Humans and CNNs see differently: Action affordances are represented in scene-selective visual cortex but not CNNs

Clemens G. Bartnik (c.g.bartnik@uva.nl)
Institute for Informatics, University of Amsterdam
Science Park 900, 1098 XH, Amsterdam, The Netherlands

Iris I.A. Groen (i.i.a.groen@uva.nl)
Institute for Informatics, University of Amsterdam
Science Park 900, 1098 XH, Amsterdam, The Netherlands

Abstract

To navigate the immediate visual environment, humans use a variety of locomotive actions, such as walking, swimming or climbing. How does the brain represent such environmental action affordances and which visual features drive these representations? Here, we compared representations of visual properties derived from human annotations, fMRI measurements, and convolutional neural networks (CNNs) on a new set of real-world scenes that afford distinct locomotive actions in a diverse set of indoor and outdoor environments. Representational similarity analysis shows that scene-selective brain regions represent information about action affordances as well as materials and objects. In contrast, CNNs trained on scene classification show comparatively lower correlation with action affordances, instead most strongly representing global scene properties. Together, these results suggest that specialized models that incorporate action affordances may be needed to fully capture representations in scene-selective visual cortex.

Keywords: scene perception, fMRI, action affordances, navigation, convolutional neural networks

Introduction

How we represent visual scenes is a question central to neuroscience, cognitive psychology and AI. While scenes are often conceptualized as collections of objects or surfaces, recent work suggests that scene understanding is strongly shaped by perceived action possibilities **(Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016)**. This idea goes back to **Gibson (1977)** who introduced the term *affordances* to describe action possibilities an environment offers an individual. Where and how might such affordances be computed in the brain?

Scene perception in humans is characterized by the activation of three scene-selective regions **(Epstein & Baker, 2019)**. **Bonner and Epstein (2017)** implicated one of these regions, OPA, in the representation of navigable space affording walking in indoor environments, and linked these representations to visual features computed in mid-to-high level layers of place-trained CNNs **(Bonner & Epstein, 2017)**. However, other work reported a dissociation between representations in the brain, CNNs and human behavior, finding strong representation of a very broad set of affordances in behavior, but not in CNNs or scene-selective cortex **(Groen et al., 2018)**. This discrepancy between studies may be due to a number of reasons, such as the range of actions and environments sampled and how affordances were operationalized.

Here, we attempt to close this gap by comparing behavioral annotations of visual scene properties, place-trained CNNs, and fMRI measurements on a novel set of natural scenes chosen to span six distinct locomotive action affordances. Our results provide evidence of action affordance representations in scene-selective cortex. However, these representations are not well captured by place-trained CNNs, which align well with behavioral annotations of objects, materials, global properties and scene category, but not action affordances.

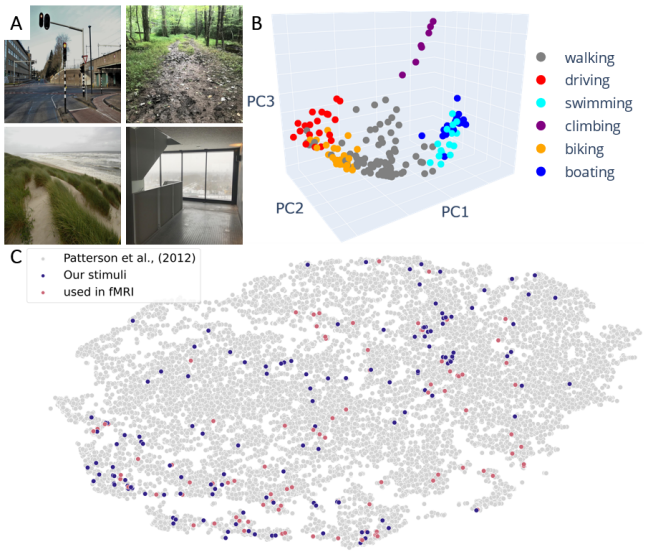

Figure 1: Stimuli overview (A) Example scenes. (B) PCA of action annotations. Stimuli separate along 3 dimensions. Walking is central, with one dimension for swimming and boating, one for biking and driving and one for climbing. (C) t-SNE visualization of the entire stimulus set and the fMRI subset covering the attribute space of the SUN Attribute Database.

Materials and Methods

Stimuli

We created a new set of 231 high-resolution (1024×1024 pixels) color photographs of daily scenes collected from Flickr. These images were carefully chosen to depict typical everyday scenes without prominent objects, humans, or animals, captured from human-scale, eye-level viewpoints (see **Fig. 1A** for examples). To ensure a balanced representation, we curated the set to include an equal number of indoor, outdoor-natural, and outdoor-manmade environments. **Fig. 1B** depicts a t-SNE visualization of our stimulus set covering the attribute space of over 12.000 images of the SUN Attribute DB **(Patterson & Hays, 2012)**. Behavioral annotations were collected for the full set of stimuli. For the fMRI experiment, we selected a subset of 90 images based on how well they captured three dimensions of scene navigability identified through PCA (**Fig. 1C**) on action affordance annotations.

Behavioral annotations

We collected behavioral annotations for five distinct scene properties (possible actions, objects, materials, global scene properties, scene category) online using Gorilla **(Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020)** and Prolific (N=152) **(Palan & Schitter, 2018)**. Each stimulus and property was labeled by on average 21.7 (SD = 0.87) participants. Representational dissimilarity matrices (RDMs) were computed using pairwise Pearson's correlation distances between the proportion of participants that annotated the presence or possibility of a given label (e.g., proportion of participants that annotated the scene as walkable).

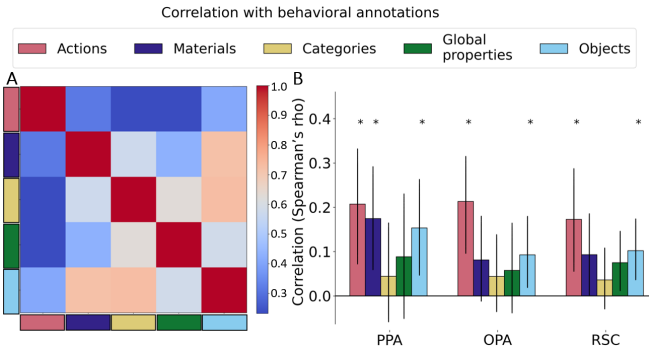

Figure 2: Correlation between behavioral RDMs for each annotated scene property (A) with each other and (B) with RDMs for each functional ROI (PPA, OPA and RSC). Error bars show 95% CI. * $p < 0.05$, two-tailed, Bonferroni corrected.

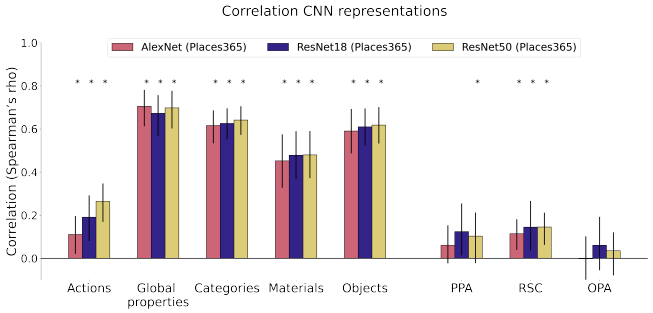

Figure 3: Correlation between CNN layer activations, average behavioral RDMs, and functional ROIs. Only the correlation for the single highest correlating layer is reported. Error bars show 95% CI. * $p < 0.05$, two-tailed, Bonferroni corrected.

fMRI experiment

12 healthy participants completed four fMRI sessions. First, scene-selective Parahippocampal Place Area (PPA), Occipital Place Area (OPA) and Retrosplenial Complex (RSC) were identified in each participant using a standard block-design functional localizer scan. The other sessions each consisted of six event-related presentations of all 90 images, under three different task instructions: action affordance labeling, object labeling, or an orthogonal task at fixation. For the current set of analyses, stimulus-specific beta coefficients were estimated per run and then averaged across all 18 image repetitions (i.e. across tasks). RDMs were created for each ROI by computing pairwise Pearson's correlation distances between z-scored t-values of the beta estimates, and averaged across subjects.

CNN representations

Computational model features were extracted from three CNNs architectures (AlexNet, ResNet18 and ResNet50) trained for scene classification on the Places365 dataset **(Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2018)**. RDMs were computed using pairwise correlation distances between feature activations to all 90 stimuli in each network layer. We here report the highest correlation with a single layer for each network, which varied between the models but was either one of the highest convolutional layers or a fully connected layer.

Between-RDM comparisons

We compared the behavioral, fMRI and CNN RDMs using Spearman correlations. Significance was determined with a two-tailed, Bonferroni-corrected permutation test (n = 10,000), and 95% confidence intervals were estimated by computing a bootstrap distribution of correlation values (n = 10,000).

Results

Affordances correlate weakly with other properties

We first examined to what extent action affordance ratings correlated with other behavioral annotations (**Fig. 2A**). Affordance ratings exhibited comparatively lower correlations (average 0.3, *SD* = 0.07) with other scene properties than these properties correlated with one another (average 0.61, *SD* = 0.1). This suggests that action affordance ratings are not immediately reducible to other scene properties, instead forming a distinct representational space.

Scene-selective cortex represents affordances

Of the five different behaviorally annotated scene properties, action affordances show the highest correlation with all scene-selective regions (**Fig. 2B**). PPA additionally correlates significantly with both material and object annotations, and OPA and RSC with object annotations. These results provide new evidence of action affordance-related representations in scene-selective regions, which appear distinct from representations of objects, materials and other scene properties.

CNNs weakly represent affordances and poorly predict fMRI responses

CNN features have significant correlations with all behavioral annotations (**Fig. 3**). However, correlations with action affordances are substantially lower than with other scene properties. We find the highest correlations between the CNN features and global property ratings of each scene. Comparisons between the CNN features and fMRI responses also show substantially lower correlations than in prior studies, e.g., **(King, Groen, Steel, Kravitz, & Baker, 2019)**.

Conclusion

Collectively, these results show that scene-selective regions are sensitive to action affordances, in addition to other scene properties. Our analyses so far show that scene-trained CNNs do not strongly represent action affordances, suggesting that different task objectives may be needed to fully capture the computations in scene-selective regions.

Further analyses of these data will focus on determining the degree of unique representations of action affordances in fMRI responses, investigating the effect of task instruction on neural representation of scene properties, and comparison with CNNs trained on other tasks than scene classification.

Acknowledgments

This work was supported by an VENI grant (VI.Veni.194030) from the Netherlands Organisation for Scientific Research (NWO) to IIAG.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020, February). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. doi: 10.3758/s13428-019-01237-x

Bonner, M. F., & Epstein, R. A. (2017, May). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798. doi: 10.1073/pnas.1618228114

Epstein, R. A., & Baker, C. I. (2019, September). Scene perception in the human brain. *Annual Review of Vision Science*, *5*(1), 373–397. doi: 10.1146/annurev-vision-091718-014809

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, *1*(2), 67–82.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82–94. doi: 10.1037/xge0000129

Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018, March). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, *7*, e32962. doi: 10.7554/eLife.32962

King, M. L., Groen, I. I., Steel, A., Kravitz, D. J., & Baker, C. I. (2019, August). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, *197*, 368–382. doi: 10.1016/j.neuroimage.2019.04.079

Palan, S., & Schitter, C. (2018, March). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. doi: 10.1016/j.jbef.2017.12.004

Patterson, G., & Hays, J. (2012). SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2751–2758). Providence, RI: IEEE. doi: 10.1109/CVPR.2012.6247998

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. doi: 10.1109/TPAMI.2017.2723009