et al., 2006). This view would predict lower detection performance for improbable images because these presumably have lower resemblance to stored templates. Such a view is also compatible with predictive-coding schemes in which the probable scene is accommodated more quickly by upstream "expectations," thus providing a smaller error signal (and fewer iterations) through the cortical hierarchy in order to achieve recognition.

More broadly, being able to rapidly assess and react to novel or unexpected events is critical to our survival. However, our results suggest that this ability is limited. We deal with the difficult perceptual situations presented in rapid visual presentations by relying on top-down knowledge gleaned from a lifetime of experience in typical environments.

Although future work will be needed to understand why particular errors occurred, the data presented here clearly indicate that not all natural images can be accurately apprehended with brief presentations. The inadequacy of a single glance for understanding unusual visual input is even typified in the well-known comedic device known as the "double take"—an actor turns around to look again at an unusual event that he initially ignored. The results of these experiments show that our initial perception of unusual events is poor, so we all can benefit from a second glance.

Author note This work was funded by NRSA Grant No. NEI F32EY019815 (to M.R.G.) and National Institutes of Health Grant No. 1 R01 EY019429 (to D.M.B and L.F.-F.). M.R.G., A.P.B., D.M.B. and L.F.-F. conceived the ideas and designed the experiments. M.R.G. and A.P.B. collected and analyzed the data. A.P.B. created the AMT analysis tool. Finally, M.R.G., A.P.B., D.M.B., and L.F.-F. interpreted the results, and M.R.G. wrote the manuscript with critical input from A.P.B., D.M.B., and L.F.-F.

References

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*, 449–454. doi:10.1073/pnas.0507062103

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 20–30. doi:10.1037/0096-1523.33.1.20

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177. doi:10.1016/0010-0285(82)90007-X

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology. General, 138,* 487–502. doi:10.1037/a0016797

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision, 11*(5), 4. doi:10.1167/11.5.4

Brewer, W. F., & Treyans, J. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13,* 207–230.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559–564. doi:10.1111/j.0956-7976.2004.00719.x

Dilks, D. D., Julian, J. B., Kubilius, J., Spelke, E. S., & Kanwisher, N. (2011). Mirror-image sensitivity and invariance in object and scene processing pathways. *Journal of Neuroscience, 31,* 11305–11312. doi:10.1523/JNEUROSCI. 1935-11.2011

Eger, E., Henson, R. N., Driver, J., & Dolan, R. J. (2007). Mechanisms of top-down facilitation in perception of visual objects studied by fMRI. *Cerebral Cortex, 17,* 2123–2133. doi:10.1093/cercor/bhl119

Ehinger, K. A., Xiao, J., Torralba, A., & Oliva, A. (2011). Estimating scene typicality from human rankings and image features. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Expanding the space of cognitive science: Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2562–2567). Austin, TX: Cognitive Science Society.

Esterman, M., & Yantis, S. (2010). Perceptual expectation evokes category-selective cortical activity. *Cerebral Cortex, 20,* 1245–1253. doi:10.1093/cercor/bhp188

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1476–1492. doi:10.1037/0096-1523.31.6.1476

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.

Gajewski, D. A., Philbeck, J. W., Pothier, S., & Chichka, D. (2010). From the most fleeting of glimpses: On the time course for the extraction of distance information. *Psychological Science, 21,* 1446–1453. doi:10.1177/0956797610381508

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20,* 464–472. doi:10.1111/j.1467-9280.2009.02316.x

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16,* 152–160. doi:10.1111/j.0956-7976.2005.00796.x

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., & Scholte, H. S. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience, 33,* 18814–18824. doi:10.1523/JNEUROSCI. 3128-13.2013

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228. doi:10.1037/0096-1523.25.1.210

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36,* 1469–1482. doi:10.1109/TPAMI.2013.200

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506. doi:10.1016/S0042-6989(99)00163-7

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*(7), 4. doi:10.1167/3.7.4

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, *8*(13), 11:1–18. doi:10.1167/8.13.11

Kaplan, S. (1992). Environmental preference in a knowledge-seeking, knowledge-using organism. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 581–598). New York, NY: Oxford University Press.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452,* 352–355. doi:10.1038/nature06713

Lampinen, J. M., Copeland, S. M., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1211–1222. doi:10.1037/0278-7393.27.5.1211

Liu, K., & Jiang, Y. (2005). Visual working memory for briefly presented scenes. *Journal of Vision, 5*(7), 650–658. doi:10.1167/5.7.5

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology. Human Perception and Performance, 4,* 565–572. doi:10.1037/0096-1523.4.4.565

Mack, M., Gauthier, I., Sadr, J., & Palmeri, T. (2008). Object detection and basic-level categorization: Sometimes you know it is there before you know what it is. *Psychonomic Bulletin & Review, 15,* 28–35. doi:10.3758/PBR.15.1.28

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (2005). The use of visual information in natural scenes. *Visual Cognition, 12,* 938–953. doi:10.1080/13506280444000599

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41,* 176–210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175. doi:10.1023/A:1011139631724

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609. doi:10.1038/381607a0

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

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509–522. doi:10.1037/0278-7393.2.5.509

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics, 76,* 270–279. doi:10.3758/s13414-013-0605-z

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2,* 79–87. doi:10.1038/4580

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Eye movements when looking at unusual/weird scenes: Are there cultural differences? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35,* 254–259. doi:10.1037/a0013508

Rayner, K., & Pollatsek, A. (1992). Eye movements and scene perception. *Canadian Journal of Psychology, 46,* 342–376. doi:10.1037/h0084328

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN Features off-the-shelf: An astounding baseline for recognition. Retrieved from arxiv.org/abs/1403.6382

Rémy, F., Saint-Aubert, L., Bacon-Macé, N., Vayssière, N., Barbeau, E., & Fabre-Thorpe, M. (2013). Object recognition in congruent and incongruent natural scenes: A life-span study. *Vision Research, 91,* 36–44. doi:10.1016/j.visres.2013.07.006

Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., & Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision, 9*(4), 29. doi:10.1167/9.4.29

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). OverFeat: Integrated recognition, localization and detection using convolutional networks. Retrieved from arxiv.org/abs/1312.6229

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science, 314,* 1311–1314. doi:10.1126/science.1132028

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522. doi:10.1038/381520a0

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14,* 391–412. doi:10.1088/0954-898X_14_3_302

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,* e58594. doi:10.1371/journal.pone.0058594

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17,* 159–170. doi:10.1016/j.concog.2006.11.008

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 24. doi:10.1167/9.3.24

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. doi:10.1523/JNEUROSCI. 0559-09.2009

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks: The Official Journal of the International Neural Network Society, 19*(9) 1395–407. doi:10.1016/j.neunet.2006.10.001

Zelinsky, G. J. (2013). Understanding scene understanding. *Frontiers in Psychology, 4,* 954. doi:10.3389/fpsyg.2013.00954