Current Biology

Report

Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes

Highlights

- Humans often miss giant targets in scenes during visual search
- Giant targets are more often missed even when humans foveate them
- Deep neural networks do not exhibit such a deficit with giant targets
- Missing giant targets is a functional brain strategy to discount distractors

Authors

Miguel P. Eckstein, Kathryn Koehler, Lauren E. Welbourne, Emre Akbas

Correspondence

miguel.eckstein@psych.ucsb.edu

In Brief

Eckstein et al. show that during visual search, humans, but not deep neural networks, often miss targets that have an atypical size relative to the surrounding objects in the scene. The authors suggest that this is not a human malfunction but a useful brain strategy to rapidly discount distractors during visual search.

Eckstein et al., 2017, Current Biology *27*, 2827–2832
September 25, 2017 © 2017 Elsevier Ltd.
http://dx.doi.org/10.1016/j.cub.2017.07.068

Current Biology
Report

Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes

Miguel P. Eckstein,[1,2,4,*] Kathryn Koehler,[1] Lauren E. Welbourne,[1,2] and Emre Akbas[1,3]
[1]Department of Psychological and Brain Sciences
[2]Institute for Collaborative Biotechnologies
University of California, Santa Barbara, Santa Barbara, CA, 93103, USA
[3]Department of Computer Engineering, Middle East Technical University, 06800 Ankara, Turkey
[4]Lead Contact
*Correspondence: miguel.eckstein@psych.ucsb.edu
http://dx.doi.org/10.1016/j.cub.2017.07.068

SUMMARY

Even with great advances in machine vision, animals are still unmatched in their ability to visually search complex scenes. Animals from bees [1, 2] to birds [3] to humans [4–12] learn about the statistical relations in visual environments to guide and aid their search for targets. Here, we investigate a novel manner in which humans utilize rapidly acquired information about scenes by guiding search toward likely target sizes. We show that humans often miss targets when their size is inconsistent with the rest of the scene, even when the targets were made larger and more salient and observers fixated the target. In contrast, we show that state-of-the-art deep neural networks do not exhibit such deficits in finding mis-scaled targets but, unlike humans, can be fooled by target-shaped distractors that are inconsistent with the expected target's size within the scene. Thus, it is not a human deficiency to miss targets when they are inconsistent in size with the scene; instead, it is a byproduct of a useful strategy that the brain has implemented to rapidly discount potential distractors.

RESULTS

Searching for a target in an unfamiliar environment can be difficult. But in their daily lives, humans visually search in familiar environments. The human brain rapidly processes a scene and utilizes relationships among objects and global properties of scenes to guide search toward likely target locations and facilitate target detection [4–6, 8, 10, 13–16]. For example, when searching for a toothbrush in a bathroom (Figure 1A), observers typically look toward the sink, where toothbrushes are often placed [8, 10, 13, 14, 17, 18]. When the toothbrush is placed at an unexpected location in the scene (Figure 1B), search becomes more difficult [7, 18, 19].

Here, we hypothesized that scene information also guides search toward likely target sizes. The human brain rapidly processes a scene and evaluates visual sensory evidence (spatial forms in the image) at spatial scales that are consistent with likely sizes of the target object relative to the rest of the scene.

We reasoned that if search is guided toward target sizes consistent with the scene, then if we scaled the target to be larger but inconsistent in size with the scene (Figure 1C), it would be missed more often during visual search. This prediction would not be expected based on classic visual search results with simple symbols (e.g., circles) on uniform backgrounds that predict lower error rates with larger targets among smaller distractors [20].

Sixty observers viewed a total of 42 rendered scenes (28.2 × 22.6 degrees visual angle), each with a unique target object that would be searched for by participants in the experimental task. There were a total of 14 target objects (toothbrush, computer mouse, parking meter, etc.), each repeated three times but never identically (i.e., color and viewing angle changed across the three instances). After looking at an initial fixation cross and being presented with a word of the target object, observers had 1 s to search through the scene (Figure 1E) and report whether the target was present or absent (half of the images contained the target). They were allowed to freely make eye movements.

In one-third of the target present scenes, the target was mis-scaled so that it was inconsistent in size with the surrounding objects and the rest of the scene. The mis-scaled targets were enlarged by a factor of 3× to 4× to ensure that the manipulation did not reduce the visibility of the low-level features defining the target. The location of the target was preserved as well as its local background to try to preserve the saliency (Figure 1C). To dissociate changes in search performance due to target and/or scene size inconsistency rather than physical change in target size, one-third of the target present scenes consisted of a control condition for which the entire image was rescaled and cropped so that the target size matched that of the target in the mis-scaled condition (Figure 1D). The remaining target present trials contained targets that were consistent in size with the surrounding scene (Figure 1A). Images from all conditions were randomly interleaved with target absent images (i.e., Figures 1A and 1D, but without the target, referred to as ''normal (target absent)'' and ''control (target absent),'' respectively). We measured observers' ability to detect the target in the scenes.

Figure 2A shows that the proportion of target present trials for which the observers correctly detected the target (hit rate) was significantly lower when the target was mis-scaled compared to when it was at normal scale within the scene ($t(59) = -3.94$, $p < 0.001$; bootstrap resampling, $p < 0.001$).

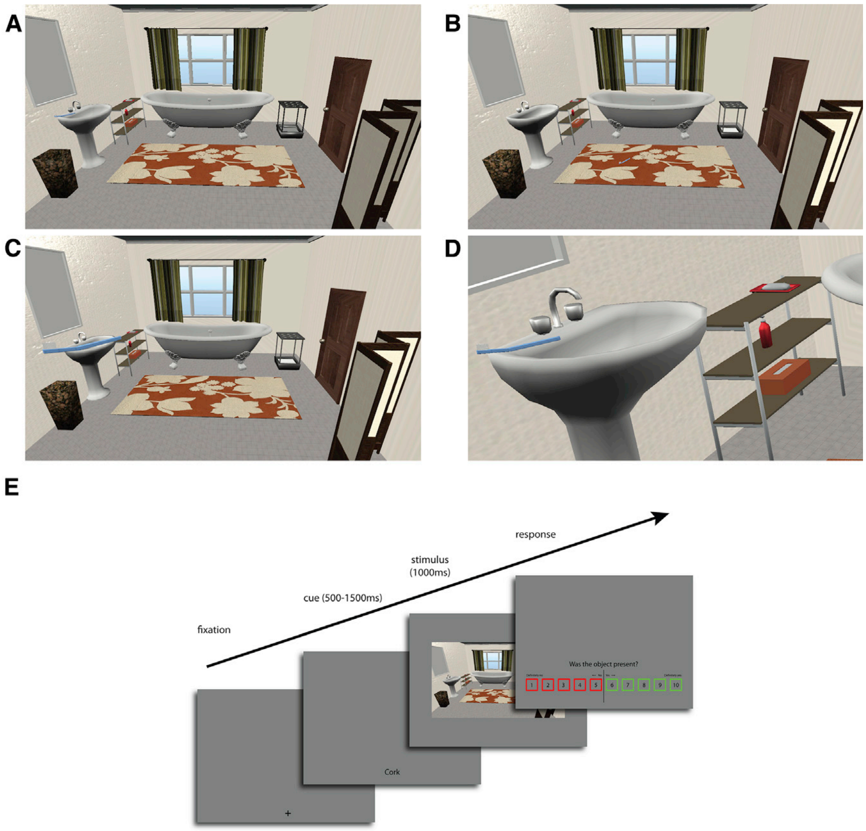

Figure 1. Visual Search for Objects in Scenes

(A) Sample image of a bathroom (28.2 × 22.6 degree visual angle) and toothbrush at a likely location (on the sink).

(B) Toothbrush at an unexpected location (on the floor rug).

(C) Mis-scaled toothbrush with a size inconsistent with the scene but placed at the expected location (sink).

(D) Entire image is rescaled so that the toothbrush is consistent in size with scene but has the same physical size as in the mis-scaled toothbrush image in (C).

(E) Timeline of a single experimental trial.

For both of these conditions (normal and mis-scaled target present), the target absent versions of the scenes were the same (normal (target absent)), and therefore the false alarm rate was the same for both conditions. The difference in hit rate between the normal and mis-scaled conditions translates to a 24% difference in the index of detectability (d′ = 2.09 for the normally scaled targets; d′ = 1.68 for the mis-scaled targets). The deficit for mis-scaled targets persisted across five blocks (containing seven trials each) but diminished in the last block (Figure 2B).

We evaluated whether higher error rates for mis-scaled targets was related to observers' failure to scrutinize the missed size-inconsistent targets with the high-resolution fovea through eye movements. For example, mis-scaled targets could potentially disrupt the guidance of eye movements toward likely target locations, forcing decisions to rely more on lower-resolution peripheral processing and increasing miss rates. We analyzed fixation patterns separately for each experimental condition.

The closest distance from the fovea to the target, in trials where the target was missed, was no different between the normal and mis-scaled target conditions (0.12° closer when mis-scaled, t(35) = 1.48, p = 0.15) (Figure 2C). Furthermore, analysis of only those trials in which the observers fixated the target (within 2 degrees, to account for eye tracking errors) still showed significantly larger miss rates for the mis-scaled targets (0.24) compared to the normal targets (0.12) (t(59) = 3.49, p < 0.001; bootstrap resampling, p < 0.001) (Figure 2D). The total difference in miss rate between normal and mis-scaled trials for foveated target trials (0.12) was similar to the overall difference in miss rate across all trials (0.13). Therefore, when observers are failing to detect the mis-scaled target, there is no indication that it is due to differences in the spatial guidance of eye movements.

The difference in target misses across size-consistent (normal) and inconsistent (mis-scaled) conditions cannot be attributed to feature-based changes from scaling the target object. A control condition that scaled the entire image to match the target object's physical size to that of the target size-inconsistent condition was detected near perfectly (see "control (target present)" condition, Figure 2A). We also evaluated the possibility that our object rendering software did not result in realistic objects that would be recognizable outside of their usual context. A separate group of 105 observers completed a task where they were shown an image of the mis-scaled target object in isolation on a gray background and were asked to name the object shown. Thirty of the 42 objects were correctly identified by more than 80% of the observers. We repeated our analysis while including only the thirty correctly identified objects to assess whether ambiguity over object identity could be driving the results, but the difference in miss rates between target size-consistent (miss rate = 0.16) and inconsistent (miss rate = 0.29) conditions remained the same (M = −0.13, t(59) = −3.95, p < 0.001; bootstrap resampling, p < 0.001). This result suggests that our effect is not due to ambiguity in the selected computer-rendered target objects.

We then asked whether failing to find mis-scaled objects is a property of the human brain or whether state-of-the-art deep neural networks (Faster R-CNN [23], R-FCN [21], and YOLO [22]) would demonstrate similar behavior. The neural networks take an image as an input and return bounding boxes outlining possible target locations with an associated probability. We analyzed a subset of 12 of the experimental images from our human study as well as 31 additional computer-rendered scenes containing targets. We assessed the detection probability of the networks for targets when these were normally scaled or mis-scaled relative to the scene. Our results (Figure 2E) show that, unlike in humans, the target probabilities for the neural networks did not decrease when the target was mis-scaled relative

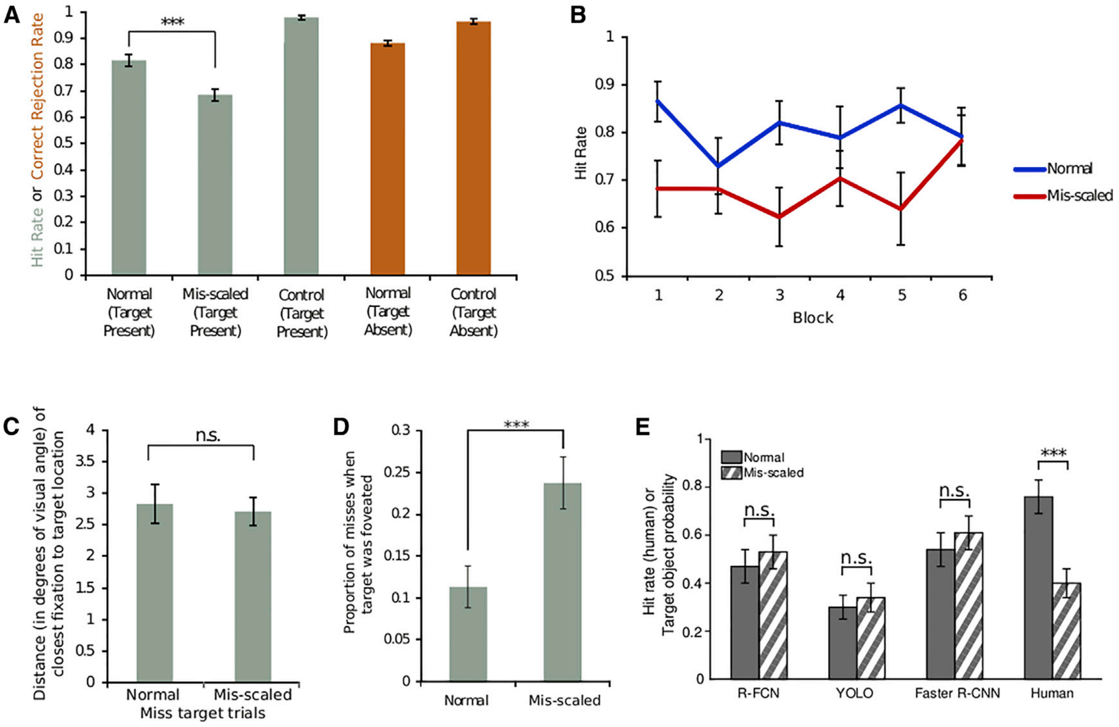

Figure 2. Target Detection by Humans and Deep Neural Networks

(A) Hit rates for images with target objects consistent (normal) or inconsistent (mis-scaled) in size with the scene, and the control condition. ***p < 0.001. The correct rejection rate for target absent images is also shown, for reference.

(B) Hit rate for normal and mis-scaled target conditions across blocks of seven trials.

(C) Distance in degrees of closest fixation to the target for missed target trials in normal and mis-scaled target conditions. n.s., not statistically significant.

(D) Proportion of trials in which the target was missed even if it was foveated for the normal and mis-scaled conditions. ***p < 0.001.

(E) Hit rate for humans detecting targets consistent (normal) and inconsistent (mis-scaled) in spatial scale with the scenes for a subset of 12 images, on which the deep neural networks were tested. Also plotted are the target probabilities associated with this subset of images (as well as 31 additional scenes) for normal and mis-scaled targets, shown for three different deep neural networks: R-FCN [21], YOLO [22], and Faster R-CNN [23].

All error bars are SEM.

to the scene. These results suggest that state-of-the-art deep neural networks do not fail to find targets that are inconsistent in size within a scene.

DISCUSSION

The guidance of eye movements and covert attention is fundamental to successful visual search [5, 7, 24–26]. Factors [7] that guide visual search include bottom-up saliency, visual attributes (features) that define the target [26–30], properties of scenes that predict the likely target location [4, 6, 8, 13, 14, 19, 31–35], and rewards [36–39]. Classic studies have demonstrated that a variety of violations in the relationship between a target and a scene can influence object detection for very briefly presented (150 ms) line drawings in the visual periphery [40, 41]. For longer presentations where observers engage in active visual search, placing targets at locations where they rarely occur naturally leads to increased misses or longer reaction times [34].

Here, we showed that observers utilize scene information to guide search toward likely target sizes, which results in increased miss rates for targets that are inconsistent in size with the scene. This phenomenon is not present in deep neural networks [23]. Do the errors in detecting mis-scaled targets reflect limitations of human visual processing, such as those shown by change blindness [42]? Or do they reflect a strategy that has some functional importance to successfully accomplish visual search? We suggest that the human brain utilizes information about the scene to give priority to objects that are at a likely spatial scale for the target, in order to rapidly discount other objects that resemble the target but are not at a consistent spatial scale. This strategy allows humans to reduce false positives when making fast decisions. To illustrate this concept, we identified images that a deep neural network (Faster R-CNN) incorrectly identified as containing a target; these targets would be easily dismissed by humans for not being at the appropriate spatial scale relative to the scene.

Figure 3A (top row) shows a series of such examples with false positives and their associated target probabilities for Faster R-CNN. Figure 3A also shows a sample of the five comparison images, where the object was correctly detected by Faster R-CNN with a target probability comparable to the images with the false positives (Figure 3A, bottom row). For example, the object detector gives a probability of 0.5 to a computer keyboard incorrectly identified as a cell phone. The location of the keyboard is in spatial proximity to a human hand, as would be expected of a cell phone, but its size is inconsistent with a cell

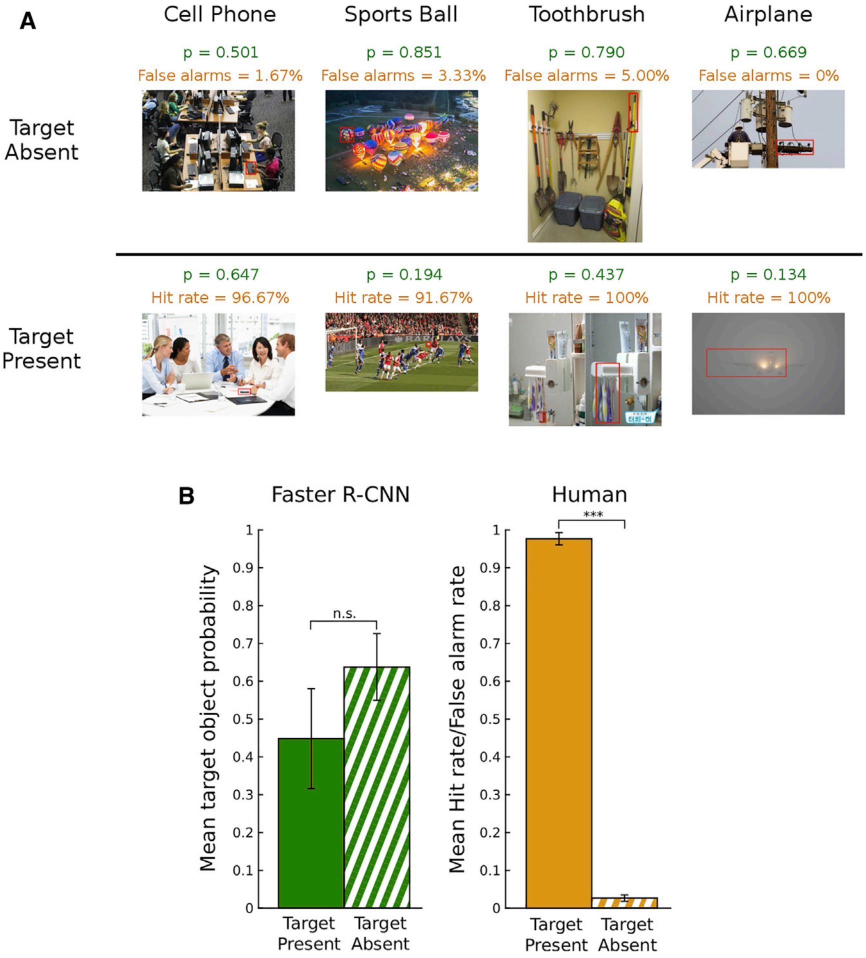

Figure 3. Deep Neural Network False Positives that Are Inconsistent in Size with the Scene

(A) Top row: sample images where the target is absent but other objects in the image produce high target probabilities for Faster R-CNN; these objects would be inconsistent in size (mis-scaled) with the scene if they were the target. Bottom row: comparison scenes where Faster R-CNN correctly identifies target objects, resulting in probabilities that are comparable or lower than the mis-scaled object false positives shown in the top row. The target object names are listed above the images and apply to both rows. Above each image, the target object probabilities are shown in green, and the false alarm or hit rates from humans are shown in orange. Image credits: Ronnie Macdonald (https://www.flickr.com/photos/ronmacphotos/), Brian Dryden, Aimee Wanner/TTN, dhgate.com, atlantacloset.com, Monkey Business Images (dreamstime.com), and Fotoamator (iStockPhoto.com).

(B) Hit rate and false alarm rate for humans (orange) and associated target probabilities for Faster R-CNN (green) for all of the real-world images that were tested (including the samples shown in A). Error bars are SEM. Humans can discount distractors that elicit large probabilities in Faster R-CNN because they are inconsistent with the expected size of the searched target relative to the scene.

phone, being five times larger than a hand. A human observer would likely easily dismiss the keyboard due to its inconsistency with typical sizes for cell phones. All of the real-world images that were used in Faster R-CNN were assessed with 60 additional subjects. This experiment confirmed that humans rarely misclassify such distracters as the target when viewing the scene for 1 s (Figure 3A shows individual hit rates/false alarm rates for the sample images; Figure 3B shows the means across all images for both humans [significantly lower false alarm rate than hit rate, bootstrap, $p < 0.001$] and Faster R-CNN). However, typical current deep neural networks do not evaluate the size of the potential cell phone (i.e., the keyboard in Figure 3A) to determine whether it has an unlikely large size for a cell phone relative to surrounding objects and thus discount it as a target. Our examples illustrated false positives for Faster R-CNN, but the same principle can be demonstrated for other deep neural networks.

An unwanted byproduct of the human brain's strategy of guiding attention toward probable target sizes is that when the targets are inconsistent in size with the scene, the observers will more often miss those targets. Yet, the incurred cost of the strategy is not high since such scenarios are not found often in the real world. And when these atypical circumstances do occur, our results show that, after repeated exposure, humans are eventually able to learn the unusual scenario. They then attend to the atypical target sizes and reduce the miss rate (see last block in Figure 2B). The ability to continuously learn changes in the statistical relationship among objects and scenes is another characteristic that distinguishes the human brain from the majority of current deep neural network frameworks, which rely on pre-training but typically do not continuously adapt through unsupervised learning from object detections in changing environments.

It is not known how the brain takes the scene into account to represent likely target sizes. In terms of computations, the human brain rapidly extracts information about the scene background [31, 43], other objects in the scene [13, 18, 35], and depth information [44, 45]. It is likely that the brain utilizes this information as a prior probability to favor sensory evidence for likely target sizes, possibly in a manner similar to that proposed for contextual locations within scenes and as a basic mechanism of spatial attention [4, 19, 46, 47]. This could be implemented in terms of increased baseline activity for neurons tuned to target sizes likely to appear in the scene [48]. Where might these neurons reside? Subpopulations of neurons in the occipitotemporal areas are probable candidates. The occipitotemporal areas of the human brain represent many important components of visual search including the category of the object being searched [49, 50], its expected location within the scene [51], and the physical size of objects in the real world [52].

Is this mechanism present in non-human animals? This is also unknown, but it is likely that, as with other statistical relationships in the visual environment, the visual systems of insects, mammals, and non-human primates also have an ability to use scene information to guide search toward the probable size of searched targets.

In conclusion, our results provide a functional explanation about why humans often miss searched objects if they are inconsistent in size within scenes. The findings also suggest a possible additional source of information from scenes, in order for deep neural networks to reduce false positives. Indeed, some studies have shown how incorporating information about likely target size can potentially reduce false positives for machine object detectors [53]. However, it is likely that humans would benefit from this additional information more than deep neural networks, because of the varying fidelity of human vision across the visual field. Objects away from the point of fixation are processed with lower resolution and are subject to crowding, making them confusable with the searched target. Thus, utilizing scene information to guide search toward likely target sizes might be an even more critical strategy for humans than machines to discount distractors and reduce false positives.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Stimuli and Design
  - Apparatus
  - Psychophysical Study Procedures
  - Deep Neural Networks
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

AUTHOR CONTRIBUTIONS

M.P.E., K.K., E.A., and L.E.W. contributed to the concept and design of the study. K.K. and L.E.W. were responsible for human data collection and analyses. E.A. was responsible for the deep neural network results. M.P.E., K.K., E.A., and L.E.W. contributed to manuscript preparation.

ACKNOWLEDGMENTS

This research was supported by the Naval Air Warfare Center Aircraft Division under Prime Grant N68335-16-C-0028, Mayachitra Incorporated, and the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the US Army Research Office. The content of the information does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

Received: May 8, 2017
Revised: July 21, 2017
Accepted: July 28, 2017
Published: September 7, 2017

REFERENCES

1. Srinivasan, M.V. (2010). Honey bees as a model for vision, perception, and cognition. Annu. Rev. Entomol. *55*, 267–284.
2. Eckstein, M.P., Mack, S.C., Liston, D.B., Bogush, L., Menzel, R., and Krauzlis, R.J. (2013). Rethinking human visual attention: spatial cueing effects and optimality of decisions by honeybees, monkeys and humans. Vision Res. *85*, 5–19.
3. Wasserman, E.A., Teng, Y., and Castro, L. (2014). Pigeons exhibit contextual cueing to both simple and complex backgrounds. Behav. Processes *104*, 44–52.
4. Torralba, A., Oliva, A., Castelhano, M.S., and Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. *113*, 766–786.
5. Eckstein, M.P. (2011). Visual search: a retrospective. J. Vis. *11*, 11.
6. Wolfe, J.M., Võ, M.L.-H., Evans, K.K., and Greene, M.R. (2011). Visual search in scenes involves selective and nonselective pathways. Trends Cogn. Sci. *15*, 77–84.
7. Wolfe, J.M., and Horowitz, T.S. (2017). Five factors that guide attention in visual search. Nat. Hum. Behav. *1*, 0058.
8. Neider, M.B., and Zelinsky, G.J. (2006). Scene context guides eye movements during visual search. Vision Res. *46*, 614–621.
9. Peterson, M.S., and Kramer, A.F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. Percept. Psychophys. *63*, 1239–1249.
10. Castelhano, M.S., and Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. Atten. Percept. Psychophys. *72*, 1283–1297.
11. Droll, J.A., Abbey, C.K., and Eckstein, M.P. (2009). Learning cue validity through performance feedback. J. Vis. *9*, 1–23.
12. Droll, J.A., Hayhoe, M.M., Triesch, J., and Sullivan, B.T. (2005). Task demands control acquisition and storage of visual information. J. Exp. Psychol. Hum. Percept. Perform. *31*, 1416–1438.
13. Mack, S.C., and Eckstein, M.P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. J. Vis. *11*, 1–16.
14. Malcolm, G.L., and Henderson, J.M. (2010). Combining top-down processes to guide eye movements during real-world scene search. J. Vis. *10*, 1–11.
15. Wolfe, J.M., Alvarez, G.A., Rosenholtz, R., Kuzmova, Y.I., and Sherman, A.M. (2011). Visual search for arbitrary objects in real scenes. Atten. Percept. Psychophys. *73*, 1650–1671.
16. Võ, M.L.-H., and Henderson, J.M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. J. Vis. *10*, 1–13.
17. Castelhano, M.S., and Henderson, J.M. (2007). Initial scene representations facilitate eye movement guidance in visual search. J. Exp. Psychol. Hum. Percept. Perform. *33*, 753–763.
18. Koehler, K., and Eckstein, M.P. (2017). Beyond scene gist: Objects guide search more than scene background. J. Exp. Psychol. Hum. Percept. Perform. *43*, 1177–1193.
19. Eckstein, M.P., Drescher, B.A., and Shimozaki, S.S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. Psychol. Sci. *17*, 973–980.
20. Hodsoll, J., and Humphreys, G.W. (2001). Driving attention with the top down: the relative contribution of target templates to the linear separability effect in the size dimension. Percept. Psychophys. *63*, 918–926.
21. Dai, J., He, K., and Sun, J. (2016). R-FCN: object detection via region-based fully convolutional networks. arXiv:1605.06409. https://arxiv.org/abs/1605.06409v2.
22. Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. arXiv:1612.08242. https://arxiv.org/abs/1612.08242.
23. Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. *39*, 1137–1149.

24. Najemnik, J., and Geisler, W.S. (2005). Optimal eye movement strategies in visual search. Nature *434*, 387–391.

25. Kunar, M.A., Flusberg, S., Horowitz, T.S., and Wolfe, J.M. (2007). Does contextual cuing guide the deployment of attention? J. Exp. Psychol. Hum. Percept. Perform. *33*, 816–828.

26. Wolfe, J.M., and Horowitz, T.S. (2004). What attributes guide the deployment of visual attention and how do they do it? Nat. Rev. Neurosci. *5*, 495–501.

27. Findlay, J.M. (1997). Saccade target selection during visual search. Vision Res. *37*, 617–631.

28. Eckstein, M.P., Beutter, B.R., and Stone, L.S. (2001). Quantifying the performance limits of human saccadic targeting during visual search. Perception *30*, 1389–1401.

29. Bravo, M.J., and Farid, H. (2009). The specificity of the search template. J. Vis. *9*, 1–9.

30. Malcolm, G.L., and Henderson, J.M. (2009). The effects of target template specificity on visual search in real-world scenes: evidence from eye movements. J. Vis. *9*, 8.1–13.

31. Greene, M.R., and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognit. Psychol. *58*, 137–176.

32. Larson, A.M., and Loschky, L.C. (2009). The contributions of central versus peripheral vision to scene gist recognition. J. Vis. *9*, 1–16.

33. Castelhano, M.S., and Heaven, C. (2011). Scene context influences without scene gist: eye movements guided by spatial associations in visual search. Psychon. Bull. Rev. *18*, 890–896.

34. Koehler, K., and Eckstein, M.P. (2017). Temporal and peripheral extraction of contextual cues from scenes during visual search. J. Vis. *17*, 16.

35. Pereira, E.J., and Castelhano, M.S. (2014). Peripheral guidance in scenes: The interaction of scene context and object content. J. Exp. Psychol. Hum. Percept. Perform. *40*, 2056–2072.

36. Navalpakkam, V., Koch, C., Rangel, A., and Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. Proc. Natl. Acad. Sci. USA *107*, 5232–5237.

37. Ackermann, J.F., and Landy, M.S. (2013). Choice of saccade endpoint under risk. J. Vis. *13*, 27.

38. Eckstein, M.P., Schoonveld, W., Zhang, S., Mack, S.C., and Akbas, E. (2015). Optimal and human eye movements to clustered low value cues to increase decision rewards during search. Vision Res. *113* (Pt B), 137–154.

39. Sullivan, B.T., Johnson, L., Rothkopf, C.A., Ballard, D., and Hayhoe, M. (2012). The role of uncertainty and reward on eye movements in a virtual driving task. J. Vis. *12*, 19.

40. Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. Cognit. Psychol. *14*, 143–177.

41. Palmer, T.E. (1975). The effects of contextual scenes on the identification of objects. Mem. Cognit. *3*, 519–526.

42. Rensink, R.A., O'Regan, J.K., and Clark, J.J. (1997). To see or not to see: the need for attention to perceive changes in scenes. Psychol. Sci. *8*, 368–373.

43. Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. *42*, 145–175.

44. Sherman, A.M., Greene, M.R., and Wolfe, J.M. (2011). Depth and size information reduce effective set size for visual search in real-world scenes. J. Vis. *11*, 1334.

45. Wolfe, J.B. (2012). Visual search. In Cognitive Search: Evolution, Algorithms, and the Brain, P.M. Todd, T.T. Hills, and T.W. Robbins, eds. (MIT Press), pp. 159–176.

46. Eckstein, M.P. (2017). Probabilistic computations for attention, eye movements, and search. Annu Rev Vis Sci *3*.

47. Kanan, C., Tong, M.H., Zhang, L., and Cottrell, G.W. (2009). SUN: Top-down saliency using natural statistics. Vis. Cogn. *17*, 979–1003.

48. Eckstein, M.P., Peterson, M.F., Pham, B.T., and Droll, J.A. (2009). Statistical decision theory to relate neurons to behavior in the study of covert visual attention. Vision Res. *49*, 1097–1128.

49. Peelen, M.V., and Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. Proc. Natl. Acad. Sci. USA *108*, 12125–12130.

50. Peelen, M.V., and Kastner, S. (2014). Attention in the real world: toward understanding its neural basis. Trends Cogn. Sci. *18*, 242–250.

51. Preston, T.J., Guo, F., Das, K., Giesbrecht, B., and Eckstein, M.P. (2013). Neural representations of contextual guidance in visual search of real-world scenes. J. Neurosci. *33*, 7846–7855.

52. Konkle, T., and Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. Neuron *74*, 1114–1124.

53. Choi, M.J., Torralba, A., and Willsky, A.S. (2012). A tree-based context model for object recognition. IEEE Trans. Pattern Anal. Mach. Intell. *34*, 240–252.

54. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2016). Speed/accuracy trade-offs for modern convolutional object detectors. arXiv:1611.10012. https://arxiv.org/abs/1611.10012.

55. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. arXiv:1405.0312v1. https://arxiv.org/pdf/1405.0312v1.pdf.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Unity 3D	Unity Technologies, Bellevue, WA, USA	https://unity3d.com
Eyelink 1000 Eyetracker	SR Research, Mississauga, ON, Canada	RRID: SCR_009602
Qualtrics	Qualtrics, Provo, UT, USA	https://www.qualtrics.com/
PsiTurk	PsiTurk.org	https://psiturk.org/
Python	http://www.python.org	RRID: SCR_008394
MATLAB R2016b	The MathWorks	RRID: SCR_001622
Faster R-CNN	[23]	https://github.com/rbgirshick/py-faster-rcnn
R-FCN	[21]	https://github.com/YuwenXiong/py-R-FCN
YOLO	[22]	https://github.com/pjreddie/darknet
Other		
Amazon Mechanical Turk	Amazon.com	https://www.mturk.com/
Deposited Data		
Raw Data	this paper	http://dx.doi.org/10.17632/rwy2nb24df.1

CONTACT FOR RESOURCE SHARING

Information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Miguel P. Eckstein (miguel.eckstein@psych.ucsb.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Eye-tracking and target detection response data were collected from 60 undergraduate students at the University of California, Santa Barbara who received course credit in exchange for participation. All participants were verified to have normal or corrected-to-normal vision. A second group of 105 observers performed an object labeling task that served as a control experiment, and a third group of 60 subjects performed an object detection task for the real-world scenes (Figure 3A). Both of the latter groups were recruited via the Amazon Mechanical Turk website, and received payment in exchange for participation, equivalent to a rate of approximately $10 per hour. The gender balance for the experimental studies at UC Santa Barbara followed an approximate breakdown of 60% females and 40% males and ages 18-22 years. The gender and age of our individual participants was not recorded for the experiments, as we had no hypotheses that related to this information. All participants provided informed written consent and were recruited and treated according to approved human subject research protocols by the University of California, Santa Barbara.

METHOD DETAILS

Stimuli and Design

The computer generated scenes were created in Unity 3D (Unity Technologies, Bellevue, WA, USA), each with a unique target object that would be searched for by participants in the experimental task. From each scene, five images were created: (1) normal scene with target scaled proportionally to surroundings, (2) scene with target scaled 3x and 4x larger than the normal scene, (3) zoomed-in image of scene where target is identical in size to (2), but all other objects are proportionally larger as well, (4) target absent version of (1), and (5) target absent version of (3). A total of 42 scenes were created in this manner, along with an additional 31 scenes that were not shown to the human observers, but used to create a larger set of images containing targets for which the deep neural networks had pre-trained detection models.

Images were divided into the five conditions using a Latin square design, resulting in 7 images per condition, with the exception of the normal target absent scenes (see (4), above), of which there were two groups of seven (fourteen total). This ensured there was an equal number of target present and target absent scenes. Participants were randomly assigned to view a particular stimulus set shown in randomized order.

The ten images of real-world scenes used in the target detection task and deep neural networks were acquired from different sources on the internet (see Figure 3 caption for credits of the images shown in Figure 3A, where credits are applicable). Five images contained one of five target objects (airplane, airplane, cell phone, sports ball, toothbrush), and five images did not contain any of these objects. These sets of images were paired, so that for each of the five target objects, there was a target present image and a target absent image.

Apparatus

Stimuli from the search task were displayed on a 1280 × 1024 pixel resolution Barco MDRC-1119 monitor. Each pixel subtended 0.022 degrees of visual angle. The display subtended an angle of 28.2 × 22.6 degrees visual angle. Eye tracking data were recorded on an Eyelink 1000 (SR Research, Mississauga, ON, Canada) monitoring gaze position at 250 Hz and was calibrated and validated using a nine-point grid system. A velocity greater than 22°/s and acceleration greater than $4000^\circ/s^2$ classified an event as a saccade.

The stimuli used in the object labeling tasks were presented using PsiTurk (https://psiturk.org/), and the stimuli for the real-world target detection task were presented in the form a questionnaire, built using Qualtrics software (Qualtrics, Provo, UT, USA). Both were distributed through Amazon Mechanical Turk (https://www.mturk.com/). All subjects therefore viewed the images on their own computers, where monitor specifications, internet speed (for presentation timings) and viewing distance could not be fully controlled.

Psychophysical Study Procedures

In the human behavioral search task, participants were instructed that they would be viewing a series of images and determining whether or not a particular object was present within them. They were told there was a 50% likelihood that the target would be present, but were not given any indication that some target objects may be abnormally sized. Each trial began with the presentation of a word, stating the target object to be searched for, followed by a fixation cross in the bottom-center of where the image would appear. Participants fixated the cross and pressed a button to initiate the trial. The image appeared after the participant maintained fixation on the cross for a random interval between 0.5-1.5 s. The participants had 1000 ms to search for the target object while their eyes were tracked before a response screen appeared where they indicated on a ten-point scale whether the target was present and how confident they were in their response.

Given that the objects in the scenes were 3D renderings of real objects, a separate group of participants completed a control task to ensure that they were able to properly identify the simulated target object as intended in the complete absence of contextual information. A total of 105 observers completed a task where they were shown an image of the target objects (physical size equivalent to that in the mis-scaled condition) in isolation on a gray background and were asked to freely name the object shown. The observers were split into three separate groups (76 observers in group 1, 17 in group 2, and 12 in group 3) so that each group would only see a single instance of each target object (recall that three versions of each target were used for a total of 42 scenes), thus each group viewed a total of 14 objects. We evaluated the proportion of observers correctly naming the object.

A final human experiment was performed using 10 real-world scenes. Five of the images were selected to illustrate examples of distractors, which are objects that are inconsistent in size compared to that of the searched target (relative to the scene). The Figure 3 caption lists the credits for these images, where applicable. Subjects were presented with the name of the target object for a maximum of 2000 ms before being shown an image for 1000 ms, which contained a red box surrounding the object in the scene that was identified as the target by Faster R-CNN. For the target present images, the red box surrounded the target object, and in the target absent images the red box surrounded the distractor object in the scene (which was identified as the target, incorrectly, by Faster R-CNN). Subjects were then asked whether the red box in the image contained the target object. The images were presented in a randomized order for each subject, with all subjects viewing all images.

Deep Neural Networks

Currently, there are three meta-architectures for state-of-the-art convolutional neural network based object detection [54]: 1) Faster R-CNN, 2) R-FCN and 3) single shot detection (SSD). Both Faster R-CNN and R-FCN are two stage networks where generic object proposals are generated in the first stage and they are recognized in the next. The models initially propose candidate image regions to contain any object type, and then compute a probability for each instance of each object category, within each region. Typically, the category with the highest probability is nominated as the object within the region and non-maxima suppression is applied so that object proposal regions with significant overlap do not produce redundant object labels. The major difference between Faster R-CNN and R-FCN lies in how they do the region-of-interest pooling of the features in the last layer of the network. R-FCN is usually faster than Faster R-CNN. On the other hand, YOLO is a single shot detection (SSD) network and directly predicts both boxes and their class probabilities without requiring a separate recognition stage.

We ran Python implementations of Faster R-CNN (https://github.com/rbgirshick/py-faster-rcnn) and R-FCN (https://github.com/YuwenXiong/py-R-FCN). We ran a C implementation of YOLO (https://github.com/pjreddie/darknet). The implementations are pre-trained on the 80 categories of objects in images from the Microsoft Common Object in Context (MSCOCO [55]) database. MSCOCO is an image database chosen to contain cluttered scenes with detailed backgrounds, as opposed to more typical databases that contain images of a single object against a mostly uniform background. Three of the MSCOCO object categories were target objects in the visual search stimuli: toothbrush, parking meter, and computer mouse. To directly compare human and model behavior, we assessed any object proposal region that overlapped with the target objects in the images used for the visual search task. We then selected the region with the highest probability associated with the target category (in all cases, the region contained at least half of the target object) and compared the average model detection probability with observer detection hit rates.

Additionally, we ran Faster R-CNN on a small set of 10 photographic images containing real-world scenes and assessed the detections of object categories (Figure 3). We compared the object probabilities to the hit rate and false alarm rates of human subjects (see psychophysics studies section, and Figure 3).

QUANTIFICATION AND STATISTICAL ANALYSIS

Parametric independent sample t tests using MATLAB (version R2016b) were used to assess statistical significance of hit and miss rates, target probabilities of the models, as well as the distance of the fovea to the target. Error bars on graphs correspond to standard error of the mean. The significance levels indicated on the figure are based on the results of these t tests. For all statistical tests, a significance criteria of $p < 0.05$ was used.

Due to possible violation of the normal distribution assumption for proportions we also tested statistical significance utilizing bootstrap re-sampling methods which do not make assumptions about the underlying distribution of the population nor the distribution of sample means. To assess differences in hit rates (HR)/false alarm (FA) rates between the normal and mis-scaled trials, we re-sampled 10,000 sets of trials for each condition. We calculated the difference in hit rate/false alarm rate across each of the 10,000 bootstrapped samples (Normal HR - Mis-scaled FA) and assessed the distribution of the differences. The reported p values correspond to the proportion of re-sampled differences that are less than/greater than a difference of zero.

Each reported experimental result in the Results section indicates the statistical test, degrees of freedom when appropriate and the associated p value. Our sample sizes varied across experiments. For the search experiment our sample size ($n = 60$ subjects) was based on the goal of having enough statistical power for hypothesis testing, estimated effect sizes for previous scene context studies we recently conducted [18], and consideration that eye tracking studies are more time consuming and limited by the availability of the equipment. The larger sample size ($n = 105$) for the object identification experiment was determined by taking into account that the objective was not hypothesis testing but obtaining reliable estimates of the mis-identifications for each individual image, the study was also less time consuming and easier to collect. The final experiment with the 10 real-world scenes included 60 subjects, and this sample size was also based on demonstrating a statistically significant difference in hit rate and false alarm rate across the two sets of images. The Results section indicates samples sizes for each study.

DATA AND SOFTWARE AVAILABILITY

All of the raw data from this article are accessible via Mendeley Data (http://dx.doi.org/10.17632/rwy2nb24df.1). The main analysis scripts are available by request to the Lead Contact. The various software implementations for the deep neural networks are publicly available and are listed in the Key Resource Table.