Evaluating view-invariant place recognition in humans and machines

Nathan C. L. Kong University of Pennsylvania
Tyler Bonnen UC Berkeley
Russell A. Epstein University of Pennsylvania

Abstract

We are able to perceive spatial structure in the world around us. This ability supports a range of downstream behaviours—from navigation to memory retrieval—and is thought to rely on a network of 'scene-selective' cortical structures. Feedforward deep learning models are often thought to provide a suitable approximation of these perceptual abilities. This human-model correspondence, however, has largely been evaluated in classification tasks. Here we develop a novel behavioural assay which reveals a profound gap between these vision models and human abilities. We collect a corpus of naturalistic scenes (panorama captures from Google maps) and format these environments into 'oddity' tasks: participants are presented with two different viewpoints from one location (A), alongside an image from a different location (B), and must identify the odd-one-out (B). Critically, we manipulate the angular difference between views and perceptual similarity between environments. Through a series of experiments, we find that humans substantially outperform models on this benchmark, and that human reaction times scale linearly with task difficulty. These data highlight the temporal dynamics of place recognition, challenging common assumptions about the feedforward underpinnings of this foundational human ability.

Keywords: vision; scene perception; place recognition; view-invariance; deep neural networks; temporal dynamics

Introduction

Our ability to perceive spatial structure supports a constellation of downstream behaviours—from navigation to episodic memory, foraging to social behaviours (Epstein & Baker, 2019). Representing this spatial structure is thought to be supported by scene-selective brain networks (Epstein & Kanwisher, 1998; Groen, Silson, & Baker, 2017). Remarkably, these neural structures have been effectively modelled by feedforward, deep neural networks (Groen et al., 2018; Conwell, Prince, Kay, Alvarez, & Konkle, 2024). However, this human-model alignment is typically evaluated in the context of classification tasks, which is not always predictive of other perceptual abilities (Jagadeesh & Gardner, 2022; Bonnen, Yamins, & Wagner, 2021). Recognizing this, here we develop a more fine-grained behavioural assay of place recognition and show a *misalignment* between humans and vision models on this task.

We operationalize "view-invariant place recognition" as the ability to recognize that different views from the same 360° panorama correspond to the same place. We adopt the oddity task experimental design that has been used to evaluate the role of the medial temporal lobe in visual object perception, to evaluate place recognition behaviour (Barense, Henson, Lee, & Graham, 2010; Lee et al., 2005). In our setup, a single oddity trial consists of three images, each taken from a panorama. Two of those images are different views from the *same* panorama, and the third image is a view from a *different* panorama (i.e., the odd-one-out). We generate a large set of oddity trials and compare the accuracy of humans (obtained via online experiments) to that of models, allowing us to ascertain the conditions under which purely visual representations are sufficient or insufficient to support view-invariant place recognition.

Results

Place recognition oddity task experimental setup We first acquired a set of 1260 panoramas of New York City from Google maps. Panoramas were taken at different locations; thus each panorama corresponds to a unique place. An example of a flattened panorama is shown in Figure 1A. We constructed oddity trials from pairs of panoramas (A and B) using three images, a, a', and b, where a and a' were sampled from panorama A and b was sampled from panorama B (Figure 1B). All images had a field-of-view of 90°. We varied the difficulty of trials by controlling the angular difference between a and a', denoted by θ in Figure 1C, with $\theta \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ\}$. Note that when the angular difference between a and a' is 90° or greater, there is no visual overlap between a and a', so that a strategy that depends solely on matching visual features may not be sufficient. For each angular difference, we randomly generated 800 problems (i.e., panorama pairs) and evaluated a computational proxy of visual processing (i.e., a convolutional neural network or vision transformer) on each problem.

Performance of a vision model decreases as the angular difference between views increases To compute computational model choice on each oddity trial, we extracted features from a penultimate layer of the model for each of the three images and computed the pairwise similarities between the three feature vectors using the Pearson's correlation metric. The model's odd-one-out choice was taken to be the image with the lowest similarity with the other two images. By computing model choices for a large number of oddity problems, we obtained a distribution of model performance across problems that use different panorama pairs. We found that performance of a task-optimized convolutional neural network (ResNet-18 trained on image categorization) decreased as the angular difference between a and a' increased, and the variability of performance across problems became larger (Figure 1D).

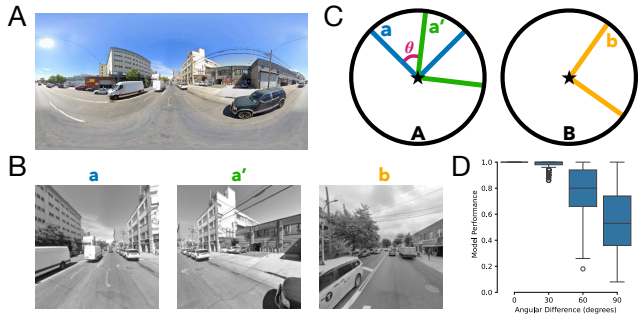

Figure 1: Schematic of a single oddity problem and model performance across problems. **A.** Example of a flattened panorama in the stimulus set. **B.** Example of an oddity trial ($60°$ angular difference between images a and a'). The correct choice is the image on the right (image b). a and a' were taken from the panorama shown in panel A. **C.** Top-down view of two panoramas (A and B) and sampled views (a, a', and b). θ denotes the angular difference between a and a', and ★ denotes the location of the observer. **D.** Model (ResNet-18 trained on image categorization) performance across 3200 problems of differing angular difference between a and a' (800 problems per angle).

Humans outperform a vision model on view-invariant place recognition We compared model performance to the performance of human participants who were administered the oddity task online ($N = 90$). On each trial, participants viewed the three images side-by-side (as in Figure 1B) for a maximum of 10 seconds and chose whether the left, the middle, or the right image was the odd-one-out (i.e., belonged to a different panorama). The three images were selected based on model performance (Figure 1D). Specifically, for each angular difference, we uniformly sampled oddity problems (i.e., panorama pairs) across the range of model performance. We predicted that problems that were easily solved by the models (i.e., model performance is close to 100%) should be easily solved by humans. In contrast, we hypothesized that problems where models struggled to solve would have a larger range of human performance, with some being easy for people, and others being hard. For each angular difference, we sampled 50 problems that spanned the performance range of ResNet-18 as uniformly as possible (Figure 2A).

We observed that humans also exhibited a range of performance values across these problems. As the problems become more difficult for people, exemplified by problems with higher angular difference between a and a' and by reduced accuracy, they spent more time on the problem (Figure 2B; $\beta = -2711.75$, $F(1, 145) = -16.93$, $p = 4 \times 10^{-36}$). Thus, the variability across problems was not just a speed-accuracy tradeoff but instead reflected a true difference in problem difficulty.

We then compared the performance (i.e., accuracy) of human participants to the performance of ResNet-18 trained on image categorization (i.e., a computational proxy of visual processing). We found that human and model performance were positively correlated ($r(145) = 0.65$, $p = 4.53 \times 10^{-19}$). This shows that problems that were more difficult for models also tended to be more difficult for humans. However, overall performance was better for humans than for ResNet-18 when the angular difference between the two views from the same panorama was $60°$ or $90°$ (paired t-tests; $60°$ angular difference: $t(48) = 6.44$, $p = 5.28 \times 10^{-8}$; $90°$ angular difference: $t(48) = 5.15$, $p = 4.84 \times 10^{-6}$). Inspection of the plot revealed that this was because human participants outperformed the model on the most difficult oddity problems (Figure 2C). We observed a qualitatively similar trend when DINOv2 was evaluated against human performance (paired t-tests; $60°$ angular difference: $t(48) = 5.86$, $p = 4.07 \times 10^{-7}$; $90°$ angular difference: $t(48) = 2.71$, $p = 0.009$; Figure 2D).

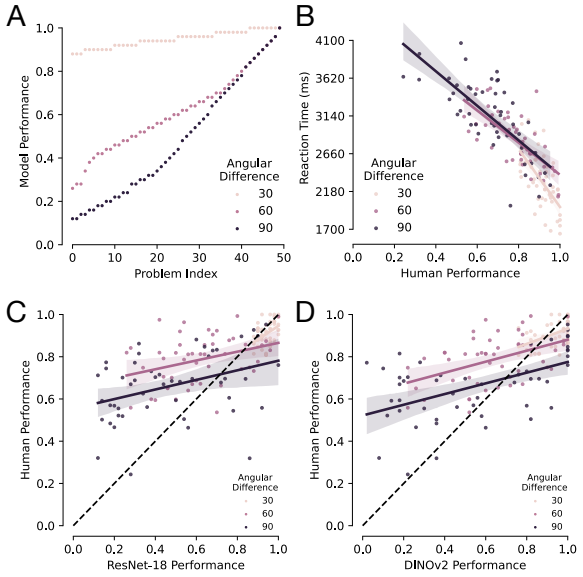

Figure 2: Comparing human and model performance on view-invariant place recognition. **A.** Oddity problems were selected to uniformly sample the range of model performance. **B.** Human reaction time is predicted by performance on the oddity problem. **C.** Humans outperform a vision model on difficult problems. That is, humans do not show as large a drop in performance as compared to models on increasingly difficult problems. Here we show performance of ResNet-18 trained on image categorization. **D.** Humans outperform DINOv2, a much larger deep learning model.

Conclusion

Our work provides a novel benchmark to evaluate place recognition in humans and machines. We find that humans substantively outperform vision models, and that there are clear temporal dynamics in human performance. Feedforward models do not capture critical aspects of place recognition—findings that are analogous to work on the temporal dynamics of human 3D shape inferences (Bonnen, Wagner, & Yamins, 2023). These data motivate us to develop novel computational frameworks that move beyond feedforward visual processing.

References

Barense, M. D., Henson, R. N., Lee, A. C., & Graham, K. S. (2010). Medial temporal lobe activity during complex discrimination of faces, objects, and scenes: Effects of viewpoint. *Hippocampus*, *20*(3), 389–401.

Bonnen, T., Wagner, A. D., & Yamins, D. L. (2023). Perirhinal cortex supports object perception by integrating over visuospatial sequences. *bioRxiv*, 2023–09.

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755–2766.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383.

Epstein, R., & Baker, C. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, *5*(1), 373–397.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, *7*, e32962.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102.

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, *119*(17), e2115302119.

Lee, A. C., Buckley, M. J., Pegman, S. J., Spiers, H., Scahill, V. L., Gaffan, D., ... others (2005). Specialization in the medial temporal lobe for processing of objects and scenes. *Hippocampus*, *15*(6), 782–797.