From Pixels to Scene Categories: Unique and Early Contributions of Functional and Visual Features

Michelle R. Greene (mgreene2@bates.edu)
Bates College Neuroscience Program, 3 Andrews Road
Lewiston, ME 04240 USA

Bruce C. Hansen (bchansen@colgate.edu)
Colgate University Department of Psychological & Brain Sciences, Neuroscience Program, 13 Oak Drive
Hamilton, NY 13346 USA

Abstract:

Human scene categorization is rapid and robust, but we have little understanding of how individual features contribute to categorization, nor the time scale of their contribution. This issue is compounded by the non-independence of the many candidate features. Here, we used singular value decomposition to orthogonalize 11 different scene descriptors that included both visual and semantic features. Using high-density EEG and regression analyses, we observed that most explained variability was carried by a late layer of a deep convolutional neural network, as well as a model of a scene's functions given by the American Time Use Survey. Furthermore, features that explained more variance also tended to explain earlier variance. These results extend previous large-scale behavioral results showing the importance of functional features for scene categorization. Furthermore, these results fail to support models of visual perception that are encapsulated from higher-level cognitive attributes.

Keywords: scene categorization, EEG, encoding, RSA

Introduction

Human scene understanding is remarkable for its speed (Greene & Oliva, 2009; Potter, Wyble, Hagmann, & McCourt, 2014), as well as its robustness in the face of limited information, such as in the far periphery (Boucart, Moroni, Thibaut, Szaffarczyk, & Greene, 2013), or with limited spatial resolution (Torralba, 2009). Understanding the representations and transformations that enable categorization is thus a fundamental goal for computational cognitive neuroscience.

Scene categories can be differentiated on the basis of many types of features, ranging from low-level visual properties such as color (Oliva & Schyns, 2000), texture (Renninger & Malik, 2004), or contour junctions (Choo & Walther, 2016), to high-level properties such as conceptual attributes (Patterson & Hays, 2012) and affordances (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016). Despite these results, we do not yet understand how each feature type contributes to the neural processing of scene category over time.

Assessing the relative contributions of low- and high-level visual features has been challenging (Groen, Silson, & Baker, 2017), primarily because these features are not independent. Recent work has used variance partitioning techniques to assess the relative contributions of information sources (Greene et al., 2016; Groen et al., 2018; Lescroart, Stansbury, & Gallant, 2015), but these methods are most interpretable when only a handful of features are considered.

In this work, we used both optimized image selection and orthogonal feature transformation in order to examine the independent contributions of eleven different visual models to the microgenesis of visual scene categorization. Our results show that both visual features and functional features primarily contribute to early image-evoked activity.

Methods

Stimulus Selection

Participants (N=13) viewed 2250 color images from 30 scene categories across two EEG recording sessions. Scene categories were chosen to maximize differences in representational dissimilarity matrices (RDMs) across three types of features: a late layer (FC7) of a pre-trained AlexNet deep neural network (DNN, (Krizhevsky, Sutskever, & Hinton, 2012), a bag-of-objects model (Lazebnik, Schmid, & Ponce, 2006), and a model of the scene's functions / affordances (Greene et al., 2016). Our iterative selection procedure was inspired by the odds algorithm of (Bruss, 2000). In each of 10,000 iterations, we created a set of 30 scene categories from the SUN database (Xiao, Ehinger, Hays, Torralba, & Oliva, 2014) that had equal representation across indoor, urban, and natural environments, and recorded the Spearman's rho

correlation between function, object, and DNN RDMs. We continued sampling scene category sets until we observed a set with lower inter-feature correlations than had been observed in the initial 10,000.

Experimental Procedure

Participants viewed scenes (20° visual angle) one at a time for 750 ms each, and engaged in a three-alternative forced choice (3AFC) task following each trial. Each trial began with a 500 ms fixation point followed by a variable duration blank screen followed by the image. Continuous high-density EEGs were recorded using EGI's Geodesic EEG acquisition system. A full description of the recoding details and pre-processing procedures is found in (Greene & Hansen, 2017).

Encoding Models

We employed eleven different encoding models of both visual and semantic features. For all features except lexical distance, representational dissimilarity matrices (RDMs) were created by computing the distance between each category pair in the feature space, using the 1-Spearman rho distance metric.

DNN features We extracted activations from two layers (Conv2 and FC6) of a pre-trained DNN. The network used the AlexNet architecture (Krizhevsky et al., 2012), and was trained on the Places database (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). These layers were chosen to reflect a lower-level and higher-level layer respectively.

GIST The spatial envelope descriptor (Oliva & Torralba, 2001) was computed for each image using 3 scales, 4-8 orientations per scale, and 64 spatial blocks for a total feature vector of length 1152.

Color We transformed each RGB image to CIE L*a*b color space, and for each image created a two-dimensional histogram from the a* and b* channels in 50 bins per channel.

Wavelets To encode structural features, we passed scenes through a Gabor filter feature bank of 3 spatial scales, 4 orientations, and two quadrature phases. Weights for each of the 1328 Gabors were obtained with ridge regression.

Texture Texture features came from (Portilla & Simoncelli, 2000), and consisted of 6495 features from four statistic types: marginal statistics of pixels, wavelet coefficient correlations, wavelet magnitude correlations, and cross-scale phase statistics.

Tiny images To serve as a baseline, images were downsampled to 32 by 32 pixels, and RDMs were created by pixel distance.

Functions Each image was rated by observers on Amazon's Mechanical Turk according to each of 227 actions from the American Time Use Survey. The resulting feature vectors consisted of the proportion of scenes in a category affording each of the actions.

Objects All objects and regions were hand-labeled using the LabelMe tool (Russell, Torralba, Murphy, & Freeman, 2008), resulting in 3,563 unique region labels. The final feature vector consisted of the proportion of scene images containing each of the labels.

Lexical We computed the lexical distance between each pair of scene category names, operationalized as the shortest path between entries in WordNet (Miller, 1995).

Attributes We included the category-averaged attribute descriptions of (Patterson & Hays, 2012) that represent attributes of objects, materials, layout, and affordances.

As shown in Figure 1, there were substantial correlations between RDMs of all features. Therefore, we used singular value decomposition to create an orthonormal feature basis that expresses the unique contributions made by each feature space relative to one another.

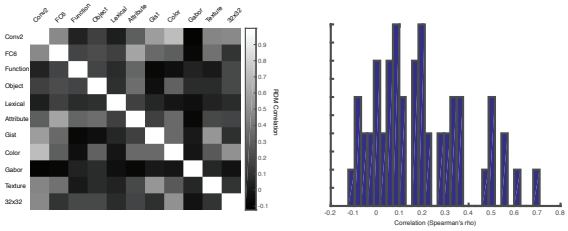

Figure 1: Correlations between all features (left), and histogram of feature correlations (right).

Time-Resolved Encoding Analysis

For each of the 11 models, we created 30-category by 30-category RDMs from the 30-category by N-feature matrices using the 1-correlation metric. All RDMs were combined in a 435-pair by 11-model matrix. Singular value decomposition was used to create a new

orthonormal basis from this set, which was used in regression analysis.

For each participant and each electrode, we extracted ERP signals within a 40 ms sliding window beginning 100 ms before stimulus presentation, and extending through the 750 ms scene duration. For each window, we created a 30x30 RDM as above. In separate regression analysis, we predicted this neural RDM from each of the orthogonalized feature predictors. Model fit was assessed with adjusted R^2.

Results

Overall

Figure 2 shows the explained variability of all 11 features included together. These variables explain significant ERP variance starting 36 ms after stimulus onset, and show two distinct peaks at 93 ms and 157 ms post-stimulus respectively.

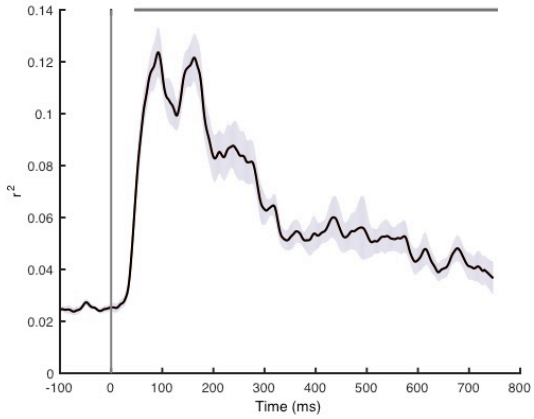

Figure 2: Explained variability for all features, averaged across electrodes. Error surface represents +/- 1 SEM.

Individual Models

The contribution from individual models is shown in Figure 3. Overall, each model started explaining ERP variability early, ranging from 26 ms post-scene onset (FC6 DNN features) to 40 ms post-stimulus for texture features. Although all models explained some ERP variability, the FC6 DNN features and the functional features, explained twice as much variance as any other model. Each model was a better predictor of ERP activity than later. Peak explained variability ranged from 80 ms post-stimulus for gist and 32x32 tiny image features, to 177 ms for the Wavelet features. Although we did not observe a significant correlation between the maximum explained variability and the peak where that maximum occurred (r=0.04, p=0.90), nor the onset of explained variability and the time of peak (r=0.18, p=.60), we did observe a striking correlation between the onset and the peak (r=-0.88, p=0.0003), demonstrating that models that had earlier onsets also explained more variability overall.

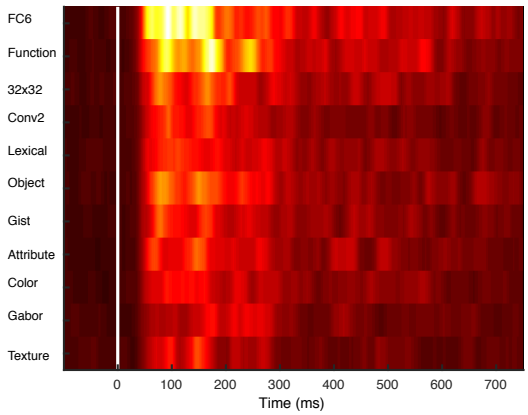

Figure 3: Explained variability over time for each of 11 features, ordered by onset of explained variability.

Discussion

Both visual and conceptual features could explain early image-evoked EEG activity, and surprisingly neither feature type seemed to be advantaged over the other. These results corroborate the findings of (Greene et al., 2016) who demonstrated that functional features explained most of the variability in scene categorization behavior. While (Groen et al., 2018) replicated this behavioral result, it was observed that most variability in scene-selective brain regions was driven by visual features rather than functions. Our results highlight the importance of functional and visual features in explaining early neural activity, which may have been missed at the time scale of fMRI.

Acknowledgments

This work was funded by National Science Foundation (1736274) to MRG and BCH, and James S McDonnell Foundation (220020439) to BCH.

References

Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., & Greene, M. (2013). Scene categorization at large visual eccentricities. *Vision Research*, *86*, 35–42.

Bruss, F. T. (2000). Sum the Odds to One and Stop. *The Annals of Probability*, *28*(3), 1384–1391.

Choo, H., & Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories

in high-level human visual cortex. *NeuroImage*, *135*, 32–44.

Fei-Fei, L., & Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02* (pp. 524–531). IEEE Computer Society.

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), 82–94.

Greene, M. R., & Hansen, B. C. (2017). Shared Spatiotemporal Category Representations in Biological and Artificial Deep Neural Networks. *BioRxiv*, 225607.

Greene, M. R., & Oliva, A. (2009). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science*, *20*, 464–472.

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Phil. Trans. R. Soc. B*, *372*(1714), 20160102.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *ELife*, *7*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc.

Lescroart, M. D., Stansbury, D. E., & Gallant, J. L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Frontiers in Computational Neuroscience*, *9*, 135.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM*, *38*(11), 39–41.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*, 176–210.

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.

Patterson, G., & Hays, J. (2012). SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, 1–10.

Renninger, L., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, *77*(1), 157–173.

Torralba, A. (2009). How Many Pixels Make an Image? *Visual Neuroscience*, *26*(01), 123–131.

Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., & Oliva, A. (2014). SUN Database: Exploring a Large Collection of Scene Categories. *International Journal of Computer Vision*, 1–20.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*(99), 1–1.