CHAPTER

41

Gist of the Scene

Aude Oliva

ABSTRACT

Studies in scene perception have shown that observers recognize a real-world scene at a single glance. During this expeditious process of seeing, the visual system forms a spatial representation of the outside world that is rich enough to grasp the meaning of the scene, recognizing a few objects and other salient information in the image, to facilitate object detection and the deployment of attention. This representation refers to the *gist* of a scene, which includes all levels of processing, from low-level features (e.g., color, spatial frequencies) to intermediate image properties (e.g., surface, volume) and high-level information (e.g., objects, activation of semantic knowledge). Therefore, *gist* can be studied at both perceptual and conceptual levels.

I. WHAT IS THE "GIST OF A SCENE"?

With just a glance at a complex real-world scene, an observer can comprehend a variety of perceptual and semantic information. The phenomenal experience of understanding everything at once, regardless of the visual complexity of the scene, can be experienced while watching television and flipping rapidly through the channels: with a mere glimpse of each picture, observers can grasp each one's meaning (a politician, a car chase, the news, cartoons, etc.) independently of the clutter and the variety of details. This refers to the *gist* of a scene (Friedman, 1979; Potter, 1976).

Behavioral studies have shown that observers can recognize the basic-level category of the scene (e.g., a street; Potter, 1976), its spatial layout (e.g., a street with tall vertical blocks on both sides (Schyns and Oliva, 1994), as well as other global structural information (e.g., a large volume in perspective) in less than 100 msec. Observers may also remember a few objects (e.g., a red car and green car), the context in which they appear (e.g., parked on the side) and other low-level characteristics of regions that are particularly salient (see Chapter 39).

Concurrent with gist development is the automatic activation of a framework of semantic information, including scripts (e.g., the actions occurring in a scene; Friedman, 1979), scene-related knowledge (e.g., the typicality or familiarity of a particular scene), as well as predictions of which objects are likely to be found in the environment. The gist benefits object detection mechanisms almost instantaneously, as well as attention deployment and gaze control in cluttered scenes (Oliva, Torralba, Castelhano, and Henderson, 2003; Torralba and Oliva, 2003 (see Chapter 96)).

II. THE NATURE OF THE GIST

Because gist includes all levels of visual information—ranging from low-level features (e.g., color, contours) to intermediate (e.g., shapes, texture regions) and high-level information (e.g., activation of semantic knowledge)—it can be represented at both perceptual and conceptual levels. Perceptual gist refers to the structural representation of a scene built during perception. Conceptual gist includes the semantic information that is inferred while viewing a scene or shortly after the scene has disappeared from view. Conceptual gist is enriched and modified as the perceptual information bubbles up from early stages of visual processing.

A. Conceptual Gist

The pioneering work of Mary Potter (1976; Potter and Levy, 1969) described the conceptual information that observers are able to quickly comprehend from a picture. In their original study, Potter and Levy (1969) allowed observers a single glance at a series of meaningful images before testing their memory of these images. When presented alone for 100 msec, each image was easily remembered. However, when embedded in a Rapid Serial Visual Presentation paradigm (e.g., RSVP pace of 125 msec exposure per image), performances deteriorated to the level of chance. In a second study (Potter, 1976), observers were cued ahead of time about the possible appearance of a picture in the RSVP stream (the cue consisted of a picture, or a short verbal description of the picture, "a picnic at the beach") and were asked to detect it. Results improved drastically, with performances of detection reaching 60 percent and 80 percent, respectively, at a pace of 125 and 250 msec per picture (Potter, 1976, Experiment 1) as compared with 12 percent and 30 percent in the control recognition memory task. Together with other experimental evidence (for reviews, see Intraub, 1999; Potter, 1999), the results obtained by Mary Potter demonstrated that during a rapid sequential visual presentation, the processing of a new picture might disrupt the consolidation in short-term memory of the previous picture. Within 100 msec, a picture is indeed instantly understood, and observers seem to comprehend a lot of visual information, but a delay of a few hundreds msec is required for the picture to be consolidated in memory. When consolidated, conceptual gist can be represented as a verbal description of the scene image, including that which was perceived and inferred.

B. Perceptual Gist

Ascertaining the perceptual content of the gist involves determining the image properties (e.g., spatial frequency, color, texture) that provide a structural representation of a scene (see, for example, Fig. 41.3). By manipulating the availability of these image properties (filtering out the edges of an image, for example), as well as imposing task constraints (e.g., duration of exposure, level of categorization, attentional demands), one can determine empirically the information needed to build a structural perceptual gist, and, therefore, enable scene identification.

Oliva and Schyns (1997, 2000; Schyns and Oliva, 1994) showed that a coarse description of the input scene (oriented blobs in a particular spatial organization at a resolution as low as 4 cycles per image) would initiate recognition before the identity of the objects was processed. Similarly, the structure of a scene can be assembled rapidly from a layout of parts (spatial arrangement of simple volumetric forms like *geons*; Biederman, 1995) or a coarse layout description of texture density (Figure 41.3; Torralba and Oliva, 2002). Common to these representations is their holistic nature: the structure of the scene is inferred, with no need to represent the shape or meaning of the objects.

In a series of articles, Oliva and collaborators (Oliva and Schyns, 1997, 2000; Schyns and Oliva, 1994, 1999; Oliva and Torralba, 2001; Torralba and Oliva, 2002, 2003) studied the image properties that enable an efficient categorization of a real-world scene. These properties included spatial frequency orientations and scales, color, and texture density. In their original study, Schyns and Oliva (1994) aimed to determine the role of spatial frequency scales for rapid scene categorization tasks. They contrasted information from blobs and edges by presenting participants with ambiguous visual stimuli (termed *hybrids*) combining the low spatial frequency (LSF) of one image with the high spatial frequency (HSF) of another. Figure 41.1 illustrates complementary hybrid stimuli: low spatial scale conveys information related to the spatial arrangement of oriented blobs, while high spatial scale conveys details, surfaces, and object contours. Two experiments revealed that hybrids were preferentially categorized in a *coarse-before-fine* sequence. In a categorization task, participants were briefly shown a hybrid scene (prime image, e.g., Fig. 41.1*a*) and asked to match the picture with a subsequent normal image (target). If the prime were the hybrid of Fig. 41.1*a*, the subsequent target image could be a city scene, matching the LSF components of the hybrid, or a hallway scene, matching the HSF components. Brief (30 msec) presentations of prime hybrids elicited matching based on their coarse structures (*city* in Fig. 41.1*a*). Longer (150 msec) presentation of the same hybrid elicited the opposite matching based on fine structures (*hallway* in Fig. 41.1*a*). This effect was reproduced in a categorization task in which an animated sequence of two hybrids (Fig. 41.1a–b) was preferentially categorized according to a *coarse-to-fine* sequence (*city*, 67% of coarse-to-fine interpretation), although the animation simultaneously presented the *fine-to-coarse* sequence of another scene (*hallway*, 29% of fine-to-coarse interpretations).

Because different spatial frequency scales transmit diverse information about the scene image (blobs preferentially convey scene spatial layout information, and edges convey surfaces and density of texture regions), an identical image may be quickly perceived at a scale that would optimize the information required to resolve a specific task: a diagnostic scale. Consequently, the allocation of attention to a specific spatial scale might determine what type of visual information enters the perceptual gist. Indeed, related experiments with hybrid stimuli (Oliva and Schyns, 1997) demonstrated a mandatory registration of multiple spatial

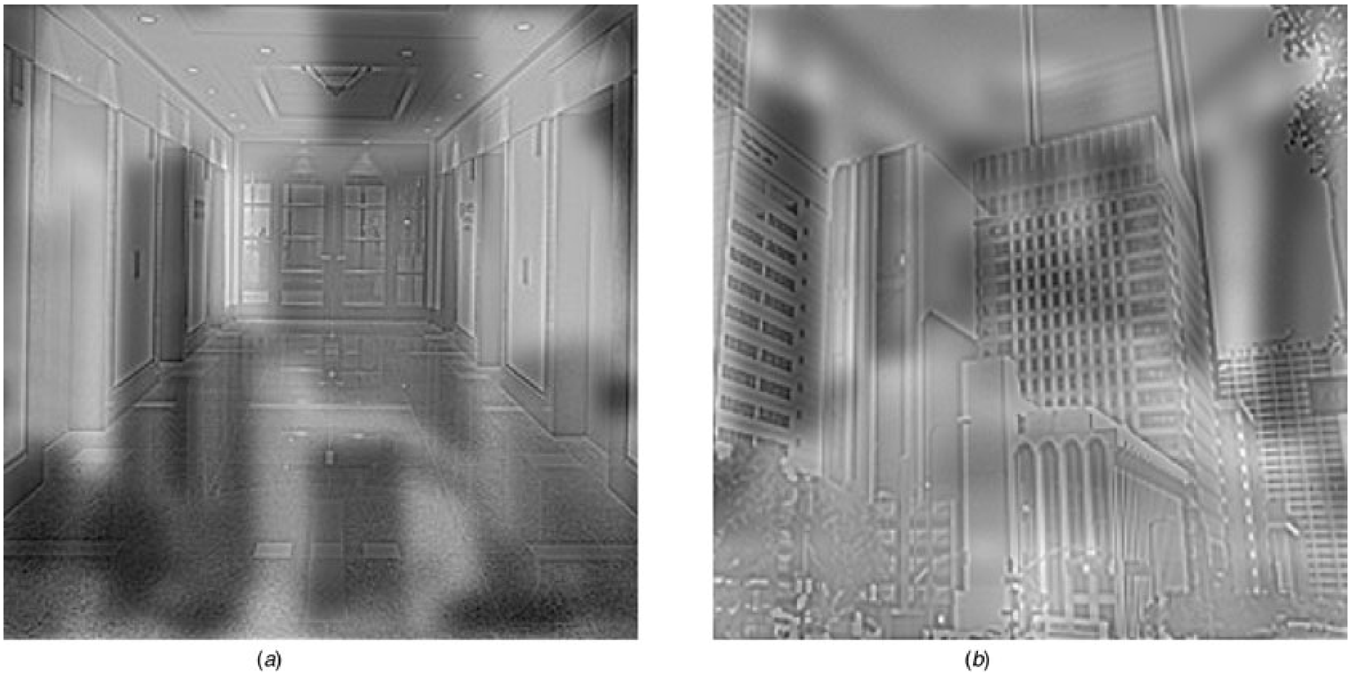

FIGURE 41.1 (*a*) A hybrid image representing a hallway scene in high spatial frequency (HSF, 24 cycles per image) and a city scene in low spatial frequency (LSF, 8 cycles per image). If you squint, blink, or defocus, the city scene should replace the hallway scene. (If this demonstration fails, step back from the image until your perception changes.) (*b*) The complementary hybrid scene, showing a city scene in HSF and the hallway scene in LSF (from Schyns and Oliva, 1994; Oliva and Schyns, 1997).

scales, but a flexible use of the information (blobs versus edges) proved most relevant to resolve a categorization task. In Experiment 2 of Oliva and Schyns (1997), the authors initially had two groups of observers (the LSF and HSF group) view, for 150 msec each, a set of hybrids that were only meaningful at one scale (either LSF, or HSF), the other scale being structured noise. The rationale was that these stimuli would sensitize categorization processes to seek scene cues on a diagnostic scale (either LSF, or HSF depending on the training). The testing phase consisted of presenting the two groups of observers with the same hybrid images, where the two scales were both meaningful (as in Fig. 41.1). With the observers unaware of the presence of two meaningful scenes in the hybrids, the two groups categorized the same images orthogonally: the group sensitized to the LSF scale categorized 73 percent of hybrids according to their LSF components (*city* in Fig. 41.1*a*), while HSF participants categorized 72 percent of the same stimuli on the basis of their HSF (*hallway* in Fig. 41.1*a*). These results demonstrated that attention could determine which information pertaining to spatial scale would be used during fast scene identification, leaving participants consciously blind to the unattended scale.

To complement these data, results from a third and a fourth experiment (Oliva and Schyns, 1997) suggested that, nevertheless, some covert processing of the neglected spatial frequency scale seemed, to have taken place: two scene representations may be simultaneously activated within the duration of a brief glance, with the covertly processed HSF scene influencing the overtly processed LSF scene.

In another study, Oliva and Schyns (2000) evaluated the role of color in express recognition of scene gist. They did a comparison between performances of fast categorization of scenes with their normal natural colors and scenes with their colors transformed (Fig. 41.2). A scene was presented for 150 msec, and participants were asked to identify it as quickly as they could (e.g., street, bedroom, forest). The rationale was that when the color of a surface is unrelated to the meaning of a scene (Fig. 41.2*a*), as is the case in most man-made environments, color information is not necessarily used by high-level cognitive processes. Consequently, color manipulation should not necessarily affect fast scene recognition. However, when color is a feature diagnostic of the meaning of a scene, as is the case in most natural environments, altering color information should impair recognition (Fig. 41.2*b*). Results demonstrated that color influences fast scene recognition when it is *diagnostic* of the scene category: the addition of normal color to a grey-level scene accelerates its recognition, whereas the additional of

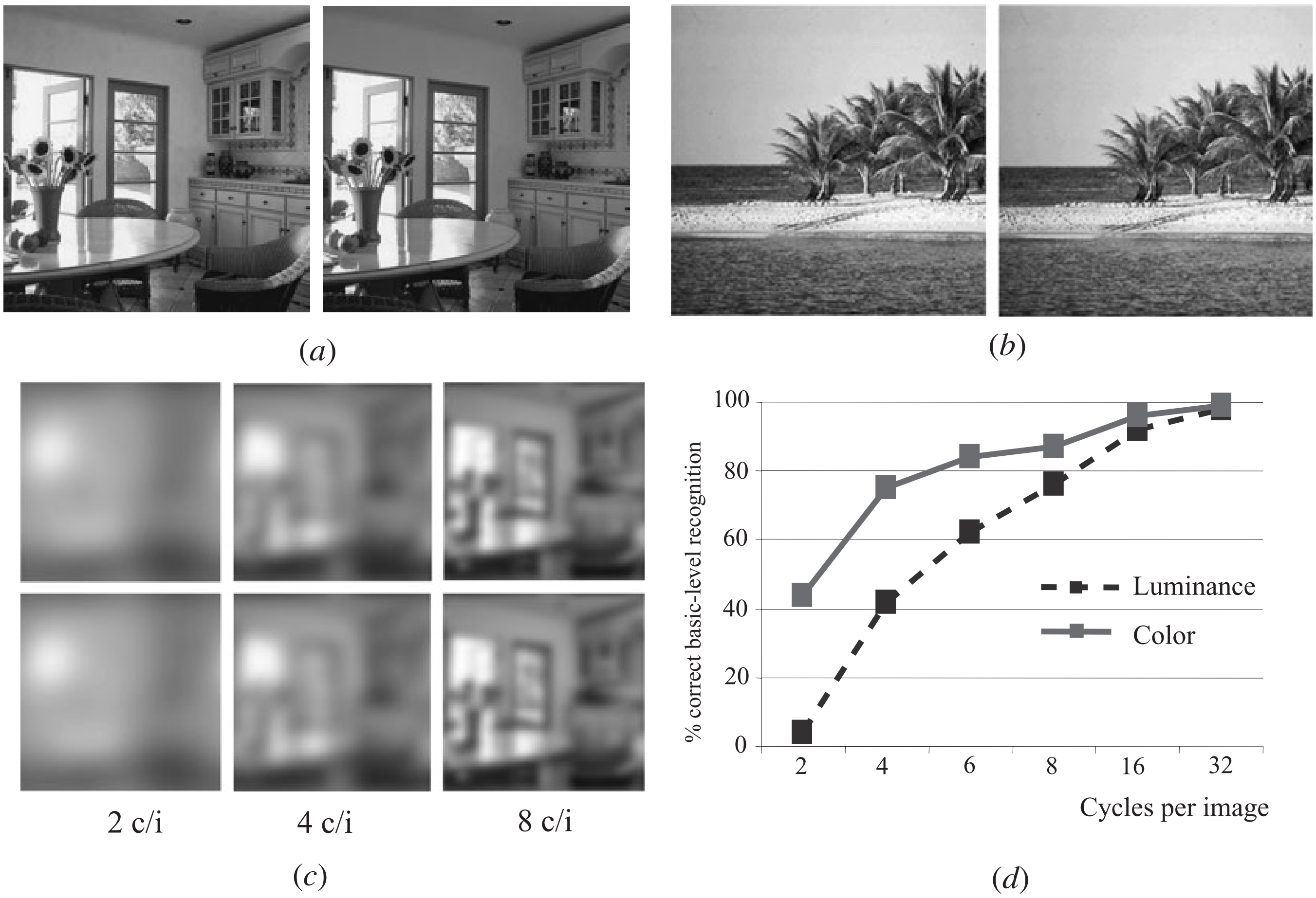

FIGURE 41.2 (*a*) Examples of color-nondiagnostic scenes: surface color is unrelated to meaning, (*b*) Examples of color-diagnostic scenes: surface color is related to the meaning of the object or region, (*c*) Illustration of color and luminance layout information available at 2, 4, and 8 cycles per image, (*d*) Performances of scene recognition as a function of cycles per image, for color and grey-level images (from Oliva and Schyns, 2000).

abnormal color impedes it. Additional experiments showed that the spatial layout of color blobs at a scale as coarse as 2 to 4 cycles per image (Fig. 41.2*c*) still mediates fast scene identification, while the grey-level counterparts do not provide enough structural cues for recognition before 6 to 8 cycles per image (Fig. 41.2*c* and *d*).

Together, these empirical results suggest that a reliable perceptual gist may be structured quickly based on coarse spatial scale information (from 4 to 8 cycles per image). At this resolution, enough structural cues are provided to allow the identification of the scene, although the local identity of objects may not be recovered. The flexible usage of spatial scale during scene perception encourages a soft-wiring view of the selection of image properties, with task constraints guiding the attentional selection of spatial scales. If a scene is unknown and must be categorized very quickly, highly salient, though uncertain, information may be more efficient for an initial rough estimate of the scene's gist. However, if one already knows what the content of the scene might be before seeing it, the informational constraints of a fast verification task may lead to a selection of fine edges before coarse blobs (Schyns and Oliva, 1994). By first attending to the coarse scale, the visual system can quickly get a rough estimate of the input to activate the conceptual part of the gist (scene schemas in memory). Attending to the fine information allows binding together local contours providing a refinement or refutation of the raw estimate. This view encourages a holistic mechanism during the perceptual gist elaboration that may be independent of object localization and identification (Oliva and Torralba, 2001).

III. A HOLISTIC REPRESENTATION OF GIST

In order to satisfy the main requirement of the gist—to deliver structural summary that is meaningful enough to permit image identification—most proposals of gist content have included spatial layout information. For instance, an estimation of the spatial layout of a scene may be built upon volumetric forms termed *geons* (Biederman, 1987, 1995), spatial arrangements of blobs of different contrasts and colors (Schyns and Oliva, 1994; Oliva and Schyns, 2000), or a representation of the coarse layout of principal contours and texture density (Oliva and Torralba, 2001, Fig. 41.3). Any description related to the diagnostic structure of a scene category (e.g., urban zones are vertically structured, forests are textured) is a likely candidate for perceptual gist.

One prominent view of scene recognition is based on the idea that a scene is built as a collection of objects. This notion has been influenced by seminal approaches in computational vision that have depicted visual processing as a hierarchical organization of modules of increasing complexity (edges, surfaces, objects), with the highest level, object identification, eventually initiating scene schema activation (Marr, 1982). However, empirical results suggest that a scene may be initially processed as a single entity and that segmentation of the scene in objects operates at a later stage during gist formation. For example, speed and accuracy in scene recognition are not affected by the quantity of objects in a scene (for a review, see Biederman, 1995), and recognition can be achieved equally well even when object information is degraded so much that objects cannot be locally recovered (Schyns and Oliva, 1994). If a scene is initially processed as a single entity, then what is the nature of this entity? An alternative approach to gist representation (Oliva and Torralba, 2001) takes advantage of the regularities found in the statistical distribution of image properties when considering a specific scene category (e.g., a highway must afford speed, so ground is a flat surface stretching to the horizon). Along these lines, perceptual and conceptual representations of gist could be initiated without processing object information. To this end, Oliva and Torralba (2001; Torralba and Oliva, 2002, 2003) reasoned that, since a scene is arranged in three-dimensional space, fast recognition could be based on image properties that are diagnostic of the space the scene subtends. The authors found that eight perceptual dimensions capture most of the three-dimensional structures of real-world scenes (naturalness, openness, perspective or expansion, size or roughness, ruggedness, mean depth, symmetry, and complexity). They observed that scenes with similar perceptual dimensions shared the same semantic category. In particular, scenes given the same basic-level name (e.g., street, beach) by observers tend to cluster within the same region of a multidimensional space in which the axes are the perceptual properties.

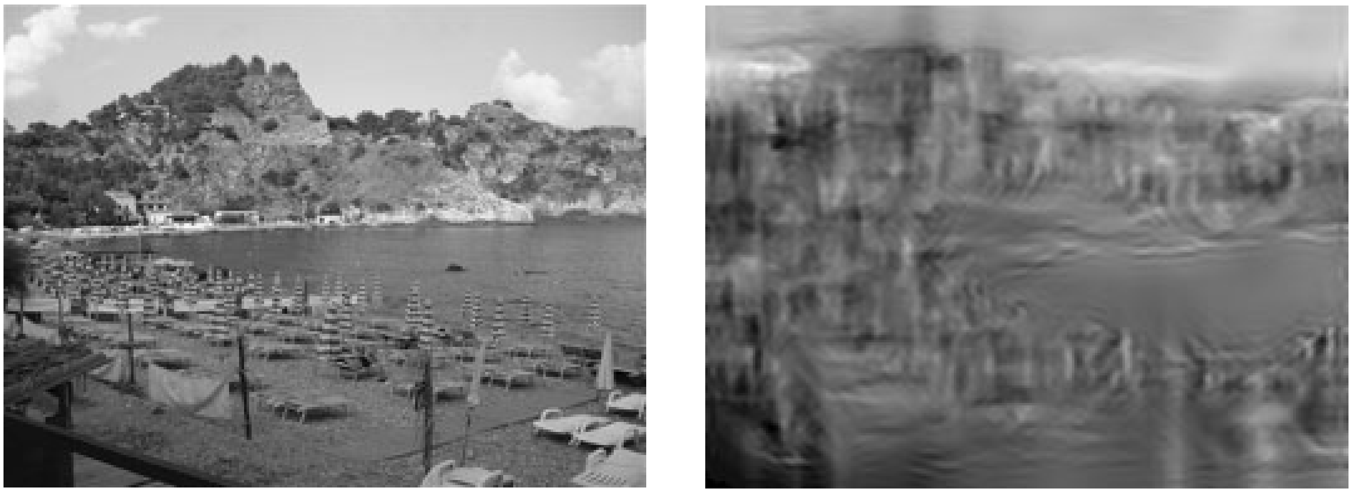

FIGURE 41.3 Illustration of a scene gist representation that conserves sufficient structural cues to infer the probable category of the scene. This global scene representation is used to determine spatial envelope properties of a scene. The information preserved by this global representation is illustrated on the right-hand image: it represents a sketch version of the original scene, computed by coercing noise images to have the same global features as the left scene (Torralba and Oliva, 2002). The scene sketch corresponds to an "unbound" spatial layout representation of contours, texture density and colors in the original scene picture.

The information about these perceptual dimensions can be extracted from the image using simple combinations of linear filters of the sort thought to exist in the early visual system. All together, the spatial perceptual dimensions form the *spatial envelope* of a scene—a holistic representation that does not require the use of objects as an intermediate representation and, therefore, is not being based on stages of region and object segmentation. For example, the spatial envelope of a forest scene could be described as "an enclosed natural environment, in the range of 100 meters, vertically structured in the background (trees) and connected to a textured horizontal and rough surface (grass)."

IV. CONCLUSION

In the attempt to explain how the brain represents the gist of a scene, the part-based approach (Marr, 1982; Biederman, 1987) depicts access to scene meaning as the last step within a hierarchical organization of modules of visual processing with increasing complexity (edges, surfaces, objects, scene). The "geon" theory put forth by Irving Biederman (1987, 1995) suggests that fast scene understanding could be achieved via a representation of the arrangement of simple volumetric forms from which the identity of the individual objects and scenes can be inferred. Alternatively, a holistic-based approach (*spatial envelope* theory; see Oliva and Torralba, 2001) constructs a meaningful representation of scene gist directly from the low-level features pool, without binding contours to form surfaces, and surfaces to form objects. Both approaches are based on the biological evidence that mechanisms of visual encoding are grounded in a multiscale and multiorientation representation. They differ in the number of steps, modules and binding procedures in order to achieve high-level scene categorization.

Models of the mandatory role of scene gist in object detection tasks (see Chapter 96) and its competence to predict the zones where attention might be allocated during object search have been proposed in the computational domain (Oliva, Torralba, Castelhano, and Henderson, 2003; Torralba, 2003). However, whether the neural correlate of the gist of a visual scene is a final product representation hosted in a dedicated cortical region (Epstein and Kanwisher, 1998), or a representation already formed in earlier cortical areas, or even a distributed topological representation at several cortical and functional levels—all are intriguing possibilities that remain to be explored.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. *Psychological Review* **94**, 115–148.

Biederman, I. (1995). Visual object recognition. *In* "An Invitation to Cognitive Science: Visual Cognition" (2nd edition). (M. Kosslyn and D. N. Osherson, Eds.), vol. **2**, 121–165.

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

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

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: implications for visual scanning and memory. *Fleeting Memories: Cognition of Brief Visual Stimuli*, V. Coltheart (ed.), 47–70.

Marr, D. (1982). *Vision*. W. H., Freeman, San Francisco, CA.

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

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

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

Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. (2003). Top-down control of visual attention in object detection. *Proceedings of the IEEE International Conference on Image Processing*, Vol. I, 253–256.

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

Potter, M. C. (1999). Understanding sentences and scenes: the role of conceptual short-term memory. *In* "Fleeting Memories: Cognition of Brief Visual Stimuli," (V. Coltheart, ed.), 13–46.

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

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

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

Torralba, A., and Oliva, A. (2003). Statistics of natural images categories. *Network: Computation in Neural Systems* **14**, 391–412.

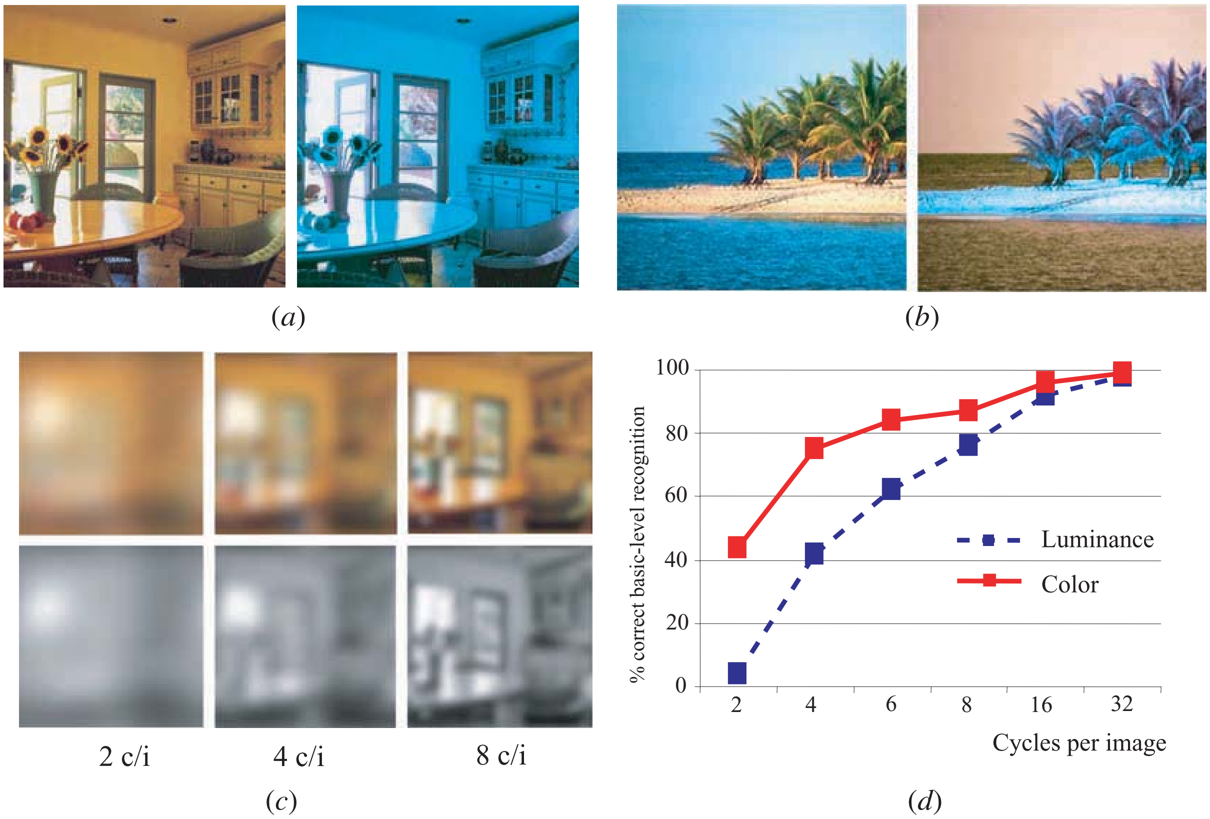

FIGURE 41.2 (*a*) Examples of color-nondiagnostic scenes: surface color is unrelated to meaning, (*b*) Examples of color-diagnostic scenes: surface color is related to the meaning of the object or region, (*c*) Illustration of color and luminance layout information available at 2, 4, and 8 cycles per image, (*d*) Performances of scene recognition as a function of cycles per image, for color and grey-level images (from Oliva and Schyns, 2000). (see color plate)

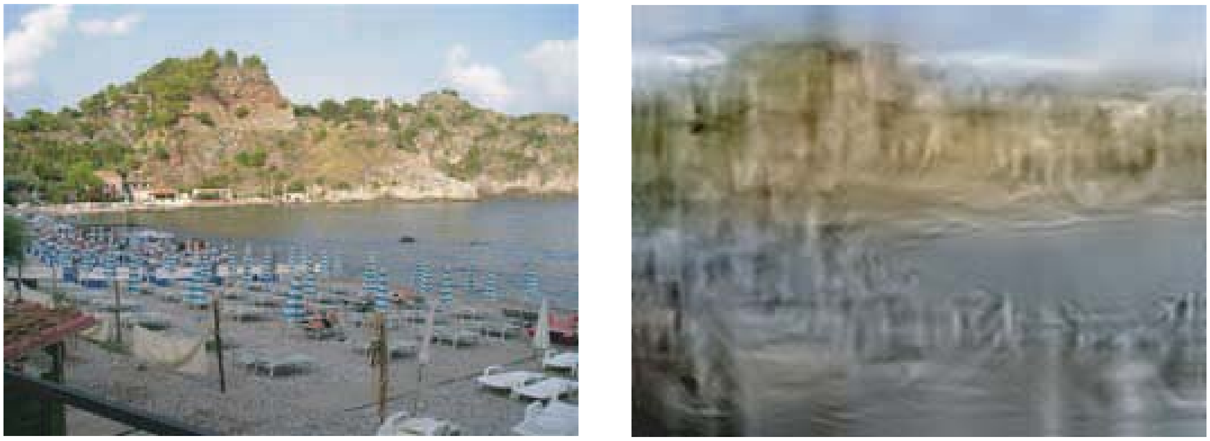

FIGURE 41.3 Illustration of a scene gist representation that conserves sufficient structural cues to infer the probable category of the scene. This global scene representation is used to determine spatial envelope properties of a scene. The information preserved by this global representation is illustrated on the right-hand image: it represents a sketch version of the original scene, computed by coercing noise images to have the same global features as the left scene (Torralba and Oliva, 2002). The scene sketch corresponds to an "unbound" spatial layout representation of contours, texture density and colors in the original scene picture.