Neural representations of implied motion in body-selective regions revealed by caption-based encoding models

Ryuto Yashiro (yashiror@zedat.fu-berlin.de)
Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
Institute of Cognitive Science, Universität Osnabrück, Osnabrück, Germany
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

Masataka Sawayama (masataka_sawayama@ist.hokudai.ac.jp)
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

Ayumu Yamashita (ayumu722@g.ecc.u-tokyo.ac.jp)
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

Kaoru Amano (kaoru_amano@ipc.i.u-tokyo.ac.jp)
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

Abstract

Recent studies using encoding models that accurately predict brain responses to natural images from corresponding captions suggest that high-level visual regions encode the rich semantics of a scene, rather than merely responding to specific categories such as faces or bodies. Here we focused on the extrastriate body area (EBA) and investigated what semantic content is represented in this region by analyzing the relationship between the co-occurrence of multiple objects in large-scale natural scene captions and EBA responses predicted by caption-based encoding models. We found that responses in EBA as well as the fusiform body area (FBA) outside motion-processing regions exhibited high correlation with the perceived motion speed of a human body implied in an image, suggesting that these body-selective regions encode the implied motion of a human body inferred from the co-occurring objects in a scene.

Keywords: EBA; FBA; implied motion; encoding model

Introduction

The visual system has a remarkable capacity to accurately recognize human bodies. One primary contributor of this capacity is the extrastriate body area (EBA), which exhibits stronger response to human body parts relative to other objects (Downing et al., 2001). Prior research has demonstrated that this body-selective region specializes in analyzing a static form of human bodies by measuring fMRI responses to simplified and well-controlled body stimuli (Peelen et al., 2006; Urgesi et al., 2004).

In our real world, however, humans need to recognize bodies in a cluttered scene with diverse objects, rather than an isolated body in a uniform background. To understand the functional property of visual regions under such circumstances, it has become increasingly popular for neuroscientists to investigate brain responses to natural images. For example, recent studies utilized large language models (LLMs), which generate fixed-length vectors (embeddings) from image captions, to predict brain responses to the corresponding natural images (Doerig et al., 2022; Luo et al., 2023). This caption-based encoding model showed higher capacity to predict brain responses in high-level visual regions including EBA compared to models using categorical embeddings, suggesting that EBA represents not just a form of human bodies but complex semantic contents of a scene.

Despite this new perspective of the neural representation in EBA, it has yet to be understood precisely what semantic information is represented in EBA due to the lack of interpretability of caption embeddings. In the present study, inspired by the fact that the semantic content of a scene is generally shaped by the co-occurrence of multiple objects, we performed a novel analysis on the relationship between co-occurrence statistics of natural scene captions and corresponding EBA responses, which led us to the hypothesis that dynamic human action implied in static images (i.e., implied motion) may be linked to EBA responses. Indeed, a subsequent experiment and analysis showed that EBA exhibited high correlation with implied motion ratings provided by humans, as well as with other body-related features.

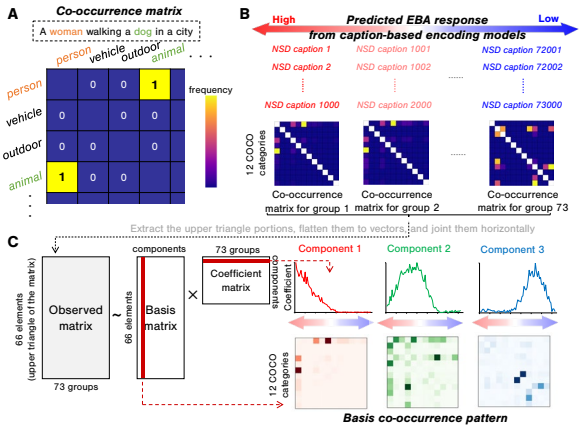

Figure 1: Overview of the co-occurrence analysis.

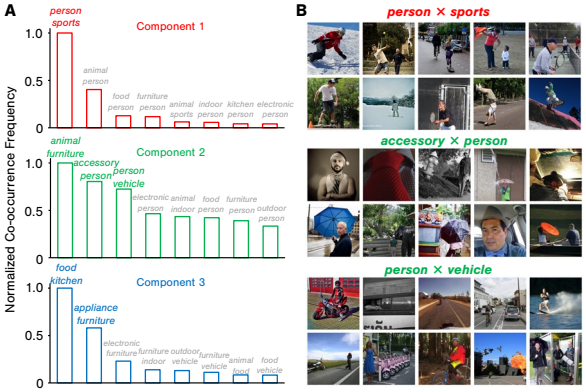

Figure 2: (A) Top-8 co-occurrent categories for each component (colored if the value exceeds 0.5). (B) Sample NSD images with the frequently co-occurring categories.

Results and Discussion

Co-occurrence analysis. We used the Natural Scenes Dataset (NSD) (Allen et al., 2021) which contains large-scale fMRI responses to 73,000 natural images for 8 human subjects. To understand how object co-occurrence is linked to EBA responses, we first trained linear encoding models on mean EBA responses to the NSD images based on the corresponding LLM (MPNet) (Song et al., 2020) embeddings of the captions from MS COCO (Chen et al., 2015). Then we sorted 73,000 captions

in descending order of predicted EBA response from the encoding models and divided these captions into 73 groups, each containing 1,000 captions. For each group, we defined a matrix using 12 MS COCO superordinate categories (person, vehicle, outdoor, animal, accessory, sports, kitchen, food, furniture, electronic, appliance, indoor; see an example in Figure 1A) and their co-occurrence frequency in a set of captions (Figure 1B). Next, we performed non-negative matrix factorization to extract 3 major co-occurrence components (Figure 1C). Importantly, each row of the decomposed coefficient matrix reflects how strongly each major co-occurrence component is associated with the 73 groups, thereby capturing its relationship to the magnitude of EBA responses (high, moderate and low response for component 1, 2, and 3 respectively, given the peak locations).

Figure 2A shows top-8 pairs of word categories with high co-occurrence frequency for each component. As expected, component 1 and 2 (high and moderate EBA response) included the "person" category. Notably, it cooccurred with different categories in each component: "sports" in component 1 while "accessory" and "vehicle" in component 2. We qualitatively assessed images whose captions contain these co-occurrent categories (Figure 2B) and found that images associated with higher EBA responses tend to imply more dynamic human body motion. Next, we tested this observation by collecting implied motion ratings of NSD images in a behavioral experiment.

Rating experiment. We subsampled 100 NSD images of a person for our experiment. 5 participants were presented with an image for 3 secs and evaluated the speed of the implied motion for the person in each image on a scale of 1 to 5. We provided participants with rating criteria: 1 indicates static and 5 corresponds to the fastest motion they had previously seen in a passive-viewing session in which all 100 images were successively presented before the main rating experiment.

Correlation analysis. We computed the rank correlation between the rating scores and mean EBA responses to the 100 images for each NSD subject. As we confirmed high between-participant reliability for the rating score (rank correlation > 0.8), we used the scores averaged across participants for our analysis. Figure 3 shows a correlation map throughout the visual cortex for one representative subject with functionally defined body-selective regions (yellow; EBA and FBA) and anatomically defined motion processing regions (green and blue; MT and MST). Consistent with previous studies on the neural representation of implied motion (Kourtzi & Kanwisher, 2000; Lu et al., 2016; Senior et al., 2000), we observed high correlation in the MT and MST. Importantly, EBA and FBA outside the MT and MST also exhibited high correlation with implied motion, corroborating our hypothesis from the co-occurrence analysis. To quantitatively compare these regions, we computed the proportion of vertices with significantly high correlation for each region and subject using a permutation test and FDR correction ($p < 0.05$). The proportion of vertices showing significantly high correlation in EBA and FBA was comparable to that in MT (Figure 4A).

Lastly, to test the effect of other body-related image features, we computed body size, number of people, and spatial distance between the body and image center for each image using a segmentation model (Ren et al., 2024), and correlated the responses with these features. We again computed the proportion of vertices with significant correlation and found that a comparable proportion of vertices represent implied motion as well as number of people in EBA and body size in FBA (Figure 4B).

These results suggest that parts of EBA and FBA represents the perceived implied motion of human bodies that can be shaped by the co-occurrence of people and objects in a scene. Future studies should extend our approach to other fMRI datasets to evaluate the generalizability of our findings.

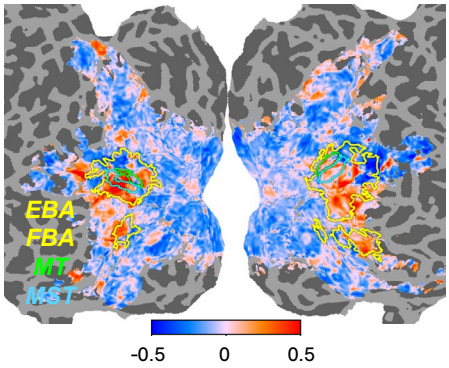

Figure 3: Correlation between responses and implied motion ratings on a surface map for NSD subject 1.

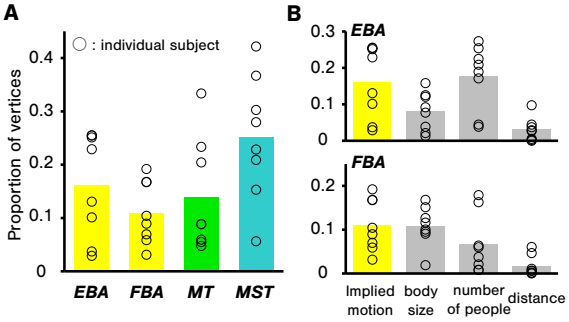

Figure 4: Proportion of vertices whose responses correlated significantly with implied motion of the images (A) and other body-related image features (B).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 23KJ0477 to RY.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2021). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1504.00325

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. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2209.11737

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science (New York, N.Y.)*, *293*(5539), 2470–2473.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience*, *12*(1), 48–55.

Lu, Z., Li, X., & Meng, M. (2016). Encodings of implied motion for animate and inanimate object categories in the two visual pathways. *NeuroImage*, *125*, 668–680.

Luo, A., Henderson, M. M., Tarr, M. J., & Wehbe, L. (2023, October 13). BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity. *The Twelfth International Conference on Learning Representations*. https://openreview.net/pdf?id=mQYHXUUTkU

Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron*, *49*(6), 815–822.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., & Zhang, L. (2024). Grounded SAM: Assembling open-world models for diverse visual tasks. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2401.14159

Senior, C., Barnes, J., Giampietro, V., Simmons, A., Bullmore, E. T., Brammer, M., & David, A. S. (2000). The functional neuroanatomy of implicit-motion perception or representational momentum. *Current Biology: CB*, *10*(1), 16–22.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.09297

Urgesi, C., Berlucchi, G., & Aglioti, S. M. (2004). Magnetic stimulation of extrastriate body area impairs visual processing of nonfacial body parts. *Current Biology: CB*, *14*(23), 2130–2134.