NeuroImage 153 (2017) 346–358

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks

Radoslaw Martin Cichy[a,b,*], Aditya Khosla[b], Dimitrios Pantazis[c], Aude Oliva[b]

[a] Department of Education and Psychology, Free University Berlin, Berlin, Germany
[b] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[c] McGovern Institute for Brain Research, MIT, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Scene perception
Spatial layout
Magnetoencephalography
Deep neural network
Representational similarity analysis

ABSTRACT

Human scene recognition is a rapid multistep process evolving over time from single scene image to spatial layout processing. We used multivariate pattern analyses on magnetoencephalography (MEG) data to unravel the time course of this cortical process. Following an early signal for lower-level visual analysis of single scenes at ~100 ms, we found a marker of real-world scene size, i.e. spatial layout processing, at ~250 ms indexing neural representations robust to changes in unrelated scene properties and viewing conditions. For a quantitative model of how scene size representations may arise in the brain, we compared MEG data to a deep neural network model trained on scene classification. Representations of scene size emerged intrinsically in the model, and resolved emerging neural scene size representation. Together our data provide a first description of an electrophysiological signal for layout processing in humans, and suggest that deep neural networks are a promising framework to investigate how spatial layout representations emerge in the human brain.

Introduction

Perceiving the geometry of space is a core ability shared by all animals, with brain structures for spatial layout perception and navigation preserved across rodents, monkeys and humans (Epstein and Kanwisher, 1998, 1998; Doeller et al., 2008, 2010; Moser et al., 2008; Epstein, 2011; Jacobs et al., 2013; Kornblith et al., 2013; Vaziri et al., 2014). Spatial layout perception, the demarcation of the boundaries and size of real-world visual space, plays a crucial mediating role in spatial cognition (Bird et al., 2010; Epstein, 2011; Kravitz et al., 2011a; Wolbers et al., 2011; Park et al., 2015) between image-specific processing of individual scenes and navigation-related processing. Although the cortical loci of spatial layout perception in humans have been well described (Aguirre et al., 1998; Kravitz et al., 2011b; MacEvoy and Epstein, 2011; Mullally and Maguire, 2011; Park et al., 2011; Bonnici et al., 2012), the dynamics of spatial cognition remain unexplained, partly because neuronal markers indexing spatial layout processing remain unknown, and partly because quantitative models of spatial layout processing are missing. The central questions of this study are thus twofold: First, what are the temporal dynamics with which representation of spatial layout emerge in the brain? And second, how can the emergence of representations of spatial layout in cortical circuits be modeled?

The temporal dynamics of spatial layout processing

Given the intermediate position of spatial layout perception in the visual processing hierarchy between image-specific processing of individual scenes and navigation-related processing, we hypothesized that a signal for spatial layout processing would emerge after signals related to low-level visual processing in early visual regions (~100 ms, (Schmolesky et al., 1998; Cichy et al., 2015a)), and before activity observed typically in navigation-related regions such as the hippocampus (~400 ms (Mormann et al., 2008)). Further, to be considered as an independent step in visual scene processing, spatial layout must be processed tolerant to changes in low-level features, including typical variations in viewing conditions, and to changes in high-level features such as scene category. We thus hypothesized that representation of spatial layout would be tolerant to changes in both low- and high-level visual properties.

To investigate, we operationalized spatial layout as scene size, that is the size of the space a scene subtends in the real-world (Kravitz et al., 2011a; Park et al., 2011, 2015). Using multivariate pattern classification (Carlson et al., 2013; Cichy et al., 2014; Isik et al., 2014) and representational similarity analysis (Kriegeskorte, 2008; Kriegeskorte and Kievit, 2013; Cichy et al., 2014) on millisecond-resolved magne-

[*] Corresponding author at: Computer Science and Artificial Intelligence Laboratory, MIT, 32-D430 Cambridge, MA, USA.
E-mail address: rmcichy@mit.edu (R. Martin Cichy).

http://dx.doi.org/10.1016/j.neuroimage.2016.03.063
Received 23 November 2015; Accepted 23 March 2016
Available online 01 April 2016

toencephalography data (MEG), we identified a marker of scene size around 250 ms, preceded by and distinct from an early signal for lower-level visual analysis of scene images at ~100ms. Furthermore, we demonstrated that the scene size marker was independent of both low-level image features (i.e. luminance, contrast, clutter, image identity) and semantic properties (the category of the scene, i.e. kitchen, ballroom), thus indexing neural representations robust to changes in viewing conditions as encountered in real-world settings.

A model of scene size representations

As an intermediate visual processing stage, spatial layout perception is likely to be underpinned by representations in intermediate- and high-level visual regions, where neuronal responses are often complex and nonlinear. To model such visual representations, complex hierarchical models might be necessary. We thus hypothesized that representation of scene size would emerge in complex deep neural networks, rather than in compact models of object and scene perception. To investigate, we compared brain data to a deep neural network model trained to perform scene categorization (Zhou et al., 2014, 2015), termed deep scene network. The deep scene network *intrinsically* exhibited receptive fields specialized for layout analysis, such as textures and surface layout information, without ever having been explicitly taught any of those features. We showed that the deep scene neural network model predicted the human neural representation of single scenes and scene space size better than a deep object model and standard models of scene and object perception HMAX and GIST (Riesenhuber and Poggio, 1999; Oliva and Torralba, 2001). This demonstrates the ability of the deep scene model to approximate human neural representations at successive levels of processing as they emerge over time.

In sum, our results give a first description of an electrophysiological signal for scene space processing in humans, providing evidence for representations of spatial layout emerging between low-level visual and navigation-related processing. They further offer a novel quantitative and computational model of the dynamics of visual scene space representation in the cortex, suggesting that spatial layout representations naturally emerge in cortical circuits learning to differentiate visual environments (Oliva and Torralba, 2001).

Materials and methods

Participants

Participants were 15 right-handed, healthy volunteers with normal or corrected-to-normal vision (mean age ± s.d.=25.87 ± 5.38 years, 11 female). The Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT approved the experiment and each participant gave written informed consent for participation in the study, for data analysis and publication of study results.

Stimulus material and experimental design

The image set consisted of 48 scene images differing in four factors with two levels each, namely two scene properties: physical size (small, large) and clutter level (low, high); and two image properties: contrast (low, high) and luminance (low, high) (Fig. 1A). There were 3 unique images for every level combination, for example 3 images of small size, low clutter, low contrast and low luminance. The image set was based on behaviorally validated images of scenes differing in size and clutter level, sub-sampling the two highest and lowest levels of factors size and

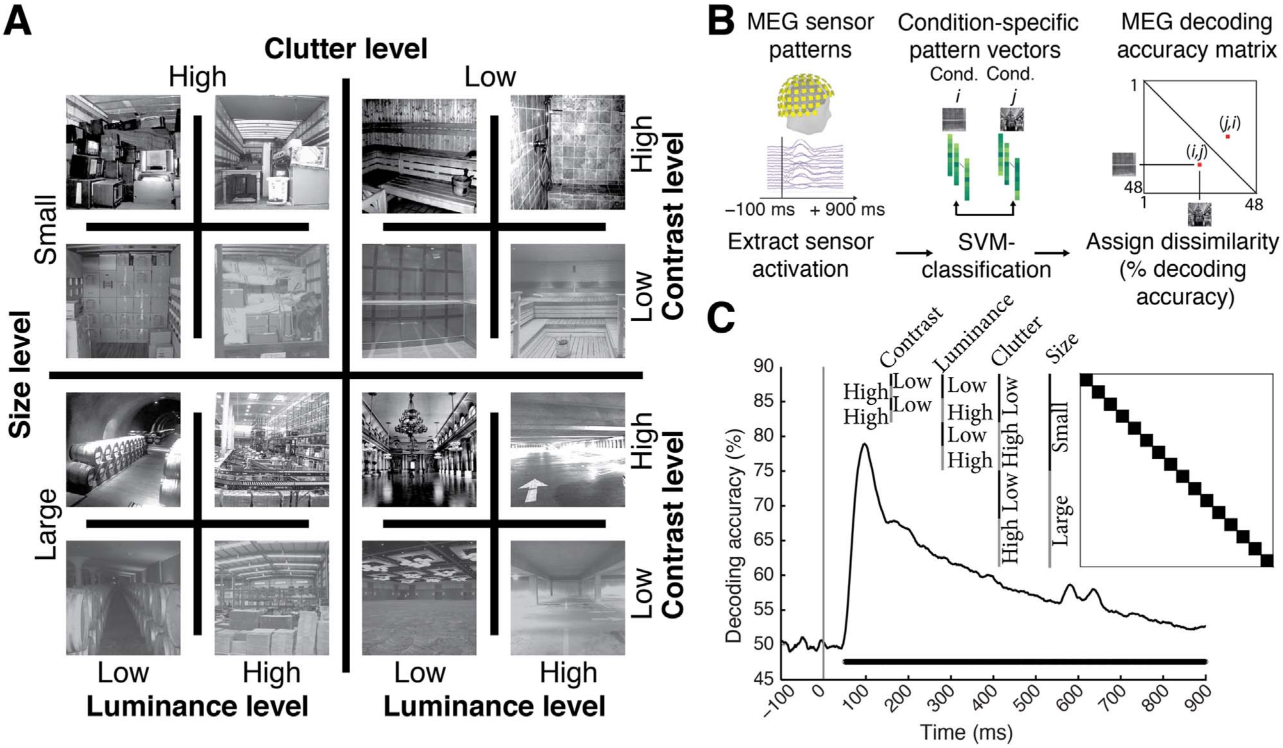

Fig. 1. *Image set and single-image decoding. A)* The stimulus set comprised 48 indoor scene images differing in the size of the space depicted (small vs. large), as well as clutter, contrast, and luminance level; here each experimental factor combination is exemplified by one image. The image set was based on behaviorally validated images of scenes differing in size and clutter level, de-correlating factors size and clutter explicitly by experimental design (Park et al., 2015). Note that size refers to the size of the real-world space depicted on the image, not the stimulus parameters; all images subtended 8 visual angle during the experiment. B) Time-resolved (1 ms steps from -100 to +900 ms with respect to stimulus onset) pair-wise support vector machine classification of experimental conditions based on MEG sensor level patterns. Classification results were stored in time-resolved 48×48 MEG decoding matrices. C) Decoding results for single scene classification independent of other experimental factors. Decoding results were averaged across the dark blocks (matrix inset), to control for luminance, contrast, clutter level and scene size differences. Inset shows indexing of matrix by image conditions. Horizontal line below curve indicates significant time points (n=15, cluster-definition threshold $P < 0.05$, corrected significance level $P < 0.05$); gray vertical line indicates image onset.