Separate Neural Representations for Physical and Communicative Social Interactions: Evidence from Data-driven Voxel Decomposition

Yuanfang Zhao (yzhao213@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
34000 N. Charles Street, Baltimore, MD 21218

Emalie McMahon (emaliemcmahon@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
34000 N. Charles Street, Baltimore, MD 21218

Leyla Isik (lisik@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
34000 N. Charles Street, Baltimore, MD 21218

Abstract

Recognizing social interactions is remarkable for its adaptive significance. Previous studies have suggested that the lateral visual cortex and superior temporal sulcus (STS) are generally involved in social interaction perception. However, it has been difficult to further disentangle neural responses of different types of social interaction with hypothesis-driven approaches, due to challenges with feature labeling, sampling and experimenter bias. Employing a data-driven voxel decomposition technique (i.e., non-negative matrix factorization, NMF) to a large-scale naturalistic fMRI dataset, our analysis of the lateral visual cortex and STS revealed two components with distinct functional profiles related to social interaction. The first component responds strongly to joint physical actions between people in the videos and weighs strongly in mid-level regions of the lateral stream, including middle temporal area (MT) and extrastriata body area (EBA). Conversely, the second component responds strongly to communicative interaction between people in the videos and weighs heavily in the anterior STS. Together, our findings suggest that joint action and communication represent two distinct forms of social interaction that are encoded differently in posterior to anterior regions along the lateral visual pathway.

Keywords: social interaction; non-negative matrix factorization; joint action; communication

Introduction

Social interaction takes many forms. For example, consider scenarios such as “one person chasing another” and “two people talking with each other”. While both scenarios are undoubtedly recognized as social interactions, they differ fundamentally in terms of characteristic visual features and abstract aspects such as the relationship between agents. Although previous studies have suggested that social interaction perception generally recruits the lateral visual cortex and STS – referred to as the “third visual pathway” (Pitcher & Ungerleider, 2021) – it remains unknown whether there are distinct features representing different types of social interactions, and if so, which neural substrates are involved.4

A potential approach to address this question is to test key features hypothesized to represent different types of social interaction. However, a major weakness of this hypothesis-driven approach is that, regardless of a hypothesis’ validity, there are always endless features outside the hypothesis not having been tested, making it possible to miss critical features.

Here, we took a hypothesis-neutral approach to investigate a rich naturalistic fMRI social interaction dataset. Specifically, we applied a data-driven voxel decomposition technique (NMF) to de-mix the underlying neural responses that are otherwise spatially intermingled within individual fMRI voxels (Figure 1, Khosla, Murty, & Kanwisher, 2022). The de-mixed neural responses represent relatively independent components that have distinct functional profiles. We identified components that responded consistently across participants and analyzed both their functional profiles and anatomical locations.

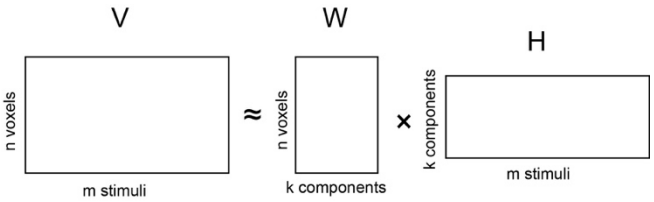

Figure 1: Illustration of NMF

Methods

fMRI dataset We used an open dataset (McMahon, Bonner, & Isik, 2023), which includes fMRI responses from 4 participants to two hundred 3-second video clips. These clips, depicting two individuals engaging in various social and nonsocial activities, were sampled from everyday scenes. Additionally, we used functional ROIs released with the dataset to provide a comprehensive view of the neuroanatomical distribution of each component. All analyses were conducted in MNI152 space.

NMF decomposition Focusing our analysis on voxels in the lateral visual cortex and STS, areas known as the “third visual pathway” and specialized for social motion, we extracted each participant’s fMRI response to the videos using a combined mask of lateral visual cortex and STS, using anatomical parcels from Natural Scenes dataset (Allen et al., 2022) and functional parcels on social perception (Deen et al., 2015). This generated a n (voxels) by m (stimuli) fMRI responses matrix (V) for each participant. After normalization and subtracting the minimum value from the responses in each voxel, the matrix was decomposed using NMF. NMF decomposes a high-dimension matrix (V) into two low-dimension matrices, W and H, under the constraint that all matrices must be non-negative (Figure 1). The resulting matrix H was a k (components) by m (stimuli) response matrix, representing the response magnitude of each component to each video. The matrix W was an n (voxels) by k (components) weight matrix, indicating the influence (i.e., weight) of each component on each voxel. In this way, fMRI responses were decomposed

into several components with distinct stimulus response profiles and voxel weights. The optimal number of components were determined as the one yielding the maximum marginal likelihood (Schmidt, Winther, & Hansen, 2009), resulting in 4 components for each participant. Finally, we selected components in each participant that showed high inter-subject consistency (> 0.5, Khosla et al., 2022) and averaged these consistent components across participants, resulting in 2 group-averaged components.

Functional profiling To identify the functional significance of each component, we used video annotations released with the dataset, ranging from low-level features like motion energy to mid-level features including agent distance and facingness, and high-level features such as joint action (i.e., whether two people are acting jointly or not) and communication. We explored which features were most predictive of a component's response magnitude. Additionally, we collected free-response captions from an independent group of participants' (N = 5) for each video. These captions were analyzed using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, which evaluates the specificity of each word in a caption relative to all the captions (Ramos, 2003). By averaging TF-IDF results for captions of the 30 top-responding videos, we were able to determine which words were most representative of those videos.

Voxel weight analysis The voxel weights of components in the later visual cortex and STS were projected onto the MNI152 template and analyzed with functional ROIs, to shed light on their anatomical locations.

Results

Functional profiling We found that the first component had the strongest correlation with the feature "joint action", even after controlling for features such as motion energy (Figure 2A). Consistently, TF-IDF analysis revealed that the most representative word used in describing the 30 top-responding videos was "dancing", and the highest-responding video shows two people "fighting", both typical forms of joint physical actions performed between people (Figure 2B). Conversely, the second component had the strongest correlation with the feature "communication" (Figure 2C). Consistently, TF-IDF analysis revealed that the most representative word used in describing the 30 top-responding videos was "talking" (Figure 2D).

Voxel weight analysis The first component was most strongly weighted in mid-level regions of the lateral stream, including the middle temporal area (MT) and extrastriate body area (EBA) (Figure 3A). In contrast,

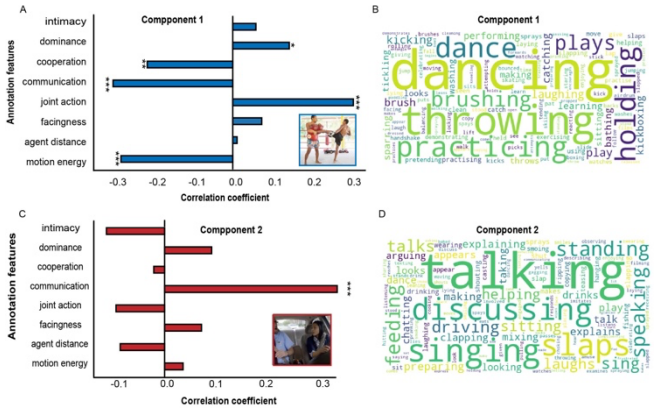

Figure 3: A, C) Partial correlation between features and response magnitude for each component. Image representative of highest-responding video. B, D) World cloud of TF-IDF results for 30 top-responding videos for each component.

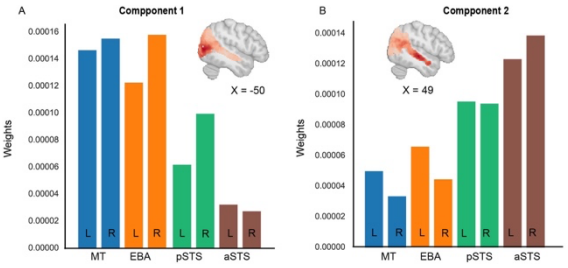

Figure 2: A, B) Voxel weight distribution and averaged weights in each ROI for each component. In the bar plot, annotations are provided to indicate the hemisphere (L: left; R: right).

the second component was highly weighted in the anterior STS (Figure 3B).

Discussion

By leveraging the hypothesis-neutral voxel decomposition technique, our results suggest that social interactions can be classified according to two distinct social features. One feature encodes the joint physical action of two people and is represented mainly in the MT and EBA, which are the visual motion and body areas. The other feature encodes whether two people are communicating with each other and is represented most strongly in the middle and anterior STS. These results are consistent with prior behavioral distinctions found between these categories (Wu et al., 2024) but provide the first neural evidence of this dichotomy. More generally, our results speak to the hierarchical structure of social interaction in the lateral pathway, from visual analysis of joint body movement in the posterior regions to the more abstract representation of communication in the anterior regions.

Acknowledgements

This work was funded by R01 grant (No. NIH R01MH132826) awarded to Dr. Leyla Isik.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116-126.

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral cortex*, *25*(11), 4596-4609.

Khosla, M., Murty, N. A. R., & Kanwisher, N. (2022). A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. Current Biology, 32(19), 4159-4171.

McMahon, E., Bonner, M. F., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. Current Biology, 33(23), 5035-5047.

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. Trends in Cognitive Sciences, 25(2), 100-110.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29-48).

Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings 8 (pp. 540-547). Springer Berlin Heidelberg.

Wu, J., Guo, Y., Chen, Z., Shen, M., & Gao, Z. (2024). Dual routes of chunking social interaction: Insights from grouping two agent actions in working memory. *Journal of Experimental Psychology: General*.