The representation of visual scenes

Helene Intraub

The visual world exists all around us, yet this information must be gleaned through a succession of eye fixations in which visual acuity is limited to the small featural region of each retina. In spite of these phyiological constraints, we experience a richly detailed and continuous visual world. Research on transaccadic memory, perception, picture memory and imagination of scenes will be reviewed. Consistent evidence suggests that the representation of visual scenes is much more schematic and abstract than our immediate experience would indicate. The visual system may have evolved to maximize comprehension of discrete views at the expense of representing unnecessary detail, but through the action of attention it allows the viewer to access detail when the need arises. This capability helps to maintain the "illusion" of seeing a rich and detailed visual world at every glance.

Visual information exists all around us, but physiological constraints prevent us from seeing it all at once. Input is relatively piecemeal as eye movements (called saccades) bring a different region of the world into view as rapidly as three times per second. Furthermore, even when the eye is fixating on an area, high resolution is limited to the relatively small area that falls on the fovea. Yet, in spite of these constraints on visual input, viewers claim to experience a clearly visible, detailed and continuous visual world. A classic question in perception has been how this limited input is capable of yielding such a rich visual experience.

Eye movements and scene perception

One traditional explanation of our visual experience is that sensory information from each fixation is integrated in a high capacity memory buffer. This information is essentially "linked together" into a detailed spatiotemporal representation of the environment that is maintained across saccades. As we will see, recent research on transaccadic memory has not supported this theoretical perspective. Instead, it supports a somewhat counterintuitive view that the representation of visual scenes is, in large part, abstract and schematic. As the eye briefly shifts from one location to another, the visual world is represented in the form of a schematic map of the scene's layout and major landmarks. Attended areas may be represented in detail, but areas outside the focus of attention are not.

Some of the most compelling evidence for an abstract transaccadic representation has been obtained in research on reading. In one study, readers viewed sentences in which the words were presented in ATEBrddBG CaSe on a computer screen. Eye movements were monitored, and during some saccades, the case of every letter was changed. Surprisingly, this did not disrupt the viewers' eye-movement patterns or ability to read. In fact, none of the viewers noticed anything unusual. Clearly information important to normal reading was retained and integrated across fixations, but it apparently did not include the actual visual characteristics of the letters. Similar results have been obtained using visual tasks other than reading.

Recently, photographs of scenes were presented using the "ansi paradigm." In this case, during some saccades, the scene was shifted horizontally or vertically by 0.3", 0.6" or 1.2" to or was expanded or contracted by either 10% or 20%. Eye movements were monitored, and subjects were required to indicate if they noticed a change. Small changes were more sensitive, suggesting that the observers were not using an independent metric of space to piece together successive views. Results suggested that detection of change relied more on local information in the region of the eye's landing position than on a detailed global representation of previously fixated areas.

Though somewhat surprising, given the observer's experience of a detailed visual world, these results fit well with the proposition that not only the transaccadic representation, but perception itself, is not uniformly detailed and concrete. Julian Hochberg has proposed that even during the time that an object is in view, all of its parts are not equally represented in our perceptual system. Specific local and global features of the visual world are fitted into a schematic map of the spatial layout. This schematic incorporates detailed attended information, vague global information (e.g., gross size and shape) and expectations about information just outside the field of view. An important aspect of this theory is that the representation is not a fixed and detailed model of the external environment, but a malleable one that continually changes as the viewer shifts attention.

He has supported this argument with many ingenious demonstrations, including the seemingly simple task shown
in Fig. 1 (see Refs 14, 17). Figure 1A shows a Necker cube, the well-known ambiguous figure that seems to shift perspective as the viewer gazes at it. The usual exploration focuses on the ambiguity of the visual stimulus, that allows the mind to switch between two equally plausible alternatives. In Fig. 1B, it is a nonambiguous version of a Necker cube, in which the interpretation at point 1 dictates a single orientation. According to a holistic model of perception, such a stimulus should not shift because when the whole object is seen, the interpretation will be seen as well. However, contrary to this prediction, if the viewer gazes at point 2, the cube will readily shift. In doing so, as the viewer can attest, point 1 is still seen, it too doesn't seem to exert any influence on the orientation. Hochberg argues that "unlike objects themselves, our perception of objects are not everywhere dense.\textsuperscript{17} That is, even when looking at the cube, it is not represented as a whole – an attended location is represented in a more concrete manner than a location that is not. This description of differences in density, is also captured in "sens descriptions of transaccadic memory and perception, that is the concept of a master map of locations and object files.\textsuperscript{16}

Applying this concept to dynamically changing scenes in motion pictures, Hochberg argues that it explains why viewers are poor at detecting the transient continuity errors in edited films.\textsuperscript{11} Recent research has experimentally tested and supported this observation. In one experiment, a continuity error was deliberately created in a videotape of two shots shown on a television, so that when the camera passed away from the subject, to focus on one actor for 4 seconds, a central item on the table (a large soda bottle that was conspicuously used) was replaced with a cardboard box. When the camera returned to the original view, remaining for 30 seconds, none of the viewers noticed that the bottle had become a box.\textsuperscript{21}

This inability to detect change was observed even when scenes were explicitly directed to do so, and an object repeatedly changed while the viewer watched a sequence of pictures.\textsuperscript{22} The changes were large and often dramatic, covering about 20 square degrees of visual angle, and including the deletion or relocation of an object, or a change in its color. Picture duration was 240 ms with an 80 ms blank interstimulus interval. The initial version was presented twice followed by two repetitions of the changed version, and the alteration pattern continued for as long as the subject needed. Subjects were remarkably poor at detecting the change – in the most difficult cases requiring over 80 alterations (more than 50 seconds) to do so. However, provision of a verbal cue indicating which object was likely to change led to detection within about five alterations or fewer. The changes were easy to see, as long as the viewer was directly attending to the critical location. Although the subjective experience is one of seeing the entire picture, performance suggests instead that at any given moment perception is not "everywhere dense".

Although it may seem odd that the visual system would yield a representation that is not 'photograph-like', upon reflection it becomes clear that such a system would be very economical. Viewers would not need to store the representation of the world from moment to moment, because whenever they needed to discern the details of a particular region, they could simply shift their eyes and look at it.\textsuperscript{16} The ability to readily do this would support the viewer's impression that the entire visual world is clearly visible at all times. The abstract nature of the representation would also allow for a seamless integration of information that is currently in the visual field with expectations about neighboring information that the next eye fixation is likely to bring into view. With this in mind, we will now consider the implications this has for the long-term representation of scenes, and for imagination.

Remembering and imagining scenes

Does the long-term representation of a scene have a schematic nature similar to what has been proposed for its short-term transaccadic counterpart, or do abstract schematic representations in some way contribute to creating a holistic memory? If the former is true, then there should be instances in which the long-term representation of a scene will reflect schematic expectations that were never actually viewed. Such cases need not be rare. For example, in one study, 24 views of the same city scene were presented: six views (that together yield a panoramic view) from each of four corners of an intersection.\textsuperscript{23} Subjects sorted the 24 pictures, indicating at which of the four locations (shown on a map) they believed the camera had been placed. They received feedback and were asked to view the first eight views for about 30 minutes. Subjects were then asked to sort the views from the source views in the original sorting task. This representation apparently contained information about the projected size, shapes and perspective of numerous landmarks as they would appear from viewpoints that had not been experienced.

Do these remembered expectations about a scene's layout require theacity of numerous views even relatively long periods of time to develop? Or as suggested by Hochberg's formulation, might a single view of a scene be enough to elicit expectations about areas that are not currently in view? Boundary extension, a spatial memory distinction for scenes, suggests that even a single view will evoke schematic expectations\textsuperscript{24,25}, and that it does so quite rapidly.
Boundary extension is a visual spatial memory distortion observed when observers are asked to remember single views of scene objects. Following presentation, they tend to remember seeing a greater expanse of the scene than was shown in the photograph. Figure 2 provides an example. Hodges' claims about schematic expectations of subjects' memory reflect not only what was physically present in the picture, but information that was understood to exist just outside the picture's boundaries. The phenomenon is evident, not only in drawings but in viewers' responses to recognition test pictures.

When the test picture is the same as the stimulus (target), subjects tend to reject it as 'old,' reporting instead that it shows a closer view than did the original. When viewing intruder pictures, subjects show an asymmetric response: wide-angle distractors are rated as looking more like the original picture than are close-up distractors, and wide-angle distractors are more frequently mistaken as being the original view. Consistent with the notion of spatial perception, both recognition and drawing tests show that boundary extension is greatest for close-ups (in which highly redundant information surrounding the amodal object is physically present), and becomes less apparent as pictures (view images) in which the expunged information is clearly shown in the picture are encountered. Wide-angle views often show no directional distortion of their boundaries.

Boundary extension suggests a seamless integration of remembered information physically presented in the picture and information that was inferred. This is consistent with the notion that a single, coherent schematic robustly underlies both types of information. Might this be the same representation that underlies trans-saccadic memory? If so, then boundary extension should be detectable very early in processing. In most research, however, picture durations were relatively long (e.g., 15 seconds) and extension intervals ranged from 3 minutes to 2 days. However, recent research supports this possibility in that boundary extension was obtained following stimulus durations as brief as a single eye fixation (e.g., 250 ms), and at rates of change that simulate rapid visual scanning (e.g., three pictures per second). In one experiment, on each of 72 trials, three complex color scenes were presented on a computer screen for 233 ms each in rapid succession, followed by a 1 s visual noise mask, and a repetition of one of the three pictures. Subjects tended to rate the repetition as looking 'closer-up' than before, indicating that boundary extension occurs very rapidly indeed.

If boundary extension is due to the activation of expectations evoked by a partial view of a continuous scene, then it should not occur for stimuli that do not present a partial view. To test this possible constraint, memory for spatial extent was tested for close-up and wide-angle pictures that clearly depicted part of a continuous world (photographs of scenes, and line drawings of the same scenes) and pictures that did not (line drawings of the main objects from each scene on a blank background). Consistent with the hypothesis, photographs and scenes and outline-scene yielded the boundary extension pattern described earlier. Close-ups yielded boundary extension, wide-angle views yielded no directional distortion, and responses to distractors yielded the expected response asymmetry. However, pictures of outline-objects did not yield a directional bias. Responses to close-up and wide-angle targets were more similar than were responses to line drawings of the same objects.
were symmetrical. Large objects (close-ups) were remembered as slightly smaller, and small objects (wide-angle views) were remembered as slightly larger (see Fig. 4). Responses to the distractors were also symmetrical. When the background was blank, memory showed evidence of a 'regression to the mean' in terms of object size, rather than a unidirectional distortion of the picture's boundaries to include 'more of the scene'. This suggests that schema activation requires the expectation of a continuous background that is understood to 'exist' just outside a given view. In a sense, activation of expectations beyond a picture's boundaries is similar to imagining the spatial layout of the scene from which the picture was taken. This raises the possibility that the same abstract mental schema that may underlie perception and memory for scenes may underlie imagination of scenes as well. Behavioral and neurophysiological evidence has supported the theory that perception and imagination share some of the same mental structures. In support of this position, the same pictures of outline-objects did yield the typical pattern of boundary errors associated with scenes, when specific 'image-scene' instructions were added to the standard instruction 16.

In this case, while memorizing the size of each outline-object, subjects were read a verbal description of the background from each associated photograph and were asked to imagine a white screen in front of it. Responses to the close-up and wide-angle ranges and distortions changed dramatically, yielding the same unidirectional pattern of errors as had the scenes (see Fig. 4). A control condition, in which subjects were read a description of the object's color to imagine, showed that this change in memory for spatial extent was not due to the introduction of an imagination task per se. In the image-color condition, as in the initial standard condition, the unidirectional bias indicative of boundary extension was eliminated (see Fig. 4).

These experiments demonstrate that the pattern of errors in memory for layout and object size were not dictated by the physical stimulus, but were determined by whether or not the viewer understood the display to be part of a continuous scene—a context that activates schematic expectations.

Conclusions

Research on transactive memory, perception, picture memory and imagination yield support for the notion of abstract schematic representation of visual scenes. However, critical questions remain to be answered before a strong theoretical stance can be taken. Foremost is the need to formally specify what a mental or abstract mental schema is. To this end, research is needed that will focus on: (a) the variability of the 'density' of information across the representation, (b) the best means of modifying the way in which this density map is stored, and (c) the extent to which this model can predict performances during perception, retrieval and imagination of scenes. At this stage, the notion of a common underlying representation for such a wide range of cognitive functions is highly speculative. Does a common representation underlie transactive memory and boundary extension? Perhaps it does, or perhaps the similarity between the two is relatively superficial. To pursue this question, it is necessary to obtain converging evidence regarding the efficacy of a number of factors on both. For example, if a picture was remembered just as a fixation to the right was about to be implemented, would there be a rightward bias in boundary memory for the scene? And if such dynamic effects are found in picture memory, will they also be manifest in an imagination task? At present, what the reviewed research does clearly show is that the representation of visual scenes is much more schematic and abstract than our immediate experience suggests. The visual system has evolved in such a way as to maximize comprehension of discrete views at the expense of uniqueness detail, but through the action of attention allows the viewer to access detail when the need arises. Trends in the literature suggest that the visual system may treat all views in the same way, regardless of their source: eye fixation, a picture, or an act of imagination.

Outstanding questions

- What is the best way to characterize the notion of an 'abstract schematic representation' of a scene?
- What type of model would be most useful for expressing the effects of attention on changes in information density across the representation over time?
- Can the concept of 'object files' be used as a means of capturing the complex relationships among objects and background elements in a scene?
- What type of tasks would be most convincing in establishing whether or not transactive memory, perception, picture memory and imagination of scenes share the same underlying representation, or are just similar processes?
- Do many 'filling-in' processes in visual cognition share the same underlying mechanisms? (e.g. amodal completion, completion across the blind-spot, and boundary extension?)
Fig. 4 Error patterns in memory for scenes and objects. Mean boundary ratings and 0.95 confidence intervals for close-up (white bar) and wide-angle (gray bar) target pictures in the outline-scene and outline-object conditions following standard (no imagery) instructions, and for the outline-object conditions when subjects either imagined scenes or imagined the objects in color. Test pictures were rated on a 5-point scale ranging from -2 to 2 (thus more of a close-up to +2) for each of a wide-angle view. Negative scores that differ from 0 indicate boundary extension, and positive scores that differ from 0 indicate boundary restriction. Bar graphs show results from Experiments 1, 3 and 6 (Ref. 30).

Acknowledgements

This review was supported by Grant MH40088 from the NIMH. I thank Carmine V. Guttman and Barbara Landis for their helpful comments on an earlier draft.

References

25. Brewster, R.A., O’Regan, J.K. and Clark, L.J. To see or not to see: To see or not to see: the need for attention to perceive change in scenes Psychol. Sci. In press