Object Real-World Size Representations in Human Brains and Artificial Neural Networks

Zitong Lu (lu.2637@osu.edu)
Department of Psychology, The Ohio State University, 1827 Neil Ave
Columbus, OH 43212 USA

Julie Golomb (golomb.9@osu.edu)
Department of Psychology, The Ohio State University, 1827 Neil Ave
Columbus, OH 43212 USA

Abstract:

Human brains have the ability to process various object features, not only color, shape, texture, and animacy, but also real-world size, during object recognition. However, studies examining representations of perceived real-world size may be confounded with a related dimension, perceived real-world depth. In this study, we aimed to isolate representations of object real-world size from both visual (image) size and perceived depth information in both human brains and artificial neural networks using the THINGS EEG2 dataset, which incorporated more naturalistic stimuli. Our results successfully differentiated various visual information and revealed a pure representation of object real-world size in human brains. Furthermore, representational comparisons with different artificial neural networks offers further insight into the dissociated mechanisms of forming real-world size, visual size, and real-world depth perception.

Keywords: real-world size; depth perception; RSA; artificial neural networks; object recognition; visual cortex

Instructions

If we are viewing an apple while walking around in the real world, as we change our perspective and distance, the apple's visual (retinal) size varies, but we can still perceive the apple as having a constant real world size. How and when is real world size represented during visual processing? Behavioral studies have demonstrated familiar-size stroop effects (Konkle & Oliva, 2012a; Long & Konkle, 2017) and canonical visual size effects (Chen et al., 2022; Konkle & Oliva, 2011) related to object real-world size. Also, human fMRI studies using univariate analysis have found that the ventral temporal cortex encodes object real-world size information (Huang et al., 2022; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b). These findings suggest real-world size is a fundamental property of object representation. However, real-world size is closely related to distance in depth. For instance, because a walnut has a smaller real-world size than a basketball, if we viewed images of those two objects with matched visual (retinal) size, we would perceive the walnut as closer to us than the basketball.

In previous neuroimaging studies of perceived real-world size, researchers controlled the visual size of objects, but did not control for perceived real-world depth, which could also serve as a dimension to explain the results. Deep convolutional neural networks (CNNs) have also been found to represent real-world size in a recent computational study (Huang et al., 2022). However, they used same fixed visual size image dataset, making it difficult to determine whether CNNs encode real-world size or depth. While an increasing number of studies confirm that CCNs exhibit representations similar to human visual systems (Cichy et al., 2016; Güçlü & van Gerven, 2015; Yamins et al., 2014; Yamins & DiCarlo, 2016), some recent works indicated that semantic embedding and multimodal neural networks could better explain human visual representations in visual areas and even hippocampus than vision-only networks (Choksi, Mozafari, et al., 2022; Choksi, Vanrullen, et al., 2022; Conwell et al., 2022; Doerig et al., 2022; Jozwik et al., 2023; Wang et al., 2022). Investigating how different artificial neural networks (ANNs) represent object real-world size and other size and depth features can not only help us understand how our brains process object features but also provide insights for developing more brain-like models.

In the current study, we aimed to used computational methods to distinguish the representations of object real-world size and other size and depth features in both human brains and artificial neural networks based on an open EEG dataset, THINGS EEG2 (Gifford et al., 2022). The images used in this dataset are more naturalistic and include objects that vary in real world size, depth, and visual size (as opposed to prior datasets where images were isolated objects all presented at the same visual size). Our results reveal that human brains indeed have pure object real-world size representations, which emerge later in processing than real-world depth and visual size representations. Additionally, although size and depth are closely related, representational results from different ANNs suggest that the perception of size and depth may arise through distinct mechanisms.

Methods

Image and EEG dataset: We utilized the open dataset from THINGS EEG2 (Gifford et al., 2022) , which includes EEG data from 10 healthy human subjects viewing 16740 images of natural scenes and objects. We specifically used the 'test' dataset portion, which includes 1600 trials corresponding to 200 images with 80 trials per image. We used already pre-processed data from 17 channels overlying occipital and parietal cortex.

ANN models: We used four pre-trained models, including one visual model (ResNet-101 (He et al., 2016)), one semantic model (Word2Vec (Mikolov et al., 2013)), one multi-modal (visual+semantic) model (CLIP with a ResNet-101 backbone (Radford et al., 2021)), and one brain-like model (CORnet-S (Kubilius et al., 2019)). We used THINGSvision (Muttenthaler & Hebart, 2021) to obtain ANN activations for the images.

Representational similarity analysis (RSA): We first constructed four hypothesis-based representational dissimilarity matrices (RDMs): (1) Real-World Size RDM based on the perceived real-world size of the objects (human ratings from Stoinski et al., 2022), (2) Visual Size RDM based on the visual size of the objects (the measured size of the segmented object in pixels), (3) Real-World Depth RDM based on the perceived depth of the objects (estimated as visual size index / real-world size index), and (4) a Low-Level Visual RDM based on low-level visual similarity (image pixel-wise correlations). For EEG, we constructed timepoint-by-timepoint neural RDMs for each subject using classification-based decoding for each pair of objects, with decoding accuracy as the dissimilarity index. For ANN models, we constructed seven RDMs: early (second convolutional layer) and late (last visual layers) for ResNet, CLIP, and CORnet, plus a single RDM for Word2Vec. For RSA (Kriegeskorte et al., 2008), we calculated partial correlations between (1) the hypothesis-based RDMs and timepoint-by-timepoint EEG neural RDMs, and (2) the hypothesis-based RDMs and the ANN RDMs. All RSA was implemented using NeuroRA (Lu & Ku, 2020).

Results

Dynamic representations in human brains

The Real-World Size RDM showed significant representational similarity with EEG neural RDMs from 169-240ms and 490-550ms. The partial correlation technique suggests that this reflects a pure representation of object real-world size in the human brain, independent from Visual Size and Real-World Depth, which showed significant representational similarity at different time windows.

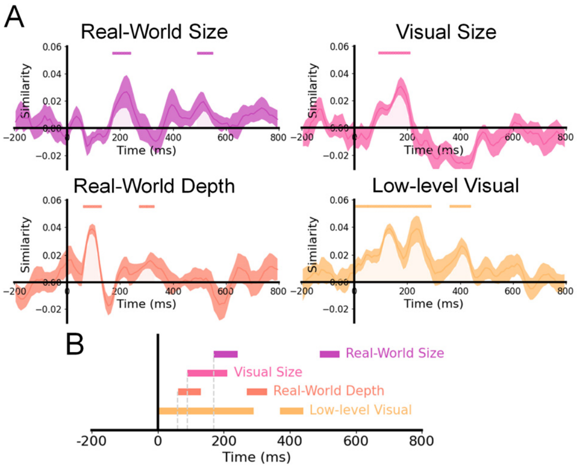

Figure 1: (A) Temporal similarities (partial correlations) between hypothesis-based RDMs and EEG RDMs. (B) Significant time-windows of partial correlation. Color-coded small squares indicate significant timepoints, cluster-corrected *p*<.05.

Representations in artificial neural networks

We found significant representations of real-world size in Word2Vec and the late layers of ResNet, CLIP and CORnet, which might suggest that object real-world size emerged later, possibly from semantic information. Visual size representation was significant in the early layers of models and only slightly in the late layer of ResNet. Real-world depth representation was significant in the early layer of ResNet, and both early and late layers of CLIP. All models represent low-level visual information.

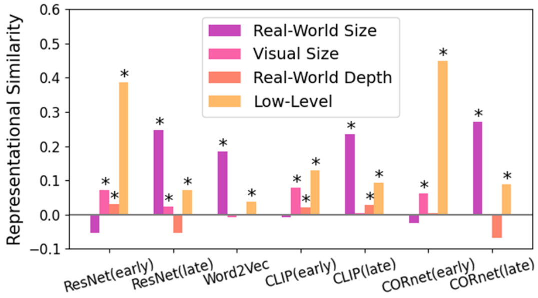

Figure 2: Real-world size, visual size, real-world depth, and low-level visual representations in different ANNs.

Conclusion

Our study applied computational methods to distinguish the representations of object real-world size and other size and depth features in both human brains and ANNs. Consistent with prior studies reporting real-world size representations (Huang et al., 2022; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b) , we found that human brains and ANNs contain significant information about real-world size. Critically, unlike the prior studies, we were able to dissociate pure real-world size representations from both visual size and real-world depth representations, Moreover, using EEG we uncovered a representational timeline for visual object processing, with low-level visual information represented first, followed by real-world depth and visual size, and finally real-world size. Finally, our ANN results offer further insight, and might suggest that although real-world object size and depth are closely related, the combination of semantic and visual information may help us perceive real-world size, while real-world depth perception may need only visual information.

Acknowledgments

This work was supported by research grants from the National Institutes of Health (R01-EY025648) and from the National Science Foundation (NSF 1848939).

References

Chen, Y. C., Deza, A., & Konkle, T. (2022). How big should this object be? Perceptual influences on viewing-size preferences. *Cognition*, *225*, 105114.

Choksi, B., Mozafari, M., VanRullen, R., & Reddy, L. (2022). Multimodal neural networks better explain multivoxel patterns in the hippocampus. *Neural Networks*, *154*, 538–542.

Choksi, B., Vanrullen, R., & Reddy, L. (2022, August 25). Do multimodal neural networks better explain human visual representations than vision-only networks? *Conference on Cognitive Computational Neuroscience 2022*.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports 2016 6:1*, *6*(1), 1–13.

Conwell, C., Prince, J. S., Alvarez, G. A., & Konkle, T. (2022). Large-Scale Benchmarking of Diverse Artificial Vision Models in Prediction of 7T Human Neuroimaging Data. *BioRxiv*.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *ArXiv*.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, T., Song, Y., & Liu, J. (2022). Real-world size of objects serves as an axis of object space. *Communications Biology 2022 5:1*, *5*(1), 1–12.

Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N., & Mur, M. (2023). Deep Neural Networks and Visuo-Semantic Models Explain Complementary Components of Human Ventral-Stream Representational Dynamics. *Journal of Neuroscience*, *43*(10), 1731–1741.

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242.

Konkle, T., & Oliva, A. (2011). Canonical Visual Size for Real-World Objects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 23–37.

Konkle, T., & Oliva, A. (2012a). A familiar-size Stroop effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 561.

Konkle, T., & Oliva, A. (2012b). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers* *in Systems Neuroscience*, *4*.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N. J., Issa, E. B., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L. K., & Dicarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*.

Long, B., & Konkle, T. (2017). A familiar-size Stroop effect in the absence of basic-level recognition. *Cognition*, *168*, 234–242.

Lu, Z., & Ku, Y. (2020). NeuroRA: A Python Toolbox of Representational Analysis From Multi-Modal Neural Data. *Frontiers in Neuroinformatics*, *14*, 61.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Muttenthaler, L., & Hebart, M. N. (2021). THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. *Frontiers in Neuroinformatics*, *15*, 45.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the International Conference on Machine Learning (ICML)*.

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2022). THINGSplus: New Norms and Metadata for the THINGS Database of 1,854 Object Concepts and 26,107 Natural Object Images. *PsyArXiv*.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2022). Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience 2016 19:3*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624.