Visual Scene Perception

A visual scene is commonly defined as a view of an environment comprised of objects and surfaces organized in a meaningful way, like a kitchen, a street or a forest path. More broadly, the domain of scene perception encompasses any visual stimulus that contains multiple elements arranged in a *spatial layout*, for example a shelf of books, an office desk, or leaves on the ground. As a rough distinction, objects are typically acted upon, while scenes are acted within. Most visual elements in the world will be categorized as either an object or scene, although an item's status may change depending on how it is used: a keyboard is readily regarded as an object when being purchased, for example, but when placed on the desk, it becomes part of a continuous surface and layout, and a scene (for your hands), when you are trying to find the right key.

Perceiving scenes presents a paradox that the movie industry, for instance, has successfully exploited for decades. The complex arrangement of objects and surfaces in natural scenes can create the impression that there is too much to see at once. However, 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! This is about the same time it takes a person to identify that a single object is a face, a dog or a car. This remarkable feat of the human brain can be experienced (and enjoyed!) at the movies: with a few rapid scene cuts from a movie trailer, it seems as if we have seen and understood much more of the story in a few instants than could be described later. We will easily remember the movie's genre and limited context (for example, a romantic story with views of Venice, or a science fiction story set in the near future), but we will have forgotten detailed information. The same phenomenon happens when quickly changing television channels or flipping pages of a magazine: one single glance is often enough to recognize a popular TV personality, a high-speed car chase, a football game, etc., but memory of the details is wiped away almost immediately. Perceiving scenes in a glance is like looking at an abstract painting of a landscape and recognizing that a "forest" is depicted without seeing necessarily the "trees" that create it.

Perceiving and Remembering Visual Scenes

Because a scene can encompass a large space, we must acquire information about its extent by navigating our bodies and moving our head and our eyes. Although we seem to experience a continuous world, the brain actually samples the visual world in a series of "snapshots", by moving the eyes about every 1/3 of a second (see *rapid serial visual presentation)*. Perceiving a visual scene is like watching a movie in which the camera cuts quickly from one view to the next. During each brief moment in which we see a particular view, we are able to understand its overall meaning or "gist" (e.g. a wedding, a dog running in a park, a busy New York street), we have a compelling perception of space, and we anticipate what is coming next. However, details and objects are quickly lost from *visual memory* and, in many cases, are not even perceived in the first place.

Figure 1: Look for one second only at each picture. Then close your eyes and flip the page to view Figure 2.

To illustrate this phenomenon, look at Figure 1. It contains photographs of 8 scenes. Spend about one second on each image, then close your eyes for a few seconds, and flip the page to view Figure 2. There you will find a set of 8 photographs. Your task is to determine which images you saw in the first figure. Come back to reading this section after doing the test. Although some of the images in Figure 2 look remarkably similar to images from Figure 1, they are, in fact, all different. However, most people misidentify some or all of the images in Figure 2 as being the same images from Figure 1. The demonstration illustrates the limitations of visual scene perception.

The boundary extension phenomenon: Fitting views into scene schemas

The pair of images which show a set of objects on a table (A-1) illustrates a systematic error in scene perception revealed by Helene Intraub's work: people often remember a wider view of a scene than was originally seen. If you look closely at the borders of the images, you will notice that Figure 2-1 is a wider view in which you can see more of the pitcher and the tablecloth. This boundary extension effect occurs as early as half a second after viewing an image, and may also occur between two views of the same place. For instance, the autumn park images (B-2) depict the same place, but in the second image, the observer has moved a few steps closer to the riverbank. In many cases, people will not notice the difference between two views of the same place that differ by 10 to 20 degrees of head movement, or a few steps of body translation.

These spatial memory errors reveal a fundamental mechanism of visual scene analysis: people rely on their previous experience and knowledge of the world to rapidly process the vast amount of detail in a real world scene. One's current view of a scene is automatically incorporated into a "scene schema," which includes stored memories of

similar places which have been viewed in the past, as well as expectations about what is likely to be seen next. Although we aren't aware of it, viewing a scene is an active process, in which images are combined with memory and experience to create an internal reconstruction of the visual world.

The change blindness phenomenon: objects and details go unnoticed

The pictures of a house and lawn (C-6) and the pictures of a city street (D-4) illustrate another well known phenomenon of visual scene perception: people are surprisingly poor at detecting a change in a scene when the change happens between two eye movements or a shift in viewpoint (like a turn of the head, or, as in this demonstration, a flip of a page). This demonstration illustrates the change blindness phenomenon studied by Ronald Rensink and Daniel Simons (see *change detection*). In Figure 2, the scene depicting a house and lawn is missing the parasols and people in the center, the two images of the European street differ actually on a dozen details! (the color of the door, the shop, the presence of pedestrian, the whole building on the right, etc.). The change blindness phenomenon further illustrates that, contrary to our subjective experience, many details of the visual world simply go unnoticed.

The Gist Phenomenon: Understanding the whole before the parts

If you thought that the canal street (E-7), the mountainous coast (F-5) or the corridors (G-8) were the same images, then you may have experienced an error based on scene gist. Different images that share a similar meaning and look similar may be falsely remembered as the same scene. This phenomenon is similar to the feeling of "déjà vu", when a novel place is experienced as familiar. In a similar vein, it should have been easy to spot that the living room of Figure 2 was a new scene, as there are no similar images, in terms of spatial layout or semantics in the set of Figure 1.

Figure 2: Which pictures did you see in Figure 1? (See text for details).

Mechanisms of Visual Scene Perception

We understand the meaning of a real-world scene in an instant, but we deform the visual input with our memories, expectations and knowledge. Clearly, scene perception is an active process linking visual input to memory and expectation. What is the mental process behind scene perception?

The prevalence of global layout

Research in scene perception has traditionally treated objects as the atoms of recognition. However, studies have shown that the speed and accuracy of scene recognition are not affected by the number of objects in the scene, which suggests an alternative: the meaning of scenes may be inferred from their spatial layout, the arrangement of surfaces and forms. Figure 3 A illustrates how the spatial arrangement of surfaces and regions drives scene perception. When looking at the blurred image on the left, viewers describe the scene as a car parked in front of a fancy building's facade. Even though the local information available in the image is insufficient to recognize the objects individually (the car, street, and facade are too blurry be recognized in isolation as illustrated in the center of Figure 3-A), viewers are confident and highly consistent in their descriptions. This is because the blurred scene has the spatial layout of a typical street.

When the image is shown in high resolution, new details reveal that the image has been manipulated and that the buildings are in fact pieces of furniture. In fact, more than half of the image depicts an indoor scene, not a street. The misinterpretation of the low-resolution image is not a failure of the visual system. Instead, it illustrates the strength of spatial layout information in determining the identity of the objects within a scene. The importance of global layout information is especially evident in degraded viewing conditions (for example, viewing a scene at a distance, or at a quick glance) in which object identities cannot be derived from local information alone.

Segmentation and Figure-ground analysis of visual scene

Figure 3-B demonstrates further the dominance of global information in scene perception. The left image seems to show a forest scene, with the observer looking up through the trees to see the sky in the background; the right image represents a snow-filled depression on a craggy mountainside. But if you study the images closely, you may notice local inconsistencies such as incorrect shadows and odd surface shapes or occlusions. Indeed, if you turn the book upside down, you should see a new interpretation of each image: the mountain scene becomes a much larger view of a snowy mountain range (a view over a cliff) under a cloudy sky, and the forest scene depicts a river receding into the distance. In this particular illusion, reversing the images changes the assignation of elements as *figure* (i.e. the object of focus) or *ground* (i.e. the rest of the perceptual field, see *perceptual segregation*). When the mountain scene is turned upside down, the snow is perceived as sky and some of the rocks even change size; in the forest scene, the sky becomes a river, and the darker foliage in the

distance (a figure element in the original view), turns into the reflection of trees in the water (a ground element in the upside down view). The attribution of scene elements as *figure* or *ground* depending on the perceived layout shows that perceiving a complex, real world scene is an interaction between *bottom-up* and *top-down processing*. The visual features extracted from the image (e.g. color, lines, and patterns of texture) are rearranged according to our expectations and knowledge of the world. For instance, we know that a blue region at the top of an outdoor scene is most likely the sky; and we know that, in a large scene, elements below the horizon line are more likely to be close-by than elements above the horizon line.

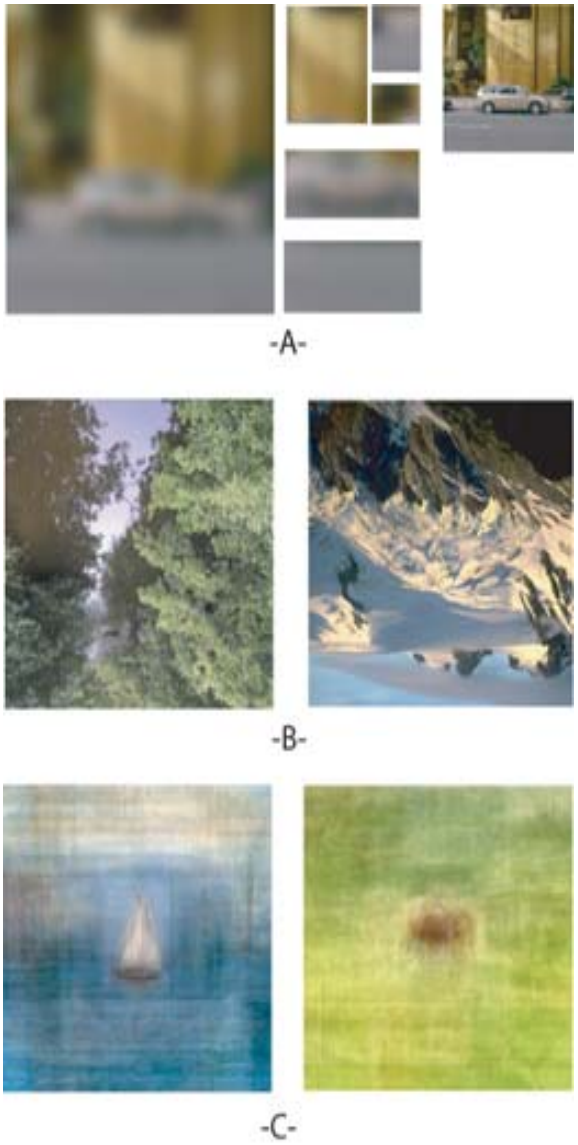

Figure 3: - A - The blurry image looks like a street scene, even if the objects cannot be recognized in isolation; -B- (left) A forest scene and (right) a mountain scene. – C - Average images centered on pictures of sailboats (left) and cows (right).

The role of statistical regularities in visual scene perception

The visual world is not random: objects, surfaces, and the gaps between them are organized in predictable fashion. Just as there are rules governing the structure of objects (for example, faces have two eyes, a nose, and a mouth), the structure of many real-world scene environments is governed by strong configural rules which predict what kind of objects and surfaces will appear near each other. This is illustrated in Figure 3-C. By averaging hundreds of images aligned on either a sailboat (left) or a cow (right), a common pattern of intensities emerges around each object. The average sailboat is surrounded by a blue background (water) with lighter patches at the top and sides (sky or shore) and vertical elements which hint at neighboring boats. The average cow, on the other hand, is surrounded by greenish patterns representing fields. This averaging illustrates the statistical regularities that are imposed on the scene by the objects within it.

Similarly, environments that share the same category and function (such as streets, beaches, or forests) tend to share similar layouts: a highway is a flat ground surface stretching to the horizon, affording speedy travel; a corridor is an enclosed, perspective space, with an unobstructed path to afford navigation; cities are made up of vertical facades; and a dining-room is a space organized around a central object, a table. Because we experience scenes as real, three dimensional spaces, Oliva and Torralba have proposed that visual scene perception is based on a global layout representation that describe the space and volume of the scene (e.g. beaches are large, open environments; closets are small, enclosed spaces) and not necessarily the objects the scene contains.

Importantly, human observers are very good at extracting summary statistics of scenes: that is, they are able to rapidly estimate the regularities (and differences) in the layout of a scene or in the objects that compose it. For example, people know the mean size of a collection of objects without knowing the precise size of each of them; they know the center of mass of a group of objects without necessarily remembering the location of each of them; and they can rapidly recognize whether a space is a small or large volume, or a busy or an empty space before identifying whether it is a street or a closet. Summary statistics and other regularities of the environment learned with experience (see *statistical learning*) are valuable for scene perception, because they provide an efficient and compact representation of the image that can provide information about other scene properties and facilitate *object perception* and search before local image information has been fully analyzed. This is experienced every day: a glance at your own kitchen or bedroom cues the location of objects that may not be immediately visible. For example, a glimpse of your bedroom should be sufficient to cue you to the location of the alarm clock in your mental representation of that place, even if the alarm clock isn't visible from your current position.

Conclusion

How do we perceive visual scenes? Decades of behavioral research suggest that scene perception begins at a global level. First, the spatial layout and observer's viewpoint are evaluated, and then the localization and recognition of parts and objects within the scene progress at a slower rate. In other words, you would know that you are in a house, and in a large kitchen, before recognizing that that particular form is a fridge and that this object is a microwave. Memory for scenes is often prone to small errors and distortions, but these false reconstructions reveal how the brain analyzes complex spaces. Environments have all sorts of regularities that we learn and store in memory. When faced with a novel scene, we use our knowledge and expectations to rapidly understand its meaning, although this comes at the cost of losing some detail from perception. Scene perception is essentially the process of reconstructing a space from a lifetime of stored representation.

Aude Oliva

Massachusetts Institute of Technology

See also: **Auditory scene analysis; Bottom-up vs. top-down processing; Change detection; Computer Vision; Navigation through spatial layout; Perceptual organization: Vision; Perceptual segregation; Rapid Serial Visual Presentation (RSVP); Spatial layout perception: neural; Spatial layout perception, psychophysical; Statistical learning; Visual Memory; Visual Scene Statistics; Visual Spatial Frequency Analysis.**

Further Readings and References

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143-177.

Chun, M. M. (2003). Scene perception and memory. In D. Irwin and B. Ross (Eds.) *Psychology of Learning and Motivation: Advances in Research and Theory: Cognitive Vision, Vol. 42* (pp. 79-108). Academic Press, San Diego, CA.

Epstein, R., and Kanwisher, N. (1998). A Cortical Representation of the Local Visual Environment. *Nature,* 392: 598-601.

Henderson, J. M., & Hollingworth, A. (1999). High-Level Scene Perception. *Annual Review of Psychology, 50,* 243-271.

Greene, M.R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58, 137-179.*

Intraub, H. (2007). Scene perception: The world through a window. In Peterson, M.A., Gillam, B., Sedgwick, H.A. (Eds), in The *Mind's Eye: Julian Hochberg on the Perception of Pictures, Films, and the World* (pp. 454-466). NY: Oxford University Press.

Navon, D. (1977). Forest before Trees: The precedence of global features in visual perception. *Cognitive Psychology*, 353-383.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Science*. 11(12), 520-527.

Potter, M.C. (1999). Understanding sentences and scenes: the role of conceptual short-term memory. In *Feeting memories: Cognition of Brief Visual Stimulus*. Veronika Coltheart (Ed.). (pp. 13-46), MIT press.

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.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1), 16-20.

Standing, L. (1973). Learning 10,000 pictures. *Quartely journal of experimental psychology*, 25, 207-222.