Short Communication

Estimations of object frequency are frequently overestimated

Michelle R. Greene

Stanford University, Department of Computer Science, Minerva Schools at KGI, School of Computational Sciences, United States

ARTICLE INFO

ABSTRACT

Real-world scenes are complex but lawful: blenders are more likely to be found in kitchens than beaches, and elephants are not generally found inside homes. Research over the past 40 years has demonstrated that contextual associations influence object recognition, change eye movement distributions, and modulate brain activity. However, the majority of these studies choose object-scene pairs from experimenters' intuitions because the statistical relationships between objects and scenes had yet to be systematically quantified. How do intuitive estimations compare to actual object frequencies? Across six experiments, observers estimated the frequency with which an object is found in a particular environment, such as the frequency of "mug" in an office. Estimated frequencies were compared to observed frequencies in two fully labeled scene databases (Greene, 2013). Although inter-observer similarity was high, observers systematically overestimated object frequency by an average of 32% across experiments. Altogether, these results speak to the richness of scene schemata and to the necessity of measuring object frequencies.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In the last 40 years, a growing body of research has highlighted our remarkable scene understanding capabilities (Biederman, 1972; Greene & Oliva, 2009; Intraub, 1981; Li, VanRullen, Koch, & Perona, 2002; Potter & Levy, 1969; Potter, Wyble, Hagmann, & McCourt, 2014; Thorpe, Fize, & Marlot, 1996). These results present challenges to leading theories of visual perception that are built on results from sparse laboratory displays. Specifically, visual search appears to be more efficient in scenes (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011) compared to unstructured displays, and our recall for complex scenes is no worse than that of single objects (Konkle, Brady, Alvarez, & Oliva, 2010). It has been proposed that stored knowledge of scene-object relations aids these abilities (Bar, 2004; Wolfe, Võ, Evans, & Greene, 2011), but testing this hypothesis requires knowing both the strength of the contextual relations, and extent to which human observers accurately encode them. While recent work has addressed the first issue by characterizing the statistical frequencies of objects in scenes (Greene, 2013), the current work aims to address the second: how accurate are human object frequency estimates?

I asked human observers to estimate the frequency with which they would find various objects in different environments. I then compared these estimates to the observed frequencies from the Greene (2013) databases. Several possible patterns of results may

be expected. On one hand, we readily learn statistical relationships and contingencies from visual displays (Brady & Oliva, 2008; Brockmole & Võ, 2010; Chun & Jiang, 1998), and our frequency estimates of physical entities is generally good (Attneave, 1953; Tversky & Kahneman, 1973). Furthermore, scene and object recognition are subject to strong contextual effects (Davenport & Potter, 2004) even though the contextual relationships are experimenter-defined. Taken together, these results would predict near-veridical estimations of object frequency.

Contrastingly, human frequency estimation in other domains can be badly skewed. Highly representative items can cause observers to discount prior probabilities (Tversky & Kahneman, 1973), we exhibit an optimistic bias when estimating task durations (Kahneman & Tversky, 1977), and we overestimate salient, low probability events such as airplane crashes (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). Within vision, our place memories are schematic, and observers often report objects in a display that, while absent, fit within the scene context (Brewer & Treyans, 1981; Lampinen, Copeland, & Neuschatz, 2001). Similarly, we tend to remember a more expansive scene view than was presented (boundary extension, (Intraub & Richardson, 1989)), suggesting that our memories reflect constructions and simulations of the environment as a whole. These results predict that our estimates of object frequency may be overestimated.

Here, I demonstrate that human estimates of object frequency are consistently overestimated, despite high agreement between observers. This holds over experiments held in the laboratory

E-mail address: mgreene@minerva.kgi.edu

and online, and in the normative rankings of previously published work. This holds over multiple labeled databases, so these results are not due to one biased sample. Together, these results highlight the deeply constructive nature of human scene representations.

2. Methods

2.1. Free response task

10 participants (5 male, ages 22–35) were provided with a 16-page packet with the name of a scene category listed on top. The categories were taken from (Greene, 2013) and included *bathroom, bedroom, conference room, dining room, kitchen, living room, office, skyscraper, city center, street, highway, coast, open country, mountain* and *forest*. The order of categories was randomized across participants. For each category, participants were asked to freely list all the objects they could think of that were found in that scene category with the following frequencies: "always", "often", "sometimes" and "never". Here, an "object" was defined as any visual, nameable entity, following (Greene, 2013). Participants were encouraged to list at least ten instances per frequency bin whenever possible. They were allowed to take the packet home and to bring it back completed one week later. Participants were not given any instructions about the ordering of objects.

The lists were transcribed into a spreadsheet, with spelling errors corrected. Using WordNet (Miller, 1995), I unified synonyms. The raw tabulations were translated into estimated frequencies by coding object frequencies from 3 to 1 for "always", "often" and "sometimes" ratings respectively, with the weight of -1 to objects listed in the "never" frequency bin. For each object, I tabulated the number of participants who listed that object along with the average value from these codes. I multiplied these two values together to get an overall weight for each object, and then divided each value by the maximal score (30, reflecting the situation in which all 10 observers rate the object as "always" present). This accounts for the number of participants naming the object as well as the estimated frequency within each participant.[1]

2.2. Online object frequency rating experiments

In the online experiments, observers directly estimated the frequency of a provided scene-object pair. I chose 30 objects for each scene category. Of these, ten objects were very frequent (most frequent objects observed in (Greene, 2013)), ten were of moderate frequency (most common objects that were the nearest to 0.5 frequency) and the remaining ten were very low frequency objects (objects that had <0.05 frequency but had at least 10 instances in the (Greene, 2013) database).

In two experiments conducted on Amazon's Mechanical Turk (AMT), observers were presented with a category name along with an object name and a frequency slider that could be set between 0 (never found in category) and 100 (always found in category). In the first experiment, participants were instructed to estimate how often the object would be found in a photograph of the target scene category using the sliding bar. In the second experiment, participants were explicitly told to think of a single, photograph-like view of a scene that might not show every object present in the environment.

43 individuals participated in the first experiment while 46 new

individuals took part in the second experiment. All were US-based workers with at least 2000 previously submitted hits with at least 98% approval. Participants completed between 1 and 480 estimations in both experiments, with a median of 32 per participant in the first experiment, and 36 per participant in the second. Participants were compensated $0.01 per trial for their time. All submitted work was analyzed.

2.3. Observed object frequencies

Observed object frequencies were taken from the labeled databases of Greene (2013). In these databases, photographs of scene environments (at least 95 per category) were hand-labeled by trusted observers such that each pixel in each scene was annotated with an object label. Object frequency was defined as the proportion of category images containing the given object. The main database contained images from the web, while the auxiliary database contained images uploaded by researchers.

3. Results

### 3.1. Free response experiment

Each participant listed an average of 39.9 (SD: 2.5) objects in each of the 16 basic-level categories. Any repeated items within a category were omitted, along with any non-visual item such as "air". This left an average of 38.9 (SD: 3) objects per participant per category. Across participants, 1,086 unique objects were listed, 390 of which were listed in the database of (Greene, 2013).

3.1.1. Consistency of objects listed as present

How similar are observers' listings? For each category, I tabulated the number of participants listing a particular object as "always", "often" and "sometimes" present. On average, an object was listed by 2.8 observers (SD: 2.4). Each category had at least one object listed by all 10 observers. To what extent does this behavior deviate from random listing? To simulate chance, I considered the case of an observer randomly listing objects, without replacement, from a set of 30,000 possible objects, estimated by (Biederman, 1987). In each of 10,000 simulations, ten simulated observers were created, and the number of unique objects listed and the maximum number of simulated observers listing a particular object were recorded. A Wilcoxan sign-rank test indicated that the maximum number of listings for each category was significantly more consistent than was observed in the chance simulation (all $p < 0.0001$, see Fig. 1A). Additionally, I examined the amount of overlap (intersection over union) in the objects listed as "always" present between each pair of observers. The mean overlap was 0.3 (range: 0.18–0.44 across categories), and this was found to be significantly higher than what would be the maximum expected from chance (0.08) from permutation tests. Therefore, human observers have more similar object listings than would be expected from chance alone.

3.1.1.1. Consistency of objects listed as never present. I next applied the same analysis to the objects listed as "never" present in each category. Overall, there was less consistency among observers, with any given object listed by an average of 1.4 observers (SD: 0.95). However, when comparing these listings to the simulated chance condition, each category demonstrated inter-subject consistency that was higher than expected from chance (all $p < 0.0001$, see Fig. 1B). Remarkably, observers share not only consistent notions of the objects that belong in a scene category, but also the objects that *never* belong in that scene category.

[1] Estimated object frequencies were also calculated in two additional ways: (1) Giving "always" a frequency of 0.95, "often" a frequency of 0.75 and "sometimes" a frequency of 0.5 with 0 given to "never"; (2) Giving "always" a frequency of 0.95, "often" a frequency of 0.6 and "sometimes" a frequency of 0.3 with 0 given to "never". However, as these other methods were not well-correlated with observed frequencies ($r = 0.39$ and $r = 0.42$ respectively), they were not used for further analysis.
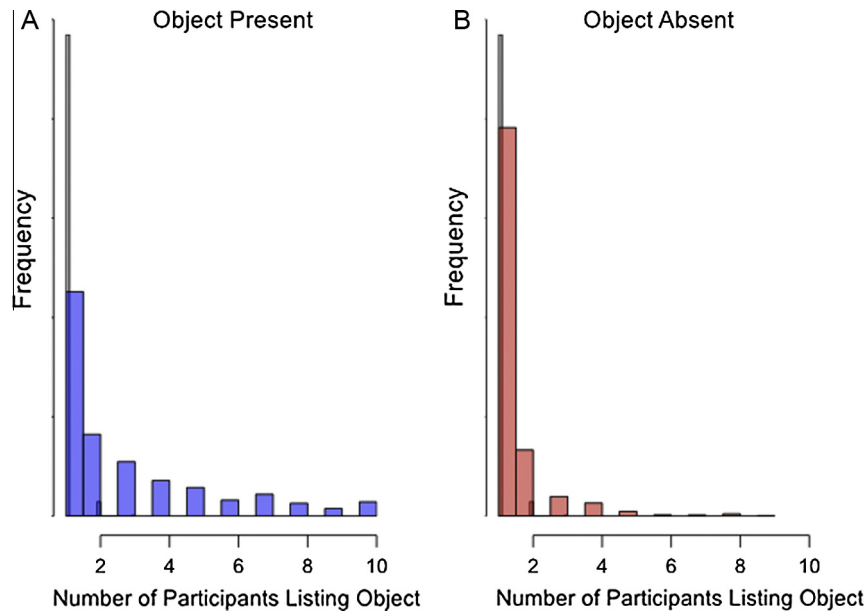
Fig. 1. (A) Histogram of participant agreement in objects listed as present in a scene category (blue). Chance-level agreement is shown in gray. (B) Histogram of participant agreement in objects listed as absent in a scene category (red) with chance-level agreement shown in gray. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Why does this consistency exist? As category order was randomized across participants, the consistency cannot be attributed to priming from a previously viewed category. Perhaps participants listed objects that are diagnostic of a different scene category (e.g. a "toilet" in a bedroom). I compared the mutual information between the never-present objects and compared it to the mutual information for all objects in the database. Objects with high mutual information give the most discriminative evidence for or against a scene category. I found that the objects listed as absent had higher mutual information for distinguishing among scene categories (Mdn = 0.04) compared to the general population of objects (Mdn = 0.006, Z = 12.4, p < 0.0001), suggesting that the consistency is driven by our contextual knowledge of diagnostic objects.

Additionally, the consistency of participants' responses in the "never present" condition could reflect the increased availability of more frequent objects in memory (Anderson & Schooler, 1991). I examined the overall frequency of objects listed as "never present" and compared these to overall object frequencies in (Greene, 2013). I found that the median frequency of objects listed as never present (Mdn = 38) was significantly larger than the overall median frequency in the database (Mdn = 5, Z = 12.5, p < 0.0001). Therefore, both overall object frequencies as well as mutual information appear to contribute to the inter-observer consistency of absent objects.

3.1.2. Consistency of object ordering

In this free-response paradigm, an additional type of inter-subject consistency could exist in the order of objects named by participants, even within a frequency bin. I examined the rank-order correlations between observers for the "always" present objects, and compared them to permutations of random orderings.[2] Remarkably, the mean Spearman's rho across observers was 0.43 (range: 0.21–0.72, see Fig. 2) while the maximum correlation that would be expected from chance was 0.08, suggesting that
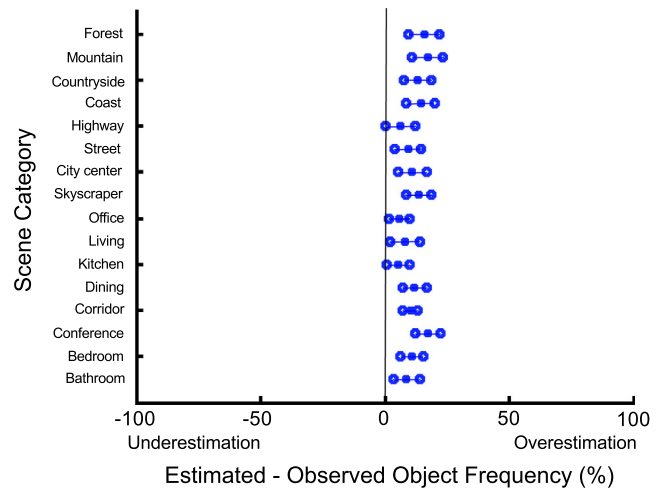

Fig. 2. Inter-subject rank-order correlations in the objects listed as "always" present in each of the sixteen basic-level scene categories. Each cell of each matrix represents the rank-order correlation between a pair of observers.

observers agree on how representative they are of the category, even without specific instructions concerning object ordering. Insofar as cognitive availability reflects ecological frequency (Anderson & Schooler, 1991; Tversky & Kahneman, 1973), these results could indicate that human observers have near-veridical estimates of object frequency. Next, I will compare estimated frequencies from this experiment with observed frequencies from the (Greene, 2013) scene database.

3.1.3. Overestimation of object frequency

To what extent are estimated object frequencies reflective of real-world frequencies? For each category, I correlated the estimated frequency with observed frequency. As shown in Fig. 3, observers overestimated object frequency in all scene categories by an average of 11% (range: 5–17%). I observed no significant difference in the overestimation for indoor and outdoor scene categories ($t(14)$ = 1.5, p = 0.15).

---

[2] For each observer, the objects listed in "always", "often", and "sometimes" were considered to be numbered 1–10. This ordering was compared to the ordering of each other observer in the same manner. If an observer listed an object not included by the original observer, it was given a rank of 11. Then, for each pair of observers, rank order correlations were computed for each category.
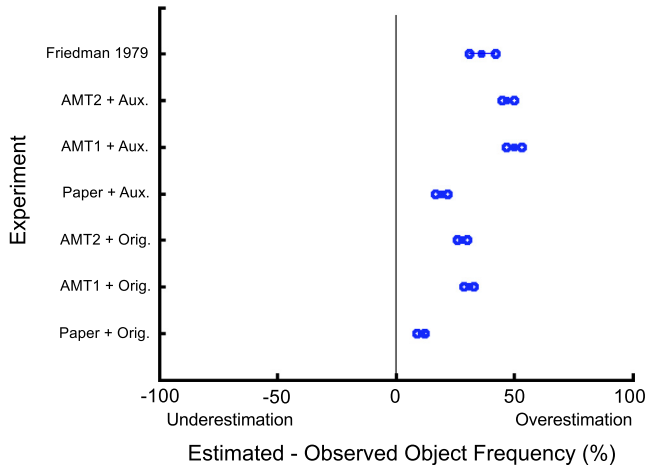
Fig. 3. Difference between estimated versus observed object frequency (in %) for each of the sixteen basic-level scene categories. Filled dots indicate the median while the open dots indicate the 95% confidence interval for each category.

To determine that this over-estimation was not solely due to bias in the dataset used by Greene (2013), I compared observers' frequency estimates to the independent auxiliary database provided by Greene (2013). I compared estimated frequency to observed frequency and found that on average, observers' estimates were overestimated by 19% (range: 3–33%).

3.2. Online object frequency rating experiments

The free response experiment aggregates frequency estimates across participants, but does not directly ask observers to rank frequencies. I conducted two online experiments asking participants to directly estimate object frequency using an analog slider. Although estimates were highly correlated with free response estimates ($r = 0.73$), participants overestimated object frequency by 31%. However, because I was unsure that participants understood that I was asking for frequency estimates in a single view of a scene, I ran the second experiment with the explicit instructions to consider frequencies in a single view, such as in a photograph. Despite the change in instructions, the two experiments produced nearly identical frequency estimates ($r = 0.92$), with an overestimation of object frequency of 28%. Comparing the overestimations to the observed frequencies in the auxiliary database, I found overestimations of 50% and 45% respectively. Therefore, participants overestimate object frequencies when asked directly about frequency, and do so even when explicitly told to consider single scene views.

3.3. Comparing to Friedman's 1979 normative rankings

Last, I used the object frequency estimates published in (Friedman, 1979) to the observed frequencies in my database. Considering the 93 scene-object pairs from her normative rankings that were observed in my database, observers in her experiment overestimated object frequencies by an average of 36%. Fig. 4 shows the extent of overestimation in all experiments. Over all experiments, observers overestimated object frequency by an average of 32%.

4. Discussion

Six experiments demonstrate that human observers have strikingly consistent estimations of object frequency. Despite this consistency, observers overestimate frequencies by nearly one third on average.
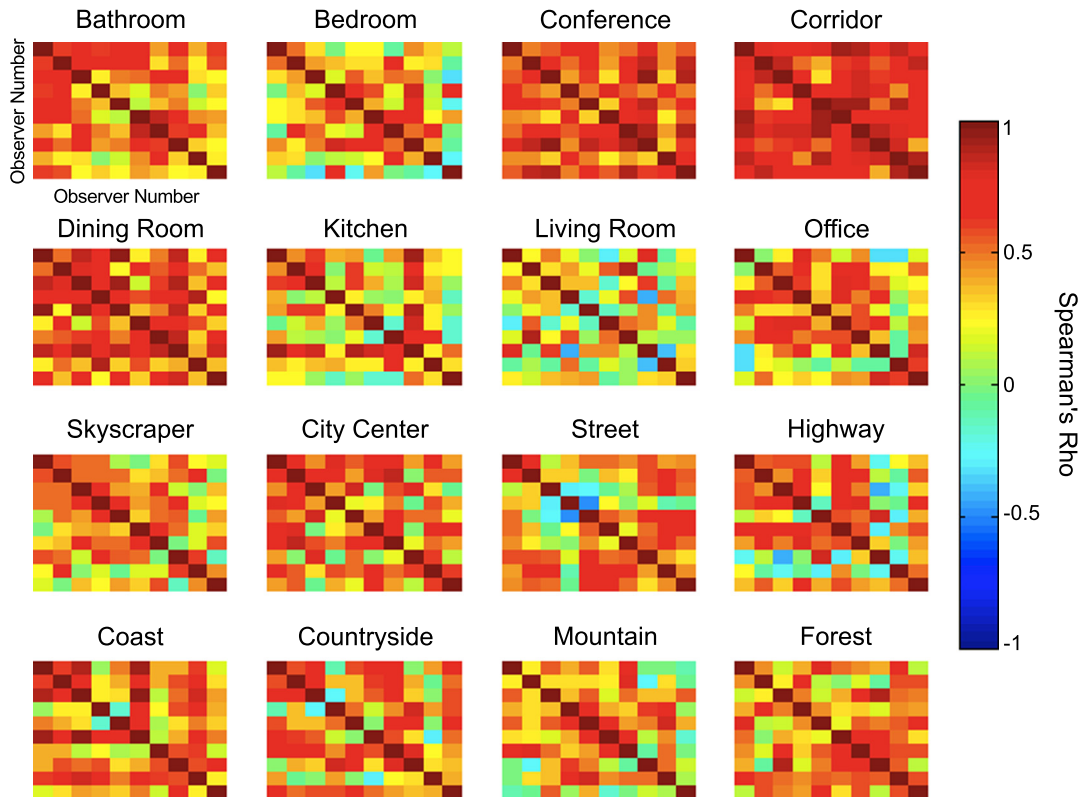
Fig. 4. Summary plot for all experiments. Orig refers to the main database in (Greene, 2013), while Aux refers to the auxiliary database in that same paper. Over each of these experiments, observers overestimate object frequency by an average of 32%. Filled dots indicate the median while the open dots indicate the 95% confidence interval for each experiment.

The overestimation is robust to task changes as it was observed in both free response and direct estimation. It was also found when observers were explicitly told to think about single scene views rather than the whole environment. These results cannot be attributed to demand characteristics, as they show up in estimates from previously published work (Friedman, 1979) that was not designed to test this question. Furthermore, the overestimation was not simply due to database bias (Torralba & Efros, 2011), as the same patterns of results were found in an independent dataset.

A large literature exists on how scene-object contextual relations can be used to aid scene and object understanding (Biederman, Mezzanotte, & Rabinowitz, 1982; Davenport & Potter, 2004; DeGraef, Christiaens, & d'Ydewalle, 1990; Hollingworth & Henderson, 1998; Võ & Henderson, 2009, for review see Bar, 2004). Yet all of these studies have defined contextual relations by the intuitions of the experimenter because we had not objectively measured contextual statistics from real world scenes. With the advent of large, labeled scene databases (Greene, 2013; Russell, Torralba, Murphy, & Freeman, 2008) we can do just that. The current work highlights the utility of mining these statistics by demonstrating that the statistical intuitions of observers are systematically skewed. Understanding how human estimates vary from objective measurements provides insights into the nature of scene representations.

Despite the consistency of the current results, they are not without limitation. Although overestimation cannot be attributed to the biases of a particular database, it could be the result of biases in photography as a whole. Photographers tend to place informative objects at the center of a photograph (Parkhurst & Niebur, 2003; Schumann et al., 2008; Tatler, 2007). It is possible that this bias systematically pushes other objects out of the frame. Aesthetics may also create bias: in Greene (2013), only 60% of *bathroom* scenes contained a "toilet", for example. Future work can examine the extent to which these biases arise from photographs by asking observers for object frequencies from real environments. To what extent are these biases implicit in object representations themselves, and to what extent do they arise while activating the representation? Future work will use implicit measures such as priming to disentangle these possibilities.

Context is only useful insofar as it provides lawful information about the world. These results challenge the view that context helps recognition because our contextual knowledge is biased. However, we experience a continuous three-dimensional world synthesized through discrete fixations (Hochberg, 1968; Potter, 1975). The striking inter-observer consistency in estimated frequencies suggests that our contextual knowledge follows our continuous percepts, rather than our discrete sensory inputs. This would allow us to use these schemata for predicting and finding objects that are not currently within view, allowing us to act efficiently in the world.

Acknowledgments

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00174.x.

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46*(2), 81–86. http://dx.doi.org/10.1037/h0057955.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5,* 617–625.

Biederman, I. (1972). Perceiving real-world scenes. *Science (New York, N.Y.), 177*(43), 77–80.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Biederman, I., Mezzanotte, R., & Rabinowitz, J. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes extracting categorical regularities without conscious intent. *Psychological Science, 19*(7), 678–685. http://dx.doi.org/10.1111/j.1467-9280.2008.02142.x.

Brewer, W., & Treyans, J. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), 207–230.

Brockmole, J. R., & Võ, M. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception, & Psychophysics, 72*(7), 1803–1813. http://dx.doi.org/10.3758/APP.72.7.1803.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention∗1, ∗2. *Cognitive Psychology, 36*(1), 28–71. http://dx.doi.org/10.1006/cogp.1998.0681.

Davenport, J., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559–564.

DeGraef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research Psychologische Forschung, 52,* 317–329.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108* (3), 316–355.

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Perception Science, 4,* 777. http://dx.doi.org/10.3389/fpsyg.2013.00777.

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

Hochberg, J. (1968). In the Mind's Eye. In *Contemporary theory and research in visual perception* (pp. 309–331).

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

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

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 15*(2), 179–187.

Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures.*

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think. *Psychological Science, 21*(11), 1551–1556. http://dx.doi.org/10.1177/0956797610385359.

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*(5), 1211–1222. http://dx.doi.org/10.1037/0278-7393.27.5.1211.

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, 99*(14), 9596–9601.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 551–578. http://dx.doi.org/10.1037/0278-7393.4.6.551.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41. http://dx.doi.org/10.1145/219717.219748.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*(2), 125–154.

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965–966.

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

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, 1–10. http://dx.doi.org/10.3758/s13414-013-0605-z.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1), 157–173. http://dx.doi.org/10.1007/s11263-007-0090-8.

Schumann, F., Einhauser, W., Vockeroth, J., Bartl, K., Schneider, E., & Konig, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision, 8*(14), 1–17. http://dx.doi.org/10.1167/8.14.12.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 1–17. http://dx.doi.org/10.1167/7.14.4.

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

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *2011 IEEE Conference on computer vision and pattern recognition (CVPR)* (pp. 1521–1528). http://doi.org/10.1109/CVPR.2011.5995347.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232. http://dx.doi.org/10.1016/0010-0285(73)90033-9.

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). http://dx.doi.org/10.1167/9.3.24.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception & Psychophysics, 73*(6), 1650–1671. http://dx.doi.org/10.3758/s13414-011-0153-3.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15*(2), 77–84. http://dx.doi.org/10.1016/j.tics.2010.12.001.