PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Opinion piece

Cite this article: Groen IIA, Silson EH, Baker CI. 2017 Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Phil. Trans. R. Soc. B* **372**: 20160102.
http://dx.doi.org/10.1098/rstb.2016.0102

Accepted: 20 July 2016

One contribution of 15 to a theme issue 'Auditory and visual scene analysis'.

Subject Areas:
neuroscience

Keywords:
natural scenes, functional magnetic resonance imaging, electro-encephalography, category-selectivity, retinotopy, image statistics

Author for correspondence:
Iris I. A. Groen
e-mail: iris.groen@nih.gov

Contributions of low- and high-level properties to neural processing of visual scenes in the human brain

Iris I. A. Groen, Edward H. Silson and Chris I. Baker

Laboratory of Brain and Cognition, National Institutes of Health, 10 Center Drive 10-3N228, Bethesda, MD, USA

IIAG, 0000-0002-5536-6128; EHS, 0000-0002-6149-7423; CIB, 0000-0001-6861-8964

Visual scene analysis in humans has been characterized by the presence of regions in extrastriate cortex that are selectively responsive to scenes compared with objects or faces. While these regions have often been interpreted as representing high-level properties of scenes (e.g. category), they also exhibit substantial sensitivity to low-level (e.g. spatial frequency) and mid-level (e.g. spatial layout) properties, and it is unclear how these disparate findings can be united in a single framework. In this opinion piece, we suggest that this problem can be resolved by questioning the utility of the classical low- to high-level framework of visual perception for scene processing, and discuss why low- and mid-level properties may be particularly diagnostic for the behavioural goals specific to scene perception as compared to object recognition. In particular, we highlight the contributions of low-level vision to scene representation by reviewing (i) retinotopic biases and receptive field properties of scene-selective regions and (ii) the temporal dynamics of scene perception that demonstrate overlap of low- and mid-level feature representations with those of scene category. We discuss the relevance of these findings for scene perception and suggest a more expansive framework for visual scene analysis.

This article is part of the themed issue 'Auditory and visual scene analysis'.

1. Introduction

Our natural visual input—what we might intuitively refer to as a visual scene—consists of a complex array of reflected light from the objects and surfaces that constitute our daily environment. This scene information impinges on our retinae, stimulating neural signals that ultimately give rise to internal representations that enable adaptive behaviour. Although the sensation of seeing occurs extremely rapidly and is seemingly effortless, the visual input is highly complex and dynamic, changing with each fixation as we move our eyes. Correspondingly, the underlying neural architecture comprises multiple stages of processing from the retina through sub-cortical structures to the cortex, where multiple distinct visual areas have been defined. In particular, functional magnetic resonance imaging (fMRI) studies in humans have identified scene-selective areas that respond more when viewing scenes compared with objects or faces and may be specialized for representing specific aspects of the environment. However, the exact contribution of these areas to scene perception is just beginning to be explored, and there is considerable debate about what scene information is processed in each. In this opinion piece, we shall discuss scene analysis in the context of the classic hierarchical view of visual processing and argue that the distinctions that are often made between low- and high-level processing may not be applicable directly to scene perception.

Despite the intuition for what constitutes a scene, the concept itself is notoriously difficult to define. For our current purposes, it is useful to distinguish the notion of a scene from that of an object. Experimentally, scenes and objects are operationalized very differently. While objects are often presented as segmented, isolated stimuli, scenes are commonly everyday photographs selected to be representative of

THE ROYAL SOCIETY PUBLISHING

real-world environments. Thus, while objects have a bounding contour, defining a shape, the boundary of a scene stimulus is typically arbitrary (e.g. circular or rectangular aperture). In some cases, a collection of objects is considered a scene (e.g. visual search arrays [1]) or objects are pasted onto a background, (e.g. [2]). However, the natural scene input is more than just a collection of objects, and objects are not always easily segmented from the background. Theoretically, scenes are thought to form a 'context' to objects, their structural embedding [3]. However, scenes also encompass these objects, i.e. the objects are an integral part of the scene. While some proposals suggest that scenes can be seen as the sum of their constituent objects (e.g. sand + water + palm tree = beach, e.g. [4]), others have pointed out that scenes cannot merely be reduced to a list of object labels [5]: they have features that are not necessarily object-bound such as spatial layouts, boundaries and textures, which are somehow combined seamlessly into a coherent scene percept.

Scenes and objects can also be distinguished in terms of relative scale. While scenes can typically be thought to encompass the whole visual field, objects are smaller components of a scene that are commonly fixated. Thus, while a specific object might dominate high acuity foveal vision, the scene in which that object is located also engages lower acuity peripheral vision, which has important implications for the type of information that can be extracted.

Given all these considerations, for the purpose of this opinion piece, we shall adopt the definition of a scene as a multi-element stimulus that encompasses both foveal and peripheral vision and contains both object and spatial features with real-world contextual associations between them. To understand how visual scene analysis is implemented in the brain, we shall next briefly review the classic hierarchical view of visual processing, before discussing scene processing in particular.

2. Hierarchical view of visual processing

Visual processing is typically considered in a hierarchical framework, comprising a series of discrete stages that successively produce increasingly abstract representations [6]. These different stages are often considered in terms of low-, mid- and high-level representations (figure 1). Low-level vision is thought to involve the representation of elementary features, such as local colour, luminance or contrast. Such processing is typically linked to the flow of information to primary visual cortex (V1) via the retinogeniculate and geniculostriate pathways (see also [7]), which translate light intensity at the retina into an orientated edge representation by means of small receptive fields (RFs) tiling the entire visual field [8].

The immediate stages beyond V1, V2–V4 are often considered to encompass mid-level vision, but are comparatively much less well understood [9]. Overall, these areas can be thought of as containing the representation of conjunctions of elementary features and properties such as surfaces, higher order image statistics, disparity and intermediate shape features [10–12]. Recent studies have linked fMRI responses from these areas to representation of locally pooled low-level representations in computational models [13–15].

Finally, high-level vision is considered to reflect abstraction of the visual input into categorical or semantic representations that enable classification or identification. One of the major challenges and goals of high-level vision is often stated to be the building of *invariant* representations that are robust to incidental viewing conditions, such as illumination, size and position [16]. One of the most striking findings in visual neuroscience is that multiple distinct brain regions exhibit selective and highly reliable responses to stimuli from particular categories. Indeed, monkey inferotemporal cortex contains neurons that respond selectively to specific objects and categories, and human neuroimaging has identified entire brain regions that respond selectively to faces, bodies and objects [17].

Interestingly, three *scene-selective* brain regions have also been identified by contrasting responses elicited by viewing scenes compared to objects or faces: the parahippocampal place area (PPA), occipital place area (OPA) and retrosplenial complex (RSC) [18]. How these regions fit into the larger hierarchical framework of visual processing is, however, a topic of considerable debate, as we shall discuss in §3.

3. Is scene perception low- or high-level?

It is often reported that at least some of the scene-selective regions contain high-level representations of scenes; for example, scene category [19,20], contextual associations [21] or familiar places and landmarks [22,23] that generalize across viewpoints [24]. On the one hand, this view makes a lot of sense: scenes can be considered the ultimate goal of abstraction, a higher level of representation than even individual objects, residing 'above' the low-, mid- and high-levels of information depicted in figure 1, for example representing object co-occurrence statistics [4]. On the other hand, however, a number of studies have shown that these areas are also sensitive to what in the classic view would be considered low-level features, such as spatial frequency, line orientation, contrast and texture [25–28]. Others have stressed the role of mid-level features [29] or spatial properties that might depend on such features [30–32].

As a result of these disparate findings, the role of scene-selective regions is heavily debated. The low-level view is supported by the finding that responses in scene-selective regions can be elicited using minimal stimuli [25,26] or can be explained by low-level models of scene representation [33]. However, responses to low-level information are not necessarily sufficient to explain scene selectivity [34], and high-level scene percepts have been reported for small sets of carefully controlled stimuli in which low-level differences were minimized [35]. Moreover, a common argument in favour of high-level representations comes from running control experiments in which low-level features are either isolated or superimposed (e.g. [32]), often contrasting response patterns from scene-selective areas with those of earlier stages in visual processing, to highlight that the representations are not the same as in V1 [19,30].

How to resolve this debate? It is important to realize that the standard framework of visual perception sketched above is largely informed by object recognition [36], where the goal of the system is typically characterized as achieving a categorical or identity label for an individual object (figure 1, left). To achieve such an invariant representation, low-level information is often assumed to be an impediment that needs to be 'thrown out' in order to distil the abstract meaning of stimuli. Although theoretical frameworks involving three-dimensional high-level models that need to be explicitly matched to low-level input [37] have become more relaxed in favour of a distributed coding framework in which objects are represented via neural

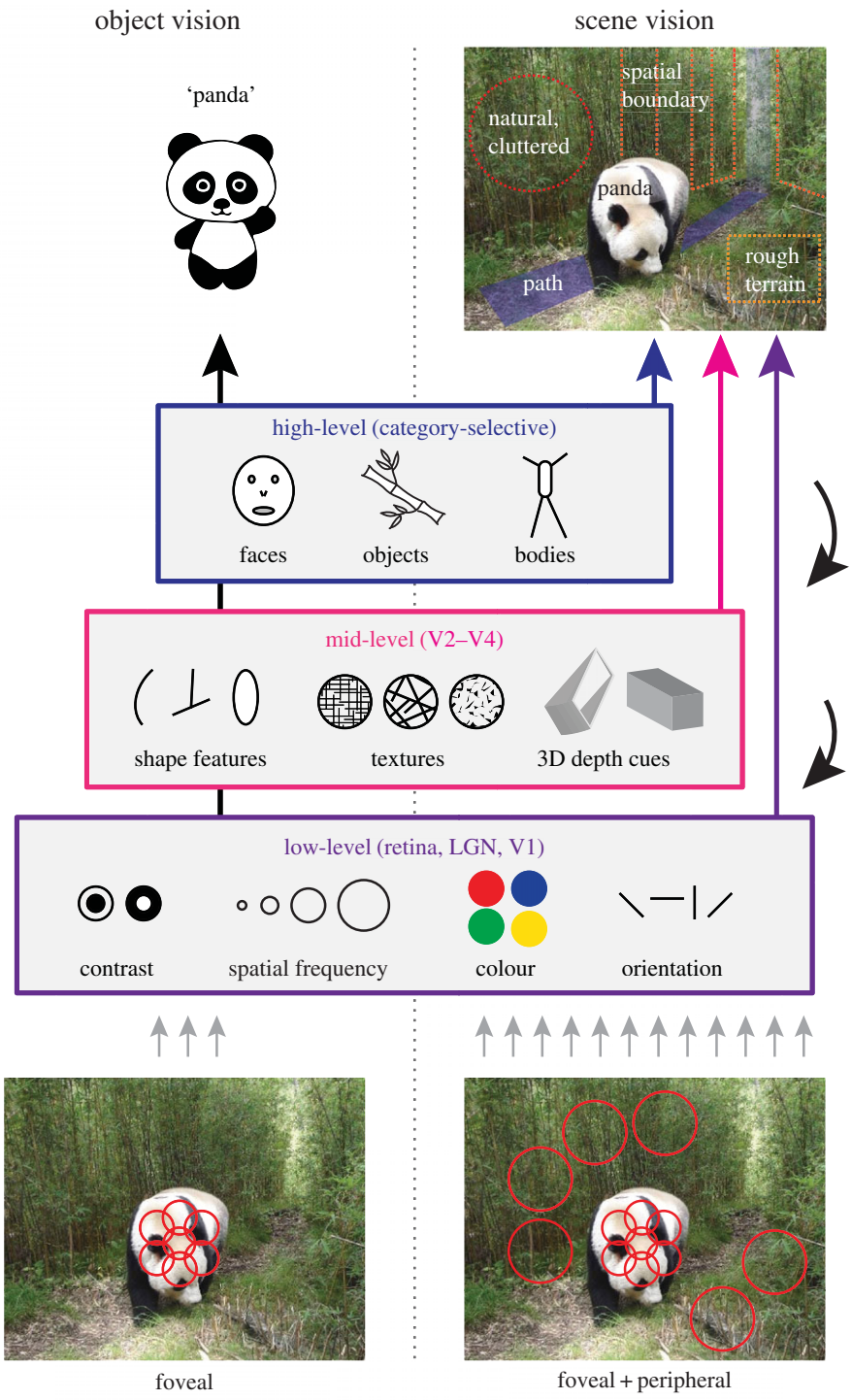

Figure 1. Hierarchical framework of visual perception. The standard view of visual perception posits that a visual percept is built from the retinal input by successive extraction of low-, mid- and high-level features. Object vision (left) typically involves foveal vision and the generation of an object label. By contrast, scene vision (right) involves both foveal and peripheral information, and the representation of multiple features including gist and navigability, which are not necessarily object-bound, and may be extracted from multiple levels of the hierarchy. Red circles (bottom) depict schematic receptive fields.

tuning to shape similarity [38], the notion that the incidental, low-level features of the input must be discarded to achieve a high-level invariant percept remains pervasive. Essentially, high-level vision is thought to *decode* the stimulus from the distributed patterns of information in lower level areas [6].

An important caveat of this view, however, is that for scenes it might not be possible to separate low- and high-level features in this way, as they are inherently correlated in our real-world environments. Beach scenes, for example, have a higher likelihood to be dominated by low spatial frequencies, due to the presence of a prominent horizontal boundary and large homogeneous sections (sky + beach), resulting in an open spatial layout. Indeed, it has recently been shown that low-, mid- and high-level features of scenes largely explain the same variance in brain responses in scene-selective regions [39].

We propose taking a step back and question whether behaviourally useful, or invariant, information in scenes can only be found at the highest level of representation. As highlighted above, scenes contain many more sources of information beyond the (foveated) objects, including texture and spatial layout, which typically extend into the periphery. It is likely that all this information, be it low-, mid- or high-level, can be

useful for some aspects of scene vision (figure 1, right). In §4, we shall highlight aspects of scene perception that may require different levels of representation.

4. The relevance of low-level information for scene perception

An extensive body of computational and behavioural research, which developed largely independently from the neuroimaging literature, has shown that many aspects of scene recognition can be carried out without explicit segmentation or labelling of objects in scenes. Instead, observers can rely on global scene features that can be derived from summary statistics of localized responses to local orientated spatial frequency filters [40,41] or texture [42]. Supporting this view, behavioural performance in rapid scene categorization is affected by manipulations of low-level features [43–47] and human participants are able to perceive global, non-object-based properties of scenes, even before they are able to categorize them [48,49]. The difficulty of some scene category distinctions compared to others, as well as apparent high-level effects in scene categorization such as the superordinate advantage, has recently been shown to be entirely predictable from differences in low-level visual features [50].

In these rapid scene perception tasks, peripheral representations appear to be at least as important as foveal information [51,52]. Thus, this literature suggests that some aspects of scene perception, such as rapid recognition, can rely on low-level features that are extracted from peripheral parts of the visual field (figure 1, bottom right). In line with this idea, it has been proposed that scene-selective regions might rely more on global representations, while object–object relations are represented in a complementary pathway comprising object-selective areas [53,54] which might nevertheless interact with information represented in scene-selective regions [55,56].

So far we have discussed scene categorization or identification, but scene perception entails more than just deriving a label of the environment. Indeed, we rarely need to rapidly categorize environments when we interact with them in our daily life. Instead, we tend to *act in* scenes. A primary example of action in scenes is navigation, getting from point A to point B. Indeed, human observers can determine scene navigability rapidly and consistently, possibly relying on a combination of global and local scene features [48,49]. It has been proposed that despite their shared category-selectivity, scene-selective regions may represent different aspects of scenes relevant to navigation, with the most posterior region OPA purportedly more involved in the representation of large-scale visual features [57], such as boundaries [58], whereas PPA and RSC are thought to undertake more complex computations relating to landmark coding [59], spatial memory [57] and navigation [60]. Importantly, however, even though navigation can be thought of as a high-level behavioural goal, it can be argued that the brain might still make use of low-level information for this task. For example, localized information in scenes is crucial to identify paths in the scene and to navigate around obstacles (figure 1, top right). Surface reflection and texture may be useful for identifying the accessibility of regions to be navigated, e.g. whether they are rough or slippery, while the spatial layout and three-dimensional surfaces may be important for identifying spatial boundaries that restrict movement.

These arguments suggest that low- or mid-level level features may in fact play a pervasive and potentially useful role in scene perception. In §5 we shall review neural evidence supporting this view in both the spatial and temporal domains as obtained with fMRI and magneto- and electroencephalography (MEG/EEG), respectively. We will show that despite earlier claims of position invariance, scene-selective areas contain plentiful information about position, evidenced by retinotopic biases towards specific parts of the visual field. Then we review recent EEG and MEG evidence of sensitivity to low- and mid-level information representation across multiple stages of processing in scene perception.

5. Visual field biases in scene-selective cortex

One of the most basic low-level properties is visual field position relative to fixation or *retinotopic location*. Historically, category-selective regions of visual cortex have been considered to either lack or contain weak retinotopic organization. Indeed, a general assumption is that visual representations become increasingly more position-invariant along the visual hierarchy. Recently, however, the presence and influence of retinotopic information, in the form of visual field biases, have been shown to extend beyond classical retinotopic cortex (V1–V4) and persist across both lateral and ventral cortex. Visual field biases can be characterized across three major retinotopic dimensions: (i) contralateral versus ipsilateral, (ii) foveal versus peripheral and (iii) upper versus lower. Below we discuss recent work demonstrating the influence of each dimension within scene-selective cortices (figure 2). While we restrict our discussion to these scene-selective regions, similar effects have been reported in other category-selective areas [61,64–66].

(a) Contralateral versus ipsilateral

During the initial stages of processing, visual inputs from the left and right visual fields are processed contralaterally within sub-cortical structures, such as the lateral geniculate nucleus (LGN), and this contralateral representation persists into V1. A similar contralateral preference is also observed in scene-selective cortices OPA, PPA and RSC, evidenced by larger fMRI responses to contralaterally over ipsilaterally presented scenes [57,61]. Beyond simple response magnitudes, more recent work has taken advantage of population receptive field mapping (pRF) techniques, which allow estimation of the effective RFs of individual voxels [67,68]. By sweeping a bar stimulus systematically across the visual field, the position (x-, y-coordinate) and extent (pRF size) of the visual field to which each voxel is maximally responsive can be estimated. Using bars filled with fragments of scenes, voxels in scene-selective regions have been shown to exhibit a strong bias for the contralateral field with more voxel pRFs centred within, and more representation of the contralateral over ipsilateral visual field [61,62].

(b) Foveal versus peripheral

Over a decade ago, Levy *et al.* [69] demonstrated that face- and scene-selective cortices are organized according to a foveal–peripheral gradient, whereby face-selective areas co-localize with more foveal representations, and scene-selective areas co-localize with representations of the periphery. Extending

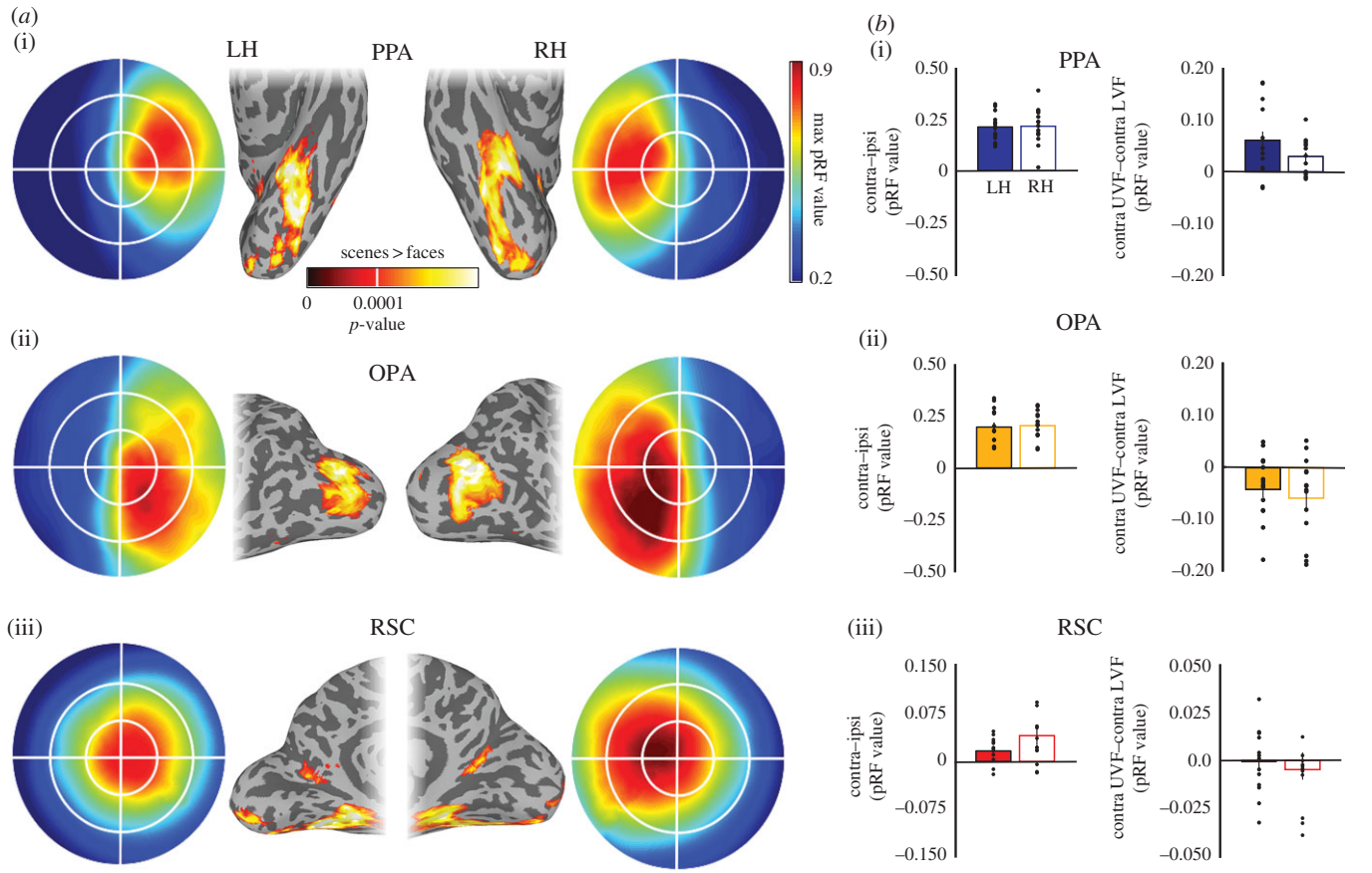

Figure 2. Retinotopic biases in scene-selective cortical regions. (*a*) Group average ($n = 16$) scene-selectivity (contrast of scenes > faces) and visual field coverage. (i) PPA on ventral temporal cortex, (ii) OPA on the lateral cortical surface and (iii) RSC in medial parietal cortex. All three scene-selective regions show a clear bias for the contralateral visual field. In addition, PPA and OPA show a bias for the upper and lower visual field, respectively. (*b*) Quantification of visual field biases in scene-selective regions. Left column: bars depict the contralateral biases (contralateral *minus* ipsilateral pRF value) exhibited by PPA (i), OPA (ii) and RSC (iii), respectively. Right column: bars depict the elevation biases (contralateral upper *minus* contralateral lower pRF value) exhibited by all regions. Dots indicate individual subjects. Adapted from [61–63].

this further, recent studies employing the pRF mapping paradigm demonstrate that the locations and sizes of pRFs within OPA, PPA and RSC are not only more eccentric, but also significantly larger than those within face-selective regions, confirming the peripheral bias within scene-selective cortices [61,62]. Functional connectivity with V1 suggests that RSC is more peripherally biased than either OPA or PPA [70], a pattern consentient with direct eccentricity measurements within these regions [61,62].

(c) Upper versus lower

Throughout retinotopic visual cortex, the representations of the contralateral visual field are segregated into upper (V1v–V3v) and lower (V1d–V3d) quadrants, and biases for the upper and lower fields extend into lateral and ventral extrastriate cortex, respectively [61,62,71]. On the lateral surface of visual cortex, OPA exhibits larger response magnitudes for stimuli presented in the contralateral lower than upper visual field, whereas on the ventral surface, PPA shows stronger response magnitudes for stimuli presented in the contralateral upper visual field [61]. pRFs within OPA and PPA also demonstrate differential representations of the contralateral lower and upper visual fields, respectively [62]. Relative to the largely quadrant representations within OPA and PPA, RSC contains a more complete hemifield representation and may therefore be able to mediate scene information from both the upper and lower visual fields [63].

(d) Retinotopic divisions of scene-selective regions

While these scene-selective regions exhibit retinotopic biases, there is no one-to-one relationship between scene selectivity and a given retinotopic map. For instance, OPA has been shown to overlap multiple maps, including V3d, V3A, V3B, LO1 and LO2 [62,72]. Similarly, PPA shows overlap with putative maps VO2, PHC1 and PHC2 [61,73]. The presence of retinotopic map divisions of scene-selective cortex, brings with it the possibility that each map undertakes multiple different computations of specific scene features and, moreover, that these computations are undertaken independently. For example, both fMRI [67] and transcranial magnetic stimulation [74] data support a dissociation between LO1 and LO2 (two adjacent maps in lateral occipital cortex with the same visual field coverage), with each computing orientation and shape independently [74].

(e) Functional significance of retinotopic biases

Taken together, the evidence discussed above (§5a–d) demonstrates the ubiquitous influence of retinotopy even within regions often thought to reflect high-level processing of visual scenes. Importantly, the presence of these retinotopic biases does not explain away the category selectivity observed. Rather, these retinotopic biases may provide important insight into their functional roles and need to be taken into account when theorizing about the computations performed by

scene- and other category-selective cortices. For example, the relatively large, peripheral pRFs in all three scene-selective regions make them sensitive to larger scale summary statistics of the input that may be particularly relevant for both rapid scene recognition and navigation. The upper field bias in PPA is consistent with a specific role in representing landmarks [75,76], large immovable objects that will typically occupy peripheral vision. Similarly, the lower field bias in OPA may reflect a role in representing the relative spatial layout of objects within a scene, which typically occurs in the lower field [77], for reaching and orientating the body in space. Finally, the retinotopy within RSC was recently proposed to reflect a functional subdomain (referred to as medial place area [63]) that is grounded in the visual input and that may be distinct from more anterior regions of medial parietal cortex that appear to be more involved in memory for scenes than immediate visual scene analysis [59,63,78].

This pervasive influence of retinotopy provides one example of how and why low-level information may be relevant for scene representation in the human brain. In §6, we shall discuss another example, by reviewing findings from time-resolved techniques such as EEG and MEG on object and scene perception.

6. Temporal dynamics of object and scene perception

Given the hierarchical view of visual processing, the different processing stages are commonly considered to be reflected in the temporal domain, with low-level vision associated with the initial responses of the system. Indeed, areas V1, V2 and V3 are often referred to as 'early' visual cortex. From detailed studies of event-related potentials (ERPs) extracted from the EEG signal, the general view has emerged that early ERP components are prominently involved in visual processing [79]. The earliest component, C1, can appear as soon as 40 ms after stimulus onset and is likely generated in V1, as it is highly sensitive to stimulus contrast and spatial frequency [80]. Subsequent components, such as the P1 (onset 60–90 ms; peak amplitude 100–130 ms) and N1 (onset 100–150 ms; peak 150–200 ms) may reflect activity in extrastriate and higher order visual regions, respectively, and become progressively less sensitive to stimulus parameters, with the N1 in particular being reflective of visual discrimination [81]. Representations of objects in natural scenes (e.g. 'animal') are thought to be present at the N1 level, arising around 150 ms after stimulus onset [82]. By contrast, activity beyond 200 ms, indexed by components such as the P2 and P3, are thought to reflect cognitive processing beyond visual encoding, such as task context, target probability or expectation [79].

(a) Dissociation of high- and low-level information in the temporal domain

Analogous to the fMRI literature, attempts have been made to dissociate high-level from low-level information in the early ERP components. Numerous studies have examined whether a face-selective component in the N1 time window, the N170, is driven by generic stimulus properties or a face-specific representation [83]. For objects, high-level categorical differences have been reported to occur as early as 80 ms, but since behaviour only correlated with ERP differences after 150 ms following stimulus onset [84,85], such early differences were suggested to be driven by low-level, task-irrelevant properties, whereas activity beyond 150 ms was thought to reflect task-relevant target representations. Distinct stages of processing were also proposed based on a study using objects embedded in phase noise [86,87]. In that study, comparison of single-trial ERP amplitudes with behavioural discrimination performance demonstrated that both the N170 and a late component at 300 ms correlated with behaviour [86]. However, this correlation was much stronger for the second component, which was also sensitive to the level of noise as reflected in an intermediate component at 200 ms, and the authors took this to reflect task difficulty [87].

However, other findings show that these three putative successive stages of (i) encoding, (ii) recognition and (iii) decision-making, are not always as easily separated. For example, stimulus luminance—arguably one of the most basic low-level properties—can affect ERP time courses well beyond the initial 100 ms of processing into 'high-level' time windows [88]. Moreover, whether or not an object-specific signal in the N1 window can be achieved depends on the degree of luminance information provided [89], indicating an interaction between low-level properties and object recognition. Furthermore, the aforementioned component at 200 ms was not observed in ERPs when task difficulty was manipulated by changing stimulus similarity rather than adding phase noise, suggesting that ERP activity around 200 ms may still be stimulus-tied, indexing visual noise instead of domain-general task difficulty [90]. Moreover, using the same object discrimination paradigm with more phase noise levels, ERPs were found to be sensitive to both noise and task in a relatively broad window between 140 and 300 ms [91], suggesting at least some overlap between processing stages.

(b) Temporal evidence for processing of low-level scene features

The majority of studies reviewed in §6a focused on detection of objects in scenes or background noise. Studies that investigated scene perception rather than object recognition cast further doubt on the discreteness of low- and high-level visual stages in the temporal dynamics of scene analysis. Intracranial recordings obtained during scene viewing indicate that PPA responds from very early on in visual processing, with initial responses distinguishing scenes from non-scenes at 80 ms, and buildings from non-buildings at 170 ms [92]. In a series of meticulous psychophysical experiments, Hansen *et al.* [93,94] demonstrated that C1, P1 and N1 amplitude were all modulated strongly and differentially by the spatial frequency content of scene patches. This ERP sensitivity was itself modulated by the structural sparseness of the scenes reflected in the overall number of edges [95], indicating that the EEG signal is highly sensitive to several different kinds of natural image statistics at multiple stages of visual processing.

In line with this suggestion, ERP modulations by two summary statistics derived from local contrast have been observed when participants view real-world scenes [96]. These two statistics, which the cortical visual system could plausibly read out from the population response of LGN neurons [97], summarize the mean and variance of the local contrast distribution and thereby index the level of stimulus energy (contrast energy) and degree of fragmentation or clutter (spatial coherence) of the scene [98]. Together, these parameters describe a feature

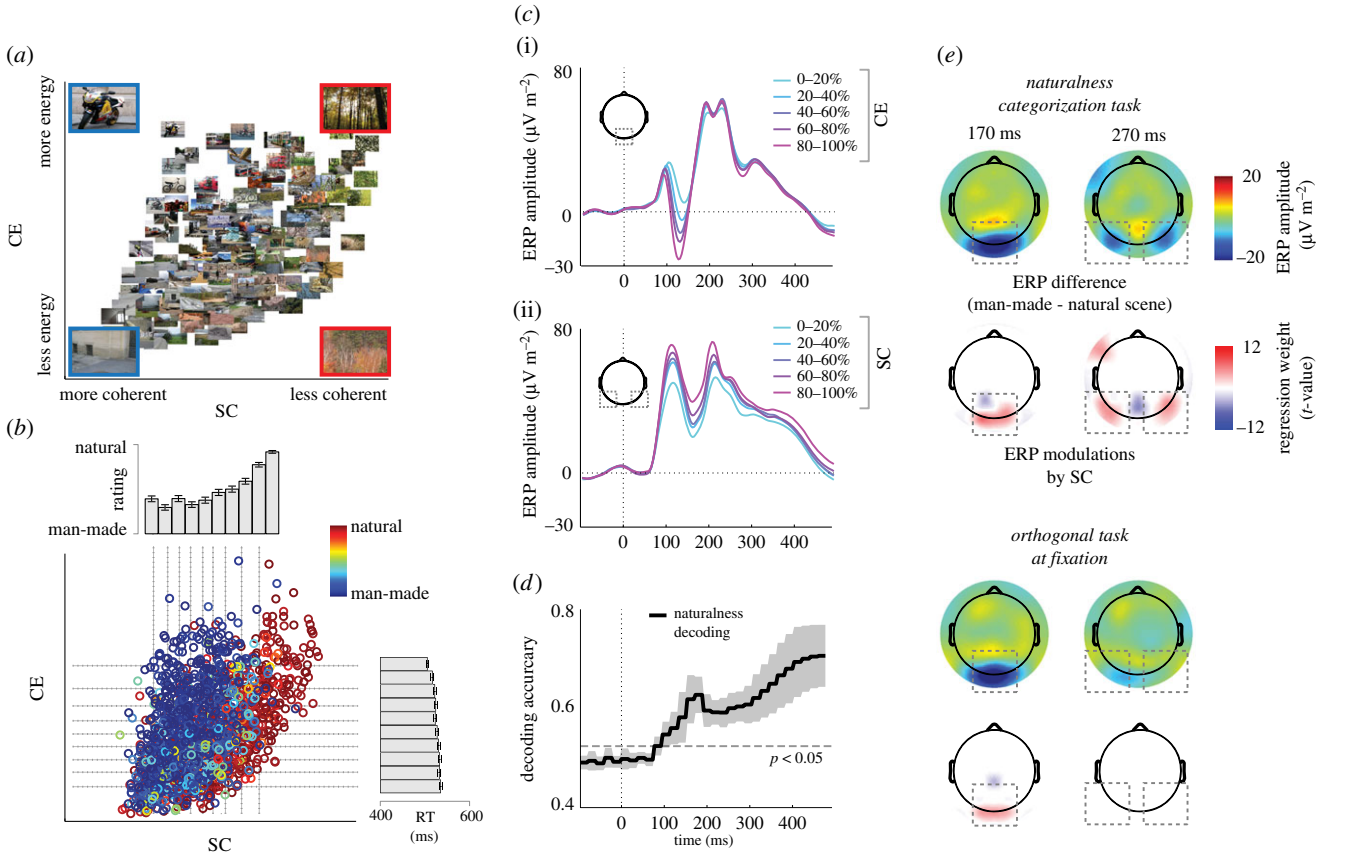

Figure 3. Image statistics modulate the temporal dynamics of scene perception. (*a*) Two summary statistics, contrast energy (CE) and spatial coherence (SC), describe a feature space in which complex, cluttered images are on the right and simple, organized images are on the left (larger scenes depict representative exemplars). (*b*) SC predicts naturalness ratings, while CE predicts reaction times. Colour index indicates average rating across 14 human observers. (*c*) At occipital electrodes (i), CE parametrically modulates ERPs in a transient time window, while SC (ii) modulates ERPs in a broad time window up to 300 ms over parietal–occipital sites. (*d*) Naturalness ratings can be decoded from the ERPs as early as 100 ms post-stimulus. (*e*) During naturalness categorization, categorical differences and SC modulations are present both early (170 ms) and late (270 ms) in time. When scenes are task irrelevant, only early effects are present. Adapted from [99,100].

space (figure 3*a*) that appears to be diagnostic of the degree of naturalness of a scene, due to the association of natural scenes with clutter (low spatial coherence) and of man-made scenes with organized structure (high spatial coherence). Indeed, these statistics have been shown to predict behavioural naturalness judgements for a large set of natural scenes ($n = 1600$) [99]: spatial coherence correlated strongly with the average naturalness rating, while contrast energy predicted reaction times (figure 3*b*). At the same time, contrast energy modulated ERPs in a discrete time window centred around the N1, while spatial coherence modulated ERPs up to 300 ms, thus extending well beyond early stages of visual processing (figure 3*c*). Importantly, the naturalness decision could already be decoded from single-trial ERPs as early as 100 ms (figure 3*d*) and the degree to which this was possible was also modulated by spatial coherence, with better decoding for scenes with more disparate spatial coherence values. Collectively, these results illustrate that diagnostic low-level summary statistics can affect the speed and accuracy of rapid scene categorization.

In a follow-up study [100], scene category distinctions beyond 200 ms disappeared when the scenes were task-irrelevant, consistent with the view that later time ERP windows are reflective of recognition and decision-making. Interestingly, however, when the scenes were task-relevant, categorical ERP differences were still associated with prolonged sensitivity to spatial coherence (figure 3*e*). Thus, rather than a separate stage of visual encoding followed by a recognition or decision phase, the presence of diagnostic stimulus information continued to modulate neural activity into later time windows, but only when it was task-relevant. Consistent with a strong interaction between low-level information and behavioural outcomes, another EEG study recently showed that behavioural ratings of naturalness and openness as well as image statistics both affected ERP amplitude at the P2 level, with these measures largely sharing the same variance in neural responses [101].

(c) Perceptual similarity overlaps with low-level feature similarity

Rather than looking at broad categorical differences in ERPs, recent studies have started to examine the time course of visual processing with novel information mapping approaches such as representational similarity analysis [102] (see also [103]). These studies confirm that representations of stimulus category occur very early in visual processing, but they also indicate that low-level features are often represented very close to or overlapping with time segments that are predictive of perceptual similarity of the visual input. Consistent with the ERP literature, decoding of isolated object categories from MEG signals was found to be possible from 80 ms onwards and peaked at 120 ms, while high-tier distinctions (e.g. animate/inanimate) peaked at later time points [104]. Comparison of representational dissimilarity matrices (RDMs) of isolated objects across MEG and fMRI showed that while early RDMs correlated most strongly with V1 and late RDMs with category-selective

cortex [105], there were also persistent correlations between early and late time points, suggesting that there was similarity in information content between early and late stages of visual processing. Another MEG study showed that neural dissimilarity of abstract patterns correlated most strongly with low-level similarity and with perceptual similarity at different time points (90 versus 145 ms, respectively), but these models also overlapped for a substantial period starting as early as 50 ms [106]. Similarly, EEG studies examining representational similarity of naturalistic image categories [107] and textures [108] found that time points with the highest correlations with image statistics (110 and 150 ms, respectively) were also predictive of perceptual similarity of the same stimuli. Importantly, when participants are engaged in a rapid scene categorization task, MEG correlations with global scene properties and behavioural categorization become near-simultaneous [109]. Similarly, decoding of scene identity and scene size overlapped with representations of both low- and high-level features represented in convolutional neural networks [110].

In sum, these studies highlight that while there are moments in scene processing at which various stimulus features are represented more strongly than others, these effects do not always occur in a neat progression from low to high level, and there is evidence for dynamic interactions between low- and high-level representations over time.

7. Low and high roads to visual scene analysis in the human brain

In §§3–6 we have highlighted how in scene vision, compared to object recognition, the contributions of low- and high-level information are more difficult to separate than suggested by the standard model of visual processing (figure 1). First, we emphasized that retinotopic information is present throughout scene-selective visual cortex. Second, we discussed how modulations of neural signals by low-level features tend to overlap and interact with those associated with recognition or categorization.

In the classic hierarchical view, low-level information tends to be dismissed as irrelevant, as it is thought to hinder the construction of invariant high-level object representations. In this opinion piece, we have therefore chosen to emphasize the utility of what is typically considered low-level information for scene vision. As alluded to above, however, we do not mean to suggest that these low-level features 'explain away' more complex neural representations that are most certainly required for many aspects of scene vision. However, in trying to reveal these representations, it is important to question the utility of classic hierarchical models for scene vision, and instead to consider which level of representation might be most fitting for a specific behavioural goal. For example, contour junctions might be particularly useful for representing scene category [29], while image statistics may be useful for rapid extraction of scene naturalness [48,99]. In line with a previous proposal made for object recognition [111], we suggest that considering the 'diagnosticity' of multiple scene properties and their representations in scene-selective regions under various tasks [112], or for specific scene functions [113], will lead to a more expansive framework for scene understanding.

We have also highlighted both (i) the highly dynamic nature of real-world scene perception, which until recently had not been studied extensively with time-resolved techniques [92,99–101,109,110] and (ii) the possibility that representations in scene-selective regions may arise via inputs from multiple pathways, including object-selective areas [54,75], possibly in addition to inputs from posterior areas carrying summary statistics [33]. Collectively, these considerations are consistent with an interactive parallel neural architecture with extensive reciprocal connections [71]. Such reciprocal connections may be particularly important for processes such as object grouping in real-world scenes [114] or to facilitate sensory processing in V1 by means of feedback from other cortical visual regions in both the dorsal and ventral stream [115], or non-visual sensory regions [116]. It is important to realize that traditional hierarchical models [117] as well as modern computational implementations of those models [118] do not typically allow for such dynamic interactions. The ubiquity of retinotopic biases in higher level visual cortex also raises the intriguing question of how information from multiple maps is combined, e.g. between left and right hemisphere representations [119]. Time-resolved techniques that are sensitive to the dynamics of scene perception may provide more insight into these questions; for example, by showing that contralateral information dominates early, but not late, visual evoked responses [120,121].

Moving forward, we propose that to understand how brain pathways and neural mechanisms subserve scene analysis, the focus should be on understanding the contribution of multiple scene properties to scene perception, rather than explicitly labelling them as either low or high level. The pertinent question then changes from 'how is low-level information discarded to achieve an invariant representation of a scene?' to 'what kind of neural representation is needed in order to achieve particular goals in scenes (e.g. recognition, navigation)?' [122]. This approach opens up exciting research questions for future work such as: in what way is the retinal input summarized or transformed to achieve these representations? How do neural representations in scene-selective areas change dynamically over time with different behavioural goals? Answering these questions will help us better understand the neural mechanisms underlying scene analysis in the human brain.

Authors' contributions. I.I.A.G., E.H.S. and C.I.B. conceived of and wrote the paper.

Competing interests. We have no competing interests.

Funding. I.I.A.G., E.H.S. and C.I.B. are supported by the Intramural Program of the National Institutes of Health (ZIA-MH-002909). I.I.A.G. is also supported by a Rubicon Fellowship from the Netherlands Organization for Scientific Research (NWO).

Acknowledgements. We thank members of the Laboratory of Brain and Cognition at the National Institutes of Mental Health for helpful discussions.

References

1. Wolfe JM, Võ ML-H, Evans KK, Greene MR. 2011 Visual search in scenes involves selective and nonselective pathways. *Trends Cogn. Sci.* **15**, 77–84. (doi:10.1016/j.tics.2010.12.001)

2. Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ. 2016 Explicit information for category-orthogonal object

properties increases along the ventral stream. *Nat. Neurosci*. **19**, 613–622. (doi:10.1038/nn.4247)

3. Oliva A, Torralba A. 2007 The role of context in object recognition. *Trends Cogn. Sci*. **11**, 520–527. (doi:10.1016/j.tics.2007.09.009)
4. Stansbury DE, Naselaris T, Gallant JL. 2013 Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* **79**, 1025–1034. (doi:10.1016/j.neuron.2013.06.034)
5. Edelman S. 2002 Constraining the neural representation of the visual world. *Trends Cogn. Sci*. **6**, 125–131. (doi:10.1016/S1364-6613(00)01854-4)
6. Yamins DLK, DiCarlo JJ. 2016 Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci*. **19**, 356–365. (doi:10.1038/nn.4244)
7. Veale R, Hafed ZM, Yoshida M. 2017 How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. *Phil. Trans. R. Soc. B* **372**, 20160113. (doi:10.1098/rstb.2016.0113)
8. 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**, 10 577–10 597. (doi:10.1523/JNEUROSCI.3726-05.2005)
9. Peirce JW. 2015 Understanding mid-level representations in visual processing. *J. Vis*. **15**, 5–9. (doi:10.1167/15.7.5)
10. 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. (doi:10.1038/nn.3402)
11. Welchman AE, Deubelius A, Conrad V, Bülthoff HH, Kourtzi Z. 2005 3D shape perception from combined depth cues in human visual cortex. *Nat. Neurosci*. **8**, 820–827. (doi:10.1038/nn1461)
12. 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. (doi:10.1016/j.neuron.2012.03.011)
13. Khaligh-Razavi SM, Kriegeskorte N. 2014 Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol*. **10**, e1003915. (doi:10.1371/journal.pcbi.1003915)
14. Güçlü U, van Gerven MAJ. 2015 Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci*. **35**, 10 005–10 014. (doi:10.1523/JNEUROSCI.5023-14.2015)
15. Ramakrishnan K, Scholte HS, Groen IIA, Smeulders AWM, Ghebreab S. 2014 Visual dictionaries as intermediate features in the human brain. *Front. Comput. Neurosci*. **8**, 168. (doi:10.3389/fncom.2014.00168)
16. DiCarlo JJ, Cox DD. 2007 Untangling invariant object recognition. *Trends Cogn. Sci*. **11**, 333–341. (doi:10.1016/j.tics.2007.06.010)
17. Kanwisher N, Dilks DD. 2013 The functional organization of the ventral visual pathway in humans. In *New visual neuroscience* (eds L Chalupa, S Werner), pp. 733–746. Cambridge, MA: MIT Press.
18. Epstein RA. 2014 Neural systems for visual scene recognition. In *Scene vision* (eds M Bar, K Kveraga), pp. 105–134. Cambridge, MA: MIT Press.
19. Walther DB, Caddigan E, Fei-Fei L, Beck DM. 2009 Natural scene categories revealed in distributed patterns of activity in the human brain. *J. Neurosci*. **29**, 10 573–10 581. (doi:10.1523/JNEUROSCI.0559-09.2009)
20. Walther DB, Chai B, Caddigan E, Beck DM, Fei-Fei L. 2011 Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proc. Natl Acad. Sci. USA* **108**, 9661–9666. (doi:10.1073/pnas.1015666108)
21. Aminoff EM, Kveraga K, Bar M. 2013 The role of the parahippocampal cortex in cognition. *Trends Cogn. Sci*. **17**, 379–390. (doi:10.1016/j.tics.2013.06.009)
22. Epstein RA, Parker WE, Feiler AM. 2007 Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *J. Neurosci*. **27**, 6141–6149. (doi:10.1523/JNEUROSCI.0799-07.2007)
23. Marchette SA, Vass LK, Ryan J, Epstein RA. 2015 Outside looking in: landmark generalization in the human navigational system. *J. Neurosci*. **35**, 14 896–14 908. (doi:10.1523/JNEUROSCI.2270-15.2015)
24. Park S, Chun MM. 2009 Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *Neuroimage* **47**, 1747–1756. (doi:10.1016/j.neuroimage.2009.04.058)
25. 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. (doi:10.1371/journal.pbio.1000608)
26. Nasr S, Echavarria CE, Tootell RBH. 2014 Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *J. Neurosci*. **34**, 6721–6735. (doi:10.1523/JNEUROSCI.4802-13.2014)
27. Kauffmann L, Ramanoël S, Guyader N, Chauvin A, Peyrin C. 2015 Spatial frequency processing in scene-selective cortical regions. *Neuroimage* **112**, 86–95. (doi:10.1016/j.neuroimage.2015.02.058)
28. Cant JS, Xu Y. 2012 Object ensemble processing in human anterior-medial ventral visual cortex. *J. Neurosci*. **32**, 7685–7700. (doi:10.1523/JNEUROSCI.3325-11.2012)
29. Choo H, Walther DB. 2016 Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *Neuroimage* **135**, 32–44. (doi:10.1016/j.neuroimage.2016.04.021)
30. Kravitz DJ, Peng CS, Baker CI. 2011 Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci*. **31**, 7322–7333. (doi:10.1523/JNEUROSCI.4588-10.2011)
31. Park S, Brady TF, Greene MR, Oliva A. 2011 Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *J. Neurosci*. **31**, 1333–1340. (doi:10.1523/JNEUROSCI.3885-10.2011)
32. Park S, Konkle T, Oliva A. 2014 Parametric coding of the size and clutter of natural scenes in the human brain. *Cereb. Cortex* **25**, 1–14. (doi:10.1093/cercor/bht418)
33. Watson DM, Hartley T, Andrews TJ. 2014 Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage* **99**, 402–410. (doi:10.1016/j.neuroimage.2014.05.045)
34. Bryan PB, Julian JB, Epstein RA. 2016 Rectilinear edge selectivity is insufficient to explain the category selectivity of the parahippocampal place area. *Front. Hum. Neurosci*. **10**, 1–12. (doi:10.3389/fnhum.2016.00137)
35. Schindler A, Bartels A. 2016 Visual high-level regions respond to high-level stimulus content in the absence of low-level confounds. *Neuroimage* **132**, 520–525. (doi:10.1016/j.neuroimage.2016.03.011)
36. Rousselet GA, Thorpe SJ, Fabre-Thorpe M. 2004 How parallel is visual processing in the ventral pathway? *Trends Cogn. Sci*. **8**, 363–370. (doi:10.1016/j.tics.2004.06.003)
37. Ullman S. 1996 *High-level vision*. Cambridge, MA: Bradford/MIT Press.
38. Edelman S. 1999 *Representation and recognition in vision*. Cambridge, MA: MIT Press.
39. 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. (doi:10.3389/fncom.2015.00135)
40. Torralba A, Oliva A. 2003 Statistics of natural image categories. *Netw. Comput. Neural. Syst*. **14**, 391–412. (doi:10.1088/0954-898X_14_3_302)
41. Oliva A, Torralba A. 2001 Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis*. **42**, 145–175. (doi:10.1023/A:1011139631724)
42. Renninger LW, Malik J. 2004 When is scene identification just texture recognition? *Vision Res*. **44**, 2301–2311. (doi:10.1016/j.visres.2004.04.006)
43. Kaping D, Tzvetanov T, Treue S. 2007 Adaptation to statistical properties of visual scenes biases rapid categorization. *Vis. Cogn*. **15**, 12–19. (doi:10.1080/13506280600856660)
44. Loschky LC, Larson AM. 2008 Localized information is necessary for scene categorization, including the natural/man-made distinction. *J. Vis*. **8**, 1–9. (doi:10.1167/8.1.4)
45. Joubert OR, Rousselet GA, Fabre-Thorpe M, Fize D. 2009 Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *J. Vis*. **9**, 2.1–216. (doi:10.1167/9.1.2)
46. Loschky LC, Hansen BC, Sethi A, Pydimarri TN. 2009 The role of higher order image statistics in masking scene gist recognition. *Atten. Percept. Psychophys*. **71**, 481–489. (doi:10.3758/APP.71.3.481)
47. Oliva A, Schyns PG. 2000 Diagnostic colors mediate scene recognition. *Cogn. Psychol*. **41**, 176–210. (doi:10.1006/cogp.1999.0728)
48. Greene MR, Oliva A. 2009 Recognition of natural scenes from global properties: seeing the forest

without representing the trees. *Cogn. Psychol*. **58**, 137–176. (doi:10.1016/j.cogpsych.2008.06.001)

49. Greene MR, Oliva A. 2009 The briefest of glances: the time course of natural scene understanding. *Psychol. Sci*. **20**, 464–472. (doi:10.1111/j.1467-9280.2009.02316.x)
50. Sofer I, Crouzet SM, Serre T. 2015 Explaining the timing of natural scene understanding with a computational model of perceptual categorization. *PLoS Comput. Biol*. **11**, e1004456. (doi:10.1371/journal.pcbi.1004456)
51. Larson AM, Loschky LC. 2009 The contributions of central versus peripheral vision to scene gist recognition. *J. Vis*. **9**, 1–16. (doi:10.1167/9.10.6)
52. Boucart M, Moroni C, Thibaut M, Szaffarczyk S, Greene M. 2013 Scene categorization at large visual eccentricities. *Vision Res*. **86**, 35–42. (doi:10.1016/j.visres.2013.04.006)
53. Kim JG, Biederman I. 2010 Where do objects become scenes? *Cereb. Cortex* **21**, 1–9. (doi:10.1093/cercor/bhq240)
54. MacEvoy SP, Epstein RA. 2011 Constructing scenes from objects in human occipitotemporal cortex. *Nat. Neurosci*. **14**, 1323–1329. (doi:10.1038/nn.2903)
55. Linsley D, MacEvoy SP. 2014 Encoding-stage crosstalk between object- and spatial property-based scene processing pathways. *Cereb. Cortex* **25**, 2267–2281. (doi:10.1093/cercor/bhu034)
56. Harel A, Kravitz DJ, Baker CI. 2012 Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cereb. Cortex* **23**, 947–957. (doi:10.1093/cercor/bhs091)
57. MacEvoy SP, Epstein RA. 2007 Position selectivity in scene- and object-responsive occipitotemporal regions. *J. Neurophysiol*. **98**, 2089–2098. (doi:10.1152/jn.00438.2007)
58. Julian JB, Ryan J, Hamilton RH, Epstein RA. 2016 The occipital place area is causally involved in representing environmental boundaries during navigation. *Curr. Biol*. **26**, 1104–1109. (doi:10.1016/j.cub.2016.02.066)
59. Marchette SA, Vass LK, Ryan J, Epstein RA. 2014 Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. *Nat. Neurosci*. **17**, 1598–1605. (doi:10.1038/nn.3834)
60. Epstein RA, Higgins JS. 2007 Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cereb. Cortex* **17**, 1680–1693. (doi:10.1093/cercor/bhl079)
61. Silson EH, Chan AW-Y, Reynolds RC, Kravitz DJ, Baker CI. 2015 A retinotopic basis for the division of high-level scene processing between lateral and ventral human occipitotemporal cortex. *J. Neurosci*. **35**, 11 921–11 935. (doi:10.1523/JNEUROSCI.0137-15.2015)
62. Silson EH, Groen IIA, Kravitz DJ, Baker CI. 2016 Evaluating the correspondence between face-, scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. *J. Vis*. **16**, 1–21. (doi:10.1167/16.6.14)
63. Silson EH, Steel AD, Baker CI. 2016 Scene selectivity and retinotopy in medial parietal cortex. *Front. Hum. Neurosci*. **10**, 1–17. (doi:10.3389/fnhum.2016.00412)
64. Kravitz DJ, Kriegeskorte N, Baker CI. 2010 High-level visual object representations are constrained by position. *Cereb. Cortex* **20**, 2916–2925. (doi:10.1093/cercor/bhq042)
65. Chan AW-Y, Kravitz DJ, Truong S, Arizpe J, Baker CI. 2010 Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nat. Neurosci*. **13**, 417–418. (doi:10.1038/nn.2502)
66. Sayres R, Grill-Spector K. 2008 Relating retinotopic and object-selective responses in human lateral occipital cortex. *J. Neurophysiol*. **100**, 249–267. (doi:10.1152/jn.01383.2007)
67. Larsson J, Heeger DJ. 2006 Two retinotopic visual areas in human lateral occipital cortex. *J. Neurosci*. **26**, 13 128–13 142. (doi:10.1523/JNEUROSCI.1657-06.2006)
68. Dumoulin SO, Wandell BA. 2008 Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647–660. (doi:10.1016/j.neuroimage.2007.09.034)
69. Levy I, Hasson U, Avidan G, Hendler T, Malach R. 2001 Center-periphery organization of human object areas. *Nat. Neurosci*. **4**, 533–539. (doi:10.1038/87490)
70. Baldassano C, Fei-Fei L, Beck DM. 2016 Pinpointing the peripheral bias in neural scene processing networks during natural viewing. *J. Vis. Revis*. **16**, 1–14. (doi:10.1167/16.2.9)
71. Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M. 2013 The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci*. **17**, 26–49. (doi:10.1016/j.tics.2012.10.011)
72. Nasr S, Liu N, Devaney KJ, Yue X, Rajimehr R, Ungerleider LG, Tootell RBH. 2011 Scene-selective cortical regions in human and nonhuman primates. *J. Neurosci*. **31**, 13 771–13 785. (doi:10.1523/JNEUROSCI.2792-11.2011)
73. Arcaro MJ, McMains SA, Singer BD, Kastner S. 2009 Retinotopic organization of human ventral visual cortex. *J. Neurosci*. **29**, 10 638–10 652. (doi:10.1523/JNEUROSCI.2807-09.2009)
74. 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. (doi:10.1038/nn.3327)
75. Troiani V, Stigliani A, Smith ME, Epstein RA. 2014 Multiple object properties drive scene-selective regions. *Cereb. Cortex* **24**, 883–897. (doi:10.1093/cercor/bhs364)
76. Auger SD, Mullally S, Maguire EA. 2012 Retrosplenial cortex codes for permanent landmarks. *PLoS ONE* **7**, e43620. (doi:10.1371/journal.pone.0043620)
77. Greene MR. 2013 Statistics of high-level scene context. *Front. Psychol*. **4**, 777. (doi:10.3389/fpsyg.2013.00777)
78. Baldassano C, Esteva A, Fei-Fei L, Beck DM. 2016 Two distinct scene processing networks connecting vision and memory. *eNeuro* **3**, 1–14. (doi:10.1523/ENEURO.0178-16.2016)
79. Luck SJ. 2005 *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
80. Di Russo F, Martínez A, Sereno MI, Pitzalis S, Hillyard SA. 2002 Cortical sources of the early components of the visual evoked potential. *Hum. Brain Mapp*. **15**, 95–111. (doi:10.1002/hbm.10010)
81. Vogel EK, Luck SJ. 2000 The visual N1 component as an index of a discrimination process. *Psychophysiology* **37**, 190–203. (doi:10.1111/1469-8986.3720190)
82. Thorpe S, Fize D, Marlot C. 1996 Speed of processing in the human visual system. *Nature* **381**, 520–522. (doi:10.1038/381520a0)
83. Rossion B, Jacques C. 2011 The N170: understanding the time course. In *Oxford handbook of event-related potential components* (eds ES Kappenman, SJ Luck), p. 115–142. Oxford, UK: Oxford University Press.
84. VanRullen R, Thorpe S. 2001 The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci*. **13**, 454–461. (doi:10.1162/08989290152001880)
85. Johnson JS, Olshausen BA. 2003 Timecourse of neural signatures of object recognition. *J. Vis*. **3**, 499–512. (doi:10.1167/3.7.4)
86. Philiastides MG, Sajda P. 2006 Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cereb. Cortex* **16**, 509–518. (doi:10.1093/cercor/bhi130)
87. Philiastides MG, Ratcliff R, Sajda P. 2006 Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *J. Neurosci*. **26**, 8965–8975. (doi:10.1523/JNEUROSCI.1655-06.2006)
88. Bieniek MM, Frei LS, Rousselet GA. 2013 Early ERPs to faces: aging, luminance, and individual differences. *Front. Psychol*. **4**, 268. (doi:10.3389/fpsyg.2013.00268)
89. Martinovic J, Mordal J, Wuerger SM. 2011 Event-related potentials reveal an early advantage for luminance contours in the processing of objects. *J. Vis*. **11**, 1–15. (doi:10.1167/11.7.1)
90. Bankó EM, Gál V, Kortvélyes J, Kovács G, Vidnyánsky Z. 2011 Dissociating the effect of noise on sensory processing and overall decision difficulty. *J. Neurosci*. **31**, 2663–2674. (doi:10.1523/JNEUROSCI.2725-10.2011)
91. Rousselet GA, Gaspar CM, Kacper P, Pernet CR. 2011 Modeling single-trial ERP reveals modulation of bottom-up face visual processing by top-down task constraints (in some subjects). *Front. Psychol*. **2**, 1–19. (doi:10.3389/fpsyg.2011.00107)
92. Bastin J, Vidal JR, Bouvier S, Perrone-Bertolotti M, Bénis D, Kahane P, David O, Lachaux J-P, Epstein RA. 2013 Temporal components in the parahippocampal place area revealed by human intracerebral recordings. *J. Neurosci*. **33**, 10 123–10 131. (doi:10.1523/JNEUROSCI.4646-12.2013)
93. Hansen BC, Jacques T, Johnson AP, Ellemberg D. 2011 From spatial frequency contrast to edge

preponderance: the differential modulation of early visual evoked potentials by natural scene stimuli. *Vis. Neurosci*. **28**, 221–237. (doi:10.1017/S095252381100006X)

94. Hansen BC, Johnson AP, Ellemberg D. 2012 Different spatial frequency bands selectively signal for natural image statistics in the early visual system. *J. Neurophysiol*. **108**, 2160–2172. (doi:10.1152/jn.00288.2012)
95. Hansen BC, Hess RF. 2007 Structural sparseness and spatial phase alignment in natural scenes. *J. Opt. Soc. Am. A Opt. Image Sci. Vis*. **24**, 1873–1885. (doi:10.1364/JOSAA.24.001873)
96. Scholte HS, Ghebreab S, Waldorp L, Smeulders AWM, Lamme VAF. 2009 Brain responses strongly correlate with Weibull image statistics when processing natural images. *J. Vis*. **9**, 1–15. (doi:10.1167/9.4.29)
97. Ghebreab S, Smeulders AWM, Scholte HS, Lamme VAF. 2009 A biologically plausible model for rapid natural image identification. In *Adv. Neural Inf. Process. Syst*. **22**, 629–637. See http://papers.nips.cc/paper/3785-a-biologically-plausible-model-for-rapid-natural-scene-identification.pdf.
98. Geusebroek J-M, Smeulders AWM. 2005 A six-stimulus theory for stochastic texture. *Intl J. Comput. Vis*. **62**, 7–16.
99. Groen IIA, Ghebreab S, Prins H, Lamme VAF, Scholte HS. 2013 From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *J. Neurosci*. **33**, 18 814–18 824. (doi:10.1523/JNEUROSCI.3128-13.2013)
100. Groen IIA, Ghebreab S, Lamme VAF, Scholte HS. 2016 The time course of natural scene perception with reduced attention. *J. Neurophysiol*. **115**, 931–946. (doi:10.1152/jn.00896.2015)
101. Harel A, Groen IIA, Kravitz DJ, Deouell LY, Baker CI. 2016 The time course of scene processing: a multi-faceted EEG investigation. *eNeuro* **3**, e0139–16.2016. (doi:10.1523/ENEURO.0139-16.2016)
102. Kriegeskorte N, Mur M, Bandettini P. 2008 Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci*. **2**, 4. (doi:10.3389/neuro.01.016.2008)
103. Cichy R, Teng S. 2017 Resolving the neural dynamics of visual and auditory scene processing in the human brain: a methodological approach. *Phil. Trans. R. Soc. B* **372**, 20160108. (doi:10.1098/rstb.2016.0108)
104. Carlson T, Tovar D, Alink A, Kriegeskorte N. 2013 Representational dynamics of object vision: the first 1000 ms. *J. Vis*. **13**, 1–19. (doi:10.1167/13.10.1)
105. Cichy RM, Pantazis D, Oliva A. 2014 Resolving human object recognition in space and time. *Nat. Neurosci*. **17**, 455–462. (doi:10.1038/nn.3635)
106. Wardle SG, Kriegeskorte N, Khaligh-Razavi S-M, Carlson TA. 2015 Perceptual similarity of visual patterns predicts the similarity of their dynamic neural activation patterns measured with MEG. *Neuroimage* **132**, 59–70. (doi:10.1016/j.neuroimage.2016.02.019)
107. Groen IIA, Ghebreab S, Lamme VAF, Scholte HS. 2012 Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLoS Comput. Biol*. **8**, e1002726. (doi:10.1371/journal.pcbi.1002726)
108. Groen IIA, Ghebreab S, Lamme VAF, Scholte HS. 2012 Low-level contrast statistics are diagnostic of invariance of natural textures. *Front. Comput. Neurosci*. **6**, 34. (doi:10.3389/fncom.2012.00034)
109. Ramkumar P, Hansen BC, Pannasch S, Loschky LC. 2016 Visual information representation and rapid scene categorization are simultaneous across cortex: an MEG study. *Neuroimage* **134**, 295–304. (doi:10.1016/j.neuroimage.2016.03.027)
110. Cichy RM, Khosla A, Pantazis D, Oliva A. 2016 Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *bioRxiv Preprint* **53**, 1–36. (doi:10.1016/j.neuroimage.2016.03.063)
111. Schyns PG. 1998 Diagnostic recognition: task constraints, object information, and their interactions. *Cognition* **67**, 147–179. (doi:10.1016/S0010-0277(98)00016-X)
112. Lowe MX, Gallivan JP, Ferber S, Cant JS. 2016 Feature diagnosticity and task context shape activity in human scene-selective cortex. *Neuroimage* **125**, 681–692. (doi:10.1016/j.neuroimage.2015.10.089)
113. Greene MR, Baldassano C, Esteva A, Beck DM, Fei-Fei L. 2016 Visual scenes are categorized by function. *J. Exp. Psychol. Gen*. **145**, 82–94. (doi:10.1037/xge0000129)
114. Korjoukov I, Jeurissen D, Kloosterman NA, Verhoeven JE, Scholte HS, Roelfsema PR. 2012 The time course of perceptual grouping in natural scenes. *Psychol. Sci*. **23**, 1482–1489. (doi:10.1177/0956797612443832)
115. Petro LS, Vizioli L, Muckli L. 2014 Contributions of cortical feedback to sensory processing in primary visual cortex. *Front. Psychol*. **5**, 1–8. (doi:10.3389/fpsyg.2014.01223)
116. Petro LS, Paton AT, Muckli L. 2017 Contextual modulation of primary visual cortex by auditory signals. *Phil. Trans. R. Soc. B* **372**, 20160104. (doi:10.1098/rstb.2016.0104)
117. Poggio T, Serre T. 2013 Models of visual cortex. *Scholarpedia* **8**, 3516. (doi:10.4249/scholarpedia.3516)
118. Kriegeskorte N. 2015 Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci*. **1**, 417–446. (doi:10.1146/annurev-vision-082114-035447)
119. Schiller PH. 1997 Past and present ideas about how the visual scene is analyzed by the brain. In *Extrastriate visual cortex in primates* (eds KS Rockland, JH Kaas, A Peters), pp. 59–90. Cerebral Cortex in Primates, vol. 12. New York, NY: Springer.
120. Smith ML, Gosselin F, Schyns PG. 2004 Receptive fields for flexible face categorizations. *Psychol. Sci*. **15**, 753–761. (doi:10.1111/j.0956-7976.2004.00752.x)
121. Rousselet GA, Ince RAA, van Rijsbergen NJ, Schyns PG. 2014 Eye coding mechanisms in early human face event-related potentials. *J. Vis*. **14**, 1–24. (doi:10.1167/14.13.7)
122. Malcolm GL, Groen IIA, Baker CI. 2016 Making sense of real-world scenes. *Trends Cogn. Sci*. **20**, 843–856. (doi:10.1016/j.tics.2016.09.003)