Advanced Review

Rethinking visual scene perception

Helene Intraub*

A classic puzzle in understanding visual scene perception is how to reconcile the physiological constraints of vision with the phenomenology of seeing. Vision captures information via discrete eye fixations, interrupted by saccadic suppression, and limited by retinal inhomogeneity. Yet scenes are effortlessly perceived as coherent, continuous, and meaningful. Two conceptualizations of scene representation will be contrasted. The traditional visual-cognitive model casts visual scene representation as an imperfect reflection of the visual sensory input alone. By contrast, a new multisource model casts visual scene representation in terms of an egocentric spatial framework that is 'filled-in' by visual sensory input, but also by amodal perception, and by expectations and by constraints derived from rapid-scene classification and object-to-context associations. Together, these nonvisual sources serve to 'simulate' a likely surrounding scene that the visual input only partially reveals. Pros and cons of these alternative views will be discussed. © 2011 John Wiley & Sons, Ltd.

How to cite this article:
WIREs Cogn Sci 2012, 3:117–127. doi: 10.1002/wcs.149

INTRODUCTION

The visual field can encompass the soaring vista of a rocky seascape or the small, intimate world of one's top desk drawer during the search for a pen. Regardless of the complexity and distance of a given view, a fundamental puzzle for vision scientists has been to reconcile the observer's visual experience with the constraints that physiology places on visual sensory input.[1,2] In spite of the observer's sense of a high fidelity world, the best visual acuity is limited to the foveal region (about 2° of visual angle), and drops precipitously the farther from the fovea light falls. To capture visual details, eye movements shift the position of the fovea as quickly as three to four times per second. While the eyes are in motion, vision is suppressed (*saccadic suppression*)[3] creating a rapid alternation between visual sensory perception and very short-term memory as the eyes move. In addition, the visual field is spatially limited, requiring frequent changes in head position to bring previously invisible areas of a surrounding scene into view. A classic question is how this rapid succession of discrete, inhomogeneous views supports perception of a coherent, continuous, and meaningful scene.

This puzzle has motivated a large body of research that explores scene perception at the level of a single fixation. In these experiments, pictures (photographs, line drawings, and computer-generated images; see Ref 4 for comparisons) serve as surrogates for views of the world. Pictures are frequently presented for brief durations that prevent more than a single fixation,[5,6] but also for longer intervals in combination with eye tracking, to gain insight into memory across successive fixations.[7–10] Three fundamental questions have guided research: (1) how quickly can the observer grasp the general meaning of a scene? (2) what details can be remembered across fixations? and (3) what is the nature of the ensuing representation? The purpose of this paper is to provide a review of this research and to consider the results in the light of two different frameworks or 'models' for characterizing visual scene representation: the *traditional visual-cognitive model* and *the multisource model*.

The models differ in terms of what is considered to be the source of scene information, the fundamental structure of scene representation, and the relationship of visual and conceptual information in memory. According to the visual-cognitive perspective, visual scene representation has a single source—the visual input. This input can be rapidly identified, making contact with abstract, semantic knowledge, but the representation of a studied view and the conceptual knowledge it evokes are distinct representations.

*Correspondence to: intraub@udel.edu

Department of Psychology, University of Delaware, Newark, DE, USA

Integration of successive views is no longer thought to rely on visual sensory integration but instead on a remembered representation that is somewhat schematic in nature.[1,11] By contrast, the *multisource model*[12,13] takes an approach to scene perception that bears similarities to the concept of situated perception (also referred to as grounded cognition).[14,15] The fundamental structure of scene perception here is not visual, but spatial. It is the observer's egocentric framework of space. Multiple sources of input 'fill-in' this framework: visual sensory input, amodal perception beyond the view boundaries,[16–18] and expectations (and constraints) about the broader surrounding layout provided by rapid-scene classification[19,20] and object recognition.[21] Scene representation is conceptualized as a 'simulation' of the world that a given view only partially reveals. Rather than being built up from the integration of successive views over time, the first fixation on a scene elicits a surrounding multisource representation that can be refined and updated by subsequent fixations.

THE TRADITIONAL VISUAL-COGNITIVE MODEL

The traditional visual-cognitive model frames the problem of scene perception in terms of the fate of visual information at different stages of processing. Various very short-term buffers have been proposed that maintain the fragile, easily disrupted representation of a view within time windows as brief as a few tens of milliseconds to a few seconds following stimulus offset. Because transient disruptions to visual input occur throughout normal oculomotor activity (e.g., saccades and eye blinks[22,23]), there is interest in understanding the interfering effects on retention caused by the onset of new information (whether another picture or a visual noise mask). Different instantiations of the visual-cognitive model have focused on different very short-term buffers, depending on the nature of the questions being asked and the type of tasks being administered. It is generally understood that in many cases the proposed stores may not actually be distinct entities but different time slices in an unfolding process. Bearing this caveat in mind, five very short-term buffers have frequently been addressed across studies.

The first is the visual sensory register (also referred to as 'iconic memory').[24] Although most research on the sensory register involves displays of letters or numbers, Loftus, Johnson, and Shimamura[25] demonstrated sensory memory for brief, unmasked photographs (usually 100 ms or less). If a briefly presented picture is masked, sensory memory is curtailed and information may be retained in *transsaccadic memory* (for the duration of an eye movement,[11,26] *visual short-term memory* (VSTM: for a couple of seconds),[27,28] *conceptual short-term memory* (CSTM: initiated about 100 ms after onset)[6,29] and finally, information can be maintained for a longer time in visual working memory (VWM)[30,31] to support ongoing tasks. All are considered to be limited-capacity buffers that place a bottleneck on processing. Depending on a variety of factors, information maintained in these buffers will either be retained in a more stable form (long-term memory) or will be lost.

Several lines of research have sought to determine how much of the original visual information can be retained at different times during processing. It is generally agreed that memory for a view, even after many fixations is not 'picture perfect' and to some extent must be schematic in nature.[11,32] Debate has centered on how much detail can survive an eye movement from one location to the next. Research on change blindness[33–35] led many to suggest that scene representation may be fairly sparse.[1,33,36] In contrast, detail description and memory for left–right orientation during rapid serial visual presentation (RSVP) of photographs,[37,38] memory across eye fixations made on longer duration pictures[7,8] and recognition memory tasks[39] suggest that specific visual details are often retained. There is, as yet, no agreed upon metric for characterizing levels of detail (a very dicey problem), and as a result, the balance between schematic versus detailed parts of the representation is not well specified.

As we will see in the next section, the bottleneck in processing appears to impact memory (how much can be retained), rather than perception of meaning. Many lines of research have demonstrated that identification of a view (e.g., 'outdoor scene', 'river scene') occurs very rapidly—likely within the first fixation. Access to concepts in the visual-cognitive model plays an important role in processing, not only by allowing the viewer to rapidly grasp the meaning of the input, but also by influencing the placement of subsequent fixations[40] and perhaps even influencing the speed or accuracy of object identification.[41,42] Within the visual-cognitive framework, conceptual knowledge is elicited by visual information but is itself an independent type of representation as suggested in Figure 1. This separation of images and their associated abstract concepts is common in models of high-level cognition[15] and picture perception.[43] As we will see later, the multisource model takes a different view of the relationship between visual information and conceptual knowledge.

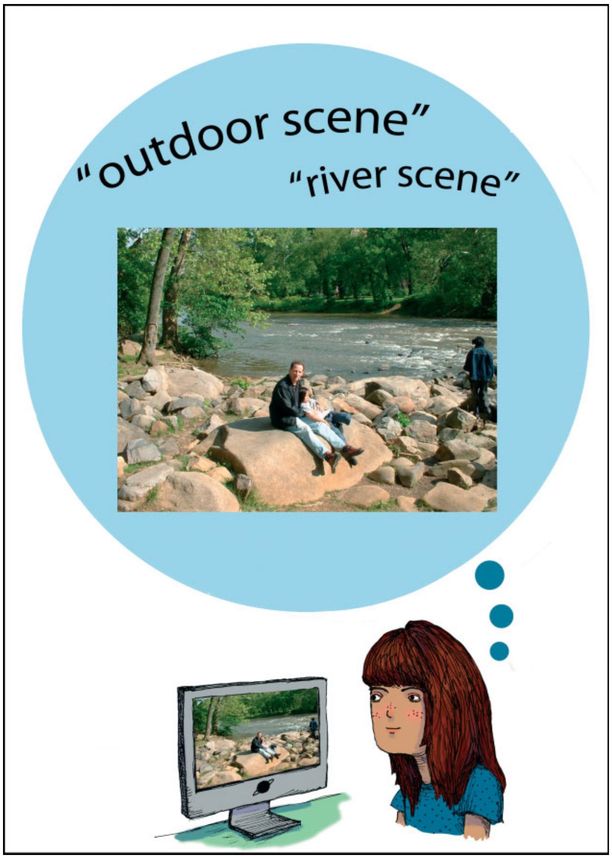

FIGURE 1 | An illustration of the visual representation of a picture and the associated abstract concepts that classify its content. Artwork by Stevie French.

SCENE CLASSIFICATION: UNDERSTANDING THE WORLD AT A GLANCE

Early research on scene perception provided what was a surprising insight at the time. Multiple fixations were not necessary for classifying a view. Biederman and colleagues[42,44,45] demonstrated that a complex view (in the form of a line drawing or photograph) presented for a brief duration that allowed no more than a single fixation, could be readily identified (e.g., classified as a 'street scene'). It was argued that the context (e.g., jumbled vs coherent; probable vs improbable object placement) could be grasped so rapidly, that it could affect the speed and accuracy of object identification. Subsequent research has both criticized[46] and supported[41,47] the idea that object identification is affected by context. However, the rapidity of scene identification (classification) has received consistent support. In early research using color photographs in RSVP tasks, pictures were presented at rates that mimicked (three to four pictures per second) or exceeded (e.g., nine per second) the fastest fixation frequency of the eye. Presentation of completely unrelated views at these rates is highly unnatural and eliminates the continuity and control of normal visual scanning, yet observers were remarkably good at identifying pictures based on a verbal title[29] (e.g., 'road with cars') or a single word that conveyed little or no specific visual information[37] (basic level, e.g., 'dog'; superordinate level, e.g., 'animal'). Detection was determined on the basis of reaction time[29] or a combination of reaction time and open-ended description (e.g., 'a reddish dog, facing right').[37] That observers could correctly detect and even describe pictures under these extreme conditions, suggested that they could do at least as well during normal free viewing of the world.

Potter[29] suggested that pictures could be identified within about 100 ms, and many subsequent studies have converged on this time frame[37,47–49] (for a comparison of different methodologies see Ref 50). Behavioral studies, however, are limited because of the long-time window between picture onset and the behavioral response. These studies demonstrate that a 100-ms exposure to a picture is sufficient to allow identification, but cannot provide insight into *when*, between stimulus onset and the response, classification had actually been achieved. Event-related potentials (ERPs) allow assessment of processing within a much tighter time window. A categorization task during RSVP revealed a divergence in the ERP signatures to pictures that matched or did not match a category as soon as 150 ms following picture onset, thus providing converging evidence that classification of a view could be completed with the first fixation on a scene.[51]

The speed with which views can be classified has led researchers to ask whether observers can identify more than a single view at a time—and do so, 'cost free'. Some studies have suggested that two to four simultaneously presented pictures can be identified without a cost.[52–54] Others have suggested that simultaneous identification has a cost (requires additional attention).[55,56] Resolution of this issue is complicated by multiple differences in stimulus sets and methods, and thus has yet to be resolved. However, whether or not there is a cost, in terms of the current discussion, these studies all demonstrate the rapidity of scene classification even under highly unnatural and demanding conditions. The next question of course is, 'How' classification is achieved. With most of the image falling well outside the small foveal region, what features of the visual information serve to support such rapid identification?

SCENE IDENTIFICATION: WHAT CHARACTERISTICS SUPPORT IDENTIFICATION?

Biederman[42] proposed that the layout of a view (rather than identification of individual objects) provides a rapid route to identification. Layout in his conceptualization referred to the relation of shapes to one another across an image. To illustrate he provided an outline drawing in which the objects (when seen in isolation) were meaningless geometric shapes; but when placed in relation to one another, created a recognizable layout (in this case a desktop) in which the geometric shapes were readily identified as a letter, a name plate, a book, and so forth. Oliva and Torralba[57] proposed that scene classification can be based on characteristics of global layout that require no specific object relations. An illustration of the power of spatial layout can be seen in Figure 2 which shows a picture discussed by Oliva and Torralba.[58] The filtered picture on the left is readily identified as a street scene. However, without filtering it becomes clear that the central region of the 'street scene' is actually part of a kitchen (cabinets and countertops). Although the picture is a mixture of indoor and outdoor scenes, the spatial layout of the picture as a whole conveys the sense of a street scene. Furthermore, Oliva and Torralba[57,58] proposed that global characteristics of layout that reflect a view as a whole (rather than reflecting a specific layout; e.g., trees on the left, river on the right) can support scene classification. They identified eight dimensions (e.g., 'openness', 'ruggedness', and 'mean depth') and argued that weightings across these dimensions create a profile, referred to as a 'spatial envelope' that is associated with specific superordinate categories (e.g., natural vs manmade). These spatial properties were computed directly from images using linear filtering techniques. Their computational model was successful in classifying views based upon spatial envelope.[57] In other research, global characteristics (e.g., mean depth) as well as ecological assessments (e.g., navigability) allowed a classifier model to achieve levels of categorization similar to human performance in classifying unfiltered photographs at the basic level (e.g., 'beach scene').[20]

This does not mean that object recognition plays no role in rapid-scene identification[41,47] or in refinements of that classification; object-to-context associations likely play an important role in scene cognition.[21] For example, specific object characteristics can provide information necessary for determining not only that something is a beach scene, but also based on an identifiable landmark—is a particular beach; or for determining not only that an image is a 'kitchen scene', but also whether or not it is a modern kitchen or older style kitchen. The importance of classification based on spatial envelopes, however, is that it provides an explanation of how, with most of an image falling outside of the high-acuity foveal region, enough information can be perceived within a single fixation to understand what one is looking at and use this to help guide subsequent fixations.[40]

MEMORY: HOW MUCH CAN BE RETAINED?

Does the 'deep' level of processing[59] that rapid conceptual identification confers ensure that a briefly glimpsed picture will be remembered? In the picture-detection experiments described in the previous section, although detection of a picture based on verbal cues was good, when the same RSVP sequences were presented to other participants followed by a recognition memory test, recognition memory was relatively poor, suggesting that more pictures could be momentarily identified than subsequently remembered.[29,37] The fragility of memory following

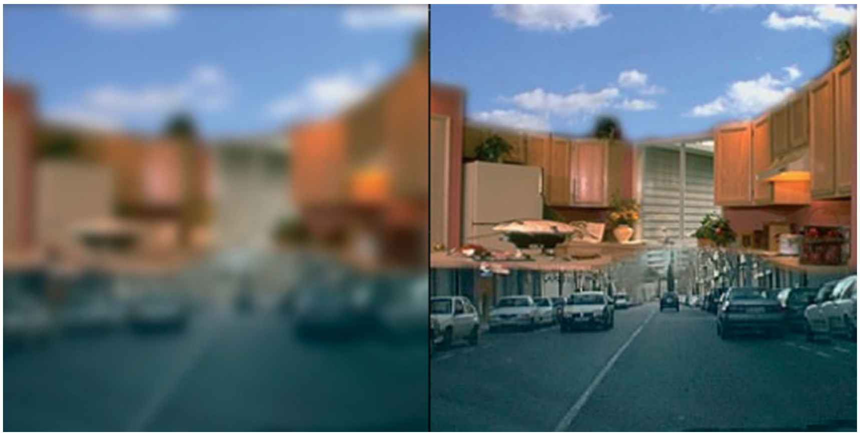

FIGURE 2 | The filtered picture on the left is interpreted as a 'street scene' based on the picture's spatial layout; the picture on the right is the unfiltered version and shows that rather than buildings, the central part of the picture is actually part of a kitchen scene (cabinets and counters). (Reprinted with permission from Ref 58. Copyright 2006 Elsevier)

RSVP could not be attributed solely to the brevity of each picture: when 150 pictures were presented for only 110 ms each, recognition memory increased from 21% correct when they were presented in a continuous RSVP sequence to 78% correct when a 1.5 s masked interval was interspersed between the pictures.

Potter[29] proposed that conceptual masking was responsible for poor memory following RSVP. An identified view could be maintained in a CSTM until it was interrupted by the onset of a new picture (i.e., a 'conceptual mask'). If the first picture was consolidated into long-term memory prior to the onset of a new picture, it would be remembered. If it was interrupted by the new picture before consolidation was complete, then the picture would be forgotten. Unlike visual masking, conceptual masking was caused by conceptual processing of a new picture. Several studies using a variety of designs have supported this distinction.[38,60–62] Potter[29] proposed that once a picture is consolidated in memory, it is no longer subject to conceptual masking.

Subsequent research, however, has demonstrated that conceptual masking does not inevitably occur upon the onset of a new picture; under some conditions, observers can voluntarily allocate attention within a stream of pictures to some pictures at the expense of others.[61] In other research, it was demonstrated that in a dual task situation, onset of a primary task that coincided with the onset of a picture in an RSVP sequence served to enhance memory for that picture.[63] More recent research by Potter et al.[64] has shown that a fragile representation of numerous pictures in a sequence can be maintained for a few seconds. Delaying an immediate recognition test by as little as 5.4 s resulted in a decrease in recognition memory performance. Thus, although memory is fragile during RSVP, consolidation is not determined solely by the presentation rate; several factors can affect the allocation of attention and thereby impact memory.

MEMORY FOR DETAILS WITHIN A VIEW: ERRORS OF OMISSION VERSUS ERRORS OF COMMISSION

In addition to overall recognition memory, there has long been interest in memory for the details within pictures. Traditionally this was studied in long-term memory paradigms. However, the phenomenon of *change blindness*,[33,34] brought widespread attention to the question of detail retention, even over very brief intervals. For example, Rensink et al's[35] flicker paradigm demonstrated that sizeable changes to an object (or region) of a picture could be missed across surprisingly brief intervals. In their study, an original photograph and a changed version of the photograph were repeatedly presented in the same location for 240 ms each with only an 80-ms blank field interspersed between them. Observers were strikingly poor at detecting the repeated change; in the most extreme case they required on average 80 alternations (50 s). However, if the to-be-changed object was named in advance, the change was relatively easy to report (usually within five alternations). Rensink et al. interpreted their results as demonstrating that very little detail can be retained across a brief transient (similar to a saccade) and that unless one happens to attend to that detail (as when it is named in advance) the change will be missed.

These errors of omission were surprising given such a brief retention interval between stimuli, but could be explained in terms of the limited capacity of early memory.[26] Alternative interpretations of this and other change blindness studies have been offered, and suggest that although some errors may reflect a failure to store information, many errors may reflect limitations on the observer's ability to make rapid comparisons across complex stimuli.[8,65] Although change blindness is likely caused by multiple factors, the point I would like to make here is that these alternative explanations do not pose problems for the visual-cognitive model as they can be explained in terms of limits of attention and the limited capacity of the early buffers. By contrast, a rapidly occurring error of *commission* in which observers remember seeing information that had not been present in the stimulus (e.g., objects, sections of the background, additional details) would pose a challenge.

Errors of commission in picture memory have long been the subject of study, but are usually studied in the context of a heavy memory loads, misleading information, or confusing stimulus sets.[66] Constructive errors are not expected across brief intervals when observers are explicitly trying to remember pictures. However, recent research has demonstrated that one such error, *boundary extension*,[67] can occur across retention intervals commensurate with a saccade.[13,68] Boundary extension is an error in which the observer confidently remembers having seen a greater expanse of a scene than was shown in a given view (an illustration is shown in Figure 3). What is interesting about boundary extension is that it appears to be an adaptive error in that anticipates upcoming layout (see Refs 69 and 70 for reviews). It occurs under conditions that would be expected to support good memory, such as low memory loads (as small as 1-3 pictures) and the presentation of no misleading or confusing information. In fact, the typical memory instruction encourages the observers to pay as much attention to the background as to the main objects in the view. Originally reported in long-term memory (following retention intervals of 35 min or 2 days),[67,71] subsequent research has demonstrated boundary extension when briefly presented pictures are disrupted by masked intervals lasting 1–2 s (consistent with VSTM)[72,73] or more surprisingly, lasting only 42 ms (consistent with transsaccadic memory).[13,68]

In the most extreme case, a single close-up view was interrupted by a 42-ms mask and then reappeared and remained on the screen.[13] Although the break in sensory input was less than 1/20th of second, observers tended to rate the identical view as being more 'close-up' than before, signifying that the pre-mask view had revealed more of the scene. Memory for an unseen expanse is challenging to explain within the context of the traditional visual-cognitive model. One would have to introduce a 'scene extrapolation' function to transsaccadic memory or VSTM; but the addition of this computational feature would seem to run counter to the limited-capacity nature of these buffers. One could propose a new buffer (the 'scene extrapolation' buffer), but this would provide no explanation. Instead, Intraub and Dickinson suggested an alternative account of scene representation, referred to as the *multisource model*.[12,13]

THE MULTISOURCE MODEL OF SCENE REPRESENTATION

According to the *multisource model*,[12,13] the fundamental structure underlying scene perception is an egocentric representation of space. In the world, observers are embedded within the scenes they perceive. Objects and spaces within a scene are 'in front of me', 'to the right or left of me', 'behind me', 'above me', and 'below me'.[74] When viewing a photograph, the observer takes the viewpoint of the camera. In typical snapshots, the view is usually 'in front' of the viewer, but the viewpoint can differ, e.g., in the case of aerial photographs, e.g., Zelinsky and Schmidt,[75] the view is 'below' the observer. Egocentric space, not the visual input, provides the framework for scene perception in the model and this framework supports information from multiple sources. The view of the trash cans against a fence shown in Figure 3 (perception panel) is understood as being 'in front' of the viewer. Visual information informs that location in the egocentric framework. At the edges of the view, amodal perception 'completes' the objects[16] and 'continues' the surfaces and textures[17,18] beyond the view boundaries. Thus, when viewing the picture, scene representation is more like the illustration shown in Figure 4. Without amodal perception, observers would interpret the photograph in Figure 3 (perception panel) as broken trash cans and a broken section of fence that just happens to line up with the edges of the view. Instead, the representation embeds the visual information within the larger expected context.

Rapid classification of a view within the first 100 ms or so provides a critical source of information. Once classified (e.g., 'outdoor scene') the category constrains expectations about the likely content of the surrounding space. In the case of the picture of the trash cans and fence, rapid identification of this as an outdoor scene elicits representation of the sky above, the ground below and more of the environment behind the 'viewer'. If the observer is familiar with these types of trash cans and this style of fence, object-to-context associations are likely to constrain this outdoor view further as a 'suburban outdoor scene'. Conceptual and contextual information both serve to 'fill-in' the egocentric structure with expectations about the surrounding world that is only partially revealed in the photograph. Comparison of Figures 3 and 4 (perception panels) illustrates the difference between the traditional visual-cognitive view of scene representation and the multisource view. In Figure 3, perception of the picture reflects only the visual-sensory input and its association with a concept. In Figure 4, perception of the picture elicits a representation of the observer's understanding of the surrounding scene (i.e., the view is situated within a larger spatial context). Conceptual knowledge is one of the sources of information that contributes to this 'simulation' of the likely content of surrounding space the visual input only partially reveals.

Just as the visual field itself is graded (from high acuity at the fovea to low acuity throughout the periphery), the fidelity of the multisource scene representation is graded. The visual sensory information shades into the amodally perceived region (which is tightly constrained by the visual information at the picture's edges), and then into the more schematic surrounding scene which is constrained based on its classification and object-to-context associations. For example, when individuals familiar with the photograph of the trash cans were asked to report (from memory) if they had a sense of where the camera was located and what was behind it, one observer reported that the camera was in a backyard with the owner's house behind it, another reported that the camera was in the street and a neighbor's house (across the street) was behind it, whereas a third stated that the camera was 'in an alley' and the 'fence on the other side of the alley' was behind the camera (this individual had lived in a region

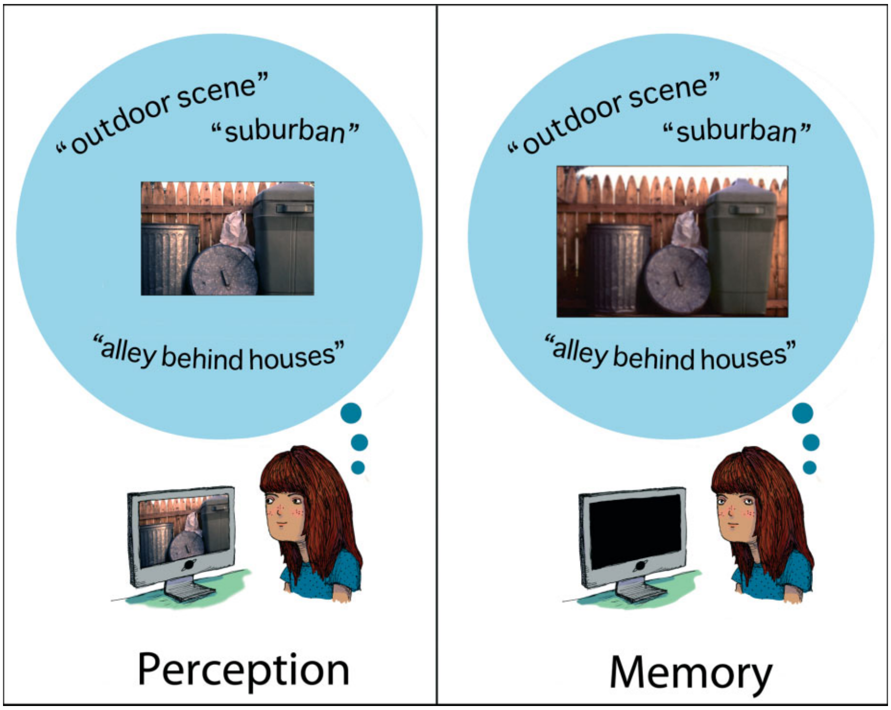

FIGURE 3 | Boundary extension in the context of the visual-cognitive model: while the picture is available (perception panel), visual input is the only source of scene representation. This representation is rapidly associated with relevant abstract concepts (shown in verbal form in this illustration). When the stimulus is removed (memory panel), the observer remembers having seen more beyond the edges of the view. Artwork by Stevie French.

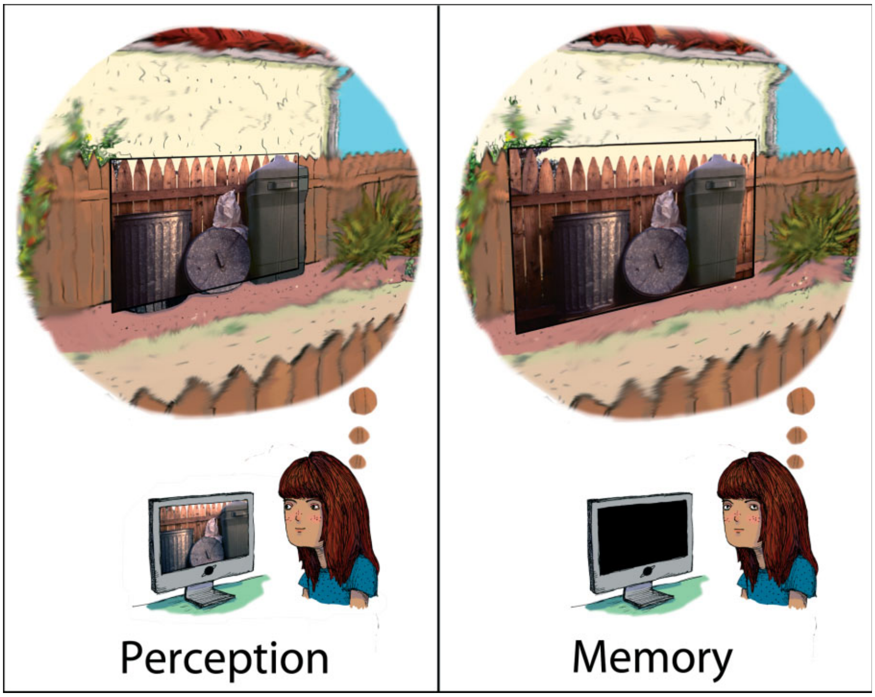

FIGURE 4 | Boundary extension in the context of the multisource model: the observer takes the camera's viewpoint (perception panel), in this case located in the center of the alley. With egocentric space as the organizing structure, visual sensory input, amodal perception beyond the boundaries of the photograph, conceptual knowledge and object-to-context expectations provide a multisource representation—including, in this case, the expectation of a fence behind the camera. When the stimulus is gone (memory panel), the observer misclassifies memory for the highly constrained amodal continuation just beyond the original boundaries as visual memory, thus causing boundary extension. Artwork by Stevie French.

of the United States in which trash cans are placed in alleyways behind suburban homes; her account is depicted in Figure 4).[12] All observers claimed that they had always thought of the scene this way. The anecdote suggests what Barsalou[14,15] has described as a 'simulation' in his theory of grounded cognition—in this case a simulation of the expected surrounding scene that the picture only partially reveals.

This characterization of scene perception provides a ready explanation for rapid boundary extension. While the visual sensory input is present (as in Figure 4, perception panel) the observer can clearly see where the visual input ends—there is no confusion between visual perception and amodal continuation beyond the view boundaries. However, once the visual sensory input is interrupted, even for as little as 1/20th of a second, the observer is left with a graded scene representation in memory, as shown in Figure 4 (memory panel). When asked to discern which part of that remembered representation reflects was had been visual, the observer is faced with a source-monitoring task[76] and misattributes memory for the highly constrained amodal region as having been seen before. Thus, the remembered boundaries are 'shifted' outward, resulting in boundary extension. In this view, the boundary-extended region is not extrapolated after the stimulus is gone—but was already part of the scene representation while the visual input was present (in the form of amodal perception). Thus, the same principles underlying source monitoring in long-term memory can explain boundary extension across a retention interval commensurate with a saccade.

Consistent with the source-monitoring hypothesis, factors that would tend to minimize memory for visual detail (and thus minimize the difference between the regions originally derived from visual and amodal sources), such as divided attention[77] or a reduction in stimulus duration[72] have led to increases in boundary extension. Supporting the idea that at its core the representation is spatial (not visual) is the observation of boundary extension in haptic exploration (without vision)[12,78] and the relation of boundary extension to activation in areas of the brain thought to be scene-selective in nature[79] [the parahippocampal place area (PPA) and retrosplenial cortex (RSC)[80,81]]. These areas have also been associated with navigation and placement of locations within larger geographic areas. The same spatial framework that underlies scene perception during viewing may underlie scene perception via other sensory modalities (e.g., haptic exploration).

The multisource approach provides an alternative way of thinking about the classic question of how we come to experience a rich coherent world in spite of the limitations on visual input. Here, the observer brings a strong egocentric framework to any viewing situation and the framework is 'filled-in' by different sources of information that become rapidly available during the first fixation on a scene. In the world, observers are embedded within the scenes they perceive. A graded representation based upon multiple sources would take advantage of rapid classification to provide expectations about parts of the world that are currently out of view. The representation would be updated on-line as the viewer makes successive fixations to different areas of the scene or moves through the environment, correcting and refining the representation. This approach to scene perception offers bridges to other areas of cognition. For example, it creates a bridge between research on early memory (milliseconds following stimulus offset) and studies of long-term memory, as shown in the example of source monitoring. Here without any change, the same source-monitoring model that addresses long-term memory errors can explain false memory over an interval lasting less than 1/20th of a second. It also suggests that some of the controversies regarding the representation of concepts and categories (often referred to as high-level cognition) may be applicable to scene perception. Because it is not a modality-specific model, but instead has at its core a spatial representation, it motivates questions about scene representation not only in the case of visual sensory input but in the case of other sensory modalities.

CONCLUSION

The visual-cognitive approach to scene perception has provided, and continues to provide important insights into the fate of visual information as we visually explore the world. Conceptual knowledge is accessed early and has been shown to play a role in guiding the placement of successive fixations within a view. Visual detail and associated conceptual knowledge are generally described as two different types of information; one concrete and visual-perceptual and the other abstract. The field has struggled with explaining the richness of our experience as we view scenes (given the physiological limitations on visual sensory input) and with addressing the underlying framework that supports view integration. Perhaps most important, visual scene perception is a cognitive activity that is always ongoing during our waking hours, and yet as a field, it has generally been isolated from other areas of cognition—considered as an independent field of study within the area of visual cognition.

The multisource model provides an alternative account that encompasses current knowledge about visual scene perception but places the central framework of scene representation squarely in the domain of spatial cognition. This motivates more general questions about scene representation, including perception across modalities and the relation of scene perception to navigation. It offers a different characterization of the visual/conceptual relationship that makes a potential connection with issues in high-level cognition (i.e., how concepts are represented) and provides a possible connection between scene perception and models of grounded cognition (situated perception). It opens a bridge between research on long-term memory and very short-term memory—as evidenced by the application of the source-monitoring model to boundary extension across a saccade. And finally it provides a new account of view integration that may better explain the coherence of scene perception. In this conceptualization, scene representation draws upon multiple sources of rapidly available information that are organized within the scaffolding of the observer's egocentric sense of space.

REFERENCES

1. O'Regan JK. Solving the "real" mysteries of visual perception: the world as an outside memory. *Can J Psychol* 1992, 46:461–488.
2. Rayner K. Eye movements and attention in reading, scene perception, and visual search. *Q J Exp Psychol (Colchester)* 2009, 62:1457–1506. doi:911217562.
3. Volkmann FC. Human visual suppression. *Vision Res* 1986, 26:1401–1416.
4. Tatler BW, Melcher D. Pictures in mind: initial encoding of object properties varies with the realism of the scene stimulus. *Perception* 2007, 36:1715–1729. doi:10.1068/p5592.
5. Intraub H. Understanding and remembering briefly glimpsed pictures: implications for visual scanning and memory. In: Coltheart V, ed. *Fleeting Memories: Cognition of Brief Visual Stimuli*. Massachusetts: MIT Press; 1999, 47–70.
6. Potter MC. Understanding sentences and scenes: the role of conceptual short-term memory. In: Coltheart V, ed. *Fleeting Memories: Cognition of Brief Visual Stimuli*. Massachusetts: MIT Press; 1999, 13–46.
7. Hollingworth A, Henderson JM. Accurate visual memory for previously attended objects in natural scenes. *J Exp Psychol Hum Percept Perform* 2002, 28:13–136. doi:10.1037/0096-1523.28.1.113.
8. Henderson JM, Hollingworth A. Eye movements and visual memory: detecting changes to saccade targets in scenes. *Percept Psychophys* 2003, 65:58–71.
9. Melcher D, Kowler E. Visual scene memory and the guidance of saccadic eye movements. *Vision Res* 2001, 41:3597–3611.
10. Zelinsky GJ, Loschky LC. Eye movements serialize memory for objects in scenes. *Percept Psychophys* 2005, 67:676–690.
11. Irwin DE. Information integration across saccadic eye movements. *Cogn Psychol* 1991, 23:420–56. doi:0010-0285(91)90015-G.
12. Intraub H. Rethinking scene perception: a multisource model. In: Ross B. ed. *Psychology of Learning and Motivation*, Vol 52. Burlington: Academic Press; 2010, 231–264.
13. Intraub H, Dickinson CA. False memory 1/20th of a second later: what the early onset of boundary extension reveals about perception. *Psychol Sci* 2008, 19:1007–1014. doi:PSCI2192.
14. Barsalou LW. Perceptual symbol systems. *Behav Brain Sci* 1999, 22:577–609.
15. Barsalou LW. Situated simulation in the human conceptual system. *Lang Cogn Process* 2003, 18:513–562.
16. Kanizsa G. *Organization in Vision*. New York: Praeger; 1979.
17. Fantoni C, Hilger JD, Gerbino W, Kellman PJ. Surface interpolation and 3D relatability. *J Vis* 2008, 8:1–19. doi:10.1167/8.7.29.
18. Yin C, Kellman PJ, Shipley TF. Surface integration influences depth discrimination. *Vision Res* 2000, 40:1969–1978. doi:S0042-6989(00)00047-X.
19. Greene MR, Oliva A. The briefest of glances: the time course of natural scene understanding. *Psychol Sci* 2009, 20:464–472.
20. Greene MR, Oliva A. Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cogn Psychol* 2009, 58:137–176. doi:S0010-0285(08)00045-5.
21. Bar M. Visual objects in context. *Nat Rev Neurosci* 2004, 5:617–629. doi:10.1038/nrn1476.
22. Deubel H, Bridgeman B, Schneider WX. Different effects of eyelid blinks and target blanking on saccadic suppression of displacement. *Percept Psychophys* 2004, 66:772–778.
23. Irwin DE. Eyeblinks and cognition. In: Coltheart V, ed. *Tutorials in Visual Cognition. Macquarie Monographs in Cognitive Science*. New York: Psychology Press; 2010, 121–141.
24. Sperling G. The information available in brief visual presentations. *Psychol Monogr: Gen Appl* 1960, 74.

25. Loftus GR, Johnson CA, Shimamura AP. How much is an icon worth? *J Exp Psychol Hum Percept Perform* 1985, 28:113–136.

26. Irwin DE. Perceiving an integrated visual world. In: Meyer DE, Kornblum S, eds. *Attention and Performance 14: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*. Cambridge, MA: The MIT Press; 1993, 121–142.

27. Phillips WA. On the distinction between sensory storage and short-term visual memory. *Percept Psychophys* 1974, 16:290–293.

28. Hollingworth A, Richard AM, Luck SJ. Understanding the function of visual short-term memory: transsaccadic memory, object correspondence, and gaze correction. *J Exp Psychol Gen* 2008, 137:163–181. doi:10.1037/0096-3445.137.1.163.

29. Potter MC. Short-term conceptual memory for pictures. *J Exp Psychol Hum Learning Mem* 1976, 2:509–522.

30. Hollingworth A, Luck SJ. The role of visual working memory (VWM) in the control of gaze during visual search. *Atten Percept Psychophys* 2009, 71:936–949. doi: 2008-01081-011.

31. Liu K, Jiang Y. Visual working memory for briefly presented scenes. *J Vis* 2005, 5:650–658. doi:10:1167/5.7.5/5/7/5/.

32. Intraub H. The representation of Visual Scenes. *Trends Cogn Sci* 1997, 1:217–221. doi:S1364-6613(97)01067-X.

33. Simons DJ, Levin DT. Change blindness. *Trends Cogn Sci* 1997, 1:261–267. doi:S1364-6613(97)01080-2.

34. Simons DJ, Rensink RA. Change blindness: past, present, and future. *Trends Cogn Sci* 2005, 9:16–20. doi:S1364-6613(04)00293-1.

35. Rensink RA, O'Regan JK, Clark JJ. To see or not to see: the need for attention to perceive changes in scenes. *Psychol Sci* 1997, 8:368–373.

36. Rensink RA. The dynamic representation of scenes. *Vis Cogn* 2000, 7:17–42.

37. Intraub H. Rapid conceptual identification of sequentially presented pictures. *J Exp Psychol Hum Percept Perform* 1981, 7:604–610.

38. Intraub H. Presentation rate and the representation of briefly glimpsed pictures in memory. *J Exp Psychol Hum Learn* 1980, 6:1–12.

39. Konkle T, Brady TF, Alvarez GA, Oliva A. Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol Sci* 2010, 21:1551–1556. Epub October 4, 2010. doi:0956797610385359.

40. Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol Rev* 2006, 113:766–786. doi:2006-12689-003.

41. Davenport JL. Consistency effects between objects in scenes. *Mem Cogn* 2007, 35:393–401.

42. Biederman I. On the semantics of a glance at a scene. In: Kubovy M, Pomerantz JR, eds. *Perceptual Organization*. Hillsdale, NJ: Erlbaum; 1981, 213–253.

43. Potter MC. Mundane symbolism: the relations among objects, names, and ideas. In: Smith NR, Franklin MB, eds. *Symbolic Functioning in Childhood*. Hillsdale, NJ: Erlbaum; 1979, 41–65.

44. Biederman I. Perceiving real-world scenes. *Science* 1972, 177:77–80.

45. Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: detecting and judging objects undergoing relational violations. *Cogn Psychol* 1982, 14:143–177.

46. Hollingworth A, Henderson JM. Does consistent scene context facilitate object perception? *J Exp Psychol Gen* 1998, 127:398–415.

47. Davenport JL, Potter MC. Scene consistency in object and background perception. *Psychol Sci* 2004, 15:559–564. doi:10.1111/j.0956-79762004.00719.x.

48. Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? *J Vis* 2007, 7:1–29. http://journalofvision.org/7/1/10/. doi:10.1167/7.1.10/7/1/10/.

49. Võ ML-H, Henderson JM. The time course of initial scene processing for eye movement guidance in natural scene search. *J Vis* 2010, 10:1–14. doi:10.1167/10.3.14/10/3/14/.

50. Bacon-Macé N, Kirchner H, Fabre-Thorpe M, Thorpe SJ. Effects of task requirements on rapid natural scene processing: from common sensory encoding to distinct decisional mechanisms. *J Exp Psychol Hum Percept Perform* 2007, 33:1013–1026. doi:2007-14662-002.

51. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *Nature* 1996, 381:520–522. doi:10.1038/381520a0.

52. Fei-Fei L, Van Rullen R, Koch C, Perona P. Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Vis Cogn* 2005, 12:893–924.

53. Li F-F, VanRullen R, Koch C, Perona P. Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci U S A* 2002, 99:9596–9601. doi:10.1073/pnas.092277599.

54. Rousselet GA, Fabre-Thorpe M, Thorpe SJ. Parallel processing in high-level categorization of natural images. *Nat Neurosci* 2002, 5:629–30. doi: 10.1038/nn866.

55. Potter MC, Fox LF. Detecting and remembering simultaneous pictures in RSVP. *J Exp Psychol Hum Percept Perform* 2009, 35:28–38. doi:2009-00768-017.

56. VanRullen R, Reddy L, Koch C. Visual search and dual-tasks reveal two distinct attentional resources. *J Cogn Neurosci* 2004, 16:4–14. doi:10.1162/089892904322755502.

57. Oliva A, Torralba A. Modeling the Shape of the scene: a holistic representation of the spatial envelope. *Int J Comp Vis* 2001, 42:145–175.

58. Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition. *Prog Brain Res* 2006, 155:23–36. doi:S0079-6123(06)55002-2.

59. Craik FIM, Lockhart RS. Levels of processing: a framework for memory research. *J Verb Learn Verb Behav* 1972, 11:671–684.

60. Loftus GR, Ginn M. Perceptual and conceptual processing of pictures. *J Exp Psychol Learn Mem Cogn* 1984, 10:435–441.

61. Intraub H. Conceptual masking: the effects of subsequent visual events on memory for pictures. *J Exp Psychol Learn Mem Cogn* 1984, 10:115–125.

62. Loschky LC, Sethi A, Simons DJ, Pydimarri TN, Ochs D, Corbeille JL. The importance of information localization in scene gist recognition. *J Exp Psychol Hum Percept Perform* 2007, 33:1431–1450. doi:72/2/427.

63. Lin JY, Pype AD, Murray SO, Boynton GM. Enhanced memory for scenes presented at behaviorally relevant points in time. *PLoS Biol* 2010, 8:000337. doi:10.1371/journal.pbio.1000337.

64. Potter MC, Staub A, Rado J, O'Connor DH. Recognition memory for briefly presented pictures: The time course of rapid forgetting. *J Exp Psychol Hum Percept Perform* 2002, 28:1163–1175. doi:10.1037/0096-1523.28.5.1163.

65. Hollingworth A. Visual memory for natural scenes: evidence from change detection and visual search. *Vis Cogn* 2006, 781–807.

66. Koriat A, Goldsmith M, Pansky A. Toward a psychology of memory accuracy. *Annu Rev Psychol* 2000, 51:481–537. doi:10.1146/annurev.psych.51.1.481.

67. Intraub H, Richardson M. Wide-angle memories of close-up scenes. *J Exp Psychol Learn Mem Cogn* 1989, 15:179–187.

68. Dickinson CA, Intraub H. Transsaccadic representation of layout: what is the time course of boundary extension? *J Exp Psychol Hum Percept Perform* 2008, 34:543–555. doi:71/6/1251.

69. Hubbard TL, Hutchison JL, Courtney JR. Boundary extension: findings and theories. *Quart J Exp Psychol* 2010, 63:1467–1494.

70. Michod KO, Intraub H. Boundary extension. *Scholarpedia* 2009, 4:3324.

71. Intraub H, Bender RS, Mangels JA. Looking at pictures but remembering scenes. *J Exp Psychol Learn Mem Cogn* 1992, 18:180–191.

72. Intraub H, Gottesman CV, Willey EV, Zuk IJ. Boundary extension for briefly glimpsed pictures: do common perceptual processes result in unexpected memory distortions? *J Mem Lang* 1996, 35:118–134.

73. Intraub H, Hoffman JE, Wetherhold CJ, Stoehs S. More than meets the eye. *Percept Psychophys* 2006, 5:759–769.

74. Tversky B. Spatial cognition: embodied and situated. In: Robbins P, Aydede M, eds. *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press; 2009, 201–216.

75. Zelinsky GJ, Schmidt J. An effect of referential scene constraint on search implies scene segmentation. *Vis Cogn* 2009, 17:1004–1028. doi:10.1016/j.visres.2009.05.017.

76. Johnson MK, Hashtroudi S, Lindsay DS. Source monitoring. *Psychol Bull* 1993, 114:3–28.

77. Intraub H, Daniels KK, Horowitz TS, Wolfe JM. Looking at scenes while searching for numbers: dividing attention multiplies space. *Percept Psychophys* 2008, 70:1337–1349. doi:70/7/1337.

78. Intraub H. Anticipatory spatial representation of 3D regions explored by sighted observers and a deaf-and-blind-observer. *Cognition* 2004, 94:19–37. doi:10.1016/j.cognition.2003.10.013.

79. Park S, Intraub H, Yi DJ, Widders D, Chun MM. Beyond the edges of a view: boundary extension in human scene-selective visual cortex. *Neuron* 2007, 54:335–342. doi:S0896-6273(07)00256-5.

80. Epstein R, Kanwisher N. A cortical representation of the local visual environment. *Nature* 1998, 392:598–601. doi:10.1038/33402.

81. Epstein RA, Higgins JS. Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cereb Cortex* 2007, 17:1680–1693. doi:bhl079.

FURTHER READING

Barsalou LW. Grounded cognition. *Annu Rev Psychol* 2008, 59: 617–645. doi:10.1146/annurev.psych.59.103006.093639.

Konkle T, Brady TF, Alvarez GA, Oliva A. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J Exp Psychol Gen* 2010, 139:558–578. doi: 10.1037/a0019165.