IMT School for Advanced Studies, Lucca
Lucca, Italy

Modeling object representations in natural vision

PhD Program in Cognitive Computational and Social Neuroscience
XXXII Cycle

by Paolo Papale
2019

Program Coordinator: Prof. Pietro Pietrini, IMT School for Advanced Studies

Thesis Advisor: Prof. Pietro Pietrini, IMT School for Advanced Studies

The dissertation of Paolo Papale has been reviewed by:

Viviana Betti, *La Sapienza, Rome, Italy*

Peter Neri, *CNRS, Paris, France*

IMT School for Advanced Studies, Lucca 2019

Contents

List of figures

List of tables

Acknowledgments

The content presented in Chapter 2 has been published in the journal *eNeuro* as:

> *Papale P, Leo A, Cecchetti L, Handjaras G, Kay KN, Pietrini P, Ricciardi E (2018)*
> *Foreground-Background Segmentation Revealed during Natural Image Viewing.*
> *eNeuro 5.* License: CC BY 4.0

The content presented in Chapter 3 has been published in the journal *Scientific Reports* as:

> *Papale P, Betta M, Handjaras G, Malfatti G, Cecchetti L, Rampinini A, Pietrini P, Ricciardi E, Turella L, Leo A (2019)*
> *Common spatiotemporal processing of visual features shapes object representation.*
> *Sci Rep 9:7601.* License: CC BY 4.0

The content presented in Chapter 4 is actually under review with the following author list and title:

> *Papale P*, Leo A*, Handjaras G, Cecchetti L, Pietrini P, Ricciardi E.*
> *Shape coding in occipito-temporal cortex relies on object silhouette, curvature and medial-axis*
> *bioRxiv,* doi.org/10.1101/814251 **shared first authorship*

Vita and Publications

Vita

1989

Born in Pisa, Italy.

2014

M.A. Degree in Architecture, School of Architecture, University of Florence, Italy.

Enrolled in the PhD program in Engineering and Architecture, University of Trieste, Italy.

2016

Enrolled in the PhD program in Cognitive Computational and Social Neuroscience, IMT School for Advanced Studies Lucca, Italy.

Publications

2016

Papale, P., *Chiesi, L., Rampinini, A. C., Pietrini, P., and Ricciardi, E. (2016). When neuroscience 'touches' architecture: from hapticity to a supramodal functioning of the human brain. Frontiers in psychology, 7, 866.*

2017

Handjaras, G., Leo, A., Cecchetti, L., ***Papale, P.****, Lenci, A., Marotta, G., ... and Ricciardi, E. (2017). Modality-independent encoding of individual concepts in the left parietal cortex. Neuropsychologia, 105, 39-49.*

2018

Papale, P.*, Leo, A., Cecchetti, L., Handjaras, G., Kay, K. N., Pietrini, P., & Ricciardi, E. (2018). Foreground-background segmentation*

revealed during natural image viewing. eneuro, 5(3).

*Benuzzi, F., Ballotta, D., Handjaras, G., Leo, A., **Papale, P.**, Zucchelli, M., ... & Sartori, G. (2018). Eight Weddings and Six Funerals: An fMRI Study on Autobiographical Memories. Frontiers in behavioral neuroscience, 12.*

2019

***Papale, P.**, Betta, M., Handjaras, G., Malfatti, G., Cecchetti, L., Rampinini, A., ... & Leo, A. (2019). Common spatiotemporal processing of visual features shapes object representation. Scientific reports, 9(1), 7601.*

Oral presentations

2018

Fast concurrent processing of object shape and category in posterior MEG sensors
at: ECVP 2018 – European conference on visual perception, Trieste, Italy

Object coding in the ventral stream is sparse and distributed
at: SIPF 2018 – Italian society of cognitive neuroscience, Turin, Italy

2019

Neural mechanisms of object segmentation in natural vision
at: AVM 2019 – Amsterdam Vision Meeting, Amsterdam, The Netherlands

Awards

2017

Perceptual salience controls natural contour coding of V1 neurons
Best poster presentation at SIPF 2017 (shared), Rome, Italy

Abstract

Object perception relies on intensive processing in the human occitotemporal cortex (OTC). While its large-scale pattern of object selectivity has been widely described, looking at the computational processes controlling the spatial organization of OTC has proven challenging, since visual dimensions are mutually correlated (e.g., object shape and identity). In the present thesis, we investigated how different object properties that are relevant to behavior and share common variance are represented in our visual cortex in natural vision.

In Chapter 1, we described how our exceptionally reliable visual system transforms continuous retinal signaling into meaningful objects. In addition, we demonstrated how this process is challenging and complex, given the unreliability of the retinal input and the widespread mutual correlations between behaviorally relevant object properties (e.g., shape and semantic category).

In Chapter 2, we described an fMRI study on scene segmentation. In this first study, we analyzed brain responses during passive natural image viewing. Subjects attended to hundreds of natural scenes and we derived brain representations from each occipital region and compared them to parametric representations, so to reveal the inner filtering operated by each brain region.

In contrast to strictly hierarchical and compartmentalized views on brain selectivity, the whole occipital lobe is involved in the high-level cognitive task of segmenting foreground and background, as early as V1. At the same time, contrast and spatial frequencies are represented also in higher visual regions such as V4 and LOC.

However, due to the low temporal resolution of fMRI, the first study alone cannot resolve if those shared representations reflect a common spatiotemporal process. Thus, in a second MEG study, presented in Chapter 3, we derived brain representations in time and space of subjects attending to

different objects. We compared these representations to model-derived representations comprising V1-like features, object shape and semantic category. We also employed a statistical approach that compute the relative weight (i.e., orthogonal component) of each model in explaining the MEG representations.

By doing so, we found that a small cluster of posterior sensors independently processes all the tested features as early as 100-150ms after stimulus onset. Thus, the same features can be retrieved in the activity of multiple regions, and orthogonal components of those features are processed by the same cortical structures at the same latencies. Thus, is there a broader organization determining these observations? What is the link between coding of mutual and orthogonal object representations?

In our third study, presented in Chapter 4, we employed fMRI to explore the spatial organization of sensitivity to mutual and orthogonal representations in the human visual cortex. Subjects attended to object pictures while performing an unrelated attentive task. By employing a variance partitioning method, we found that the weight of mutual representations increases along the visual hierarchy, from posterior to anterior regions.

Overall, these results depict a complex picture of our visual cortex. First, there is not a clear selectivity hierarchy, but information spreads between regions: early cortical areas access to high-level representations and are involved in complex cognitive tasks (i.e., object segmentation). Second, concurrent processing of orthogonal object shape, contrast and category representations is fast (100-150ms) in the right posterior brain. And, third, the visual cortex encodes mutual relations between different features in a topographic fashion while object shape is encoded along different dimensions, each representing orthogonal features.

Chapter 1

Background

The human visual system

Our brain continuously processes the information coming from our eyes, extracting knowledge about the objects and events around us. While we perceive effortlessly, vision relies on intensive and complex computations we are not aware of.

Most of the complexity of vision arises from the heterogeneity and uncertainty of visual inputs: there is no rule of thumb to detect the relevant points of an image. To illustrate, suppose to write a simple computer algorithm aimed at identifying the objects in a scene, a process known as *segmentation*. As the information that is captured from our eyes is local luminance and color, we would start by trying to detect the edges of the image, as defined by the local difference in luminance or color (i.e., contrast). This is indeed one of the first processes occurring in our visual system.

However, consider the picture in Figure 1.1, depicting a lamb on rocks. The patch enclosed in the orange box is centered on the border between the lamb's head and the rocks on the background, while the blue box is centered on the shadow of the rocks. If we observe the two patches in isolation (right inset), the blue patch has a higher contrast with respect to the orange one. Likely, however, most of us would simply neglect the presence of the high-contrast edge in the blue box and would pay attention to the lamb only. Consequently, our simple program would fail to discriminate which is the most relevant patch between orange and blue, and in general, would not capture the essence of what we perceive as salient when looking at that image.

Thus, how can our visual system cope with the uncertainty of its most basic source information in such a reliable and seemingly effortless way?

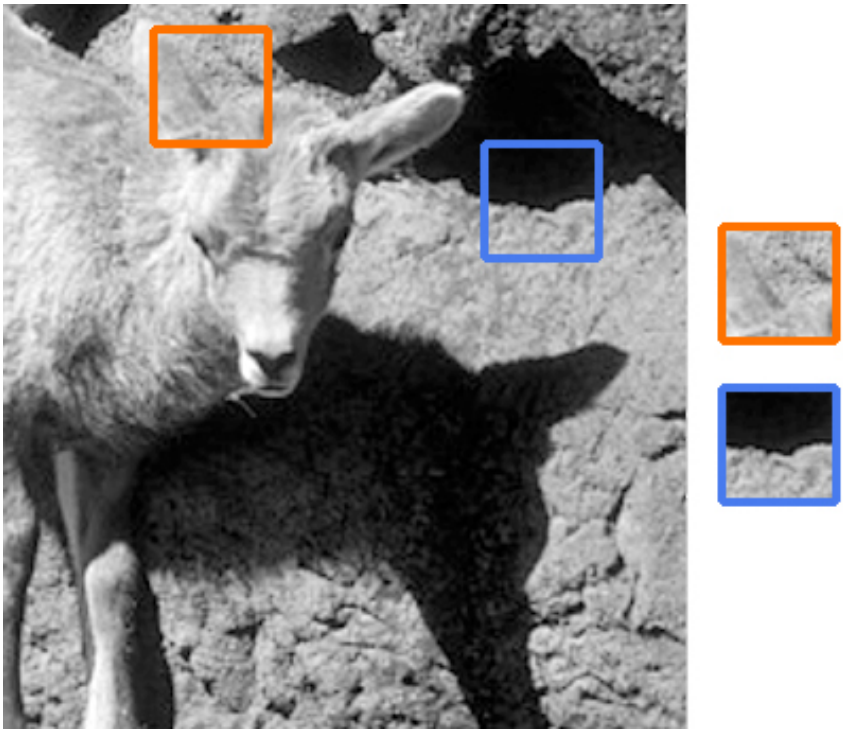


Figure 1.1 | Uncertainty in object segmentation
The boxes identify two patches of the image centered on either a salient (orange) or not salient (blue) location, having different luminance properties. A simple program aimed at making sense of the image content, inspired by low-level brain representations, would probably look for contrast defined edges in the image. However, as evident, contrast is not a good predictor of what is relevant to our perception.

Visual pathways: from retina to cortex

The complexity of this computational task is mirrored in its physical implementation in our brain, as vision requires a huge portion of gray matter. Visual processing is hierarchical,

comprising several anatomical structures and the white matter tracts linking them (Figure 2.1). And each of those structures operates a transformation of the image.

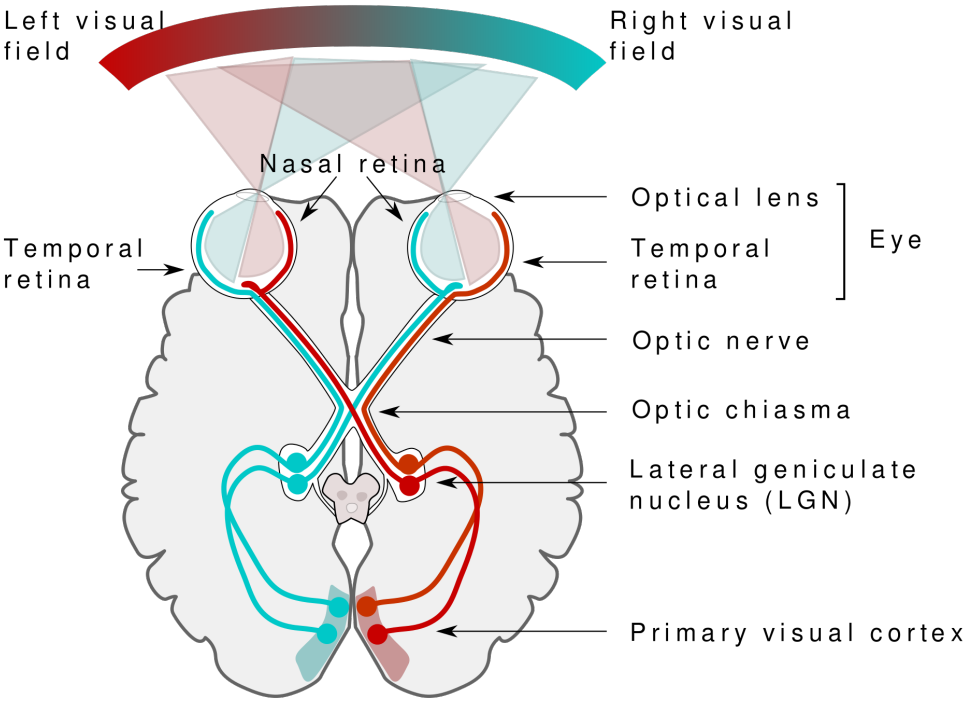


Figure 1.2 | Visual signaling from retina to cortex
Local visual information caught by the retina in the eye is then pushed to the cortex after an intermediate stage in the lateral geniculate nucleus (LGN). Image: Miquel Perello Nieto [CC BY-SA 4.0].

Cornea and lens perform the first optical transformations, adjusting the retinal image that is later captured by the photoreceptors, the light-sensitive cells of the retina. Our ability

to see is ultimately determined and constrained by the properties and arrangement of photoreceptors (Figure 1.3).

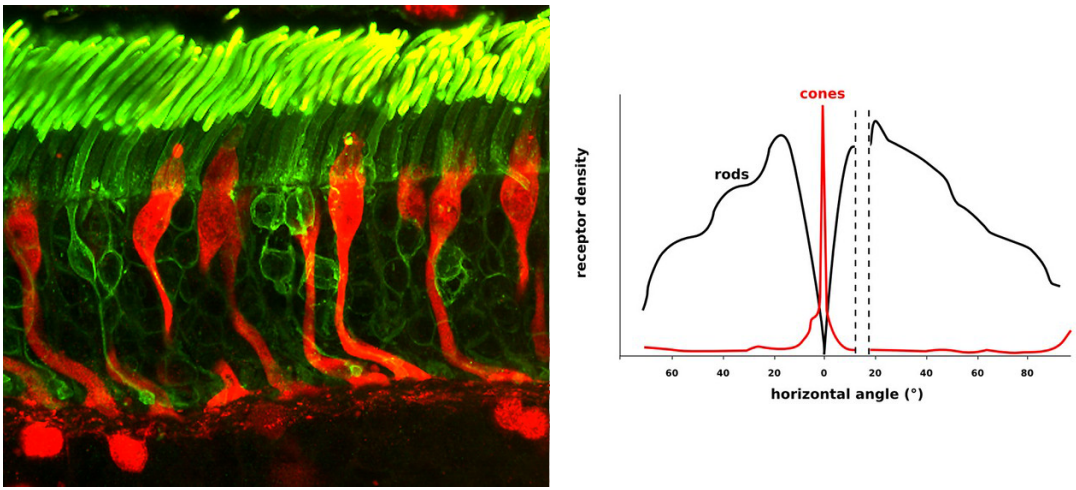


Figure 1.3 | Photoreceptors in the human retina
Left: rods (green) and cones (red) in the human retina. Confocal microscope image by Robert Fariss, National Eye Institute, NIH [CC BY 2.0].
Right: rods and cones density. Image: Johannes Ahlmann [CC BY 2.0].

The two types of photoreceptor cells are rods and cones. They differ in their number, spatial arrangement and encoding role. There are 20 times more rods than cones in the retina (about 5 vs. 100 million in each eye). Cones are concentrated in the fovea, a small portion of the retina with highest visual acuity, comprising no rods. And even if rods are more sensitive to light than cones, signaling from many rods usually reaches a single neuron, while several neurons collect the responses of foveal cones. Thus, while rods mainly improve sensitivity, information coming from the cones encodes finer spatial resolution. Furthermore, cones are also responsible for color vision, since they may have three possible spectral sensitivities (L-, M- and S- cones).

All the information received from photoreceptors streams through retinal ganglion cells, whose axons form the optic nerve, to the lateral geniculate nucleus (LGN), a sub-cortical relay station located in the thalamus, and then hits the cortex in layer IV of the primary visual area (V1). In the process, several transformations occur. First, as in our previous example, retinal neurons compute local contrast from light intensity. Then, information is relayed by LGN (about 1 million of cells) and finally decomposed along orientation selectivity columns in V1 (about 100 million of cells: Figure 1.4).

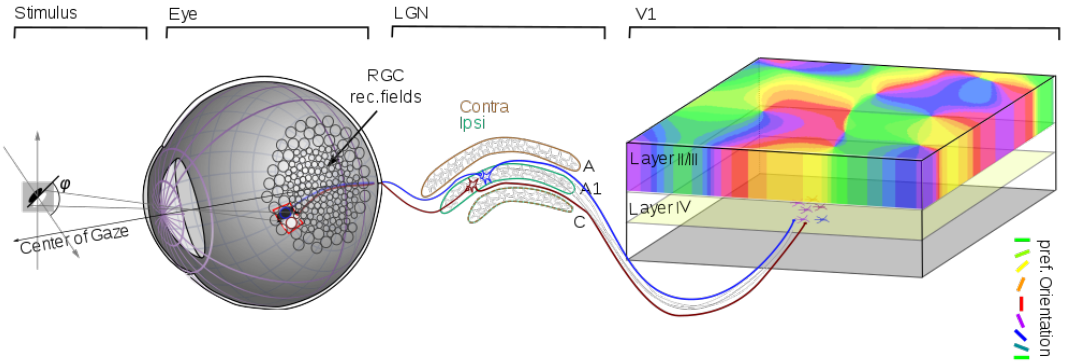


Figure 1.4 | Visual transformations from retina to cortex
Ganglion cells in the retina target neurons in LGN (optic nerve). From there, through the optic radiation, visual information reaches layer IV in V1, where several neurons receive signaling from the same LGN neuron. Image adapted from: (Schottdorf et al., 2015)[CC BY-SA 4.0].

The 2D spatial selectivity of neurons, known as the classical *receptive field* (RF), is central to this process: each region maps the visual field in an organized fashion, and neurons from each stage connect with neurons with similar RFs in downstream regions. At first, information from rods and cones is compressed by retinal neurons (but note that signaling from foveal cones is

likely over-represented: Wassle et al., 1990), which develop a center-surround RF organization and project to LGN and V1 neurons in layer IV with matching RF properties. Then, more superficial layers in V1 build orientation selectivity on top of this center-surround representation.

The primary visual cortex

V1 organization and coding properties have been studied extensively. At the micro-scale, V1 has a high neuronal density (Rockel et al., 1980), and its neurons are organized along maps of ocular dominance and orientation selectivity (i.e., the typical pinwheel-like pattern depicted in Figure 1.4). At the macro-scale, V1 is retinotopically arranged, meaning that V1 maps the visual field in a topographic fashion (Figure 1.5B). Other downstream regions are retinotopically organized too.

These maps mirror the shape of V1 neurons RF, thus determining their stimulus selectivity. Signaling from LGN neurons is sampled so to detect oriented edges (Figure 1.5A), centered at fixed positions in the visual field (i.e., the classical RF) and with a specific spatial frequency. Moreover, V1 neurons are also selective for the direction of the stimulus, the color and its temporal frequency (Carandini, 2012).

V1 neurons can be classified as either simple or complex cells depending on their RF. The latter are sensitive to oriented edges with specific spatial frequencies as simple cells, but are insensitive to the phase of the edge (Figure 1.5C).

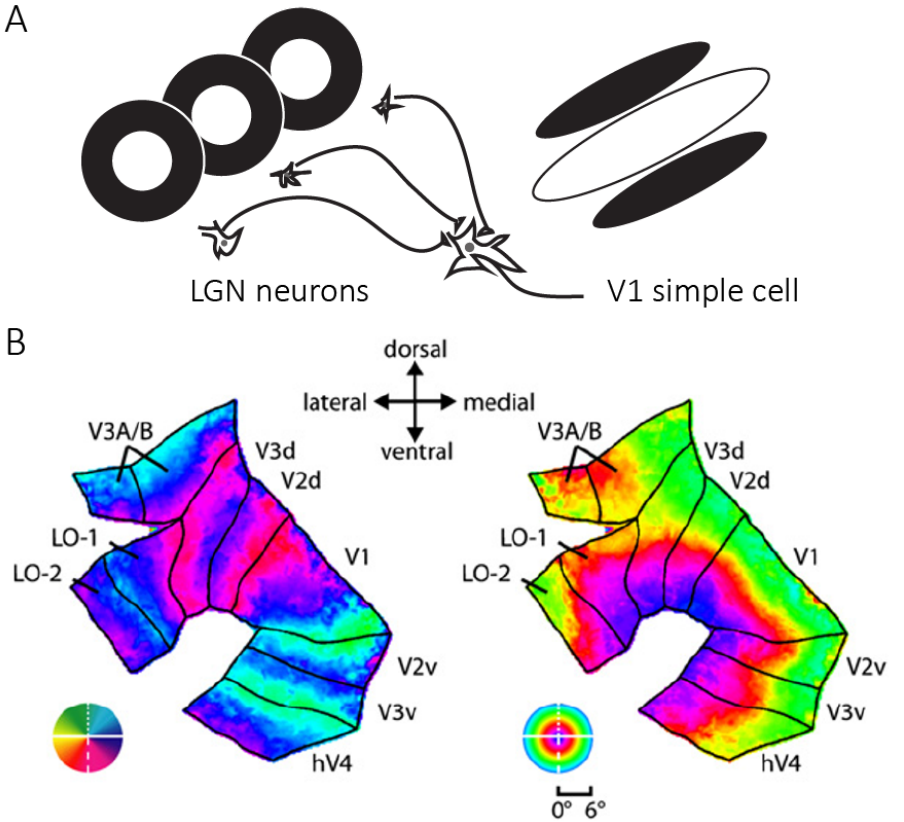

A
LGN neurons
V1 simple cell
B
dorsal
lateral
medial
ventral
V3A/B
V3d
V2d
LO-1
LO-2
V1
V2v
V3v
hV4
0°
6°
C
D
# of neurons
100
200
500
1000
2000
5000
10000

Figure 1.5 | V1 organization and coding properties

A) The relationship between LGN neurons and V1 simple cells is constrained by their relative RFs.

B) Retinotopic organization in the human visual system. Image adapted from: (Larsson and Heeger, 2006) [CC BY 3.0].

C) Complex cell selectivity is insensitive to phase.

D) Image reconstructions using 100 to 10,000 V1 simulated neurons.

Of note, there are between 25 (Olshausen, 2003) and 100 times (Leuba and Kraftsik, 1994) more V1 neurons than retinal ganglion cells, and around 100 times less ganglion cells than rods and cones (Curcio and Allen, 1990; Schottdorf et al., 2015). This organization favors the hypothesis of a compression-transmission-expansion strategy (Babadi and Sompolinsky, 2014). This is thought to arise from evolutionary adaptation. Actually, a popular theory claims that V1 optimally encodes the regularities in the environment, as reflected in the statistical structure of natural images (Olshausen and Field, 1996; Vinje and Gallant, 2000). Indeed, it has been shown that V1 can reliably represent the retinal input with only a few neurons active in any given moment. Figure 1.5D shows how an image can be reliably reconstructed summing a few V1-like edges.

Overall, from this rough and simplified description of the first stages of the visual system one may start thinking at it as a cascade of fixed bottom-up transformations. However, V1 neuronal representation and selectivity are significantly shaped by signaling from nearby neurons (horizontal connections) and from neurons in downstream regions (top-down interactions). For instance, neurons with different orientation tuning but overlapping classical RF inhibit each other (a mechanism known as divisive normalization), thus increasing the response gain (Carandini and Heeger, 2011). Also, top-down mechanisms control V1 activity: for instance, signaling from higher-level regions increases the activity of V1 neurons when co-linear flankers are placed outside their canonical RF (Li et al., 2006) or when they are placed in a portion of the image perceived as a figure with respect to the ground (Lamme, 1995; Poort et al., 2012).

Cortical visual pathways: ventral and dorsal stream

Several cortical visual areas compose the occipital lobe, in the posterior part of the brain. As all of them are retinotopically organized, they are identifiable by looking at boundaries between visual field maps (Figure 1.5C). Those cortical visual regions are named after their position with respect to V1. V1 neurons project to V2 neurons, and them to V3 and V4 neurons. Later stages are the lateral occipital complex (LOC), and areas V3A and V3B.

The visual cortex also extends beyond the occipital lobe to temporal and parietal lobes (Figure 1.6). And even frontal regions are sensitive to visual information (Squire et al., 2012). A popular view, the two-stream hypothesis (Goodale and Milner, 1992), considers temporal and parietal regions to be specialized for different tasks. The temporal lobe (i.e., ventral stream) processes object category, while the parietal cortex (i.e. dorsal stream) is tuned to object position and actions.

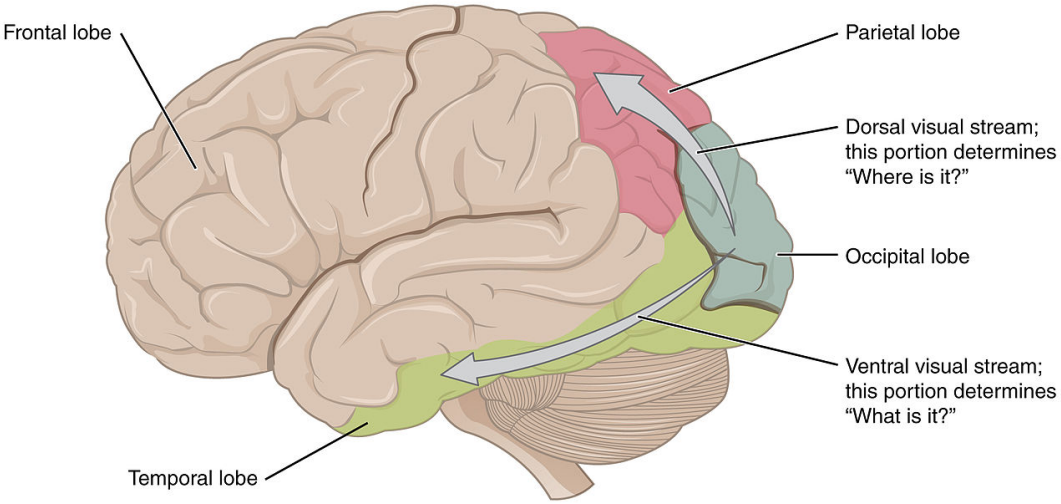


Figure 1.6 | Ventral and dorsal stream

The two-stream hypothesis suggests that after V1, information flows both to dorsal regions –where information about objects position is processed – and ventral areas – where information about objects identity is computed. Image: OpenStax College [CC BY 3.0].

The sensitivity of neurons in each region is also different. The RF shape increases moving across the visual hierarchy. And neurons in higher regions develop category-selectivity and viewpoint invariance (e.g., Zoccolan et al., 2007; Figure 1.7). V2 neurons are tuned to simple textures and combinations of edges (Freeman et al., 2013), while V4 and neurons are selective for local shape properties and complex textures (Cadieu et al., 2007; Okazawa et al., 2015). Object global shape modulates LOC activity, while inferotemporal (IT) neurons are tuned to specific semantic categories (Kriegeskorte et al., 2008b).

The increasing sensitivity to complex features has been interpreted as the result of a feedforward hierarchy (Riesenhuber and Poggio, 2002), as illustrated in the example model in Figure 1.7A. This looks a rather complete and satisfying picture of our visual system. However, if we look back at the image of the lamb in Figure 1.1, we would spot a clear problem in this strict feedforward description: its outputs are as reliable as its first module. An edge that goes unseen by V1, will not eventually reach V2 and the other downstream modules. Many visual tasks are not compatible with this description, especially object segmentation and shape processing.

Indeed, the amount of feedforward and feedback projections between higher regions and the visual cortex in non-human primates are comparable (Markov et al., 2011; Roelfsema and de Lange, 2016). And already in LGN, only a small portion (about 10%) of connections comes from the retina while 60% of them originate in the cortex (Wandell, 1995).

Overall, mounting evidence shows that these interactions between early and late visual areas have a causal role in determining high-level cognitive processes as discriminating between what is figure and what is ground (Roelfsema and de Lange, 2016).

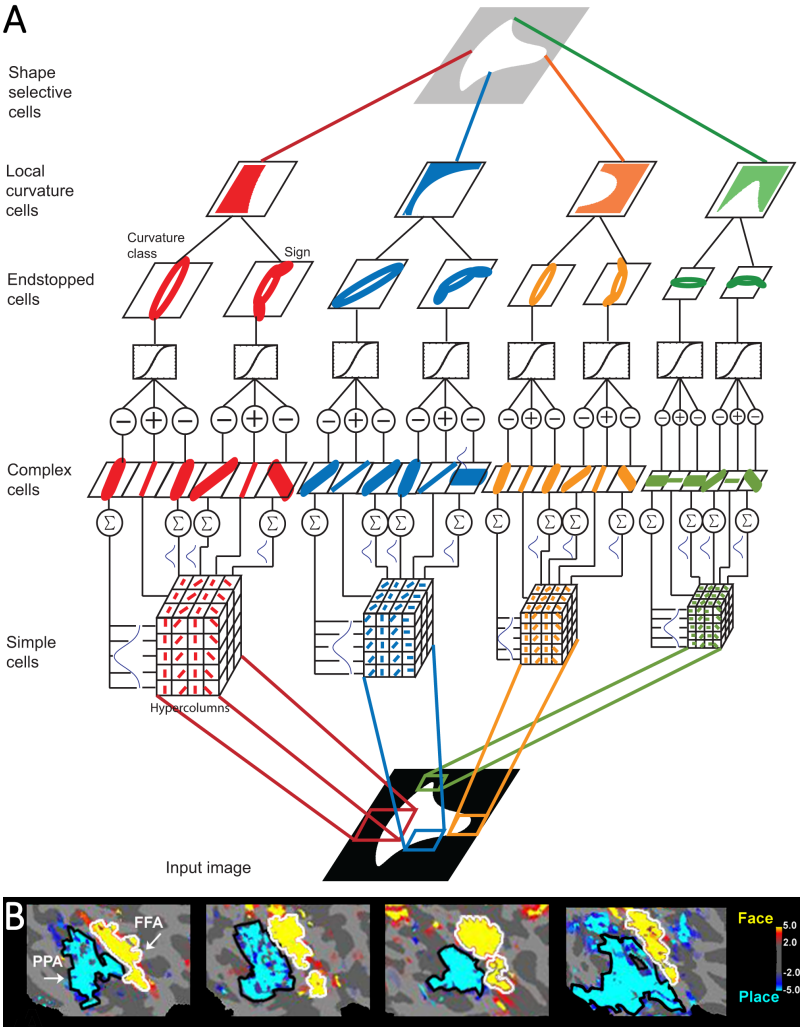


Figure 1.7 | Visual hierarchy and selectivity

A) A hierarchical model of neural selectivity. Each level samples information from a lower region, developing sensitivity to more global and complex features. Adapted from: (Rodriguez-Sanchez and Tsotsos, 2012) [CC BY 4.0]

B) Images of faces and places activate different cortical regions in

the temporal lobe. This organization is stable across subjects. Adapted from: (Rajimehr et al., 2011) [CC BY 4.0]

Modeling brain representations in natural vision

Downstream and upstream approaches

Most of what we know about the visual system has been discovered by carefully designing artificial stimuli aimed at maximizing the activity of a single neuron or a brain region in response to a specific variable of interest (Felsen and Dan, 2005; Kay, 2011). This approach has proven undoubtedly useful, but has likely contributed to a compartmentalized and strictly hierarchical view of object processing.

As a result, after decades of research, the visual system is still far from being fully understood. This is reflected in our current (in)ability to predict neural activity: the best model of V1 responses to natural images can hardly account for 50% of the explainable variance (Cadena et al., 2019). In this light, we may argue that we are halfway from knowing what V1 does in natural vision, and the same holds for higher-level regions.

Actually, the study of brain functioning in ecological conditions has to face the challenging methodological limitation of variable collinearity (Kay, 2011). To clarify, let's suppose to be interested in studying shape processing in area V4 with natural stimuli. The outline of objects pertaining to the same semantic category are always more similar to items of the same category rather than to objects belonging to different categories. For instance, trees have a typical shape, with thin branches on top of a thick trunk, while graspable objects often have an elongated central element. Thus, how would you resolve if a neuron is truly selective for elongated shapes in natural images and does not merely respond to pictures of graspable objects?

There are two possible ways to deal with collinearity in natural vision: an *upstream* approach that controls for collinearity before data acquisition, and a *downstream* approach that mitigates its effects after data recording.

The upstream approach is any different from conventional experimental design. However, if designing a set of artificial stimuli that varies along a dimension of interest, but is constant

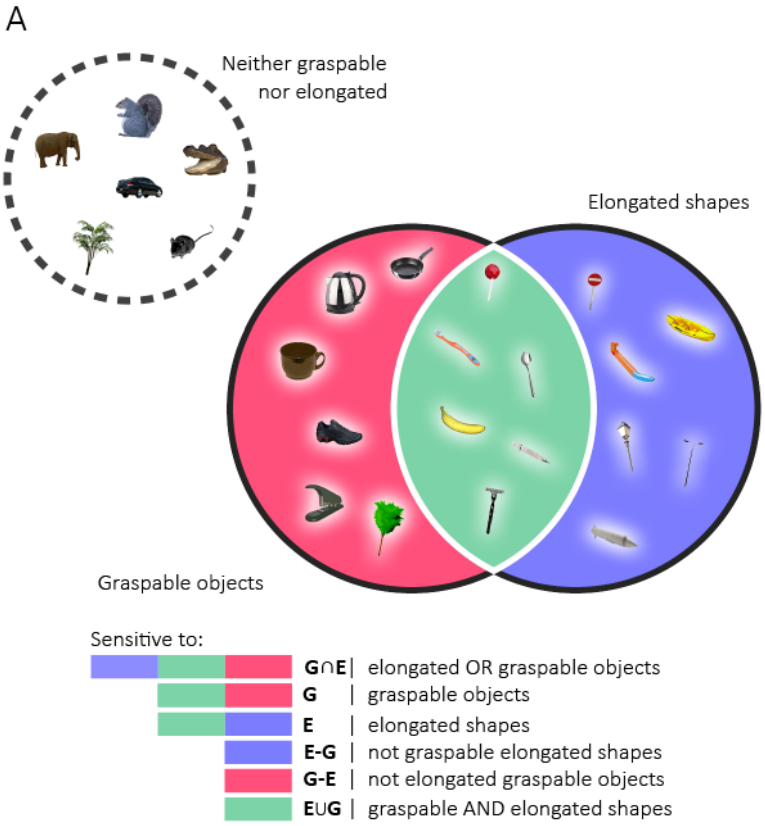
along all possible confounding dimensions, is an intellectually demanding process, when it comes to natural vision the set of controlled stimuli must also capture the statistical structure of natural images and their wide variance, to be ecologically valid. To exemplify, Freeman and Simoncelli (2011) investigated how information is represented in the periphery of the visual field using synthetic textures matching the local organization of real images. Similarly, Neri (2017) studied the influence of object segmentation on image reconstruction by grafting into natural images a fully controlled small patch of oriented elements.

Instead, the downstream approach tests competing feature vectors and looks at which one is a better predictor of brain responses to intact natural images (Wu et al., 2006). More recently, several statistical methods have been introduced to control for collinearity, considering the relative impact of each variable of interest. Three examples of the downstream approach are presented in the following chapters.

Overall, both methods have strengths and weaknesses. On one hand, the upstream approach offers a full control over the distributions of the variables of interest, while the same is not true for the downstream procedure. In fact, even when using the largest possible set of stimuli, different variables will likely have different variance. This is particularly true when looking for high-level dimensions (e.g., emotions: Lettieri et al., 2018). On the other hand, if both upstream and downstream approaches can only control for what is known to represent a potential confound, in the downstream approach it is still possible to perform a *post hoc* control analysis as far as one become aware of visual dimensions collinear with the variables of interest.

How does the brain face collinearity?

In the last few pages, we mentioned the role of feedback projections and stressed the importance of natural stimuli to overcome compartmentalized descriptions of neural selectivity. However, when considering the collinearity of visual attributes, another broader question emerges: how does the brain face the mutual variance between properties that are relevant to us? Do neurons or cortical regions encode more than a single dimension?


A
Neither graspable
nor elongated
Elongated shapes
Graspable objects
Sensitive to:
G∩E | elongated OR graspable objects
G | graspable objects
E | elongated shapes
E-G | not graspable elongated shapes
G-E | not elongated graspable objects
E∪G | graspable AND elongated shapes

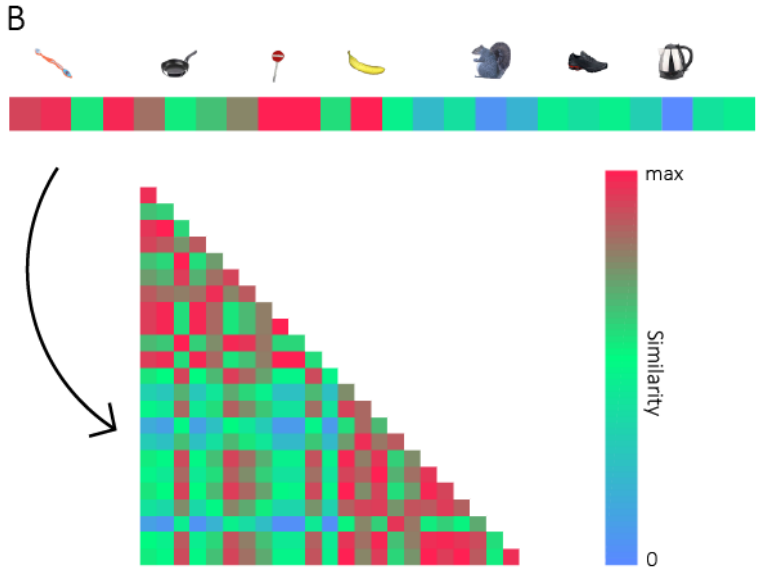

B
max
Similarity
0

Figure 1.8 | Coding of object representations
A) A brain area may show sensitivity to different combinations of features (see text). Object pictures from: (Wilf et al., 2013) [CC BY 3.0].
B) To make quantitative comparisons of feature vectors differing in type (e.g., discrete or continuous) or dimensionality, it is possible to abstract away and compare their representational structure, i.e. their (dis)similarity matrices.

Consider the example in Figure 1.8A. Many possible combinations of selectivity may exist. A brain area may be sensitive to object graspability and also to the extent to which its shape is elongated. That region would respond as strongly to both objects that are known to be graspable, but do not have an elongated shape, and to elongated shapes that are not graspable objects. On the other hand, different components of selectivity may occur: e.g., a region that is selective for graspable object with elongated shapes, or to elongated shapes that are not graspable objects.

The present work explores exactly this subject: are object attributes that are relevant to our behavior processed by the same regions and/or at the same temporal latencies? What is the link between orthogonal and mutual features?

Tackling these questions requires to find a common ground between different featural descriptions, ranging from low-level properties, as V1-like filter responses, to categorical variables, as the semantic class of the objects in a scene. In this light, a handy methodological solution is to abstract away from featural dimensions and to rely on representational geometries (Kriegeskorte et al., 2008a). This approach makes possible quantitative comparisons between units of different model and brain-activity (Figure 1.8B).

In the following chapters, we present three different studies, employing a downstream approach to model cortical representations in humans. In chapters 2 and 4, we present two fMRI (functional magnetic resonance imaging) studies. The fMRI is an *in vivo* technique that measures differences in the BOLD (blood oxygen level–dependent) signal, sampling synaptic activity of thousands of neurons at once. It has a high spatial resolution (up to the sub-millimeter scale) but a low temporal

resolution (up to few hundreds of millisecond). In the study described in Chapter 3, we employed the MEG (magnetoencephalography) technique. The MEG has a better temporal resolution (up to the millisecond scale) but a lower spatial resolution (up to 306 channels) than fMRI. It is sensitive to small differences in the magnetic field determined by the electrical activity of the brain.

References

Babadi B, Sompolinsky H (2014) Sparseness and expansion in sensory representations. Neuron 83:1213-1226.

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, Ecker AS (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. PLOS Comput Biol 15:e1006897.

Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T (2007) A model of V4 shape selectivity and invariance. J Neurophysiol 98:1733-1750.

Carandini M (2012) Area V1. Scholarpedia 7:12105.

Carandini M, Heeger DJ (2011) Normalization as a canonical neural computation. Nat Rev Neurosci 13:51-62.

Curcio CA, Allen KA (1990) Topography of ganglion cells in human retina. J Comp Neurol 300:5-25.

Felsen G, Dan Y (2005) A natural approach to studying vision. Nat Neurosci 8:1643-1646.

Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nat Neurosci 14:1195-1201.

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA (2013) A functional and perceptual signature of the second visual area in primates. Nat Neurosci 16:974-981.

Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends in neurosciences 15:20-25.

Kay KN (2011) Understanding visual representation by developing receptive-field models. Visual population codes: Towards a common multivariate framework for cell recording and functional imaging:133-162.

Kriegeskorte N, Mur M, Bandettini P (2008a) Representational similarity analysis–connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008b) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.
Lamme VA (1995) The neurophysiology of figure-ground segregation in primary visual cortex. J Neurosci 15:1605-1615.
Larsson J, Heeger DJ (2006) Two retinotopic visual areas in human lateral occipital cortex. J Neurosci 26:13128-13142.
Lettieri G, Handjaras G, Ricciardi E, Leo A, Papale P, Betta M, Pietrini P, Cecchetti L (2018) Emotionotopy: Gradients encode emotion dimensions in right temporo-parietal territories. bioRxiv:463166.
Leuba G, Kraftsik R (1994) Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. Anat Embryol (Berl) 190:351-366.
Li W, Piech V, Gilbert CD (2006) Contour saliency in primary visual cortex. Neuron 50:951-962.
Markov NT, Misery P, Falchier A, Lamy C, Vezoli J, Quilodran R, Gariel MA, Giroud P, Ercsey-Ravasz M, Pilaz LJ, Huissoud C, Barone P, Dehay C, Toroczkai Z, Van Essen DC, Kennedy H, Knoblauch K (2011) Weight consistency specifies regularities of macaque cortical networks. Cereb Cortex 21:1254-1272.
Neri P (2017) Object segmentation controls image reconstruction from natural scenes. PLoS Biol 15:e1002611.
Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci U S A 112:E351-360.
Olshausen BA (2003) Principles of image representation in visual cortex. The visual neurosciences 2:1603-1615.
Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607-609.
Papale P, Leo A, Cecchetti L, Handjaras G, Kay KN, Pietrini P, Ricciardi E (2018) Foreground-Background Segmentation Revealed during Natural Image Viewing. eNeuro 5.
Papale P, Betta M, Handjaras G, Malfatti G, Cecchetti L, Rampinini A, Pietrini P, Ricciardi E, Turella L, Leo A (2019)

Common spatiotemporal processing of visual features shapes object representation. Sci Rep 9:7601.
Poort J, Raudies F, Wannig A, Lamme VA, Neumann H, Roelfsema PR (2012) The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. Neuron 75:143-156.
Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH (2011) The "Parahippocampal Place Area" Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. PLoS Biol 9:e1000608.
Riesenhuber M, Poggio T (2002) Neural mechanisms of object recognition. Current Opinion in Neurobiology 12:162-168.
Rockel AJ, Hiorns RW, Powell TP (1980) The basic uniformity in structure of the neocortex. Brain 103:221-244.
Rodriguez-Sanchez AJ, Tsotsos JK (2012) The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. PLoS ONE 7:e42058.
Roelfsema PR, de Lange FP (2016) Early Visual Cortex as a Multiscale Cognitive Blackboard. Annu Rev Vis Sci 2:131-151.
Schottdorf M, Keil W, Coppola D, White LE, Wolf F (2015) Random Wiring, Ganglion Cell Mosaics, and the Functional Architecture of the Visual Cortex. PLOS Comput Biol 11:e1004602.
Squire RF, Steinmetz NA, Moore T (2012) Frontal eye field. Scholarpedia 7:5341.
Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287:1273-1276.
Wandell B (1995) Foundations of vision.
Wassle H, Grunert U, Rohrenbeck J, Boycott BB (1990) Retinal ganglion cell density and cortical magnification factor in the primate. Vision Res 30:1897-1911.
Wilf M, Holmes NP, Schwartz I, Makin TR (2013) Dissociating between object affordances and spatial compatibility effects using early response components. Front Psychol 4:591.
Wu MC, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29:477-505.

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ (2007) Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci 27:12292-12307.

Chapter 2

Foreground-background segmentation revealed during natural image viewing

Introduction

In the scientific journey toward a satisfying understanding of the human visual system, scene segmentation represents a central problem "for which no theoretical solution exists" (Wu et al., 2006). Segmentation into foreground and background is crucial to make sense of the surrounding visual environment, and its pivotal role as an initial step of visual content identification has long been theorized (Biederman, 1987). Indeed, according to Fowlkes and colleagues (2007), humans can produce consistent segmentations of natural images. However, even though more recent approaches based on deep convolutional networks produced promising results (He et al., 2017), both the computational and neurophysiological processes that underlie scene segmentation are still a matter of debate.

To date, numerous studies found evidence of texture segmentation and figure-ground organization in the early visual cortex of nonhuman primates (Lamme, 1995; Lee et al., 1998; Poort et al., 2012; Self et al., 2013) and humans (Kastner et al., 2000; Scholte et al., 2008; Kok and de Lange, 2014). It has been showed that the identification of salient visual attributes arises from a region-filling mechanism, that targets neural populations mapping relevant points in space (Roelfsema, 2006). In particular, a recent study on monkeys attending artificial stimuli revealed an early enhancement of V1 and V4 neurons when their receptive fields covered the foreground, and a later response suppression when their receptive fields were located in the stimulus background (Poort et al., 2016) – extending results from a previous study (Lamme et al., 1999). Thus, the primate brain groups together image elements which belong to the figure, showing an enhanced activity for the foreground and a concurrent suppression of the background.

However, from an experimental viewpoint, the role of figure-ground segmentation has primarily been demonstrated by means of non-ecological stimuli (e.g., binary figures, random dots, oriented line segments and textures). It should be noted that previous reports demonstrated how models of brain responses to artificial stimuli are suboptimal in predicting responses to natural images (David et al., 2004; Felsen and Dan, 2005). Although two recent studies investigated border-

ownership in monkeys with both artificial and natural stimuli (Hesse and Tsao, 2016; Williford and von der Heydt, 2016), a proof of the occurrence of foreground-background segmentation in the human brain during visual processing of naturalistic stimuli (e.g., natural images and movies) is still lacking. This pushes towards the development of novel methods specifically designed for testing segmentation in ecological conditions.

In light of this, we investigated foreground enhancement and background suppression, as specific processes involved in scene segmentation during passive viewing of natural images. We used fMRI data, previously published by Kay and colleagues (Kay et al., 2008), to study brain activity patterns from seven visual regions of interest (ROIs): V1, V2, V3, V3A, V3B, V4 and lateral occipital complex (LOC) in response to 334 natural images, whose "ground-truth" segmented counterparts have been included in the Berkeley Segmentation Dataset (BSD) (Arbelaez et al., 2011).

To this aim, we developed a novel pre-filtering modeling approach to study brain responses to complex, natural images without relying on explicit models of scene segmentation, and adopting a validated and biologically plausible description of activity in visual cortices. Our method is similar to other approaches where explicit computations are performed on representational features, rather than on the original stimuli (Naselaris et al., 2011). For instance, these methods have been recently used to investigate semantic representation (e.g. Huth et al., 2012; Handjaras et al., 2017) or boundary and surface-related features (Lescroart et al., 2016). However, as opposed to the standard modeling framework – according to which alternative models are computed from the stimuli to predict brain responses – here, low-level features of the stimuli are parametrically modulated and simple descriptors of each filtered image (i.e., edges position, size and orientation) are aggregated in a fixed model (Figure 2.1). The correspondence between the fixed model and fMRI representational geometry related to intact images, was then evaluated using representational similarity analysis (RSA) (Kriegeskorte et al., 2008). Notably, this approach can also be exploited to obtain highly informative "correlation images" representing the putative computations of different brain regions

and may be generalized to investigate different phenomena in visual neuroscience.

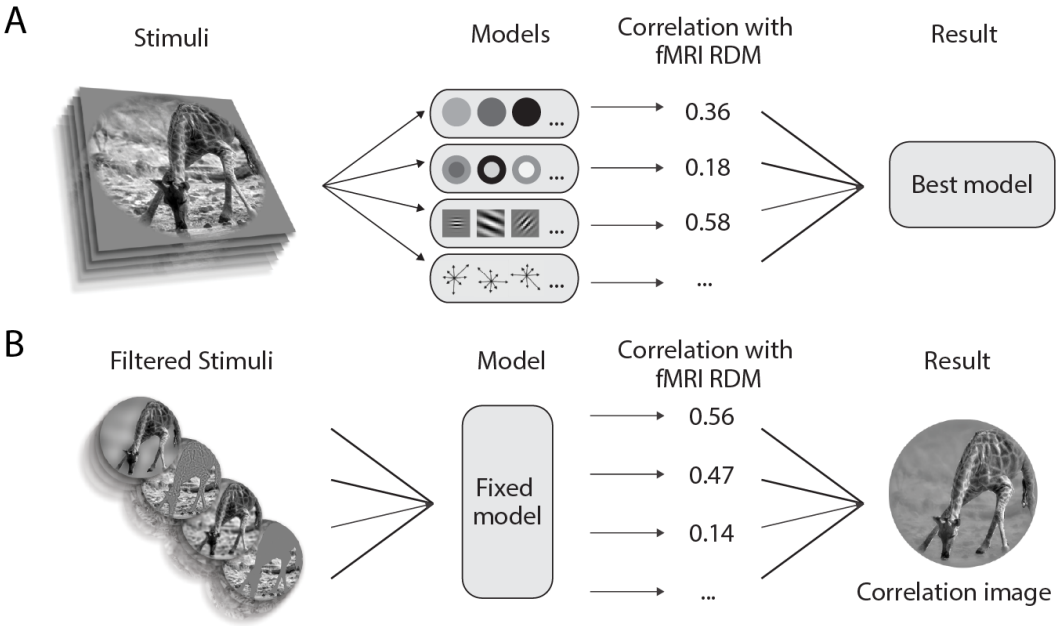


Figure 2.1 | Comparing the standard modeling approach and the pre-filtering modeling approach

A) In the standard modeling pipeline, different models are compared. After extracting features from the stimuli, competing feature vectors can be used in order to predict brain activity in an encoding procedure, whereas their dissimilarities can be used in a representational similarity analysis. Finally, the model that better predicts brain responses is discussed. B) In our pre-filtering modeling approach, different filtered versions of the original stimuli are compared. Various biologically plausible filtering procedures are applied to the stimuli prior to compute a unique feature space according to a given fixed and easily interpretable model. In our approach a single model is employed and the step showing the highest correlation with brain activity (or

representational geometry) of each filtering procedure is used to build a *post-hoc* "correlation image". While the standard modeling approach is theoretically more advantageous, as its output is a fully computable model of brain activity, it cannot be applied when reliable explicit models of perceptual processes do not exist yet, as in the case of scene segmentation. Alternative attempts to reconstruct visual stimuli from brain activity have been previously reported using multivariate techniques (e.g. Stanley et al., 1999; Thirion et al., 2006; Miyawaki et al., 2008; Nishimoto et al., 2011).

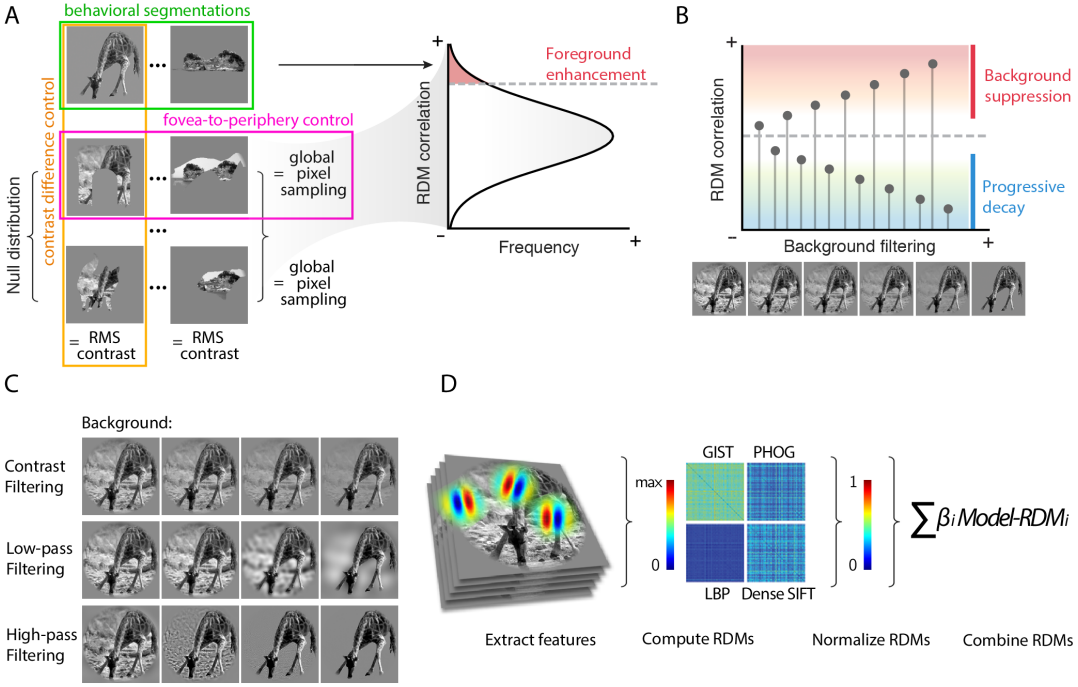


Figure 2.2 | Analytical pipeline

A) Foreground enhancement test: the set of segmented stimuli is tested against a null distribution of 1,000 permutations. Each

permutation is built by randomly shuffling the 334 behavioral foreground masks, and matching the root mean square (RMS) contrast of the behaviorally segmented counterpart. This analysis controls for size, location and contrast of the foreground when testing whether behavioral segmentations explain each ROI representational dissimilarity matrix (RDM) better than chance. B) Background suppression test: the correlation between brain RDMs and each step of the background filtering procedure is tested against the correlation determined by the intact stimuli. While information is filtered out, correlation can increase or decrease, depending on the sensitivity for background related information in each ROI. A progressive decay indicates that a region actually processes the background, while a significant increase suggests that background is suppressed C) Filtering steps for the contrast or spatial frequencies filtering. D) In clockwise order: features for each model were extracted from the stimuli; the dissimilarity (1 - Pearson's r) between each stimulus pair was computed and aggregated in four RDMs; the obtained RDMs were normalized in a 0-1 range; finally, the four RDMs were linearly combined in the fixed model, which was then correlated to the fMRI RDM obtained from each ROI.

Results

Foreground enhancement and background suppression can be tested in ecological conditions following a simple argument: when attempting to predict brain activity of a visual ROI with a specific model, the goodness-of-fit depends on the model inputs, e.g., the spatial information provided. Thus, the correlation between filtered images and fMRI representational patterns evoked by their intact counterpart can be used to verify specific hypotheses on visual processing (Figure 2.2). In this study, we posit that evidence of preferential processing (i.e., enhancement) should depend on the shape of the foreground instead of the size, the location or the contrast of the segmented region processed through the model. In this regard, a random sampling procedure of foreground segmentations across stimuli would offer a proper choice to account for all these aspects, ultimately testing whether behavioral segmentations provide a better prior

for enhancement. On the other hand, background filtering can lead either to a decay or an increase in correlation with brain representational patterns. The former indicates that background-related information is – at least to some extent – processed, whereas the latter denotes that background information is suppressed, since embedding it in the model is not different from adding noise.

Comparison of intact and behaviorally segmented images

The correlation between RDMs computed using the fMRI patterns from each of the seven visual ROIs and three descriptions of the stimuli (intact, isolated background and isolated foreground) were tested (Figure 2.3 and Table 2.1). Results show significant correlations ($p < 0.05$, Bonferroni corrected) between the intact description of images and fMRI RDMs in V1, V2 and V3. The segmented foreground RDM shows a significant correlation in V2, V4 and LOC, while the segmented background achieves significant correlations in V1 and V2 only. Of note, the correlation yielded by one of the descriptions approaches the ROI-specific SNR estimation (i.e., the maximum reachable correlation given the noise of the data), thus confirming the validity of the fixed model employed (Wu et al., 2006).

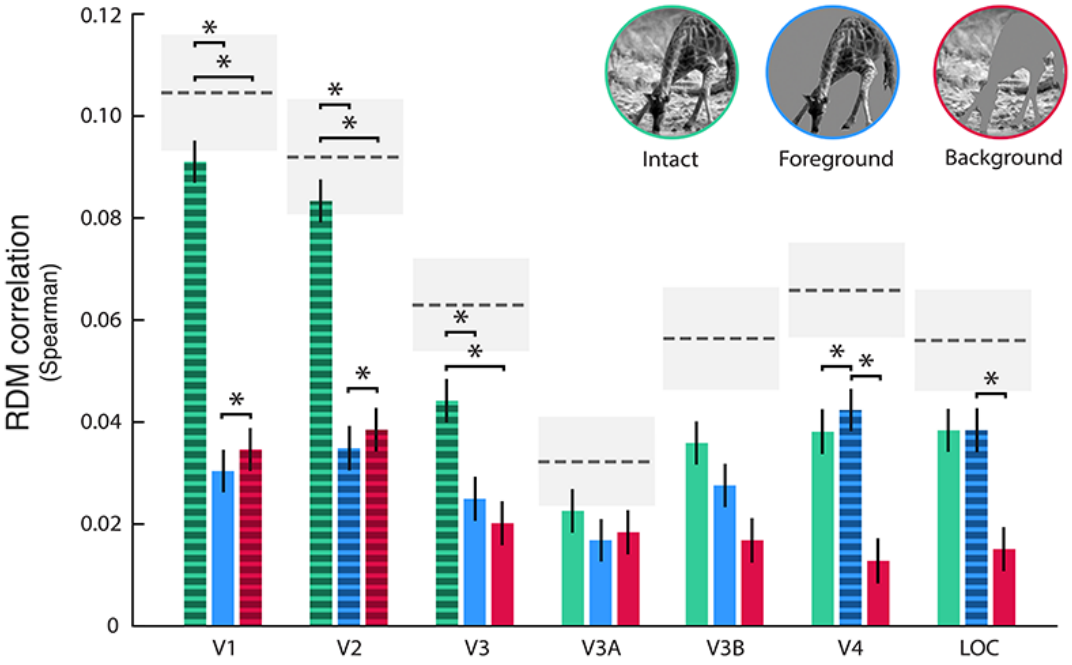


Figure 2.3 | Comparison of intact and behaviorally segmented images
The graphs show the correlation between the intact (green) and segmented versions (blue: isolated foreground; red: isolated background) of the images and brain RDMs (n = 55611). Dashed bars stand for significant correlations as resulting from the permutation test ($p < 0.05$, Bonferroni corrected; 1000 iterations). Asterisks indicate significant differences between correlation values ($p < 0.05$, Bonferroni corrected). Error bars represent the standard error estimated with bootstrapping. Dashed lines represent the SNR estimate for each ROI, while gray shaded regions indicate its standard error.

Table 2.1 | Comparison of intact and behaviorally segmented images

	Intact		Foreground		Background	
ROI	Spearman's ρ	p-value	Spearman's ρ	p-value	Spearman's ρ	p-value
V1	0.091±0.008	< 0.001*	0.03±0.008	0.006	0.035±0.008	< 0.001*
V2	0.084±0.005	< 0.001*	0.035±0.007	< 0.001*	0.039±0.004	< 0.001*
V3	0.044±0.007	< 0.001*	0.025±0.008	0.08	0.02±0.007	0.29
V3A	0.023±0.004	0.34	0.017±0.005	0.256	0.018±0.007	0.127
V3B	0.036±0.009	0.017	0.028±0.006	0.02	0.017±0.009	0.186
V4	0.038±0.015	0.027	0.043±0.006	< 0.001*	0.013±0.007	0.915
LOC	0.038±0.008	0.015	0.038±0.012	< 0.001*	0.015±0.009	0.543

* = p < 0.05 Bonferroni corrected

Foreground is enhanced in all the tested regions

We tested whether the behavioral foreground segmentation from BSD represented a better predictor of RDMs derived from fMRI activity, as compared to alternate configurations obtained by shuffling the segmentation patterns across stimuli (Figure 2.2A). The correct foreground configuration yielded a significantly higher correlation as compared to the examples from the shuffled dataset (i.e., a null distribution obtained with a permutation test), thus suggesting that the enhancement of foreground-related information occurs during passive perception of natural stimuli in all the tested ROIs (V1: $p = 0.006$; V2: $p < 0.001$; V3: $p = 0.014$; V3A: $p = 0.002$; V3B: $p = 0.005$; V4: $p < 0.001$; LOC: $p < 0.001$).

In addition, this analysis rules out two potential confounding effects. One related to a "fovea-to-periphery bias" in our image set. In fact, as already observed in literature, natural images are typically characterized by objects located at the center of the scene - see for instance the object location bias represented in figure 3B in (Alexe et al., 2010). However, since the spatial distribution and number of pixels were kept constant at each permutation step, we replicated the same "fovea-to-periphery

bias" in the null distribution. The other confound was related to potential differences in contrast between foreground and background. To account for this, in the permutation test, we matched the root mean square (RMS) contrast of each random segmentation to that of the "ground truth" segmentation obtained from BSD. Overall, these control procedures minimize the chance that the observed enhancement is driven by location, size or contrast of the foreground.

Background suppression occurs in higher cortical areas

As the correlation between the background RDM and RDM derived from fMRI activity is significant in V1 and V2 only (Figure 2.3), we hypothesized that background-related information is suppressed in "higher" visual cortices. Notably, Poort and colleagues (2016) described background suppression as a different, but associated, phenomenon with respect to foreground enhancement. Thus, in order to better characterize where and how background suppression occurs in humans attending to natural images, a further analysis was performed by parametrically filtering out the background of each image, varying its contrast or spatial frequencies (low- and high-pass filtering; Figure 2.2C). As the correlation between the representational model of V3A, V3B and those derived from intact, isolated foreground and isolated background images is not significant ($p > 0.05$ Bonferroni corrected), these ROIs were not further investigated.

When comparing the correlation value of the intact version of the stimuli and the correlation value of each background filtering step, we found that V1, V2 and V3 show a progressive decay, indicating that the background is actually processed by these regions ($p < 0.05$, Bonferroni corrected). On the other hand, in V4 and LOC filtering the background produces significantly higher correlations ($p < 0.05$ Bonferroni corrected), thus indicating that background information is not different from noise (Figure 2.4). These findings suggest that background suppression is actually performed by higher cortical areas, as also depicted in correlation images (Figure 2.5).

Of note, to validate the proposed method, we performed a simulation of the fMRI experiment using a fully connected layer

of a pre-trained convolutional neural network (AlexNet fc6: Krizhevsky et al., 2012). The RDM correlation between each contrast filtering step and the representational geometry from the net (i.e., responses to intact images) was computed as in the fMRI analyses (i.e., fixed model). Then, we assessed ground truth computation of the net by showing it the images at each filtering level, thus checking its sensitivity to background manipulation. Results (not shown) demonstrate that our pre-filtering modeling approach correctly reveals the ground-truth computation of the net.

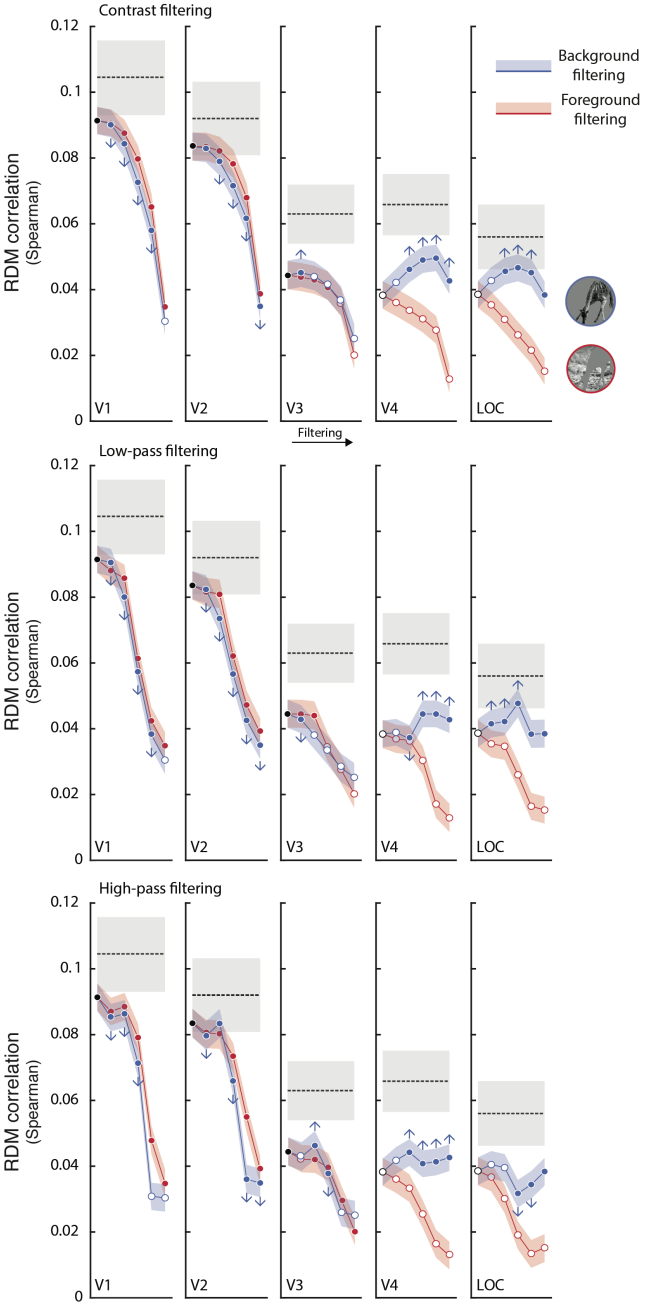

Contrast filtering
Low-pass filtering
High-pass filtering
RDM correlation (Spearman)
0.12
0.1
0.08
0.06
0.04
0.02
0
V1
V2
V3
V4
LOC
Filtering
Background filtering
Foreground filtering

Figure 2.4 | Background suppression in the human visual system
Correlation between brain activity and contrast, low- and high-pass filtering applied to the background (blue) and, as a control, to the foreground (red). Filled dots mark significant correlations ($p < 0.05$, Bonferroni corrected) while colored shaded areas represent the standard error estimates. Dashed lines represent the SNR estimate for each ROI, while gray shaded regions indicate its standard error. Arrows stand for significant differences ($p < 0.05$, Bonferroni corrected) between each filtering step and correlation values for the intact version (up: background suppression; down: progressive decay). Results show that for early regions (V1-3) background-related information is relevant, since the correlation significantly decays due to filtering ($p < 0.05$, Bonferroni corrected); on the other hand, V4 and LOC show an opposite effect, suggesting that background is suppressed in those regions.

Figure 2.5 | Correlation images
To visually represent these results, we combined the different filtering procedures (contrast, low- and high-pass filtering) of the step showing the highest correlation with the representational model from each ROI.

Table 2.2 | Statistical analysis

	Data structure	Type of test	Power
a	Single correlation values	Nonparametric permutation test	$p < 0.05$ Bonferroni corrected
b	Single correlation values	Nonparametric permutation test	$p < 0.05$
c	Single correlation values	Nonparametric permutation test	$p < 0.05$ Bonferroni corrected

Discussion

In the present study, we illustrated how the manipulation of low-level properties of natural images, and the following correlation with patterns of brain responses during passive viewing of the intact stimuli, could disclose the behavior of different regions along the visual pathway.

Employing this pre-filtering modeling approach, we tested whether scene segmentation is an automatic process that occurs during passive perception in naturalistic conditions, even when individuals are not required to perform any particular tasks, or to focus on any specific aspect of images. Here, we were able to collect three different pieces of evidence confirming our hypothesis on the mechanisms involved in scene segmentation.

First, by using RSA, we demonstrated that representational models built from fMRI patterns show a significant correlation with isolated foreground in V2, V4 and LOC, while a significant correlation with isolated background is achieved in V1 and V2 only.

Second, our analyses specifically found that foreground enhancement is present in all the selected visual ROIs, and that this effect is driven neither by the foreground contrast, nor by its size or location in the visual field. Thus, indirect evidence of figure-ground modulation of natural images could be retrieved in the activity of multiple areas of the visual processing stream (Roelfsema, 2006; Roelfsema and de Lange, 2016). This is consistent with a recent study, which reported that border-ownership of natural images cannot be solved by single cells, but

requires a population of cells in monkey V2 and V3 (Hesse and Tsao, 2016).

Finally, a proof of segmentation can be represented by the significant suppression of background-related information in V4 and LOC. On the contrary, earlier regions across the visual stream - from V1 to V3 – have a uniform representation of the whole image, as evident at first glance in the obtained correlation images (Figure 2.5). Overall these results further support the idea that foreground enhancement and background suppression are distinct, but associated, processes involved in scene segmentation of natural images.

Foreground segmentation as a proxy for shape processing

Of note, our proposed pre-filtering modeling approach produces a visual representation (i.e., correlation image) of how information is selectively coded by a specific population of interest (e.g., LOC). Further interpretations on the obtained visual representation may result more empirical and, similarly to other computational neuroimaging methods (e.g. Inverted Encoding Models: Liu et al., 2018), should be grounded on previous neurophysiological knowledge. For instance, the correlation image of LOC could be interpreted as resulting from two alternative mechanisms: LOC could preferentially process the foreground as a whole, while suppressing the background, or it could act as a 'feature detector', whose neurons are selectively tuned towards a single visual attribute (e.g., the whiskers of a cat), without actively performing any suppression. Either way, what our method clearly reveals is that LOC is selective for object texture and shape properties, and is unaffected by background-related information. At the same time, previous knowledge suggests that an active process, rather than a passive feature-matching mechanism, determines the observed results (Roelfsema and De Lange, 2016).

Furthermore, the observed behavior of V4 and LOC is consistent with several investigations on shape features selectivity in these regions, and in their homologues in monkey (Carlson et al., 2011; Hung et al., 2012; Lescroart and Biederman, 2013; Vernon et al., 2016). In fact, the extraction of shape properties requires segmentation (Lee et al., 1998), and

presumably occurs in brain regions where background is already suppressed. As mentioned before, "correlation images" reconstructed from V4 and LOC are characterized by a strong background suppression, while the foreground is preserved. This is consistent with a previous neuropsychological observation: a bilateral lesion within area V4 led to longer response times in identifying overlapping figures (Leek et al., 2012). Hence, this region resulted to be crucial for accessing foreground-related computations, and presumably plays a role in matching the segmented image with stored semantic content in figure recognition. In accordance with this, a recent hypothesis suggests the role of V4 in high-level visual functions, such as features integration or contour completion (Roe et al., 2012).

The preserved spatial resolution of foreground descriptive features (i.e., texture) in V4 and LOC – as shown in Figure 2.5 - represents an additional noteworthy aspect that arises from our data. The progression from V1 towards higher-level regions of the cortical visual pathway is associated with a relative increase in receptive fields size (Gattass et al., 1981; Gattass et al., 1987; Gattass et al., 1988; Dumoulin and Wandell, 2008; Freeman and Simoncelli, 2011; Kay et al., 2015). However, it should be kept in mind that regions such as V4 demonstrate a complete representation of the contralateral visual hemifield, rather than selective responses to stimuli located above or below the horizontal meridian (Wandell and Winawer, 2011). The evidence that the foreground portion of "correlation images" maintains fine-grained details in V4 and LOC seems to contrast a popular view according to which these regions are more tuned to object shape (i.e., silhouettes), instead of being selective for the internal configuration of images (e.g. Malach et al., 1995; Grill-Spector et al., 1998; Moore and Engel, 2001; Stanley and Rubin, 2003). However, it has been shown that foveal and peri-foveal receptive fields of V4 do accomodate fine details of the visual field (Freeman and Simoncelli, 2011) and that the topographic representation of the central portion of this area is based on a direct sampling of the primary visual cortex retinotopic map (Motter, 2009). Therefore, given the "fovea-to-periphery" bias found in our stimuli and in natural images, it is reasonable that an intact configuration of the foreground may be more tied to the activity of these brain regions, and that a richer representation of

the salient part may overcome simplistic models of objects shape (e.g., silhouettes). Our results are also consistent with a recent study on monkeys that demonstrates the role of V4 in texture perception (Okazawa et al., 2015).

Moreover, it is well known that selective attention represents one of the cognitive mechanisms supporting figure segmentation (Qiu et al., 2007; Poort et al., 2012), as suggested, for instance, by bistable perception phenomena (Sterzer et al., 2009), or by various neuropsychological tests (e.g., De Renzi et al., 1969; Bisiach et al., 1976). In the present experiment, participants were asked to simply gaze a central fixation point without performing any overt or covert tasks related to the presented image. Nonetheless, we found evidence of a clear background suppression and foreground enhancement, suggesting that scene segmentation is mediated by an automatic process that may be driven either by bottom-up (e.g., low-level properties of the foreground configuration), or top-down (e.g., semantic knowledge) attentional mechanisms. Neurophysiological studies suggest that segmentation is more likely a bottom-up process, as border-ownership assignment occurs as early as 70 ms (Williford and von der Heydt, 2016), followed by later region-filling mechanisms (i.e., enhancement and suppression) (Self et al., 2013). A limit of our study is that we cannot provide any further information related to these mechanisms and their temporal dynamics, given the limited temporal resolution of fMRI and the passive stimulation task. However, a recent study (Neri, 2017) investigated behavioral and electrophysiological responses to BSD images - intact or manipulated in several different ways, including spatial frequencies filtering and warping – in subjects who were asked to reconstruct a corrupted image region. Results showed that reconstruction of patches elicits enhanced responses when masking targeted the behaviorally segmented contours, rather than the contrast energy of the images. Moreover, this effect occurs earlier than 100ms and is not altered by semantic processing or spatial attention.

Facing the challenge of explicit modeling in visual neuroscience

One of the major goals of visual neuroscience is to predict brain responses in ecological conditions (Felsen and Dan, 2005).

In this sense, the standard approach in investigating visual processing implies testing the correlation of brain responses from a wide range of natural stimuli with features extracted by different alternative computational models. This approach facilitates the comparison between performances of competing models and could ultimately lead to the definition of a fully computable model of brain activity. However, the development of explicit computational models for many visual phenomena in ecological conditions is difficult. Indeed, many current theories, especially those concerning mid-level processing, have been hardly tested with natural images, as testified by the extensive use of artificial stimuli (e.g. Carandini et al., 2005; Wu et al., 2006). As a matter of fact, it is often impossible both to extract and to control for relevant features in natural images, and thus, there is no way to compute a predicted response from complex stimuli.

Moreover, even if computer vision is a major source of computational models and feature extractors, often its objectives hardly overlap with those of visual neuroscience. Computer scientists are mainly interested in solving single, distinct tasks (e.g., segmentation, recognition, etc.), while, from the neuroscientific side, the visual system is considered as a general-purpose system that could retune itself to accomplish different goals (Medathati et al., 2016). Consequently, while computer science typically employs solutions that rely only seldom on previous neuroscientific knowledge, and its goal is to maximize task accuracy (e.g., with deep learning), visual neuroscience somehow lacks of solid computational models and formal explanations, ending up with several arbitrary assumptions in modeling, especially for mid-level vision processing, such as scene segmentation or shape features extraction (for a definition see: Kubilius et al., 2014).

In light of all this, we believe that the manipulation of a wide set of natural images, and the computation of a fixed model based on low-level features, can offer a simple and biologically plausible tool to investigate brain activity related to higher-order computations, and that representational models offer an easily accountable link between brain activity patterns and continuous stimuli descriptions (Nili et al., 2014).In fact, the results of this exploratory approach can be depicted and are as intuitive as

descriptions obtained through formal modeling (Figure 2.5), highlighting interpretable differences rather than data predictions.
Moreover, our study indicates that the sensitivity of representational models built on fMRI patterns can represent an adequate tool to investigate complex phenomena through the richness of natural stimuli. Representational models fit this purpose: even if are summary statistics obtained from the dissimilarities between actual brain activity patterns, they are independent from a priori assumptions on anatomical relationships between brain regions, or on correspondences between voxels and units of computational models, as in the case of voxelwise encoding or decoding (Kriegeskorte and Kievit, 2013).

Methods

To assess differences between cortical processes involved in foreground-background segmentation, we employed a low-level description of images, defined by a weighted sum of the representational dissimilarity matrices (RDMs) of four well-known computational models (Figure 2.2D). These models are based on simple features – edge position, size and orientation – whose physiological counterparts are well known (Marr, 1982). The model was kept constant while the images were parametrically filtered and iteratively correlated with representational measures of brain activity through RSA. For each ROI, this pre-filtering modeling approach led to a pictorial and easily interpretable representation of the optimal features (contrast and spatial frequencies) of foreground and background of natural images (i.e., "correlation images"). The analytical pipeline is schematized in Figure 2.2.

Stimuli and behavioral segmentation of foreground and background

We selected from the 1870 images used by (Kay et al., 2008) a sub-sample of 334 pictorial stimuli which are also represented in the Berkeley Segmentation Dataset 500 (BSD) (Arbelaez et al., 2011). For each BSD image, 5-7 subjects manually performed an individual "ground-truth" segmentation, which is provided by

the authors of the dataset (http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html). Although figure-ground judgment is rather stable across subjects (Fowlkes et al., 2007), we selected the largest patch - manually labeled as foreground - among the behavioral segmentations, in order to build a foreground binary mask. For each image, this mask was then down-sampled and applied to the original stimulus to isolate the foreground and the background pixels (Kay et al., 2008).

fMRI Data

The fMRI data used in this study are publicly available at http://crcns.org/data-sets/vc/vim-1 (Kay et al., 2011). Two subjects (Males, age: 33 and 25) were acquired using the following MRI parameters: 4T INOVA MR, matrix size 64x64, TR 1s, TE 28ms, flip angle 20°, spatial resolution 2 x 2 x 2.5 mm3. For each subject five scanning sessions (7 runs each) were performed on five separate days. The stimuli were 1870 greyscale natural images with diameter 20° (500px), embedded in a grey background, and were presented for 1s, flickering at 5Hz, with an ISI of 3s. Subjects were asked to fixate a central white square of 0.2°(4px). Seven visual regions of interest (ROIs) - V1, V2, V3, V3A, V3B, V4 and LOC - were defined and brain activity patterns related to stimulus presentation was extracted from these regions. For additional details on pre-processing, retinotopic mapping and ROIs localization, please refer to (Kay et al., 2008).

Computational Models

In accordance with a previous fMRI study that, to the best of our knowledge, has tested the highest number of computational models, we selected four untrained models: two showing highest correlations with brain activity patterns in early visual areas, and the others, showing highest correlations with LOC (Khaligh-Razavi and Kriegeskorte, 2014). All these models are based on biologically inspired features, such as Gabor filters and image gradient and comprise: GIST (Oliva and Torralba, 2001), Dense SIFT (Lazebnik et al., 2006), Pyramid Histograms of Gradients

(PHOG) (Bosch et al., 2007) and Local Binary Patterns (LBP) (Ojala et al., 2001). For an exhaustive description of the four models – and links to Matlab codes – see the work by Khaligh-Razavi (2014) and Khaligh-Razavi and Kriegeskorte (2014). Our model choice was also motivated by the fact that the stimuli were grayscale and had a fixed circular aperture. Thus, we excluded descriptions based on color or silhouette information, as well as pre-trained convolutional neural networks which are biased towards the global shape of the image (Kubilius et al., 2016).

Representational Similarity Analysis (RSA)

For each filtered image, we collected feature vectors from the four computational models (PHOG, GIST, LBP and Dense SIFT), and RDMs were then obtained (1 minus the Pearson correlation metric). These four RDMs were normalized in a range between 0 and 1, and combined to obtain the fixed biologically plausible model of the stimuli (for a graphical representation of the process, see Figure 2.2D). The four model RDMs were combined through a weighted sum, based on an estimation of their correlation with the representational model of brain activity. Single subject RDMs were similarly computed using fMRI activity patterns for each of the seven ROIs, and then averaged across the two subjects. We used Spearman's rho (ϱ) to assess the correlation between the RDM from each step of the image filtering procedures and the RDM of each brain ROI. To obtain unbiased estimations of the correlation between models and fMRI, a 5-fold cross-validation procedure based on a weighted sum of the models was developed: model weights were first estimated trough linear regression on a portion (80%) of the RDMs, and the correlation with fMRI data was then computed based on the remainder of the RDMs (20%). The correlation values derived from this procedure were averaged across the five folds, to obtain a unique estimate of the similarity between image features and brain activity. This analysis was performed independently in each of the seven ROIs, and the standard error for each correlation value was estimated with bootstrapping of the stimuli – 1,000 iterations (Efron and Tibshirani, 1993).

In addition, as each ROI may show a distinct signal-to-noise ratio, we computed a noise estimation by correlating the brain RDMs extracted from the two subjects. This procedure allows for qualitative comparison between different ROIs and could help in estimate how well each model explains fMRI RDMs given the noise in the data.

Foreground enhancement testing

A permutation test was performed to statistically assess the enhancement of the information retained in the behavioral segmented foreground. In this test both the "fovea-to-periphery" bias that characterizes natural images, and possible differences in contrast between foreground and background were controlled (Figure 2.2A). For each iteration, the 334 foreground masks were shuffled and a random foreground segmentation was associated to each stimulus. The RMS contrast of each obtained segmented image was matched to that of the behaviorally segmented counterpart. Of note, this set of randomly-segmented images had the same distribution of masked portions of the visual field as the one from the behavioral segmentation, so the same amount of information was isolated at each permutation step. This procedure was repeated 1,000 times, to build a null distribution of alternative segmentations: four examples of random segmentation are shown in Figure 2.2A. For each permutation step, features were extracted from each randomly segmented image and RSA was performed using the procedure described above.

Parametric filtering procedures

In order to investigate differential processing of foreground and background in the visual system, we employed three different filtering procedures (contrast - through alpha channel modulation - low- and high-pass filtering of spatial frequencies) applied parametrically (4 steps each) to the foreground or the background. For each filtering procedure, the four manipulated images are represented in Figure 2.2C. For low- and high-pass filtering, we employed a Butterworth filter (5th order), linearly sampling from a log-transformed distribution of frequencies

ranging from 0.05 to 25 cyc/°, while keeping the RMS contrast fixed.

Background suppression testing

To test background suppression, we performed a two-tailed permutation test. In each ROI, we computed the difference between the correlation of the intact version of the stimuli and each step of the background filtering procedures (Figure 2.2B). Afterwards, a permutation test (10,000 iterations) was performed by random sampling two groups from the bootstrap distributions, obtaining a null distribution of correlation differences. Reported results are Bonferroni corrected (for the 13 comparisons in each ROI).

Correlation images

For each ROI, the effects of the filtering procedures were combined, to build "correlation images". To this aim we used the filtering step with the highest correlation between the fixed model and RDMs from fMRI data, for foreground and background respectively. In detail, we averaged the best images for the low- and high-pass filters, and multiplied each pixel for the preferred alpha-channel value (contrast).

Significance testing

To assess the statistical significance of the correlations obtained with RSA in all the above mentioned filtering procedures, we built a robust ROI-specific permutation test (1,000 iterations), by randomly sampling voxels of the occipital lobe not located in any of the seven ROIs. We labeled these voxels as 'control-voxels'. This procedure has the advantage to be resilient to biases in fMRI data (Schreiber and Krekelberg, 2013), instead of simply taking into account the distribution of the RDM values, as in (Khaligh-Razavi and Kriegeskorte, 2014). In addition, the procedure that we developed is also useful to control for the effects related to number of voxels and to the signal-to-noise ratio of each ROI.

First, for each ROI we computed the standard error of the ROI-specific noise estimation with bootstrap resampling of the stimuli (1,000 iterations). Second, a number of control voxels equal to the number of voxels was randomly selected within each ROI, and the activity of these control voxels in response to the stimuli were used to build a null RDM. Third, the correlation between the null RDMs of the two subjects was computed. However, since we aimed at matching the signal-to-noise ratio of the null distribution to that of each ROI, the null RDM was counted as a valid permutation only if the single subject RDMs correlated to each other within a specific range (i.e., ROI-specific noise estimation ± standard error). Finally, for each step of the filtering procedures, each of the 1,000 ROI-specific null RDMs were correlated with the fixed model RDM to obtain a null distribution of 1,000 ϱ values. A one-tailed rank test was used to assess the significance of the ϱ of the fixed model with brain RDMs. For each ROI, we controlled for multiple comparisons (27 tests), through Bonferroni correction.

Code accessibility

All analyses have been implemented in Matlab (The Mathworks Inc.) using in-house developed code (available at the following link: https://bit.ly/2rC27hY).

References

Alexe B, Deselaers T, Ferrari V (2010) ClassCut for Unsupervised Class Segmentation. Lect Notes Comput Sc 6315:380-393.

Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33:898-916.

Biederman I (1987) Recognition-by-Components - a Theory of Human Image Understanding. Psychological Review 94:115-147.

Bisiach E, Capitani E, Nichelli P, Spinnler H (1976) Recognition of overlapping patterns and focal hemisphere damage. Neuropsychologia 14:375-379.

Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In, pp 401-408: ACM.

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC (2005) Do we know what the early visual system does? J Neurosci 25:10577-10597.
Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area V4. Curr Biol 21:288-293.
David SV, Vinje WE, Gallant JL (2004) Natural stimulus statistics alter the receptive field structure of v1 neurons. J Neurosci 24:6991-7006.
De Renzi E, Scotti G, Spinnler H (1969) Perceptual and associative disorders of visual recognition. Relationship to the side of the cerebral lesion. Neurology 19:634-642.
Dumoulin SO, Wandell BA (2008) Population receptive field estimates in human visual cortex. Neuroimage 39:647-660.
Efron B, Tibshirani R (1993) An introduction to the bootstrap. New York: Chapman & Hall.
Felsen G, Dan Y (2005) A natural approach to studying vision. Nat Neurosci 8:1643-1646.
Fowlkes CC, Martin DR, Malik J (2007) Local figure-ground cues are valid for natural images. J Vis 7:2.
Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nat Neurosci 14:1195-1201.
Gattass R, Gross CG, Sandell JH (1981) Visual topography of V2 in the macaque. J Comp Neurol 201:519-539.
Gattass R, Sousa AP, Rosa MG (1987) Visual topography of V1 in the Cebus monkey. J Comp Neurol 259:529-548.
Gattass R, Sousa AP, Gross CG (1988) Visuotopic organization and extent of V3 and V4 of the macaque. J Neurosci 8:1831-1845.
Grill-Spector K, Kushnir T, Edelman S, Itzchak Y, Malach R (1998) Cue-invariant activation in object-related areas of the human occipital lobe. Neuron 21:191-202.
Handjaras G, Leo A, Cecchetti L, Papale P, Lenci A, Marotta G, Pietrini P, Ricciardi E (2017) Modality-independent encoding of individual concepts in the left parietal cortex. Neuropsychologia 105:39-49.
He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. arXiv preprint arXiv:170306870.
Hesse JK, Tsao DY (2016) Consistency of Border-Ownership Cells across Artificial Stimuli, Natural Stimuli, and Stimuli with Ambiguous Contours. J Neurosci 36:11338-11349.

Hung CC, Carlson ET, Connor CE (2012) Medial axis shape coding in macaque inferotemporal cortex. Neuron 74:1099-1113.
Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210-1224.
Kastner S, De Weerd P, Ungerleider LG (2000) Texture segregation in the human visual cortex: A functional MRI study. J Neurophysiol 83:2453-2457.
Kay KN, Naselaris T, Gallant JL (2011) fMRI of human visual areas in response to natural images. CRCNS org.
Kay KN, Weiner KS, Grill-Spector K (2015) Attention reduces spatial uncertainty in human ventral temporal cortex. Curr Biol 25:595-600.
Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352-355.
Khaligh-Razavi S-M (2014) What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. arXiv preprint arXiv:14072776.
Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10:e1003915.
Kok P, de Lange FP (2014) Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. Curr Biol 24:1531-1535.
Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17:401-412.
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In, pp 1097-1105.
Kubilius J, Wagemans J, Op de Beeck HP (2014) A conceptual framework of computations in mid-level vision. Front Comput Neurosci 8:158.

Kubilius J, Bracci S, Op de Beeck HP (2016) Deep Neural Networks as a Computational Model for Human Shape Sensitivity. PLOS Comput Biol 12:e1004896.
Lamme VA (1995) The neurophysiology of figure-ground segregation in primary visual cortex. J Neurosci 15:1605-1615.
Lamme VA, Rodriguez-Rodriguez V, Spekreijse H (1999) Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. Cereb Cortex 9:406-413.
Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In, pp 2169-2178: IEEE.
Lee TS, Mumford D, Romero R, Lamme VA (1998) The role of the primary visual cortex in higher level vision. Vision Res 38:2429-2454.
Leek EC, d'Avossa G, Tainturier MJ, Roberts DJ, Yuen SL, Hu M, Rafal R (2012) Impaired integration of object knowledge and visual input in a case of ventral simultanagnosia with bilateral damage to area V4. Cogn Neuropsychol 29:569-583.
Lescroart M, Agrawal P, Gallant J (2016) Both convolutional neural networks and voxel-wise encoding models of brain activity derived from ConvNets represent boundary-and surface-related features. J Vis 16:756-756.
Lescroart MD, Biederman I (2013) Cortical representation of medial axis structure. Cereb Cortex 23:629-637.
Liu T, Cable D, Gardner JL (2018) Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width. J Neurosci 38:398-408.
Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, Tootell RB (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci U S A 92:8135-8139.
Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W.H. Freeman.
Medathati NVK, Neumann H, Masson GS, Kornprobst P (2016) Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. Computer Vision and Image Understanding 150:1-30.

Miyawaki Y, Uchida H, Yamashita O, Sato M-a, Morito Y, Tanabe HC, Sadato N, Kamitani Y (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60:915-929.
Moore C, Engel SA (2001) Neural response to perception of volume in the lateral occipital complex. Neuron 29:277-286.
Motter BC (2009) Central V4 receptive fields are scaled by the V1 cortical magnification and correspond to a constant-sized sampling of the V1 surface. J Neurosci 29:5749-5757.
Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400-410.
Neri P (2017) Object segmentation controls image reconstruction from natural scenes. PLoS Biol 15:e1002611.
Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.
Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology 21:1641-1646.
Ojala T, Pietikäinen M, Mäenpää T (2001) A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. In, pp 399-408: Springer.
Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci U S A 112:E351-360.
Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42:145-175.
Poort J, Self MW, van Vugt B, Malkki H, Roelfsema PR (2016) Texture Segregation Causes Early Figure 2.Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex. Cereb Cortex 26:3964-3976.
Poort J, Raudies F, Wannig A, Lamme VA, Neumann H, Roelfsema PR (2012) The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. Neuron 75:143-156.
Qiu FT, Sugihara T, von der Heydt R (2007) Figure-ground mechanisms provide structure for selective attention. Nat Neurosci 10:1492-1499.

Roe AW, Chelazzi L, Connor CE, Conway BR, Fujita I, Gallant JL, Lu H, Vanduffel W (2012) Toward a unified theory of visual area V4. Neuron 74:12-29.
Roelfsema PR (2006) Cortical algorithms for perceptual grouping. Annu Rev Neurosci 29:203-227.
Roelfsema PR, de Lange FP (2016) Early Visual Cortex as a Multiscale Cognitive Blackboard. Annu Rev Vis Sci 2:131-151.
Scholte HS, Jolij J, Fahrenfort JJ, Lamme VA (2008) Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging. J Cogn Neurosci 20:2097-2109.
Schreiber K, Krekelberg B (2013) The statistical analysis of multi-voxel patterns in functional imaging. PLoS ONE 8:e69328.
Self MW, van Kerkoerle T, Super H, Roelfsema PR (2013) Distinct roles of the cortical layers of area V1 in figure-ground segregation. Curr Biol 23:2121-2129.
Stanley DA, Rubin N (2003) fMRI Activation in Response to Illusory Contours and Salient Regions in the Human Lateral Occipital Complex. Neuron 37:323-331.
Stanley GB, Li FF, Dan Y (1999) Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. J Neurosci 19:8036-8042.
Sterzer P, Kleinschmidt A, Rees G (2009) The neural bases of multistable perception. Trends Cogn Sci 13:310-318.
Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline JB, Lebihan D, Dehaene S (2006) Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage 33:1104-1116.
Vernon RJ, Gouws AD, Lawrence SJ, Wade AR, Morland AB (2016) Multivariate Patterns in the Human Object-Processing Pathway Reveal a Shift from Retinotopic to Shape Curvature Representations in Lateral Occipital Areas, LO-1 and LO-2. J Neurosci 36:5763-5774.
Wandell BA, Winawer J (2011) Imaging retinotopic maps in the human brain. Vision Res 51:718-737.
Williford JR, von der Heydt R (2016) Figure-Ground Organization in Visual Cortex for Natural Scenes. eNeuro 3.
Wu MC, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29:477-505.

Chapter 3

Common spatiotemporal processing of visual features shapes object representation

Introduction

To make sense of the surrounding environment, our visual system relies on different transformations of the retinal input (Malcolm et al., 2016). Just consider Figure 3.1A. As any natural scene, this image is defined by a specific content of edges and lines. However, biological vision evolved to disclose the layout of discrete objects, hence the two giraffes in the foreground emerge as salient against the background, and the distinct contents pertaining to edges, shape, texture, and category contribute together to object perception.

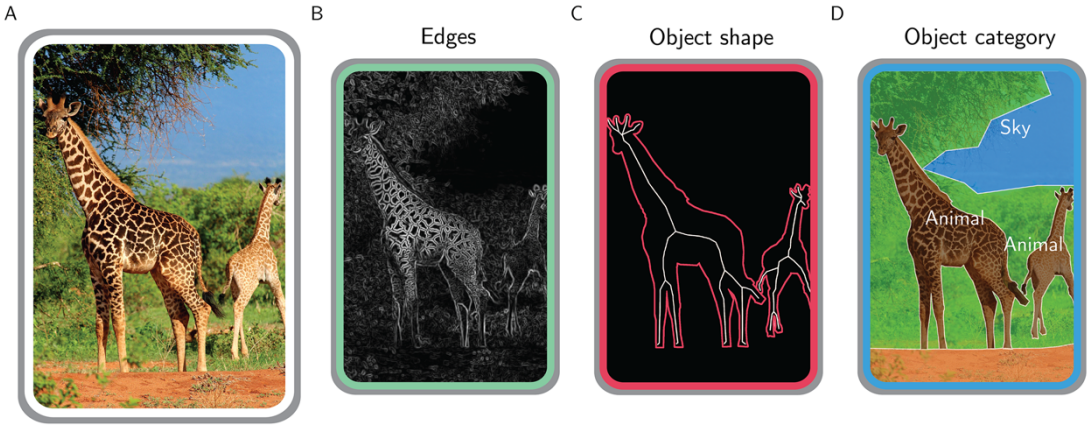


Figure 3.1 | Different representations of a natural image
A real-world scene (A), depicting two giraffes in the savannah, can be defined by its edges (B), by the shape of the giraffes (C) and also by the categorical information it conveys (D). Photo taken from http://pixabay.com, released under Creative Commons CC0 license.

Actually, each feature of Figures 1B-D is processed across the whole visual system. The primary visual cortex (V1) provides an optimal encoding of natural image statistics based on local contrast, orientation and spatial frequencies (Olshausen and Field, 1996a; Vinje and Gallant, 2000), and these low-level features significantly correlate with brain activity in higher-level

visual areas (Rice et al., 2014; Groen et al., 2018). Nonetheless, occipital, temporal and parietal modules also process object shape (Carlson et al., 2011; Hung et al., 2012; Lescroart and Biederman, 2013; Handjaras et al., 2017) and categorical knowledge (Haxby et al., 2001; Kriegeskorte et al., 2008; Handjaras et al., 2016).

Although all these features are relevant to our brain, their relative contribution in producing discrete and coherent percepts has not yet been clarified. In general, these different dimensions are interrelated and share common biases (i.e., are collinear), thus limiting the capability to disentangle their specific role (Kay, 2011). For instance, categorical discriminations can be driven either by object shape (e.g., tools have peculiar outlines) or spatial frequencies (e.g., faces and places have specific spectral signatures: Torralba and Oliva, 2003). Consequently, object shape and category are processed by the same regions across the visual cortex, even when using a balanced set of stimuli (Bracci and Op de Beeck, 2016). Even so, the combination of multiple feature-based models describes brain object representations better than the same models tested in isolation. For instance, a magnetoencephalography (MEG) study found that combining low-level and semantic features improves the prediction accuracy of brain responses to viewed objects, suggesting that semantic information integrates with visual features during the temporal unfolding of object representations (Clarke et al., 2015).

To investigate the spatiotemporal dynamics of object processing, we combined model-based descriptions of pictures, MEG brain activity patterns and a statistical procedure (Relative Weights Analysis (RWA): Johnson, 2000) that mitigate the effects of common biases across different dimensions. We ultimately determine the relative contribution across space and time of multiple feature-based representations – i.e., low-level, shape and categorical features - in producing the structure of what we perceive. First, a low-level description of the stimuli was grounded on features extracted by the early visual cortex (i.e., image contrast and spatial frequencies). Second, since shape is critical to interact with the surrounding environment (Kubilius et al., 2014), we relied on a well-assessed, physiologically-motivated description of shape, i.e., the medial axis (Blum, 1973).

Finally, objects were also distinctively represented according to their superordinate categories.

To anticipate, we observed fast (100-150ms) and overlapping representations of low-level properties (contrast and spatial frequencies), shape (medial-axis) and category in posterior sensors. These results may be interpreted as macroscale dynamics resulting in independent parallel processing, and may also suggest a role for shape in the refinement of categorical matching.

Results

We employed the Relative Weights Analysis[16] to reveal the proportional contribution of low-level, shape and category feature models in predicting time resolved representational geometries derived from MEG data, recorded from subjects attending to pictures representing thirty different stimuli from six semantic categories (Figure 3.2).

The possible transformations of retinal input were described at three canonical steps of the object processing hierarchy, grounded on previous neurophysiological investigations. A first low-level model was computed by filtering the stimuli with a bank of Gabor filters: this model captures the arrangement of spatial frequencies in a V1-like fashion (Olshausen and Field, 1996a). Then, as in previous neuroimaging investigations on the same topic (Leeds et al., 2013; Handjaras et al., 2017), we described object shape as its medial-axis transform (Blum, 1973), that roughly describes an object as its skeleton, with each object part captured by a different branch. And finally, objects were identified by the semantic category they belong to (Kriegeskorte et al., 2008).

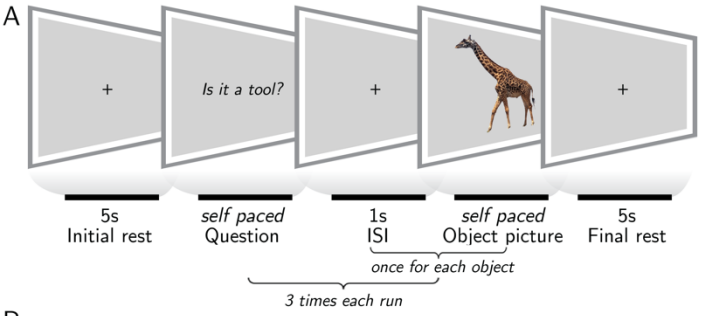
First, we assessed the collinearity between the three models, expressed as the Spearman correlation between the model RDMs (Figure 3.2B). The low-level and categorical models have a correlation of $\rho = 0.16$, the shape model has $\rho = 0.09$ correlation with the categorical model, and $\rho = 0.08$ correlation with the low-level one.

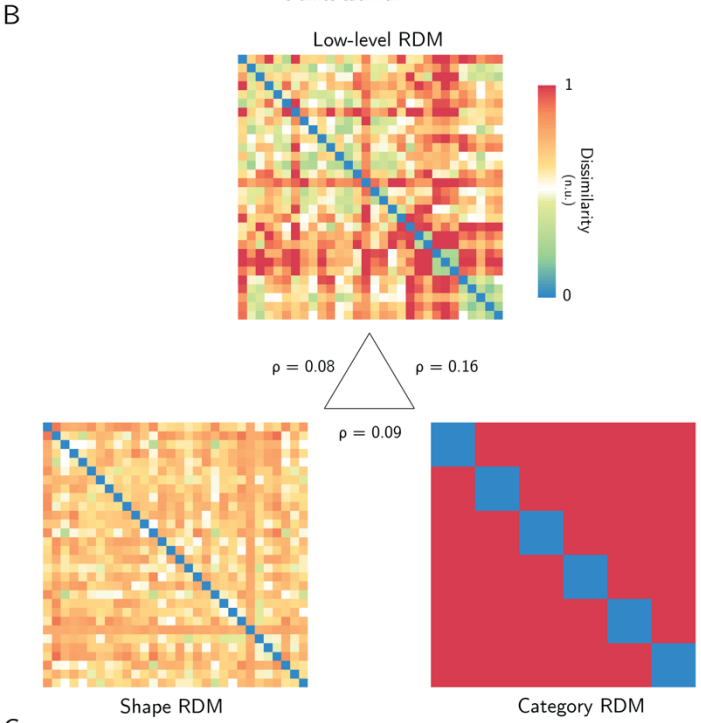
Then, RWA was performed within a sensor space searchlight, resulting for each subject in three maps that report the time courses of the metric ε for each sensor, i.e., the proportional

contribution of each model across time. RWA controls for model multicollinearity in multiple regression: its metric (ε) does not identify the impact of each model to the prediction of a dependent variable in isolation (i.e., beta weight), as in common multiple linear regressions, but considers also how each model relates to (i.e., is correlated with) the others. Thus, it reflects in a suitable manner the proportional impact of each variable on the prediction of brain activity (Figure 3.2C). The single-subject maps were aggregated in group-level z-maps for each model, corrected for multiple comparisons and divided in 50ms-long time bins for displaying purposes. Only the sensors whose corrected z-values were significant in the entire bin were retained, as displayed in Figure 3.3 (black dots mark significant sensors: $p < 0.05$, rank test, 100,000 permutations, TFCE corrected).

Results show that the model based on low-level features (contrast and spatial frequencies) is significant at early stages after stimulus presentation (0-50ms) in a cluster of posterior and medial sensors. This cluster expands in the lateral and anterior directions, reaching a maximum in the 100-150ms interval, when most of the posterior sensors are significant. Shape features are instead restricted to right posterior location in the 100-150ms interval, and do not reach significance in the remainder of sensors and time bins. The category-based model is significant in medial and posterior sensors starting at 50-100ms. The cluster expands to most of the posterior and lateral sensors, with a maximum spatial extent between 100 and 200ms, then restricting to the posterior and lateral sensors in the 200-250ms time bin. A cluster of right posterior sensors shows significant weights for the three models in the 100-150ms time bin only. None of the models was significant in the remaining parts of the time course (before stimulus onset and after 300ms).

Even if the task was intended to orient subjects' efforts specifically towards high-level semantic processing, attention towards local features could account for the observed results. To this aim, we compared the responses between semantically similar and dissimilar stimuli and found no significant difference ($p > 0.20$; see Figure 3.3-1). Thus, results are likely not driven by task demand.


A
+
Is it a tool?
+
+
5s
Initial rest
self paced
Question
1s
ISI
self paced
Object picture
5s
Final rest
once for each object
3 times each run


B
Low-level RDM
1
0
Dissimilarity
(a.u.)
ρ = 0.08
ρ = 0.16
ρ = 0.09
Shape RDM
Category RDM

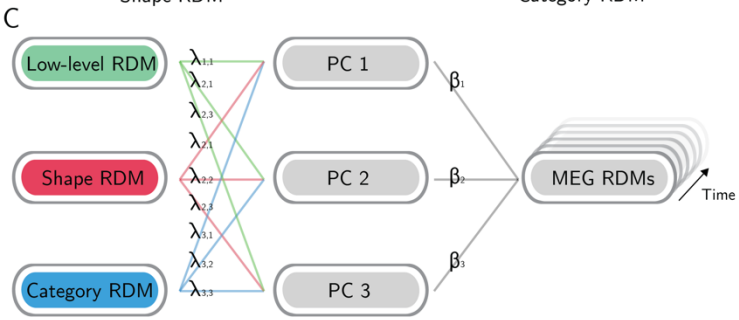

C
Low-level RDM
Shape RDM
Category RDM
PC 1
PC 2
PC 3
MEG RDMs
Time

Figure 3.2 | Methodological pipeline

A) Experimental design: subjects were asked to attend thirty object pictures during a semantic judgment task.

B) Representational dissimilarity matrices (RDMs) of three models (low-level features, shape and category) were employed to predict the MEG representational geometry – in the central triangle, Spearman correlation values between models are reported.

C) With Relative Weights Analysis, MEG RDMs were predicted using three orthogonal principal components (PCs 1-3) obtained from the models, and the resulting regression weights were back-transformed to determine the relative impact of each model on the overall prediction when controlling for the impact of model collinearity (see Methods). Photo taken and edited from http://pixabay.com, released under Creative Commons CC0 license.

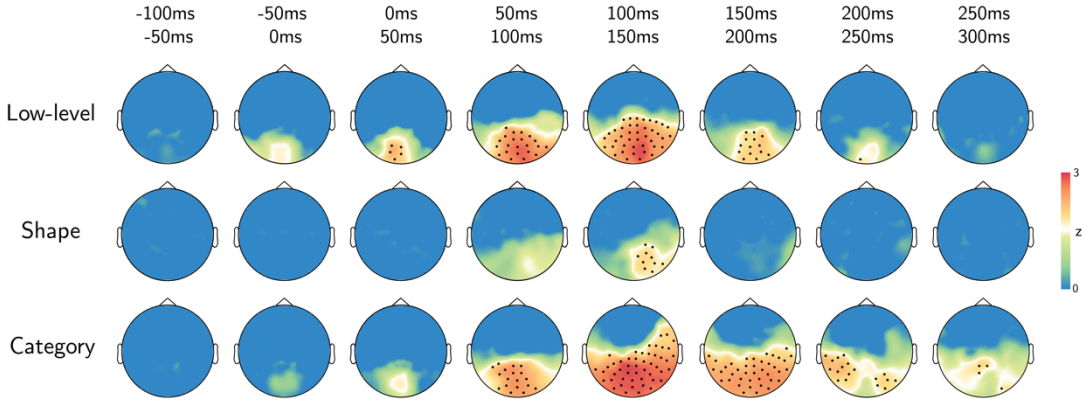


Figure 3.3 | Results

Topographic plots of the group-level z-maps. Top-row reports the time bin. Black dots stand for significant channels within all the time-bin ($p < 0.05$, rank test, 100,000 permutations, TFCE corrected).

Discussion

The visual machinery is a general-purpose system, relying on different representations that often are collinear or interact to each other. Here, by taking into account model collinearity, the spatiotemporal dynamics of joint feature processing within the human visual system were revealed to assess the relative contribution of low-level, shape and category features in predicting MEG-based representations. We observed both a temporal and spatial co-occurrence of low-level, shape and categorical processing, early in time (100-150ms) in posterior sensors. Specifically, we showed that a) low-level features (i.e., contrast and spatial frequencies) are processed early (0-50ms) after stimulus onset within posterior MEG sensors, spreading in time from medial to lateral locations; b) shape coding is limited within a few right posterior sensors in a brief time window (100-150ms) and co-occurs with low-level and categorical processing; c) categorical representation emerges later than the onset of low-level processing and is more prolonged, but spreads within a similar pattern of sensors.

Our results demonstrate that within 100-150ms after stimulus onset, these features are processed concurrently, suggesting that object discrimination may result from independent parallel processing (i.e., orthogonal feature-based descriptions processed with similar temporal dynamics), rather than from a strict feed-forward hierarchy. The observed spatiotemporal overlap is in line with previous neuroimaging evidence showing that category and shape are processed within the same visual regions (Bracci and Op de Beeck, 2016; Proklova et al., 2016), and can be decoded in the 130-200ms time window within the high-level visual cortex, as shown in a combined fMRI-MEG study, which focused on body parts and clothes (Kaiser et al., 2016). Here we employed a model-based approach which also embedded low-level features, and sampled stimuli from a broader set of categorical classes. In addition, we introduced RWA to overcome multicollinearity, which was not explicitly addressed in previous studies.

Of note, our results raise questions concerning the role of shape in categorization. The synchronization between the three models in our data occurs in a time window (100-150ms) that overlaps with those of perceptual organization (70-130ms) and

categorical recognition of visual information (>130ms), as indicated by previous neurophysiological and functional studies in both human and nonhuman primates (Bar, 2003; Johnson and Olshausen, 2003; DiCarlo et al., 2012; Poort et al., 2016; Williford and von der Heydt, 2016; Neri, 2017).

Whether shape processing is needed to recognize and classify objects in a scene has not been clarified yet. The classical view that considered shape essential to recognition (Biederman, 1987) has, however, being challenged by the success of several appearance-based computational models that could perform object recognition by relying on low-level features only (Oliva and Torralba, 2001). Since object segmentation occurs during passive natural image viewing (Papale et al., 2018) and controls scene reconstruction (Neri, 2017), shape analysis can be similarly triggered by object viewing also in a task for which shape is not explicitly relevant. Thus, our observation has at least two possible explanations: a) shape processing is to some extent necessary for categorization or, alternatively, b) it is not, but it is an automatic process occurring even when not overtly required by the task. The former hypothesis may, however, not be consistent with our results that show categorical representations occurring earlier than shape-based representations. In addition, the latter case would be in line with evidence suggesting that the extraction of object affordances – i.e., shape-related features which are able to facilitate or even trigger actions – is a fast and automatic process (Craighero et al., 1996; Grezes et al., 2003). However, a conclusion on this topic can be reached only by further studies involving task modulation (Harel et al., 2014). Of note, task is able to influence the strength of object processing late in time (>150ms: Hebart et al., 2018).

Another interesting result is the early emergence (50-100ms) of categorical processing within the same pattern of sensors that also encode contrast and spatial frequencies. As mentioned before, object recognition has been described as occurring at 150ms or later (Bar, 2003). We observed category representations within posterior sensors well before (even accounting for the temporal smoothing potentially introduced by the searchlight procedure). Early occurrence of categorical processing has been observed also in previous MEG studies (Clarke et al., 2015; Hebart et al., 2018).

In the past years, mounting evidence revealed a top-down control of neurons in the early visual cortex (Lamme, 1995; Lamme and Roelfsema, 2000; Qiu et al., 2007; Poort et al., 2012; Hesse and Tsao, 2016; Williford and von der Heydt, 2016). Moreover, in a series of elegant studies (2011, 2014, 2017), Neri found psychophysical evidence of a top-down predictive mechanism, comprising a progressive refinement of local image reconstruction driven by global saliency or semantic content. At the macroscale, the effects of this mechanism imply that both local (i.e., low-level features) and global (i.e., object-related) representations should be retrieved early in time (<150ms) within the visual cortex. Our results, show early (from 50 to 200ms), overlapping patterns for low-level and categorical processing in posterior MEG sensors, in line with this view. However, further research is needed to directly test the causal role of top-down feedbacks in controlling low-level processing within the occipital cortex, which falls beyond the original scope of this work.

A further general remark should be made. As mentioned before, multicollinearity is a pervasive property of our surrounding environment. Indeed, one of the most fascinating features of our visual system is the way it deals with correlated statistics within the natural domain, to optimally represent the retinal input (Olshausen and Field, 1996b), and to make sense of the external world, through the mean of learning and generalization. Indeed, visual correspondences are the mechanism we used to evolve more abstract, categorical representations (Tenenbaum et al., 2011). However, from the researcher perspective, this leads to an extreme effort in balancing dimensions of interest, or in developing orthogonal models. In addition, two further aspects should be considered: first, as shown empirically (Kay, 2011), since different stimuli typically vary within multiple dimensions simultaneously, it is almost impossible to isolate a single dimension of interest; second, the effort in building orthogonal competing descriptions increases with the number of tested models.

Several methods have been proposed to overcome model collinearity issues when studying brain activity (for a review, see: Nimon and Oswald, 2013). Within the field of neuroimaging, Lescroart et al. (2015) employed a variance partitioning approach

(the same method, in the domain of multiple linear regression, is known as commonality analysis – as also employed in the MEG field (Hebart et al., 2018), which aims at determining the explained variance for any possible subset of the models. While this analysis is able to estimate the variance unique to each partition, its main drawback is that partitions grow exponentially with the number of models: since there are $2^p - 1$ subsets for p predictors, just exploring the impact of 5 models generates 31 different subsets. In light of this, even comparing a low number of models would end up in a computationally intensive process and in the challenging task of interpreting and discussing a huge number of sub-models. Moreover, the partitions related to variance shared by different models can occasionally be negative, and the interpretation of these negative components is still matter of debate (Ray-Mukherjee et al., 2014). From this perspective, RWA is an attractive alternative, as it estimates the relative, non-negative weight of each model and does not imply to discuss more models or components than those initially considered.

Indeed, relative weights reflect in a suitable manner the proportional impact of each variable on the prediction of brain activity and - if the predictors are standardized - sum up to the total explained variance (Johnson, 2000). However, some limitations also affect RWA: the most relevant is that estimated weights are not invariant to the orthogonalization procedure employed. Though, it has been proven that, the more the orthogonal variables approximate the original variables, the more reliable the estimated weights become (for a deeper treatment of the topic, see: Johnson, 2000). Therefore, RWA may represent a fast and appealing recipe to deal with model multicollinearity within the neuroimaging field, especially when three or more models are compared.

In conclusion, this study reveals the spatiotemporal dynamics of object processing from a model-based perspective, providing evidence in favor of an integrated perceptual mechanism in object representation.

Methods

Participants

Sixteen healthy right-handed volunteers (5F, age 27 ± 2) with normal or corrected to normal visual acuity participated in the study. All subjects gave informed consent to the experimental procedures and received a monetary reward for their participation. The study was approved by the Ethics Committee for research involving human participants at the University of Trento, and all the experimental procedures were conducted in accordance with the Declaration of Helsinki.

Stimuli

Visual stimuli were color pictures representing thirty different objects from six semantic categories (fruits, vegetables, animals, birds, tools, vehicles). The set of stimuli were used in two previous fMRI studies from our group (Handjaras et al., 2016; Handjaras et al., 2017), and were controlled for psycholinguistic features and familiarity (for details, see: Handjaras et al., 2016). Stimuli were presented using MATLAB and the Psychophysics Toolbox (Brainard, 1997), and were projected on a translucent screen placed at about 130cm from the participant, using a Propixx DLP projector (VPixx technologies), with a refresh rate of 60 Hz and a resolution of 1280x1024 pixels (21.7 x 13.16°).

Task and design

The experiment was organized in eight runs, each consisting of three blocks (see Figure 3.2A). In each block, the thirty images were presented in randomized order, and participants were engaged in a semantic judgment task to ensure that they focused the attention on the stimuli (Sudre et al., 2012). At the beginning of each block, a binary target question (e.g., "Is it a tool?") was shown; once subjects read the questions, they prompted the start of the block by pressing a button on a keyboard. Within each block, subjects answered (yes/no) to the question presented at the beginning using the keyboard. All pictures were presented 24 times, with a different target question for each repetition. 5s-long resting periods preceded and followed each block, and 1s-long resting periods followed the behavioral response to each

stimulus within a block. During the resting periods, subjects had to fixate a black cross, displayed in the center of the screen. The order of the questions was randomized across participants.

Models

In order to predict MEG representational geometries, three different descriptions were built, representing different physiologically relevant properties of the objects seen by the subjects (see Figure 3.2B). First, a low-level model, which captures the arrangement of spatial frequencies in a V1-like fashion, was employed: a GIST (Oliva and Torralba, 2001) descriptor for each stimulus was derived by sampling (in a 4x4 grid) the responses to a bank of isotropic Gabor filters (8 orientations and 4 scales). The descriptor (consisting of a vector with 512 elements) of each stimulus was then normalized and compared to each other stimulus using the pairwise correlation distance (1 – Pearson's r). Second, a shape model was computed. Similarly to previous neuroimaging investigations on the same topic (Leeds et al., 2013; Handjaras et al., 2017), the medial-axis transform (Blum, 1973) was extracted from each manually segmented and binarized object silhouette. Then, shock-graphs skeletal representations were built, and their pairwise dissimilarity was computed using the ShapeMatcher algorithm (http://www.cs.toronto.edu/~dmac/ShapeMatcher/; Van Eede et al., 2006), which estimates the minimum deformation needed in order to match two different shapes (Sebastian et al., 2004). Finally, the thirty stimuli were described based on their semantic category, obtaining a binary categorical model.

MEG data acquisition

MEG data were recorded using an Elekta VectorView system with 306-channels, 204 first order planar gradiometers and 102 magnetometers (Elekta-Neuromag Ltd., Helsinki, Finland), located in a magnetically shielded room (AK3B, Vakuumschmelze, Hanau, Germany). The sampling rate was 1kHz. Head shapes were recorded from each participant immediately before the experiment, using a Polhemus Fastrak digitizer (Polhemus, Vermont, USA) recording the position of

fiducial points (nasion, pre-auricular points) and around 500 additional points on the scalp. MEG data were synchronized with experiments timing by sending four different triggers at question presentation, first button press (after question), stimulus presentation and stimulus-related behavioral responses (button presses), respectively.

MEG data pre-processing

MEG data pre-processing was performed using the Fieldtrip toolbox (Oostenveld et al., 2011). First, a bandpass (1-80 Hz) and a notch (50 Hz) 4th order Butterworth IIR filters were applied to the data (Gross et al., 2013). Filtered signals were then cut in epochs from 500ms before to 1s after stimulus onset and resampled at 400 Hz. Subsequently, data were visually inspected according to a set of summary statistics (range, variance, maximum absolute amplitude, maximum z-value) to search for trials and channels affected by artefacts, using the procedure for visual artefact identification implemented in Fieldtrip; trials marked as bad were rejected and noisy sensors were reconstructed by interpolating their spatial neighbors. On average, 8% of the trials and 10% of the channels were rejected for each subject.

Searchlight analysis

A searchlight analysis was performed using CoSMoMVPA (Oosterhof et al., 2016), retaining the MEG data from the gradiometers only. First, the time-locked patterns for the individual trials were reduced to thirty pseudo-trials (one for each stimulus: Guggenmos et al., 2018). Searchlights were then defined for each time point of the pseudo-trials using a spatial and temporal neighboring structure (Su et al., 2012). Each searchlight included 10 dipoles (pairs of combined gradiometers) in the spatial domain, and each time point plus the ten preceding and following it (i.e., 21 time points, 52.5ms) in the temporal domain. Within each spatiotemporal searchlight, a time-varying representational dissimilarity matrix (RDM) was derived for the MEG data by computing the pairwise correlation distances between pattern of responses to the thirty stimuli (Kocagoncu et

al., 2017); prior to computing the RDM, stimulus-specific activity patterns were normalized (z-scored).

Relative Weights Analysis (RWA)

In order to estimate how well each model RDM was related to MEG representational geometries, a multiple linear regression for each subject and each spatiotemporal searchlight was performed. Since some of the three models RDMs are significantly correlated the Relative Weights Analysis (RWA), introduced by Johnson (2000), was adopted. The metric on which RWA relies is called epsilon (ε) and reflects both the unique contribution of each model and its impact when all the other models are considered.

The RWA procedure is graphically synthetized in Figure 3.2C. Basically, the models RDMs were first orthogonalized, by performing a Principal Component Analysis (PCA), and the RDMs from each spatiotemporal searchlight were regressed on the so obtained orthogonal versions of the models RDMs. Then, the regression coefficients were related back to the original model RDMs by regressing the orthogonal RDMs also on the models RDMs. Finally, for the *j-th* model, epsilon was calculated as:

$$\varepsilon_j = \sum_{k=1}^{p} \lambda_{jk}^2 \beta_k^2$$

where p is the number of models, β_k^2 is the variance (i.e., the squared standardized regression coefficient) in each searchlight RDM accounted for by the *k-th* orthogonal RDM, and λ_{jk}^2 is the variance in the *j-th* model accounted for by the *k-th* orthogonal RDM.

Statistical analyses

The RWA analysis, performed within the spatiotemporal searchlights as described above, provided a time course of the metric (ε) for each sensor and time point. To estimate the group-level spatiotemporal distribution of weights for each of the three models, a one sample non-parametric test was performed, using a null distribution generated with 100,000 permutations (rank

test), as implemented in CoSMoMVPA. Correction for multiple comparisons was made at cluster-level using a threshold-free method (TFCE: Smith and Nichols, 2009; Pernet et al., 2015). Z-values corresponding to a corrected p-value of 0.05 (one-tailed) were considered significant.

Extended Data

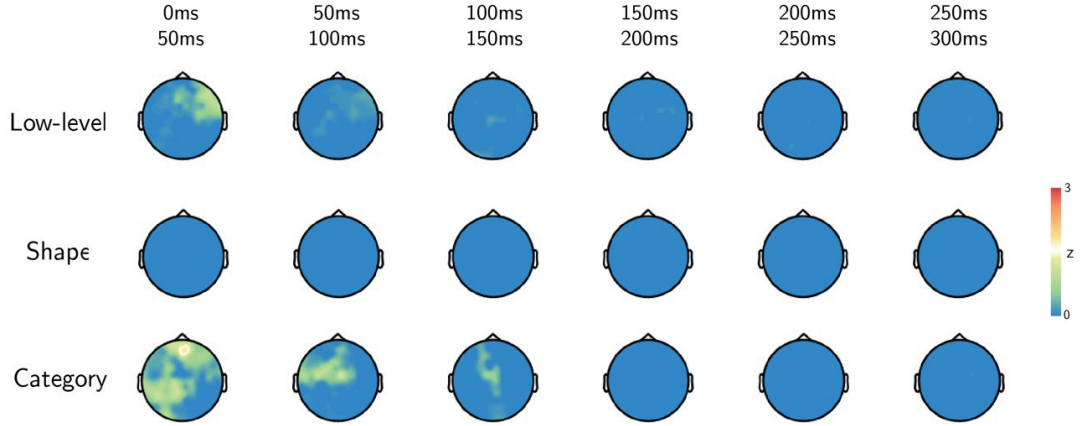


Figure 3.3-1 | Differences between the representation of stimulus features in MEG activity for visually and semantically similar as compared to dissimilar items

The maps report the sensor patterns and time bins in which the weights for each of the three models were higher for similar than dissimilar items (paired rank test, $p<0.2$). Our experimental paradigm aimed to reveal *whether* and *when* low-level, global shape or categorical information are independently processed, while participants were engaged in a semantic judgment task that does not require explicit attention to object shape. However, even if the task was intended to orient subjects' efforts specifically towards high-level semantic processing, it is important to rule out the potential bias on participants' attention towards local features. To this purpose, we

compared the responses between semantically similar and dissimilar stimuli. We partitioned the MEG representation dissimilarity matrices (RDMs) as pertaining to visually and semantically similar (e.g., *fruits vs. vegetables, animals vs. birds, tools vs. vehicles)* or dissimilar (e.g., *vegetables vs. animals, birds vs. fruits, animals vs. vehicles*) comparisons. Since the number of dissimilar comparisons is greater than the number of similar ones, we randomly selected three dissimilar comparisons from the RDMs, to balance the similar ones. Then, we replicated the RWA and the identification of group-level spatiotemporal clusters, as described in the main text, on these partitions of RDMs, and performed a paired rank test between similar and dissimilar conditions for each model. As evident in the topographic plots, there are no significant differences between similar and dissimilar comparisons in any combination of sensors and time-bins. For this reason, we conclude that our results are likely not driven by the perceptual or semantic (dis)similarity between stimuli, excluding therefore a role of task demand.

References

Bar M (2003) A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci 15:600-609.

Biederman I (1987) Recognition-by-Components - a Theory of Human Image Understanding. Psychological Review 94:115-147.

Blum H (1973) Biological shape and visual science. I. J Theor Biol 38:205-287.

Bracci S, Op de Beeck H (2016) Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. J Neurosci 36:432-444.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433-436.

Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area V4. Curr Biol 21:288-293.

Clarke A, Devereux BJ, Randall B, Tyler LK (2015) Predicting the Time Course of Individual Objects with MEG. Cereb Cortex 25:3602-3612.

Craighero L, Fadiga L, Umilta CA, Rizzolatti G (1996) Evidence for visuomotor priming effect. Neuroreport 8:347-349.
DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415-434.
Grezes J, Tucker M, Armony J, Ellis R, Passingham RE (2003) Objects automatically potentiate action: an fMRI study of implicit processing. Eur J Neurosci 17:2735-2740.
Groen, II, Greene MR, Baldassano C, Fei-Fei L, Beck DM, Baker CI (2018) Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7.
Gross J, Baillet S, Barnes GR, Henson RN, Hillebrand A, Jensen O, Jerbi K, Litvak V, Maess B, Oostenveld R, Parkkonen L, Taylor JR, van Wassenhove V, Wibral M, Schoffelen JM (2013) Good practice for conducting and reporting MEG research. Neuroimage 65:349-363.
Guggenmos M, Sterzer P, Cichy RM (2018) Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. Neuroimage 173:434-447.
Handjaras G, Ricciardi E, Leo A, Lenci A, Cecchetti L, Cosottini M, Marotta G, Pietrini P (2016) How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. NeuroImage 135:232-242.
Handjaras G, Leo A, Cecchetti L, Papale P, Lenci A, Marotta G, Pietrini P, Ricciardi E (2017) Modality-independent encoding of individual concepts in the left parietal cortex. Neuropsychologia 105:39-49.
Harel A, Kravitz DJ, Baker CI (2014) Task context impacts visual object processing differentially across the cortex. Proc Natl Acad Sci U S A 111:E962-971.
Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.
Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM (2018) The representational dynamics of task and object processing in humans. eLife 7:e32816.
Hesse JK, Tsao DY (2016) Consistency of Border-Ownership Cells across Artificial Stimuli, Natural Stimuli, and Stimuli with Ambiguous Contours. J Neurosci 36:11338-11349.

Hung CC, Carlson ET, Connor CE (2012) Medial axis shape coding in macaque inferotemporal cortex. Neuron 74:1099-1113.
Johnson JS, Olshausen BA (2003) Timecourse of neural signatures of object recognition. J Vis 3:499-512.
Johnson JW (2000) A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression. Multivariate Behav Res 35:1-19.
Kaiser D, Azzalini DC, Peelen MV (2016) Shape-independent object category responses revealed by MEG and fMRI decoding. J Neurophysiol 115:2246-2250.
Kay KN (2011) Understanding visual representation by developing receptive-field models. Visual population codes: Towards a common multivariate framework for cell recording and functional imaging:133-162.
Kocagoncu E, Clarke A, Devereux BJ, Tyler LK (2017) Decoding the Cortical Dynamics of Sound-Meaning Mapping. J Neurosci 37:1312-1319.
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.
Kubilius J, Wagemans J, Op de Beeck HP (2014) A conceptual framework of computations in mid-level vision. Front Comput Neurosci 8:158.
Lamme VA (1995) The neurophysiology of figure-ground segregation in primary visual cortex. J Neurosci 15:1605-1615.
Lamme VA, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23:571-579.
Leeds DD, Seibert DA, Pyles JA, Tarr MJ (2013) Comparing visual representations across human fMRI and computational vision. J Vis 13:25.
Lescroart MD, Biederman I (2013) Cortical representation of medial axis structure. Cereb Cortex 23:629-637.
Lescroart MD, Stansbury DE, Gallant JL (2015) Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front Comput Neurosci 9:135.

Malcolm GL, Groen IIA, Baker CI (2016) Making Sense of Real-World Scenes. Trends Cogn Sci 20:843-856.
Neri P (2011) Global properties of natural scenes shape local properties of human edge detectors. Front Psychol 2:172.
Neri P (2014) Semantic control of feature extraction from natural scenes. J Neurosci 34:2374-2388.
Neri P (2017) Object segmentation controls image reconstruction from natural scenes. PLoS Biol 15:e1002611.
Nimon KF, Oswald FL (2013) Understanding the results of multiple linear regression: Beyond standardized regression coefficients. Organizational Research Methods 16:650-674.
Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42:145-175.
Olshausen BA, Field DJ (1996a) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607-609.
Olshausen BA, Field DJ (1996b) Natural image statistics and efficient coding. Network 7:333-339.
Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:156869.
Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front Neuroinform 10:27.
Papale P, Leo A, Cecchetti L, Handjaras G, Kay KN, Pietrini P, Ricciardi E (2018) Foreground-Background Segmentation Revealed during Natural Image Viewing. eNeuro 5.
Pernet CR, Latinus M, Nichols TE, Rousselet GA (2015) Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. J Neurosci Methods 250:85-93.
Poort J, Self MW, van Vugt B, Malkki H, Roelfsema PR (2016) Texture Segregation Causes Early Figure 3.Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex. Cereb Cortex 26:3964-3976.
Poort J, Raudies F, Wannig A, Lamme VA, Neumann H, Roelfsema PR (2012) The role of attention in figure-ground

segregation in areas V1 and V4 of the visual cortex. Neuron 75:143-156.
Proklova D, Kaiser D, Peelen MV (2016) Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate-Inanimate Distinction. J Cogn Neurosci 28:680-692.
Qiu FT, Sugihara T, von der Heydt R (2007) Figure-ground mechanisms provide structure for selective attention. Nat Neurosci 10:1492-1499.
Ray-Mukherjee J, Nimon K, Mukherjee S, Morris DW, Slotow R, Hamer M (2014) Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. Methods in Ecology and Evolution 5:320-328.
Rice GE, Watson DM, Hartley T, Andrews TJ (2014) Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. The Journal of Neuroscience 34:8837-8844.
Sebastian TB, Klein PN, Kimia BB (2004) Recognition of shapes by editing their shock graphs. IEEE Trans Pattern Anal Mach Intell 26:550-571.
Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83-98.
Su L, Fonteneau E, Marslen-Wilson W, Kriegeskorte N (2012) Spatiotemporal searchlight representational similarity analysis in EMEG source space. In, pp 97-100: IEEE.
Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, Mitchell T (2012) Tracking neural coding of perceptual and semantic features of concrete nouns. Neuroimage 62:451-463.
Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to Grow a Mind: Statistics, Structure, and Abstraction. Science 331:1279-1285.
Torralba A, Oliva A (2003) Statistics of natural image categories. Network: computation in neural systems 14:391-412.
Van Eede M, Macrini D, Telea A, Sminchisescu C, Dickinson SS (2006) Canonical skeletons for shape matching. In, pp 64-69: IEEE.

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287:1273-1276.
Williford JR, von der Heydt R (2016) Figure-Ground Organization in Visual Cortex for Natural Scenes. eNeuro 3.

Chapter 4

Mutual object representations increase along the visual hierarchy

Introduction

Since the advent of neuroimaging, much effort has been devoted to characterize object-selectivity patterns in the human occitotemporal cortex (OTC; Haxby et al., 2001). Several possible organizing principles have been proposed to explain the large-scale topography of OTC, ranging from the tuning to low-level visual features, e.g., contrast and spatial frequencies (Rajimehr et al., 2011; Rice et al., 2014; Papale et al., 2018), to the processing of broad semantic dimensions, such as object size or the animate-inanimate distinction (Konkle and Caramazza, 2013; Coggan et al., 2016; Julian et al., 2017).

There is little doubt, however, that distinct visual dimensions, ranging from local orientation to global meaning, equally contribute to the striking coherency of our object perception. Thus, to establish the origins of the intrinsic organization in human visual cortex, we need to understand how these dimensions are coded, and how they mutually interact.

Remarkable evidence from previous studies suggests that visual dimensions are indeed highly correlated (Kay, 2011; Bracci and Op de Beeck, 2016; Papale et al., 2019). Consequently, addressing the extent to which brain regions represent these dimensions along the visual hierarchy has proven challenging. In a previous experiment, Bracci and Op de Beeck (2016) employed a set of stimuli in which shape silhouette and category were dissociated (i.e., by selecting objects similar in shape but pertaining to different categories), and demonstrated that object-selectivity in OTC cannot be merely ascribed to visual properties, such as shape silhouette. In another study, Long *et al.* (2018) showed that mid-level features, such as texture and curvature, covary with high-level semantic dimensions, and explain the representations in OTC even when using synthetic and unidentifiable stimuli that hinder object recognition. Hence, even if we acknowledge that shape silhouette (Bracci and Op de Beeck, 2016) and curvature (Long et al., 2018) are relevant to OTC, what is their relative contribution in explaining its activity patterns? For instance, shape silhouette and curvature may capture different aspects of object selectivity, and thus independently contribute to cortical representations, or,

alternatively, only their mutual representation is relevant to OTC that neglects their fine differences.

Another question emerges from the existing literature. Both orthogonal (Bracci and Op de Beeck, 2016) and mutual (Long et al., 2018) representations between different visual dimensions explain to a large extent the patterns of brain responses evoked by viewed objects. However, are different brain regions encoding more orthogonal or mutual information? For instance, high level associative regions may preferentially encode mutual object representations, in order to integrate fragmented descriptions into coherent percepts, while the opposite may hold for early sensory regions, aimed at representing the incoming signal with the highest fidelity.

To answer these questions, we recorded functional MRI (fMRI) activity while participants passively attended to objects. We employed a statistical approach that partitions orthogonal and mutual shape representations revealing their relative impact on brain processing, while controlling at the same time for low- and high-level confounds (Figure 4.1; Lescroart et al., 2015). We found both distinct and overlapping clusters of selectivity in OTC and in parietal regions independently explained by different shape representations (i.e., silhouette, curvature and medial-axis: Figure 4.2). Moreover, we showed that mutual representations linearly increase moving from posterior to anterior regions along the visual hierarchy (Figure 4.3).

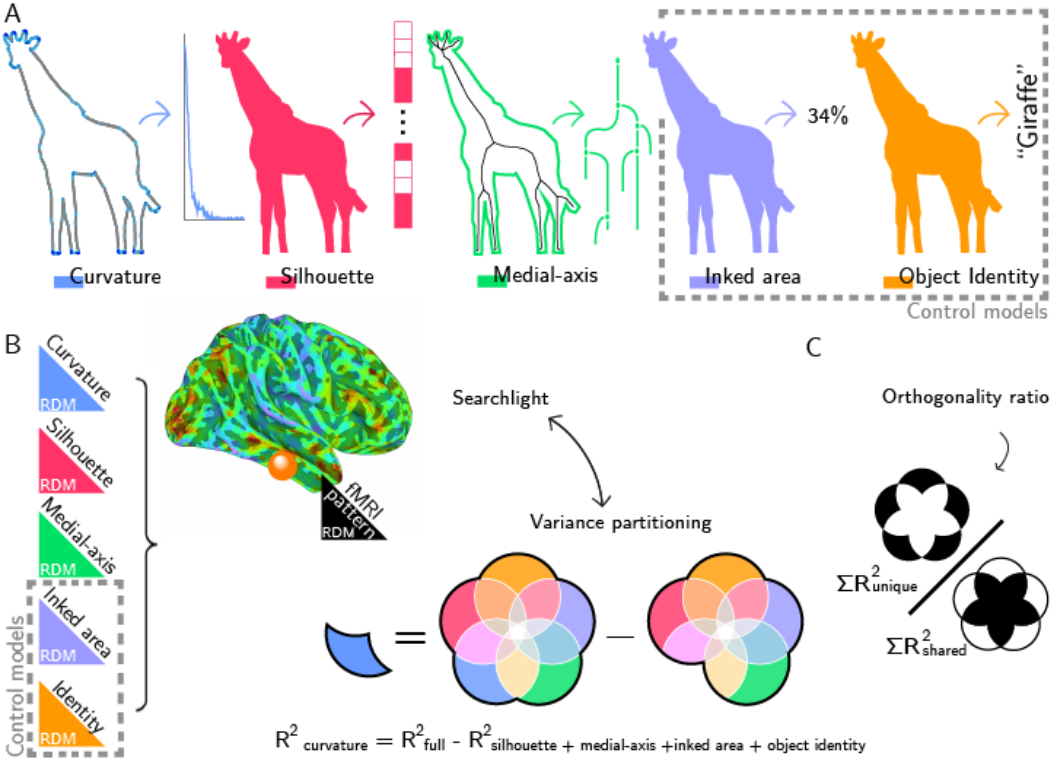


Figure 4.1 | Schematic of the shape models and experiment

A) Five different object representations are employed, three shape models and two further controls. From left: silhouette, medial axis, curvature, inked area (low-level control) and object identity (high-level control). See Figure 4.1-1A for each model representational dissimilarity matrix (RDM).

B) Methodological pipeline. The link between the five model RDMs and each brain activity RDM is computed combining a searchlight procedure with a variance partitioning analysis: within each searchlight, the brain activity RDM is predicted as a combination of the impact of the five models and of their mutual variance. See also Table 4.1-1 and Figure 4.1-1B.

C) For each searchlight, we computed the ratio between the sum of variance explained uniquely by the five models and the sum of variance explained by their mutual components. We called this metric orthogonality ratio (OR) and employed it to reveal the degree to which different brain regions are more tuned to orthogonal over mutual objects representations.

Results

As expected from both theoretical and experimental investigations on this topic (Kay, 2011; Bracci and Op de Beeck, 2016; Papale et al., 2019), the five models show moderate-to-high degrees of collinearity (Figure 4.1-1; Table 4.1-1). Thus, we used a method that accounts for multicollinearity before considering the significance of the association of each model with brain representations. Combining the variance partitioning analysis (Lescroart et al., 2015) and a searchlight procedure (whole brain, 6mm radius: Kriegeskorte et al., 2006), we identified group-level clusters significantly explained by three physiologically validated shape models (i.e., curvature, silhouette and medial-axis) independently from competing representations or confounds (i.e., object category and inked area; Figure 4.1 and 1-1). In addition, we computed the ratio between the sum of variance explained uniquely by the models and variance explained by their mutual components. We called this metric as orthogonality ratio (OR) and employed OR to reveal the degree to which different brain regions encode orthogonal or mutual object-specific representations.

The human visual cortex encodes orthogonal shape representations

Group-level results show both distinct and overlapping clusters of shape selectivity in OTC, mildly extending also to posterior dorsal regions ($p < 0.05$ one-tailed, TFCE corrected). The silhouette model (Figure 4.2, top in red; Table 4.2-1) shows a significant association with brain representations along the Calcarine sulcus (CalcS), the occipitotemporal sulcus (OTS), the right collateral sulcus (CollS), the right inferior temporal sulcus (ITS), the right fusiform gyrus (FusG), the cuneus (Cun) and in

posterior portions of the middle temporal gyrus (pMTG) and intraparietal sulcus (pIPS). The medial-axis (Figure 4.2, middle in green; Table 4.2-1) explains a significant portion of unique variance in the right lateral occipital area (LO) only. Finally, curvature (Figure 4.2, bottom in blue; Table 4.2-1) significantly predicts fMRI representational geometries in the left lingual gyrus (LinG), in the bilateral FusG, along bilateral OTS and ITS, along the right CollS, in the right MTG, bilaterally in the Cun and along the right IPS.

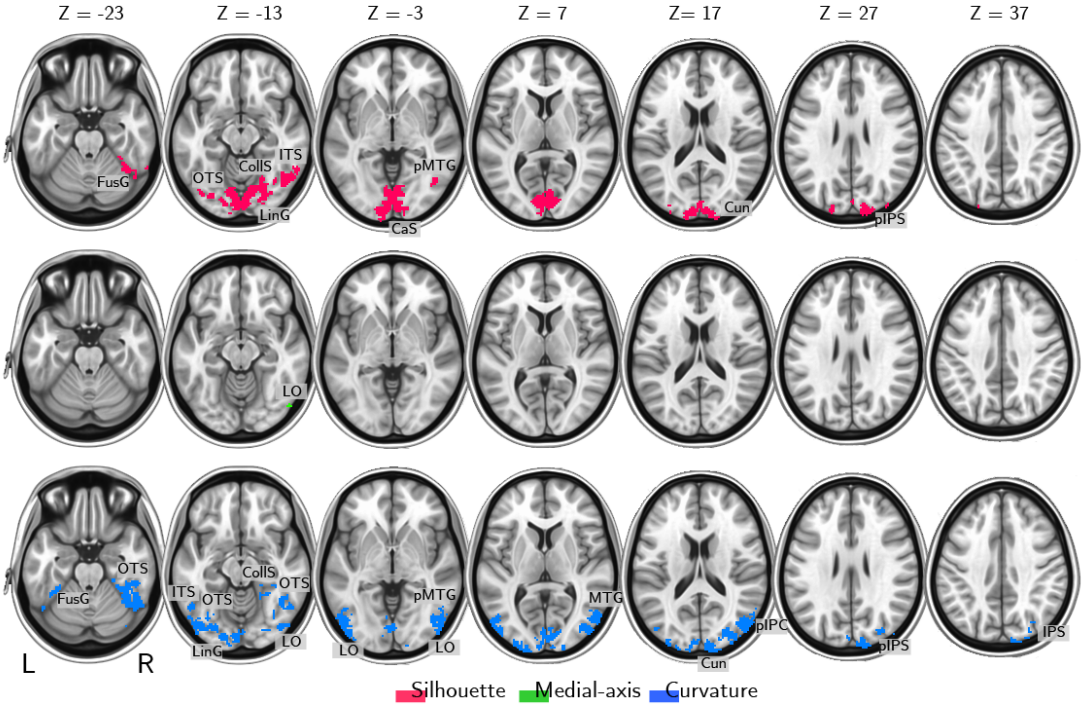


Figure 4.2 | The human visual cortex encodes orthogonal shape representations

Group-level maps showing clusters of shape selectivity in OTC and in posterior dorsal regions (one-tailed $p < 0.05$, TFCE corrected). Selectivity to silhouette (red, top row), medial-axis (green, middle row)

and curvature (blue, bottom row). See also Table 4.2-1 and Figure 4.2-1.

Coding of orthogonal object representations decreases from posterior to anterior regions

The group-level OR map reveals a linear trend in the coding of orthogonal object representations (Figure 4.3). The early visual cortex is biased towards unique components and the OR decreases from posterior to anterior regions in both left (Spearman's $\varrho = -0.67$, $p < 0.001$, parametric test) and right ($\varrho = -0.82$, $p < 0.001$, parametric test) hemispheres. The same trend is present also in left ($\varrho = -0.63$, $p = 0.004$, parametric test) and right ventral ($\varrho = -0.66$, $p < 0.001$, parametric test) regions and in right ($\varrho = -0.75$, $p < 0.001$, parametric test), but not left ($\varrho = -0.48$, $p = 0.178$, parametric test), dorsal areas (Figure 4.3, left and right insets).

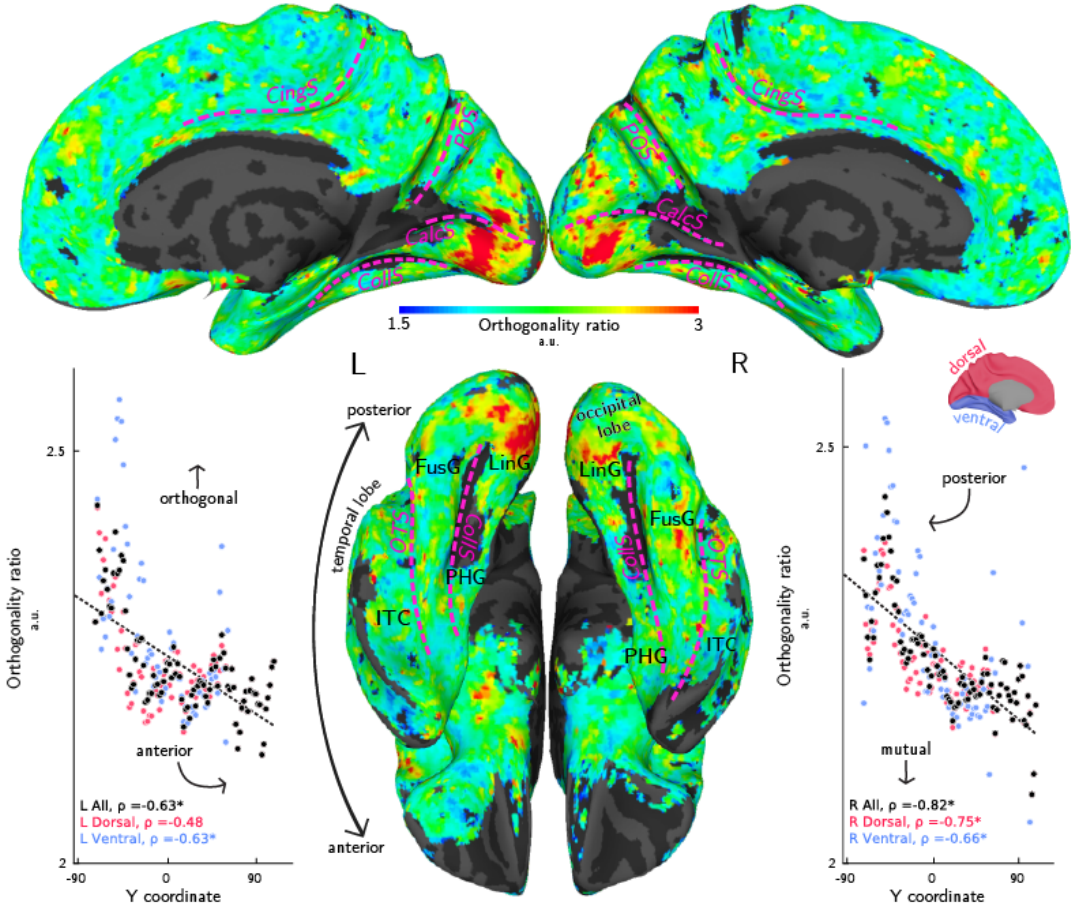


Figure 4.3 | Coding of orthogonal object representations decreases from posterior to anterior regions

Group-level OR map shows that coding of orthogonal descriptions is higher in the early visual cortex, in the left (black dots in left inset) and right (black dots in right inset) hemispheres and in right dorsal (red dots) and right and left ventral portions (blue). Spearman's correlation coefficients are reported at the bottom of left and right insets, while asterisk stands for significant correlations ($p < 0.01$, parametric test).

Discussion

In the present study, we revealed that orthogonal shape representations (i.e., silhouette, medial-axis and curvature) independently contribute to object coding in the human occitotemporal cortex (OTC; Figure 4.2). Moreover, we showed that the brain encodes orthogonal object representations in a topographic fashion: the early visual cortex is biased towards unique components of variance, while mutual representations are progressively more relevant in anterior regions (Figure 4.3).

In line with previous studies, we found that object silhouette is mainly encoded in early visual areas (Figure 4.2, top in red; Table 4.2-1; Khaligh-Razavi and Kriegeskorte, 2014; Bracci and Op de Beeck, 2016; Kaiser et al., 2016; Proklova et al., 2016). This result can be explained by top-down figure-dependent mechanisms that modulate V1 activity both in monkeys (Poort et al., 2016; Self et al., 2019) and humans (Kok and de Lange, 2014; Muckli et al., 2015), and enhances the processing of object-related information in early visual areas also during natural vision (Chapter 2: Papale et al., 2018). However, another possibility may be that the silhouette model better captures the object physical appearance (Table 4.1-1; Kubilius et al., 2016;).

Instead, the variance component unique to the medial-axis model – which is the most transformation-resistant shape description (Yang et al., 2008) – was significant in a smaller extent of cortex comprising only a subset of voxels in right LO (Figure 4.2, middle in green). This can be due to a higher spatial inter-subject variability of this representation that has been already observed by Leeds *et al.* (2013), or to a higher collinearity with the control models we employed (Figure 4.1-1) that prevents from disentangling its contribution from competing representations. Anyway, our result fits previous evidence of medial-axis coding in monkey IT (Hung et al., 2012; putative homologue of human LO), and is consistent with the MEG study presented in Chapter 3, showing that medial-axis processing is limited to a small cluster of right posterior sensors, when controlling for collinearity with low-level and categorical representations (Papale et al., 2019).

Finally, IT (Kayaert et al., 2005a; Yue et al., 2014), LO (Vernon et al., 2016) and FusG (Caldara et al., 2006) were bilaterally tuned to contour curvature (Figure 4.2, bottom in blue), in accordance

with previous neuroimaging investigations. Actually, LO has a pivotal role in object processing (Grill-Spector et al., 1999; Grill-Spector et al., 2001; Kourtzi and Kanwisher, 2001), as IT in monkeys (Desimone et al., 1984; Op de Beeck et al., 2001; Tanaka, 2003; Brincat and Connor, 2004; Kayaert et al., 2005b; Zoccolan et al., 2007). In addition, while we focus our discussion on the ventral stream, we also observed few significant clusters in dorsal visual regions (R pIPS; see Figure 4.2), both for curvature and silhouette, which confirm previous observations (Freud et al., 2017).

Closed shapes can be easily and reliably generated by combining simple elements (i.e., geons or medial axes), by connecting few salient points with acute curvature or by modulating its radial frequency. This may suggest that a unique featural dimension – and maybe a single brain region as V4 or LOC - could critically account for cortical shape representations. However, the evidence that all the tested dimensions independently contribute to shape representation in the human visual cortex favors the hypothesis of a multi-dimensional coding of object shape (Silson et al., 2013; Silson et al., 2016), similarly to what observed for texture processing (Okazawa et al., 2015; Ziemba et al., 2016).

Long *et al.* (2018) suggested that mid-level computations, including curvature extraction, covarying with high-level semantic processing, control the organization of OTC. In the present study, however, we observed overlapping selectivity to orthogonal features in LO (medial-axis and curvature), IT, right FusG, Cun, right pMTG and right pIPS (silhouette and curvature). Since we controlled for collinearity between models, this result could not be merely ascribed to the variance shared by those features. While this may apparently result in contrast with the proposal by Long *et al*. (2018), we also observed that coding of mutual descriptions in OTC is topographically arranged and linearly increases from posterior to anterior regions (Figure 4.3). This observation, consistently with the core finding of Long *et al.* (2018), suggests that the hierarchy of visual processing is not only shaped by specificity to increasingly complex features, but also by a higher selectivity to mutual representations.

This observation complements what has been already observed on the two extremes of the ventral visual pathway: V1

and IT. Representations in V1 are over-complete relative to the retinal input (Olshausen and Field, 1996; Vinje and Gallant, 2000). In addition, inhibitory interactions in V1 are specifically targeted at neurons with similar tuning properties (Chettih and Harvey, 2019). Both these factors increase V1 representational capacity and may ultimately lead to a higher selectivity to orthogonal features, as we observed in posterior regions. On the other hand, higher sensitivity to mutual information in more anterior areas may be produced by populations of neurons that are not tuned to a specific property but that encode multiple dimensions at once. Indeed, shared featural selectivity has been proposed as the mechanism responsible to achieve dimensionality reduction of the sensory input in IT (Lehky et al., 2014), where both neural density and surface are much lower than in V1 (Van Essen et al., 1992; Cahalane et al., 2012). In line with this, the highest dimensional among our three shape models (i.e. silhouette) is also represented in posterior regions (Figure 4.2).

It should be noted that due to the low fMRI temporal resolution we cannot resolve which mechanisms support the different tuning for mutual representations. Moreover, while the selected models capture visual transformations, many alternative descriptions exist (e.g., Khaligh-Razavi and Kriegeskorte, 2014). Overall, however, our results hint at the existence of a multi-dimensional coding of object shape, and reveal that selectivity for mutual object representations is topographically arranged and increases along the visual hierarchy. Future research will identify how different tasks (e.g., determining object similarity vs. extracting affordances), and alternative descriptions impact on the observed patterns of selectivity.

Methods

Subjects

Seventeen subjects were enrolled for the study. Two subjects participated as pilot subjects with a different version of the experimental protocol and their data were not used for the subsequent analyses; data from a subject who abruptly terminated the experiment were discarded. Fourteen subjects

were further considered. The final sample comprised six females, age was 24 ± 3 years, all subjects were right-handed with normal or corrected-to-normal vision and were recruited among the students at the University of Pisa, Italy. Signed informed consent was acquired from all subjects and all the experimental procedures were performed according to the Declaration of Helsinki, under a protocol (1616/2003) approved by the Ethical Committee at the University of Pisa, Italy.

Task

For this study, an event-related design was adopted. Stimulus set consisted of 42 static images of grayscale unfamiliar and common objects, presented against a fixed gray background, with a superimposed fixation cross (size: 2x2°), followed by a baseline condition characterized by a gray screen with a red fixation cross.

A set of stimuli was selected, consisting of 24 common (animate and inanimate) and 18 unfamiliar objects. The latter group represented existing objects which combine the function and the shape of two of the common objects (e.g., a fish-shaped teapot). Of note, a similar criterion has been employed for stimuli selection also in a recent study (Bracci et al., 2017). To build the final set of stimuli, pictures of existing objects were found on internet, resized, normalized for luminance and root-mean-square contrast.

Stimuli were presented with the Presentation software (Neurobehavioral Systems, Albany, CA, USA) on MR-compatible goggles (VisuaStim, Resonance Technology Inc, CA, USA), with a LCD at the resolution of 800x600 pixels (32°x24°). The study was organized in six runs, comprising 56 trials which consisted of 500ms of stimulus presentation and 7000ms of inter-stimulus interval; each run started and ended with 15 seconds of rest, to estimate baseline levels of BOLD signal, and lasted 7:20 minutes. The total duration of the experiment, including anatomical scans, was about 55 minutes.

During the functional runs, subjects were asked to fixate the cross at the center of the screen. On selected trials, the cross changed its color from red to green, and subjects were asked to detect such changes by pressing a key on a MR-compatible

keyboard with the index finger of their dominant hand. Order of trials was randomized across runs, and a different randomization schema was used for each participant.

Functional MRI data acquisition

Data were acquired with a 3-Tesla GE Signa scanner (General Electric Inc., Milwaukee, WI, USA) equipped with an 8-channel phased-array coil. For functional images, a gradient-echo echo-planar imaging sequence (GE-EPI) was used, with TE = 40ms, TR = 2500ms, FA = 90°, 160 volumes with four additional dummy scans, acquisition time 6'50"; image geometry parameters were: Field-Of-View 258x258mm, 128x128 in-plane matrix, voxel size 2.03x2.03x4mm, 37 axial slices for total brain coverage (z-axis extent = 148mm). To acquire detailed information of subject anatomy, a 3D Fast Spoiled Gradient Echo T1-weighted sequence was also acquired (TE = 3.18ms, TR = 8.16ms, FA = 12°, Field-Of-View 256x256mm, 256x256 matrix size, $1mm^3$ isotropic voxels, 256 axial slices, z-axis extent 256mm).

Functional MRI data processing

Data preprocessing was carried out with AFNI (Cox, 1996) and FSL 5.0 (Jenkinson et al., 2012). Preprocessing of functional data comprised slice timing correction with Fourier method (*3dtshift*), rigid-body motion correction using the first volume of the third run as reference (*3dvolreg*), spike removal (elimination of outliers in the functional time series, *3dDespike*), smoothing with a Gaussian filter (fixed FWHM 4 mm, *3dmerge*), scaling of BOLD time series to percentage of the mean of each run (*3dTstat*, *3dcalc*). Processing of anatomical images consisted of brain extraction (*bet*), segmentation for bias-field estimation and removal (*FAST*, *fslmaths*), linear (*FLIRT*) and nonlinear registration (*FNIRT*) to MNI152 standard space.

For each subject, data from the six concatenated runs (960 time points) were used for a GLM analysis (*3dDeconvolve*) with the responses for each stimulus – modeled with 1 seconds-long block functions convolved with a canonical HRF – as predictors of interest, and the six motion parameters plus polynomial trends up to 4th order as predictors of no-interest.

Responses for individual stimuli were converted to MNI152 space by applying the transformation matrices estimated as explained above, and resampled to a resolution of 2x2x2mm.

Shape models and controls

Five different representations of the 42 stimuli were developed: three shape-based descriptions of interest and two further controls. For each model, we obtained a stimulus-specific feature space, and pairwise dissimilarities between stimuli were computed to obtain a representational dissimilarity matrix (RDM). Before computing shape-related information, stimuli were binarized.

A first shape description was computed by extracting the silhouette, consisting of a simple stimulus vectorization. The link between shape silhouette and OTC representations has been extensively investigated in neuroimaging studies (Khaligh-Razavi and Kriegeskorte, 2014; Bracci and Op de Beeck, 2016; Kaiser et al., 2016; Proklova et al., 2016). Pairwise dissimilarity was computed using correlation distance (1 - Pearson's rho). Second, a skeletal representation of each stimulus was extracted by performing the medial axis transform (Blum, 1973). It controls the spike rate of IT neurons in monkey (Hung et al., 2012), captures behavioral ratings of shape similarity (Lowet et al., 2018) and its spatiotemporal association with brain activity in humans has been described in several neuroimaging studies (Leeds et al., 2013; Lescroart and Biederman, 2013; Handjaras et al., 2017; Papale et al., 2019). Pairwise distance between skeletal representations was computed using the ShapeMatcher algorithm (http://www.cs.toronto.edu/~dmac/ShapeMatcher/; (Van Eede et al., 2006)). In sum, the ShapeMatcher algorithm builds the shock-graphs of each shape and then estimates their dissimilarity as the deformation required to match different objects (Sebastian et al., 2004). A third description was obtained by computing the curvature distribution for each object's contour. It has been showed that V4 neurons in monkey are selective to a specific degree of curvature (Cadieu et al., 2007; Connor et al., 2007; Carlson et al., 2011). Moreover, the pivotal role of contour curvature in object perception has been extensively demonstrated both by behavioral (Wolfe et al., 1992;

Elder and Velisavljevic, 2009; Lawrence et al., 2016; Long et al., 2017) and neuroimaging studies in humans (Caldara et al., 2006; Yue et al., 2014; Vernon et al., 2016; Long et al., 2018). Curvature was computed as the chord-to-point distance (Monroy et al., 2011) in a 40-pixels window. Pairwise dissimilarity was computed using correlation distance. Finally, two further control RDMs were built (Figure 4.1-1). The area (in pixels) of each stimulus was computed to account for the inked-area bias – a problem that is almost unavoidable when using complex objects in isolation (but see Bracci and Op de Beeck, 2016 for an elegant stimulus design). For the inked-area bias, pairwise dissimilarity was computed as the Euclidean distance. In addition, to get rid of high-level biases that could affect the prediction performance of the three shape models, object category was included as a further control. For identity, a binary representation was employed (Kriegeskorte et al., 2008; Khaligh-Razavi and Kriegeskorte, 2014). Unfamiliar stimuli were considered as belonging to categories according to both their function and shape (Figure 4.1-1).

Shape selectivity

A variance partitioning analysis (Lescroart et al., 2015) was performed to determine whether the three shape models in this study significantly predict unique components of the variance of brain representations, as computed in 6 mm-radius spherical searchlights (Kriegeskorte et al., 2006). To this aim, explained variance coefficient (R^2) was computed for each model RDM in independent linear regressions, and then all the different combinations of models were tested in further multiple linear regressions. The final statistic reporting the partial goodness of fit for unique and shared components was computed following the work by Nimon and colleagues (2008). To exemplify, the unique variance explained by the curvature model in a specific searchlight was determined as the difference between the full-model R^2 and the variance explained by the combination of all other models (i.e., $R^2_{\text{curvature}} = R^2_{\text{full}} - R^2_{\text{silhouette + medial-axis +inked area + identity}}$). In the context of multiple linear regression, this approach is better known as 'commonality analysis' (Nimon and Oswald,

2013), and its popularity is growing in neuroimaging (Lescroart et al., 2015; de Heer et al., 2017; Groen et al., 2018).

Correlation distance was used to compute the RDM of fMRI activity patterns in each searchlight and only voxels pertaining to the cerebral cortex with a probability higher than 50% were included in the procedure. The z-scored partial correlation coefficient (de Heer et al., 2017) for each component of unique and shared variance were then assigned to the center of the searchlight, so obtaining a map for each subject and component. For each model, threshold free cluster enhancement (TFCE: Smith and Nichols, 2009) was used to detect group-level clusters significantly predicted by the corresponding unique variance component (5000 randomizations with 6mm variance smoothing, as implemented in FSL's *randomise*: www.fmrib.ox.ac.uk/fsl/randomise). Statistical maps were then thresholded at one-tailed $p < 0.05$, corrected for multiple comparison across gray matter voxels (minimum cluster size = 10 voxels; Figure 4.2 and 2-2).

Orthogonality ratio

Orthogonality ratio (OR) was computed by dividing the group-averaged sum of variance explained uniquely by the five models with the group-averaged sum of variance explained by their shared components for each searchlight (cortical map in Figure 4.3). In addition, a linear trend between the Y coordinate and the mean OR in each XZ-slice was computed in each hemisphere and 4 ROIs using the Spearman's correlation (left and right insets in Figure 4.3) and significance was then computed with a parametric test. The ROIs comprised the left and right hemispheres, and ventral and dorsal regions within the two hemispheres. Voxels above the center of the CalcS and superior to the Sylvian fissure were considered part of the dorsal ROIs, while occipital regions below the center of the CalcS and the whole temporal lobe were considered part of the ventral ROIs.

Second-level analyses were performed using custom-made code written in MATLAB (MathWorks Inc.).

Extended Data

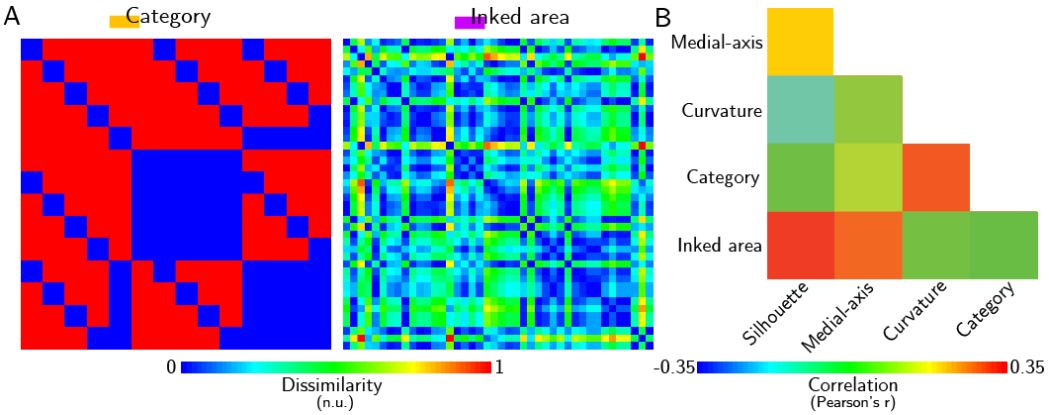


Figure 4.1-1 related to Figure 4.1 | Control RDMs and model collinearity

A) RDMs of the two control models (category and inked area)

B) Similarity matrix showing the pairwise correlation between models (Pearson's r).

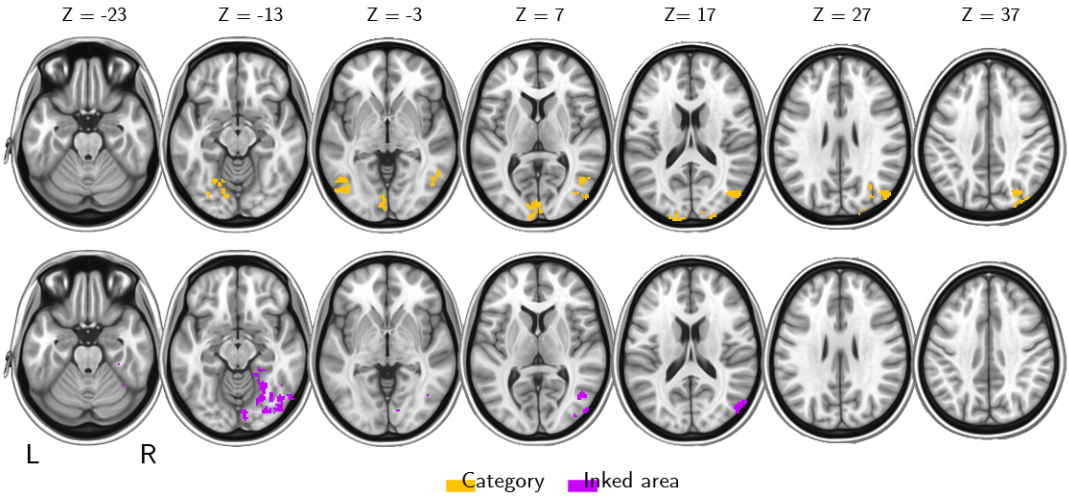


Figure 4.2-1 related to Figure 4.2 | Control model selectivity
Group-level statistical maps thresholded at $p < 0.05$ (one-tailed, TFCE corrected). From above: sensitivity to object category (yellow) and inked area (purple).

Table 4.1-1 related to Figure 4.1 | Correlation with low-level features
Pearson's correlation coefficient between the five models and the GIST model (Oliva and Torralba, 2001).

Silhouette	Medial-axis	Curvature	Category	Inked area
0.75	0.12	-0.14	0.11	0.11

Table 4.2-2 related to Figure 4.2 | Shape selectivity

MNI coordinates of the centers of mass of significant clusters for the shape models.

	Voxels	CMass x	CMass y	CMass z
Silhouette	3556	+9.7	-80.5	-2.7
	71	-19.9	-90.5	+30.0
	21	-31.0	-91.5	+20.4
	20	-40.4	-77.3	-10.7
	15	+36.1	-87.6	+29.7
	12	-36.3	-75.7	-18.7
Medial-axis	12	-50.0	-73.7	-12.8
Curvature	3387	+32.0	-72.7	+2.4
	1255	-40.3	-76.6	-6.2
	193	-26.1	-96.8	+13.4
	72	+26.0	-80.3	-10.8
	51	-36.3	-42.7	-23.1
	44	+28.4	-67.7	+34.2
	11	-42.7	-50.2	-16.0
	10	+14.2	-97.4	-4.6

References

Blum H (1973) Biological shape and visual science. I. J Theor Biol 38:205-287.

Bracci S, Op de Beeck H (2016) Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. J Neurosci 36:432-444.

Bracci S, Kalfas I, de Beeck HO (2017) The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. bioRxiv:228932.

Brincat SL, Connor CE (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci 7:880-886.

Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T (2007) A model of V4 shape selectivity and invariance. J Neurophysiol 98:1733-1750.

Cahalane DJ, Charvet CJ, Finlay BL (2012) Systematic, balancing gradients in neuron density and number across the primate isocortex. Front Neuroanat 6:28.
Caldara R, Seghier ML, Rossion B, Lazeyras F, Michel C, Hauert CA (2006) The fusiform face area is tuned for curvilinear patterns with more high-contrasted elements in the upper part. Neuroimage 31:313-319.
Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area V4. Curr Biol 21:288-293.
Chettih SN, Harvey CD (2019) Single-neuron perturbations reveal feature-specific competition in V1. Nature 567:334-340.
Coggan DD, Liu W, Baker DH, Andrews TJ (2016) Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. Neuroimage 135:107-114.
Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. Curr Opin Neurobiol 17:140-147.
Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162-173.
de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The Hierarchical Cortical Organization of Human Speech Processing. J Neurosci 37:6539-6557.
Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051-2062.
Elder JH, Velisavljevic L (2009) Cue dynamics underlying rapid detection of animals in natural scenes. J Vis 9:7.
Freud E, Culham JC, Plaut DC, Behrmann M (2017) The large-scale organization of shape processing in the ventral and dorsal pathways. Elife 6.
Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. Vision Res 41:1409-1422.
Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R (1999) Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24:187-203.

Groen, II, Greene MR, Baldassano C, Fei-Fei L, Beck DM, Baker CI (2018) Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7.
Handjaras G, Leo A, Cecchetti L, Papale P, Lenci A, Marotta G, Pietrini P, Ricciardi E (2017) Modality-independent encoding of individual concepts in the left parietal cortex. Neuropsychologia 105:39-49.
Hung CC, Carlson ET, Connor CE (2012) Medial axis shape coding in macaque inferotemporal cortex. Neuron 74:1099-1113.
Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012) Fsl. Neuroimage 62:782-790.
Julian JB, Ryan J, Epstein RA (2017) Coding of Object Size and Object Category in Human Visual Cortex. Cereb Cortex 27:3095-3109.
Kaiser D, Azzalini DC, Peelen MV (2016) Shape-independent object category responses revealed by MEG and fMRI decoding. J Neurophysiol 115:2246-2250.
Kay KN (2011) Understanding visual representation by developing receptive-field models. Visual population codes: Towards a common multivariate framework for cell recording and functional imaging:133-162.
Kayaert G, Biederman I, Vogels R (2005a) Representation of regular and irregular shapes in macaque inferotemporal cortex. Cereb Cortex 15:1308-1321.
Kayaert G, Biederman I, Op de Beeck HP, Vogels R (2005b) Tuning for shape dimensions in macaque inferior temporal cortex. Eur J Neurosci 22:212-224.
Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation.
Kok P, de Lange FP (2014) Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. Curr Biol 24:1531-1535.
Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. J Neurosci 33:10235-10242.

Kourtzi Z, Kanwisher N (2001) Representation of perceived object shape by the human lateral occipital complex. Science 293:1506-1509.
Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103:3863-3868.
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.
Kubilius J, Bracci S, Op de Beeck HP (2016) Deep Neural Networks as a Computational Model for Human Shape Sensitivity. PLOS Comput Biol 12:e1004896.
Lawrence SJ, Keefe BD, Vernon RJ, Wade AR, McKeefry DJ, Morland AB (2016) Global shape aftereffects in composite radial frequency patterns. J Vis 16:17.
Leeds DD, Seibert DA, Pyles JA, Tarr MJ (2013) Comparing visual representations across human fMRI and computational vision. J Vis 13:25.
Lehky SR, Kiani R, Esteky H, Tanaka K (2014) Dimensionality of object representations in monkey inferotemporal cortex. Neural Comput 26:2135-2162.
Lescroart MD, Biederman I (2013) Cortical representation of medial axis structure. Cereb Cortex 23:629-637.
Lescroart MD, Stansbury DE, Gallant JL (2015) Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front Comput Neurosci 9:135.
Long B, Stormer VS, Alvarez GA (2017) Mid-level perceptual features contain early cues to animacy. J Vis 17:20.
Long B, Yu CP, Konkle T (2018) Mid-level visual features underlie the high-level categorical organization of the ventral stream. Proc Natl Acad Sci U S A 115:E9015-E9024.
Lowet AS, Firestone C, Scholl BJ (2018) Seeing structure: Shape skeletons modulate perceived similarity. Atten Percept Psychophys 80:1278-1289.
Monroy A, Eigenstetter A, Ommer B (2011) Beyond straight lines—object detection using curvature. In, pp 3561-3564: IEEE.

Muckli L, De Martino F, Vizioli L, Petro LS, Smith FW, Ugurbil K, Goebel R, Yacoub E (2015) Contextual Feedback to Superficial Layers of V1. Curr Biol 25:2690-2695.
Nimon K, Lewis M, Kane R, Haynes RM (2008) An R package to compute commonality coefficients in the multiple regression case: an introduction to the package and a practical example. Behav Res Methods 40:457-466.
Nimon KF, Oswald FL (2013) Understanding the results of multiple linear regression: Beyond standardized regression coefficients. Organizational Research Methods 16:650-674.
Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci U S A 112:E351-360.
Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42:145-175.
Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607-609.
Op de Beeck H, Wagemans J, Vogels R (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat Neurosci 4:1244-1252.
Papale P, Leo A, Cecchetti L, Handjaras G, Kay KN, Pietrini P, Ricciardi E (2018) Foreground-Background Segmentation Revealed during Natural Image Viewing. eNeuro 5.
Papale P, Betta M, Handjaras G, Malfatti G, Cecchetti L, Rampinini A, Pietrini P, Ricciardi E, Turella L, Leo A (2019) Common spatiotemporal processing of visual features shapes object representation. Sci Rep 9:7601.
Poort J, Self MW, van Vugt B, Malkki H, Roelfsema PR (2016) Texture Segregation Causes Early Figure 4.Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex. Cereb Cortex 26:3964-3976.
Proklova D, Kaiser D, Peelen MV (2016) Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate-Inanimate Distinction. J Cogn Neurosci 28:680-692.
Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH (2011) The "Parahippocampal Place Area" Responds

Preferentially to High Spatial Frequencies in Humans and Monkeys. PLoS Biol 9:e1000608.
Rice GE, Watson DM, Hartley T, Andrews TJ (2014) Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. The Journal of Neuroscience 34:8837-8844.
Sebastian TB, Klein PN, Kimia BB (2004) Recognition of shapes by editing their shock graphs. IEEE Trans Pattern Anal Mach Intell 26:550-571.
Self MW, Jeurissen D, van Ham AF, van Vugt B, Poort J, Roelfsema PR (2019) The Segmentation of Proto-Objects in the Monkey Primary Visual Cortex. Curr Biol 29:1019-1029 e1014.
Silson EH, Groen, II, Kravitz DJ, Baker CI (2016) Evaluating the correspondence between face-, scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. J Vis 16:14.
Silson EH, McKeefry DJ, Rodgers J, Gouws AD, Hymers M, Morland AB (2013) Specialized and independent processing of orientation and shape in visual field maps LO1 and LO2. Nat Neurosci 16:267-269.
Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83-98.
Tanaka K (2003) Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cereb Cortex 13:90-99.
Van Eede M, Macrini D, Telea A, Sminchisescu C, Dickinson SS (2006) Canonical skeletons for shape matching. In, pp 64-69: IEEE.
Van Essen DC, Anderson CH, Felleman DJ (1992) Information processing in the primate visual system: an integrated systems perspective. Science 255:419-423.
Vernon RJ, Gouws AD, Lawrence SJ, Wade AR, Morland AB (2016) Multivariate Patterns in the Human Object-Processing Pathway Reveal a Shift from Retinotopic to Shape Curvature Representations in Lateral Occipital Areas, LO-1 and LO-2. J Neurosci 36:5763-5774.

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287:1273-1276.
Wolfe JM, Yee A, Friedman-Hill SR (1992) Curvature is a basic feature for visual search tasks. Perception 21:465-480.
Yang M, Kpalma K, Ronsin J (2008) A survey of shape feature extraction techniques: In-Tech.
Yue X, Pourladian IS, Tootell RB, Ungerleider LG (2014) Curvature-processing network in macaque visual cortex. Proc Natl Acad Sci U S A 111:E3467-3475.
Ziemba CM, Freeman J, Movshon JA, Simoncelli EP (2016) Selectivity and tolerance for visual texture in macaque V2. Proc Natl Acad Sci U S A 113:E3140-3149.
Zoccolan D, Kouh M, Poggio T, DiCarlo JJ (2007) Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci 27:12292-12307.

Chapter 5

Conclusions

Summary

In the present thesis, we investigated how different object properties are represented in our visual cortex in natural vision. In the first introductory chapter, we described how our reliable visual system transforms continuous retinal signaling into meaningful objects. In addition, we showed why this process is challenging: the retinal input is often an unreliable source of information about object shape and location; and in the surrounding environment, behaviorally relevant properties are often mutually correlated (e.g., shape and semantic category).

In Chapter 2, we presented an fMRI study on scene segmentation (Papale et al., 2018). Segmentation has been considered critical to form discrete object representations from continuous sensory percepts. In this first study, we analyzed brain responses during passive natural image viewing. Subjects attended to hundreds of natural scenes and we derived brain representations from each occipital region and compared them to parametric representations, so to reveal the inner filtering operated by each brain region. In contrast to strictly hierarchical and compartmentalized views on brain selectivity, our findings provide novel support to the hypothesis that foreground-background segmentation of natural scenes occurs during passive perception, sustained by the distributed activity of multiple areas in the occipital lobe. In fact, while foreground information is enhanced along the entire visual pathway, mid-level regions show a background suppression effect, though retaining low-level information from the foreground.

However, due to the low temporal resolution of fMRI, the first study alone cannot resolve if those shared computations reflect a common spatiotemporal process. Thus, in a second MEG study, presented in Chapter 3, we revealed the temporal dynamics of feature processing in human subjects attending to objects from six semantic categories (Papale et al., 2019). We mitigated collinearity between model-based descriptions of stimuli and showed that low-level properties (contrast and spatial frequencies), shape (medial-axis) and category are represented within the same spatial locations early in time: 100-150ms after stimulus onset. This fast and overlapping processing may result from independent parallel computations, with

categorical representation emerging later than the onset of low-level feature processing, yet before shape coding. Categorical information is represented both before and after shape, suggesting a role for this feature in the refinement of categorical matching.

Accordingly, the same features can be retrieved in the activity of multiple regions, and orthogonal components of those features are processed by the same cortical structures at the same latencies. Consequently, is there a broad organization determining these observations? What is the link between coding of mutual and orthogonal object representations? This overlap may be explained either by mere collinearity across representations, or may instead reflect the coding of multiple dimensions by the same cortical population. Moreover, independent and mutual components may differently impact on distinctive stages of the visual hierarchy. To answer those questions, we performed a third experiment, presented in Chapter 4: we recorded fMRI activity while participants passively attended to objects, and employed a statistical approach that partition orthogonal and mutual object representations to reveal their relative impact on brain processing. Orthogonal shape representations (i.e., silhouette, curvature and medial-axis) independently explain distinct and overlapping clusters of selectivity in occitotemporal and parietal cortex. Moreover, we showed that the relevance of mutual representations linearly increases moving from posterior to anterior regions.

Searching for an explanation

Visualizing and understanding the complex mental representations our brain relies upon is an open and demanding challenge. In the present work, we introduced a novel method to (literally) figure out how each cortical region is transforming the retinal input (i.e., the correlation images in Chapter 2). Moreover, we implemented existing methods to the context of neuroimaging (i.e., the relative weight analysis in Chapter 3) and adapted them to answer novel questions (the orthogonality ratio in Chapter 4).

Overall, these results depict a complex picture of our visual cortex. First, there is not a clear selectivity hierarchy, but information spreads between regions: early cortical areas access to high-level representations and are involved in complex cognitive tasks (i.e., object segmentation), and vice versa. Second, concurrent processing of orthogonal object shape, contrast and category representations is fast (100-150ms) in the right posterior brain. And third, the visual cortex encodes mutual relations between different features in a topographic fashion while object shape is encoded along different dimensions, each representing orthogonal features.

A tentative and preliminary explanation of these results may be the following. At first, each region performs its specific inner transformation (e.g., edge detection and normalization) of the incoming image, signaling it to downstream regions in a feedforward manner – similarly to artificial neural networks, these transformations are probably highly collinear in nearby levels. Then, accessing different representations thru feedback projections quickly reshapes the local selectivity of each region. And finally, the visual system may take advantage of those interactions to increase the sensitivity to (mutual) object representations matching prior knowledge or behavioral needs.

This process may either be static, i.e. depending on global structural properties like the lower neuronal density in higher regions (Van Essen et al., 1992; Cahalane et al., 2012), or more likely, may be task-dependent. In this view, earlier regions may compute full, reliable and redundant representations of the retinal input, that are then decomposed and passed on to downstream regions depending on a broader task-dependent gating mechanism, that filters out (or actively suppress) all the irrelevant pieces of information that are orthogonal to the task (similarly to what proposed in: Roelfsema and de Lange, 2016). In this view, different components of variance may be selected by different task configurations.

However, future research is needed to validate/confute this perspective, involving more direct measures of *in vivo* neural activity (e.g., extracellular recordings) and possibly, combinations of both upstream and downstream approaches.

In this journey to a full understating of the visual system, a long road ahead awaits us.

References

Cahalane DJ, Charvet CJ, Finlay BL (2012) Systematic, balancing gradients in neuron density and number across the primate isocortex. Front Neuroanat 6:28.

Papale P, Leo A, Cecchetti L, Handjaras G, Kay KN, Pietrini P, Ricciardi E (2018) Foreground-Background Segmentation Revealed during Natural Image Viewing. eNeuro 5.

Papale P, Betta M, Handjaras G, Malfatti G, Cecchetti L, Rampinini A, Pietrini P, Ricciardi E, Turella L, Leo A (2019) Common spatiotemporal processing of visual features shapes object representation. Sci Rep 9:7601.

Roelfsema PR, de Lange FP (2016) Early Visual Cortex as a Multiscale Cognitive Blackboard. Annu Rev Vis Sci 2:131-151.

Van Essen DC, Anderson CH, Felleman DJ (1992) Information processing in the primate visual system: an integrated systems perspective. Science 255:419-423.