51 Scene Perception

AUDE OLIVA

Visual scene perception is the gateway to many of our most valued behaviors, including navigation, recognition, and reasoning with the world around us. What is a "visual scene"? What are its properties? Is scene perception different from object perception? Operationally, a visual scene can be defined as a view in which objects and surfaces are arranged in a meaningful way, for example a kitchen, a street, or a forest. Scenes contain elements arranged in a *spatial layout* and can be viewed at a variety of spatial scales (e.g., the up-close view of an office desk or the view of the entire office). As a rough distinction, one generally takes action *on* an object, whereas one usually acts *within* a scene.

One paradoxical feature of visual scene analysis is that the complex arrangement of objects and surfaces in the world creates the impression that there is too much to see at once. How can so much visual information be processed and understood in a timely manner? Remarkably, we are able to interpret the meaning of multifaceted and complex scene images—a wedding, a birthday party, or a stadium crowd—in a fraction of a second (Potter, 1975)! This is about the same time it takes a person to identify that a single object is a face, a dog, or a car (Grill-Spector & Kanwisher, 2005; Intraub, 1981; Thorpe, Fize, & Marlot, 1996). An unmistakable demonstration of the brain's prowess in visual scene understanding can be experienced at the movies: With a few rapid scene cuts from a movie to form a trailer, it seems as if we have perceived and understood much more of the story in a few seconds than could be described later in the same amount of time. Perceiving scenes in a glance is like looking at an abstract painting of a landscape and recognizing that a "forest" is depicted before seeing the "trees" that create it (Navon, 1977).

This chapter reviews research in the behavioral, computational, and cognitive neuroscience domains that describe how the human visual system analyses real-world scenes. Although we typically experience scenes in a three-dimensional physical world, most studies are conducted using two-dimensional pictures. There are likely important differences between perceiving the world and perceiving visual scenes via pictures, and this chapter describes principles that are likely to apply to both mediums (for a review, see Cutting, 2003).

BEHAVIORAL STUDIES OF SCENE PERCEPTION

Historical Perspective

Pioneering work done by Mary Potter (1975), David Navon (1977), and Irving Biederman (1981) has shown that the overall scene meaning is invariably captured within a glance regardless of the complexity of the image. These landmark studies have laid down much of the empirical and theoretical foundation for modern inquiries into how the human brain perceives complex, real-world scenes.

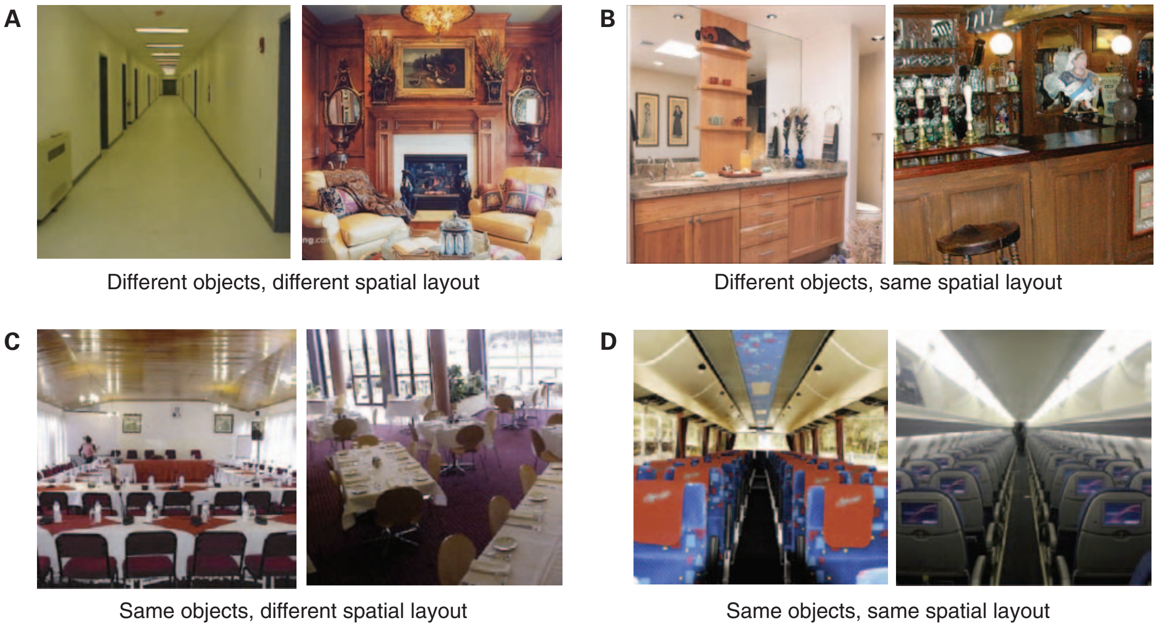
Figure 51.1 illustrates what scenes are made of: surfaces of different materials laid out within a three-dimensional physical space. Some surfaces define the boundaries of the space (e.g., walls, tall objects), and other surfaces are created by objects of specific identities and functions (e.g., a dining room table). Importantly, scene perception is not merely the sum of its parts; it is paramount to consider the scene as a whole. In figure 51.1, for example, scenes can be similar in layout (B, D) or contain similar objects (see C, D) yet belong to different semantic categories. Scene analysis involves perceiving the type of surfaces, objects, their placement, and their quantity.

In their original study Potter and Levy (1969) allowed observers a single brief glance at a series of real-world images and subsequently tested their memory of these images. They observed that understanding happens fast, very fast: When presented alone for 100 ms, each image was easily remembered and described. Together with other experimental evidence, these results showed that most of the information from an image could be captured in tenths of a second. So what type of representation could possibly be built so fast?

Theoretical Perspectives

Biederman (1981) proposed three levels of representations observers could build quickly, either sequentially or in parallel, to reach a rich description of the scene by the end of a glance. According to this framework, scene understanding may follow from (1) a prominent object or surface, (2) a multiple-component

K2



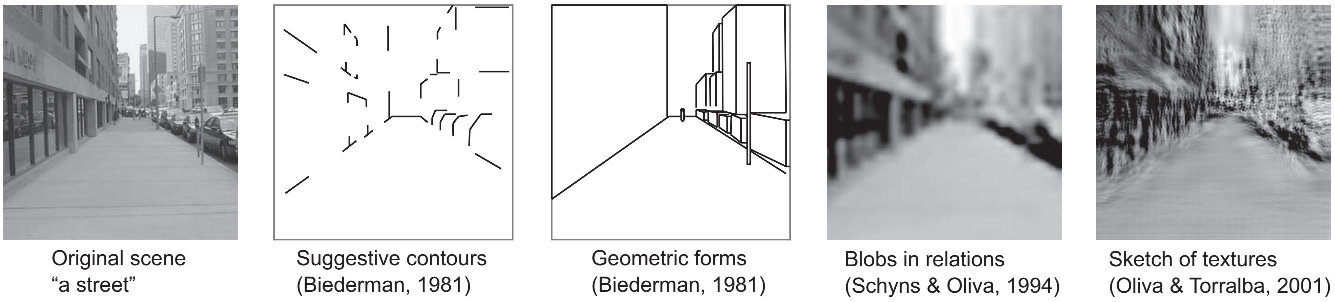
Different objects, different spatial layout

Different objects, same spatial layout

Same objects, different spatial layout

Same objects, same spatial layout

Figure 51.1 Each image pair contains scenes of different semantic categories. Yet notice how the scenes can differ in spatial layout (A, C) or object content (A, B) or have similar layouts (B, D) and similar objects (C, D).



Original scene
“a street”

Suggestive contours
(Biederman, 1981)

Geometric forms
(Biederman, 1981)

Blobs in relations
(Schyns & Oliva, 1994)

Sketch of textures
(Oliva & Torralba, 2001)

Figure 51.2 Illustrations of global scene emergent features. These representations have formed the basis of object-free computational models of scene perception.

representation incorporating a layout of distinct objects or surfaces, and (3) a more global representation of scene-emergent features, which does not necessarily depend on object and surface identification. The *prominent object* representation capitalizes on a region of the scene (Friedman, 1979)—often, a large and diagnostic object (a bed in a bedroom, a sofa in a living room, a large region of grass in a field)—to quickly activate contextual information from stored knowledge. The *multiple-component* representation is based on objects and regions segmented from the background and organized in a coherent layout. Finally, the *global scene-emergent features* representation is built from grasping components from all over the available percept that do not necessarily correspond to segmented objects or meaningful regions. Figure 51.2 illustrates some of the forms these emergent global scene features haven taken in various works.

Inspired by this early proposal studies in behavioral, computational, and cognitive neuroscience of the past decade have converged to describe two complementary paths of scene perception: an *object-centered approach* in which components are segmented and function as the scene descriptors (i.e., this is a street because there are buildings and cars); and a *scene or space-centered approach* in which spatial layout and global properties of the whole image or place act as the scene descriptors (i.e., this is a street because it is an outdoor, urban environment flanked with tall frontal vertical surfaces with squared patterned textures). How do these different levels of scene information unfold over the course of a glance?

Time Course of Scene Analysis

Although a complete picture of the time course of the components contributing to scene perception has yet to emerge, most experimental work distinguishes between early stages (before 100 ms) and late stages (from 200 to 300 ms, before the observer moves her eye) of scene analysis. When an image is briefly presented, there is a temporal progression to how an observer perceives a scene's content: As image exposure increases, observers are better able to fully perceive the details of an image such as high spatial frequencies (Schyns & Oliva, 1994), texture (Walker-Renninger & Malik, 2004), or object identities (Fei-Fei et al., 2007; Rayner et al., 2009). This is known as global-to-local (Navon, 1977) or coarse-to-fine (Schyns & Oliva, 1994) scene analysis. For instance, in Fei-Fei et al. (2007), observers were presented with briefly masked pictures depicting various events and scenery (e.g., a soccer game, a busy hair salon, a choir, a dog playing fetch) and later asked to describe in detail what they saw in the picture. The authors found that observers perceived global scene information, such as whether the picture was outdoor or indoor, well above chance with less than 100 ms of exposure, whereas details about objects were reported with a couple of hundred milliseconds more exposure time. Similarly, Greene and Oliva (2009a) found that global scene properties (i.e., the volume, openness, naturalness of a scene) explain early scene categorization better than representation of a scene solely by its objects (i.e., trees, grass, rock). In fact, an exposure of only 20 to 30 ms is sufficient to know whether the scene depicts a natural or an urban place (Greene & Oliva, 2009b; Joubert et al., 2007) or whether the scene has a small or large volume (e.g., cave vs. lake). Yet it takes twice that much time to determine the basic level category of the scene (Greene & Oliva, 2009b), for example mountain versus beach. It follows that, during early visual processing, there is a time point at which a scene may be classified as a large or navigable landscape but not yet as a mountain or lake.

The Building Blocks of Scene Perception

One important question in visual analysis is the role of temporal history in perception: How does what we have recently perceived influence what is presently perceived? This refers to the notion of *adaptation*: When observers are overexposed to certain visual features, the adaptation of those features affects the conscious perception of a subsequently presented stimulus. Which representations of scene information, if any, adapt? Our daily life experiences suggest that spatial layout characteristics might be susceptible to aftereffects. For example, after spending all day working in a small office or cubicle, the first sight of the expansive world right outside of the office building might appear much larger than it did on entry.

Using an adaptation paradigm Greene and Oliva (2010) found that prolonged exposure to scenes with shared global properties can influence how observers perceive that property in later scenes, and, interestingly, these aftereffects even influence semantic categorization of the scene. In their experiment observers adapted to a stream of open and panoramic views of natural images (openness) or to a stream of pictures depicting scenes enclosed with frontal and lateral surfaces (closeness). When ambiguous probe images (e.g., a field with trees) were then tested, they were more likely to be judged as a *closed* space if the observer had adapted to openness or as an *open* space if adapted to closeness. Similar aftereffects are found after adapting to natural versus urban spaces (see also Kaping, Tzvetanov, & Treue, 2007). Furthermore, adapting to an open or closed spatial layout even shifted the categorical recognition of a briefly perceived scene. This design took advantage of the fact that fields are usually open scenes, whereas forests are typically closed scenes, but importantly there is a continuum between field and forest scenes, with some scenes existing ambiguously between the two categories that can be perceived both as a field or a forest (see figure 51.3). Indeed, Greene and Oliva (2010) found that adapting to a stream of closed natural

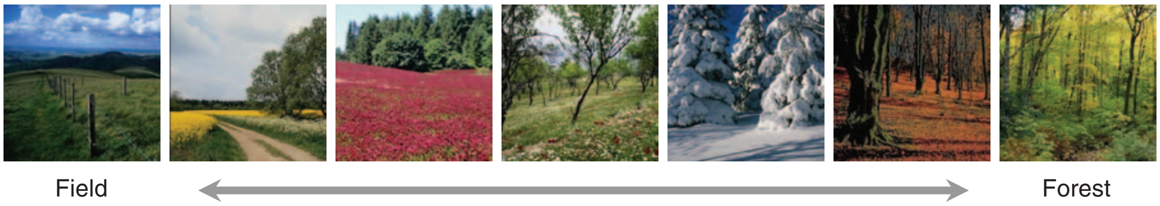


FIGURE 51.3 Continuum between fields and forests. Scenes in the middle of continuum have an ambiguous category and can be perceived as either a field or a forest, after adapting to closed versus open scenes, respectively. (From Greene & Oliva, 2010.)

images caused an ambiguous field/forest to be judged as more likely to be a field. In summary, the observation of aftereffects for global scene properties implies that, during natural scene processing, the visual system extracts statistics and becomes attuned to recently processed global properties such as scene layout and volume.

Discovering what scene properties become available over the course of a glance and learning what representational role global properties have in scene recognition provide critical insights into the possible computations underlying scene perception.

COMPUTATIONAL FRAMEWORK OF SCENE PERCEPTION

Just as external shape and internal features are separable dimensions of face encoding, Oliva and Torralba (2001, 2002) proposed a framework in which a scene, whether a physical space or its projection onto a two-dimensional image, can be represented by two separable and complementary descriptors: its *spatial boundary* (i.e., the external shape, size, and scope of the space the scene represents) and its *content* (the internal elements, encompassing textures, colors, materials, and objects). As illustrated in figure 51.4, the shape of an outdoor scene may be expansive and open to the horizon, as in field and parking lot, or closed and bounded by frontal and lateral surfaces, as in forests and streets. Importantly, the spatial boundary is independent of the scene's content, which may contain natural or manufactured elements. This framework is general enough so that it does not make assumptions regarding the type of features used to represent the scene. For instance, one can identify a scene as a landscape by using colors, textures, materials, or objects. The size or scope of a scene can be captured by geometrical relations between its boundaries or by low-level image features that are correlated with scene size (Torralba & Oliva, 2002, 2003) and semantic categories (Oliva & Torralba, 2001; Xiao et al., 2010). Therefore, this framework is orthogonal to object- and scene-centered views to scene perception: Both boundaries and content descriptors can be built up, in theory, from segmented objects, from relations between components, or from global "scene-emergent" features (Oliva & Torralba, 2001; Ross & Oliva, 2010). Given that visual scene analysis generates a rich percept with multiple representations of description, how does the brain accomplish these diverse functions of scene understanding?

NEUROIMAGING STUDIES OF SCENE PERCEPTION

A growing body of research spanning behavioral, computational, and neuroimaging methods has shown that scene representations are not unitary per se. Rather, natural scene processing appears to involve a combination of many visual feature and scene property representations that, in parallel, create the scene image representation. Vision scientists have made significant progress in identifying several candidate brain regions responsible for processing different scene properties. These brain regions are distributed across low-, mid-, and high-level visual areas, with different areas supporting different types of scene information.

For a given semantic scene category (e.g., forest), images are correlated with a plethora of low- and midlevel visual features that help to distinguish that category from other semantic categories. Recent neuroimaging studies have shown that the activity from

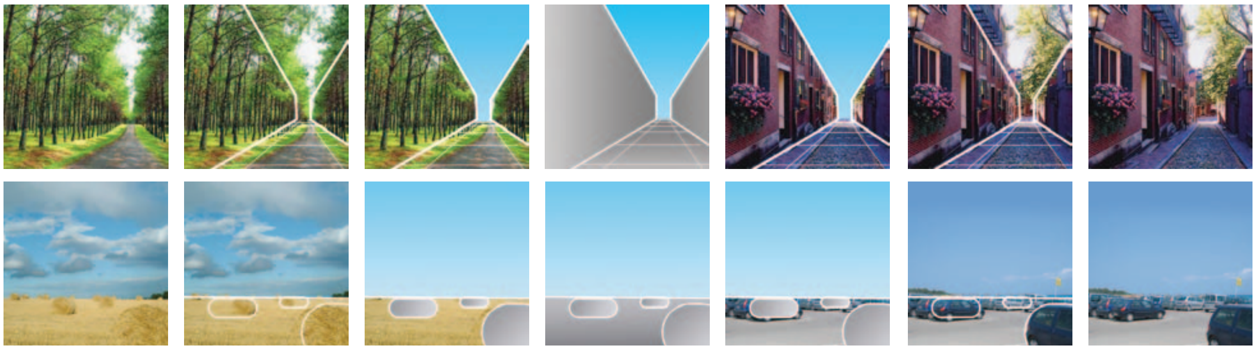


FIGURE 51.4 A schematic illustration of how pictures of real-world scenes can be represented uniquely by their spatial boundaries and content. Keeping the enclosed spatial layout, if we strip off the natural content of a forest and fill the space with urban contents, then the scene becomes a street. Keeping the open spatial layout, if we strip off the natural content of a field and fill the space with urban contents, then the scene becomes a parking lot. (Adapted from Park et al., 2011.)

early visual areas (V1 to V4) can be used to predict which particular image was being viewed by an observer and, even more impressively, to reconstruct some of the visual feature content of the scenes themselves (Kay et al., 2008; Naselaris et al., 2009). Furthermore, the pattern of responses in these regions also contains the information to allow scene classification into a handful of semantic categories (Naselaris et al., 2009; Walther et al., 2009).

Which constituent representations are necessary and sufficient to support our remarkable ability to rapidly understand complex real-world scenes? Although this remains an open question, human functional neuroimaging investigations of the last decade have made significant progress in identifying brain regions important for higher-level aspects of scene and space perception (figure 51.5). Akin to the empirical and theoretical proposals described in the previous sections, recent neuroimaging studies (Kravitz, Peng, & Baker, 2011; MacEvoy & Epstein, 2011; Park et al., 2011) have found that visual scene analysis recruits distinct and complementary high-level representations, indicating distinct neural pathways for the representation of scene/space-centered versus content/object-centered information.

Scene- and Space-Centered Cortical Regions

The two most studied scene-selective regions so far have been the parahippocampal place area (PPA), a region of the collateral sulcus near the parahippocampal–lingual boundary (Epstein & Kanwisher, 1998) and the retrosplenial complex (RSC) (Bar & Aminoff, 2003). Both of these regions respond preferentially to pictures depicting scenes, spaces (like those shown in figures 51.1 and 51.3), and landmarks more than to pictures of faces or single movable objects (for a review, see Epstein & MacEvoy, 2011). A third scene-selective functional region, the occipital place area (OPA) (Dilks, Julian, Paunov, & Kanwisher, in press) is found around the transverse occipital sulcus (see chapter 52 by Kanwisher and Dilks). Interestingly, neither the PPA nor the RSC responses are modulated by the quantity of objects in the scene (i.e., both regions are equally active when viewing an empty room or a room with clutter) (Epstein & Kanwisher, 1998); however, they both show selectivity to the spatial layout of the scene in various tasks (Aguirre, Zarahn, & D'Esposito, 1998; Epstein & Kanwisher, 1998; Janzen & Van Turennout, 2004; Park et al., 2011). So what distinguishes PPA from RSC?

Several studies have shown that the PPA and RSC have different selectivities to perceiving a scene from an observer's viewpoint or perceiving the place the scene is embedded in. For instance, Park and Chun (2009) showed observers different views of scenes that were part of the same panoramic scene, simulating the perception of an observer moving her head, taking snapshots of views in a spatiotemporally continuous way. They found that the PPA treated each view of the panoramic scene as a different "image," suggesting a view-specific representation in PPA (see also Epstein, Graham, & Downing, 2003; Epstein, Parker, & Feiler 2007). By contrast, Park and Chun found that RSC treated different views of a panorama as the same stimulus, suggesting that this region may hold a larger representation of the place beyond the current view (see also Park et al., 2007; Park, Chun, & Johnson, 2010; Epstein, Parker, & Feiler, 2007, for supporting evidence; see Baumann & Mattingley, 2010, for a related topic in heading direction). Interestingly, however, another recent study (Dilks et al., 2011) found that scene representations in PPA were in fact tolerant to more severe transformations (i.e., reflections about the vertical axis—a transformation of 180°). Thus, the further question of whether PPA representations are only tolerant to mirror reversals is an interesting one.

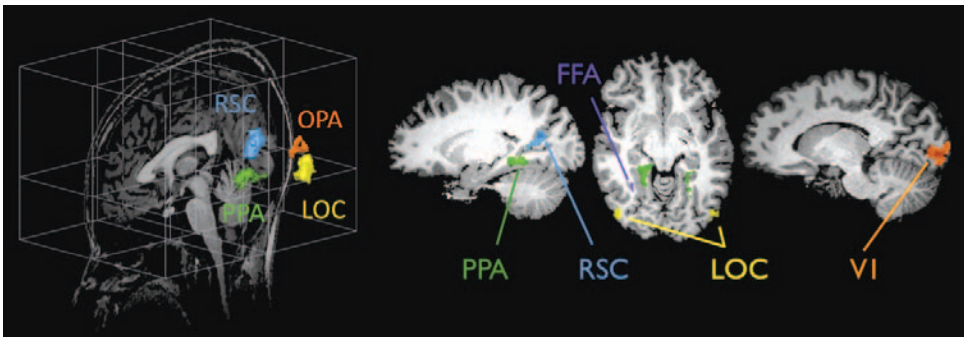


FIGURE 51.5 Several functionally defined regions involved in scene perception are shown for two individuals. PPA, parahippocampal place area; RSC, retrosplenial complex; OPA, occipital place area; LOC, lateral occipital complex. FFA (fusiform face area) and V1 (primary visual cortex) are shown here for comparison. (See chapter 52 by Kanwisher and Dilks.)

K2

Content- and Object-Centered Cortical Regions

Beyond spatial layout information, perceiving objects is an important part of scene processing. Much of the time identifying the objects in a scene will dictate the scene's function (figure 51.1). For that reason the lateral occipital complex (LOC) is another candidate region of the scene perception network. The LOC is an area of the ventral visual pathway that specializes in representing object shapes and object categories (Grill-Spector et al., 1998; see chapter 52 by Kanwisher and Dilks for a complete review). Among the many neuroimaging studies examining the nature of object representations in the LOC, the section below describes findings with implications for an object-centered pathway to scene understanding.

Because scenes typically contain several objects rather than single isolated objects, it stands to reason that object-processing brain regions should be encoding the content of a scene. Recent studies have shown that the presence of multiple objects in the visual field can be decoded from averaging the activity for individual objects in LOC (MacEvoy & Epstein, 2009, 2011). Furthermore, the pattern of neural responses in the LOC is sufficient to distinguish among a few scene categories (e.g., beach and city) (Walther et al., 2009) as well as to decode whether certain objects were present within the scenes (Peelen, Fei-Fei, & Kastner, 2009).

The LOC is not the only brain region involved in object processing. Large objects—landmarks and buildings, for example—have been shown to activate the PPA (Epstein & Kanwisher, 1998). Importantly, certain physical and experience-based properties of real-world objects evoke selective brain responses based on properties such as their real-world size (e.g., paperclip vs. car) (Konkle & Oliva, 2012), their contextual strength within a scene (e.g., a fire hydrant vs. a book) (Bar, 2004), and whether the objects define a local space in a larger scene (e.g., a sofa) (Mullally & Maguire, 2011). An example of selective activity for object properties was shown by Konkle and Oliva (2012), in which the authors identified a region of interest in the parahippocampal cortex that showed peaks of activity for large objects that our bodies typically interact with (e.g., by physically interacting with that object as with a bed or a sofa or walking toward it as to a piano or a dishwasher). Similarly, a left-lateralized region in the occipitotemporal sulcus showed peaks of selectivity for objects of a small physical size in the world that can be typically handled (e.g., a strawberry or a hat). Thus, as multiple regions of the brain encode different spatial aspects of a scene, multiple regions may represent different types of content and objects encountered in a scene.

Complementary Representations of Spatial Boundary and Object Content

How can these multiple regions, and perhaps others, work together to form a complete representation of a scene? Although this remains an open question for future neuroimaging investigation, some recent lines of research have started to shed light on how the brain analyzes an input natural image using a visual feature representation and ultimately creates a higher-level representation of the space and content of the scene. Inspired by the fact that spatial boundaries and scene content are orthogonal properties of a real-world scene (see figure 51.4), Park et al. (2011) designed an experimental paradigm to empirically test the underlying nature of the representations in the PPA and LOC. As illustrated in figure 51.4 the shape of a scene may be expansive and open to the horizon, as in a field or parking lot, or closed and bounded by frontal and lateral surfaces, as in forests or streets. Furthermore, a scene may be comprised of natural or urban (manufactured) objects, independent of its spatial boundary.

Park et al. (2011) used pattern analysis of neural responses in the PPA and LOC to classify whether a scene belonged to a particular class (i.e., open, closed, natural, or urban space). By analyzing the types of classification errors that occurred when each region's response was used, the authors could probe questions relating to whether the regions represent a scene in an *overlapping* fashion (e.g., both produce similar errors when classifying different scenes) or in a *complementary* fashion (e.g., each region shows a specialization in representing either the boundaries or content of a scene). They found a dissociation between the types of classification errors made in two brain regions. The PPA confused scenes with similar spatial boundaries, regardless of the type of content, whereas the LOC made the opposite errors (confusing scenes with the same content, independent of their spatial layout). In the PPA, scenes representing closed urban environments, such as streets and buildings, were most confused with closed natural scenes, such as forests and canyons (see also Kravitz, Peng, & Baker, 2011). On the other hand, the LOC confused scenes that contained similar objects and surfaces, such as mistaking fields, deserts, or ocean with forest, mountain, or canyons; LOC accuracy was insensitive to the spatial layout.

Using multivoxel pattern analysis as well, MacEvoy and Epstein (2011) found a similar result. In their study neural responses evoked in the LOC by pictures of scenes (e.g., a kitchen, a street) were well predicted by the response patterns elicited by their sole prominent objects (e.g., refrigerator, traffic light), although this was not the case for PPA activity. Altogether, the current state of the art in neuroimaging studies suggests that scene analysis in the brain recruits distinct regions that perform different and complementary computations to create a singular and unique representation of the scene in view.

CONCLUSION

Altogether, studies across disciplines and methods provide evidence for the existence of a distributed scene representation, encoding different levels of information, as a basis for scene perception. The existence of neural pathways for the representation of scene/space-centered versus content/object-centered information corroborates the experimental and computational frameworks of scene perception that have emerged in the last decade. Whereas many discoveries remain to be found regarding how the brain computes this immediate "understanding" of the world, and which levels of features are used to perform particular operations, the common theoretical framework in scene perception between disciplines should provide fast-track progress in the years to come.

ACKNOWLEDGMENTS

The writing of this chapter was supported by a National Eye Institute grant EY020484 to A.O. The author would like to thank Barbara Hidalgo-Sotelo, Daniel Dilks, and Soojin Park for insightful comments.

REFERENCES

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.

Bar, M. (2004). Visual objects in context. *Nature Reviews. Neuroscience*, *5*, 617–629.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*, 347–358.

Baumann, O., & Mattingley, J. B. (2010). Medial parietal cortex encodes perceived heading direction in humans. *Journal of Neuroscience*, *30*, 12897–12901.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–263). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cutting, J. E. (2003). Reconceiving perceptual space. In H. Hecht, M. Atherton, & R. Schwartz (Eds.), *Perceiving pictures: An interdisciplinary approach to pictorial space* (pp. 215–238). Cambridge, MA: MIT Press.

Dilks, D. 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*, *31*(31), 11305–11312.

Dilks, D. D., Julian, J. B., Paunov, A., & Kanwisher, N. (in press). The occipital place area (OPA) is causally and selectively involved in scene perception. *Journal of Neuroscience*.

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

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

Epstein, R. A., & MacEvoy, S. P. (2011). Making a scene in the brain. In L. Harris & M. Jenkin (Eds.), *Vision in 3D environments* (pp. 255–279). Cambridge: Cambridge University Press.

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*, 6141–6149.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 1–29.

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

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

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

Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene property. *Journal of Experimental Psychology. Human Perception and Performance*, *36*, 1430–1442.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*, 152–160.

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.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology. Human Perception and Performance*, *7*, 604–610.

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

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

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition*, *15*, 12–19.

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

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

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*, 7322–7333.

MacEvoy, S. P., & Epstein, R. A. (2009). Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Current Biology*, *19*, 943–947.

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

Mullally, S. L., & Maguire, E. A. (2011). A new role for the parahippocampal cortex in representing space. *Journal of Neuroscience*, *31*, 7441–7449.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *53*, 902–915.

Navon, D. (1977). Forest before the trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

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

Oliva, A., & Torralba, A. (2002). Scene-centered description from spatial envelope properties. In H. Bulthoff, S. W. Lee, T. Poggio, & C. Wallraven (Eds.), *Computer Science Series Procedure, Second International Workshop on Biologically Motivated Computer Vision* (pp. 263–272). Tübingen: Springer-Verlag.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from its 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 panoramic scene perception. *NeuroImage*, *47*, 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*, 2813–2822.

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

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*, 94–97.

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

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*, 10–15.

Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological Science*, *20*, 6–10.

Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1), 2, 1–25. doi:10.1167/10.1.2.

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

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

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Pattern Analysis and Machine Intelligence*, *24*, 1226–1238.

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

Walker Renninger, L., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*, 2301–2311.

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*, 10573–10581.

Xiao, J., Hayes, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). SUN Database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492). San Francisco, CA: IEEE Computer Society.