3

The Constructive Nature of Scene Perception

Soojin Park and Marvin M. Chun

Humans have the remarkable ability to recognize complex, real-world scenes in a single, brief glance. The *gist*, the essential meaning of a scene, can be recognized in a fraction of a second. Such recognition is sophisticated, in that people can accurately detect whether an animal is present in a scene or not, what kind of event is occurring in a scene, as well as the scene category, all in as little as 150 ms (Potter, 1976; Schyns & Oliva, 1994; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). With this remarkable ability, the experience of scene perception feels effortless. It is ubiquitous as it is fundamental—after all, every image that comes into our brain is a scene. Scene perception directly impacts our actions in a 3D world by providing information about where we are as well as where we should navigate. This requires the integration of views across eye movements and across time to connect the present view with past memory. Thus, as effortless as it seems, scene perception involves many different levels of computation that integrate space, time, and memory. In this chapter we demonstrate the constructive nature of scene perception involving different brain regions to achieve a meaningful experience of the visual world.

The human visual system has three overarching goals in processing the visual environment. First, at the moment of physical input, the visual system must rapidly compute diagnostic properties of space and objects contained in the scene. Aside from recognizing faces and communicating with people, our daily activities require comprehension of the environment's spatial layout for the purposes of navigation as well as recognition of objects contained within that environment. As you view a particular scene, you are rapidly computing its spatial structure: determining where buildings are located and identifying paths through which you might navigate. At the same time, you can recognize a scene as a part of the broader environment and as a familiar scene in your memory. Visual scene understanding thus involves integrating a series of computations to enable coherent and meaningful scene perception.

Spatial structure, landmarks, and navigational paths are the major structural properties that define a scene. Recognizing these different structural properties is central to scene perception. Scenes with similar sets of structural properties will be grouped

into similar scene categories. For example, if two scenes both have an open spatial layout, natural content, and strong navigability, they will both be categorized as fields. On the other hand, if two scenes have some overlapping structural properties but differ largely in other properties, they will be categorized differently. For example, if both scenes have an open spatial layout, but one has urban content and the other has natural content, they will be categorized differently (e.g., a highway vs. a field). Thus, a scene category is defined by combinations of different structural features (e.g., spatial layout and objects), and these structural features dictate how the viewer will recognize the space and function within it. In the first part of the chapter we examine how the brain represents structural property dimensions of scenes.

If the initial problem of scene recognition involves integrating multiple structural properties into a representation of a single view of a scene, then the second major challenge for the visual system is the problem of perceptual integration. To describe this problem, we should define the following terms—*view, scene,* and *place*—which depend on the observer's interactions with the environment (Oliva, Park, & Konkle, 2011). When an observer navigates in the real world, the observer is embedded in a space of a given "place," which is a location or landmark in the environment and often carries semantic meaning (e.g., the Yale campus, my kitchen). A "view" refers to a particular viewpoint that the observer adopts at a particular moment in one fixation (e.g., a view of the kitchen island counter when standing in front of the refrigerator), and a "scene" refers to the broader extension of space that encompasses multiple viewpoints. For example, a scene can be composed of multiple viewpoints taken by an observer's head or eye movements (e.g., looking around your kitchen will reveal many views of one scene). Visual input is often dynamic, as the viewer moves through space and time in the real environment. In addition, our visual field is spatially limited, causing the viewer to sample the world through constant eye and head movement. Yet, in spite of this succession of discrete sensory inputs, we perceive a continuous and stable perceptual representation of our surroundings. Thus, the second challenge for scene recognition is to establish coherent perceptual scene representations from discrete sensory inputs. Specifically, this involves the balancing of two opposing needs: each view of a scene should be distinguished separately to infer the viewer's precise position and direction in a given space, but these disparate views must be linked to surmise that these scenes are part of the same broader environment or "place." In the second part of this chapter we discuss how the human visual system represents an integrated visual world from multiple discrete views that change over time. In particular, we focus on different functions of the parahippocampal place area (PPA) and retrosplenial complex (RSC) in representing and integrating multiple views of the same place.

A third challenge for the visual system is to mentally represent a scene in memory after the viewer moves away from a scene and the perceptual view of the scene has disappeared. We often bring back to our mind what we just saw seconds ago, or need

to match the current view with those in memory that reflect past experience. Such memory representations can closely reflect the original visual input, or they may be systematically distorted in some way. In the last part of the chapter we describe studies that test the precise nature of scene memory. In particular, we show that the scene memory is systematically distorted to reflect a greater expanse than the original retinal input, a phenomenon called *boundary extension*.

These complex visual and memory functions are accomplished by a network of specialized cortical regions devoted to processing visual scene information (figure 3.1, plate 2). Neuroimaging studies of scene recognition have provided insight about the functioning of these specialized cortical regions. Among them, the most well-known region is the parahippocampal place area (PPA) near the medial temporal region, which responds preferentially to pictures of scenes, landmarks, and spatial layouts depicting 3D space (Aguirre, Zarahn, & D'Esposito, 1998; Epstein, Harris, Stanley, & Kanwisher, 1999; Epstein & Kanwisher, 1998; Janzen & Van Turennout, 2004). The PPA is most sensitive to the spatial layout or 3D structure of an individual scene, although some recent work suggests that the PPA also responds to object information such as the presence of objects in a scene (Harel, Kravitz, & Baker, 2013), large real-world objects (Konkle & Oliva, 2012), and objects with strong context (Aminoff, Kveraga, & Bar, 2013). The complexity and richness of the PPA representation are discussed further under Representing Structural Properties of a Scene.

The PPA has been one of the most studied regions to represent "scene category-specific" information; however, more recent findings suggest that there is a family of regions that respond to scenes beyond the PPA, including the retrosplenial cortex and the transverse occipital sulcus. The retrosplenial complex (RSC), a region superior to the PPA and near the posterior cingulate, responds strongly to scenes compared to other objects (just as the PPA does). Yet, the RSC shows unique properties that may be important for spatial navigation rather than visual analysis of individual scenes (Epstein, 2008; Park & Chun, 2009; Vann, Aggleton, & Maguire, 2009). For example, the RSC shows relatively greater activations than the PPA for route learning in a virtual environment, mentally navigating in a familiar space, and recognizing whether a scene is a familiar one in memory (Epstein, 2008; Ino et al., 2002; Maguire, 2001). The section on Integrating a View to a Scene focuses on comparing the different functions of the PPA and RSC. The transverse occipital sulcus (TOS) also responds selectively to scenes compared to other visual stimuli. Recent findings suggest that the TOS is causally involved in scene recognition and is sensitive to mirror-reversal changes in scene orientation, whereas the PPA is not (Dilks, Julian, Kubilius, Spelke, & Kanwisher, 2011; Dilks, Julian, Paunov, & Kanwisher, 2013). Finally, in contrast to the regions above that prefer scenes over objects, the lateral occipital complex (LOC) represents object shape and category (Eger, Ashburner, Haynes, Dolan, & Rees, 2008; Grill-Spector, Kushnir, Edelman, Itzchak, & Malach, 1998; Kourtzi &

Kanwisher, 2000; Malach et al., 1995; Vinberg & Grill-Spector, 2008). Because scenes contain objects, we also consider the role of the LOC in representing the object contents and object interactions in a scene.

The goal of this chapter is to review studies that characterize the nature of scene representation within each of these scene-sensitive regions. In addition, we address how the functions of scene-specific cortical regions are linked at different stages of scene integration: structural construction, perceptual integration, and memory construction.

We propose a theoretical framework showing distinct but complementary levels of scene representation across scene-selective regions (Park & Chun, 2009; Park, Chun, & Johnson, 2010; Park, Intraub, Yi, Widders, & Chun, 2007), illustrated in figure 3.1 (plate 2). During navigation and visual exploration different physical views are perceived, and the PPA represents the visuostructural property of each view separately (Epstein & Higgins, 2007; Epstein & Kanwisher, 1998; Goh et al., 2004; Park, Brady, Greene, & Oliva, 2011; Park & Chun, 2009), encoding the geometric properties of

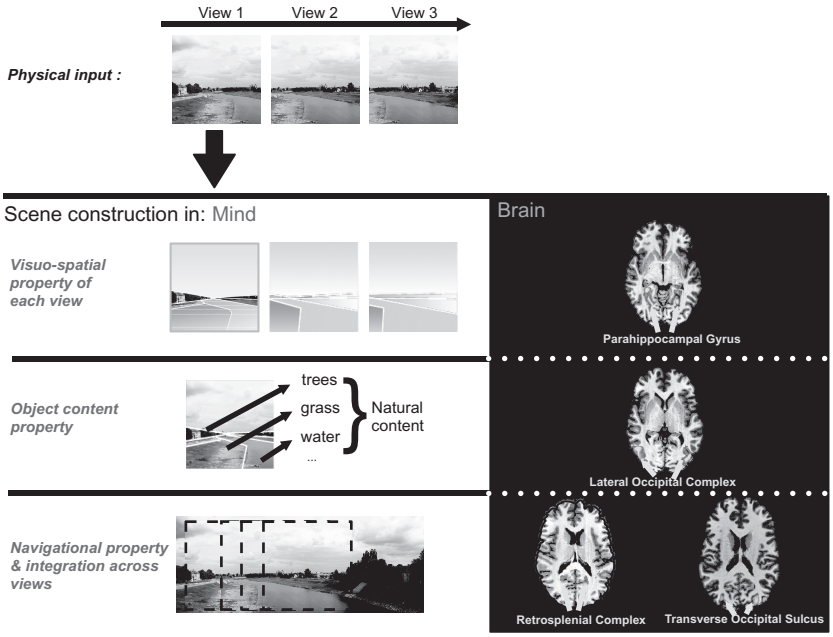

Figure 3.1 (plate 2)
A schematic illustration of three levels of scene processing. As the viewer navigates in the world, different views of scenes enter the visual system (view 1, view 2, view 3). The PPA treats each view of scenes as different from the others and is involved in analyzing the spatial properties of each specific view, such as the spatial layout and structure. The LOC processes object content properties in a scene, such as whether scenes have natural or urban content. The RSC and TOS analyze the navigationally relevant functional properties of a scene, creating an integrated representation of a scene across views.

scenes such as perspective, volume, and open/closed spatial layout, regardless of what types of objects fill in the space (Kravitz, Peng, & Baker, 2011; Park, Brady, et al., 2011; Park, Konkle, & Oliva, 2014). In parallel, the LOC represents the object properties in a scene, such as whether the scene has natural content (e.g., trees and vegetation) or whether the scene has urban content (e.g., buildings and cars; Park, Brady, et al., 2011). None of these regions represents scenes solely based on semantic category; for example, a city street and a forest will be represented similarly in the PPA as long as they have similar spatial layout, regardless of their differing semantic categories (Kravitz et al., 2011; Park, Brady, et al., 2011). The RSC represents scenes in an integrated/view-independent manner, treating different views that are spatiotemporally related as the same scene (Epstein & Higgins, 2007; Park & Chun, 2009). Given its involvement in spatial navigation in humans and rodents (Kumaran & Maguire, 2006), the RSC may also represent a scene's functional properties, such as how navigable a scene is, how many possible paths there are, or what actions the observer should take within the environment. The TOS may also represent the navigability of a scene, given that this region is sensitive to mirror-reversal changes of scenes, which alter the direction of a path (e.g., a path originally going to the left now will become a path going to right; Dilks et al., 2011). This pattern of response is similar to that of the RSC but different from that of the PPA, which does not show any sensitivity to mirror-reversal changes.

In the current chapter we present evidence that demonstrates how the distinct regions illustrated in figure 3.1 (plate 2) play a complementary role in representing the scene at the visuostructural level, perceptual integration level, and memory level.

Representing Structural Properties of a Scene

People are good at recognizing scenes, even when these scenes are presented very rapidly (Potter, 1975; also see chapter 9 by Potter in this volume). For example, when a stream of images is presented at a rapid serial visual presentation rate of around 100 ms per item, people can readily distinguish if a natural forest scene appeared among a stream of urban street images (Potter, 1975; Potter, Staub, & O'Connor, 2004). Even though people are able to recognize objects in rapidly presented scenes such as "trees," what subjects often report is in the basic-level category of a scene, such as a forest, beach, or a field (Rosch, 1978). Thus, one might assume that scenes are organized in the brain according to basic-level categories, with groups of neurons representing forest scenes, field scenes, and so on. However, recent computational models and neuroimaging studies suggest that the visual system does not classify scenes as belonging to a specific category per se but rather according to their global properties, that is, their spatial structure (Hoiem, Efros, & Hebert, 2006; Torralba & Oliva, 2003; Torralba, Oliva, Castelhano, & Henderson, 2006).

Object information and the spatial structure of a scene are extracted separately but in parallel (Oliva & Torralba, 2001) and then are later integrated to arrive at a decision about the identity of the scene or where to search for a particular object. In other words, when the visual system confronts a scene, it first decomposes the input into multiple layers of information, such as naturalness of object contents, density of texture, and spatial layout. This information is later combined to give rise to a meaningful scene category (in this example, a forest). Behavioral studies also suggest that object and scene recognition take place in an integrated manner (Davenport & Potter, 2004; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). Target objects embedded in scenes are more accurately identified in a consistent than an inconsistent background, and scene backgrounds are identified more accurately when they contain a consistent rather than inconsistent object (Davenport & Potter, 2004; Loftus & Mackworth, 1978; Palmer, 1975). We also almost never see objects devoid of background context, and many scenes are in fact defined by the kinds of objects they contain—a pool table is what makes a room a pool hall, and recognizing a pool hall thus involves the recognition of the pool table in it, in addition to the indoor space around it. Taken together, these facts indicate that objects and scenes usefully constrain one another and that any complete representation of a visual scene must integrate multiple levels of these separable properties of spatial layout and object content.

Natural scenes can be well described on the basis of global properties such as different degrees of openness, expansion, mean depth, navigability, and others (Greene & Oliva, 2009b; Oliva & Torralba, 2006). For example, a typical "field" scene has an open spatial layout with little wall structure, whereas a typical "forest" scene has an enclosed spatial layout with strong perspective of depth (figure 3.2, plate 3). In addition, a field has natural objects or textures such as grass and trees, and a forest scene typically has natural objects such as trees, rocks, and grass. Similarly, urban scenes such as a street or highway can also be decomposed according to whether the scene's horizon line is open and visible (e.g., highway) or enclosed (e.g., street), in addition to its manmade contents (e.g., cars, buildings). We recognize a field as belonging to field category and a street as belonging to a street category because the visual system immediately computes the combination of structural scene properties (e.g., spatial layout and object content). The combination of such scene properties thus constrains how we interact with scenes or navigate within them.

In the example above we mentioned the spatial and object dimensions of a scene, but it is worth noting that real-world scenes have much higher degrees of complexity and dimensionality of structural information (Greene & Oliva, 2009a, 2009b; Oliva & Torralba, 2006). In a complex real-world scene these numerous properties are often entangled and are difficult to examine separately. Indeed, most investigations concerning the neural coding of scenes have focused on whether brain regions respond to one type of category-specific stimulus compared to others (e.g., whether the PPA

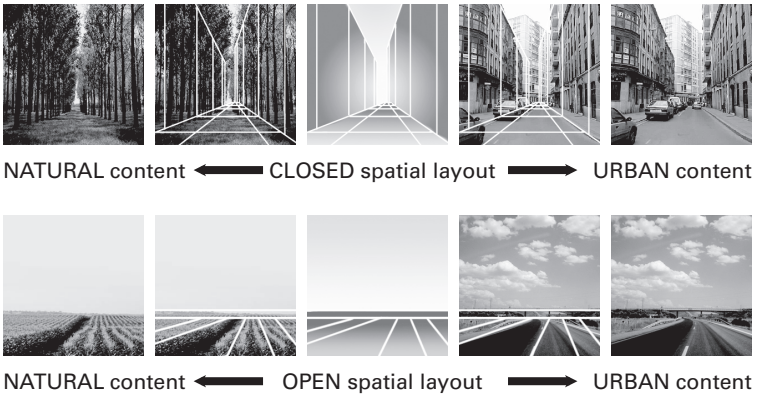

Figure 3.2 (plate 3)
A schematic illustration of spatial layout and content properties of scenes. Note that the spatial layout can correspond between natural and urban scenes. If we keep the closed spatial layout and fill in the space with natural contents, the scene becomes a forest, whereas if we fill in the space with urban contents, the scene becomes an urban street scene. Likewise, if we keep the open spatial layout and fill in the space with natural contents, the scene becomes a field; if we fill in the space with urban contents, the scene becomes a highway scene. Figure adapted from Park et al. (2011).

responds to a field vs. forest or whether LOC responds to a cut-out tree on a blank background). However, such category-specific representation may be a product of how the visual system reduces the complex dimensionality of a visual scene into a tractable set of scene categories. Thus, it is important to identify the precise dimensions in which neurons in scene-selective visual areas encode scene information.

An initial step to study scene processing in the brain should involve examining if scene categories are even represented to start with. After all, scenes in the same category (e.g., two scenes in the field category) are the scenes that share the most similar spatial and object properties (e.g., both scenes have open spatial layout, similar expansion, and natural contents and textures). Research has demonstrated that scene-responsive cortical regions such as the PPA and RSC represent the semantic category of scenes. Walther, Caddigan, Fei-Fei, and Beck (2009) used multivoxel analysis to test if patterns of fMRI activity in scene-selective cortices could classify six different natural scene categories (beach, buildings, forests, highways, industry, and mountains). Analysis of patterns of neural activity can offer more precise information about representation in a particular brain region compared to conventional methods, which average activity across voxels (Cox & Savoy, 2003; Kamitani & Tong, 2005). Machine learning methods, such as support-vector machine (SVM) classification, enable classification of different patterns of activity associated with different categories of scenes. Walther et al. (2009) found high classification performance in the PPA and RSC for distinguishing scene categories. Interestingly, they ran a separate behavioral study to measure errors in categorizing these scenes when presented very briefly (e.g.,

miscategorizing a highway scene as a beach). These behavioral error patterns were then compared to fMRI multivoxel classification error patterns, and a strong correlation was found between the two. In other words, scenes that had similar patterns of brain activity (e.g., beaches and highways) were scenes that were often confused in the behavioral scene categorization task. This elegant study showed that scene representations in the PPA reflect semantic categories and that scenes that are behaviorally confusable have similar patterns of voxel activity in this region.

What are the similarities across scene categories that made particular scenes highly confusable both behaviorally and at the neural level? The confusability between scene categories may be due to similarity in their spatial layouts (e.g., open spaces with a horizontal plane), similarity among the types of objects contained in these scenes (e.g., trees, cars, etc.), or similarity in the everyday function of scenes (e.g., spaces for transportation, spaces for social gatherings). Determining what types of scenes are systematically confused with one other can reveal whether a brain region represents spatial properties or object properties. Park et al. (2011) directly tested for such confusion errors using multivoxel pattern analysis. They asked whether two different properties of a scene, such as its spatial layout and its object content, could be dissociated within a single set of images. Instead of asking whether the PPA and LOC could accurately represent different categories of scenes, they focused on the confusion errors of a multivoxel classifier to examine whether scenes were confused based on similarity in spatial layout or object contents. There were four types of scene groups defined by spatial layout and object content (figure 3.3, plate 4: open natural scenes, open urban scenes, closed natural scenes, and closed urban scenes). Open versus closed defined whether the scene had an open spatial layout or a closed spatial layout. The natural versus urban distinction defined whether the scene had natural or urban object contents. Although both the PPA and LOC had similar levels of accurate classification performance, the patterns of confusion errors were strikingly different. The PPA made more confusion errors across images that shared the same spatial layout, regardless of object contents, whereas the LOC made more confusion errors across images that shared similar objects, regardless of spatial layout. Thus, we may conclude that a street and a forest will be represented similarly in the PPA as long as they have similar spatial layout, even though a street is an urban scene and a forest is a natural scene. On the other hand, a forest and field scene will be represented similarly in the LOC because they have similar natural contents.

Another study computed a similarity matrix of 96 scenes and also found that PPA representations are primarily based on spatial properties (whether scenes have open spatial layout vs. closed spatial layout), whereas representations in early visual cortex (EVC) are primarily based on the relative distance to the central object in a scene (near vs. far; Kravitz et al., 2011). Using a data-driven approach, the authors measured multivoxel patterns for each of 96 individual scenes. They then cross-correlated these response patterns to establish a similarity matrix between each pair of scenes.

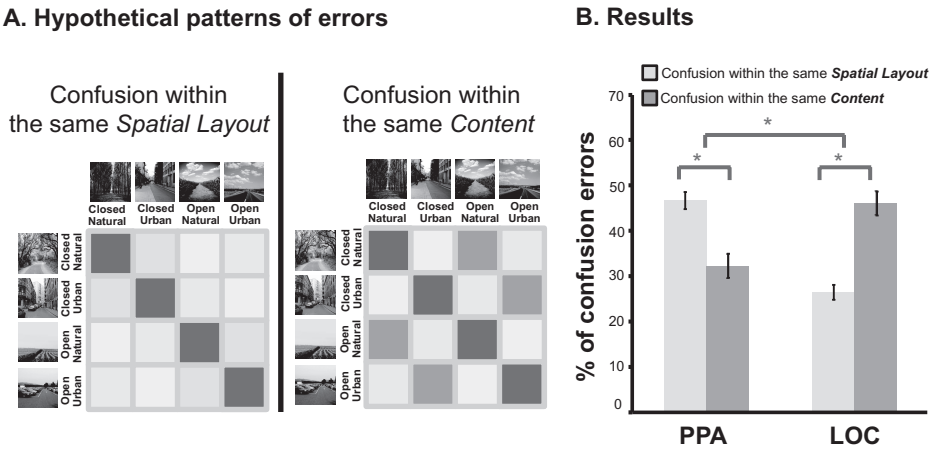

Figure 3.3 (plate 4)
(A) Hypothetical patterns of confusion errors based on the spatial layout or object content similarity. The rows represent the scene image conditions as presented to the participants, and the columns represent the scene condition that the classifier predicted from the fMRI patterns of activity. If spatial layout properties of scenes are represented in a particular brain area, we expect confusion within scenes that share the same spatial layout (marked in light gray). If content properties of scenes are important for classification, we expect confusion within scenes that share the same content (dark gray cells). (B) Confusion errors (percentage) are shown for the PPA and the LOC. Figure adapted from Park et al. (2011).

When the matrix was reorganized according to dimensions of space (open vs. closed), objects (natural vs. urban) and distance (near vs. far), there was a high correlation in the PPA for scenes that shared dimensions of space (figure 3.4A, plate 5), and high correlation in EVC for scenes that shared the dimension of distance. These results highly converge with those of Park et al. (2011), together suggesting that scene representations in the PPA and RSC are primarily based on spatial layout information and not scene category per se.

Park et al. (2011) and Kravitz et al. (2011) indicate that the PPA and LOC have relatively specialized involvement in representing spatial or object information. However, one should be careful in drawing conclusions about orthogonal or categorical scene representations across the PPA and LOC. The PPA does not exclusively represent spatial information, and the LOC does not solely represent object information. For example, Park, Brady et al. (2011) found above-chance levels of classification accuracy for four groups of scene types (open natural, open urban, closed natural, and closed urban) in both the PPA and LOC. To accurately classify these four groups of scenes, the PPA and LOC must encode both spatial layout (open vs. closed) and object information (natural vs. urban). Thus, even though the confusion error patterns suggest a preference for information concerning spatial layout in the PPA and a preference for object content information in the LOC, these functions are not exclusively specialized. In fact, scene information spans a gradient across ventral visual regions.

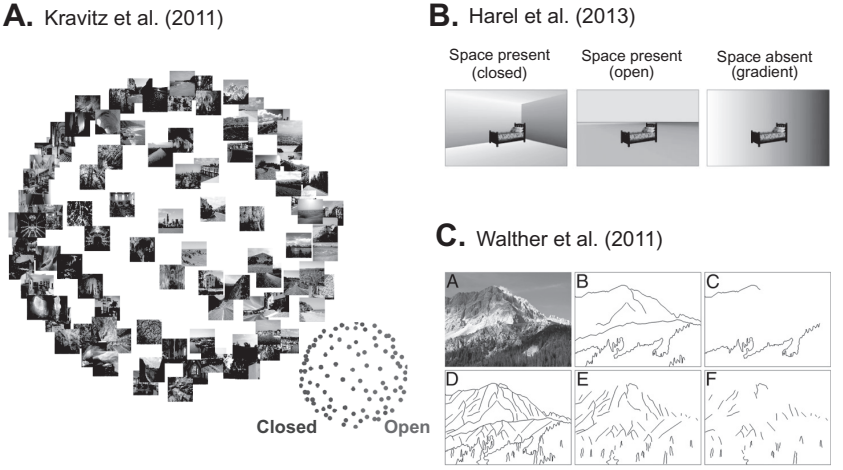

Figure 3.4 (plate 5)
(A) Multidimensional scaling plot for the PPA. Scenes are shown in a two-dimensional plane with the distance between pairs of scenes reflecting the correlation between their response patterns. Pairs of images that had higher correlations are shown closer together. Here, you can see that scenes that had similar spatial layout (closed or open) are clustered closely together (Kravitz, Peng, & Baker, 2011). (B) Stimuli used in Harel et al. (2013). Participants saw minimal scene stimuli that are composed of objects (e.g., furniture) combined with three different types of background stimuli (closed-space present, open-space present, and gradient-space absent). (C) Examples of line drawing images used in Walther et al. (2011). A corresponding line drawing is shown (D) for a photograph of a scene (A). This line drawing scene was degraded by either removing 50% of its pixels by removing local contours (short lines) (B) or global contours (long lines) (E); or by removing 75% of pixels by removing short (C) or long contours (F). The category identification performance was significantly impaired when global contours were removed (E and F) compared to when local contours were removed (B and C), suggesting that global spatial layout information is important. Figures adopted from Kravitz, Peng, and Baker (2011), Harel, Kravitz, and Balker (2013), and Walther et al. (2011).

Harel, Kravitz, and Baker (2013) manipulated spatial layout (e.g., open spatial layout, closed spatial layout, or no spatial layout) and object content (furniture present or absent; figure 3.4B, plate 5). They tested if the PPA, LOC, and RSC could correctly decode whether a scene background contained spatial layout information (space absence decoding) and whether a scene contained an object (object absence decoding). Multivoxel pattern analysis showed that RSC was able to decode whether a scene included spatial layout information but not whether a scene contained objects. In contrast, the LOC was able to decode whether a scene contained objects but not whether a scene's background included spatial layout information. The PPA was able to decode both whether the scene contained spatial layout information or object information. These results suggest that there is a gradient of representation: strong spatial representation with little object representation in the RSC; some spatial and some object representation in the PPA; and strong object representation with little spatial representation in the LOC.

Other studies have tested whether different cues for defining spatial layout information matter for scene categorization. Walther et al. (2011) suggests that scene categorization is largely based on the global structure of a scene, such as its global contours. To test whether global or local contours have different degrees of impact, Walther et al. (2011) selectively removed equal pixel amounts of global (long) or local (short) contours from a line drawing scene (figure 3.4C, plate 5). Participants performed significantly worse in identifying the categories of scenes that had global contours removed compared to scenes that had local contours removed, suggesting that global spatial layout information is more important for scene identification.

Although the studies described above have investigated spatial representation, other studies have focused on the representation of object properties in scenes. MacEvoy and Epstein (2011) were able to predict a scene category from multiple objects in the lateral occipital cortex (LO) but not in the PPA. That is, multivoxel patterns in the LO for a scene category (e.g., kitchen) were highly correlated with the average of the patterns elicited by signature objects (e.g., stove or refrigerator). These results support earlier views of scene perception, which held that real-world scene identification emerges by identifying a set of objects in it (Biederman, 1981; Friedman, 1979). However, a scene is not just a bag of objects or linear combinations of them but reflects the semantic co-occurrence or spatial composition between these objects. Objects often appear in meaningful spatial arrangements based on the functional or semantic relationship between them (e.g., a cup on a table; a pot pouring water into a cup). This interacting relationship enhances the identification of individual objects (Green & Hummel, 2006) and scenes (Biederman, 1981). Kim and Biederman (2011) tested how a collection of objects may be processed together as a scene. They asked whether a particular brain region encodes meaningful relationships among multiple objects. They showed that the LOC responds strongly to a pair of objects presented in an interacting position (e.g., a bird in front of a bird house) compared to a pair of objects presented side by side and not interacting in a semantically meaningful way. They did not find any preference for interacting objects in the PPA, consistent with the idea that the PPA does not care about object information (MacEvoy & Epstein, 2011). These studies suggest that the LO represents more than simple object shape and should be considered a scene-processing region, representing multiple objects and their relationships to one another. On the other hand, the PPA seems to represent geometric space beyond an object or multiple objects consistent with recent computational findings that suggest parallel processing of objects and spatial information (Fei-Fei & Perona, 2005; Lazebnik, Schmid, & Ponce, 2006; Oliva & Torralba, 2001).

Thus, both the PPA and LO contribute to scene processing: the PPA represents geometric aspects of space, and the LO represents multiple objects and their relationships. Although this suggests that the spatial and object properties of scenes are represented differently in distinctive brain regions, defining what constitutes an object

property or spatial property can be ambiguous. Objects sometimes define a scene's spatial layout—for example, a fountain can be categorized as an object or as a landmark. Size, permanence, and prior experience with a particular object modulate whether it is treated as an object or a scene. When objects gain navigational (landmark) significance, they may be treated as scenes and activate the PPA. Janzen and Van Turennout (2004) had subjects view a route through a virtual museum while target objects were placed at either an intersection of a route (decision point relevant for navigation) or at simple turns (nondecision point). High PPA activity was found for objects at intersections, which were critical for navigation, compared to objects at simple turns, which were equally familiar but did not have navigational value. This finding suggests that prior experience with an object in a navigationally relevant situation transforms these objects to relevant landmarks, which activates the PPA. Konkle and Oliva (2012) also showed that large objects such as houses activate the PPA region more than small objects.

One can also ask whether the PPA and RSC differentiate landmark properties. Auger, Mullally, and Maguire (2012) characterized individual landmarks by multiple properties such as size, visual salience, navigational utility, and permanence. They found that the RSC responded specifically to the landmarks that were consistently rated as permanent, whereas the PPA responded equally to all types of landmarks. In addition, they showed that poor navigators, compared to good navigators, were less reliable and less consistent in their ratings of a landmark's permanence. Thus, the primary function of the RSC may be processing the most stable or permanent feature of landmarks, which is critical for navigation. Altogether, the above studies suggest that object and scene representations are flexible and largely modulated by object properties or prior interactions with an object, especially when the objects may serve a navigational function, which we discuss further in the next section.

Integrating a View to a Scene

Once an immediate view is perceived and a viewer moves through the environment, the visual system must now confront the problem of integration. There are two seemingly contradictory computational problems that characterize this process. First, the visual system has to represent each individual view of a scene as unique in order to maintain a record of the viewer's precise position and heading direction. At the same time, however, the visual system must recognize that the current view is a part of a broader scene that extends beyond the narrow aperture of the current view. Constructing such an integrated representation of the environment guides navigation, action, and recognition from different views. How does the brain construct such stable percepts of the world? In this section, we discuss how the human visual system perceives an integrated visual world from multiple specific views that change over time.

For this purpose, we focus on two scene-specific areas in the brain, the PPA and the RSC. Both of these regions may be located by using a scene localizer, exhibiting strong preference to scenes over other visual stimuli. However neurological studies with patients suggest that the PPA and the RSC may play different roles in scene perception and navigation. Patients who have damage to the parahippocampal area cannot identify scenes such as streets or intersections and often rely on identification of small details in a scene such as street signs (Landis, Cummings, Benson, & Palmer, 1986; Mendez & Cherrier, 2003). However, these patients are able to draw a map or a route that they would take in order to navigate around these landmarks (Takahashi & Kawamura, 2002). Another patient with PPA damage showed difficulty learning the structure of new environments but had spared spatial knowledge of familiar environments (Epstein, DeYoe, Press, Rosen, & Kanwisher, 2001). This contrasts with patients with RSC damage, who were able to identify scenes or landmarks but had lost the ability to use these landmarks to orient themselves or to navigate through a larger environment (Aguirre & D'Esposito, 1999; Maguire, 2001; Valenstein et al., 1987). For example, when patients with RSC damage saw a picture of a distinctive landmark near their own home, they would recognize the landmark but could not use this landmark to find their way to their house. These neurological cases suggest that the parahippocampal and retrosplenial areas encode different kinds of scene representations: the parahippocampal area may represent physical details of the view of a scene, and the retrosplenial area may represent navigationally relevant properties such as the association of the current view to other views of the same scene in memory. These functional differences in the PPA and RSC may account for two different approaches taken to explain visual integration across views. The PPA, with higher sensitivity to perceptual details of a scene, may encode specific features of each view individually. On the other hand, the RSC, with its involvement in navigationally relevant analysis of a scene, may encode spatial regularities that are common across views, representing the scene in a view-invariant way.

Park and Chun (2009) directly tested viewpoint specificity and invariance across the PPA and RSC. When the same stimulus is repeated over time, the amount of neural activation for the repeated stimulus is significantly suppressed in comparison to the activity elicited when it was first shown. This robust phenomenon, called repetition suppression, may be used as a tool to measure whether a particular brain region represents two slightly different views of scenes as the same or different (see Grill-Spector, Henson, & Martin, 2006). Park and Chun (2009) presented three different views from a single panoramic scene to mimic the viewpoint change that may occur during natural scanning (for example, when you move your eyes from the left to the right corner of a room; figure 3.5, plate 6). If scene representations in the brain are view specific, then physically different views of the same room will be treated differently, so that no repetition suppression will be observed. Conversely, if scene

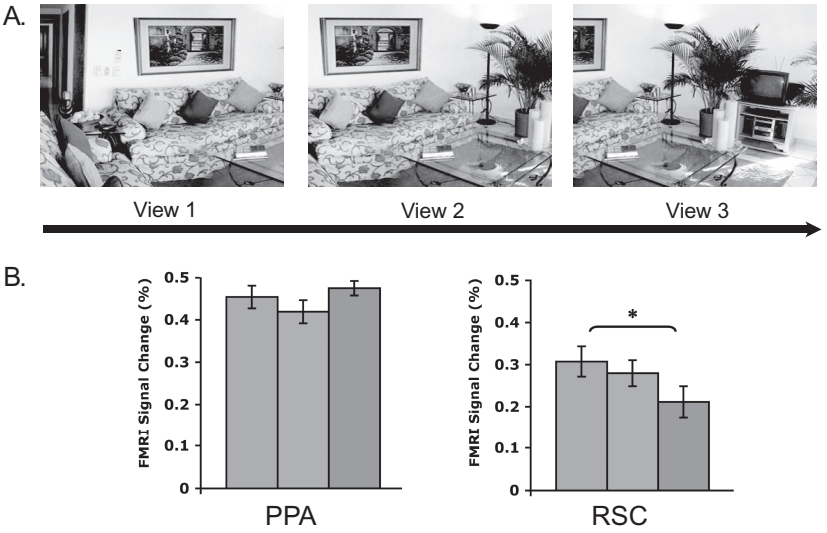

Figure 3.5 (plate 6)
(A) Example of panoramic first, second, and third images. These views were taken from a single panoramic scene. These panoramic scenes were presented in order at fixation. The PPA panoramic third image was taken from a single panoramic view. Panoramic first, second, and third images were sequentially presented one at a time at fixation. (B) Mean peak hemodynamic responses for panoramic first, second, and third in the PPA and RSC. The PPA showed no repetition suppression from the first to the third panoramic image, suggesting view specificity, whereas the RSC showed a significant repetition suppression, suggesting scene integration. Figure adapted from Park and Chun (2009).

representations in the brain are view invariant, then these views will be integrated into the representation of a single continuous room, yielding repetition suppression for different views from the same scene. The results revealed that the PPA exhibits view specificity, suggesting that this area focuses on selective discrimination of different views, whereas the RSC shows view-invariance, suggesting that RSC focuses on the integration of scenes under the same visual continuity. Viewpoint specificity in the PPA is supported by previous literature (Epstein, Graham, & Downing, 2003; Epstein & Higgins, 2007), and viewpoint integration in RSC fits with its characterization as an area that is important in navigation and route learning in humans and rodents (Burgess, Becker, King, & O'Keefe, 2001; Aguirre & D'Esposito, 1999; see also Vann et al., 2009 for review). This finding of two distinct but complementary regions in scene perception suggests that the brain develops ways to construct our perception with both specificity and stability from fragmented visual input. In addition, the experiment showed that spatiotemporal continuity across multiple views is critical to build an integrated scene representation RSC. When different views of panoramic scenes were presented with a long lag and intervening items, the RSC no longer showed patterns of neural attenuation consistent with scene integration. Thus, the

continuous percept of time and space across changing views provides important cues for building a coherent visual world.

Other researchers have also found that the PPA and RSC distinctively represent individual scenes as components of broader unseen spaces. Epstein, Parker, and Feiler (2007) tested whether a specific view of a scene (e.g., a view of school library) is represented neurally as part of a broader real-world environment beyond the viewer's current location (e.g., the whole campus). In their study they presented participants from the University of Pennsylvania community with views of familiar places around the campus or views from a different, unfamiliar campus. Participants judged either the location of the view (e.g., whether the view of a scene is on the west or east of a central artery road through campus) or its orientation (e.g., whether the view is facing west or east of the campus). The PPA responded equally to all conditions regardless of the task, but the RSC showed stronger activation to location judgments compared to orientation judgments. The location judgment required information about the viewer's current location as well as the location of the current scene within the larger environment. The RSC also showed much higher activity for familiar scenes than for unfamiliar scenes. Thus, the RSC is involved in the retrieval of specific location information of a view and how this view is situated relative to the surrounding familiar environment.

In a related vein, researchers found different levels of specificity and invariance across other scene selective areas including the transverse occipital sulcus (TOS). The TOS specifically responds to scenes compared to objects and often shows up along with the PPA and RSC in scene localizers. It is more posterior and lateral and is also often referred to as an occipital place area. Dilks et al. (2011) tested mirror-viewpoint change sensitivity in object- and scene-specific brain areas. When a scene image is mirror-reversed, the navigability of the depicted scene changes fundamentally as a path in the scene will reverse direction (e.g., a path originally going to the left now will become a path going to the right). Using repetition suppression they found that the RSC and the TOS were sensitive to mirror-reversals of scenes, treating two mirror-reversed scenes as different from each other. On the other hand, they found that the PPA was invariant to mirror-reversal manipulations, which challenges the idea that the PPA is involved in navigation and reorientation. Although these results seemingly contradict other findings showing viewpoint specificity in the PPA, they fit with the idea that the PPA represents the overall spatial layout of a given view, which is unchanged by mirror-reversal, as an image that has a closed spatial layout will remain as a closed scene; an open spatial layout will remain the same. What the mirror reversal changes is the functional navigability or affordance within a scene, such as in which direction the viewer should navigate. Thus, it makes sense that mirror-reversals did not affect the PPA, which represents visuospatial layout, but they affected the RSC, which represents the navigational properties of a scene. The function of the TOS is

still not well known, although recent research with transcranial magnetic stimulation (TMS) over the TOS suggests that the TOS is causally involved in scene perception (Dilks et al., 2013). Dilks et al. delivered TMS to the TOS and to the nearby face-selective occipital face area (OFA) while participants performed discrimination tasks involving either scenes or faces. Dilks et al. found a double dissociation, in that TMS to the TOS impaired discrimination of scenes but not faces, whereas TMS to the OFA impaired discrimination of faces but not scenes. This finding suggests that the TOS is causally involved in scene processing, although the precise involvement of TOS is still under investigation.

Another related question is whether scene representations in the PPA, RSC, and TOS are specific to retinal input. MacEvoy and Epstein (2007) presented scenes to either the left or right visual hemifields. Using repetition suppression, they tested if identical scenes repeated across different hemifields are treated as the same or differently in the brain. They found position invariance in the PPA, RSC, and TOS, suggesting that these scene-selective regions contain large-scale features of the scene that are insensitive to changes of retinal position. In addition, Ward et al. (2010) found that when stimuli are presented at different screen positions while fixation of the eyes is permitted to vary, the PPA and TOS respond equally to scenes that are presented at the same position relative to the point of fixation but not to scenes that are presented at the same position relative to the screen. This suggests an eye-centered frame of reference in these regions. In another study that controlled fixations within a scene, Golomb et al. (2011) showed that active eye movements by the viewer play an important role in scene integration. Stimuli similar to those depicted in figure 3.5A (plate 6) were used. The PPA showed repetition suppression to successive views when participants actively made saccades across a stationary scene (e.g., moving their eyes from left, middle, and right fixation points embedded in a scene) but not when the eyes remained fixed and a scene scrolled in the background across fixation, controlling for local retinal input between the two manipulations. These results suggest that active saccades may play an important role in scene integration, perhaps providing cues for retinotopic overlap across different views of the same scene.

So far in this chapter, we have focused on the parahippocampal and retrosplenial cortices. However, it is important to mention the role of the hippocampus in scene and space perception. Functional connectivity analysis suggests that parahippocampal and retrosplenial regions have strong functional connectivity with the hippocampus and other medial temporal regions such as the entorhinal and perirhinal cortices (Rauchs et al., 2008; Summerfield, Hassabis, & Maguire, 2010). A long history of rodent work has demonstrated hippocampal involvement in spatial cognition, such as maze learning and construction of a "cognitive map," a mental representation of one's spatial environment, in the hippocampus (Knierim & Hamilton, 2010; O'Keefe & Nadel, 1979). In particular, hippocampal neurons provide information about both the rat's external and internal coordinate systems. Place cells are one type of hippocampal

neuron that fires when a rat is at a specific location defined by an external coordinate system. Head-direction cells fire when the rat's head is oriented at a certain direction in the rat's internal coordinate system. The hippocampus also contains boundary cells, which respond to the rat's relative distance to an environmental boundary (Bird & Burgess, 2008; O'Keefe & Burgess, 1996). These findings suggest that the hippocampus is a critical region that represents the viewer's position in the external world. In addition, the division of labor described above for the PPA and RSC is interesting to think about in relation to computational models of hippocampal function in rats. Recent studies suggest that pattern separation, which amplifies differences in input, and pattern completion, which reconstructs stored patterns to match with current input, occur in different parts of the hippocampus: CA3/DG is involved in pattern separation, whereas CA1 is involved in pattern completion (see Yassa & Stark, 2011, for review). Even though it is difficult to make a direct comparison between hippocampal subregions and outer cortical regions such as the parahippocampal and retrosplenial regions, these complementary functions found in the rodent hippocampus seem to correspond to the complementary functions found in the PPA and RSC. For example, the PPA may rely on pattern separation to achieve view specificity, and the RSC may perform pattern completion to enable view integration.

Recently, fMRI studies have probed for cognitive map-like representations in human hippocampus. Morgan et al. (2011) scanned participants while viewing photographs of familiar campus landmarks. They measured the real-world (absolute) distance between pairs of landmarks and tested whether responses in the hippocampus, the PPA, and the RSC were modulated by the real-world distance between landmarks. They found a significantly attenuated response in the left hippocampus for a pair of landmarks that are closer in the real world compared to a pair of landmarks that are farther from one another in the real world. In contrast, the PPA and RSC encoded the landmark identity but not the real-world distance relationship between landmarks (Morgan et al., 2011). These results suggest that the hippocampus encodes landmarks in a map-like representation, reflecting relative location and distance between landmarks. Another study using multivoxel pattern analysis with high-spatial-resolution fMRI found that the position of an individual within an environment was predictable based on the patterns of multivoxel activity in the hippocampus (Hassabis et al., 2009). In this experiment participants navigated in an artificially created room that had four different corners (corners A–D). In each trial the participants navigated to an instructed target position (e.g., go to the corner A). When they reached the corner, they pressed a button to adopt a viewpoint looking down, which revealed a rug on the floor. This rug view visually looked the same across four corners; thus, the multivoxel activity collected during this period was based on the viewer's position in a room and not on any visual differences between the four corners. Multivoxel patterns in the hippocampus enabled classification of which of the four corners the participant was positioned. These results are similar to the rat's place cells, which fire in response to

the specific location of the rat within an environment. With high-resolution fMRI techniques and computational models that enable segmentation of hippocampal subregions, future research should be aimed at identifying whether the human hippocampus, like that of the rat, also contains both external and internal coordinate systems facilitated by place cells and head direction cells as well as boundary cells that encode the viewer's distance to environmental boundaries.

Representing Scenes in Memory

We sample the world through a narrow aperture that is further constrained by limited peripheral acuity, but we can easily extrapolate beyond this confined window to perceive a continuous world. The previous section reviewed evidence that coherent scene perception is constructed by integrating multiple continuous views. Such construction can occur online but can also occur as we represent scene information in memory that is no longer in view. In this section we discuss how single views are remembered and how the visual system constructs information beyond the current view. Traditionally, the constructive nature of vision has been tested in low-level physiological studies, such as filling in of the retinal blind spot or contour completion. However, less is known about what type of transformations or computations are performed in higher-level scene-processing regions. Yet, expectations about the visual world beyond the aperture-like input can systematically distort visual perception and memory of scenes. Specifically, when people are asked to reconstruct a scene from memory, they often include additional information beyond the initial boundaries of the scene, in a phenomenon called *boundary extension* (Intraub, 1997, 2002; also see chapter 1 by Intraub in this volume). The boundary-extension effect is robust across various testing conditions and various populations, such as recognition, free recall, or directly adjusting borders of the boundary both visually and haptically (Intraub, 2004, 2012). Boundary extension occurs in children and infants as well (Candel, Merckelbach, Houben, & Vandyck, 2004; Quinn & Intraub, 2007; Seamon, Schlegel, Hiester, Landau, & Blumenthal, 2002). Interestingly, boundary extension occurs for scenes with background information but not for scenes comprising cutout objects on a black screen (Gottesman & Intraub, 2002). This systemic boundary extension error suggests that our visual system is constantly extrapolating the boundary of a view beyond the original sensory input (figure 3.6).

Boundary extension is a memory illusion, but this phenomenon has adaptive value in our everyday visual experience. It provides an anticipatory representation of the upcoming layout that may be fundamental to the integration of successive views. Using boundary extension, we can test whether a scene is represented in the brain as it is presented in the physical input or as an extended view that observers spatially extrapolated in memory. Are there neural processes in memory that signal the spatial

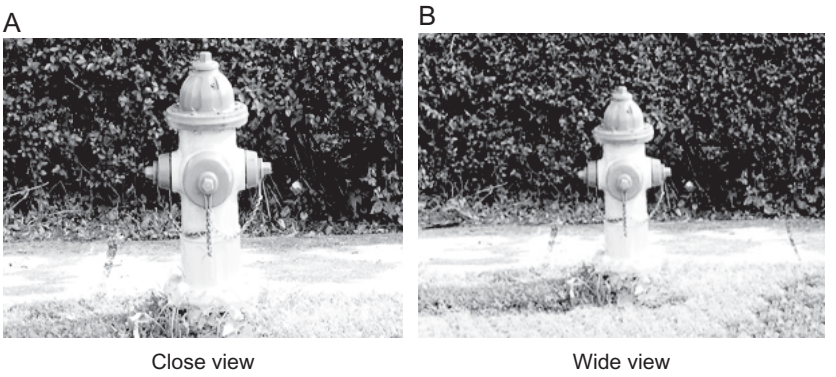

Figure 3.6
Example of boundary extension. After viewing a close-up view of a scene (A), observers tend to report an extended representation (B).

extrapolation of physically absent but mentally represented regions of a scene? If so, this would demonstrate that higher-level scene-processing areas such as the PPA and RSC facilitate the perception of a broader continuous world through the construction of visual scene information beyond the limits of the aperture-like input.

Park et al. (2007) tested for such effects of boundary extension. They used fMRI repetition suppression for close views and wide views of scenes to reveal which scene pairs were treated as similar in scene regions of the brain. When the original view of a scene is a close-up view, boundary extension predicts that this scene will be extended in memory and represented as a wider view than the original. Thus, if the same scene is presented with a slightly wider view than the original, this should match the boundary-extended scene representation in scene-selective areas and should result in repetition suppression. On the other hand, if a wide view of a scene is presented first, followed by a close view (wide-close condition), there should be no repetition suppression even though the perceptual similarity between close-wide and wide-close repetitions is identical. This asymmetry in neural suppression for close-wide and wide-close repetition was exactly what Park et al. (2007) observed (figure 3.7). Scene-processing regions such as the PPA and RSC showed boundary extension in the form of repetition suppression for close-wide scene pairs but not for wide-close scene pairs. In contrast, there were no such asymmetries in the LOC. This reveals that the brain's scene-processing regions reflect a distorted memory representation, and such boundary extension is specific to background scene information and not to foreground objects. Such extended scene representations may reflect an adaptive mechanism that allows the visual system to perceive a broader world beyond the sensory input.

Another fMRI study on boundary extension points to further involvement of the hippocampus (Chadwick, Mullally, & Maguire, 2013). Online behavioral boundary

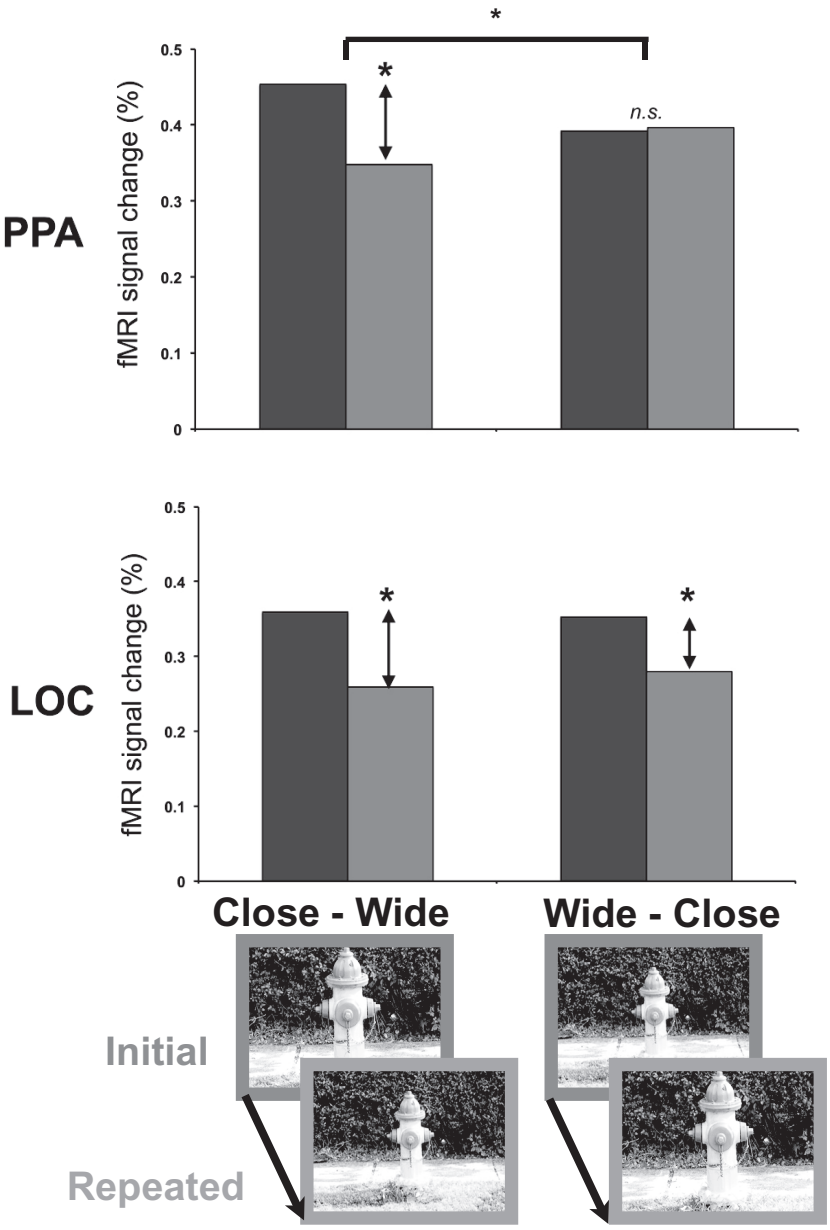

Figure 3.7
Peaks of the hemodynamic responses for close-wide and wide-close conditions are shown for the PPA and LOC. Examples of close-wide and wide-close condition are presented at the bottom. An interaction between the close-wide and wide-close condition activation, representing boundary extension asymmetry, was observed in the PPA but not in the LOC. Figure adapted from Park et al. (2007).

effects were measured for individual scenes as participants viewed scenes in the scanner. When scenes that showed the boundary extension effect were compared to scenes that did not show it, there was a significant difference in activity in the hippocampus. Moreover, functional connectivity analysis showed that scenes with boundary extension had high connectivity between the hippocampus and the parahippocampal cortex, whereas scenes without boundary extension effect did not. A neurological study with patients who have hippocampal damage also found that these patients had less or no boundary extension compared to a control group (Mullally, Intraub, & Maguire, 2012). For example, when the same close view of a scene was repeated following a close view, normal controls would respond that the second close view was different from the original, showing the usual boundary extension distortion. However, patients with hippocampal damage were more accurate at rating the second close view as identical to the original, showing no distortion from boundary extension. These are intriguing results because patients with hippocampal damage actually showed more accurate scene memory than controls, immune from the boundary extension error. These results suggest that the hippocampus may play a central role in boundary extension, and hence, the boundary extension effect found in the parahippocampal cortex in Park et al. (2007) may reflect such feedback input from the hippocampus. Because boundary extension is an example of constructive representations of scenes in memory, these results further support the role of the hippocampus in the anticipation and construction of memory (Addis, Wong, & Schacter, 2007; Buckner & Carroll, 2007; Turk-Browne, Scholl, Johnson, & Chun, 2010). In addition, the boundary extension distortion should not simply be viewed as a memory error but rather as a successful adaptive mechanism that enables anticipation of a broader perceptual world from limited input.

Because the amount of visual information a human can see at one time is limited, we have also evolved mechanisms for bringing to mind recent visual information that is no longer present in the current environment. Such acts are called *refreshing* and occur when one briefly thinks back to a stimulus one just saw. The act of refreshing may facilitate scene integration by foregrounding the information to facilitate the binding of the previous and the current views. Given the potential role of refreshing in scene integration, Park, Chun, and Johnson (2010) asked whether discrete views of scenes are integrated during refreshing of these views. Similar to results found with physical scenes, when participants refreshed different views of scenes, the PPA showed view-specific representations, and the RSC showed view-invariant representations. Research directly comparing cortical activity for perception and refreshing showed that activity observed in the RSC and precuneus for refreshing closely mirrored the activity for perceiving in these regions (Johnson, Mitchell, Raye, D'Esposito, & Johnson, 2007). Thus, the act of refreshing in these high-level regions might play an important role during perceptual integration by mirroring the activity of perceiving panoramic views of scenes in continuation.

Conclusion

Constructing a rich and coherent percept of our surroundings is essential for navigating and interacting with our environment. How do we recognize scenes? In this chapter we reviewed key studies in cognitive neuroscience that investigated how we construct a meaningful scene representation at the structural, perceptual, and memory levels. Multiple brain regions play distinctive functions in representing different properties of scenes, and the PPA, RSC, and LOC areas represent a continuum of specialized processing for spatial properties, from navigational features (RSC and PPA) to that of diagnostic objects (LOC).

However, even the useful distinction between spatial and object representation may be an oversimplification. Real-world objects or scenes have enormous complexity and vary along an exceptionally high number of dimensions such as layout, texture, color, depth, density, and so on. The next major goal in the field will be to understand the precise neural processing mechanisms in the PPA, RSC, and LOC areas. An essential first step is to identify the dimensions in which neurons encode scene information. The enormous, megapixel dimensionality of a visual scene must be reduced by the ventral pathway to a tractable set of dimensions for encoding scene information. These coding dimensions must be flexible enough to support robust categorization but also sensitive to parametric variations necessary for discriminating different exemplars and specific views.

Scene research lags behind that for faces and objects. Electrophysiological recordings reveal neurons in the face-selective cortex that encode specific and parametric components of face parts, geometry, and configuration (Freiwald & Tsao, 2010; Freiwald, Tsao, & Livingstone, 2009) and object-selective regions that encode parametric dimensions of 2D contour, 3D surface orientation, and curvature of objects (Hung, Carlson, & Connor, 2012). Research in human IT cortex has also begun to show not only categorical but continuous representations of individual objects (Kriegeskorte et al., 2008). Similarly detailed representations at the neuronal level have yet to be discovered for scene perception. Hence, one of the next goals in the field of scene perception should be to test precise coding dimensions of scene-selective neurons reflecting continuous and parametric changes in the coding dimension (e.g., varying degrees of the size of space; varying degrees of openness in spatial layout). For example, do the PPA and RSC discriminate the size of space independent of the clutter or density of objects within a given space? Estimating the size of space and the level of clutter in a scene is central to our interactions with scenes—for example, when deciding whether or not to take a crowded elevator or when driving through downtown traffic. Park et al. (2011) varied both the size of space and levels of object clutter depicted within scenes and discovered that the anterior portions of the PPA and RSC responded parametrically to different sizes of space in a way that generalized across scene categories.

Another major goal in the field of scene understanding is to describe how information from multiple scene-selective regions, representing different properties of scenes, is synthesized. Scene categorization and scene gist recognition are so rapid and efficient, some of this information must be combined at very early stages of visual processing. How and where are these properties weighted and combined to enable scene categorization that rapidly occurs within 200 ms? Future research should aim to reveal the interaction across the family of scene-selective regions within the rapid time course of scene recognition.

How do we represent a coherent scene from constantly changing visual input? Converging evidence throughout this chapter suggests that the brain overcomes multiple constraints of our visual input by constructing an anticipatory representation beyond the frame of the current view. The visual system assumes what may exist just beyond the boundaries of a scene or what may exist when our eyes are successively moved to the next visual frame. Such assumptions are represented in high-level visual areas and produce a rich and coherent perceptual experience of the world. Is there a functional architecture in the brain that enables such extrapolation of scene information? The PPA represents scenes not just based on the current visual stimulus but within the temporal context in which these scenes were presented (Turk-Browne, Simon, & Sederberg, 2012). A scene that was embedded in a predictable temporal context had greater PPA repetition suppression than a scene that did not have any predictable temporal context preceding it. These results show that the PPA not only represents the present input but integrates predictable contexts created from the past. Such predictive coding found in high-level visual cortex may support navigation by integrating past and present input. Related to this, an important future direction for scene perception research would be to show how real-world scene perception unfolds over time. In many real-world circumstances, information at the current moment becomes meaningful only in the context of a past event. For example, if you present frames of a movie trailer in a randomized order, the whole trailer will be incomprehensible. Indeed, our brain is sensitive to sequences of visual information across different time scales (Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Honey et al., 2012). When a meaningful visual event is presented over time (e.g., a movie clip), early visual areas such as V1 are involved in frame-by-frame analysis of single snapshots; midlevel areas such as the FFA and the PPA are involved in integration over a short time scale (e.g., a few seconds); and higher-order areas such as the temporoparietal junction (TPJ) are involved in integration and reasoning of an event over a longer time scale, creating a hierarchy of temporal receptive fields in the brain. Although these studies tested higher-level understanding of the meaning of complex event sequences, one can imagine a similar hierarchy of temporal receptive fields in daily navigation. For example, recognizing that the current view is continuous from the previous view (e.g., integrating panoramic views over time) might require integration over a short time

scale, whereas recognizing where you are in a city may require integration of the route you took over a longer time scale. More research on scene and spatial navigation should integrate how our brain combines scene information presented over different temporal contexts and scales.

Altogether, the rich and meaningful visual experience that we take for granted relies on the brain's elegant functional architecture of multiple brain regions with complementary functions for scene perception. Research in the field of scene understanding has grown rapidly over the past few years, and the field has just begun to distinguish which structural and conceptual properties of visual scenes are processed at different stages of the visual pathway. In combination with fMRI multivoxel pattern analysis and computational models of low-level visual systems, we are at the stage of being able to roughly reconstruct what the viewer is currently seeing (akin to "mind reading") (Kay, Naselaris, Prenger, & Gallant, 2008). The next major goal in the field is to discover the precise dimensions of scenes that are encoded in multiple scene-selective regions, to figure out how these dimensions are synthesized to give rise to the perception of a complete scene, and to understand how these representations change or integrate over time as the viewer navigates in the world.

References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*(7), 1363–1377.

Aguirre, G. K., & D'Esposito, M. (1999). Topographical disorientation: A synthesis and taxonomy. *Brain*, *122*, 1613–1628.

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: Evidence and implications. *Neuron*, *21*, 373–383.

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, *17*(8), 379–390.

Auger, S. D., Mullally, S. L., & Maguire, E. A. (2012). Retrosplenial cortex codes for permanent landmarks. *PLoS ONE*, *7*(8), e43620.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–253). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bird, C. M., & Burgess, N. (2008). The hippocampus and memory: Insights from spatial processing. *Nature Reviews Neuroscience*, *9*, 182–194.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57.

Burgess, N., Becker, S., King, J. A., & O'Keefe, J. (2001). Memory for events and their spatial context: Models and experiments. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *356*, 1493–1503.

Candel, I., Merckelbach, H., Houben, K., & Vandyck, I. (2004). How children remember neutral and emotional pictures: Boundary extension in children's scene memories. *American Journal of Psychology*, *117*, 249–257.

Chadwick, M. J., Mullally, S. L., & Maguire, E. A. (2013). The hippocampus extrapolates beyond the view in scenes: An fMRI study of boundary extension. *Cortex*, *49*(8), 2067–2079.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*, 261–270.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.

Dilks, D., Julian, J. B., Kubilius, J., Spelke, E. S., & Kanwisher, N. (2011). Mirror-image sensitivity and invariance in object and scene processing pathways. *Journal of Neuroscience*, *33*(31), 11305–11312.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area (OPA) is causally and selectively involved in scene perception. *Journal of Neuroscience*, *33*(4), 1331–1336.

Eger, E., Ashburner, J., Haynes, J., Dolan, R. J., & Rees, G. (2008). fMRI activity patterns in human LOC carry information about object exemplars within category. *Journal of Cognitive Neuroscience*, *20*, 356–370.

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, *12*(10), 388–396.

Epstein, R., DeYoe, E. A., Press, D. Z., Rosen, A. C., & Kanwisher, N. (2001). Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cognitive Neuropsychology*, *18*(6), 481–508.

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint specific scene representations in human parahippocampal cortex. *Neuron*, *37*, 865–876.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, *23*(1), 115–125.

Epstein, R. A., & Higgins, J. S. (2007). Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cerebral Cortex*, *17*(7), 1680–1693.

Epstein, R. A., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Epstein, R. A., Parker, W. E., & Feiler, A. M. (2007). Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *Journal of Neuroscience*, *27*(23), 6141–6149.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *Computer Vision and Pattern Recognition*, *2*, 524–531.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*(9), 1187–1196.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology. General*, *108*, 316–355.

Goh, J. O. S., Siong, S. C., Park, D., Gutchess, A., Hebrank, A., & Chee, M. W. L. (2004). Cortical areas involved in object, background and object-background processing revealed with functional magnetic resonance adaptation. *Journal of Neuroscience*, *24*(45), 10223–10228.

Golomb, J. D., Albrecht, A., Park, S., & Chun, M. M. (2011). Eye movements help link different views in scene-selective cortex. *Cerebral Cortex*, *21*(9), 2094–2102.

Gottesman, C. V., & Intraub, H. (2002). Surface construal and the mental representation of scenes. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(3), 589–599.

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology. Human Perception and Performance*, *32*, 1107–1119.

Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472.

Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron*, *21*, 191–202.

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex*, *23*(4), 947–957.

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P. D., & Maguire, E. A. (2009). Decoding neural ensembles in the human hippocampus. *Current Biology*, *19*, 546–554.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*, 2539–2550.

Hoiem, D. H., Efros, A. A., & Hebert, M. (2006). Putting objects in perspective. *International Journal of Computer Vision*, *80*, 3–15.

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., et al. (2012). Slow dynamics in human cerebral cortex and the accumulation of information over long timescales. *Neuron*, *76*(2), 423–434.

Hung, C.-C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*(6), 1099–1113.

Ino, T., Inoue, Y., Kage, M., Hirose, S., Kimura, T., & Fukuyama, H. (2002). Mental navigation in humans is processed in the anterior bank of the parieto-occipital sulcus. *Neuroscience Letters*, *322*, 182–186.

Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences*, *1*(6), 217–222.

Intraub, H. (2002). Anticipatory spatial representation of natural scenes: Momentum without movement? *Visual Cognition*, *9*, 93–119.

Intraub, H. (2004). Anticipatory spatial representation of 3D regions explored by sighted observers and a deaf-and-blind observer. *Cognition*, *94*(1), 19–37.

Intraub, H. (2012). Rethinking visual scene perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(1), 117–127.

Janzen, G., & Van Turennout, M. (2004). Selective neural representation of objects relevant for navigation. *Nature Neuroscience*, *7*(6), 673–677.

Johnson, M. R., Mitchell, K. J., Raye, C. L., D'Esposito, M., & Johnson, M. K. (2007). A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. *NeuroImage*, *37*, 290–299.

Joubert, O. R., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*, 3286–3297.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? *Cerebral Cortex*, *21*, 1738–1746.

Knierim, J. J., & Hamilton, D. A. (2010). Framing spatial cognition: Neural representations of proximal and distal frames of reference and their roles in navigation. *Physiological Reviews*, *91*, 1245–1279.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipito-temporal cortex. *Neuron*, *74*(6), 1114–1124.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *Journal of Neuroscience*, *20*(9), 3310–3318.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, *31*(20), 7322–7333.

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Kumaran, D., & Maguire, E. A. (2006). An unexpected sequence of events: Mismatch detection in the human hippocampus. *PLoS Biology*, *4*(12), e424.

Landis, T., Cummings, J. L., Benson, D. F., & Palmer, E. P. (1986). Loss of topographic familiarity. An environmental agnosia. *Archives of Neurology*, *43*, 132–136.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Computer Vision and Pattern Recognition*, *2*, 2169–2178.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology. Human Perception and Performance*, *4*(4), 565–572.

MacEvoy, S. P., & Epstein, R. A. (2007). Position selectivity in scene and object responsive occipitotemporal regions. *Journal of Neurophysiology*, *98*, 2089–2098.

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329.

Maguire, E. A. (2001). The retrosplenial contribution to human navigation: A review of lesion and neuroimaging findings. *Scandinavian Journal of Psychology*, *42*, 225–238.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., et al. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 8135–8139.

Mendez, M. F., & Cherrier, M. M. (2003). Agnosia for scenes in topographagnosia. *Neuropsychologia*, *41*, 1387–1395.

Morgan, L. K., MacEvoy, S. P., Aguirre, G. K., & Epstein, R. A. (2011). Distances between real-world locations are represented in the human hippocampus. *Journal of Neuroscience*, *31*(4), 1238–1245.

Mullally, S. L., Intraub, H., & Maguire, E. A. (2012). Attenuated boundary extension produces a paradoxical memory advantage in amnesic patients. *Current Biology*, *22*(4), 261–268.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*, 425–428.

O'Keefe, J., & Nadel, L. (1979). The hippocampus as a cognitive map. *Behavioral and Brain Sciences*, *2*, 487–494.

Oliva, A., Park, S., & Konkle, T. (2011). Representing, perceiving and remembering the shape of visual space. In L. R. Harris & M. Jenkin (Eds.), *Vision in 3D Environments* (pp. 107–134). Cambridge: Cambridge University Press.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research: Visual Perception*, *155*, 23–36.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*, 519–526.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the PPA and LOC in representing real-world scenes. *Journal of Neuroscience*, *31*(4), 1333–1340.

Park, S., & Chun, M. M. (2009). Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in scene. *NeuroImage*, *47*(4), 1747–1756.

Park, S., Chun, M. M., & Johnson, M. K. (2010). Refreshing and integrating visual scenes in scene-selective cortex. *Journal of Cognitive Neuroscience*, *22*(12), 2813–2822.

Park, S., Intraub, H., Yi, D. J., Widders, D., & Chun, M. M. (2007). Beyond the edges of a view: Boundary extension in human scene-selective visual cortex. *Neuron*, *54*(2), 335–342.

Park, S., Konkle, T., & Oliva, A. (2014). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral Cortex*, doi: 10.1093/cercor/bht418.

Potter, M. C. (1975). Meaning in visual scenes. *Science*, *187*, 965–966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory*, *2*(5), 509–522.

Potter, M. C., Staub, A., & O'Connor, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. *Journal of Experimental Psychology. Human Perception and Performance*, *30*, 478–489.

Quinn, P. C., & Intraub, H. (2007). Perceiving "outside the box" occurs early in development: Evidence for boundary extension in three- to seven-month-old infants. *Child Development*, *78*(1), 324–334.

Rauchs, G., Orban, P., Balteau, E., Schmidt, C., Degueldre, C., Luxen, A., et al. (2008). Partially segregated neural networks for spatial and contextual memory in virtual navigation. *Hippocampus*, *18*(5), 503–518.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hilldale, NJ: Lawrence Erlbaum Associates.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Seamon, J. G., Schlegel, S. E., Hiester, P. M., Landau, S. M., & Blumenthal, B. F. (2002). Misremembering pictured objects: People of all ages demonstrate the boundary extension illusion. *American Journal of Psychology*, *115*(2), 151–167.

Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2010). Differential engagement of brain regions within a "core" network during scene construction. *Neuropsychologia*, *48*(5), 1501–1509.

Takahashi, N., & Kawamura, M. (2002). Pure topographical disorientation—the anatomical basis of landmark agnosia. *Cortex*, *38*, 717–725.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network (Bristol, England)*, *14*(3), 391–412.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of eye movements in real-world scenes: the role of global features on object search. *Psychological Review*, *113*(4), 766–786.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*, 11177–11187.

Turk-Browne, N. B., Simon, M. G., & Sederberg, P. B. (2012). Scene representations in parahippocampal cortex depend on temporal context. *Journal of Neuroscience*, *32*, 7202–7207.

Valenstein, E., Vowers, D., Verfaellie, M., Heilman, K. M., Day, A., & Watson, R. T. (1987). Retrosplenial amnesia. *Brain*, *110*, 1631–1646.

Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, *10*(11), 792–802.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Vinberg, J., & Grill-Spector, K. (2008). Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex. *Journal of Neurophysiology*, *99*(3), 1380–1393.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience*, *29*(34), 10573–10581.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(23), 9661–9666.

Ward, E. J., MacEvoy, S. P., & Epstein, R. A. (2010). Eye-centered encoding of visual space in scene-selective regions. *Journal of Vision, 10*(14), 1–12.

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525.