Exploratory visual and social relationship models correlate with neural responses to face identity during naturalistic viewing

Junaid S Merchant*, Shawn A Rhoads*, Oliver Xie*, and Sarah Dziura*
*Equal contribution

Introduction

Naturalistic stimuli such as movies or television shows elicit rich, complex, and ecologically valid responses akin to everyday experiences (Hasson, Malach, & Heeger, 2010). Due to the unconstrained nature of the stimuli, there are myriad ways that the resulting data can be analyzed, as well as a number of limitations that arise due to difficulties in controlling for unintended effects (Grall & Finn, 2022). Previous work has examined the neural representation of person perception in both the spatial and temporal domains (Ambrus et al., 2019; Anzellotti et al., 2014; Dobs et al., 2019; Dziura & Thompson, 2020). Convolutional Neural Networks (CNNs) such as AlexNet (Krizhevsky, Sutskever, & Hinton, 2012; Russakovsky et al., 2015) have been previously demonstrated to reflect different layers of the hierarchical visual system using static image stimuli (Güçlü & van Gerven, 2015; Horikawa, & Kamitani, 2017). However, many of these studies used identical images to create visual system target representations and to elicit neural responses. Determining whether such representations are found in the visual system when viewing more unconstrained and dynamic information is an ongoing effort. One recent study moved toward a more naturalistic direction by eliciting neural data from videos of people, although these were only limited to faces and the target representations were created using still images taken from the videos themselves (Tsantani et al., 2021). Additionally, externally generated representations of relationships among people are observed in neural patterns from controlled experimental paradigms (Dziura & Thompson, 2020; Parkinson, Kleinbaum, & Wheatley, 2017). In the present study, we investigated whether patterns of neural responses to faces of characters in a 45 minute-long dynamic film involving multiple people (viewed in a variety of orientations, lighting, movement, and clothing) correspond to representations of low-level visual features (e.g., visual similarity) and higher-level social features (e.g., social relationships) created from independent sources.

Methods

Rekognition Celebrity Face Detection. The pilot episode of the TV series Friday Night Lights was fed through Amazon Rekognition, an image and video analysis tool, that recognizes celebrity faces and returns on-screen timecodes with a high degree of accuracy (https://aws.amazon.com/rekognition/). Events of interest were limited to the top 5 most frequently occurring characters: Coach Eric Taylor, Jason Street, Matt Saracen, Brian "Smash" Williams, and Tim Riggins. After extracting the timestamps of these 5 main characters, we excluded the events when multiple characters were detected. We then estimated the duration of characters' appearance for each event through these independent onset timestamps. Thus, we used 297 events in total: 142 events corresponding to Coach Taylor, 48 events corresponding to Jason Street, 39 events corresponding to Matt Saracen, 34 events corresponding to Smash Williams, and 34 events corresponding to Tim Riggins.

Least-Squares-Sum Event-wise Activation Estimation. Unsmoothed, preprocessed data from 34 subjects was downloaded from OpenNeuro (https://openneuro.org/datasets/ds003521/versions/2.1.0) (see Chang et al., 2021 for methods). To estimate trial-wise BOLD activity to the five characters' faces, we performed temporal data reduction using Least Squares-Sum (LSS) estimation (Mumford et al., 2012) using 3dLSS in AFNI (Cox, 1996). This approach runs a separate GLM for each event, where the trial is modeled as the regressor of interest and all other events are combined into a single nuisance regressor, and the event duration was also included in the model. Thus, for each participant, we estimated 297 different activation maps that corresponded to each time the face of a character of interest was displayed across the episode. Following best practices, we subtracted ½ TR (one second) from event onsets to align default parameters between fMRIprep and AFNI modeling. We also included five nuisance regressors that controlled for all faces detected by Rekognition throughout the episode, face size (i.e., area of rectangle as proportion of total area of screen), face brightness, face sharpness, and confidence of the Rekognition detection outputs.

Visual Similarity Model from AlexNet. We collected promotional images of the five characters used in the LSS analysis from a google image search. These images were cropped to 227 x 227 px sze with a 150 x 200 px oval surrounding the face and a gray background (RGB 195 195 195). We fed the images into a pretrained Convolutional Neural Network (AlexNet: Russakovsky et al., 2015) and extracted the output of two convolutional layers ("conv1" and "conv5"). We then calculated the Euclidean distance between the feature outputs of each convolutional layer for each pair of images. These pairwise distance vectors were used as target dissimilarity matrices in further analyses.

Social Relationship Model from Wikipedia. Using text data from each character's Wikipedia page, we generated a model of social associations. This approach used the number of times one character was referenced in another character's Wikipedia page (and vice-versa) in a pairwise fashion as a proxy for the characters' social relationships. These counts were then normalized to represent the number of references as a proportion of the maximum number of references. Directed values (i.e., Jason Street's page referencing Tim Riggins, Tim Riggins' page referencing Jason Street) were averaged across pairs of characters to yield a symmetric, weighted dissimilarity matrix approximating the strength of the social relationships.

Representational Similarity Analysis. We used the AlexNet and the Social Relationship representational dissimilarity matrices (RDMs) as target models for neural response similarity between characters using an exploratory searchlight approach implemented using a modified version of the CoSMoMVPA MATLAB Toolbox (Oosterhof, Connolly, & Haxby, 2016). Neural RDMs were derived for each 100-voxel searchlight using the unsmoothed, subject-level beta-maps corresponding with the 297 character events, and calculated as unbiased, cross-validated Euclidean distances between characters (i.e., Willet et al., 2020; https://github.com/fwillett/cvVectorStats). Similarity between lower-triangles of the neural and each model RDM was assigned to the center voxel of each searchlight, yielding whole-brain maps for each model consisting of voxel-level similarity values (quantified as Spearman's Rho to Z). Group-level statistical inference was conducted with a non-parametric one-sample t-test with 10,000 permutations using AFNI's 3dttest++ (Cox, 1996). Resulting maps were thresholded using cluster-correction with the -Clustsim option.

Results

Visual Similarity Model. Face similarity estimates from convolutional layer 1 of AlexNet revealed wide-spread correlations throughout the brain extending across ventral and dorsal stream visual processing, mentalizing/social cognition networks, & affective and reward systems. We did not observe any significant associations in the primary visual cortex, although this may be due to regressing out general visual features (i.e., brightness, sharpness, and presence of faces) in the LSS model. Face similarity estimates from convolutional Layer 5 of AlexNet revealed correlations within clusters of the dorsomedial prefrontal cortex (dmPFC), dorsolateral prefrontal cortex (dlPFC), and posterior cingulate extending into precuneus. Together, these results suggest that simple visual features of characters' faces contribute to how numerous higher-order neurocognitive systems represent individuals' identities.

Social Relationship Model. The social relationship model revealed that relationship strength is related to neural similarity in regions associated with social cognition, salience and reward processing, including clusters in the dmPFC and precuneus that overlap with the results from AlexNet's layer 5 output, as well as bilateral superior temporal sulcus (STS), left temporoparietal junction (TPJ), left insular cortex, and the midbrain/periaqueductal gray (PAG). Results highlight the roles of these higher-order social-cognitive regions representing interpersonal relationships.

Conclusions and Future Directions

In sum, we successfully demonstrated one way in which neural response data from long naturalistic viewing sessions can be segmented and compared to models of different types of information content. We found that both visual and social relationship information was represented in regions of the brain linked to social perception and cognition during naturalistic viewing. This is consistent with previous findings demonstrating that neural responses to static and more limited video face stimuli index different types of similarity structure, including visual and non-visual social information (Tsantani et al., 2021). More work can be done to test and validate the results discussed here, such as partial correlation analyses with multiple models at once, and examining responses to individual events independently. However, this is an important first step toward understanding the neural signatures underlying face identities and the social information they hold (such as the relationships they have with each other) in a naturalistic setting. With similar methodologies, further directions could involve hypothesis-driven examinations of higher order representations of person characteristics, social roles, and relationships with each other, including how all of these develop while people learn new information about individuals. For example, the social relationship RDM was a coarse measure taken from Wikipedia pages of the characters throughout the entire series, rather than from the pilot episode. Future work could probe perceived relationships through subject-generated data during the videos themselves, which would allow for examining how the brain tracks changes in relationships over time.

References

Ambrus, G. G., Kaiser, D., Cichy, R. M., & Kovács, G. (2019). The neural dynamics of familiar face recognition. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhz010

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2014). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, *24*(8), 1988–1995. https://doi.org/10.1093/cercor/bht046

Chang, L. J., Jolly, E., Cheong, J. H., Rapuano, K., Greenstein, N., Chen, P.-H. A., & Manning, J. R. (2021). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Science Advances*, *7*(17), eabf7129. https://doi.org/10.1126/sciadv.abf7129

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, *10*(1), 1258. https://doi.org/10.1038/s41467-019-09239-1

Dziura, S. L., & Thompson, J. C. (2020). Temporal dynamics of the neural representation of social relationships. *The Journal of Neuroscience*, *40*(47), 9078–9087. https://doi.org/10.1523/JNEUROSCI.2818-19.2020

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Ayse, I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., Gorgolewski, K. J., Isik, A. I., Erramuzpe Aliaga, A., Kent, J. D., … Gorgolewski, K. J.

(2018). FMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. https://doi.org/10.1101/306951

Güçlü, U. & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience, 35,* 10005–10014.

Hasson, U., Malach, R., & Heeger, D.J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences, 14,* 40. doi:10.1016/j.tics.2009.10.011.

Horikawa, T. & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications, 8,* 1–15.

Krizhevsky, A. Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60,* 84-90.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Oosterhof, N.N., Connolly, A.C., & Haxby, J.V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics, 10,* PMC4956688. doi: 10.3389/fninf.2016.00027.

Parkinson C., Kleinbaum, A.M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behavior, 1,* 0072.

Russakovsky, O., Deng, J., Su, H., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*, 211–252.

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA encode distinct types of face identity information. *The Journal of Neuroscience*, *41*(9), 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020