Scene Perception

Ronald A. Rensink

Cambridge Basic Research

Nissan Research & Development, Inc.

Cambridge, Massachusetts, USA

In AE Kazdin (ed.), *Encyclopedia of Psychology*. vol. 7. (pp. 151-155).
New York: Oxford University Press. 2000.

Scene Perception is the visual perception of an environment as viewed by an observer at any given time. It includes not only the perception of individual objects, but also such things as their relative locations, and expectations about what other kinds of objects might be encountered.

Given that scene perception is so effortless for most observers, it might be thought of as something easy to understand. However, the amount of effort required by a process often bears little relation to its underlying complexity. A closer look shows that scene perception is a highly complex activity, and that any account of it must deal with several difficult issues: What exactly is a scene? What aspects of it do we represent? And what are the processes involved? Finding the answers to these questions has proven to be extraordinarily difficult.

However, answers are being found, and a general understanding of scene perception is beginning to emerge. Interestingly, this emerging picture shows that much of our subjective experience as observers is highly misleading, at least in regards to the way that scene perception is carried out. In particular, the impression of a stable picture-like representation somewhere in our heads turns out to be largely an illusion.

To see how this comes about, imagine a seashore where there is a sailboat, some rocks, some clouds, and perhaps a few other objects (see Figure 1). How do we perceive this scene? Intuitively, it seems that the set of objects in the environment would give rise to a corresponding set of representations in the observer. Thus, there would be detailed representations of the sailboat, clouds, etc., with each representation describing the identity, location, and 'meaning' of the item it refers to. In this view, the goal of scene perception is to form a literal re-presentation of the world, with all of its visible structure represented concurrently and in great detail everywhere. This representation then serves as the basis for all subsequent visual processing.

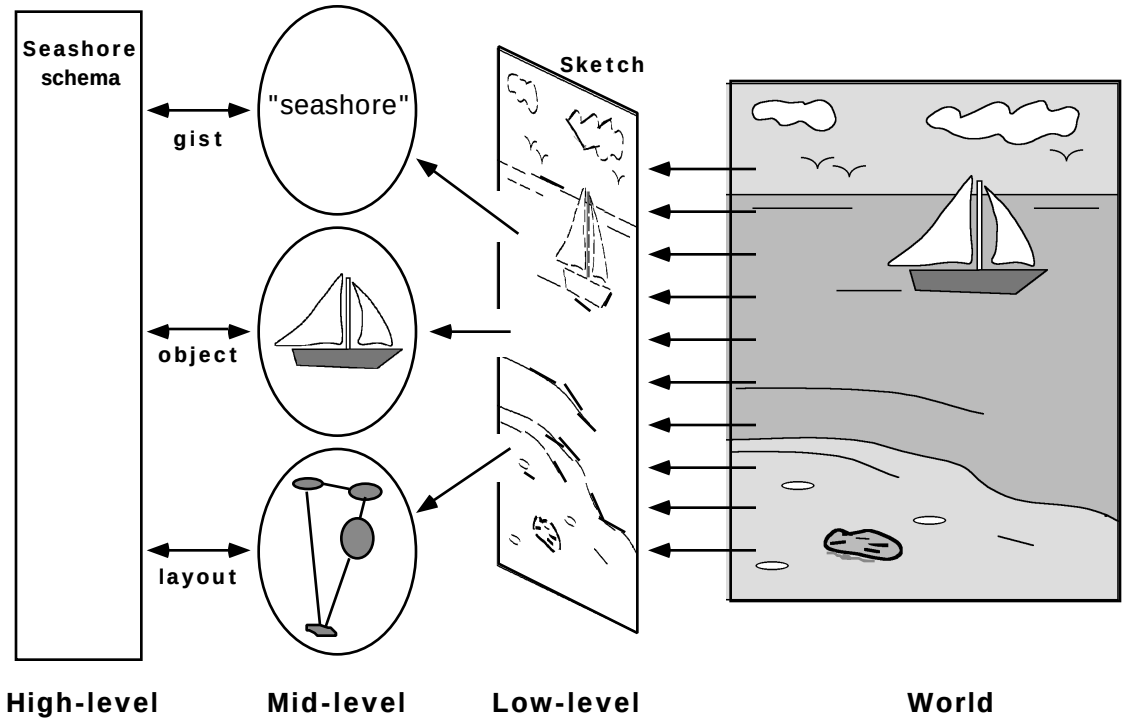

Figure 1: Representation of a scene.

As it turns out, however, memory for visual detail is generally quite short-lived (maybe 100 ms, Irwin, 1996). And since successive eye fixations are usually separated by at least 150-200 ms, it follows that their contents cannot be integrated into a complete, detailed representation. Conversely, it has also been found that a complete, detailed representation is not necessary—the meaning of a scene (e.g., whether or not it is a seashore) can be determined within 100-120 ms (Biederman, 1981; Potter, 1976), a time that allows recognition of only a few objects. Evidently, a small set of object and scene properties is enough to provide us with an impression of a scene that is complete and detailed everywhere.

This realization causes a shift in perspective: scene representations are no longer structures *built up* from eye movements and attentional shifts, but rather are rapidly-formed structures that can *guide* such activities. More generally, the goal of scene

perception appears to be the establishment of an immediate context for various aspects of visual processing, as well as for visuo-motor operations such as reaching or locomotion.

How might this be done? Scene perception is a special case of visual perception, and so likely involves the same processing levels as vision generally. [See VISION.] The first of these is *low-level* processing, which uses the incoming light to recover simple properties of the environment visible to the observer, such as the color of the sky or the texture of the clouds. The second is *mid-level* processing, concerned with more complex tasks, such as separating the sailboat from its background, and representing it as a distinct object with its own size, shape, and colors. Finally, there is *high-level* processing, concerned with issues of meaning. For example, high-level processes might identify a mid-level object as a sailboat and a scene as a seashore, and so allow us to expect such things as seagulls, whitecaps, and fishing vessels.

The exact nature of the processes involved in scene perception is largely unknown. However, at least some understanding—summarized in the following sections—has been obtained of the kinds of operations carried out at each processing level, and their interactions with each other.

i) Low-level Processing

Scene perception begins with the creation of a detailed map-like representation, or *sketch*, obtained from the pattern of light that falls on the retina (Marr, 1982). This sketch describes the properties of the scene at every point in the visual field (see Figure). These can be simple image-based features (color, size, etc.) or more complex scene-based properties (three-dimensional slant, surface curvature, etc.) obtained by "quick and dirty" interpretation processes. [See EARLY VISION.] Thus, when looking

at the seashore, the sketch will indicate "white" at regions corresponding to the sails and the surrounding clouds, and "blue" at regions corresponding to the sky. It will also describe the distribution of other properties as well, for example indicating a degree of surface curvature at regions corresponding to the sail and the clouds, and no curvature at regions corresponding to the sky.

In the absence of attention, these low-level representations are volatile, their contents being overwritten by subsequent stimuli or else fading away a few hundred milliseconds after light stops entering the eyes (Rensink, 1999). Thus, the sketch at any particular fixation effectively exists only as long as the eyes do not move. As such, although the sketch is detailed, it is not stable—it must be constantly regenerated, with a new sketch formed with each new fixation.

ii) Mid-level Processing

The volatile, ever-changing representations of low-level vision cannot support the stable perception we have of a scene. Instead, this must be accomplished by a different set of representations owing their stability to some form of short-term memory. Because such memory is costly, only a few aspects of scene structure can be given stable representation. These appear to include object structure, scene layout, and scene gist (see Figure). The extent to which these aspects are determined independently of each other is unknown.

Object structure. Although visual detail is generally volatile, information about a small number of objects can be held across an eye movement or temporal interruption (Irwin, 1996; Rensink, 1999). This is likely done by attentional mechanisms that can store several attributes of an object—such as its shape, location, or the arrangement of its parts—in visual short-term memory (Irwin, 1996). Thus, if attention is given to the

sailboat at the seashore, its representation would contain information such as the shape of its hull and sails, and perhaps the relative location of the sails with respect to each other and to the hull. This representation will remain stable (in short-term memory) as long as attention is directed to the sailboat. [See VISUAL ATTENTION.]

The capacity of short-term memory is severely limited, so that little accumulation of object structure is possible (Irwin, 1996). And although facilitation of processing can occur between related objects viewed in succession, this is restricted to the past few objects viewed (Henderson, 1992). Consequently, only a few object representations are in play at any time, with a limited amount of information in each. This strongly suggests that object representation is *dynamic*: although all the objects in a scene cannot be represented simultaneously, eye movements and attentional shifts are coordinated such that a stable representation of any selected object can be formed whenever needed (Rensink, 1999).

Scene layout. Just as there appears to be a stable representation of the arrangement of parts within an object, so does there appear to be a stable representation of the arrangement, or *layout*, of objects within a scene. The layout of the seashore, for example, might be described by the relative position of the sailboat, rocks, clouds, and shoreline with respect to each other. Such information is vital if the limited information obtained from individual eye fixations is to be integrated into a structure capable of directing subsequent eye movements and attentional shifts. [See EYE MOVEMENTS.]

Layout is sometimes thought to be represented by a *schematic map*, which describes the locations of various objects in the scene without a detailed description of their structure or identity (Hochberg, 1978). Such a minimal description is sufficient for the guidance of many actions, such as reaching or obstacle avoidance. And the lack of detailed description allows the map to be constructed quickly and with minimal

memory. However, little is known about the particular aspects of layout that are represented or about how a layout representation might be formed.

Scene gist. The most abstract aspect of a scene is its meaning, or *gist*. In the case of the example considered above, the gist would simply be "seashore". Similarly, a scene with several boats and a dock might be perceived as "harbor". Other examples of gist would be "farmyard", "shopping center", or "city". Gist is a highly invariant quantity, remaining constant over many different eye positions and viewpoints, as well as over many changes in the composition and layout of objects in an environment. As such, it can potentially provide a stable context for other processes, such as object recognition.

Experiments based on naming and categorization show that gist can be determined within 100-120 ms of presentation (Potter, 1976). Only a few objects can be perceived within this time, suggesting that the perception of gist may be based on the perception of two or three key objects. For example, if an object were recognized as a sailboat, it would suggest that the scene is a seashore, harbor, or open sea; if a beach were also recognized, it would reduce the set of candidates to "seashore". Another possibility is that gist can be determined without perceiving objects at all (Henderson, 1992). This position is supported by the finding that different gists can be determined simultaneously at different spatial scales, without any need for attention (Schyns & Oliva, 1994). This suggests that gist may be invoked by low-level features diagnostic of scene category, such as the distributions of line orientations or colors in the image.

iii) High-level Processing

The considerable invariance of object structure, scene layout, and scene gist over different viewing positions implies that they often remain constant over relatively long stretches of time. This allows the long-term learning of their occurrence in various

scenes. The result of this is a *scene schema*, an interlinked collection of representations in long-term memory that describes such things as the kinds of objects that occur together, and how they might be positioned relative to each other. The information in a schema can constrain the kinds of objects expected, and perhaps also indicate their importance for the task at hand (Friedman, 1979).

In contrast to the simple structures that invoke it, the contents of a schema can be relatively sophisticated. For instance, schemas are believed to include an inventory of objects likely to be present in the scene. In the case of a seashore schema, this inventory could include rocks, clouds, a beach, boats, and possibly a few other objects. Various aspects of layout information may also be stored, such as the relative locations of the inventory objects (Mandler & Ritchey, 1977).

iv) Interaction of Systems

Although details are far from clear, it appears that a relatively simple set of interactions may underlie most of scene perception (see Figure). When viewing a scene, the low-level processes provide a constantly-regenerating sketch of the properties visible to the viewer. A subset of these properties could determine scene gist and layout, which then invoke a scene schema. Subsequent processes could attempt to verify the schema and supply it with the information needed to carry out any required actions (Friedman, 1979). When an unexpected object is encountered, more sophisticated (attentional) processes could reevaluate the object, reevaluate the gist, or learn a new association between the two. Meanwhile, the layout could be used as a direct check on the current interpretation, as well as providing spatial guidance of attention to appropriate items.

Such an account of scene perception creates an apparent paradox: if detailed representations are short-lived, and stable representations contain little detail, how can our impression of a scene be both detailed and stable? The solution to this lies in the "just in time" nature of object representation, where a stable representation of an object can be formed whenever needed via the coordinated use of attentional shifts and eye movements. Provided that only a few objects need to be represented at any one time, such a dynamic representation has all the power of a static one, while requiring much less in the way of processing resources (Rensink, 1999).

Bibliography

Biederman, I. (1981). "On the semantics of a glance at a scene". In M. Kubovy and J.R. Pomerantz (Eds.), *Perceptual Organization* (pp. 213-253). Hillsdale, NJ: Erlbaum.

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

Henderson, J.M. (1992). Object identification in context: The visual processing of natural scenes. *Canadian Journal of Psychology*, **42**:319-341.

Hochberg, J.E. (1978). *Perception (2nd ed.)* (ch. 6). Englewood Cliffs, NJ: Prentice-Hall.

Irwin, D.E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, **5**, 94-100.

Mandler, J. & Ritchey, G.H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, **3**, 386-396.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

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

Rensink, R.A. (1999). The dynamic representation of scenes. *Visual Cognition*, **6**, xx-xx. (To appear late 1999.)

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

Ronald A. Rensink

Hypertext terms

Key terms:

Early vision

Eye movements

Object perception

Visual attention

Visual memory

Names:

Antes, James R.

Biederman, Irving

DeGraef, Peter

Friedman, Alinda

Henderson, John M.

Hochberg, Julian

Intraub, Helene

Irwin, David E.

Loftus, Geoffrey R.

Mandler, Jean M.

Pollatsek, Alexander

Oliva, Aude.

Potter, Mary C.

Rensink, Ronald A.

Schyns, Philippe G.

Simons, Daniel J.

Publications:

Eye Movements and Visual Cognition: Scene Perception and Reading. K. Rayner, ed. New York: Springer. 1992.

Canadian Journal of Psychology, 46(3). 1992. (Special Issue on Scene Perception).

Visual Cognition, 6(?). 1999. (Special Issue on Change Detection and Visual Memory).

Eye Movements: Cognition and Visual Perception. D.F. Fisher, R.A. Monty, and J.W. Senders, eds. Hillsdale, NJ: Erlbaum. 1981.

Eye Movements and Psychological Processes. R.A. Monty, and J.W. Senders, eds. Hillsdale, NJ: Erlbaum. 1976.