Semantic action representations in the mind and brain

Diana C Dima (ddima@uwo.ca)
Department of Computer Science, Western University
1151 Richmond St, London ON N6A 3K7, Canada

Jody C Culham (jculham@uwo.ca)
Department of Psychology, Western University
1151 Richmond St, London ON N6A 3K7, Canada

Yalda Mohsenzadeh (ymohsenz@uwo.ca)
Department of Computer Science, Western University
1151 Richmond St, London ON N6A 3K7, Canada

Abstract:

Understanding others' actions is an essential part of our everyday visual experience, yet the underlying computations are not well understood. Natural actions pose a complexity challenge, varying along many perceptual features. To address this, we annotated natural videos of everyday actions with a rich set of visual, social, and semantic features. In particular, we tested four models of action categorization defining actions at different levels of abstraction from specific (action verb) to broad (action target: an object, a person, or the self). We combined behavioral similarity judgments, EEG, and fMRI to investigate action representations in the mind and brain. Using variance partitioning, we found that the target of actions uniquely explained behavioral similarity judgments, as well as EEG patterns starting at 200 ms after video onset. EEG-fMRI fusion linked this processing stage to representations in the lateral occipitotemporal cortex. Together, our results show that actions are categorized primarily according to their target, and reveal the underlying spatiotemporal dynamics.

Keywords: action perception; action categorization; representational similarity analysis; EEG-fMRI fusion

Introduction

As we navigate the visual world, we rely on our understanding of others' actions to infer social information, make predictions, and decide on our course of action. How do we extract conceptual information from a wide variety of actions across different contexts? In the real world, actions are visually complex, involving interactions between people, scenes, and objects, and thus pose a challenge in disentangling the contributing features. Previous fMRI work suggests that semantic features like action goals and action categories defined at different levels of abstraction are extracted in the lateral occipitotemporal cortex (LOTC; Wurm & Caramazza, 2022; Zhuang et al., 2023). However, most previous studies relied on small, controlled stimulus sets and tested a limited range of features, making it difficult to disentangle different types of semantic information. Furthermore, few studies have investigated the underlying temporal dynamics. Action processing is thought to unfold in stages, with semantic, invariant information processed within 200 ms (Dima et al., 2022; Isik et al., 2018); yet it remains unclear what features support this invariant neural response.

To address this, we combined a set of naturalistic action videos with rich feature annotations. In particular, we tested four hypothesis-driven semantic features, defining actions at different levels of abstraction (Figure 1a): action verb (e.g. *swimming*), everyday activity (e.g. *sports*: broad categories based on the American Time Use Survey), action class (e.g. *locomotion*: behaviorally relevant categories inspired by primate research; Graziano & Aflalo, 2007); and action target (whether an action is directed towards an object, another person, or the self; Wurm et al., 2017). Using multimodal data (behavior, EEG, and fMRI), we disentangled the contributions of semantic features with variance partitioning, and characterized the underlying neural dynamics with EEG-fMRI fusion.

Methods

Stimuli. We curated 95 two-second videos from the Moments in Time dataset (Monfort et al., 2020), representing a variety of actions defined at four levels of abstraction (Figure 1a). We annotated the videos with the four semantic features (action verb, activity, class, and target), as well as perceptual features (CORnet-S deep neural network activations, motion energy, body parts involved in each action), and social features (number and gender of agents).

Behavior. Thirty-nine healthy adult participants (22 female) completed an online multiple arrangement

experiment. Participants arranged the videos according to the actions' semantic similarity, resulting in representational dissimilarity matrices (RDM) that quantified the distances between stimuli (Kriegeskorte & Mur, 2012).

EEG. Twenty participants (13 female) viewed the videos and performed a one-back action task while EEG data were recorded using a BioSemi 64-channel system. We performed time-resolved linear decoding of the videos using each participant's whole-brain EEG patterns, and used the pairwise decoding accuracies to create time-resolved neural RDMs.

fMRI. Eight participants (six female) viewed the videos and performed a one-back action task while fMRI data were collected in a rapid event-related design. Single-trial responses were extracted using *GLMsingle* (Prince et al., 2022). With a searchlight approach, we identified a left LOTC region encoding behavioral similarity, and created a participant-averaged LOTC RDM.

Analysis. We linked stimulus features to the data using representational similarity analysis (RSA). To evaluate the unique and shared contributions of semantic features to the behavioral and EEG RDMs, we applied a variance partitioning approach. Finally, we linked LOTC patterns to specific processing stages with EEG-fMRI fusion (Cichy & Oliva, 2020).

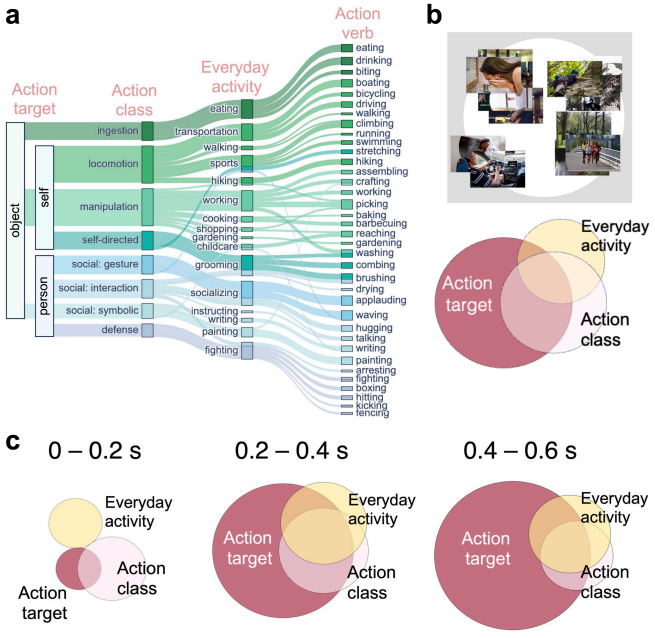

Figure 1: Semantic feature representations in the mind and brain. **a,** Actions were defined at four levels of abstraction. **b,** Semantic similarity estimated via multiple arrangement was best predicted by the target of actions. **c,** Among semantic features, the action target best predicted EEG patterns starting at ~200 ms.

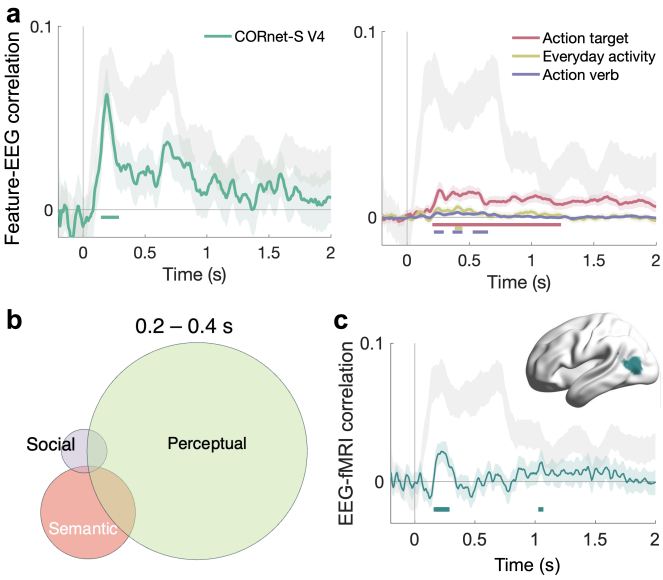

Figure 2: Neural dynamics of action representations. **a,** Average RSA correlation (ρ_A) time courses for the best-performing deep neural network layer (left) and significant semantic features (right). The noise ceiling is shown in gray, error bars (light patches) are SEM, and horizontal bars denote significance (cluster-corrected p<0.05). **b,** Variance partitioning results showing a unique contribution of semantic features after ~200 ms. **c,** EEG-fMRI (LOTC) correlation time course.

Results

Variance partitioning analyses revealed that the target of actions best explained behavioral similarity (p<0.001, randomization testing, maximum-statistic-corrected; Figure 1b) and EEG patterns starting at ~200 ms after stimulus onset (p<0.002; Figure 1c). The action class and everyday activity features contributed little unique variance. Note that similar results were obtained when including the action verb feature as a predictor.

Feature correlations with the EEG patterns revealed a temporal hierarchy from early perceptual features (~100 ms) to later semantic features (~200 ms; Figure 2a). This was confirmed in a variance partitioning analysis showing a significant unique contribution of semantic features after 200 ms (p<0.005), alongside perceptual features (p<0.001; Figure 2b).

Finally, using EEG-fMRI fusion, we found a correspondence with LOTC patterns between ~200-300 ms (Figure 2c). Together, these results highlight the actions' target as an organizing semantic feature, and provide multimodal evidence of the processing of actions after ~200 ms in LOTC.

Discussion

Using variance partitioning applied to behavioral and EEG data, we show that the actions' target explains both behavioral and neural representations better than, and independently of, other semantic features such as the action class. This extends recent findings highlighting the importance of action goals and social features in action perception (Dima et al., 2022; Tarhan et al., 2021). Furthermore, this finding suggests that representations of action category in the brain may reflect a broader goal-based organization.

Time-resolved analyses revealed that this semantic representation emerges in the EEG patterns after 200 ms, independently of other perceptual, social, and action features, including essential features like the body parts involved in actions. Using EEG-fMRI fusion, we linked this processing stage (200-300 ms) to patterns in LOTC. Together, our results elucidate the nature of semantic representations extracted during action observation, and show how they unfold over time in the brain. These insights suggest that socially-relevant, high-level goals might be key to human event recognition and its successful replication in computer vision models.

Acknowledgments

This work was supported through the Canada First Research Excellence Fund, a Western Interdisciplinary Development Initiatives Grant, a Vector Institute for Artificial Intelligence Research Grant, and a Western Postdoctoral Fellowship

References

Cichy, R. M., & Oliva, A. (2020). A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. *Neuron, 107*(5), 772–781.

Dima, D. C., Tomita, T. M., Honey, C. J., & Isik, L. (2022). Social-affective features drive human representations of observed actions. *eLife*, *11*.

Graziano, M. S. A., & Aflalo, T. N. (2007). Mapping behavioral repertoire onto the cortex. *Neuron 56*(2), 239-251.

Isik, L., Tacchetti, A., & Poggio, T. (2018). A fast, invariant representation for human action in the visual system. *Journal of Neurophysiology*, *119*(2), 631-640.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements, *Frontiers in Psychology, 3*.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., & Oliva, A. (2020). Moments in Time Dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 502-508.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022) Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife, 11*.

Tarhan, L., De Freitas, J., & Konkle, T. (2021). Behavioral and neural representations en route to intuitive action understanding. *Neuropsychologia, 163.*

Wurm, M., & Caramazza, A. (2022). Two 'what' pathways for action and object recognition. *Trends in Cognitive Sciences, 26*(2), 103-116.

Wurm, M., Caramazza, A., & Lingnau, A. (2017). Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *Journal of Neuroscience, 37*(3), 562-575.

Zhuang, T., Kabulska, Z., & Lingnau, A. (2023). The representation of observed actions at the subordinate, basic, and superordinate level. *Journal of Neuroscience, 43*(48), 8219-8230.