High-Level Features Organize Perceived Action Similarities

Leyla Tarhan (ltarhan@g.harvard.edu)
Department of Psychology, Harvard University, 33 Kirkland St.
Cambridge, MA 02138 USA

Talia Konkle (tkonkle@fas.harvard.edu)
Department of Psychology, Harvard University, 33 Kirkland St.
Cambridge, MA 02138 USA

Abstract:
Other people's actions fill our visual worlds – we watch others run, dance, cook, and laugh on a daily basis. Together, these make up a repertoire of visual actions that we can recognize and reason about. How is this repertoire organized in our minds, so that some actions appear more similar than others? To answer this, we measured the perceived similarity among a large set of everyday actions. We then used a modeling framework to explore which kinds of features predict that similarity. We found that the mental action similarity space is organized primarily by relatively high-level features relating to semantic category and body part involvement. Further, neural similarity within regions that tile the visuo-motor cortex does not predict these judgments well, suggesting that they do not directly support this higher-level space. These results echo recent findings that human similarity judgments in the object and scene domains are best predicted by high-level feature spaces not grounded in the ventral visual stream (Groen et al., 2017; Jozwik et al., 2017), a pattern that has now been observed across three domains of vision and may reflect a broader principle of the perceptual system.

Keywords: action perception; representational similarity

Introduction

Watching other people's actions is a hallmark of our visual experience. Among the many ways we see others move around the world, some actions are intuitively more similar than others. What causes us to see running and walking as related, but cooking as different than both? Does the answer lie simply in perceptual similarities, or do higher-level features like semantic content and inferred mental states also play a role?

The structure of the action similarity space has been studied before under both behavioral and neural approaches. For example, Watson and Buxbaum (2014) found that tools are naturally sorted into groups that reflect kinematic aspects of how they are used. In the neural domain, recent work suggests that some of these representational spaces are organized by sociality and transitivity in the lateral temporal lobe (Wurm, Caramazza & Lingnau, 2017), semantic content in the ventral visual stream (Huth et al., 2016), and kinematics in the lateral temporal and motor cortices (Kemmerer et al., 2008; Lingnau & Downing, 2015).

However, a central challenge for characterizing the visual action repertoire is balancing stimulus control with diversity. Most studies that examine action representations focus on a small subset of highly-controlled actions. This allows them to test highly specific hypotheses, but limits their conclusions to a small corner of our visual experience. In addition, few studies have investigated how we naturally perceive action similarity and what kind of a similarity space we draw on during that perception. As such, this study has two main aims. First, what are the major organizing dimensions that structure the mental similarity space among a wide range of everyday actions? Second, which neural regions house this representational space?

Materials and Method

Stimuli

120 short (2.5 s.) action videos were selected based on what a large sample of Americans reported performing on a daily basis in the American Time Use Survey (U.S. Bureau of Labor Statistics, 2014) (Figure 1). These actions span several broad categories such as personal care, eating and

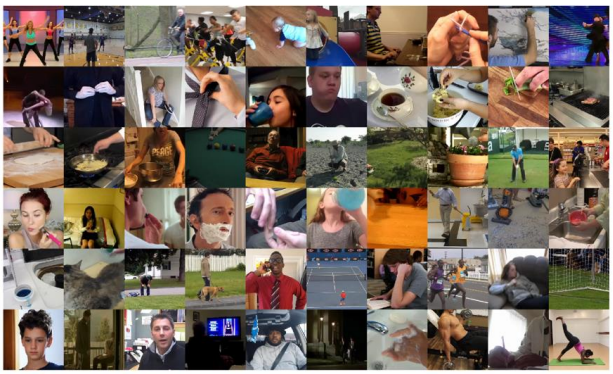

Figure 1. The 60 everyday actions portrayed in the stimulus set (key frames from one of two video sets).

drinking, socializing, and exercise. The videos were divided into two sets (test and validation) depicting the same 60 actions.

Behavioral Paradigm

To measure perceived similarities among these actions, 36 human participants (Set 1: 20, Set 2: 16) completed an unguided action similarity task. Participants arranged 60 action videos (1 set) within a circular arena so that their distances reflected their perceived similarity (following the paradigm developed by Kriegeskorte & Mur, 2012; Figure 2). Across trials, we used an inverse multidimensional scaling method to obtain pairwise distances among all 60 videos, which reflect perceived dissimilarities.

Hypothesized Feature Spaces

What features might determine whether we see walking and running as similar, but cooking as different? Are they low- or high-level, and what kinds of information do they carry? We investigated a broad range to answer these questions.

We considered four low- to high-level feature spaces that might organize the mental action similarity space (Figure 3a). Although we didn't expect that action similarity judgments are based on visual statistics, we included gist features as a low-level control (Oliva & Torralba, 2001). Gist features capture the visual statistics present in the videos, and in prior work we have found that they predict a large extent of the action responses in early visual regions (Tarhan & Konkle, CCN 2017). Next, we considered "mid-level" features: the body parts involved in performing each action and what the actions were directed at (e.g., an object or a person). Similar "means" and "ends" features have been hypothesized to organize action representations in the lateral

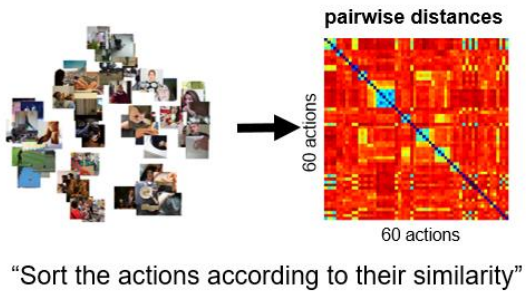

Figure 2. Behavioral Paradigm. Across many iterations, subjects arranged still frames from 60 action videos so that the distances among all stills reflected the videos' perceived similarities (method adapted from Kriegeskorte & Mur, 2012).

occipital cortex, and are therefore good candidates for properties that also organize the mind (Lingnau & Downing, 2015). Finally, we captured the actions' high-level semantic categories using the super- and sub-ordinate labels provided by the American Time Use Survey, aligning with similar approaches used in modeling the similarity structure of scene categories (Greene et al., 2016).

Finally, we considered a neural feature space based on neural responses to the action videos collected in a previous fMRI experiment (N = 13). These responses were used to segment the visuo-motor cortex into 13 functionally distinct networks, using k-means clustering. Only responses in voxels that responded reliably (split-half reliability > 0.3) were included in the analysis.

For modeling purposes, pairwise similarity among the actions was calculated along each dimension in each feature space using squared Euclidean distance.

Modeling Approach

To predict behavioral action similarity using these different possible feature spaces, we used a predictive modeling approach. Specifically, for each feature space, we used a weighted combination of similarities along each dimension within a feature space (for example, hands, legs, etc. for the body part involvement feature space) to predict action similarity judgments. This was done using a leave-1-condition-out cross-validation method to predict similarity judgments (averaged across subjects), and these predictions were then correlated with each subject's actual judgments using Kendall's tau-a.

Results

Model performance is plotted in Figure 3b. All feature spaces performed better than chance, with the exception of the visual gist features (two-sided Wilcoxon signed-rank test p<0.05), with the highest predictions from both semantic category and body part involvement models (mean leave-1-out τ_A: body parts = 0.15, category = 0.14, action target = 0.09, gist = -0.01) Combining these perceptual features into a single model further boosted performance (mean leave-1-out τ_A = 0.22), suggesting that a combination of high- and mid-level features best captures the structure of the mental action similarity space. In contrast, similarity in action responses within visual cortex regions did not predict similarity judgments well (τ_A = 0.08). This pattern of results held across both stimulus sets.

Conclusions

In the current study, we found that similarity judgments for everyday actions are best predicted by higher-level features that capture what they are for, rather than how they look. In addition, the visual system does not predict this mental similarity space well, raising the possibility that these judgments are not directly represented within the visual cortex.

When interpreting this neural result, it is necessary to bear two caveats in mind. First, this finding does not preclude all

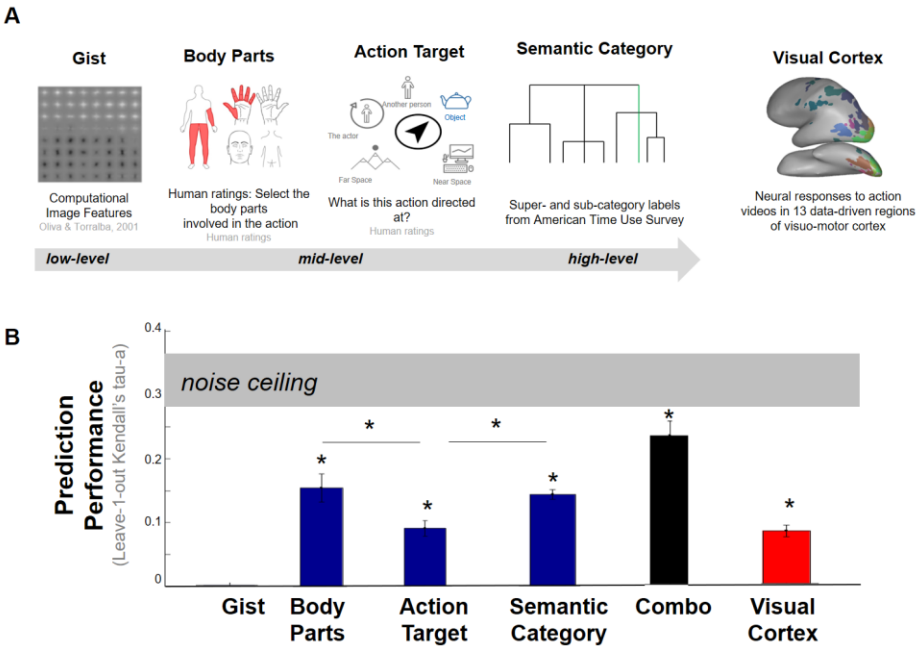

Figure 3. Predictive Modeling Results. (a) Hypothesized feature spaces used to predict behavioral similarity judgments. (b) Prediction performance for each hypothesized feature space. Asterisks denote significant Wilcoxon signed-rank tests ($p < 0.05$).

regions of the visual cortex. Because the analysis only included voxels that respond reliably (and therefore with some variance across videos), we lacked coverage in some visual regions, such as superior temporal sulcus, that have been implicated in high-level and social reasoning. Second, more work is needed to determine how well this neural model can predict other kinds of action similarity, and therefore what features form the major organizing dimensions of its representations.

Overall, these findings echo recent work showing that judgments of object and scene similarity rely on high-level feature spaces outside of the ventral visual stream (Groen et al., 2017; Jozwik et al., 2017). Given such similar results across these visual domains, it is possible that most mental similarity spaces for visual items are organized based on higher-level features, rather than those strictly governing visual appearance. However, future work is required to determine how well which this principle generalizes across different mental similarity spaces (i.e., measured with different tasks), and to clarify the distinction between high- and low-level features.

Are action similarities categorical?

The behavioral action similarity spaces in the current study had a relatively low noise ceiling range (r=0.29 – 0.36). Although these values are typical for studies using this multi-arrangement paradigm (Jozwik, Kriegeskorte & Mur, 2016; Jozwik, Kriegeskorte, Storrs & Mur, 2017), they raise the interesting possibility that continuous distances may not be the best way to characterize the underlying representations. One alternative is that that action similarity is represented in a more categorical manner.

To investigate this possibility, we are developing a grouping paradigm. In a pilot study, participants viewed each action video and then divided them into discrete groups. The number of groups was not constrained. The number of groups formed different dramatically across participants (range: 3 – 14). However, the same items fell into the same groups across participants (average pairwise d': Set 1 = 0.78, Set 2 = 0.67). This preliminary evidence suggests that when action similarity is measured via discrete grouping, inter-subject reliability is higher than when it is measured via continuous distances. It is therefore possible that the mental action similarity space is categorical rather than continuous, and is thus better captured using categorical paradigms.

Acknowledgments

Funding for this project was provided by NIH grant S10OD020039 to Harvard University Center for Brain Science, NSF grant DGE1144152 to L.T., and the Star Family Challenge Grant to T.K.

References

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology:*

General, *145*(1), 82.

Groen, I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2017). Distinct contributions of functional and deep neural network features to scene representation in brain and behavior. *bioRxiv*, 207530.

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., & Gallant, J. L. (2016). Decoding the semantic content of natural movies from human brain activity. *Frontiers in systems neuroscience*, *10*, 81.

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, *83*, 201-226.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in psychology*, *8*, 1726.

Kemmerer, D., Castillo, J. G., Talavage, T., Patterson, S., & Wiley, C. (2008). Neuroanatomical distribution of five semantic components of verbs: evidence from fMRI. *Brain and language*, *107*(1), 16-43.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in psychology*, *3*, 245.

Lingnau, A., & Downing, P. E. (2015). The lateral occipitotemporal cortex in action. *Trends in cognitive sciences*, *19*(5), 268-277.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, *42*(3), 145-175.

Watson, C. E., & Buxbaum, L. J. (2014). Uncovering the architecture of action semantics. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1832.

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