The role of motion cues in object representation: A study of visual area specializations using deep learning

Nastaran Darjani (Nastaran.darjani@umontreal.ca)
Department of Computer Science, University of Montreal, Montreal, Canada

Maryam Vaziri-Pashkam (mvaziri@udel.edu)
Movement and Visual Perception Lab, Department of Psychological and Brain Sciences, University of Delaware, Newark, DE, USA

Shahab Bakhtiari (shahab.bakhtiari@umontreal.ca)
Department of Psychology, University of Montreal, Mila (Quebec AI Institute), Montreal, Canada

Abstract

The human brain processes both static and motion-defined visual cues for object representation, yet most computational models emphasize static information. We investigated neural responses to motion-defined object stimuli (”object kinematograms”) by comparing brain activity with a dual-pathway artificial neural network that separates slow- and fast-varying visual information. Our findings show that while this dual-stream network captures aspects of biological motion processing, integration of slow and fast information improves similarity to brain in some regions but not others. These results highlight the functional diversity across visual areas in dynamic object representation.

Keywords: object categorization, object motion, kinematogram, dual-pathway model, Artifical Neural Network

Introduction

While motion information is essential for tracking objects and understanding environmental dynamics, its role in object categorization remains understudied. Recent work by Robert, Ungerleider, and Vaziri-Pashkam (2023) investigated how motion cues inform object categorization and contribute to object representation in the visual system. They used object kinematograms in which the object structure is defined solely by motion cues. Their findings revealed that motion-defined object information is processed by a broad network of visual areas that extends beyond classical motion-sensitive regions, including regions traditionally associated with static object processing. These results challenge the classical view that motion processing is strictly separate from object form perception and suggest that motion cues may influence object form representations across multiple areas of the visual system. Nevertheless, how visual areas process and integrate fast- and slow-varying visual information towards object representation is unclear. To address this gap, here we use a deep learning approach using an artificial neural network (ANN) with two parallel specialized pathways (Feichtenhofer, Fan, Malik, & He, 2019). This ANN processes visual information at different temporal resolutions by directing slow and fast-varying motion information through separate pathways, while allowing information integration via cross-pathway connections. We input object kinematograms into the model and compare its layer-wise representations with fMRI responses to the same stimuli in multiple visual regions, probing the distinct slow and fast representational properties of each region. Through ablation experiments on the model, we further characterize the temporal processing profiles of individual visual areas.

Results

To examine how motion cues contribute to object representations in the brain, we analyzed fMRI responses to “object kinematograms”. These stimuli were generated by extracting motion vectors from videos of single moving objects and mapping them onto random dot patterns. A prior behavioral study confirmed clear categorical distinctions between three categories of animate (human, mammal, reptile) and three inanimate (tool, ball, pendulum/swing) objects (Robert et al., 2023). These neural activations were compared with activations from the SlowFast network, a two-stream ANN trained on action recognition (Kay et al., 2017) that processes visual input through parallel slow (low temporal resolution) and fast (high temporal resolution) pathways. We computed representational dissimilarity matrices (RDMs) from ANN ReLU activations and fMRI responses, then used Kendall’s Tau correlation and cluster permutation-based significance tests to measure representational similarity (RSA) between the model and the brain, across visual areas (Nili et al., 2014; Maris & Oostenveld, 2007).

Comparisons with the Slow and Fast Pathways

We examined how each network pathway aligned with brain activity of seven visual areas (Figure 1A) using RSA. The regions were selected based on previous studies on object kinematograms (Robert et al., 2023). Comparisons with the slow (S_{wx}) and fast (F_{wx}) pathways of the model suggest a diverse pattern of specializations across visual areas (Figure 1B). In the early visual cortex (V1), neither the slow (S_{wx}) nor fast (F_{wx}) pathway showed significant correlation, consistent with the low-level feature processing of V1. Similarly, neither pathway was aligned with the object-selective posterior fusiform sulcus (pFS), suggesting limited sensitivity to motion-defined objects as captured by our ANN. The remaining areas displayed variable similarities, predominantly with S_{wx}. The object-selective lateral occipital cortex (LO) showed weak sim-

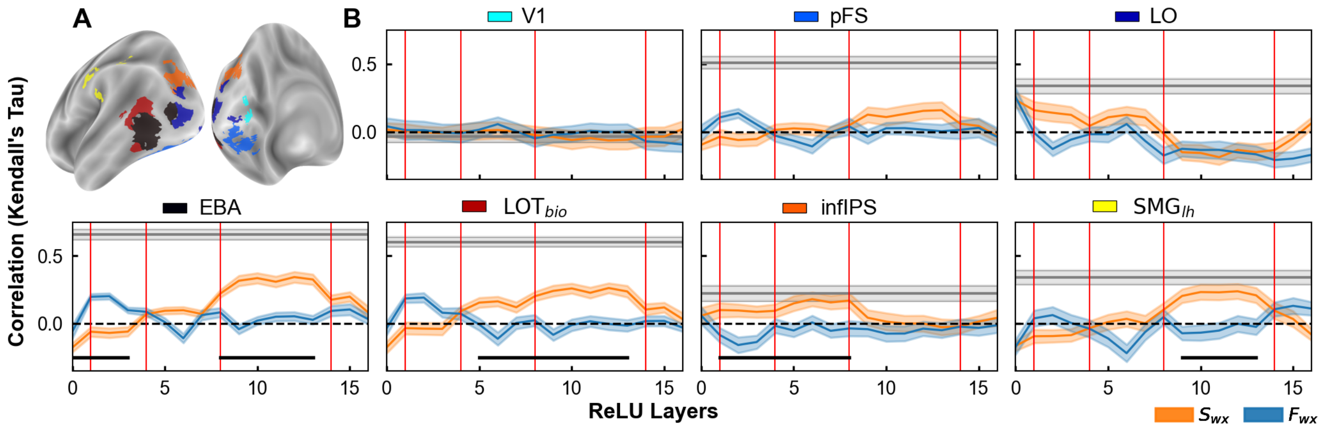

Figure 1: RSA between networks and brain regions. Error bars show SEM, black lines significant RSA differences, gray areas the noise ceiling and red vertical lines the cross-pathway connections.

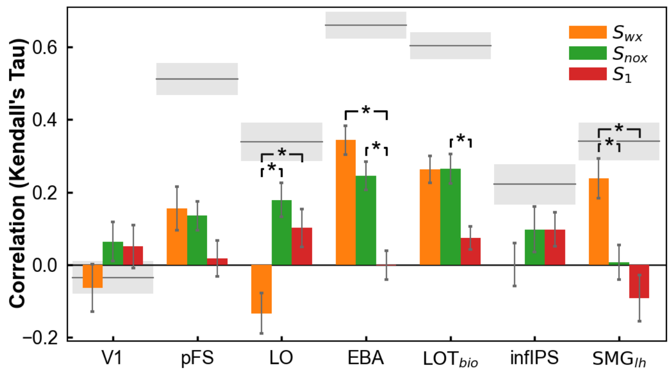

Figure 2: Results of the RSA comparing layer 12 of networks (layer with the largest difference in Figure 1) with the brain. Error bars show SEM, Asterisks significance, and the gray area the noise ceiling SEM.

ilarity to both pathways in the early layers, but the deeper layers exhibited a negative correlation, suggesting that the LO organizes information about object kinematograms in a manner that differs systematically from the representational structure of this ANN. inFIPS correlated more strongly with S_{wx} than F_{wx}, especially in the early layers. However, this similarity declined after the third cross-pathway connection. Extrastriate body area (EBA) and LOT-bio, associated with perception of body and biological motion, showed significantly higher similarity to S_{wx} than F_{wx}, especially in the deep-intermediate layers. Furthermore, the left supramarginal gyrus (SMG) showed stronger alignment with S_{wx}. The areas included in this study appear to cluster into three groups: those that showed no significant similarity to the model (V1 and PFS), those exhibiting decreasing similarity throughout the layers of the model (LO and inFIPS), and those demonstrating significant and relatively high similarity, particularly with the higher intermediate layers of S_{wx} (EBA, LOT-bio, SMG_{lh}). In the following sections, we focus on the 12th layer, which showed the largest RSA difference between pathways across all regions, as the representative layer (Figure 2).

Effect of Cross-pathway Connections

In the SlowFast architecture, there are occasional cross-pathway connections from the fast to the slow pathway. Due to these cross-pathway connections, it remained unclear whether the observed (dis)similarities with the slow pathway stemmed from slow-varying or was related to the fast-varying information channeled through the fast pathway to the slow pathway. To assess the impact of motion integration, we used an ablation experiment to examine a model without cross-pathway connections (S_{nox}). Removing the connections between the slow and fast pathways did not have a significant effect on alignment with V1, PFS, EBA, and LOT-bio. However, S_{nox} showed similarity with LO, whereas S_{wx} did not (orange vs. green in Figure 2), suggesting that integration with the fast pathway reduces alignment with LO. This indicates that LO's representational geometry aligns more closely with slow-varying information, and that incorporating fast-varying motion cues decreases similarity to this region. Removing the cross-pathway connection also improved alignment with inFIPS, but this effect did not reach statistical significance. Interestingly, unlike LO and inFIPS, S_{wx} showed significantly higher alignment with the left SMG than S_{nox} in later layers, suggesting that integrating fast-varying motion information increases representation similarity with this region.

Importance of Motion During Training

Finally, we asked whether the concurrent training of the interconnected slow and fast pathways had influenced the observed similarities between brain activity and the slow pathway (e.g., with EBA). To address this question, we compared S_{nox} with an independently trained slow-only model (S_1). S_{nox} significantly outperformed S_1 in EBA and LOT-bio, while both models performed similarly in other regions. This suggests that the representations of these two regions do not depend on the input of direct fast-varying motion during inference, as indicated by similar correlations with S_{wx} and S_{nox}. However, exposure to fast motion during learning had a significant impact on the development of their functional specialization. This is reflected in the higher similarity with S_{nox} compared to S_1.

Conclusion

Our findings demonstrate distinct patterns across visual areas in alignment with the slow and fast pathways of a dual-pathway ANN. Our results highlight functional specificities across visual areas in processing motion-defined objects, suggesting variability in how different brain regions handle slow and fast temporal cues during dynamic object perception.

Acknowledgment

We gratefully acknowledge Sophia Robert (Carnegie Mellon University) for sharing her fMRI dataset and results, which were essential to this study. Her contributions significantly enhanced the quality and impact of our research. This project was supported by funding from NSERC Discovery Grant RGPIN-2023-03875 to S.B. and by Digital Research Alliance of Canada (DRAC) and Calcul Québec.

References

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6202–6211).

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, *10*(4), e1003553.

Robert, S., Ungerleider, L. G., & Vaziri-Pashkam, M. (2023). Disentangling object category representations driven by dynamic and static visual input. *Journal of Neuroscience*, *43*(4), 621–634.