How Does the Brain Represent Visual Scenes? A Neuromagnetic Scene Categorization Study

Pavan Ramkumar[1], Sebastian Pannasch[1], Bruce C. Hansen[2], Adam M. Larson[3], and Lester C. Loschky[3]

[1] Brain Research Unit, O.V. Lounasmaa Laboratory, School of Science, Aalto University P.O. Box 15100, FI-00076, Finland
[2] Department of Psychology and Neuroscience Program, Colgate University, Hamilton, NY, USA
[3] Department of Psychology, Kansas State University, Manhattan, KS, USA
pavan@neuro.hut.fi

Abstract. How are visual scenes represented in the brain during categorization? We acquired magnetoencephalography (MEG) data from nine healthy subjects who participated in a rapid natural scene categorization task. Scenes were presented in two different perspectives (aerial vs. terrestrial) and two different orientations (upright vs. inverted). We applied multivariate pattern classification to categorize scene categories from computational (spatial envelope (SpEn): [6]) and neural representations (MEG responses). Predictions of both types of classifiers (1) exceeded chance but performed worse than human subjects, and (2) were significantly correlated in their pattern of predictions, suggesting the relevance of low-level visual features during scene categorization. In general, the pattern of predictions and errors were not correlated with behavioral predictions. We also examined the influence of perspective and orientation on neural and computational representations by studying the generalization performance of classifiers across perspective and orientation. We compared within-perspective-and-orientation classifiers (trained and tested on the same perspective and orientation) with across-perspective (trained on one perspective and tested on another) and across-orientation classifiers (trained on one orientation and tested on another). We report several interesting effects on category-level and identity-level (dis)agreement between neural, computational, and behavioral "views". To our knowledge, this is the first study to examine natural scene perception across scene perspectives and orientations from neural, computational, and behavioral angles.

Keywords: natural scene categorization, neural representations, spatial envelope, magnetoencephalography, multivariate pattern analysis, aerial views, terrestrial views, scene viewing orientation.

1 Introduction

Visual percepts arise from neural represention of the visual environment. Computational candidates mathematically describe how neural representations are

G. Langs et al. (Eds.): MLINI 2011, LNAI 7263, pp. 93–100, 2012.
© Springer-Verlag Berlin Heidelberg 2012

formed from natural images. By examining (1) the extent to which candidate computational representations agree with observed neural representations, and (2) how well computational and neural candidates predict behavior, we can begin to discover the true candidate neural mechanisms and computations underlying perception. Here, we propose and apply such an approach to the cognitive neuroscience of visual scene perception.

Henderson and Hollingworth [1] define the concept of a scene as "a semantically coherent (and often name-able) view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner." It is known from an early study that semantic information from scenes is available only from a single fixation [2]. Further, behavioral studies have suggested that it is even possible to infer scene category from scenes presented at durations much shorter than a typical fixation [7,8], or at low spatial resolution, where the level of detail is too coarse to accurately identify constituent objects [3,4].

Since the discovery of the parahippocampal place area [16], a dedicated brain region for scene perception, the computational role of related regions in the ventral visual stream have been under active study [17,15,18,19] using functional magnetic resonance imaging (fMRI). Recently, natural scene categories have been successfully decoded from fMRI data [13,12,14] suggesting that neural representations of scenes are accessible using non-invasive functional imaging techniques.

In the past decade, computational candidates have been put forward for how we represent scenes, and these have been shown to explain various aspects of behavioral scene categorization. Oliva and Torralba [6] have proposed low-level localized power spectra, that they called the "spatial envelope" (SpEn). They showed that the SpEn representation is sufficient to obtain good classification accuracies on scene categories. Recent work from the same group [5] showed that ratings of scene typicality correlated with the likelihood of correct scene classification based on a related low-level image feature representation called the all global feature space. Taken together, an attractive hypothesis for scene perception emerges: from brief exposures to a complex natural scene, humans categorize scenes on the basis of low-level scene statistics. Yet only little is known about whether such computations are carried out in the brain, and if so, how.

What is the nature of information represented by the ventral stream regions? Although it is known that the ventral visual stream contains high-level object representations invariant to size, pose, and occlusion, it is as yet unclear what sort of knowledge about scenes is represented neurally: high-level visual features, or semantic/conceptual knowledge. To probe the nature of information represented during scene categorization, Loschky and colleagues [11] characterized the behavioral consequences of a drastic change in scene perspective. They showed that the confusion matrices of behavioral scene categorization from aerial and terrestrial views were highly correlated, suggesting a possible semantic neural representation.

To follow up on their study, here, we addressed the following questions about neural scene representation. First, we asked how neurally realistic is a popular

computation representation: the spatial envelope. To this end, we quantified the similarity between the predictions of scene category classifiers on computational and neural representations. Second, we asked whether high-level visual or semantic representations of scenes are accessible from MEG responses. To this end we studied the ability of a classifier trained to predict scene categories from one perspective (aerial vs. terrestrial) or orientation (upright vs. vertically inverted) to predict them from another perspective or orientation.

2 Materials and Methods

2.1 Experimental Details

Stimuli were 736×736 pixel grayscale images of natural scenes from one of six possible categories, viz. airports, cities, coasts, forests, mountains, or suburbs, presented for 33 ms. We replicated the design from [11] but did not mask the stimuli. Nine healthy volunteers (2 females; mean age 32 years) were asked to categorize each scene using an eye-gaze-based response interface. Each category comprised 60 unique images: 30 aerial and 30 terrestrial scenes. There was no one-to-one correspondence between aerial and terrestrial scenes. Each image was presented in upright and vertically inverted orientations, resulting in 180 unique trials across 4 conditions: aerial upright (AERup), aerial inverted (AERdn), terrestrial upright (TERup), and terrestrial inverted (TERdn). We acquired MEG data (filtered at 0.03–330 Hz; sampled at 1000 Hz) using a 306-channel Elekta Vectorview system.

2.2 Representation of Neural Signals and Stimuli

The MEG data were preprocessed using temporal Signal Space Separation (tSSS) [9], downsampled to 500 Hz and low-pass filtered to 45 Hz. The evoked responses were separated from the trigger signals and a baseline correction was applied using a time window of 150 ms preceding the stimulus onset. Data from a post-stimulus window of 600 ms from 204 planar gradiometer channels were used. To reduce temporal redundancies, we applied a discrete cosine transform (DCT) to each channel and retained only the 50 coefficients corresponding to the lowest frequencies. For each trial, we concatenated DCT coefficients from each channel to constitute a feature vector.

For each stimulus image, we normalized local contrast and computed the SpEn features [10]. The SpEn features are localized energy spectra obtained by computing the energies of the input image convolved with Gabor filters at multiple scales and orientations. We precomputed Gabor filters at 8 orientations and 6 scales in the Fourier domain, multiplied each filter with the Fourier transform of the input image, and subsequently inverted the Fourier transform. We divided each filtered image into a coarse 4×4 grid and averaged the Fourier energies across the pixels in each block of the coarse grid, resulting in $8 \times 6 \times 4 \times 4 = 768$ image features.

2.3 Scene Classification

We built within-perspective-and-orientation classifiers (trained and tested on the same perspective and orientation; W_k's), across-perspective (trained on one perspective and tested on another; P_k's), and across-orientation classifiers (trained on one orientation and tested on another; O_k's).

For each subject, we trained one W_k for each condition viz. AERup, AERdn, TERup and TERdn on a random half of the data, i.e. 90 trials and tested them on the remaining 90 trials. The classifier was a multiclass support vector machine (SVM) which performed a majority voting on pairwise binary classifiers. We trained and tested W_k's separately on MEG and SpEn features. To obtain error estimates, we repeated the classification on 10 randomized cross-validation (CV) repeats separately for each subject. Next, we trained two P_k's: AERup $\rightarrow$ TERup and TERup $\rightarrow$ AERup, and two O_k's: AERup $\rightarrow$ AERdn, and TERup $\rightarrow$ TERdn. As before, we performed this classification on MEG features and on the SpEn features. We refer to the MEG and SpEn classification accuracies as α_n and α_c respectively, and the accuracy of behavioral reports as α_b. Table 1 gives the list of classifiers and their source and target conditions.

Table 1. List of classifiers implemented separately on MEG responses and SpEn features

Name	Source $\rightarrow$ Target	Train:Test	CV repeats
W_1	AERup $\rightarrow$ AERup	90 : 90	10
W_2	AERdn $\rightarrow$ AERdn	90 : 90	10
W_3	TERup $\rightarrow$ TERup	90 : 90	10
W_4	TERdn $\rightarrow$ TERdn	90 : 90	10
P_1	AERup $\rightarrow$ TERup	180 : 180	1
P_2	TERup $\rightarrow$ AERup	180 : 180	1
O_1	AERup $\rightarrow$ AERdn	180 : 180	1
O_2	TERup $\rightarrow$ TERdn	180 : 180	1

For each subject and randomized split, we computed the confusion matrix (CM) on the test set: each column of the CM represents one predicted category, while each row represents one true category, with correct categorization responses on the main diagonal and confusions in the off-diagonal cells. In addition to the MEG and SpEn CMs, we also computed CMs corresponding to the behavioral responses.

We quantified similarity of predictions from neural (MEG) features, computational (SpEn) features, and behavioral responses in two ways. First, we computed Spearman's rank correlation coefficients (ρ) between the entries of each pair of CMs (viz. neural-computational: ρ_{nc}, computational-behavioral: ρ_{cb}, and neural-behavioral: ρ_{nb}) concatenated over CV repeats, separately for each subject. Second, we computed the agreement fraction for a pair of classifiers, (θ), defined as the fraction of images from the test set for which both classifiers predict the same category.

3 Results and Discussion

3.1 Classification Accuracies

Figure 1A shows classification accuracies α_n, α_c and α_b along with their standard errors of mean, for P_k's, O_k's and W_k's. For neural classifiers, we found greater generalization across orientations than perspectives (O_k's were larger than P_k's) for terrestrial, but not aerial, scenes and that both performed worse than within-perspective-and-orientation classifiers (W_k's). For computational classifiers (based on SpEn), those generalized across orientations were as good as within-perspective-and-orientation classifiers (O_k's were equal to W_k's). However, generalization across perspectives was very poor (P_k's were low). We make the following remarks about these accuracy measures.

1. In general, all classifiers exceeded chance level. However, neither computational nor neural classifiers were as accurate as human subjects' behavioral responses.
2. For behavioral responses, and to a lesser extent the computational classifiers, the within-orientation-and-perspective accuracies were higher for terrestrial (W_3 and W_4) than aerial scenes (W_1 and W_2). However, the neural classifiers showed a reverse trend: aerial conditions had higher accuracies than terrestrial ones.

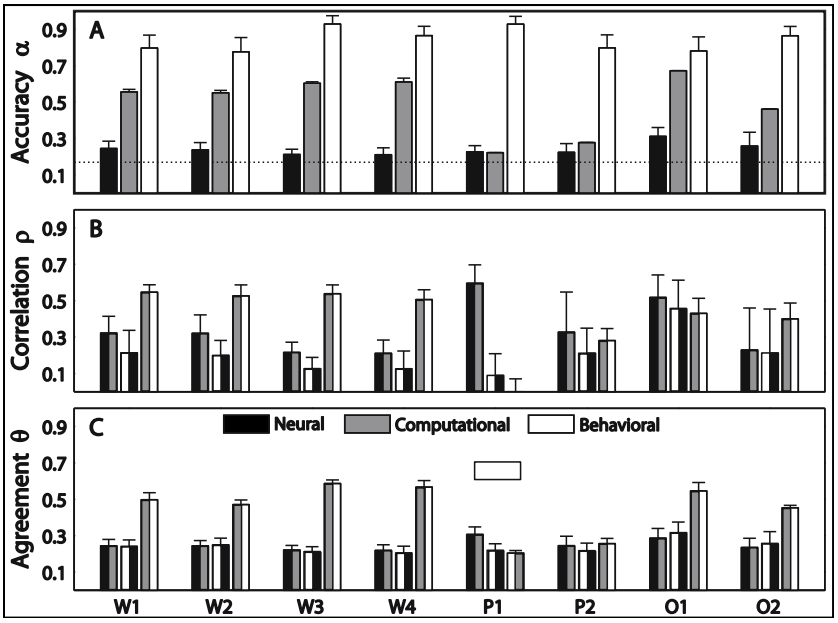

Fig. 1. A. Accuracies α (dotted line represents chance level), B. Spearman's correlation coefficient ρ, and C. agreement fraction θ for the various classifiers. Error bars represent standard errors of mean across subjects.

3. Computational classifiers were more accurate than neural classifiers for within-orientation-and-perspective classifiers and across-orientation classifiers (W_k's and O_k's) but not for the across-perspective classifiers (P_k's). Among across-perspective classifiers (P_k's) the neural classifiers behaved similarly in both directions of generalization, but the computational classifiers did not. In particular, the computational classifiers from terrestrial to aerial perspectives (P_2) generalized better than the reverse direction (P_1), suggesting that low-level statistics of terrestrial upright scenes are more predictive of scene categories from aerial upright scenes, than vice versa.
4. For across-orientation classifiers, the computational classifiers for aerial scenes (O_1) performed better than those for terrestrial scenes (O_2), suggesting that computational representations of aerial scenes generalize better across viewing orientations than the computational representations of terrestrial scenes do. A similar trend was observed for the neural O_k's but they were not significantly different. These findings agree with our observation that aerial scene accuracies are not affected by inversion, whereas terrestrial scene accuracies are greatly reduced by inversion [11]. The findings are also understandable if aerial scenes tend to have more cardinal orientations than terrestrial scenes (i.e., the orientations are more symmetrically biased) since cardinal orientations are preserved by inversion (i.e., 180 deg rotation), but we did not test for this explicitly.

3.2 Correlation and Agreement between Classifiers

Figure 1B shows Spearman's rank correlation coefficients between pairs of neural, computational or behavioral confusion matrices ρ_{nc}, ρ_{nb} and ρ_{cb} along with their standard errors of mean, for P_k's, O_k's and W_k's. The correlation between confusion matrices is a measure of how similarly two classifiers err at the level of categories. Figure 1C shows the agreement fractions θ_{nc}, θ_{nb} and θ_{cb}. The agreement fractions are a stronger measure of similarity between two classifiers because they measure the extent to which classifiers agree at the level of each individual stimulus. We make the following remarks about these measures.

1. For all within-orientation-and-perspective classifiers (W_k's), neural classifiers were weakly correlated with both computational and behavioral classifiers. In comparison, computational and behavioral classifiers were more strongly correlated. We found no clear differences between any of the orientations or perspectives. The agreement measure seems to confirm this general trend although the computational-behavioral agreement metrics tended to be lower for the aerial (W_1, W_2) than the terrestrial (W_3, W_4) scenes.
2. For across-perspective classifiers from aerial to terrestrial upright scenes (P_1), neural classifiers are strongly correlated with computational classifiers, but there is almost no correlation between neural vs. behavioral and

computational and behavioral classifiers. However, the difference between the classifier pairs is less pronounced for the agreement metric. This suggests that athough neural and computational classifiers err similarly at the category level, they err quite differently at the level of individual images, while attempting to generalize from aerial to terrestrial perspectives.

3. Although the correlations are comparable between classifiers across orientations and perspectives (O_k's and P_k's), the computational-behavioral agreement metrics for the O_k's clearly exceeded the P_k's. This suggests that computational classifiers perform similarly as humans when they generalize across orientations rather than perspectives.
4. Correlations between computational and neural classifiers were higher for the across-perspective classifiers from aerial to terrestrial scenes (P_1), and the across-orientation classifier from upright to inverted aerial scenes (O_1), than all other classifiers. This observation, together with almost equivalent accuracies for all neural classiers suggests that low-level visual information in the MEG response contributes more towards classification than high-level visual or semantic information.

4 Conclusion

Using MVPA of MEG responses to natural scenes, we showed that for both upright aerial and terrestrial perspectives, it was possible to decode scene categories above chance level. We also found that the pattern of scene category predictions from brain activity were weakly but significantly correlated with the pattern of predictions from low-level image statistics. While our result is not causal evidence, given its basis in correlational analyses, it supports the possibility that low-level statistics of scenes such as the spatial envelope are robustly represented in MEG responses.

The presented framework—comparing the pattern of errors in a classification task across neural and computational representations—is widely applicable to experimentally test computational theories of perceptual and semantic representation. More broadly, constraining neural and computational representations to agree, and constraining these in turn to predict behavioral observations (see eg. [14] for a study comparing behavioral categorization and fMRI-based classification of natural scenes) will help us elucidate the computational strategies and neural mechanisms underlying cognition.

Acknowledgments. We gratefully acknowledge the Finnish Graduate School of Neuroscience, the ERC Advanced Grant #232946 (to R. Hari), the FP7-PEOPLE-2009-IEF program #254638 (to S. Pannasch), and and the Office of Naval Research GRANT #10846128 (to L. Loschky) for their generous funding.

References

1. Henderson, J.M., Hollingworth, A.: High-level scene perception. Annual Review of Psychology 50, 243–271 (1999)
2. Potter, M.C.: Meaning in visual scenes. Science 187, 965–966 (1975)
3. Oliva, A., Schyns, P.: Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology 34, 72–107 (1997)
4. Oliva, A., Schyns, P.: Diagnostic colors mediate scene recognition. Cognitive Psychology 41, 176–210 (2000)
5. Ehinger, K.A., Xiao, J., Torralba, A., Oliva, A.: Estimating scene typicality from human ratings and image features. In: Proceedings of the 33rd Annual Cognitive Science Conference, Boston (2011)
6. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Intl. J. Comp. Vis. 42, 145–175 (2001)
7. Greene, M.R., Oliva, A.: The briefest of glances: the time course of natural scene understanding. Psychological Science 20, 464–472 (2009)
8. Loschky, L.C., Larson, A.M.: The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing. Visual Cognition 18, 513–536 (2010)
9. Taulu, S., Simola, J.: Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. 41, 1759–1768 (2006)
10. `http://people.csail.mit.edu/torralba/code/spatialenvelope/`
11. Loschky, L., Ellis, K., Sears, T., Ringer, R., Davis, J.: Broadening the Horizons of Scene Gist Recognition: Aerial and Ground-based Views. J. Vis. 10, 1238 (2010)
12. Walther, D.B., Caddigan, E., Fei-Fei, L., Beck, D.M.: Natural scene categories revealed in distributed patterns of activity in the human brain. J. Neurosci. 29, 10573–10581 (2009)
13. Peelen, M., Fei-Fei, L., Kastner, S.: Neural mechanisms of rapid scene categorization in human visual cortex. Nature 460, 94–97 (2009)
14. Walther, D.B., Chai, B., Caddigan, E., Beck, D.M., Fei-Fei, L.: Simple line drawings suffice for fMRI decoding of natural scene categories. PNAS USA 108, 9661–9666 (2011)
15. Kravitz, D.J., Peng, C.S., Baker, C.I.: Real-World Scene Representations in High-Level Visual Cortex: It's the Spaces More Than the Places. J. Neurosci. 31, 7322–7333 (2011)
16. Epstein, R.A., Harris, A., Stanley, D., Kanwisher, N.: The parahippocampal place area: recognition, navigation, or encoding? Neuron 23, 115–125 (1999)
17. Epstein, R.A., Higgins, J.S.: Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. Cereb. Cortex 17, 1680–1693 (2007)
18. Park, S., Brady, T.F., Greene, M.R., Oliva, A.: Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. J. Neurosci. 31, 1333–1340 (2011)
19. MacEvoy, S.P., Epstein, R.A.: Constructing scenes from objects in human occipitotemporal cortex. Nat. Neurosci. 14, 1323–1329 (2011)