Modeling the Neural Structure Underlying Human Action Perception

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

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

Abstract:

When humans view others' movements, large swathes of the visual cortex are activated (Kilner, 2011), but the major representational divisions organizing this neural activity are not well understood. To explore this architecture, 13 observers underwent functional neuroimaging while observing 120 2.5s videos depicting everyday actions. Using voxel-wise modeling (Mitchell et al., 2008), we found that a variety of encoding models—with features for the body parts involved in the actions, what the actions were directed at (e.g., object, person, space), and the visual image features present—all successfully predicted visual cortex responses to individual actions (leave-2-out accuracies: 43-79%). Prediction accuracy for these models varied across the cortex, revealing divisions of ventral and dorsal streams with different underlying action representations. We also used data-driven clustering to discover natural parcellations in visual cortex based on voxels' response profiles (Lashkari et al., 2010), providing a convergent approach to the dataset. These analyses reveal several meaningful functional divisions within the regions involved in action perception, including two networks linking ventral and dorsal stream representations, and begin to formalize the neural representational structure underlying our visual understanding of everyday actions.

Keywords: action perception; action representation; visual cortex; encoding models

Introduction

Humans can recognize the meaning behind a diverse set of actions – from eating and hammering to dancing and cooking – as well as parse the crucial difference between a person running for exercise and one running from a bear. What features does the visual system use to process others' actions, and how is this feature-processing organized across the visual cortex?

These questions have gone largely unanswered, in part because most research on neural action representations has studied a small subset of human visual experience (hand- and tool-based actions). The current study leverages functional neuroimaging (fMRI), a rich video stimulus set, and a variety of analytic approaches to gain better traction on these questions.

Materials and Methods

Stimuli

120 short (2.5s) video clips were used to depict 60 everyday activities, sampled from the American Time Use Survey. Videos were resized to a 512x512px frame and presented at 30 frames per second using MATLAB.

Feature Dimensions

To collect ratings for each activity video along several hypothesized organizing dimensions, Amazon Mechanical Turk workers provided the following ratings for the action depicted in each video: the body parts involved, the target of the action (e.g., directed at an object or a person), and the amount of effort required. In addition, to collect low-level image features, each video frame was passed through a gist model (Oliva & Torralba, 2001), which calculates the spatial distributions of low-level image statistics.

fMRI Procedure

13 human observers completed an event-related fMRI experiment. During the main task, participants watched the activity videos and detected an occasional red frame to maintain attention. Each video appeared four times across eight functional runs.

Results

Neural Encoding Modeling

Analysis Voxel-wise encoding models were used to model each voxel's responses to individual videos based on each of the feature spaces listed above. Model fits were cross-validated using a leave-two-out procedure: each voxel's response pattern was fit using data from 118 videos, then predicted and actual responses to the two held-out videos were correlated to yield a single *r*-value for each voxel. Only voxels with a split-half reliability ≥ 0.3 were included in the analysis.

Results A considerable portion of the occipital, temporal, and parietal cortices was moderately well fit by all models (Figure 1). Several patterns were evident in the data. First, the model based on low-level features (gist) out-performed the three higher-level models in early visual cortex, while the higher-level models outperformed the gist model in the temporal and parietal cortices. Second, some models showed complementary patterns within their cross-validation performance: specifically, the effort model succeeds almost exclusively in the lateral occipito-temporal cortex, an area where the target model predicted poorly. These results suggest that actions are represented with different underlying feature spaces across different sub-regions of the visual cortex.

Response Profile Clustering

Analysis To discover functional divisions within the visual cortex in a data-driven manner, we used a response profile clustering method (e.g., Lashkari, Vul, Kanwisher & Golland, 2010). In this analysis, K-means clustering was used to group voxels based on their overall response to the set of videos.

Results Figure 2 shows the response profile clustering solution with *k*=7 clusters. We found two networks of regions (dark blue and purple), in the parietal and lateral temporal cortices with response profiles related to activities' tool-relevance (consistent with a known "tool network"; Johnson-Frey, 2004). Additional networks parcellate early visual and temporal areas, and may relate to other activity features, such as scene-relevance (dark pink). This solution remained stable when the data were split into two sets, and supports the existence of several meaningful functional networks across the ventral and dorsal streams that are involved in activity perception.

Conclusions

Through a novel use of encoding models, we have shown that it is possible to model neural responses to actions using both low-level image statistics and high-level features such as body part involvement, action target, and effort. Further, the predictive ability of these feature spaces varies across the cortex: low-level image features are prominent in early visual areas; but, as information is passed forward to the parietal and temporal cortices, higher-level features have more explanatory value. Finally, we found evidence for several functional networks that respond differently to the action videos and that connect regions across ventral and dorsal streams. These findings take significant steps toward understanding action perception in the brain.

Figures

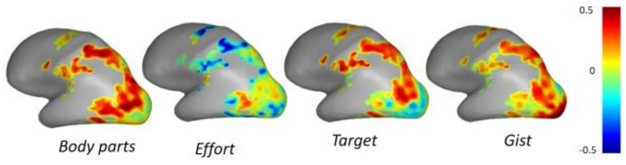

Figure 1: Voxel-wise leave-2-out cross-validation results for all models. The correlation between predicted and actual neural response to the held-out videos is plotted in each voxel on an example subject's left hemisphere.

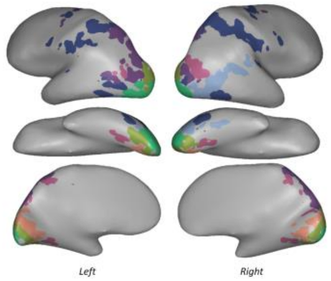

Figure 2: Response Profile Clustering results with seven clusters, shown on both hemispheres of an example subject's brain.

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

Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends in cognitive sciences*, *8*(2), 71-78.Lashkari, D., Vul, E., Kanwisher, N., & Golland, P. (2010). Discovering structure in the space of fMRI selectivity profiles. *Neuroimage*, *50*(3), 1085-1098.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191-1195.

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.