NeuroImage 179 (2018) 102–116

Contents lists available at ScienceDirect

NeuroImage

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

Spatial frequency supports the emergence of categorical representations in visual cortex during natural scene perception

Diana C. Dima [*], Gavin Perry, Krish D. Singh

Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, CF24 4HQ, United Kingdom

ARTICLE INFO

Keywords:
Multivariate pattern analysis (MVPA)
Representational similarity analysis (RSA)
Magnetoencephalography (MEG)
Convolutional neural network (CNN)
Scene categorization

ABSTRACT

In navigating our environment, we rapidly process and extract meaning from visual cues. However, the relationship between visual features and categorical representations in natural scene perception is still not well understood. Here, we used natural scene stimuli from different categories and filtered at different spatial frequencies to address this question in a passive viewing paradigm. Using representational similarity analysis (RSA) and cross-decoding of magnetoencephalography (MEG) data, we show that categorical representations emerge in human visual cortex at ~180 ms and are linked to spatial frequency processing. Furthermore, dorsal and ventral stream areas reveal temporally and spatially overlapping representations of low and high-level layer activations extracted from a feedforward neural network. Our results suggest that neural patterns from extrastriate visual cortex switch from low-level to categorical representations within 200 ms, highlighting the rapid cascade of processing stages essential in human visual perception.

Introduction

Classic models of natural vision entail a hierarchical process transforming low-level properties into categorical representations (VanRullen and Thorpe, 2001; Yamins and DiCarlo, 2016). During early stages of natural scene perception, the primary visual cortex processes low-level stimulus properties using inputs from the retina via the lateral geniculate nucleus (Hubel and Wiesel, 1962). Extrastriate and scene-selective areas are associated with mid-level and high-level properties, with categorical, invariant representations considered the final stage of abstraction (Felleman and Van Essen, 1991; Ungerleider and Haxby, 1994). Scene-selective brain regions such as the parahippocampal place area (PPA), the retrosplenial cortex (RSC), and the occipital place area (OPA) are often thought to represent such categories (Walther et al., 2009) and have been found to respond to high-level stimuli in controlled experiments (Schindler and Bartels, 2016; Walther et al., 2011).

However, this model has been challenged by evidence of low- and mid-level features being processed in scene-selective areas (Kauffmann et al., 2015b; Kravitz et al., 2011; Nasr et al., 2014; Nasr and Tootell, 2012; Rajimehr et al., 2011; Watson et al., 2014, 2016). Studies of temporal dynamics have found overlapping signatures of low-level and high-level representations (Groen et al., 2013; Harel et al., 2016), suggesting co-occurring and co-localized visual and categorical processing (Ramkumar et al., 2016). Such evidence casts doubt on the hierarchical model and on the usefulness of the distinction between low-level and high-level properties (Groen et al., 2017).

In particular, spatial frequency is thought to play an important part in natural scene perception, with low spatial frequencies mediating an initial rapid parsing of visual features in a “coarse-to-fine” sequence (Kauffmann et al., 2015b). Its role in the processing speed of different features, as well as evidence of its contribution to neural responses in scene-selective areas (Rajimehr et al., 2011), makes spatial frequency a particularly suitable candidate feature for teasing apart the temporal dynamics of low and high-level natural scene processing.

Recent neuroimaging studies of scene perception have used multivariate pattern analysis (MVPA) to highlight the links between low-level processing and behavioural goals (Ramkumar et al., 2016; Watson et al., 2014). In particular, Ramkumar et al. (2016) showed successful decoding of scene gist from MEG data and linked decoding performance to spatial envelope properties, as well as behaviour in a categorization task.

In the present study, we aimed to dissociate the role of low-level and high-level properties in natural scene perception, in the absence of behavioural goals that may influence visual processing (Groen et al., 2017). In order to do so, we recorded MEG data while participants passively viewed a controlled stimulus set composed of scenes and scrambled stimuli filtered at different spatial frequencies. Thus, we were

[*] Corresponding author. Cardiff University Brain Research Imaging Centre, CUBRIC Building, Maindy Road, Cardiff, CF24 4HQ, United Kingdom.
E-mail address: DimaDC@cardiff.ac.uk (D.C. Dima).

https://doi.org/10.1016/j.neuroimage.2018.06.033
Received 21 December 2017; Received in revised form 1 June 2018; Accepted 9 June 2018
Available online 11 June 2018

able to first contrast responses to scenes with responses to matched control stimuli (which, to the extent of our knowledge, have not yet been used in the M/EEG literature on natural scenes); and second, we were able to assess the presence of a categorical response to scenes invariant to spatial frequency manipulations.

We used multivariate pattern analysis (MVPA) and representational similarity analysis (RSA) to explore representations of scene category in space and time and to assess their relationship to low-level properties. Multivariate analyses are sensitive to differences in overlapping patterns (Norman et al., 2006) and can describe the spatiotemporal dynamics and structure of neural representations through information mapping (Kriegeskorte et al., 2008, 2006).

We successfully decoded scene category from MEG responses in the absence of an explicit categorization task, and a cross-decoding analysis suggested that this effect is driven by low spatial frequency features at ~170 ms post-stimulus onset. We also show that categorical representations arise in extrastriate visual cortex within 200 ms, while at the same time representations in posterior cingulate cortex correlate with the high-level layers of a convolutional neural network. Together, our results suggest that scene perception relies on low spatial frequency features to create an early categorical representation in visual cortex.

Methods

Participants

Nineteen participants took part in the MEG experiment (10 females, mean age 27, standard deviation SD 4.8), and fourteen in a control behavioural experiment (13 females, mean age 26, SD 4.4). All participants were healthy, right-handed and had normal or corrected-to-normal vision (based on self-report). Written consent was obtained in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). All procedures were approved by the ethics committee of the School of Psychology, Cardiff University.

Stimuli

Stimuli (Supplementary Figure 2) were 20 natural scenes (fields, mountains, forests, lakes and seascapes) and 20 urban scenes (office buildings, houses, city skylines and street views) from the SUN database (Xiao et al., 2010). Stimuli were 800 × 600 pixels in size, subtending 8.6 × 6.4 degrees of visual angle.

All the images were converted to grayscale. Using the SHINE toolbox (Willenbockel et al., 2010), luminance and contrast were normalized to the mean luminance and SD of the image set. Spatial frequency was matched across stimuli by equating the rotational average of the Fourier amplitude spectra (the energy at each spatial frequency across orientations).

To assess the similarity of image amplitude spectra between categories, we calculated pairwise Pearson's correlation coefficients based on pixel intensity values between all images (mean correlation coefficient 0.14, SD 0.27, minimum-maximum range 1.33). Next, we performed an equivalence test (two one-sided tests; Lakens, 2017) in order to compare within-category correlation coefficients from both conditions (i.e., pairwise correlation coefficients between each image and each of the 19 images belonging to the same category) to between-category correlation coefficients (i.e., pairwise correlation coeffients between each image and each of the 20 images belonging to the other category). We assumed correlation coefficients to be similar if the difference between them fell within the [-0.1, 0.1] equivalence interval (Cohen, 1992). Within-category and between-category correlation coefficients were found to be equivalent ($P_1 = 5.3*10^{-11}$, $P_2 = 2.4*10^{-4}$, 90% CI [-0.0025, 0.063]).

Prior to spatial frequency filtering, the mean of each image was set to 0 to avoid DC artefacts and effects induced by zero-padding. To obtain low spatial frequency (LSF) and high spatial frequency (HSF) stimuli, we applied a low-pass Gaussian filter with a cutoff frequency of 3 cycles per degree (25.8 cycles per image) and a high-pass filter with a cutoff of 6 cycles per degree (51.6 cycles per image). Root mean square (RMS) contrast (standard deviation of pixel intensities divided by their mean) was only normalized within and not across spatial frequency conditions, in order to maintain the characteristic contrast distribution typical of natural scenes, which has been shown to influence responses to spatial frequency in the visual system (Field, 1987; Kauffmann et al., 2015b, 2015a).

To produce control stimuli, we scrambled the phase of the images in the Fourier domain, ensuring equivalent Fourier amplitude spectra across the original and scrambled images (Perry and Singh, 2014). For each spatial frequency condition, we randomly selected 10 of the 20 phase-scrambled images for use in the experiment in order to maintain an equal number of stimuli across conditions (natural, urban and scrambled). The final stimulus set contained 180 images (filtered and unfiltered scenes and scrambled stimuli; Fig. 1, Supplementary Figure 1).

Behavioural experiment

Design and data collection

To assess potential differences in the recognizability of different scenes, participants in the behavioural experiment viewed the stimuli and were asked to categorize them as fast as possible. The design of the behavioural experiment was similar to the MEG experiment, but included a practice phase (10 trials) before each block. Participants underwent two blocks in which they had to judge whether stimuli were scenes or scrambled stimuli, or whether scene stimuli were natural or urban respectively. Blocks were separated by a few minutes' break and their order was counterbalanced across subjects.

Images were presented on an LCD monitor with a resolution of 1920 × 1080 pixels and a refresh rate of 60 Hz. Participants were required to make a keyboard response (using the keys ‘J’ and ‘K’, whose meanings were counterbalanced across subjects), as soon as each image appeared on screen. We recorded responses and reaction times using Matlab R2015a (The Mathworks, Natick, MA, USA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007).

Data analysis

To assess the effect of spatial frequency filtering on performance in the categorization task, one-way repeated-measures ANOVAs were performed on individual accuracies (after performing a rationalized arcsine transformation; Studebaker, 1985) and on mean log-transformed reaction times for each categorization task (four tests with a

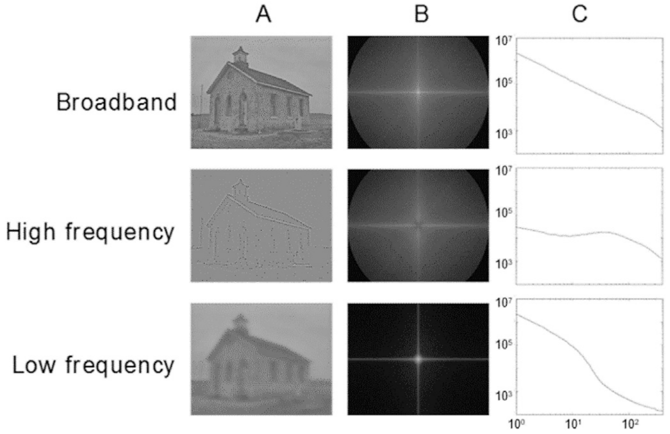

Fig. 1. Examples of urban scene stimuli filtered at different spatial frequencies (A), together with the average Fourier spectra (B) and frequency power spectra (C) for each stimulus set (log spectral power on the y-axis plotted against log spatial frequency on x-axis).