NeuroImage 178 (2018) 172–182

Contents lists available at ScienceDirect

NeuroImage

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

The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks

B.B. Bankson [*,1], M.N. Hebart [1], I.I.A. Groen, C.I. Baker

Section on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

Visual object representations are commonly thought to emerge rapidly, yet it has remained unclear to what extent early brain responses reflect purely low-level visual features of these objects and how strongly those features contribute to later categorical or conceptual representations. Here, we aimed to estimate a lower temporal bound for the emergence of conceptual representations by defining two criteria that characterize such representations: 1) conceptual object representations should generalize across different exemplars of the same object, and 2) these representations should reflect high-level behavioral judgments. To test these criteria, we compared magnetoencephalography (MEG) recordings between two groups of participants ($n = 16$ per group) exposed to different exemplar images of the same object concepts. Further, we disentangled low-level from high-level MEG responses by estimating the unique and shared contribution of models of behavioral judgments, semantics, and different layers of deep neural networks of visual object processing. We find that 1) both generalization across exemplars as well as generalization of object-related signals across time increase after 150 ms, peaking around 230 ms; 2) representations specific to behavioral judgments emerged rapidly, peaking around 160 ms. Collectively, these results suggest a lower bound for the emergence of conceptual object representations around 150 ms following stimulus onset.

Introduction

There is enormous variability in the visual appearance of objects, yet we can rapidly recognize them without effort, even under difficult viewing conditions (DiCarlo & Cox, 2007; Potter et al., 2014). Evidence from neurophysiological studies in human suggests the emergence of visual object representations within the first 150 ms of visual processing (Thorpe et al., 1996; Carlson et al., 2013; Cichy et al., 2014). For example, the specific identity of objects can be decoded from the magnetoencephalography (MEG) signal with high accuracy around 100 ms (Cichy et al., 2014). However, knowing when discriminative information about visual objects is available does not inform us about the nature of those representations, in particular whether they primarily reflect (low-level) visual features or (high-level) conceptual aspects of the objects (Clarke et al., 2014). To address this issue, in this study we employed multivariate MEG decoding and model-based representational similarity analysis (RSA) to elucidate the nature of object representations over time.

Previous studies have demonstrated increasing category specificity (van de Nieuwenhuijzen et al., 2013; Cichy et al., 2014), tolerance for position and size (Isik et al., 2014) and semantic information (Clarke et al., 2013) over the first 200 ms following stimulus onset, suggesting some degree of abstraction from low-level visual features. However, identifying the nature of object representations is an inherently difficult problem: low-level features may be predictive of object identity, making it hard to disentangle the relative contribution of low and high-level properties to measured brain signals (Groen et al., 2017). In this study, we addressed this problem by combining tests for the generalization of object representations with methods to separate the independent contributions of low- and high-level properties. We focused on two specific criteria that would need to be fulfilled for a representation to be considered conceptual. First, a conceptual representation should generalize beyond the specific exemplar presented, not just variations of the same exemplar. Second, a conceptual representation should also reflect high-level behavioral judgments about objects (Clarke and Tyler, 2015; Wardle et al., 2016). We consider fulfillment of these two properties to provide a lower bound at which a representation could be considered conceptual.

We collected MEG and behavioral data from 32 participants allowing us to probe the temporal dynamics of conceptual object representations according to the two criteria above. To test for generalization across specific exemplars, we assessed the reliability of object representations across two independent sets of objects. Further, we assessed the relation of those object representations to behavior by comparing participants' behavioral judgments with the MEG response patterns using RSA. Importantly, to isolate the relative contributions of low-level and

* Corresponding author. Laboratory of Cognitive Neurodynamics, UPMC Presbyterian, Suite B-400, 200 Lothrop Street, Pittsburgh, PA 15213, USA.
E-mail address: bbb17@pitt.edu (B.B. Bankson).

[1] Equal contribution.

https://doi.org/10.1016/j.neuroimage.2018.05.037
Received 22 November 2017; Received in revised form 10 May 2018; Accepted 14 May 2018
Available online 17 May 2018

conceptual properties to those MEG responses, we identified the variance uniquely explained by behavioral judgments, isolating low-level representations using early layers of a deep neural network, which have been shown to capture low-to mid-level responses in fMRI and monkey ventral visual cortex (Cadieu et al., 2014; Cichy et al., 2016a; Eickenberg et al., 2017; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Wen et al., 2017). Finally, to achieve a more interpretable understanding of the contribution of behavior to MEG responses, we identified the unique and shared variance explained in the MEG response by behavior and two high-level conceptual models, one perceptual (upper layers in a deep neural network) and one semantic (based on word co-occurrence statistics).

Methods

Participants

32 healthy participants (18 female, mean 25.8, range 19–47) with normal or corrected-to-normal vision took part in this study. As a part of a pilot experiment used for purely illustrative purposes (see Figure 4a), 8 participants (5 overlap) completed the same behavioral task with a different set of stimuli. All participants gave written informed consent prior to participation in the study as a part of the study protocol (93-M-0170, NCT00001360). The study was approved by the Institutional Review Board of the National Institutes of Health and was conducted according to the Declaration of Helsinki.

Stimuli

We created two independent sets of 84 object images each that were cropped and placed on a grey background. Each stimulus set contained a unique exemplar for each of the 84 object concepts, as shown in Fig. 1a. We selected object concepts by using a combination of two word databases, one of word frequency (Corpus of Contemporary American English, Davies, 2008) and the other of word concreteness (Brysbaert et al., 2014). First, based on our corpus we selected the 5000 most frequent nouns in American English. From this set of words, we then selected nouns with concreteness ratings >4/5. Finally, for words that would be difficult or impossible to distinguish when presented as an image (e.g. 'woman', 'mother', 'wife'), we used only the most frequent entry. This selection left us with a set of 112 objects.

To evaluate whether those categories would be labeled consistently, we generated three distinct images of each object concept and asked three individuals who were not involved in the study to provide a verbal label for each of the three versions of the 112 objects. Images that were not labeled correctly by all raters were discarded, leaving us with 84

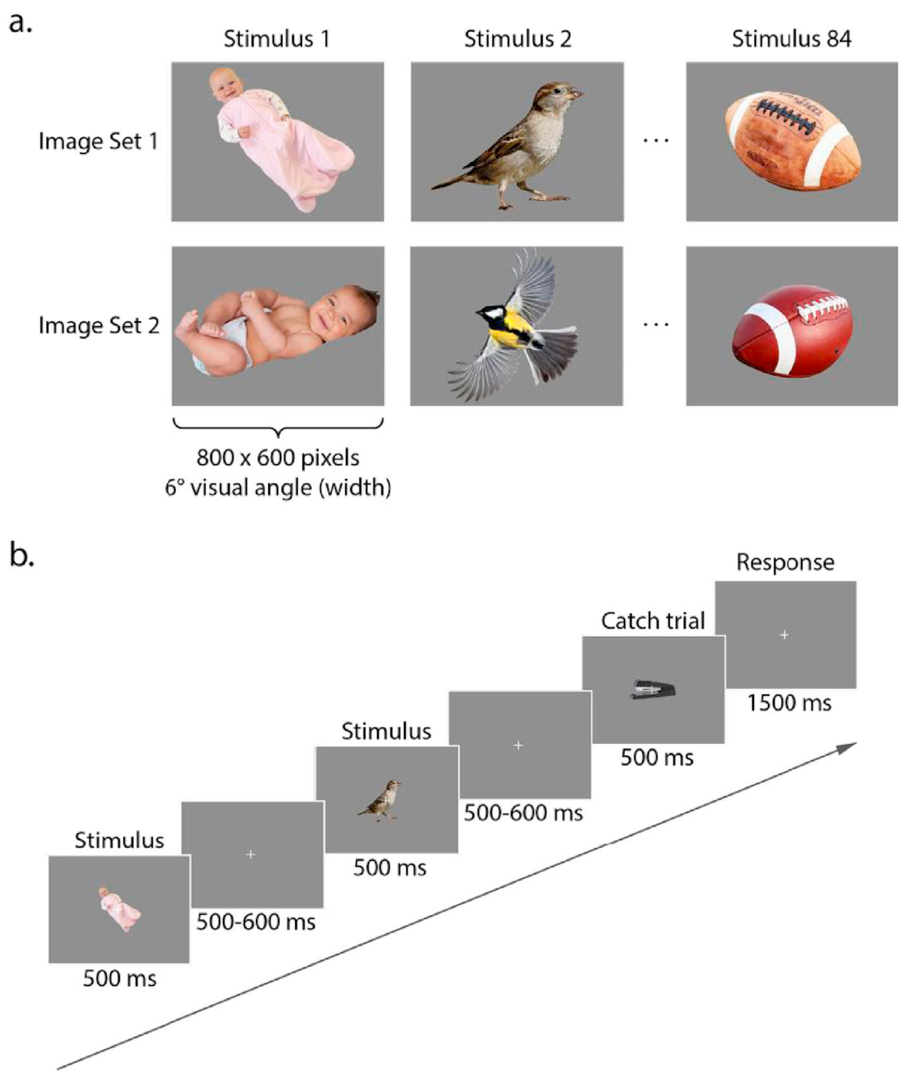

Fig. 1. Stimulus format and trial progression. **a.** Two unique object exemplars were selected for each of the 84 object concepts used in the study. **b.** Stimuli were presented on a grey background for 500 ms, followed by fixation for 500–600 ms (catch trials: 1500 ms). All 84 stimuli from both image sets are shown in Supplemental Fig. S1.