Making Sense of Real-World Scenes

George L Malcolm (1), Iris I A Groen (2), Chris I Baker (2)

[1] University of East Anglia, UK; [2] National Institute of Mental Health, MD

Correspondence: George L Malcolm, g.malcolm@uea.ac.uk

Abstract

To interact with the world, we have to make sense of the continuous sensory input conveying information about our environment. A recent surge of studies has investigated the processes enabling scene understanding, using increasingly complex stimuli and sophisticated analyses to highlight the visual features and brain regions involved. However, there are two major challenges to producing a comprehensive framework for scene understanding. First, scene perception is highly dynamic, subserving multiple behavioral goals. Second, a multitude of different visual properties co-occur across scenes and may be correlated or independent. We synthesize the recent literature and argue that for a complete view of scene understanding, it is necessary to account for both differing observer goals and the contribution of diverse scene properties.

Interacting with real-world scenes

Making a cup of tea is an easy task that requires minimal concentration, yet the composition of behaviors involved is deceptively complex: recognizing the room next door as a kitchen, navigating to it while manoeuvring around obstacles, locating and handling objects (e.g., teabag, kettle), and manipulating those objects until they are in a correct state (e.g., filling the kettle). In addition, it requires knowledge of relative locations and future destinations within the environment (e.g., take the kettle to the mug that is currently out of sight). Such interactions with the environment require the selective processing of task-relevant information, as well as the continual storage and retrieval of information from memory. Despite the seeming simplicity of the task, a multitude of scene properties across a range of different dimensions need to be processed, requiring the engagement of distributed brain regions.

The diversity of processes and goals engaged has led research on visual scene perception to progress in many directions over the past 50 years without necessarily much overlap. In the present article, we review this literature under the overarching context of scene understanding (Box 1). We argue that such a broad perspective is necessary to produce a comprehensive, theoretical framework to help elucidate the underlying cognitive and neural mechanisms. In particular, we focus on four of the most commonly studied behavioural goals – categorization, search, navigation and action – and consider the natural overlap among scene properties and the neural mechanisms involved.

Toward a Comprehensive Framework for Scene Understanding

There are two major challenges to producing a comprehensive theoretical framework that would outline the cognitive and neural mechanisms of scene understanding. The first is that while the physical characteristics of our surrounding environment are generally stable, our immediate goals are not. At any given moment, different visual aspects of an environment will be prioritized based on our current goal (Figure 1A; see also Box 2). However, the dynamic nature of scene understanding is often neglected as studies typically focus on individual tasks. The **diagnostic** (see Glossary) scene properties isolated in these studies may differ across task, making it hard to determine the extent to which such findings generalize to scene understanding in the real world.

The second challenge arises from the complexity of the input. Any scene can be described at multiple levels, from basic stimulus properties such as edges, **spatial frequency**, and color, to more complex, high-level characteristics such as object identity, spatial layout and action affordance (Figure 1B). Though individually each property might predict a given behavior, their inherent co-occurrence across scene categories [1] is often not considered: properties can either be correlated (e.g., beach scenes will generally contain large open spaces, far away objects, low spatial frequencies, etc.) or independent (e.g. both city scenes and forests may be characterized by high spatial frequency, vertical edges, etc.). It is therefore difficult, and even potentially misguided, to tease out specific contributions of any individual property to scene understanding without considering potential interactions.

These challenges are intimately linked: many different properties can be processed to complete a single goal, and conversely a single property can be used to facilitate many different goals. The diagnostic value of a visual property depends on a combination of the current goal and prior experience of the observer, as well as its availability within the scene and relationship to other properties [2,3]. In order to determine the cognitive and neural processes enabling scene understanding, it is critical to clarify how observer goals affect the weighting of different properties. In the next section, we will bring together research that addresses four of the main goals of scene understanding, and discuss how the potential utility of multiple properties varies across them.

Goal 1: What is this scene?

Scenes can be categorized at varying levels of hierarchical detail. Relative to the basic level, most commonly used in discourse (e.g., a city) [4], we can also make more generalizable, superordinate judgments (the scene is outdoors), or more detailed, subordinate judgments (a gothic city), or even identify the scene as a particular place (I am approaching North Bridge, Edinburgh). Understanding the meaning of an environment is important as it can facilitate the selection of subsequent behaviors such as actions (approach a crosswalk), searching (the time can be found by looking at the clock tower), deducing relative locations (the University is 10 minutes South), and so on.

The most heavily researched concept in scene categorization is **gist**, the initial representation of a scene that can be obtained in a brief glance. Studies typically make use of backwards-masking, or rapid serial visual presentation of, scene images and record performance as a function of presentation duration (note that this does not necessarily reflect the timing of relevant brain processes [5]). A large body of work suggests that as little as a ~13ms presentation duration allows for an initial scene percept [6], potentially including conceptual meaning [e.g., ,7] [but see ,8]. Though this duration by itself has limited ecological validity compared to the gradually changing, predictable world we experience, it serves as an important demonstration that an initial conceptual representation of a scene requires only a small subset of available visual properties to be processed. This epoch is too short to make a **saccade**, so only a single percept of a scene is afforded.

Global analysis of low-level features can facilitate the initial representation of a scene. For example, spectral features [9] and summary statistics of local contrast [10] can characterize the spatial properties of a scene: an efficient low-dimensionality resource bypassing computationally expensive processes of recognizing and integrating local components such as objects. Such global properties can accurately predict human scene recognition performance [11] and are processed early enough to facilitate the gist recognition epoch [12,13]. Other features include properties such as contour junctions [14], and color [15,16], which can facilitate initial scene understanding, although research suggests color facilitation occurs after initial processing of spatial structure [17]. Furthermore, some higher-level object information is available very rapidly [18,19] and objects and their co-occurrence may be diagnostic (e.g. the presence of a sink and an oven are highly predictive that the scene is a kitchen) [20].

While these properties and many more provide a quick summary of a scene's meaning, gist perception is limited for two reasons. First, interpreting the rapidly extracted gist depends on stored representations of typically occurring patterns [21], developed over experiences (e.g., a couch is commonly found in a living room). When scenes are less typical, such as when they contain inconsistent objects [e.g., a boulder in a living room, ,3], or contain atypical action relationships between individuals [22], the scene requires longer to process. Scene processing is therefore not entirely stimulus-driven, but is dependent on matching a percept to prior experiences. Secondly, scene recognition extends beyond gist, as we often interact with our environment at greater levels of detail. The more detailed the judgment, the longer the scene needs to be examined [23,24] as viewers supplement gist with goal-driven diagnostic details [2]. Thus, when the scene is an infrequently experienced situation or does not provide enough information relative to the viewer's goal, additional information must be acquired.

Goal 2: Where is X?

To gain information beyond gist – whether to support detailed recognition, search, or something else – eye movements are essential. This is necessitated by the retina's inhomogeneity: the central ~2°(fovea) of the visual field is processed in high resolution, but acuity drops off in the surrounding parafoveal (~4.5° into the periphery) and peripheral regions [25]. Appreciating the surrounding scene with full acuity would take roughly 16 minutes for the fovea to be directed to each location in our environment

[26]. To overcome this limitation, the visual system directs eye movements in an efficient manner by integrating low-resolution peripheral information with the current goal and knowledge of the environment [27,28], constrained by eye movement tendencies that produce stereotypical scan patterns [29]. Information falling within the foveal or parafoveal regions is optimized for detailed processing, while information in the periphery informs efficient saccadic distribution [27].

Several factors determine where the eyes are guided during scene viewing. Eye movements are strongly biased to direct **fixations** toward objects instead of backgrounds [30,31], with a preferred landing position within objects [32,33]. However, determining which objects in a scene are fixated, in what order, and for how long, relies heavily on the interplay between the viewer's goals and available visual information. When the viewer's goal is non-specific (e.g., memorizing a scene), image-based properties can predict where people fixate: edge density, visual clutter and homogenous segments predict fixation probability, while luminance and contrast play more minor roles [34]. The features used to select fixation sites are also determined by distance from the previous fixation, with shorter saccades (<8°) relying more on specific image features, particularly high-spatial frequencies, compared to longer saccades [35]. Fixation locations in free-viewing tasks can also be predicted based on eye movement tendencies, which act independently of the scene percept yet outperform some image-based models in predicting fixation locations [29].

When there is a more specific top-down goal (e.g., where is the kettle?) the visual system can utilize various scene properties depending on how diagnostic they are for that particular goal. Viewers might rely on matching low-level scene features, such as color [36] or shape [37], with target properties. High-level factors can also bias gaze by using the semantic relationship between gist and object meaning [38,39], as well as the relationships between objects [31,40–42], the spatial dependency between objects [43] or an object's relationship to the spatial layout [44]. These various guidance factors can be combined to direct attention to the most likely target location [45,46, see Figure 2B].

Gaze allocation is thus the result of a bi-directional relationship between scene properties, ranging from low-level features to high-level semantics, and the viewer's goal. However, as mentioned above, there is not necessarily a one-to-one mapping of visual properties to a goal. The diagnosticity of a particular property is also dependent on availability. Looking for a kettle in a kitchen would rely primarily on semantic knowledge (e.g., kettles are typically found on the stove), yet if the scene does not provide clear semantic cues, gaze guidance can be driven by episodic memory instead [e.g., where did I last see the kettle?, ,47]. Similarly, if the target has no specific spatial dependency, other factors dominate [48]: searching for a fly would rely on low-level feature matching; looking for a banana in a well-lit room would primarily rely on color, while in the same dimly lit room it would rely more on shape.

Additionally, as the representation of a scene changes over time, so too does the information gathered by eye movements. Once recognized, the scene category rarely changes without significant locomotion from the viewer; thus gist becomes a less relevant guiding factor over time [49,50] and other properties become more pertinent – e.g., fixated objects stored in short- and long-term memory [51–53] – enabling the development of more detailed scene representations [2], improved action efficiency [54,55], and so forth.

Goal 3: How do I get from A to B?

Accomplishing a particular goal may require moving from one location to another [e.g., ,56,57]. Scene understanding therefore not only entails what a scene is or where specific scene elements are, but also involves representing the information that enables navigation through them.

We discuss two forms of navigation here. The first is how we move from one point to another in **vista space** [58], which we refer to here as navigability, and is generally concerned with paths and obstacles [11]: for instance, crossing from one side of North Bridge to the other while avoiding buses. This relies on a dynamic representation of our position within a stable spatial layout, and prioritizes updating the location and movement of discrete objects more than their meaning. Clear paths and obstacles can potentially be processed from the same global properties that facilitate the initial labelling of a scene through spatial characteristics [11,12]. As observers move through the world, they can regularly update their spatial location and continue to sample information concerning the state and position of obstacles. For instance, some of the most commonly fixated obstacles when moving through an environment are people [59], who are generally fixated at a distance to determine their heading and avoid collisions [55,59]. Similarly, eye movements are made to gauge the distance of approaching objects [e.g., a car, ,60]. Viewers also sample ground information in a goal-driven manner, directing fixations ahead to check for changing terrain, as well as to closer regions that will be stepped on, and surface transitions to avoid [e.g., a curb, ,61].

The second form of position-based scene understanding is knowing where you are relative to an unseen location in **environmental space** [58], which we refer to here as navigation. For example, identifying your location as on North Bridge, Edinburgh, and knowing your relative direction and distance from the University. Navigation relies more on the meaning of a scene and less on the dynamic position of its elements. An observer's position during navigation can be discerned from a process relying on path-integration – the integration of an observer's translations and rotations from a start point in order to estimate the current position [62] – working in concert with two informative scene properties. The first is landmarks, which are persistent visual stimuli that have both distinct perceptual features and occur at decision points along a route [63–65]. Three types of information are needed to utilize a landmark: it needs to be recognized (e.g., clock tower), its position relative to other points needs to be retrieved from long-term spatial knowledge (e.g., the clock tower is six blocks from the University), and the heading of the viewer relative to the landmark needs to be determined (e.g., the clock tower is in front of me). At this point a route can be planned [e.g., carry on straight ahead, ,66]. The second informative property is scene boundaries, which are extended surfaces that separate one environment from another. Unlike the single point of a landmark, boundaries are made up of multiple points from which directions can be discerned [67]. Though typically thought of as large structures (e.g., walls), a boundary need not necessarily restrict movement and even a subtle geometric property – such as a small ridge – is enough to act as an informative spatial cue [68]. The relative contributions of landmarks and boundaries to navigation can be directly compared using a virtual arena paradigm, in which target location is tethered to a landmark or boundary which participants learn over time [e.g., ,69,70, see Figure 2C].

Locomotion through an environment is experienced as continuous episodes of immediate spaces (walk across the bridge, down the street, around the corner, etc.). Thus, more than recognition or searching for items, the behavioral goal of movement is made up of a sequence of smaller goals. At each new stage, the internal scene representation must change with the observer's needs, whether it is a decision based more on physical locations in vista space (e.g., head to the gap between obstacles), or recognition in environmental space (e.g., recognize the landmark and interpret location), or the two simultaneously. As such, navigation is an ecologically relevant behavioral goal that emphasizes the dynamic nature of scene understanding.

Goal 4: What can I do here?

Navigation can be seen as one example of a behavioral goal that highlights the strong link between vision and action in scene understanding. This functional perspective on vision was long ago recognized in theoretical frameworks emphasizing scenes as environments that provide possibilities for action, i.e. 'affordances' [4,71]. Action affordance has shown to be an important factor in how we understand objects [72], and object affordance influences how we search for items in visual scenes [73]. While several studies have considered how visual scenes serve as a context facilitating recognition of both objects [18,74–76] and actions [77], the idea that affordances determine how we understand the scene itself is relatively unexplored. However, a recent study has shown that descriptions of actions which might occur in a scene predict their categorization better than objects or visual features. In other words, a kitchen scene is understood as a kitchen "because it is a space that affords cooking" [78, p93] (Figure 2D).

An unresolved question is how such action affordances might be computed from a scene. While action descriptions explained most of the variance in the behavioral categorization [78], some of this variance was shared with objects and visual features. Presumably, action affordances can be deduced from a scene by a non-linear combination of multiple scene properties, which may include those depicted in Figure 1B, as well as representations stored in memory, potentially including complex sociocultural aspects of scenes [79]. Another novel line of research suggests that our ability to make physical inferences about our visual environment, e.g., predict possible movements by objects in scenes [80], involves cognitive mechanisms also used in action planning [81]. As such, the concept of action affordances highlights the importance of considering multiple scene properties simultaneously, allowing for potential combination of these properties with action goals. Generally speaking, action affordances could serve as a useful broader concept highlighting the interactive components of scene understanding, and encompassing more complex functional scene properties such as navigability, which 'afford' movement within a space.

Mapping Properties to Goals

Based on the discussion of these four main goals of scene understanding, it should be clear that the goals themselves are not mutually exclusive. For example, recognition facilitates search and navigation processes; navigation sometimes requires searching for specific information (e.g. objects, boundaries); scene affordance must consider navigability within a space, and so forth. Similarly, informative properties overlap various goals: spatial layout facilitates the early stages of recognition as well as navigability, edge information can help recognition and obstacle detection, etc. This means that there is no simple way to map between goals and properties. In this context, elucidating the neural mechanisms of scene processing can provide additional insight by demonstrating which properties are represented in different parts of the brain.

The Neural Mechanisms of Scene Understanding

In general, visual scene processing in humans has been characterized by a trio of **scene-selective** regions: occipital place area (OPA), parahippocampal place area (PPA) and retrosplenial complex (RSC), on the lateral occipital, ventral temporal and medial parietal cortical surfaces, respectively [82]. Studies in non-human primates have also reported scene-selective regions [83–85] as well as regions responsive to spatial landscapes [86]. Much of the research on humans has focused on establishing what specific properties each scene-selective region is sensitive to. For example, responses in PPA have been reported to reflect a wide range of properties, including, i) low-level properties, such as spatial frequency [87–90], orientation [91], texture [92], rectilinearity [93] [, but see ,94], and contour junctions [95]; ii) object properties, such as identity [96], size [97], space diagnosticity [98], co-occurrence [99], and object ensembles [92]; iii) 3D layout, such as size of a space [100], spatial expanse [i.e., open or closed, ,96,101,102], distance [102], and boundaries [103]; and iv) high-level properties, such as semantic category [104,105], contextual associations [106,107], and knowledge of scene correspondences [108]. Sensitivity to some of these properties is shared by both OPA and RSC, but in contrast to PPA, they show greater sensitivity to egocentric distance [109] and sense [i.e., left versus right mirror views, ,110]. Further, OPA has been associated with the local elements of scenes [111] and transcranial magnetic stimulation over OPA selectively impairs scene discrimination and categorization [112], as well as disrupting navigation relative to boundaries [70, Figure 2C]. RSC may have a particular sensitivity to landmarks [64,113,114], and, in addition to visual responsiveness, RSC (and to some extent anterior PPA) has been associated with spatial memory and imagery, particularly in the context of navigation [108,115–117]. In fact, there may be separate perceptual and memory scene networks [118,119] and systematic organization [e.g., ,95,99] within medial parietal cortex, with posterior regions showing more visual selectivity and anterior regions more related to memory [120]. Finally, there are basic response characteristics of the three scene-selective regions that may help inform their functional role and the critical scene properties they represent. In particular, these regions show a peripheral visual field bias with relatively large population receptive fields that make them well placed to capture summary information across large portions of the field of view [120–122]. Further, OPA and PPA show retinotopic biases to the lower and upper visual field, that may make them particularly well suited for capturing information relevant for navigability or processing landmarks, respectively [121].

While these studies have provided much insight into the neural processing of scenes, it is clear that each region is sensitive to multiple scene properties and determining specific critical properties for each region is difficult. In part, this may reflect the two challenges we have highlighted. First, most studies have focused on individual tasks only, and their results may be specific to that task. Indeed, a recent study [123] carefully and systematically manipulated multiple features of computer-generated scenes and compared brain representation of these features across multiple tasks revealing dynamic coding of scene properties (Figure 3A). For example, differentiation of spatial boundary in PPA was affected by a task instruction requiring participants to attend to either texture or spatial layout. Task effects have also been reported to change sensitivity to diagnostic properties in EEG signals for real-world scenes [124].

Second, the different properties identified in different studies may in fact be correlated, and reflect sensitivity to the same underlying dimension. One approach to address this problem is to use large numbers of naturalistic images and model multiple dimensions simultaneously. For example, a comprehensive analysis [125; Figure 3B] revealed that the dimensions of spatial frequency, subjective distance and object category all explained variance in scene-selective regions. However, most of the variance explained was shared across the models suggesting that, for example, the apparent sensitivity to scene category could just as easily be interpreted as reflecting differences in spatial frequency. The difficulty of this approach, however, is that there are many possible models that could be tested. Further, these models might differ in their biological plausibility (see Box 3) and we should also be careful not to assume that any sensitivity to low level properties explains away the sensitivity to high level properties - the interaction between low and high level features may actually be informative about how the brains transforms the retinal input into a task-relevant representation [126].

Despite the many goals of scene understanding, many of the neuroimaging studies we have discussed emphasize recognition or use simple tasks not necessarily related to real-world goals such as passive fixation [e.g., ,95,99], orthogonal tasks [e.g., changes in the fixation cross, ,102], simple discrimination [e.g., 1-back repetition, ,100,111], or familiarity judgments [108]. The major exceptions are the increasing number of studies that focus on navigation [e.g., ,108,113] and such an approach is necessary to help elucidate the neural mechanisms of scene understanding. The distinction we have highlighted between recognition and interaction is reminiscent of the division of cortical visual processing into dorsal and ventral pathways which have been characterized as reflecting separate processing of dynamic spatiotemporal relationships and stable visual qualities, respectively [127]. However, the role of the dorsal pathway in scene understanding has been relatively little explored and, navigation aside, there has been little focus on more immediate interactions with the environment such as guidance of eye movements in scenes, which might depend heavily on the dorsal pathway.

Moving forward, we suggest that to build on the current literature and further elucidate the neural mechanisms underlying scene understanding research should continue to emphasize: the use of naturalistic images (reflecting the diverse properties and their correlations in real-world scenes), meaningful tasks (that reflect real world goals), generalizability across tasks, simultaneous consideration of multiple scene properties (avoiding a priori assumptions about specific properties) and an understanding of how those scene properties relate to each other. While integrating all these elements together is certainly ambitious, we believe it paves the way forward for elucidating the neural representation of scenes.

Concluding Remarks and Future Directions

Scene understanding entails representing information about the properties and arrangement of the world to facilitate ongoing needs of the viewer. By focusing on four major goals of scene understanding – recognizing the environment, searching for information within the environment, moving through the environment, and determining what actions can be performed – we have demonstrated how different goals use similar properties and, conversely, how many properties can be used for different goals. Further, the brain regions implicated in scene processing appear to represent multiple different properties and might be capable of supporting multiple goals. While studying single tasks and the contribution of isolated properties has elucidated important sub-components of scene understanding, we advocate a more comprehensive scene understanding framework. Allowing for dynamic representation of multiple scene properties across multiple tasks opens up many exciting new research questions (see Outstanding Questions) for which experiments will be required that combine strong, hypothesis-driven manipulations of top-down goals with sophisticated, data-driven measurements of scene properties. While it is challenging to adopt real-world goals for experimentation in a laboratory setting, we believe that as a whole, the research community has developed the tools enabling the power of multiple approaches to be combined in order to help understand how we make sense of real-world scenes.

Acknowledgements

GLM and CIB are supported by the intramural research program of NIMH (ZIAMH002909). IIAG is supported by a Rubicon Fellowship from the Dutch Organization of Scientific Research (NWO). We thank Wilma Bainbridge, Michelle Greene, Antje Nuthmann and Edward Silson for commenting on earlier versions of this manuscript.

Text Boxes

Box 1: Scene Understanding

There is no single way to define a scene, and previous work has often focused on two key considerations [e.g. ,82,6,128]. The first consideration is primarily stimulus-based, focusing on the perceived properties of an image. Very generally, any collection of objects or shapes (e.g. a texture, or an array of search items) can be considered a scene. However, real-world scenes differ from such stimuli because they typically contain a large variety of items that are arranged in a meaningful manner, containing a *spatial layout* that organizes the scene into foreground objects and background elements (e.g., walls, ground plane). As such, a scene is often defined as consisting of a specific viewpoint of a real-world environment (e.g. a beach photograph). Such stimulus-based definitions are most commonly adopted in studies on scene recognition and categorization. To the extent that these tasks require processing an image as a single nameable entity, one could argue that processing of a scene stimulus is not unlike processing an object. However, an important second distinction that has been made is that observers act *upon* objects but act *within* scenes [82,6]. In this light, the second way we can consider scenes is as a 3D environment the observer is embedded in and interacts with. Under this interaction-based view, those aspects of a scene allowing the observer to carry out specific behavioural goals, e.g. locomotion or motor interaction, become critical. Daily tasks are generally comprised of several smaller goals occurring in quick succession, or potentially overlapping in time, while the world gradually unfolds around us. The interaction-based view on scenes thus incorporates not only what is visible, but also the memorized (or predicted) arrangement of elements that are involved in a larger, continuous environment.

These considerations are not mutually exclusive: many of the visual properties identified under the scene-as-stimulus view are relevant for interacting with scenes, and in turn individual behavioural goals that involve interaction with scenes may require the representation of different scene properties. The framework we propose explicitly incorporates both of these considerations under the umbrella term of “scene understanding”, which emphasizes both how individual goals affect one another and the respective weighting of visual properties.

Box 2: Goal driven models of visual processing

The importance of ethological goals has been emphasized in computational work employing hierarchical convolutional neural networks (CNNs) [129]. These models, which in essence are a generalized form of principles first proposed by Hubel and Wiesel, have garnered increasing attention because they approach human levels of performance on tasks such as object recognition for real-world images [129,130] and show a degree of correspondence between individual layers of the networks and different levels of neural visual object processing [131–133]. CNNs that perform better on object-recognition tasks are found to be better predictors of neuronal spiking data in monkey inferotemporal cortex [134], suggesting the importance of a goal driven approach for creating a model of a given sensory system. Although CNNs have primarily been trained on object recognition tasks, some have focused on scene categorization [135], also revealing a correspondence to human MEG data [136]. Importantly, performance on scene-related recognition tasks was found to be better for a network trained on scene-centric compared to one trained on object-centric data [137]. Further, the estimated receptive fields of a scene network appeared to exhibit features consistent with properties such as surface layout, and comparison with an object network showed stronger representational similarity with scene size [136].

While we have so far emphasized the importance of the task goal in training CNNs, this is not to say that there is no generalization, that the features developed for one task are not also applicable to other tasks. For example, a CNN optimized for object classification showed some generalization to other recognition tasks such as attribute detection (e.g. presence of specific part or material) and scene recognition [138]. Similarly, a CNN trained to perform scene categorization appeared to develop object detectors [137]. An issue that is so far unexplored is whether the degree to which CNNs trained for different tasks develop similar feature representations depends on the inherent correlations in the visual input (Figure 1B), i.e. whether generalization across tasks may occur because of shared information between features in the real-world. One possibility suggested by findings from CNNs is that different processing pathways in the brain could have developed as a direct result of the differential task constraints imposed by the required classification of the respective visual input [127,134,136].

Box 3: Biological plausibility of scene properties

An important factor to consider when comparing multiple models of scene properties is whether and how these properties can be computed plausibly by the visual system. One example highlighting this issue is the role of spatial frequency in visual scene perception. In the influential *Spatial Envelope* model [9], spatial frequency regularities of scenes are quantified based on principal components of their power spectra. However, it has long been recognized that the brain cannot perform a computation akin to a whole-scene Fourier transformation necessary to derive the principal components [139,140]. Visual processing in the brain occurs via conversion of light intensities at the retina to local contrast responses in small receptive fields in the lateral geniculate nucleus and primary visual cortex: neurons in these areas thus only “see” a small part of the visual scene. While higher-order visual regions have larger receptive fields (which are thought to integrate the information being fed from preceding regions), recent reports of extensive retinotopic biases in high-level, category-selective regions [e.g., ,122] question the idea these areas are capable of representing whole-scene information. Importantly, some global scene properties reflected in the spatial frequency decomposition, such as naturalness, are also reflected in the local contrast distribution, which thus potentially forms a more biologically plausible image statistic that can be computed across a selected portion of the visual field [10,141]. Direct comparison of statistics derived from the Fourier transform versus the local contrast distribution has shown that the latter better predicts visual evoked responses in humans to natural scenes [13,142].

This issue is also relevant considering that the Fourier transformation is commonly used in cognitive neuroscience experiments as an image manipulation tool, e.g. to demonstrate spatial frequency biases in higher-order visual regions with filtered stimuli that contain exclusively high or low spatial frequencies [e.g., ,143]. However, brain responses obtained using manipulated images do not necessarily generalize to intact natural scenes [144,145], and it is important not to attribute the operation performed to achieve an image manipulation to a neural computation without considering its biological plausibility. Instead, it might be more useful to think of image manipulation as emphasizing specific aspects that are more or less diagnostic for a given task [146,147].

Glossary

Convolutional Neural Network: A computer vision model with a multi-layer architecture performing hierarchical computations that can be trained to perform classification of visual images.

Diagnosticity: The relative usefulness of a specific subset of perceptual information in facilitating an observer's goal. For example, an oven is highly diagnostic in helping to categorize a scene as a kitchen, while an apple is less so.

Environmental Space: A physical space that is too large to be appreciated without locomotion, requiring the integration of information over time. Examples include buildings, neighborhoods, golf courses, etc.

Fixation: A period of relative eye movement stability, usually on an object so that its image falls on the fovea, allowing for the perception of local details.

Gist: The perceptual and semantic information comprehended in a single glance of a scene (generally ranging from 13-250ms in presentation duration). The content of gist usually includes a conceptual understanding of a scene (e.g., birthday party), the spatial layout of the environment, and recognizing a few objects.

Saccade: A ballistic eye movement that quickly moves the fovea from one fixated location to another. The eyes rotate up to speeds of 500° per second, and no visual information is extracted during this time.

Scene Selectivity: By comparing the response elicited by visually presented scenes with that for objects or faces, researchers have identified three scene-selective cortical regions, termed the parahippocampal place area, the occipital place area, and the retrosplenial complex.

Spatial Frequency: A measure of how often sinusoidal components of image structure repeat across units of distance, which captures the level of detail present in a scene per degree of visual angle. A scene with small details and sharp edges contains more high spatial frequency information than one composed of large coarse stimuli.

Vista Space: A physical space that can be viewed in its entirety from a single location without locomotion. Examples include a classroom, town square, field, etc.

Figures

Figure 1. Observer goals and scene properties. A) Examples of possible observer goals in scene understanding. In this article, we focus on four general task domains that involve scene understanding: 1) recognition, i.e. determining whether a visual scene belongs to a certain category (e.g., beach scene), or depicts a particular place (the park, my living room); 2) visual search, which involves locating specific objects or other scene elements; 3) navigation, which involves determining both the navigability of the immediate space and one's position relative to an unseen location; and 4) action goals, which may involve navigation but also encompass a broader set of activities such as cooking or playing baseball. B) Examples of scene properties that may be relevant for constructing mental representations necessary to achieve various observer goals. Properties that can be computed from scene images with relatively simple computational models, such as edges, spatial frequency and color, are considered 'low-level features'. More complex properties are the scene's constituent objects and 3D properties reflective of the layout of the scene, or the distance of the observer to salient elements in scenes. Finally, semantic category and action affordances can be seen as 'high-level' features of scenes that are not easily computed from scenes but may inform multiple observer goals. Note that scenes may differ from one another at multiple levels; for example, the beach scene can be distinguished from the park and living room based on virtually all dimensions, whereas the park and living room image share some but not all properties. Due to the inherent correlations between scene features, assessing their individual contributions to scene representations is challenging [125].

Figure 2. Mapping properties to goals. A) Recognizing a scene. Scenes that were easily categorized as man-made or natural resided at opposite ends of a low-level feature space described by two summary statistics of local contrast (see Box 3): contrast energy and spatial coherence, while ambiguous scenes were found in the middle of the space. These statistics also modulate evoked EEG responses in early stages of visual processing. Redrawn from data published in [13]. B) Locating information within a scene. Fixation density heat map during a visual search task. Participants combined precise search templates (object image, bottom row) and reliable spatial expectations (normal vs. switched arrangements, columns) to improve oculomotor efficiency. Reproduced, with permission, from [46]. C) Navigating through scenes. A virtual arena paradigm used to test contributions of landmark and boundary information. Participants learn target locations, which are tethered to a boundary or landmark location. Disruption of OPA with transcranial magnetic stimulation affected navigation with respect to boundaries but not with respect to landmarks. Modified with permission from [70]. D) Actions afforded by a scene. An empirically derived scene function feature space (containing, for example, a dimension separating solitary outdoor activities from social indoor activities) correlated more strongly with scene categorization behaviour than various other models of scene properties, including object labels, CNN representations, and low-level feature models. The variance explained by the function space was partly unique and partly shared with the other models. Modified, with permission, from [78].

Figure 3. Investigating multiple properties and goals. A) Dynamic coding of scene properties. Four different scene categories were created by systematically varying spatial boundary (open vs. closed) and scene content (natural vs. manufactured), and each category contained twelve unique structural layouts and twelve textures. Both scene content and a task manipulation (attend to layout or texture) modulated whether spatial boundary could be decoded from fMRI responses across multiple scene-selective areas. Modified, with permission, from [123]. B) Correlations between scene properties. Three models of scene properties were compared in terms of their inter-correlations and ability to predict fMRI responses in scene-selective cortex: 1) Fourier power at four major orientations, subdivided in low versus high frequencies, as well as a total energy measure; 2) Subjective distance to salient objects in the scene, divided in five different bins; and 3) Object labels, binned in 19 categories. Dashed white outlines indicate an example of high feature correlation between models: pictures containing sky tend to have far distance ratings and relatively high spatial frequency in the horizontal dimension, potentially due to the presence of a thin horizon line and tiny objects in faraway scenes (e.g., beaches; see also Figure 1B). As a result, most of the variance in response magnitude in scene-selective areas is shared across models: Venn diagram colors indicate variance explained by each model and their combinations, and grey shows shared variance across all three models. Adapted, with permission, from [125].

1 Torralba, Antonio and Oliva, A. (2003) Statistics of natural image categories. *Netw. Comput. Neural Syst.* 14, 391–412

2 Malcolm, G.L. *et al.* (2014) Beyond gist: strategic and incremental information accumulation for scene categorization. *Psychol. Sci.* 25, 1087–1097

3 Greene, M.R. *et al.* (2015) What you see is what you expect: rapid scene understanding benefits from prior experience. *Atten. Percept. Psychophys.* 77, 1239–1251

4 Tversky, B. and Hemenway, K. (1983) Categories of environmental scenes. *Cogn. Psychol.* 15, 121–149

5 VanRullen, R. (2011) Four common conceptual fallacies in mapping the time course of recognition. *Front. Psychol.* 2, 1–6

6 Oliva, A. (2013) Scene Perception. *New Vis. Neurosci.* DOI: 10.1111/1467-9280.02459

7 Potter, M.C. *et al.* (2014) Detecting meaning in RSVP at 13 ms per picture. *Atten. Percept. Psychophys.* 76, 270–9

8 Maguire, J.F. and Howe, P.D.L. (2016) Failure to detect meaning in RSVP at 27 ms per picture. *Attention, Perception, Psychophys.* DOI: 10.3758/s13414-016-1096-5

9 Oliva, A. and Torralba, A. (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175

10 Scholte, H.S. *et al.* (2009) Brain responses strongly correlate with Weibull image statistics when processing natural images. *J. Vis.* 9, 1–15

11 Greene, M.R. and Oliva, A. (2009) Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cogn. Psychol.* 58, 137–179

12 Greene, M.R. and Oliva, A. (2009) The briefest of glances: The time course of natural scene understanding. *Psychol. Sci.* 20, 464–472

13 Groen, I.I.A. *et al.* (2013) From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *J. Neurosci.* 33, 18814–24

14 Walther, D.B. and Shen, D. (2014) Nonaccidental properties underlie human categorization of complex natural scenes. *Psychol. Sci.* 25, 851–60

15 Oliva, A. and Schyns, P.G. (2000) Diagnostic colors mediate scene recognition. *Cogn. Psychol.* 41, 176–210

16 Goffaux, V. *et al.* (2005) Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Vis. cogn.* 12, 878–892

17 Castelhano, M.S. and Henderson, J.M. (2008) The influence of color on the activation of scene gist. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 660–675

18 Davenport, J.L. and Potter, M.C. (2004) Scene consistency in object and background perception. *Psychol. Sci.* 15, 559–564

19 Joubert, O. *et al.* (2007) Processing scene context: Fast categorization and object interference. *Vision Res.* 47, 3286–3297

20 Gagne, C.R. and MacEvoy, S.P. (2014) Do simultaneously viewed objects influence scene recognition individually or as groups? Two perceptual studies. *PLoS One* 9,

21 Greene, M.R. *et al.* (2014) Visual Noise from Natural Scene Statistics Reveals Human Scene Category Representations. 5134,

22 Glanemann, R. *et al.* (2016) Rapid apprehension of the coherence of action scenes. *Psychon. Bull. Rev.* DOI: 10.3758/s13423-016-1004-y

23 Kadar, I. and Ben-Shahar, O. (2012) A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *J. Vis.* 12, 1–17

24 Fei-Fei, L. *et al.* (2007) What do we perceive in a glance of a real-world scene? *J. Vis.* 7, 1–29

25 Rayner, K. *et al.* (1981) Masking of foveal and parafoveal vision during eye fixations in reading. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 167–179

26 Tatler, B.W. *et al.* (2016) Looking at Domestic Textiles: An Eye-Tracking Experiment Analysing Influences on Viewing Behaviour at Owlpen Manor. *Text. Hist.* 47, 94–118

27 Nuthmann, A. (2014) How do the regions of the visual field contribute to object search in real-world scenes? Evidence from eye movements. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 342–360

28 Castelhano, M.S. *et al.* (2009) Viewing Task Influences on Eye Movements During Scene Perception. *J. Vis.* 9, 1–15

29 Tatler, B.W. and Vincent, B.T. (2009) The prominence of behavioural biases in eye guidance. *Vis. cogn.* 17, 1029–1054

30 Malcolm, G.L. and Shomstein, S. (2015) Object-Based Attention in Real-World Scenes. *J. Exp. Psychol. Gen.* 144, 257–263

31 Xu, J. *et al.* (2014) Predicting human gaze beyond pixels. *J. Vis.* 14, 28

32 Foulsham, T. and Kingstone, A. (2013) Optimal and preferred eye landing positions in objects and scenes. *Q. J. Exp. Psychol. (Hove).* 66, 1707–28

33 Pajak, M. and Nuthmann, A. (2013) Object-based saccadic selection during scene perception: Evidence from viewing position effects. *J. Vis.* 13, 1–21

34 Nuthmann, A. and Einhauser, W. (2015) A new approach to modeling the influence of image features on fixation selection in scenes. *Ann. N. Y. Acad. Sci.* 1339, 82–96

35 Tatler, B.W. *et al.* (2006) The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Res.* 46, 1857–1862

36 Nuthmann, A. and Malcolm, G.L. (2016) Eye-guidance during scene search: The role color plays in central and peripheral vision. *J. Vis.* 16(2), 1–16

37 Reeder, R.R. and Peelen, M. V (2013) The contents of the search template for category-level search in natural scenes. *J. Vis.* 13, 1–13

38 Eckstein, M.P. *et al.* (2006) Attentional Cues in Real Scenes, Saccadic Targeting, and Bayesian Priors. *Psychol. Sci.* 17, 973–980

39 Pereira, E.J. and Castelhano, M.S. (2014) Peripheral Guidance in Scenes: The Interaction of Scene Context and Object Content. *J. Exp. Psychol. Hum. Percept. Perform.* 40, Advance online publication.

40 Mack, S.C. and Eckstein, M.P. (2011) Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *J. Vis.* 11, 1–9

41 Hwang, A.D. *et al.* (2011) Semantic guidance of eye movements in real-world scenes. *Vision Res.* 51, 1192–1205

42 Coco, M.I. *et al.* (2014) The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Q. J. Exp. Psychol.* 67, 1096–120

43 Wu, C.C. *et al.* (2014) The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Res.* 105, 10–20

44 Castelhano, M.S. and Heaven, C. (2011) Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychol. Bull. Rev.* 18, 890–896

45 Spotorno, S. *et al.* (2014) How context information and target information guide the eyes from the first epoch of search in real-world scenes. *J. Vis.* 14(2), 1–21

46 Spotorno, S. *et al.* (2015) Disentangling the effects of spatial inconsistency of targets and distractors when searching in realistic scenes. *J. Vis.* 15, 1–21

47 Võ, M.L.-H. and Wolfe, J.M. (2013) The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition* 126, 198–212

48 Ehinger, K.A. *et al.* (2009) Modelling search for people in 900 scenes: A combined source model of eye guidance. *Vis. cogn.* 17, 945–978

49 Navalpakkam, V. and Itti, L. (2005) Modeling the influence of task on attention. *Vision Res.* 45, 205–231

50 Hillstrom, A.P. *et al.* (2012) The effect of the first glimpse at a scene on eye movements during search. *Psychon. Bull. Rev.* 19, 204–210

51 Draschkow, D. *et al.* (2014) Seek and you shall remember: Scene semantics interact with visual search to build better memories. *J. Vis.* 14, 10–10

52 Josephs, E.L. *et al.* (2016) Gist in time: Scene semantics and structure enhance recall of searched objects. *Acta Psychol. (Amst).* 169, 100–108

53 Olejarczyk, J.H. *et al.* (2014) Incidental memory for parts of scenes from eye movements. *Vis. cogn.* 22, 975–995

54 Hayhoe, M.M. and Ballard, D. (2014) Modeling task control of eye movements. *Curr. Biol.* 24, R622–R628

55 Jovancevic, J. *et al.* (2006) Control of attention and gaze in complex environments. *J. Vis.* 6, 1431–1450

56 Ehinger, K.A. and Wolfe, J.M. (2016) When is it time to move to the next map? Optimal foraging in guided visual search. *Atten. Percept. Psychophys.*

57 Wolfe, J.M. (2013) When is it time to move to the next raspberry bush? Foraging rules in human visual search. *J. Vis.* 13, 10

58 Montello, D.R. (1993) Scale and multiple psychologies of space. In *Spatial information theory: A theoretical basis for GIS* (Frank, A. U. and Campari, I., eds), pp. 312–321, Springer

59 Foulsham, T. *et al.* (2011) The where, what and when of gaze allocation in the lab and the natural environment. *Vision Res.* 51, 1920–1931

60 Zito, G.A. *et al.* (2015) Street crossing behavior in younger and older pedestrians: an eye- and head-tracking study. *BMC Geriatr.* 15, 176

61 Marigold, D.S. and Patla, A.E. (2007) Gaze fixation patterns for negotiating

complex ground terrain. *Neuroscience* 144, 302–313

62 Loomis, J.M. *et al.* (1999) Human Navigation by Path Integration. In *Wayfinding behavior: Cognitive mapping and other spatial processes* pp. 125–151

63 Miller, J. and Carlson, L. (2011) Selecting landmarks in novel environments. *Psychon. Bull. Rev.* 18, 184–191

64 Auger, S.D. *et al.* (2012) Retrosplenial Cortex Codes for Permanent Landmarks. *PLoS One* 7, e43620

65 Stankiewicz, B.J. and Kalia, A.A. (2007) Acquisition of structural versus object landmark knowledge. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 378–390

66 Epstein, R.A. and Vass, L.K. (2014) Neural systems for landmark-based wayfinding in humans. *Philos. Trans. R. Soc. B* 369, 1–7

67 Mou, W. and Zhou, R. (2013) Defining a Boundary in Goal Localization: Infinite Number of Points or Extended Surfaces. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1115–1127

68 Lee, S.A. and Spelke, E.S. (2011) Young children reorient by computing layout geometry, not by matching images of the environment. *Psychon. Bull. Rev.* 18, 192–8

69 Doeller, C.F. *et al.* (2008) Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proc Natl Acad Sci U S A* 105, 5915–5920

70 Julian, J.B. *et al.* (2016) The Occipital Place Area Is Causally Involved in Representing Environmental Boundaries during Navigation. *Curr. Biol.* DOI: 10.1016/j.cub.2016.02.066

71 Gibson, J.J. (1986) *The Ecological Approach To Visual Perception*, Taylor & Francis Group.

72 Bainbridge, W.A. and Oliva, A. (2015) Interaction envelope: Local spatial representations of objects at all scales in scene-selective regions. *Neuroimage* 122, 408–416

73 Castelhano, M.S. and Witherspoon, R.L. (2016) How You Use It Matters: Object Function Guides Attention During Visual Search in Scenes. *Psychol. Sci.* 1, 16

74 Oliva, A. and Torralba, A. (2007) The Role of Context In Object Recognition. *Trends Cogn. Sci.* 11, 520–527

75 Wokke, M.E. *et al.* (2016) Conflict in the kitchen : Contextual modulation of responsiveness to affordances. *Conscious. Cogn.* 40, 141–146

76 Bar, M. *et al.* (2006) Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci.* 103, 449–454

77 Wurm, M.F. and Schubotz, R.I. (2016) What's she doing in the kitchen? Context helps when actions are hard to recognize. *Psychon. Bull. Rev.* DOI: 10.3758/s13423-016-1108-4

78 Greene, M.R. *et al.* (2016) Visual scenes are categorized by function. *J. Exp. Psychol. Gen.* 145, 82–94

79 Rietveld, E. and Kiverstein, J. (2014) A Rich Landscape of Affordances. *Ecol. Psychol.* 26, 325–352

80 Battaglia, P.W. *et al.* (2013) Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18327–32

81 Fischer, J. *et al.* (2016) The functional neuroanatomy of intuitive physical inference. *Prep.* DOI: 10.1073/pnas.1610344113

82 Epstein, R.A. (2014) Neural systems for visual scene recognition. In *Scene vision: making sense of what we see* (Kveraga, K. and Bar, M., eds), pp. 105–134, MIT Press

83 Kornblith, S. *et al.* (2013) A network for scene processing in the macaque temporal lobe. *Neuron* 79, 766–781

84 Lafer-Sousa, R. and Conway, B.R. (2013) Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nat. Neurosci.* 16, 1–13

85 Verhoef, B.-E. *et al.* (2015) Functional architecture for disparity in macaque inferior temporal cortex and its relationship to the architecture for faces, color, scenes, and visual field. *J. Neurosci.* 35, 6952–68

86 Vaziri, S. *et al.* (2014) A channel for 3D environmental shape in anterior inferotemporal cortex. *Neuron* 84, 55–62

87 Rajimehr, R. *et al.* (2011) The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLOS Biol.* 9, e1000608

88 Watson, D.M. *et al.* (2014) Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage* 99, 402–410

89 Watson, D.M. *et al.* (2016) Patterns of neural response in scene-selective regions of the human brain are affected by low-level manipulations of spatial frequency. *Neuroimage* 124, 107–117

90 Kauffmann, L. *et al.* (2015) The Neural Bases of the Semantic Interference of Spatial Frequency-based Information in Scenes. *J. Cogn. Neurosci.* 27, 2394–2405

91 Nasr, S. and Tootell, R.B.H. (2012) A Cardinal Orientation Bias in Scene-Selective Visual Cortex. *J. Neurosci.* 32, 14921–14926

92 Cant, J.S. and Xu, Y. (2012) Object Ensemble Processing in Human Anterior-Medial Ventral Visual Cortex. *J. Neurosci.* 32, 7685–7700

93 Nasr, S. *et al.* (2014) Thinking Outside the Box: Rectilinear Shapes Selectively Activate Scene-Selective Cortex. *J. Neurosci.* 34, 6721–6735

94 Bryan, P.B. *et al.* (2016) Rectilinear Edge Selectivity Is Insufficient to Explain the Category Selectivity of the Parahippocampal Place Area. *Front. Hum. Neurosci.* 10, 137

95 Choo, H. and Walther, D.B. (2016) Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *Neuroimage* DOI: 10.1016/j.neuroimage.2016.04.021

96 Harel, A. *et al.* (2013) Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cereb. Cortex* 23, 947–957

97 Konkle, T. and Oliva, A. (2012) A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron* 74, 1114–1124

98 Mullally, S.L. and Maguire, E.A. (2011) A New Role for the Parahippocampal Cortex in Representing Space. *J. Neurosci.* 31, 7441–7449

99 Stansbury, D.E. *et al.* (2013) Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron* 79, 1025–

1034

100 Park, S. *et al.* (2015) Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain. *Cereb. Cortex* 25, 1792–1805

101 Park, S. *et al.* (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

102 Kravitz, D.J. *et al.* (2011) Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci.* 31, 7322–33

103 Ferrara, K. and Park, S. (2016) Neural representation of scene boundaries. *Neuropsychologia* 89, 180–190

104 Walther, D.B. *et al.* (2011) Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proc. Natl. Acad. Sci.* 108, 9661–9666

105 Walther, D.B. *et al.* (2009) Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *J. Neurosci.* 29, 10573–10581

106 Baumann, O. and Mattingley, J.B. (2016) Functional Organization of the Parahippocampal Cortex: Dissociable Roles for Context Representations and the Perception of Visual Scenes. *J. Neurosci.* 36, 2536–2542

107 Aminoff, E.M. *et al.* (2013) The role of the parahippocampal cortex in cognition. *Trends Cogn. Sci.* 17, 379–390

108 Marchette, S.A. *et al.* (2015) Outside Looking In: Landmark Generalization in the Human Navigational System. *J. Neurosci.* 35, 14896–908

109 Persichetti, A.S. and Dilks, D.D. (2016) Perceived egocentric distance sensitivity and invariance across scene-selective cortex. *Cortex* 77, 155–163

110 Dilks, D.D. *et al.* (2011) Mirror-Image Sensitivity and Invariance in Object and Scene Processing Pathways. *J. Neurosci.* 31, 11305–11312

111 Kamps, F.S. *et al.* (2016) The occipital place area represents the local elements of scenes. *Neuroimage* 132, 417–424

112 Dilks, D.D. *et al.* (2013) The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. *J. Neurosci.* 33, 1331–1336

113 Auger, S.D. *et al.* (2015) A central role for the retrosplenial cortex in de novo environmental learning. *Elife* 4,

114 Troiani, V. *et al.* (2014) Multiple object properties drive scene-selective regions. *Cereb. Cortex* 24, 883–897

115 Marchette, S.A. *et al.* (2014) Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. *Nat. Neurosci.* 17, 1598–1606

116 Vass, L.K. and Epstein, R.A. (2013) Abstract Representations of Location and Facing Direction in the Human Brain. *J. Neurosci.* 33, 6133–6142

117 Vass, L.K. and Epstein, R.A. (2016) Common Neural Representations for Visually Guided Reorientation and Spatial Imagery. *Cereb. Cortex* DOI: 10.1093/cercor/bhv343

118 Baldassano, C. *et al.* (2016) Two distinct scene processing networks connecting vision and memory. *bioRxiv* DOI: 10.1101/057406

119 Baldassano, C. *et al.* (2013) Differential connectivity within the Parahippocampal Place Area. *Neuroimage* 75, 236–245

120 Silson, E.H. *et al.* (2016) Scene-selectivity and retinotopy in medial parietal cortex. *Front. Hum. Neurosci.* 10, 1–17

121 Silson, E.H. *et al.* (2015) A Retinotopic Basis for the Division of High-Level Scene Processing between Lateral and Ventral Human Occipitotemporal Cortex. *J. Neurosci.* 35, 11921–11935

122 Silson, E.H. *et al.* (2016) Evaluating the correspondence between face- , scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. 16, 1–21

123 Lowe, M.X. *et al.* (2016) Feature diagnosticity and task context shape activity in human scene-selective cortex. *Neuroimage* 125, 681–692

124 Groen, I.I.A. *et al.* (2015) The time course of natural scene perception with reduced attention. *J. Neurophysiol.* DOI: 10.1152/jn.00896.2015

125 Lescroart, M.D. *et al.* (2015) Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Front. Comput. Neurosci.* 9, 1–20

126 Groen, I.I.A. *et al.* Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philos. Trans. R. Soc. B* DOI: 10.1098/not

127 Kravitz, D.J. *et al.* (2013) The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* 17, 26–49

128 Henderson, J.M. and Hollingworth, A. (1999) High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271

129 Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365

130 Kriegeskorte, N. (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* 1, 417–446

131 Güçlü, U. and van Gerven, M. a. J. (2015) Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* 35, 10005–10014

132 Khaligh-Razavi, S.-M. *et al.* (2014) Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput. Biol.* 10, e1003915

133 Cichy, R.M. *et al.* (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6, 27755

134 Yamins, D.L.K. *et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8619–24

135 Zhou, B. *et al.* Learning Deep Features for Scene Recognition using Places Database. . (2014) , 487–495

136 Cichy, R.M. *et al.* (2016) Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage* DOI: 10.1016/j.neuroimage.2016.03.063

137 Zhou, B. *et al.* (2015) , Object Detectors emerge in Deep Scene CNNs. , in *International Conference on Learning Representations*

138 Razavian, A.S. *et al.* (2014) , CNN Features off-the-shelf: an Astounding Baseline for Recognition. , in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813

139 Graham, N. (1979) Does the brain perform a Fourier analysis of the visual scene? *Trends Neurosci.* 2, 207–208

140 Field, D.J. (1987) Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.* 4, 2379–94

141 Ghebreab, S. *et al.* (2009) , A biologically plausible model for rapid natural image identification. , in *Advances in Neural Information Processing Systems*, pp. 629–637

142 Groen, I.I.A. *et al.* (2012) Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLoS Comput. Biol.* 8, e1002726

143 Canário, N. *et al.* (2016) Distinct preference for spatial frequency content in ventral stream regions underlying the recognition of scenes, faces, bodies and other objects. *Neuropsychologia* 87, 110–119

144 Talebi, V. and Baker, C.L. (2012) Natural versus Synthetic Stimuli for Estimating Receptive Field Models: A Comparison of Predictive Robustness. *J. Neurosci.* 32, 1560–1576

145 David, S. V *et al.* (2004) Natural Stimulus Statistics Alter the Receptive Field Structure of V1 Neurons. *J. Neurosci.* 24, 6991–7006

146 Oliva, A. and Schyns, P.G. (1997) Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cogn. Psychol.* 34, 72–107

147 Awasthi, B. *et al.* (2013) Distinct spatial scale sensitivities for early categorization of faces and places: neuromagnetic and behavioral findings. *Front. Hum. Neurosci.* 7, 1–11

Outstanding Questions Box

- To what extent are a combination of properties that facilitate a single goal (e.g., edges for scene recognition), generalizable to other tasks (e.g., edges for navigation)?
- To what extent do co-occurring features within a scene category interact, or contribute independently to scene understanding?
- Can multiple visual goals be carried out simultaneously, or do they interfere with each other? If simultaneously, can a single property be used for multiple purposes or can simultaneous goals occur when they require different features? If they interfere, is a new internal representation generated each time the goal momentarily changes?
- Is the neural representation of scenes modulated by the degree of co-occurrence between different levels of description (e.g., do typical scene exemplars have higher consistency between features and therefore more reliable neural representations?)
- What is the nature of the stored representation that a recognized scene is matched with? Is it based on physical characteristics or a set of rules? Does this change with recognized hierarchical level of detail?
- How do the representations in scene-selective cortex compare with the representations in other regions implicated in navigation, such as hippocampus and entorhinal cortex?

Key Figure

Click here to download Key Figure Figures_TICS_revision_Fig1.pdf

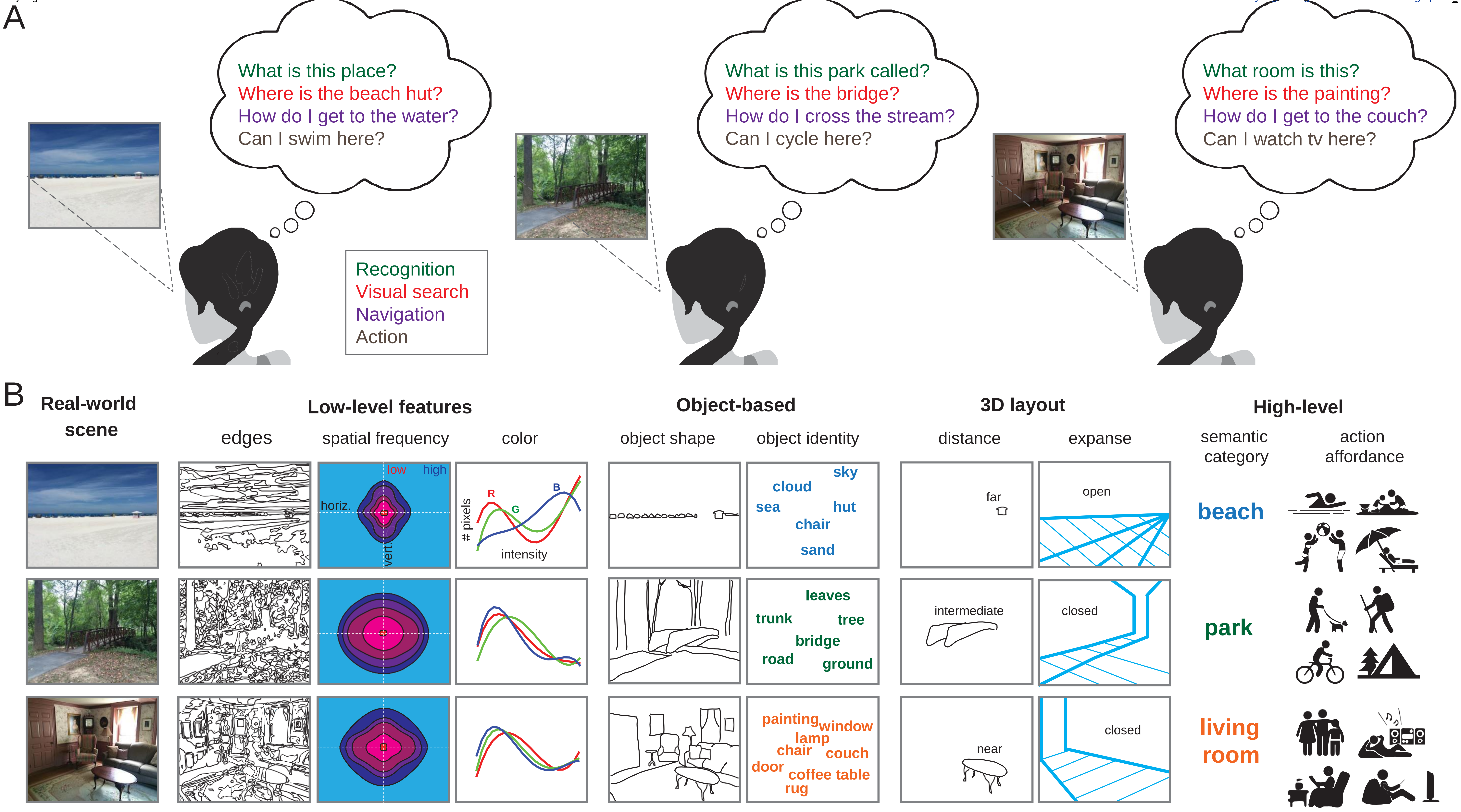
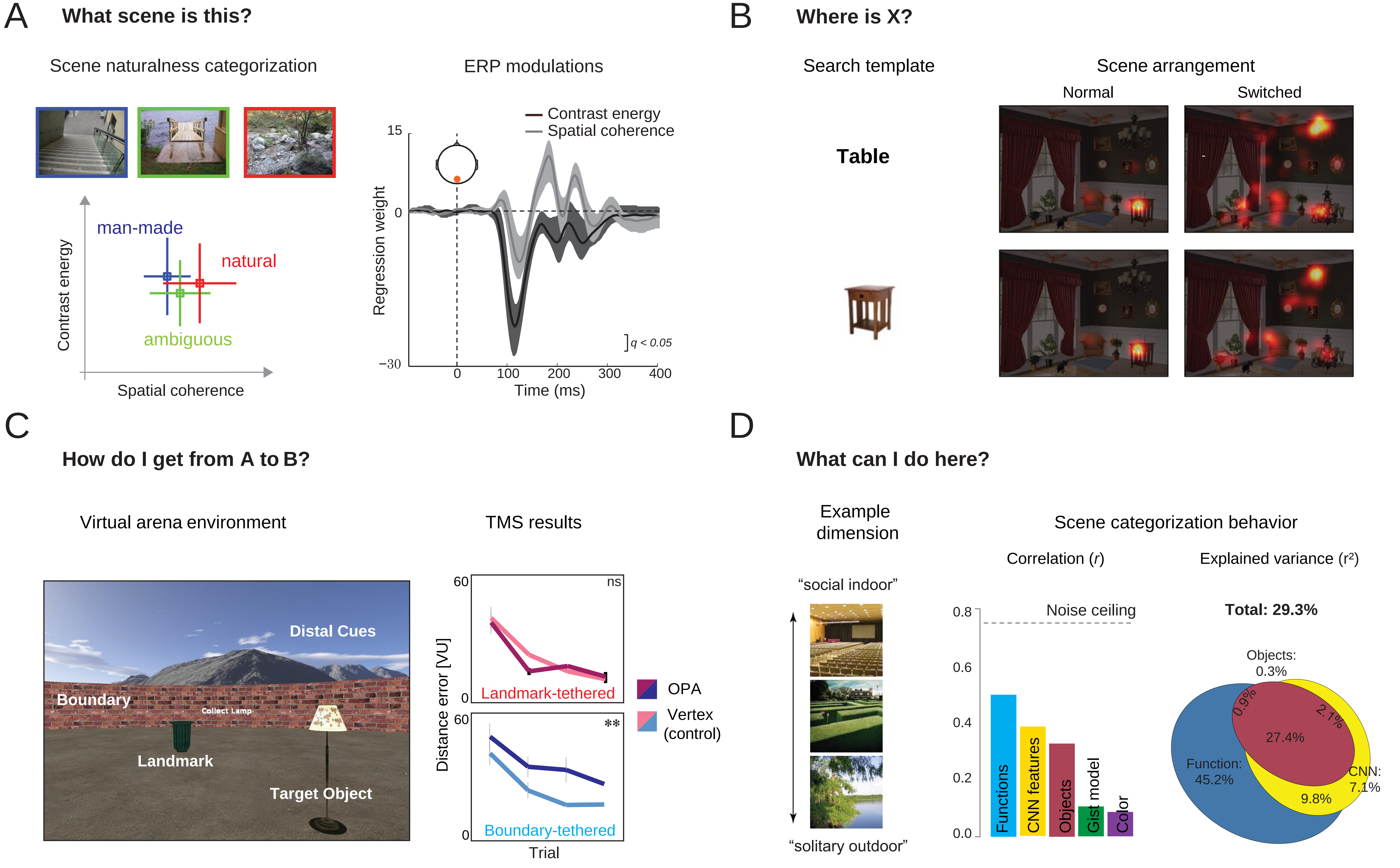
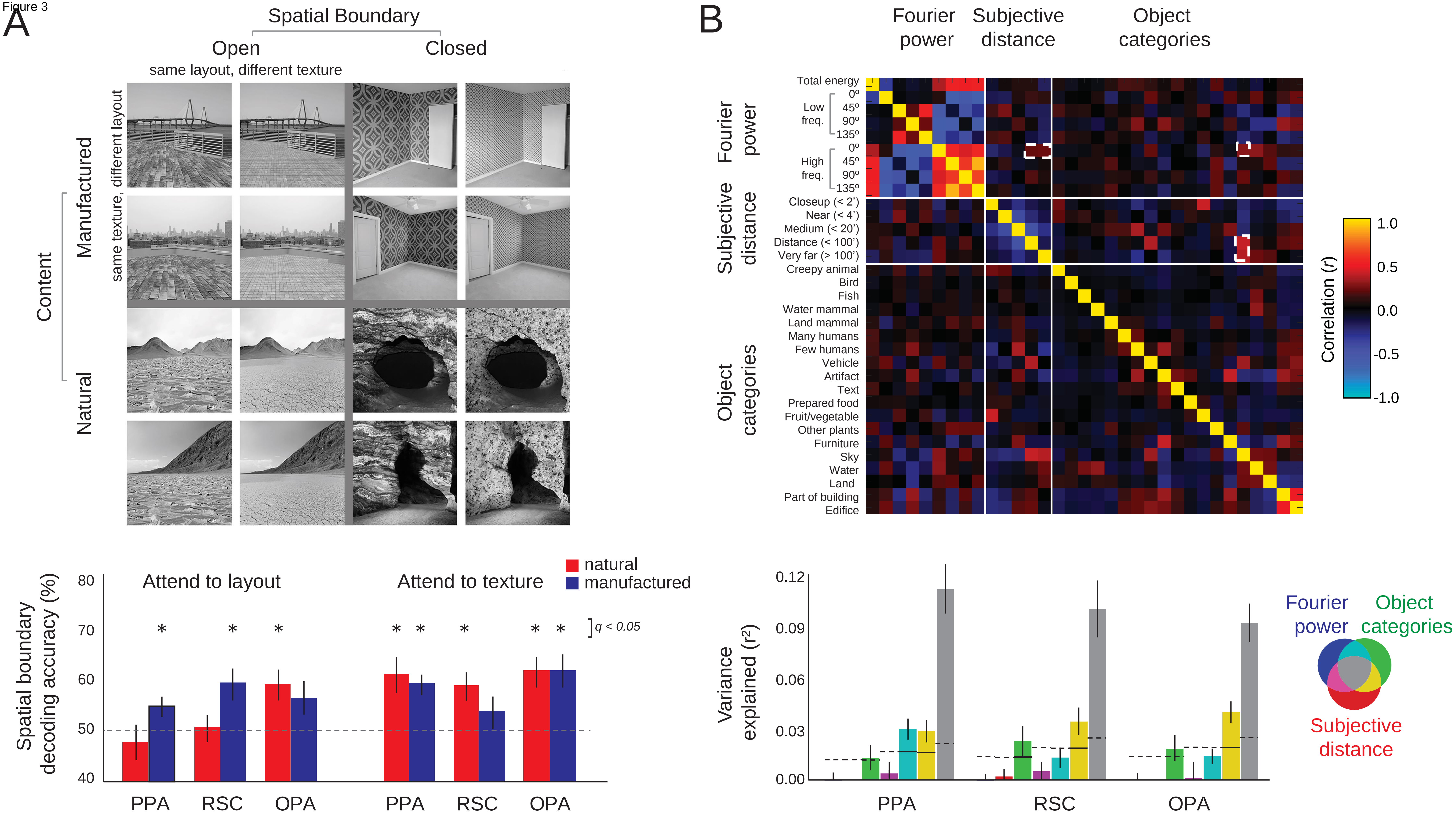


Figure 2



A
What scene is this?
Scene naturalness categorization
Contrast energy
Spatial coherence
man-made
natural
ambiguous
ERP modulations
Contrast energy
Spatial coherence
15
0
−30
Regression weight
0
100
200
300
400
Time (ms)
q < 0.05
B
Where is X?
Search template
Table
Scene arrangement
Normal
Switched
C
How do I get from A to B?
Virtual arena environment
Distal Cues
Boundary
Collect Lamp
Landmark
Target Object
TMS results
60
0
ns
Landmark-tethered
60
0
**
Boundary-tethered
Distance error [VU]
Trial
OPA
Vertex (control)
D
What can I do here?
Example dimension
"social indoor"
"solitary outdoor"
Scene categorization behavior
Correlation (r)
Noise ceiling
0.8
0.6
0.4
0.2
0.0
Functions
CNN features
Objects
Gist model
Color
Explained variance (r²)
Total: 29.3%
Objects: 0.3%
0.9%
2.1%
27.4%
Function: 45.2%
CNN: 7.1%
9.8%

Figure 3


A
Spatial Boundary
Open
Closed
same layout, different texture
same texture, different layout
Content
Manufactured
Natural
Spatial boundary decoding accuracy (%)
80
70
60
50
40
Attend to layout
Attend to texture
natural
manufactured
*
*
*
*
*
*
*
*
] q < 0.05
PPA
RSC
OPA
PPA
RSC
OPA
B
Fourier power
Subjective distance
Object categories
Fourier power
Subjective distance
Object categories
Total energy
Low freq.
0º
45º
90º
135º
High freq.
0º
45º
90º
135º
Closeup (< 2')
Near (< 4')
Medium (< 20')
Distance (< 100')
Very far (> 100')
Creepy animal
Bird
Fish
Water mammal
Land mammal
Many humans
Few humans
Vehicle
Artifact
Text
Prepared food
Fruit/vegetable
Other plants
Furniture
Sky
Water
Land
Part of building
Edifice
Correlation (r)
1.0
0.5
0.0
-0.5
-1.0
Variance explained (r²)
0.12
0.09
0.06
0.03
0.00
PPA
RSC
OPA
Fourier power
Object categories
Subjective distance