VISUAL COGNITION, 2000, *7* (1/2/3), 17–42

The Dynamic Representation of Scenes

Ronald A. Rensink

Cambridge Basic Research, Nissan Research & Development, Inc., Cambridge, MA, USA

One of the more powerful impressions created by vision is that of a coherent, richly detailed world where everything is present simultaneously. Indeed, this impression is so compelling that we tend to ascribe these properties not only to the external world, but to our internal representations as well. But results from several recent experiments argue against this latter ascription. For example, changes in images of real-world scenes often go unnoticed when made during a saccade, flicker, blink, or movie cut. This "change blindness" provides strong evidence against the idea that our brains contain a picture-like representation of the scene that is everywhere detailed and coherent.

How then do we represent a scene? It is argued here that focused attention provides spatiotemporal coherence for the stable representation of one object at a time. It is then argued that the allocation of attention can be co-ordinated to create a "virtual representation". In such a scheme, a stable object representation is formed whenever needed, making it appear to higher levels as if all objects in the scene are represented in detail simultaneously.

One of our most compelling impressions as observers is that we are surrounded by a coherent, richly detailed world where everything is present simultaneously. Although our environment is certainly this way, this impression is so compelling that we tend to believe these properties true of our *representations* as well—that is, we believe that somewhere in our brain is a stable and detailed representation of the stable and detailed world around us.

Please address all correspondence to R.A. Rensink, Cambridge Basic Research, Nissan Research & Development, Inc., 4 Cambridge Center, Cambridge, MA 02142-1494 USA. Email: rensink@pathfinder.cbr.com

I would like to thank Jack Beusmans, Dan Simons, Ian Thornton, and Carol Yin for their comments on an earlier draft of this paper. Portions of this manuscript were presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology, Ft. Lauderdale, FL, May 11-16, 1997.

But does such a representation really exist? The gathering of visual information is done using a retina that has high resolution only over a few degrees of visual angle. A complete representation of a scene therefore requires the contents of individual eye fixations to be integrated via a high-capacity visual buffer (e.g. Feldman, 1985; Trehub, 1991, 1994). But the inhomogenous nature of retinal representation rules out a simple superposition of fixations, so that any integration process would be less than straightforward (Yeshurun & Schwartz, 1989). Furthermore, even if integration *could* be done, it is not clear that it *should* be done—the computational requirements for representing our surroundings with a detail equal to that of foveal vision are overwhelming, even for human nervous systems (Rojer & Schwartz, 1990). And even if all this information could somehow be held in our brain, the sheer amount of it would cause severe problems with its efficient access by processes at higher levels (Tsotsos, 1990).

These theoretical reservations are supported by several lines of experimental work, all of which fail to find evidence for an integrative visual buffer (see Irwin, 1996; Simons & Levin, 1997). For example, changes in an image of a real-world scene become difficult to detect when made during a flicker, blink, eye movement, movie cut, or other such interruption. This "change blindness" suggests that little detailed information is being accumulated—otherwise, change detection would be easy, either by comparing immediate visual input with the contents of the buffer, or by detecting the anomalous structures formed by superimposing the original and the changed images. The fact that change blindness can be induced under a variety of conditions—together with the strength and robustness of the basic effect—indicates that the failure to accumulate detailed information is not an aberrant phenomenon occurring only under a special set of circumstances. Rather, it is central to the way we represent the world around us.

But if we are so bad at accumulating visual detail, how is it that we can see change at all? And if we do not have representations that are everywhere detailed and coherent, why do we have such a strong impression that these kinds of representations underlie our visual experience?

This paper suggests some possible answers to these questions. It begins by outlining a *coherence theory* of attention, which describes how focused attention can form a stable structure that enables the perception of change in an object. The notion of a *virtual representation* is then introduced to explain how attention can be co-ordinated so that a sparse set of stable structures can give the impression that scene representations are stable and detailed everywhere. Finally, a *triadic architecture* is suggested, showing how a virtual representation could be created in a way consistent with what is known of visual processing.

COHERENCE THEORY

If we are unable to accumulate visual detail, how is it that we can see change at all? Why do some conditions induce change blindness, but not others? The answer suggested here is based on the proposal that *focused attention is needed to see change* (Rensink, 1997; Rensink, O'Regan, & Clark, 1997). Under normal circumstances, any change in the world is accompanied by a motion signal, which attracts attention to its location (e.g. Klein, Kingstone, & Pontefract, 1992). It is only when this local signal is swamped (via the transients associated with a saccade, flicker, eyeblink, splat, etc.) that this guidance of attention is lost and change blindness then induced.

This explanation, however, creates an apparent paradox. Attention is thought to "weld" visual features into relatively long-lasting representations of objects (Kahneman, Treisman, & Gibbs, 1992; Kanwisher & Driver, 1992). It is also thought to operate relatively quickly, at a rate of 20–40 items per second (e.g. Julesz, 1984; Wolfe, 1994). But if this is so, why should the swamping of motion signals induce change blindness? Why doesn't attention simply weld all the visible items within the first few seconds of viewing and thereby enable the easy detection of change under all conditions?

The answer to this goes to the very heart of what it means to be attended. It is proposed here that attentional effects are largely concerned with *coherence*. As used here, this term denotes not only consistency in a set of representational structures,[1] but also logical interconnection, that is, agreement that the structures refer to parts of the same spatiotemporal entity in the world. Thus, two adjacent structures are spatially coherent if they refer to the same object, extended over space. Likewise, two successive structures are temporally coherent if they refer to the same object, extended over time.

Furthermore, rather than assuming that the structures formed by attention last indefinitely, it may be that their lifetimes are actually quite brief. In particular, attention may endow a structure with a coherence lasting only as long as attention is directed to it. Developing this line of thought leads to a *coherence theory* of attention:

---

[1]Discussions of representation often fail to distinguish between two rather different usages of the term. The first refers to the abstract coding scheme used when describing something, for example a Gabor function of a given frequency and orientation. The second refers to the particular instance used when describing something in the actual input, for example a Gabor function of a particular frequency and orientation that describes a particular edge fragment at a particular location. This distinction is essentially that between "type" and "token". To make clear which usage is intended, "representation" will be used when referring to form (type), and "representational structure"—or just "structure"—when referring to instance (token). Note that representational structure should not be confused with "mechanism", that is, the *way* in which representations are implemented (e.g. Marr, 1982, p. 342).

(1) Prior to focused attention, low-level "proto-objects" are continually formed rapidly and in parallel across the visual field. These proto-objects can be fairly complex, but have limited coherence in space and time. Consequently, they are volatile, being *replaced* when any new stimulus appears at their retinal location.
(2) Focused attention acts as a metaphorical hand that grasps a small number of proto-objects from this constantly regenerating flux. While held, these form a stable object, with a much higher degree of coherence over space and time. Because of temporal continuity, any new stimulus at that location is treated as the *change* of an existing structure rather than the appearance of a new one.
(3) After focused attention is released, the object loses its coherence and dissolves back into its constituent proto-objects. There is little or no "after-effect" of having been attended.

According to coherence theory, a change in a stimulus can be seen only if it is given focused attention at the time the change occurs. Since only a small number of items can be attended at any time (e.g. Pashler, 1988; Pylyshyn & Storm, 1988), most items in a scene will not have a stable representation. Thus, if attention cannot be automatically directed to the change, the changing item is unlikely to be attended, and change blindness will likely follow.

Limited coherence of unattended proto-objects

The lowest levels[2] of visual perception are generally believed to provide a detailed map-like representation, or "sketch", of the scene-based properties visible to the viewer (Fig. 1). These representational structures are thought to be retinotopic, and formed rapidly (i.e. within a few hundred msec) and in parallel across the image, without the need for focused attention (e.g. Marr, 1982; Rensink, 1992; Rensink & Enns, 1998). Given that focused attention is not involved, coherence theory states that these structures have only limited coherence in space and time.

[2] As used here, "low-level vision" refers to those processes concerned with separating out the various physical factors in the scene that give rise to the pattern of luminance intensities in the image. These processes are generally considered to be retinotopic, and carried out by a set of processors operating in parallel across the visual field. According to the theory proposed here, they are also volatile, either fading away a few hundred milliseconds after the proximal stimulus disappears, or else being replaced by a new stimulus that appears in the same retinal location. Meanwhile, "early vision" refers to that aspect of low-level processing carried out regardless of higher-level context, or perceptual set. Since context generally requires at least a hundred milliseconds of viewing to get established, early vision can be equivalently defined as that aspect of vision that is both low-level *and* rapid (e.g. Rensink & Enns, 1998).

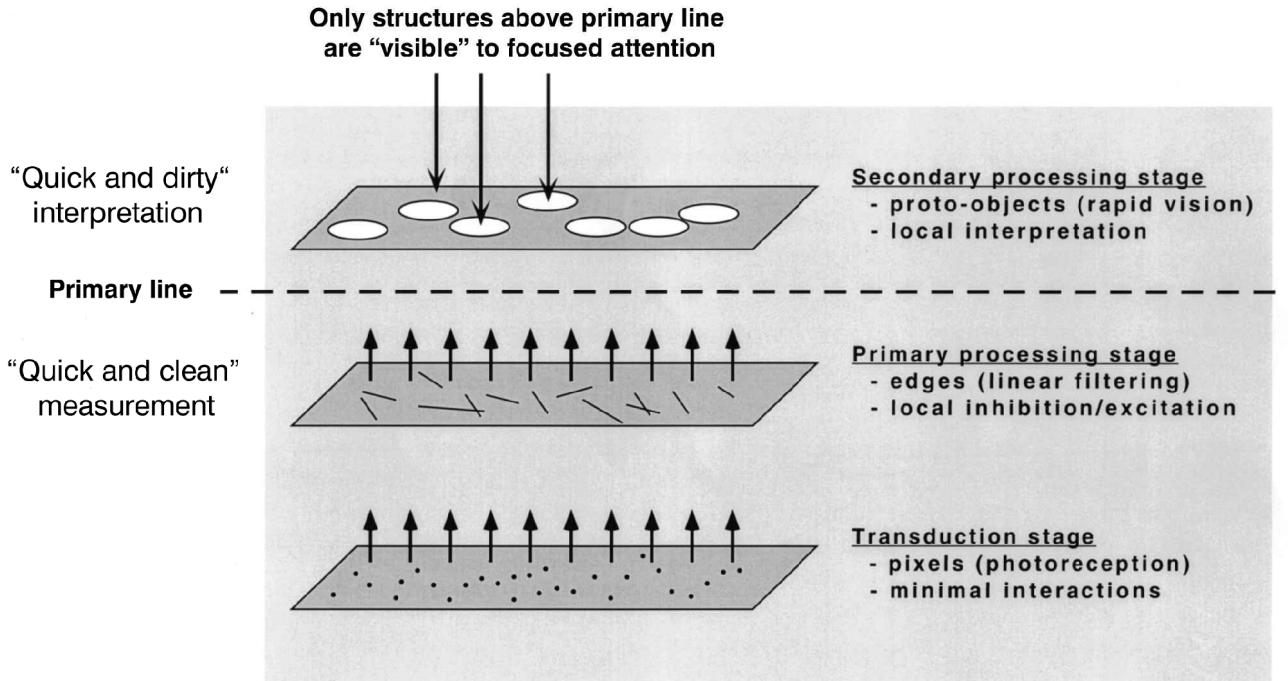

FIG. 1. Schematic of low-level vision. Three main stages are distinguished here: (1) the *transduction* stage, where photoreception occurs, (2) the *primary* processing stage, where linear or quasi-linear filters measure image properties, and (3) the *secondary* processing stage of rapid non-linear interpretation. Operations at all three stages are carried out in parallel across the visual field. The transduction and primary stages obtain their speed at the expense of complexity; in essence, they perform "quick and clean" measurements. The limits to these kinds of operations are given by the *primary line*. The secondary stage obtains its speed at the expense of reliability, opting for "quick and dirty" interpretations that may not always be correct. The outputs of this stage are proto-objects that become the operands for attentional processes.

Much of the evidence for limited spatial coherence comes from visual search experiments. The results of early studies suggested that spatial coherence was minimal, with non-attended structures limited to simple blobs and bars (e.g. Julesz, 1984). But although this *primary stage* reflects the limits of reliable measurement, it does not reflect the limits of low-level processing. Later experiments showed the existence of a *secondary stage*, where local interactions enable the "quick and dirty" recovery of various scene-based properties, such as surface curvature (Ramachandran, 1988), slant (Enns & Rensink, 1991), and shadows (Rensink & Cavanagh, 1993). But even in this stage, limits exist on the extent over which information is collected—for example, rapid line interpretation fails for items larger than 4° (von Grünau & Dubé, 1994).

Several types of rapid grouping have been found to occur at the secondary stage (Elder & Zucker, 1993; Rensink & Enns, 1995), as has the rapid completion of occluded figures (Enns & Rensink, 1992; He & Nakayama, 1992; Rensink & Enns, 1998). Thus, it appears that low-level processes are concerned not only with the recovery of scene-based properties, but also with their formation into "proto-objects", that is, relatively complex assemblies of fragments that correspond to localized structures in the world. Recent work also indicates that proto-objects are the lowest-level structures directly accessible to attention, with much of their underlying detail being accessed only by deliberate effort (e.g. Rensink & Enns, 1995, 1998). As such, proto-objects have a "two-sided" nature, being not only the highest-level outputs of low-level vision, but also the lowest-level operands upon which higher-level attentional processes can act.

Evidence for the limited temporal coherence of proto-objects comes largely from studies of visual integration, which show that stimuli can be integrated over time only if they are at the same retinal location and arrive within about 100msec of each other (e.g. DiLollo, 1980). Beyond these limits, it appears that the details of successive presentations cannot be added, compared, or otherwise combined (e.g. Irwin, 1996). These results, together with those from change-blindness studies, provide strong evidence that early level structures are inherently *volatile*: They are either overwritten by subsequent stimuli or else fade away within a few hundred milliseconds (Rensink, O'Regan, & Clark, 1997, this issue). Note that this volatility is at the level of proto-objects and not pixels—if a new stimulus has an empty space in its midst, the contents at that location will be part of the new proto-object, and so will still be overwritten (Enns & DiLollo, 1997; Rensink, this issue). In summary, then, the sketch formed at any particular fixation can be highly detailed, but will have little coherence, constantly regenerating as long as light continues to enter the eyes, and being largely created anew after each eye movement.[3]

[3]The creation of a new low-level description may not need to proceed completely from scratch for proto-objects that are attended. As described in the section on attentional coherence, attended structures have temporal continuity. Thus, there may be some carry-over when these structures are re-formed at a new retinal location.

Extended coherence of attended objects

Given that unattended structures have only limited spatial and temporal coherence, it follows that focused attention must provide the coherence that knits them into larger-scale objects and allows them to retain their continuity over time. Note that this latter property[4] is particularly important in regards to the perception of change, for continuity allows the appearance of a new stimulus to be treated as the transformation of an existing structure rather than the formation of a completely new one.

In this view, then, focused attention is intimately involved with the perception of objects.[5] Essential properties of an object include the requirement that it be discrete, be differentiated from its background, and have a coherent unity across space and time. It must also be an individual, literally something that cannot be divided without losing its integrity—if an object is taken apart, the result is a set of parts rather than a set of objects similar to the original (e.g. Smith, 1998).

To capture these properties, coherence theory asserts that focused attention is involved with the representation of only one object at a time. This limit is taken from studies showing that for some tasks attention appears to operate on just one object (e.g. Deubel & Schneider, 1996; Garavan, 1998; Rensink, 1998a). Attentional interaction with lower-level structures is posited as taking place via a *nexus*, a single structure containing a summary description of the attended object, for example, its size, overall shape, and dominant colour. Within the nexus, internal connections enable the computation of these summary properties, as well as providing a way to briefly store them.[6]

---

[4]The property of temporal continuity could also be called "identity" (cf. Kahneman et al., 1992; Pylyshyn & Storm, 1988). However, there are several problems with this term (e.g. Smith, 1998). Most prominent of these is its ambiguity—"identity" can also denote semantic recognition (as in "identification"), and so may cause connotations not wanted here. As such, "identity" will not be used to denote spatiotemporal continuity; "semantic identity" will always be used when referring to issues of semantic recognition.

[5]Several effects (e.g. the improvement of performance when uncertainty is reduced) are believed to occur via attention that is allocated to particular feature spatial locations or—more generally—feature "channels" (e.g. Davis & Graham, 1981; Shulman & Wilson, 1987). The relation between space-, feature-, and object-based attention is not yet clear. Available evidence suggests that different—but interacting—systems may be involved (e.g. Lavie & Driver, 1996). The account here is intended only for object-based attention.

[6]One meaning of the term "nexus" is *connection* or *tie*; another is *connected series* or *group*. Both meanings apply to the concept of nexus proposed here. First, the nexus is characterized as the sole connection between attention and the visual structures at lower levels. Second, the nexus can have links to several proto-objects as well as having a differentiated internal structure, thereby forming a connected series that is treated in many respects as a single entity (i.e. an object).

When a proto-object is attended, a *link* is established between it and the nexus, enabling a two-way transmission of information between these structures (Fig. 2). Information going up the link allows the nexus to obtain descriptions of selected properties from the attended proto-object. Information going down the link can in turn provide stability to the volatile proto-object, allowing it to be maintained or to rapidly regenerate when it is briefly occluded or when the eye moves. Among other things, links also enable a mapping between the ever-changing retinotopic co-ordinates of the proto-object and the more stable viewer- (or object-)centered co-ordinates of the nexus. When the link establishes a recurrent flow of information between the nexus and its proto-object, the resulting circuit is referred to as a *coherence field*.

Experiments also indicate that for some tasks four to six items can be attended at the same time (e.g. Pylyshyn & Storm, 1988; Rensink, this issue). In such a situation, the coherence field still contains a single nexus, but is now expanded to include several proto-objects, with nexus properties determined via links to these structures (Fig. 2). The recurrent flow of information between the nexus and its proto-objects not only establishes coherence over space, but also forms a type of memory, establishing coherence over time as well. Describing this in a more metaphorical way, attention may be seen as a hand that "grasps" the proto-objects with about four to six "fingers", the selected structures then forming a single coherent object as long as they are "held".

Note that the nexus and its proto-objects form a "local hierarchy", with only two levels of description (object- and part-level). Such a hierarchy is an extremely useful device, and is a natural way to represent objects (Marr, 1982, pp. 305–307). For example, a proto-object could be attentionally subdivided and the links assigned to its parts; this would correspond to a traversal down one level of the part-whole hierarchy of that object. Conversely, the links could be assigned to several widely separated proto-objects, forming a group that would correspond to a (coarsely coded) object one level up. Thus, even though the capacity of focused attention may be limited (e.g. Pashler, 1988; Rensink, this issue), the ability to quickly traverse a part–whole hierarchy gives it rapid access to any aspect of an object's structure.

The limited amount of information that can be attended at any one time explains why observers can fail to detect changes in "attended" objects (Levin & Simons, 1997). When focused attention is directed to something in the world, it will not generally be possible to represent all of its detail in a coherence field—only a few of its aspects can be represented in the nexus at any one time. If one of the aspects being represented is one of the aspects changing in the world, the change will be seen; otherwise, change blindness will still result.

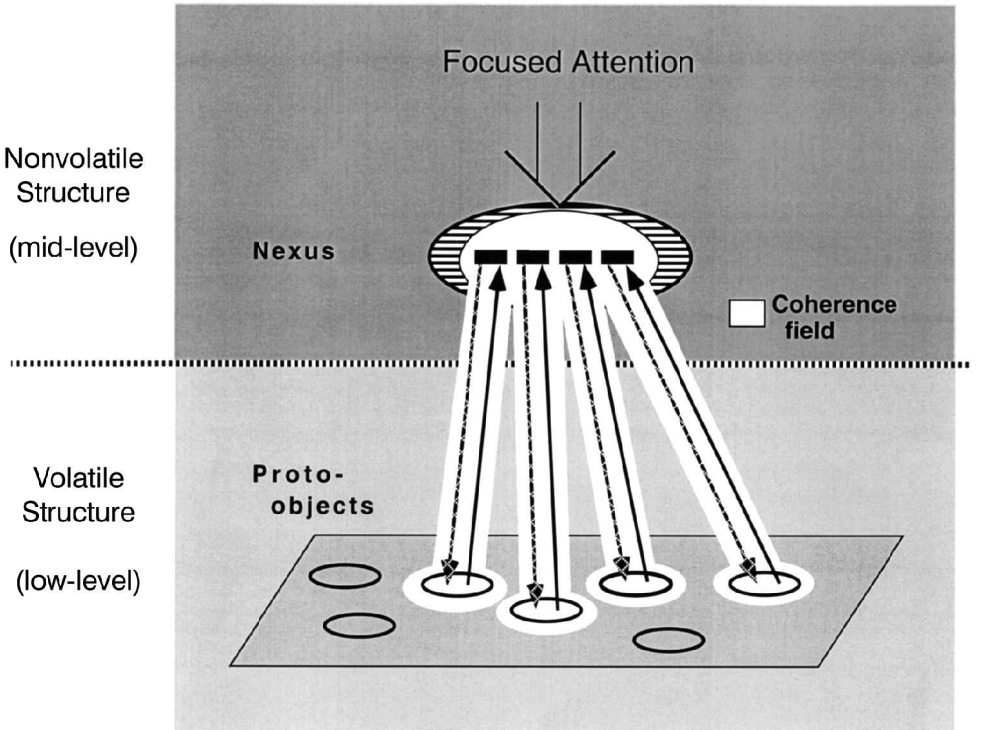

FIG. 2. Schematic of a coherence field. This structure is composed of three kinds of components: (1) a nexus, corresponding to a single object, (2) a set of 4–6 proto-objects, corresponding to object parts, and (3) bidirectional links between the nexus and the proto-objects. Coherence is established when a recurrent flow of information exists between the nexus and its proto-objects, as well as within the nexus itself. Selected information is transmitted up the links to enter into the description of the object. Information from the nexus is also transmitted back down the links to provide stability (and perhaps refinement) to the proto-objects.

Lack of attentional aftereffect

The final part of coherence theory concerns the fate of the coherence field once focused attention is removed. Given that only one object can be represented at a time, a coherence field cannot be maintained if attention is switched to another object. In such a case, the links are dissolved and the previously attended parts revert to their original status as volatile proto-objects. To appeal to the hand metaphor again: The release of focused attention is like the release of the items held by the hand, with these items returning to the "primal ooze", that is, the flux of constantly regenerating low-level structures.

In this view, then, there is little or no aftereffect of having directed attention to a structure, at least in terms of the ability to detect change. There is of course a short-term memory (STM) for items that have been previously attended (e.g. Cowan, 1988). But in the view taken here, STM is an abstract memory concerned with object *types*; in contrast, a coherence field is believed to embody visual short-term memory (vSTM), a purely visual memory supporting the formation of object *tokens*, and so may contain little or no information after attention is withdrawn.

In the extreme, vSTM can be identified as the memory formed by a coherence field, which leads to the position that there is no vSTM apart from what is attended. In other words, attending to an item is both necessary and sufficient for it to be in vSTM. Evidence for this position is mixed (e.g. Pashler & Carrier, 1996). But physiological studies show that the mechanisms for short-term (or working) memory are similar to those for focused visual attention—so similar, in fact, that there may be no difference at all (Awh & Jonides, 1998; Desimone, 1996). Furthermore, psychophysical studies indicate that there may be a complete lack of memory for items previously attended in a visual search task (Wolfe, 1996). As such, there appears to be at least some degree of support for this position.

Relation to other work

One of the most influential of early models of attention was that of Shiffrin and Schneider (1977). In their view, STM was an "activated" sub-set of LTM, with attention selecting the particular LTM items to be activated. Coherence theory differs from this in several ways. First, it views attention as concerned with the formation of immediate spatiotemporal structures (or tokens) rather than the activation of long-term categories (or types). Second, Shiffrin and Schneider believed that any process could operate without attention after sufficient practice, whereas coherence theory asserts that the perception of change always requires attention. Finally, the activation posited by Shiffrin and Schneider can last after attention is withdrawn, whereas a coherence field collapses as soon as this occurs. It is important to note that according to coherence theory such lasting activation is still possible for STM. The proposal here is that this is not

possible for vSTM (or visual working memory), a rather different system concerned entirely with spatiotemporal structure.

The notion of a coherence field is closer to the proposal of Kahneman et al., (1992) that spatiotemporal structures are represented by "object files" in which various properties are bound together. Both views agree that an attended representational structure only needs to describe a spatiotemporal entity, and that it does not need to be matched to stored descriptions in long-term memory. But whereas object files may contain information on non-visual properties (such as the appropriate response to make), nexus properties are limited to the purely visual, or to abstract properties derivable from the visual (such as semantic identity). More importantly, once set up, an object file may or may not be attended, so that several files can be maintained at a time; in contrast, there can be only one nexus (perhaps linked to several structures), and its associated field collapses as soon as attention is withdrawn.

The idea of a coherence field—in particular, its set of links—also has some similarity to the concept of FINSTs ("fingers of instantiation") proposed by Pylyshyn and Storm (1988). FINSTs are pointers that provide access paths to attended objects, continually informing higher-level processes of their positions; it is posited that about five FINSTs can be used at a time. Both FINSTs and links provide higher-level processes with information about lower-level structures, and both can stabilize these structures to get continuity over time. But FINSTs transmit only the position of an item to higher levels, whereas links transmit several kinds of visual information, and do so in a recurrent fashion. Also, FINSTs are assigned to completely independent objects, whereas links are brought into a nexus corresponding to a single object (although this object may have several parts).[7] Because there are as many links as FINSTs, links can potentially explain all the results explained by FINSTs, such as tracking and subitizing (Pylyshyn & Storm, 1988; Trick & Pylyshyn, 1993). Furthermore, the constraint of a single nexus explains why the tracking of isolated dots in a display may be better interpreted as the tracking of corners of a single virtual object (Yantis, 1992).

Coherence theory is also compatible with studies showing that when observers attend to particular objects or events in a scene, they often fail to report the appearance of other, unexpected items (Mack & Rock, 1998). Recent work suggest that this may be not so much a failure to *see* these items as it is to *remember* them (Moore & Egeth, 1997; Wolfe, 1997). This interpretation is consistent with a perceptual "here and now" in which volatile representations

[7]Each link is a finger in the metaphorical attentional hand, and so in some respects corresponds to a FINST (Pylyshyn & Storm, 1988). To have a direct correspondence with the attentional hand itself, FINSTs would need to be augmented by other mechanisms. This would presumably result in a "HANST".

of considerable detail and sophistication are continually built and rebuilt in the absence of attention.

VIRTUAL REPRESENTATION

The theory of attention proposed previously has a rather counterintuitive implication: Only one object in an environment, or scene, can be given a coherent representation at any time. Moreover, this representation is limited in the amount of information it can contain. But if this is so, why do we not notice these limitations? Why do we feel that somewhere in our brain is a complete, detailed representation of all the objects in the scene?

To answer this, consider how objects are used in everyday life. For most tasks, only one object is in play at any time: A cup is grasped, a friend recognized, a speeding cyclist avoided. A detailed representation may be required for this "target" object, but it is not required for the others. Although there appear to be tasks (e.g. juggling) that are exceptions to this, these tasks are generally handled by quickly switching back and forth, so that there is only a single target at any one time. Thus, although we may need to represent various aspects of a scene (such as the background), it would appear that we never need a detailed representation of more than one of the objects in it at any particular time.

This realization gives rise to the idea of a *virtual representation*: Instead of forming a detailed representation of all the objects in our surroundings, represent only the object needed for immediate purposes. If attention can be coordinated so that a coherent, detailed representation of an object can be formed whenever it is needed, the representation of a scene will appear to higher levels as if it is "real", that is, as if all objects are represented in great detail simultaneously. Such a representation will then have all the power of a real one, while requiring far less in the way of processing and memory resources.

Example: Accessing a computer network

To get a better feel for what is meant by virtual representation, consider the problem of accessing the data contained in a large network, such as the World Wide Web (Fig. 3). On one hand is the browser workstation, limited in the amount of information it can hold in its memory. On the other is the network, with information held by thousands of machines. Suppose now that we want the workstation to access data contained at various sites. How should this be handled?

Given sufficient memory, the workstation could contain a complete copy of all the data contained in all the computers of the network. But the amount of memory needed for this would be enormous. Furthermore, each time data was added or deleted from one of the machines on the network, it would have to be broadcast to all the others, resulting in huge transmission costs.

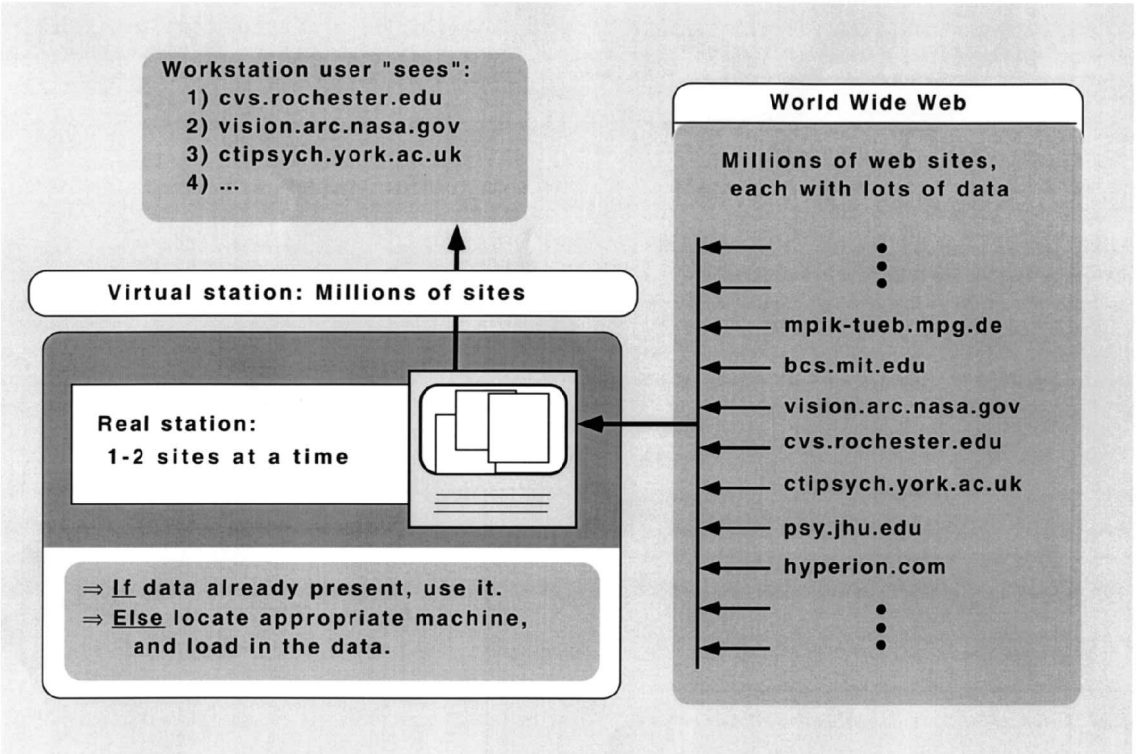

FIG. 3. Virtual representation—Computer network. If a limited-capacity workstation can access information from the computer network whenever requested, it will appear to contain all the information from all sites on the network.

Consequently, network designs generally favour a more dynamic approach to data access. If we want to see the information at a particular site, our workstation checks to see if it is already in its memory. If so, nothing more needs to be done. Otherwise, it sends out a request to the appropriate site and has the requested information loaded in (Fig. 3). If this transfer can be done sufficiently quickly, our workstation will appear to contain all the information in the network. But in reality, this information will have only a virtual representation: It is not all present simultaneously in the workstation, but is simply accessed whenever requested.[8]

To see how this strategy can explain scene perception with a limited-capacity attentional mechanism, consider the parallels between the two problems:

- A workstation can hold the contents of one (or at most a few) sites.
- There are thousands of sites on the network, containing enormous amounts of information.
- The workstation cannot hold all of this information.

- Attention can hold the contents of one (or at most a few) objects.
- There are thousands of objects in the visible scene, containing enormous amounts of information.
- Attention cannot hold all of this information.

Given the similar structure of the problems, a similar solution can be used (Fig. 4):

- If the information from a site is already held in memory, use it.
- Otherwise, locate required site, and load in information.
- Result is a virtual representation of the contents of the network.

- If the information from an object is already being attended, use it.
- Otherwise, locate required proto-objects, and make them coherent.
- Result is a virtual representation of the contents of the visible scene.

In the case of the computer network, competent co-ordination of network requests makes it appear as if a low-capacity workstation (limited to one or two sites) simultaneously holds all the information on the network. Similarly,

[8] A virtual representation is often used *within* an single computer to make it appear to have more memory that is physically present. This is referred to as "virtual memory" (e.g. Tanenbaum, 1976).

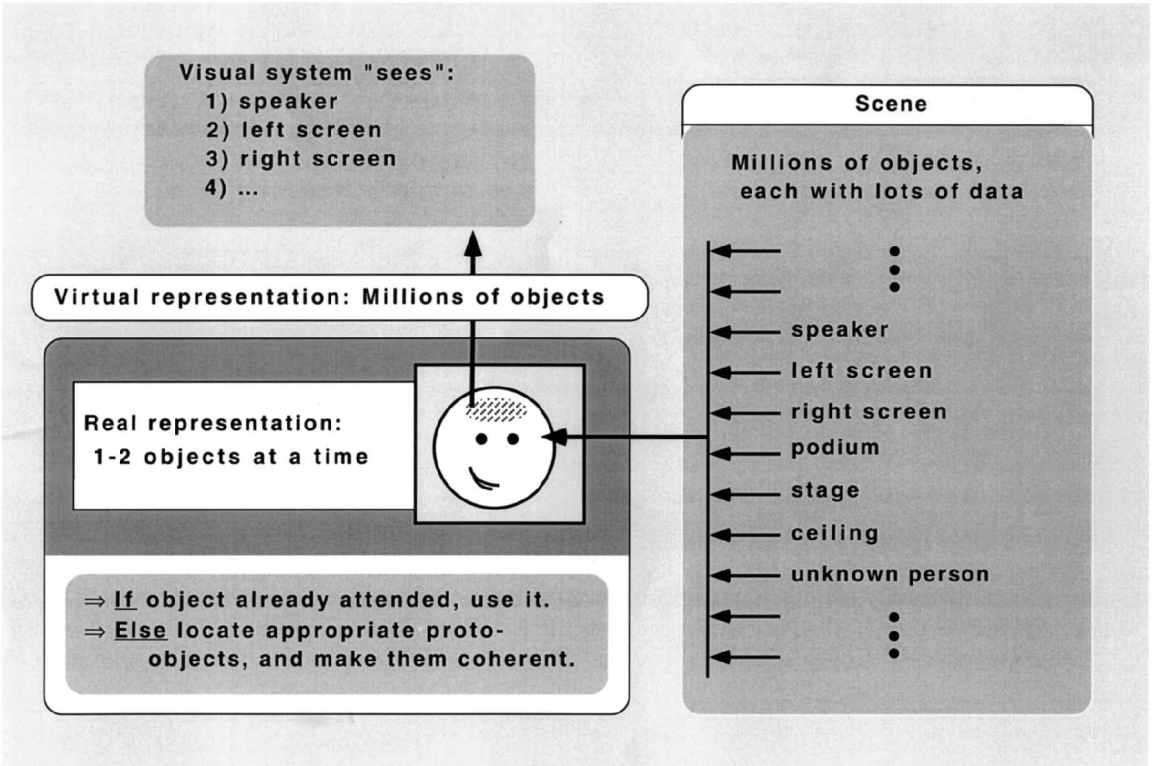

FIG. 4. Virtual representation —Human vision. If a limited-capacity attentional system can access information from the visible scene whenever requested, it will appear to contain all the information from all objects in the visible scene.

competent co-ordination of attentional requests could make it appear to higher-level processes as if a limited-capacity coherence field (limited to one or two objects) simultaneously held all the information about all objects in the scene.

Thus, even though our conscious minds may have the impression that all the objects in front of us are simultaneously given a detailed, coherent representation somewhere in our brain, this need not be the case. Instead, this can result from a much sparser "just in time" system that simply provides the right object representation at the right time.

It is important to note that this feeling of completeness does not necessarily mean that the representation really *is* complete, that is, that it represents all the objects in view. It also does not mean that it represents them all correctly. Just like a static "real" representation, a dynamic virtual representation may fail to represent particular objects, or may represent them incorrectly. As such, the extent to which a representation is virtual or real is unrelated to its accuracy or completeness.

Conditions for successful operation

Although a virtual representation can result in an enormous savings of computational resources, these savings do not come for free. Virtual representations reduce complexity in space by trading off for an increased complexity in time. Only particular types of information-processing tasks can take advantage of this trade-off. Is visual perception one of these?

The keys to the successful operation of a virtual representation are: (1) only one (or at most a few) objects need to have a "real" representation at any one time, and (2) detailed information about any object must be made available when requested. The first requirement is easily met for most (if not all) visual tasks. We usually need to attend to only one object at a one time, for example, to grasp it, or to see where it is headed. Tasks where several target objects are involved can generally be handled by "time-sharing", that is, by rapidly switching attention back and forth between the objects.

The requirement of access on request is also met under most conditions of normal viewing. Provided that there is a way to guide eye movements and attentional shifts to the location of the requested object, visual detail can be obtained from the stream of incoming light. Consequently, a high-capacity visual memory for objects is not needed—the information is generally available from the world itself. As pointed out long ago by Stroud (1955): "Since our illumination is typically continuous sunlight and most of the scenery stays put, the physical object can serve as its own short-term memory". Stroud's insight has recently been revived, with several proposals further supporting the idea that much of perception is best understood in terms of using the world as its own best model (e.g. Brooks, 1991; Dennett, 1991; Grimes, 1996; O'Regan, 1992).

Note that problems arise with this scheme when light is not available to carry the information from the object to the eyes, or when the objects themselves are somehow occluded. But these conditions also interfere with object perception itself, regardless of the memory scheme used, and so do not form a serious obstacle to the use of virtual representation.

What is more important is to consider illumination and occlusion from the point of view of short-term (or *working*) perception[9], that is, the perception of events over durations of several seconds. Since illumination is fairly constant during the daytime, an object seen at one particular time will almost always be illuminated a short time later. As such, fluctuation in illumination is highly unlikely to interfere with short-term perception. Likewise, sudden occlusions of previously seen objects are relatively rare over the span of several seconds. As such, information about an object seen at one particular time is almost always available several seconds later. Consequently, virtual representation can be a reliable and effective way of handling the large amounts of information contained in a real-world scene, at least for purposes of short-term perception.

General applicability

Virtual representation is a powerful information-processing strategy, one that is central to much of modern computer and network design (e.g. Tanenbaum, 1976). Unfortunately, this form of representation was long neglected as a way of explaining perceptual and cognitive processing—only recently has work begun to explore its potential in this regard (e.g. Brooks, 1991; Dennett, 1991). And even many of these studies do not focus on biological information processing, being instead demonstrations of its effectiveness in artificial systems.

At the most general level, work on virtual representation illustrates the power of deictic (or indexical) representation. In deictic representation, there is little memory of detailed information; instead, emphasis is placed upon the extraction of a few key "aspects", which then serve as pointers to entities in the world (e.g. Ballard, Hayhoe, Pook, & Rao, 1997; Clancey, 1997). Interactions between sub-systems can also be handled this way, using a few key behavioural aspects rather than a detailed representation of the goals or the information in the other sub-systems (e.g. Brooks, 1991). In all these cases, the power of a deictic representation is jointly determined by the representational structure and its context (i.e. the world itself, or the set of interacting sub-systems). In this view, representations do not *construct* a copy of the world or of their neighbors—rather, they simply *co-ordinate* the actions of the various systems involved.

[9] I thank Dan Simons for suggesting this term.

TRIADIC ARCHITECTURE

The successful use of virtual representation in human vision requires that eye movements and attentional shifts be made to the appropriate object at the appropriate time. But what directs these movements and shifts? How can the location of an object be known before attention has been directed to it? And if attention has no aftereffects, how can there be any memory of a scene after attention has been withdrawn?

An unequivocal, detailed answer to all these questions would largely amount to a complete theory of vision, something not yet in existence. Consequently, this section will provide only a sketch of one possibility. This sketch is not meant to be definitive. Rather, it is simply meant to show that an answer can be given to these questions, an answer that allows virtual representation to be carried out in a manner compatible with what is known about human visual processing.

The solution proposed here begins by discarding the assumption that all visual processing passes through a single attentional locus. Although having great intuitive appeal, such an "attento-centric" model of vision may not correspond to reality. Recent studies indicate that a single locus of attention may not exist: Capacity-limited processes loosely identified as "attentional" may be found at different levels of the visual system, and possibly even in different processing streams (e.g. Allport, 1992). If so, the attentional system used for object perception would simply be one system among many, the others operating concurrently and largely independently of it.

Developing this view further leads to a *triadic architecture* with three largely independent systems (Fig. 5). The first is a low-level system that rapidly creates highly-detailed, volatile structures. The second is a limited-capacity attentional system that forms these structures into stable object representations. These two systems are already part of coherence theory. What is now added is a limited-capacity non-attentional system that provides a *setting* to guide attention[10]. This "setting system" involves at least three aspects of scene structure:

(1) Perception of the abstract meaning, or *gist* of a scene (e.g. whether the scene is a harbour, city, picnic, barnyard, etc.). This could provide a

[10]Various systems have been proposed for guiding focused attention (e.g. Treisman & Sato, 1990; Wolfe, 1994). But these tend to emphasize the mechanics of guidance, being concerned with issues such as whether various features are selectively excited or inhibited, or whether selection can be done on groups or individual items. The system sketched here is not concerned with such issues; rather, it involves a more functional perspective, investigating the kinds of information that could be used to direct a guidance mechanism, however implemented.

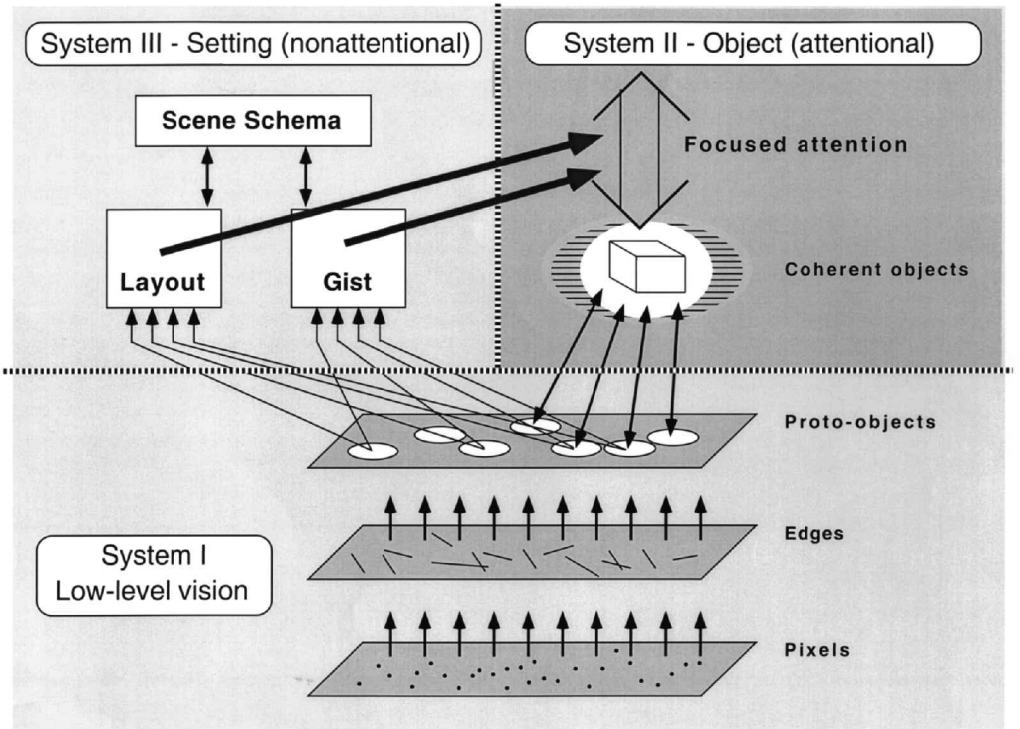

FIG. 5. Triadic architecture. It is suggested that the visual perception of scenes may be carried out via the interaction of three different systems. System I: Early-level processes produce volatile proto-objects rapidly and in parallel across the visual field. System II: Focused attention acts as a hand to "grab" these structures; as long as these structures are held, they form an individuated object with both temporal and spatial coherence. System III: Setting information—obtained via a non-attentional stream—guides the allocation of focused attention to various parts of the scene, and allows priorities to be given to the various possible objects.

useful way to prioritize attention, directing it to the objects that are most important in that context.

(2) Perception of the spatial arrangement, or *layout* of the objects in the scene. This could provide a non-volatile representation of the locations of various structures, which could then be used when attention is to be directed to particular objects in the scene.

(3) Invocation of an abstract *scene schema* stored in long-term memory, presumably via gist or layout information. Once invoked, this could facilitate the perception of these two quantities, and ultimately—through the associated interactions—facilitate the perception of objects.

This architecture is somewhat similar to early proposals that scene perception involves an initial extract of gist and subsequent refinement of detail (e.g. Loftus, 1976). However, it differs from these in that a complete scene representation is never constructed—although the representation of gist and layout may be improved over the course of viewing, there always remains only one coherent object represented at any one time. As such, this architecture embodies a fundamental change of perspective: scene representations are no longer structures *built up* from eye movements and attentional shifts, but rather, are structures that *guide* such activities.

Gist

The most abstract aspect of a scene is its meaning, or *gist*. This quantity remains constant over many different eye positions and viewpoints, as well as changes in the composition and layout of objects in an environment. As such, it can provide a stable constraint on the kinds of objects expected, and perhaps even indicates their importance for the task at hand (Friedman, 1979).

Gist can be determined within 120msec of presentation (Biederman, 1981; Intraub, 1981; Potter, 1976), a time insufficient for attending to more than two or three items. Furthermore, it can be extracted from highly blurred images, and without attention—indeed, two different gists can be determined simultaneously (Oliva & Schyns, 1997). In accord with these findings, gist does not appear to be determined by objects, which are perceived concurrently or even afterwards (e.g. Henderson, 1992). As such, it may be determined by simple measures such as the distributions of line orientations or colours in the image (e.g. Guérin-Dugué, Bernard, & Oliva, 1998), or other properties of the proto-objects that exist at low levels.

Layout

Another important aspect of scene structure is *layout*, the spatial arrangement of the objects in a scene, without regard to visual properties or semantic identity (Hochberg, 1968). This quantity—at least from an allocentric frame of

reference—is invariant with changes in eye position; as such, it could be useful for directing eye movements and attentional shifts.

The visual system appears to extract at least some layout information within a few seconds of viewing, and to be able to hold it across brief temporal gaps (Sanocki & Epstein, 1997; Simons, 1996). But the memory involved differs from that of a coherence field—it holds spatial location rather than visual properties, and concerns an entire scene (or at least an array of objects) rather than just a single object. It also appears to be non-volatile, enduring even in the absence of attention (Chun & Nakayama, this issue; Haber, 1985). It is important to note that even if layout is *held* in a non-attentional memory, this does not imply that it is *obtained* non-attentionally. It is possible, for example, that it is extracted from the scene and entered into memory via a series of attentional shifts or eye movements.

Scene schema

The invariance of gist and layout information not only allows these quantities to provide a relatively stable context for other operations—it also facilitates the long-term learning of scene constraints. Long-term memory for scenes appears to involve not only the scene category, but also an associated collection of representations, or *scene schema* (e.g. Arbib, 1990; Friedman, 1979; Intraub, 1997).

Whereas gist and layout involve short-term (or working) representations with a limited lifetime, scene schemas are long-term structures that may last indefinitely. This allows them to accumulate information, so that their contents can be more detailed and sophisticated than the perceptual structures that invoke them. For example, scene schemas are believed to include an inventory of objects likely to be present in the scene, along with various aspects of layout, such as the relative locations of the inventory objects (e.g. Mandler & Parker, 1976).

Interaction of systems

In the triadic architecture proposed here, the representation of a scene involves the dynamic interaction of three different systems. How might this be carried out?

When a scene is viewed, rapid low-level processes provide a constantly-regenerating sketch of the properties visible to the viewer. Gist may be determined by a subset of these, with subsequent processes attempting to verify the schema invoked (Antes & Penland, 1981; Friedman, 1979). Items consistent with the schema need not be encoded in detail, since verification may involve a simple checking of expected features. In other words, objects only need to be *detected*—a coherent representation of their structure need not be constructed (Henderson, 1992). If an unexpected structure in the image is encountered,

more sophisticated (attentional) processes could form a coherent representation of its structure, attempt to determine its semantic identity, or re-evaluate the gist. Meanwhile, the perceived layout of items could be used as a check on the current interpretation, as well as helping to guide attention to a requested object.

This set of interactions therefore provides a way of creating a virtual representation of all the objects in the scene. It may also help explain not only why we have an impression of all objects being present simultaneously (via the virtual representation), but also why we have a concurrent impression of detail at all background locations (possibly via the volatile set of proto-objects).

It is also worth pointing out that several interesting effects may be explained in terms of these interactions breaking down. For example, if focused attention is occupied with the formation of an object, an unattended stimulus can cause priming (Shapiro, Driver, Ward, & Sorensen, 1997), or enter long-term memory (Chun & Jiang, 1998), even though it is not consciously perceived. This can be explained in terms of the attentional system "locking on" to a particular object, with the unseen information travelling along other, non-attentional streams. Indeed, if one of these non-attentional streams can detect (but not support perception of) some kinds of change, it may explain how observers can sometimes guess that change has occurred even though they have no explicit awareness of it (Fernandez-Duque & Thornton, this issue). Another effect is "mindsight", where observers have a strong feeling that something is changing but have no accompanying visual experience (Rensink, 1998b). Here, it may be that the change is detected by a non-attentional sub-system, which then alerts the attentional system. The information transmitted in an alert need not be large, which may explain why observers have little idea of what or where the change is. Note that this kind of explanation has a parallel with that proposed for blindsight,[11] which relies on a similar separation of processing streams (Milner & Goodale, 1995).

SUMMARY

It is proposed that a dynamic representation underlies our perception of scenes. One component of this proposal is a *coherence theory* of attention, which asserts that unattended structures are volatile, and that focused attention is needed to stabilize them sufficiently to allow the perception of change. The other component is the assertion that vision makes use of *virtual*

[11] Although related, blindsight and mindsight are different phenomena. In blindsight, the observer has no conscious experience of the stimulus whatsoever, being completely surprised, for example, when their hand is able to grasp an "unseen" object. In mindsight, the observer has no conscious *visual* experience. But they still have a conscious "mental" (i.e. non-visual) experience—hence the name.

representation, a dynamic form of representation in which attention provides detailed, coherent descriptions of objects that are needed exactly when they are needed. A *triadic architecture* is proposed as one possible way to create such a representation. This architecture uses representations that are stable and representations that contain large amounts of visual detail. But at no point does it use representations that are both stable *and* contain large amounts of detail.

In this view, the impression of having representations that are both stable and detailed is due to the careful co-ordination of attention. To the extent that the resulting descriptions contain the information needed, our impression as observers will be of a richly detailed environment, with accurate representation of those aspects of greatest importance. It is only when low-level transients are masked or are disregarded due to inappropriate high-level control that attentional co-ordination will break down, causing the true nature of the virtual representation to intrude into our perceptual awareness.

REFERENCES

Allport, A. (1992). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In D.E. Meyer & S. Kornblum (Ed.), *Attention and performance XIV* (pp. 183–218). Cambridge, MA: MIT Press.

Antes, J.M., & Penland, J.G. (1981). Picture context effects on eye movement patterns. In D.F. Fisher, R.A. Monty, & J.W. Senders (Ed.), *Eye movements: Cognition and visual perception* (pp. 157–170). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Arbib, M.A. (1990). Schemas for high-level vision: The problem of instantiation. In E.L. Schwartz (Ed.), *Computational neuroscience* (pp. 340–351). Cambridge, MA: MIT Press.

Awh, E., & Jonides, J. (1998). Spatial working memory and spatial selective attention. In R. Parasuraman (Ed.), *The attentive brain* (pp. 353–380). Cambridge, MA: MIT Press.

Ballard, D.H., Hayhoe, M.M., Pook, P.K., & Rao, R.P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20,* 723–767.

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

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence*, *47, 139–159.*

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36,* 28–71.

Chun, M.M., & Nakayama, K. (this issue). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition, 7,* 65–81.

Clancey, W.J. (1997). *Situated cognition* (pp. 101–132). Cambridge, UK: Cambridge University Press.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints with the human information-processing system. *Psychological Review*, *104,* 163–191.

Davis, E.T., & Graham, N. (1981). Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision Research*, *21,* 705–712.

Dennett, D.C. (1991). *Consciousness explained*. Boston: Little, Brown & Co.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences, USA*, *93,* 13,494–13,499.

Deubel, H., & Schneider, W.X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36,* 1827–1837.

DiLollo, V. (1980). Temporal integration in visual memory. *Journal of Experimental Psychology: General*, *109,* 75–97.

Elder, J.H., & Zucker, S.W. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research*, *33,* 981–991.

Enns, J.T., & DiLollo, V. (1997). Object substitution: A new form of masking in unattended visual locations. *Psychological Science*, *8,* 135–139.

Enns, J.T., & Rensink, R.A. (1991). Preattentive recovery of three-dimensional orientation from line drawings. *Psychological Review*, *98,* 101–118.

Enns, J.T., & Rensink, R.A. (1992). An object completion process in early vision. *Investigative Ophthalmology and Visual Science*, *33,* 1263.

Feldman, J.A. (1985). Four frames suffice: A provisional model of vision and space. *Behavioral and Brain Sciences*, *8,* 265–289.

Fernandez-Duque, D., & Thornton, I.M. (this issue). Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition, 7,* 323–344.

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

Garavan, H. (1998). Serial attention within working memory. *Memory and Cognition*, *26,* 263–276.

Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins (Ed.), *Vancouver Studies in Cognitive Science: Vol. 5. Perception* (pp. 89–109). New York: Oxford University Press.

Guérin-Dugué, A., Bernard, P., & Oliva, A. (1998). Search for scale-space salient orientations in real-world scenes. *Perception*, *27* (Suppl.), 151.

Haber, R.N. (1985). Toward a theory of the perceived spatial layout of scenes. *Computer Vision, Graphics, and Image Processing*, *31,* 282–321.

He, Z.J., & Nakayama, K. (1992). Surfaces versus features in visual search. *Nature*, *359,* 231–233.

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

Hochberg, J.E. (1968). "In the mind's eye." In R.N. Haber (Ed.), *Contemporary theory and research in visual perception* (pp. 309–331). New York: Holt, Rinehart, & Winston.

Intraub, H. (1981). Identification and processing of briefly glimpsed visual scenes. In D.F. Fisher, R.A. Monty, & J.W. Senders (Eds.), *Eye movements: cognition and visual perception* (pp. 181–190). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences, 1,* 217–222.

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

Julesz, B. (1984). A brief outline of the texton theory of human vision. *Trends in Neuroscience*, *7,* 41–45.

Kahneman, D., Treisman, A., & Gibbs, B. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24,* 175–219.

Kanwisher, N., & Driver, J. (1992). Objects, attributes, and visual attention: Which, what, and where. *Current Directions in Psychological Science*, *1,* 26–31.

Klein, R., Kingstone, A., & Pontefract, A. (1992). Orienting of visual attention. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 46–65). New York: Springer.

Lavie, N., & Driver, J. (1996). On the spatial extent of attention in object-based visual selection. *Perception and Psychophysics*, *58,* 1238–1251.

Levin, D.T., & Simons, D.J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin and Review*, *4,* 501–506.

Loftus, G.R. (1976). A framework for a theory of picture representation. In R.A. Monty and J.W. Senders (Eds.), *Eye movements and psychological processes* (pp. 499–513). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge MA, MIT Press.

Mandler, J., & Parker, R.E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 38–48.

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

Milner, A.D., & Goodale, M.A. (1995). *The visual brain in action*. Oxford, UK: Oxford University Press.

Moore, C.M., & Egeth, H. (1997). Perception without attention: Evidence of grouping under conditions of inattention. *Journal of Experimental Psychology: Human Perception and Performance*, *23,* 339–352.

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

O'Regan, J.K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, *46,* 461–488.

Pashler, H. (1988). Familiarity and visual change detection. *Perception and Psychophysics*, *44,* 369–378.

Pashler, H., & Carrier, M. (1996). Structures, processes, and the flow of information. In E.L. Bjork & R.A. Bjork (Eds), *Memory* (pp. 3–29). San Diego: Academic.

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

Pylyshyn, Z.W., & Storm, R.W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 179–197.

Ramachandran, V.S. (1988). Perceiving shape from shading. *Scientific American*, *259, 76–83.*

Rensink, R.A. (1992). *The rapid recovery of three-dimensional orientation from line drawings.* Unpublished Ph.D. thesis (also Tech. Rep. 92–25), Department of Computer Science, University of British Columbia, Vancouver, BC, Canada.

Rensink, F.A. (1997). How much of a scene is seen? The role of attention in scene perception. *Investigative Ophthalmology and Visual Science, 38,* S707.

Rensink, R.A. (1998a). Limits to attentional selection for orientation. *Perception*, *27* (Suppl.), 36.

Rensink, R.A. (1998b). Mindsight: Visual sensing without seeing. *Investigative Ophthalmology and Visual Science*, *39,* S631.

Rensink, R.A. (this issue). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition, 7,* 345–376.

Rensink, R.A., & Cavanagh, P. (1993). Processing of shadows at preattentive levels. *Investigative Ophthalmology and Visual Science*, *34,* 1288.

Rensink, R.A., & Enns, J.T. (1995). Preemption effects in visual search: Evidence for low-level grouping. *Psychological Review*, *102,* 101–130.

Rensink, R.A., & Enns, J.T. (1998). Early completion of occluded objects. *Vision Research, 38,* 2489–2505.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8,* 368–373.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. (this issue). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition, 7,* 127–145.

Rojer, A.S., & Schwartz, E.L. (1990). Design considerations for a space-variant visual sensor with complex-logarithmic geometry. In *Proceedings of the 10th International Conference on Pattern Recognition* (pp. 278–285). Washington, DC: IEEE Computer Society Press.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, *8,* 374–378.

Shapiro, K., Driver, J., Ward, R., & Sorensen, R.E. (1997). Priming from the attentional blink. *Psychological Science*, *8,* 95–100.

Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84,* 127–190.

Shulman, G.L., & Wilson, J. (1987). Spatial frequency and selective attention to spatial location. *Perception*, *16,* 103–111.

Simons, D.J. (1996). In sight, out of mind: When object representations fail. *Psychological Science*, *7,* 301–305.

Simons, D.J., & Levin, D.T. (1997). Change blindness. *Trends in Cognitive Sciences, 1,* 261–267.

Smith, B.C. (1998). *On the origin of objects.* Cambridge, MA: MIT Press.

Stroud, J.M. (1955). The fine structure of psychological time. In H. Quastler (Ed.), *Information theory in psychology: Problems and methods* (pp. 174–207). Glencoe, IL: Free Press.

Tanenbaum, A.S. (1976). *Structured computer design*. Englewood Cliffs, NJ: Prentice-Hall.

Trehub, A. (1991). *The cognitive brain*. Cambridge, MA: MIT Press.

Trehub, A. (1994). What does calibration solve? *Behavioral and Brain Sciences*, *17,* 279–280.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance* , *16,* 495–478.

Trick, L.M., & Pylyshyn, Z. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, *19,* 331–351.

Tsotsos, J.K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, *13,* 423–445.

von Grünau, M., & Dubé, S. (1994). Visual search asymmetry for viewing direction. *Perception and Psychophysics*, *56,* 211–220.

Wolfe, J.M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1,* 202–238.

Wolfe, J.M. (1996). Post-attentive vision. *Investigative Ophthalmology and Visual Science, 37,* S214.

Wolfe, J.M. (1997). In the blink of the mind' s eye. *Nature*, *387,* 756–757.

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, *24,* 295–340.

Yeshurun, Y, & Schwartz, E.L. (1989). Shape description with a space-variant sensor: Algorithms for scan-path, fusion and convergence over multiple scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , *11,* 1217–1222.