Vision & Attention. M. Jenkin & L. Harris, eds. New York: Springer. 2001.

9

Change Blindness: Implications for the Nature of Visual Attention

Ronald A. Rensink

In the not-too-distant past, vision was often said to involve three levels of processing: a *low level* concerned with descriptions of the geometric and photometric properties of the image, a *high level* concerned with abstract knowledge of the physical and semantic properties of the world, and a *middle level* concerned with anything not handled by the other two.[1] The negative definition of mid-level vision contained in this description reflected a rather large gap in our understanding of visual processing: How could the here-and-now descriptions of the low levels combine with the enduring knowledge of the high levels to produce our perception of the surrounding world?

A number of experimental and theoretical efforts have been made over the past few decades to solve this "mid-level crisis". One of the more recent of these is based on the phenomenon of *change blindness*—the difficulty in seeing a large change in a scene when the transients accompanying that change no longer convey information about its location (Rensink, O'Regan, & Clark, 1997; Rensink, 2000a). Phenomenologically, this effect is quite striking: the change typically is not seen for several seconds, after which it suddenly snaps into awareness.[2] During the time the change remains "invisible", there is an apparent disconnection of the low-level descriptions (which respond to the change) from subjective visual experience (which does not). As such, this effect would seem to have the potential to help us understand how mid-level mechanisms might knit low- and high-level processes into a coherent representation of our surroundings.

---

[1]More precisely, low-level vision determines the scene-based properties (e.g., surface color and orientation) that give rise to the pattern of illumination on the retina, separating out the effects of extraneous factors such as lighting, occlusion, and noise. The result is a retinotopic sketch containing a detailed description of the visible scene at that moment in time (Marr, 1982). High-level vision, in contrast, involves issues of meaning. Among other things, it involves knowledge of object types, both in regards to how each type related to the others, and how each appeared visually (e.g., automobile tires are almost always black and have an outer perimeter that is approximately round). Mid-level vision must somehow describe the scene using high-level knowledge about the types of object present, and information about particular parameters (location, time, size, etc.) obtained from the low-level descriptions.

[2]To experience this effect, see the QuickTime examples on the accompanying CD-ROM. Examples can also be found at `http://www.cs.ubc.ca/~rensink/flicker`.

It is argued here that this potential can indeed be realized, and that change blindness can teach us much about the nature of mid-level vision.[3] A number of studies are first reviewed showing that the perception of a scene does not involve a steady buildup of detailed representation: rather, it is a dynamic process, with focused attention playing one of the main roles, viz., forming coherent object representations whenever needed. It is then argued that change blindness can also shed considerable light on the nature of focused attention itself, such as its speed, capacity, selectivity, and ability to bind together visual properties into coherent structures.

9.1 Visual Attention: Role in Scene Perception

9.1.1 Change blindness

Change blindness can be defined as the induced failure of observers to detect large changes in a visual display (Rensink et al., 1997). Under normal viewing conditions it is usually easy to notice when an item suddenly changes its color, location, or other attribute. But if the motion transients that accompany the change cannot provide information about its location, even large changes can become difficult to detect.

Consider the *flicker paradigm* of Rensink et al. (1997), shown in Figure 9.1. Here, an original image of a scene alternates with the same image modified in some way (e.g., an item changes color or is moved). If the images alternate without intervening blank fields, the change is easily seen. But if a brief blank (about 80 ms or more) is interposed between them, detection of the change is dramatically impeded. For example, detection of the change shown in Figure 9.1 requires over 40 alternations on average, even though the change is large, made repeatedly, and the observers are actively searching for it. This does not depend on the particular change or the particular image—similar results have been found for different kinds of changes on a wide variety of images (Rensink et al., 1997). In addition, this effect is robust, occurring for a wide range of blank intervals and colors (Rensink, O'Regan & Clark, 2000; see Figure 9.2).

---

[3]It should be mentioned that there is at the moment some non-uniformity in terminology. Some authors (e.g., Ullman, 1996) use "high-level" to refer to object perception, leaving "mid-level" to refer to the perception of relatively unstructured surface properties. Others (e.g., Henderson & Hollingworth, 1999) use "high-level" to refer to the perception that includes meaning (i.e., identifying an object or a scene *as* something), with "mid-level" vision referring to the formation of particular objects. As is hopefully evident, usage here is similar to that of Henderson & Hollingworth: here, "mid-level" refers to the formation of coherent object representations, with the rapid interpretation of scene properties (i.e., the formation of proto-objects, surfaces, etc.) regarded as the final stage of low-level processing.

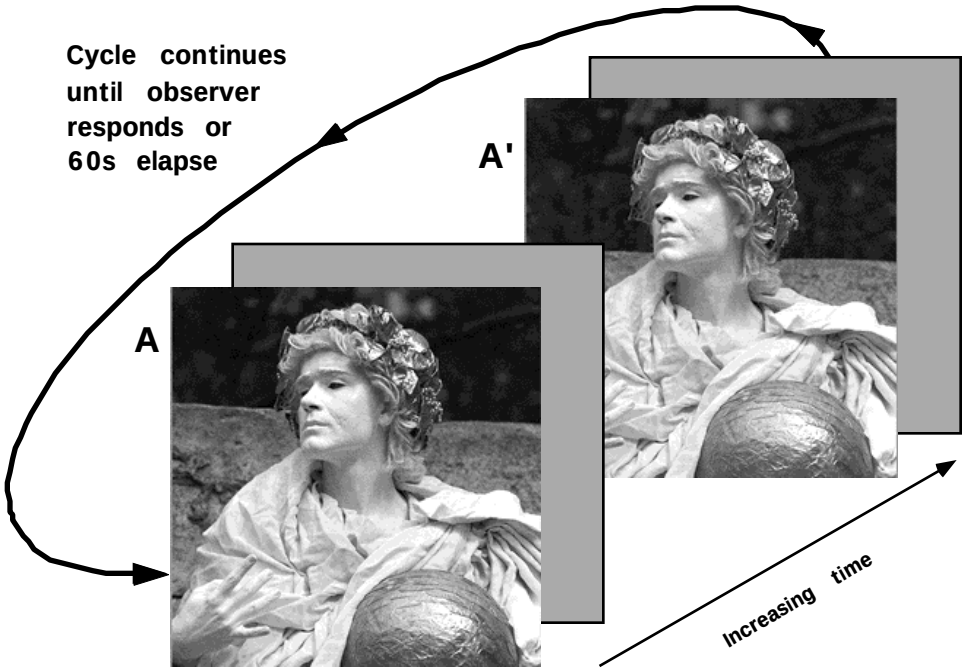

FIGURE 9.1. Example of flicker paradigm. Sequence alternates between original and modified image until observer responds. Display times (on-times) are typically 200 - 600 ms, while blank times (off-times) are 80-800 ms. In the stimulus here, original image A (statue with wall in background) and modified image A' (statue with wall lowered) appear in the order A, A', A, A',...with gray fields placed between successive images. For this example, over 40 alternations are required on average before observer detects the change.

Indeed, change blindness can be induced by a large number of techniques, such as making the change during:

- an eye movement (Bridgeman, Hendry, & Stark, 1975; Grimes, 1996),
- an eye blink (O'Regan, Deubel, Clark, & Rensink, 2000)
- a movie cut (Levin & Simons, 1997)
- occlusion of the changing item (Simons & Levin, 1998)
- small transient "splats" elsewhere in the image (Rensink et al., 2000).

(See Rensink, 2000c or Simons & Levin, 1997 for a more complete review of various change-blindness studies.) Given that change blindness can be induced in many ways and that it has a strong phenomenological effect, it follows that change blindness is not an aberrant phenomenon occurring only under a special set of conditions. Rather, it appears to touch on something important, something central to the way that the world is perceived.

9.1.2 Coherence theory

Why is it that change blindness can be so easily induced? And if it is so easy to induce, why are we nevertheless so good at seeing changes in everyday life?

All the techniques used to induce change blindness share a comment element: the transients associated with the change are swamped (or otherwise neutralized) so that there is little information about its location in the display. This leads to the suggestion that *focused attention is necessary to see change* (Rensink et al., 1997). The world is by and large a quiet

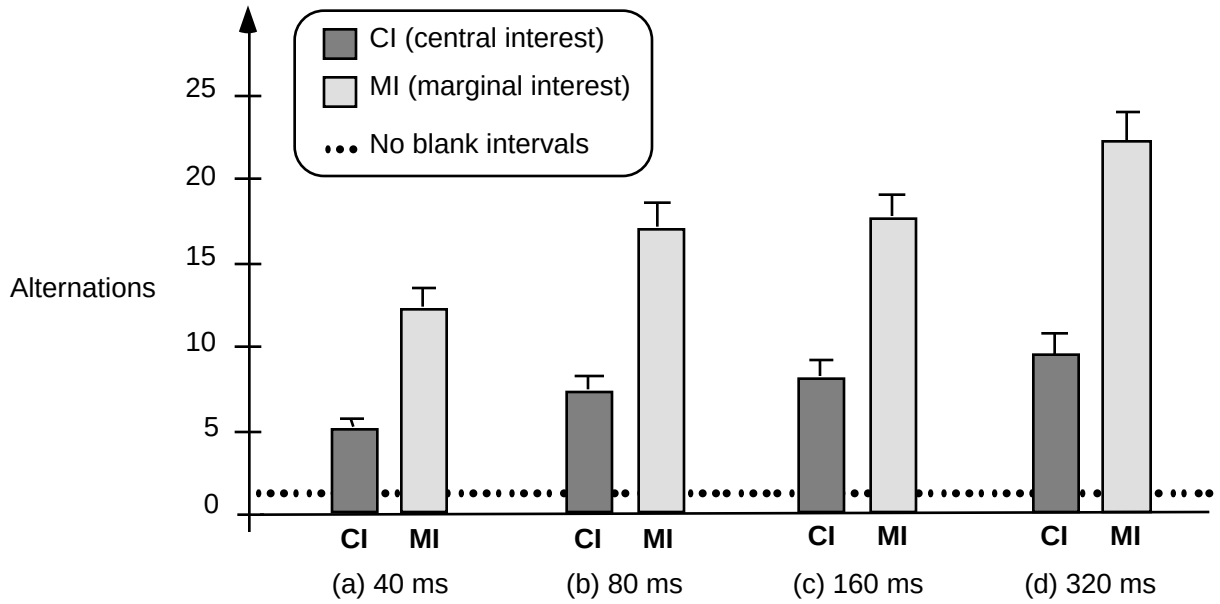

FIGURE 9.2. Time to detect change for various durations of blank intervals. Data from Rensink et al. (2000). CIs are *central interests*, items often mentioned in brief verbal descriptions of the scene; MIs are *marginal interests*, items never mentioned. (See Rensink et al, 1997 for more discussion of these terms.) Error bars indicate one standard error; dotted line indicates performance when no blank intervals present. (a) 40 ms durations. Although intervals were very brief, changes took far longer to detect than when no blanks were present. (b) 80 ms durations. Detection (analyzed in terms of number of alterations to see the change) took even longer than in the 40 ms condition. (c) 160 ms durations. No significant differences were found between this and the 80 ms condition. (d) 320 ms durations. Detection was reliably slower than for the other conditions, although the amount of slowdown was not large.

place—any change will likely be the only one occurring at that moment. As such, it will generate a motion signal in the image that will automatically attract attention to its location (see, e.g., Klein et al., 1992). However, if this signal is swamped by others occurring at the same time, attention will not be automatically sent to the location of the change; instead, a more effortful attentional scan must be used. Since attention can operate on only a few items at a time (e.g., Pashler, 1988; Pylyshyn & Storm, 1988), and since there are many items in most real-world scenes, this scanning will usually take considerable time. The result is change blindness.

Note that this explanation requires that attention cannot "weld" visual features into a detailed, relatively long-lasting coherent[4] representation (Kahneman et al., 1992). Indeed, it suggests that focused attention may endow a structure with a coherence that lasts only as

---

[4]In this chapter, "coherent" refers to consistency and logical interconnection, i.e., agreement that the structures refer to parts of the same system. This term is used in two ways: in its spatial aspect, it denotes a set of representations at different locations that refer to the same object; in its temporal aspect it denotes a set of representations at different times that refer to the same object. (See Rensink, 2000a).

long as attention is directed towards it. This viewpoint is given a more precise formulation as a *coherence theory* of focused attention (Rensink, 2000a):

- Prior to focused attention, structures are formed rapidly and in parallel across the visual field. These *proto-objects* can be quite complex, but have coherent structure only within a limited area of space (Rensink & Enns, 1995, 1998). Their temporal coherence is likewise limited—they are volatile, being constantly regenerated. As such, they are simply *replaced* by any new stimuli appearing in their retinal location.
- Focused attention acts as a metaphorical hand that grasps a small number of these proto-objects from the constantly-regenerating flux. While held, these form an individuated object with a high degree of coherence over time and space.[5] Such coherence is obtained via feedback between the proto-objects and a mid-level *nexus*, a locus where lower-level information is collected via a set of *links*. This allows the object to retain its identity across brief interruptions; as such, it is *transformed* rather than replaced by new stimuli arriving at its location.
- After focused attention is released, the object loses its coherence and dissolves back into its constituent set of proto-objects. This implies that there is little short-term memory apart from what is being attended. Such a position is also consistent with results indicating a lack of attentional aftereffect in visual search (see also Wolfe, 1999).

This view of visual processing is illustrated in Figure 9.3. Note that the separation between low- and mid-level vision is a true divide, with both levels qualitatively different from each other. The low-level processes involve retinotopic representations with a considerable amount of visual detail[6], but which are also *volatile*, lasting only a fraction of a second. Unattended proto-objects are therefore in constant flux, being built anew as long as light continues to enter the eyes. In contrast, the mid-level attentional processes involve a much sparser set of structures that are stable[7], both in time (lasting as long as attention is directed to them) and in space (remaining invariant with eye movements).

According to coherence theory, then, low-level representations of considerable detail—proto-objects—exist as long as the scene continues to be projected to the eyes. Change

---

[5]In this view, focused attention is taken to be object-based rather than spaced-based. Part of the justification for this can be seen from the example in Figure 9.1. Here, it takes considerable time before the change to the background wall is seen. If focused attention involved a spotlight of the type proposed in most space-based theories (e.g., Treisman & Gormican, 1988), some of it ought to "spill over" onto the wall while examining the statue, resulting in relatively fast detection of change. However, this does not occur, indicating that attention is allocated to relatively discrete structures (proto-objects) that correspond to various objects in the scene.

[6]The amount of detail in the representation will fall off with eccentricity from fixation. However, there is still considerable resolution even at eccentricities of several degrees (see e.g., Woodhouse & Barlow, 1982).

[7]In this chapter, "stable" is used in two ways: In its spatial aspect, it denotes invariance over eye movements; in its temporal aspect it denotes invariance over time, i.e., a representation that is not volatile. Note that both properties are required for a buffer that collects information into a representation independent of any particular viewing position.

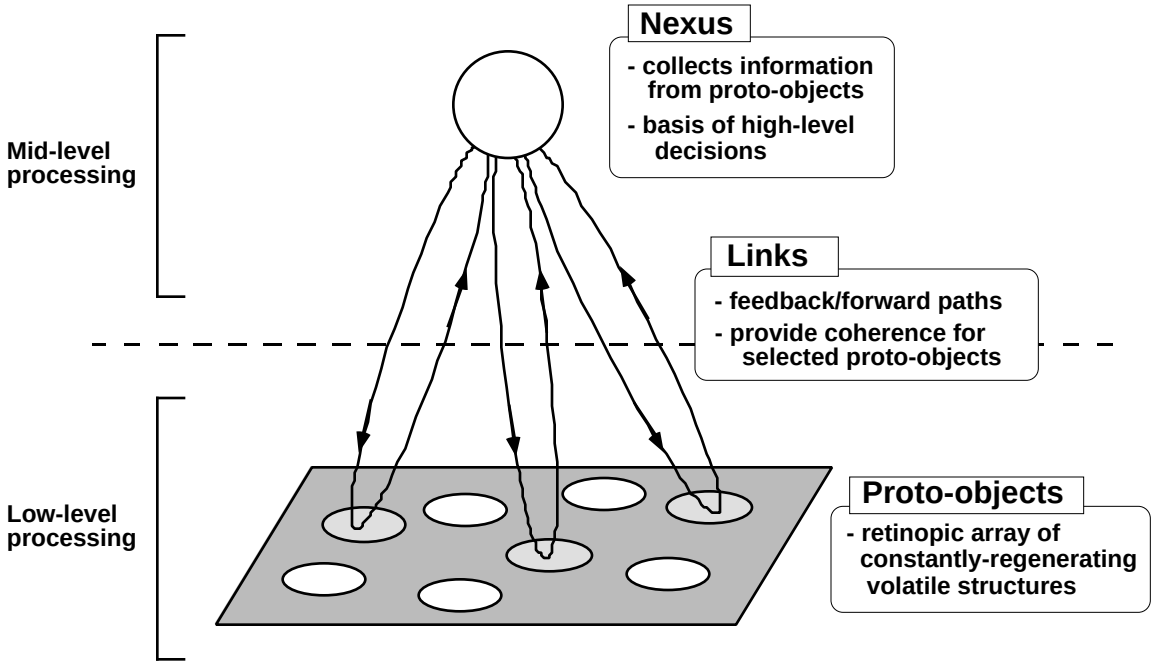

FIGURE 9.3. Relation of low- and mid-level processing. In the absence of focused attention, low-level structures (proto-objects) are volatile. Attention acts by establishing feedback links between proto-objects and a mid-level nexus. The set of interacting proto-objects, links, and nexus is a coherence field. The interaction among the various parts of the field allows the establishment of coherence in the properties of the selected proto-objects, both in space and in time.

blindness stems not from a lack of detailed representation, but rather from an inability to make the entire set of proto-objects coherent enough to support the perception of change. It is important to note that although unattended proto-objects may be volatile and so cannot be directly reported, they can still provide an immediate context that influences the perception of the attended structures that *are* reported (Moore & Egeth, 1997).

9.1.3 Virtual representation

Change blindness shows that observers are poor at combining information from successive images—they can neither detect a difference directly, nor detect the superposition of items that would occur if information was accumulated. This indicates that there is no large-scale, detailed buffer into which information is collected (see Irwin, 1991). Indeed, coherence theory claims that no more than a few coherent object representations exist at any time. But if this is so, how can a scene be represented? And why do we have the impression that we observe a large number of coherent objects simultaneously?

The answer to these questions centers around the idea of a *virtual representation*: instead of creating a detailed coherent representation of all the objects in the scene, do so only for those few objects needed for the task at hand. If a coherent representation of an object can be formed whenever requested, the resultant representation will appear to higher levels as if "real", i.e., as if all objects simultaneously have a coherent representation. Such a scheme will

have almost all the power of a complete set of object representations, while requiring far less in the way of processing and memory resources (Rensink, 2000a).

The success of a virtual representation depends on its ability to form a coherent representation whenever requested. Fortunately this is easy to do, at least in principle: when a request is made, shift attention and the eyes to the location of the object in the image; then obtain detailed information from the incoming light and incorporate it into a coherent representation. Note that a high-capacity memory is not needed—detailed information about any object in the scene can almost always be obtained from the world itself. As pointed out by Stroud (1955, p. 199):

> *"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…"*

Provided that attention and eye movements are properly co-ordinated, the result will be a virtual representation whose dynamic nature is transparent to processes at higher levels.

Given this, the question remains of how such co-ordination might be carried out. One possibility is an architecture in which there exists not only an attentional stream for the processing of objects, but also a concurrent nonattentional stream that forms stable structures for attentional guidance (Rensink, 2000a). The resulting system—shown in Figure 9.4—has three subsystems. Each of these obtains its input from low-level vision, and is informed by the abstract knowledge available at high levels:

- a limited-capacity attentional system that forms low-level proto-objects into coherent, low-capacity object representations.
- a limited-capacity nonattentional system that uses the statistics of the proto-objects to determine the gist of the scene (i.e., its abstract meaning.
- a limited-capacity nonattentional system that uses the locations of the most significant proto-objects to determine the layout of objects in the scene.

In this architecture, attention retains the role assigned to it by coherence theory, viz., the temporary formation of coherent objects. What has now been added is a *setting system* that maintains both scene layout and gist, but involves little (if any) focused attention. The feasibility of such a system is suggested by several studies showing that the extraction—or at least the maintenance—of gist and layout information does not require attentional processing (see e.g., Henderson & Hollingworth, 1999; Rensink, 2000a). In this view, the mid-level processes that link low-level structures with high-level knowledge do not form a unitary system. Rather, they form a heterogeneous collection of subsystems, each with its own particular characteristics.

Given this architecture, a relatively simple set of interactions could carry out the visual perception of the scene. First, low-level processes would provide a constantly-regenerating description of the scene-based properties visible to the viewer. A subset of these could determine scene gist and layout, which could then invoke high-level knowledge about the objects and events that might be expected. These expectations could be tested via attention, which could provide detailed, coherent descriptions of selected objects, with the expected structure (and importance) of these objects being obtained from high-level knowledge. The perceived layout of the scene could facilitate this process, helping to guide attention to the appropriate low-level items.

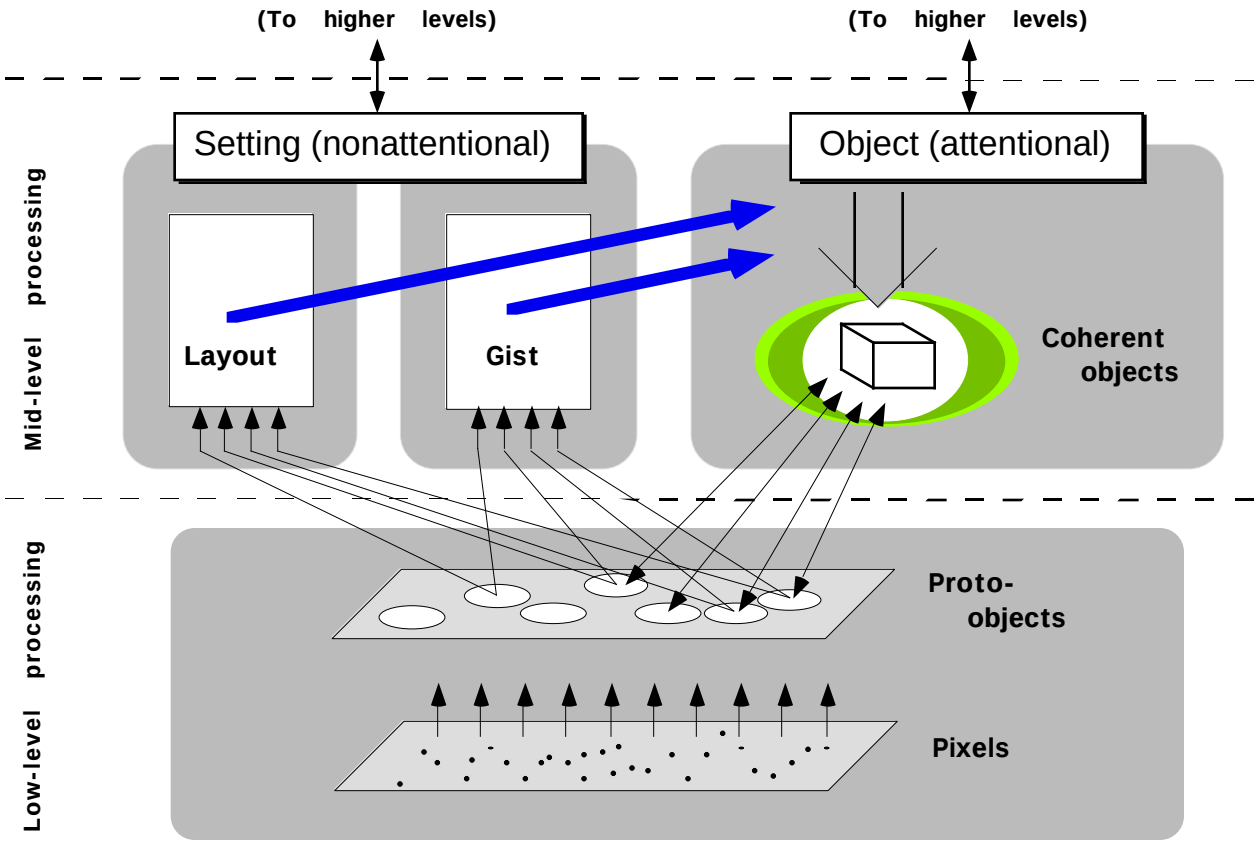

FIGURE 9.4. Architecture implementing virtual representation. Three different mid-level systems are involved: (i) an attentional system that gives selected proto-objects temporal and spatial coherence; (ii) a nonattentional setting system that uses low-level information to obtain the gist of the scene; (iii) a nonattentional setting system that uses low-level information to obtain the spatial layout of the objects in the scene. The latter two systems operate in tandem with the attentional system, and provide information that helps guide it, so that it can form the appropriate coherent representation when requested.

Note that in this view of mid-level vision, focused attention still plays an important role, viz., the formation of coherent object representations. But the role of these representations has been reduced. Coherent representations are no longer needed for the recognition of various types of scenes—this can be largely done via the statistics of relatively simple properties. Indeed, coherent representations may not even be needed to recognize objects—at least in regard to object *type*. In other words, it may be possible to determine the presence of an object simply via the statistics of the properties in some part of the image. Coherent representations would then be primarily involved in describing particular *instantiations* of these types, assigning them temporal and spatial co-ordinates, as well as—via the links to the proto-objects—properties such as size and color. The formation of these kinds of descriptions would be necessary only for a more restricted set of operations, such as the intensive verification (scrutiny) of perceptual hypotheses about objects and scenes, and perhaps the selection and guidance of actions.

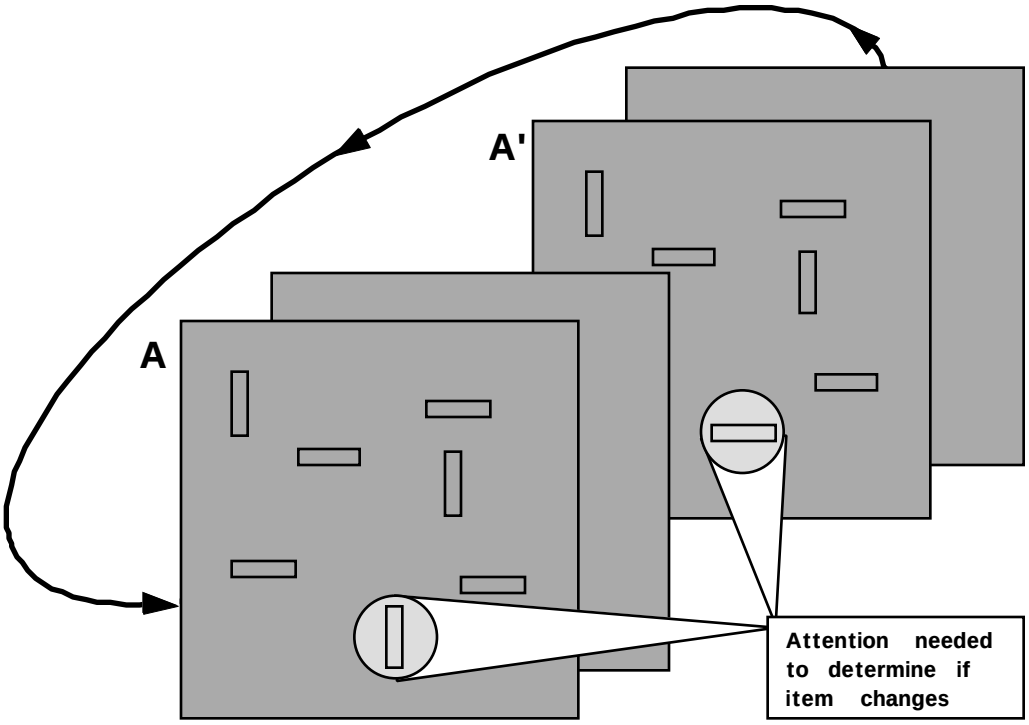

FIGURE 9.5. Example of flicker paradigm with controlled displays. Here, displays are arrays of rectangles. In half the displays, one item (the target) changes orientation while the others (the distractors) remain constant. According to coherence theory, the only way for an observer to determine if a change is occurring is to carry out an attentional scan of each item in the display.

9.2 Visual Attention: Mechanisms

9.2.1 Methodology

Change blindness not only provides a way to determine how focused attention relates to the rest of vision—it can also provide a way to explore the nature of attentional processing itself.

One particularly powerful way to do this is by replacing the images of scenes in the original flicker paradigm (Section 9.1) with arrays of simple figures (Rensink, 2000b). To see how this works, consider an array of rectangles, where one item (the target) changes its orientation on half the trials while the other items (the distractors) do not; the observer must then report for each trial whether or not a change was present (Figure 9.5). If roughly half the items are horizontal and half vertical in both images, the target cannot be detected from any single display—both displays must be compared. If the interstimulus interval (ISI) between displays is sufficiently long, the transients due to the changing target will be swamped by the transients produced by the flickering distractors, requiring an attentional scan to carry out the task.

This approach therefore extends the "classic" visual search paradigm on static displays (e.g., Treisman & Gormican, 1988) into a more dynamic realm. All the power of the static techniques (e.g., investigating different shapes, different features, search asymmetries) is retained, while allowing manipulation of two more degrees of freedom: display time (on-

time) and ISI (off-time). This methodological power allows various aspects of attentional processing to be investigated, including:

- **speed**. This can be determined from the search rate, i.e., the time required per item in the display. Note that the influence of salience can often be eliminated by proper choice of items in the display, such as when half the items have one value of a property (e.g., vertical) and the other half the other value (e.g., horizontal).

- **capacity**. This can be determined by increasing the amount of on-time in each cycle. As on-time increases, more items can be "grabbed" by attention, until saturation is reached. The value of this asymptote is a measure of attentional capacity (Rensink, 2000b). Note that this estimate is an upper bound on the number of attentional links involved, since grouping factors may lead to chunking, causing more than one stimulus item to be assigned to each link.

- **selectivity**. This can be determined by comparing the speed when all items must be examined against the speed when the change occurs in a selected subset. For example, the speed for orientation change can be measured for a set of black and white items; it can then be measured for the same items, but with the change occurring only in the black ones. The ratio of these two speeds is a measure of the selectivity for black (Rensink, 1998).

- **basic codes**. The basic "building blocks" (or codes) for coherent objects can be determined by comparing different types of changes and different kinds of items (see Rensink, 2000b). Changes in basic codes are indicated in three ways: (i) speed is relatively high, since the least amount of coding and comparison is required; (ii) capacity is relatively high, since only a minimal description needs to be stored; (iii) selectivity is efficient, since a minimum number of codes need to be excited or inhibited. It remains an interesting empirical matter to determine if the set of codes is equivalent to the set of "features" obtained from studies of search speed on static displays (e.g., Treisman & Gormican, 1988).

- **task dependence**. Different tasks can be done using these flickering displays. These include not only detection (reporting *if* there is a change somewhere in the display), but also identification (reporting *what* the change is), and localization (reporting *where* is it). Although it might be thought that all these tasks should lead to similar estimates of attentional ability, such is not the case—for example, capacity estimates for identification are always below those for detection (Wilken, Mattingley, Korb, Webster, & Conway, 1999).

In addition to this, the stimuli themselves can have different levels of complexity. Importantly, the level of complexity can be varied smoothly, allowing experiments to progress in a straightforward way from simple tasks on highly-controlled arrays to more natural tasks on images of complex, real-world scenes.

9.2.2 *Experimental results: Capacity*

To illustrate how the approach described above can help map out the mechanisms of visual attention, consider the issue of attentional capacity, i.e., how many items can be attended at any one time. One way to determine this is by measuring the speed of detecting a changing target among non-changing distractors. If only one item can be attended at a time, search can only examine one item per alternation; the search rate then equals the alternation rate. If two items can be held, search should be twice as fast as the alternation rate. More generally, if h items can be held, search and alternation rates will be related by

$$\text{search rate} = (\text{alternation rate}) \; / \; h \; .$$

This can be rewritten as

$$h = (\text{alternation rate}) \; / \; (\text{search rate}),$$

where the measure h (attentional hold) describes how many items on average are held and compared across the temporal gap. The attentional capacity C is the asymptotic value of h as display time increases.

Note that h is an average measure—it does not, for example, allow us to distinguish between h items processed at each alternation, $2h$ items every other alternation, or $3h$ items every third alternation. Only when h reaches its asymptotic value (i.e., the capacity) is it possible to state unequivocally that h items are being processed at each alternation.

In a set of experiments on search for changing orientation (Rensink, 1999, 2000b), search rates (and therefore values of h) were measured for a set of on-times. The results (Figure 9.6) show the existence of two different realms. The first is the set of on-times of less than about 600 ms. Here, h increases linearly with on-time, corresponding to a constant search rate of about 100 ms/item); performance is evidently limited by the processing needed to read in and compare the items. As such, this realm is called the *processing range*. The second realm is the set of on-times of more than 600 ms. Here, h reaches an asymptote, indicating that performance is now governed primarily by the number of items that attention can hold. The particular value of the asymptote is about 5.5, indicating that information is collected from no more than 5 or 6 items at a time.[8]

Consider now the reverse situation, where search is for a non-changing target among changing distractors (Rensink, 1999). As Figure 9.7 shows, two realms are again found: (i) for on-times of less than about 300 ms, h increases linearly with on-time (corresponding to a rate of about 330 ms/item); (ii) for on-times of more than 300 ms, h is at an asymptotic value of about 1.4. Although this estimate of capacity is above 1, it never reaches 2, suggesting that

---

[8]Capacity measured in this way may include effects of grouping (or chunking). If used to estimate the number of stimulus items that can be held, this measure is perfectly suitable, being a simple operational description. However, if used to estimate the number of independent links involved, things are less straightforward, since grouping will often cause more than one stimulus item to be assigned to a link. In this latter case, the measured capacity will be an upper limit on the number of links involved.

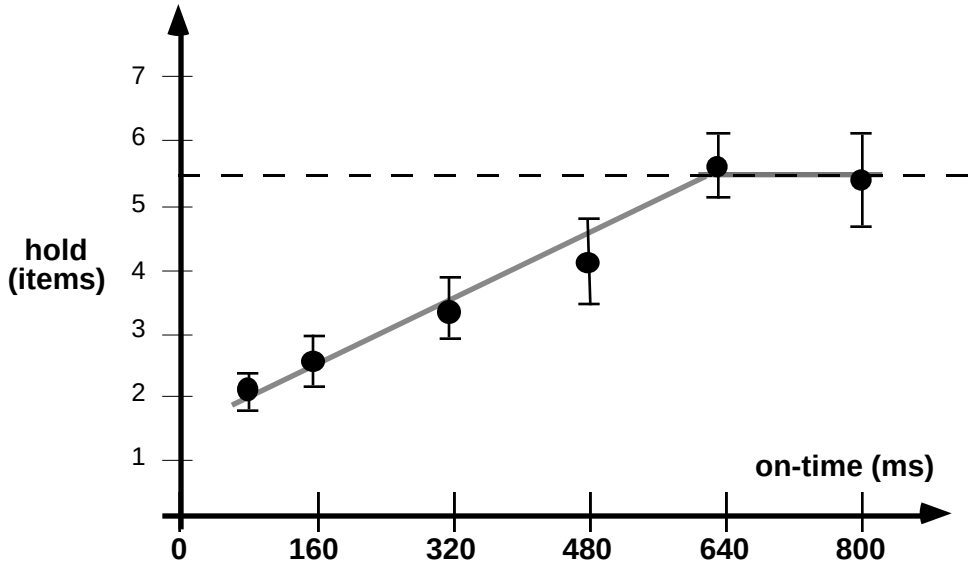

FIGURE 9.6. Attentional search for presence of change (orientation). Data from Rensink (2000b); off-time for these conditions is 120 ms. Hold increases linearly with on-time up to about 600 ms, indicating that search rates are approximately constant in this realm. (Search speed is approximately 100 ms/item.) For longer on-times, hold reaches an asymptotic value of 5.5 items.

only one link at a time is being used. The excess of 0.4 items could result from a variety of factors, such as the grouping of neighboring structures.[9]

Evidence in support of this latter possibility is found in conjunction experiments. Here, the target changes in two dimensions (orientation and contrast polarity) simultaneously, whereas each distractor changes in just one (orientation or polarity). For this situation, the grouping of neighboring structures is less helpful. The pattern of two realms again emerges, as shown in Figure 9.8: (i) for on-times of less than about 300 ms, *h* increases with on-time (corresponding to a search rate of about 450 ms/item); (ii) for on-times of more than 300 ms, *h* again has an asymptote. However, the asymptote is now closer to 1 (Rensink, 1999), in accord with the reduced influence of grouping.

Bringing these results together, it appears that search for the presence of a change leads to a relatively high estimate of capacity (5-6 items), whereas search for the absence or conjunction of a change leads to a much lower estimate (1 item). Note that this pattern—obtained for attended structures—has an interesting similarity to that found for preattentive structures, where search for the presence of a distinctive feature is relatively easy, whereas search for the absence or the conjunction of a feature requires a slower, more effortful scan of the display (Treisman & Gormican, 1988).

---

[9]For example, consider the case where two adjacent items have the same orientation. It might be that these two items form a single group (pair) containing parallel items; if so, the pair could maintain its structure when the orientation of both its constituents changes. In this case, then, the detection of a nonchanging item could be signalled by a transition from a pair of parallel items to a pair of orthogonal ones.

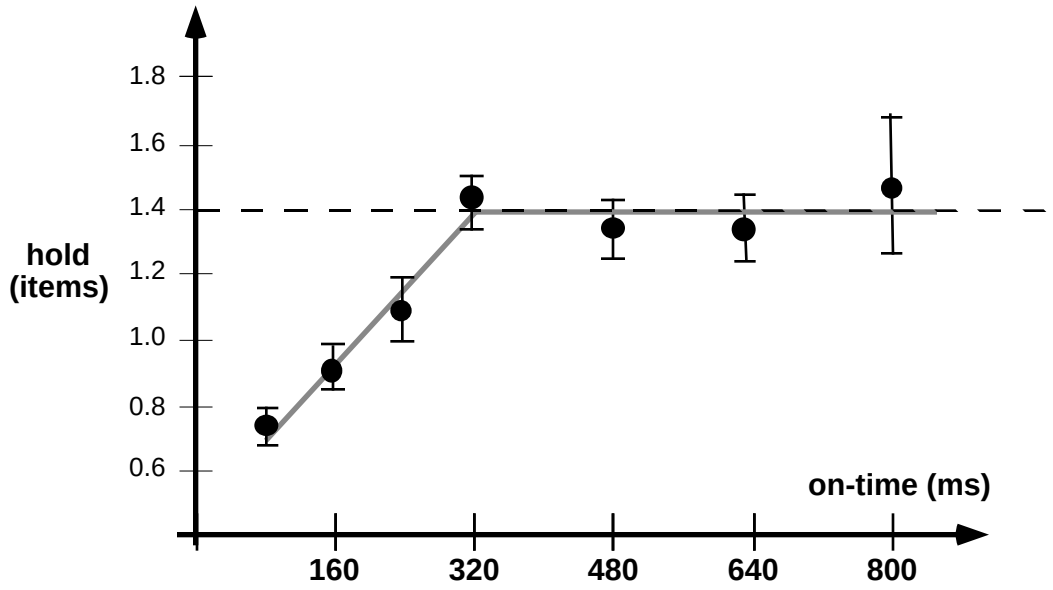

FIGURE 9.7. Attentional search for absence of change (orientation). Data from Rensink (1999); off-time for these conditions is 120 ms. Hold increases linearly with increasing on-time up to about 300 ms. (Search speed in this realm is approximately 300 ms/item.) The asymptote here is 1.4 items.

9.2.3 Implications for attentional mechanisms

The pattern of results for attentional capacity supports two proposals about the nature of focused attention: (i) the *singularity thesis*, the thesis that only one object is attended at any time (i.e., that there is only one nexus), and (ii) the *aggregation thesis*, the thesis that some of the information from each attended item is not kept separate (i.e., bound to that item), but is pooled into an aggregate description at the nexus.[10]

The justification for the singularity thesis rests on the finding that search for the absence of change has a severe capacity limit: 1.4 items. Although this value is slightly more than one, it is clearly less than two. Importantly, it stays constant over a fairly wide range of on-times (300-800 ms). It therefore appears that this task involves only one link at a time, with the excess of 0.4 items likely due to factors such as grouping. It is worth pointing out that search in the processing range (i.e., search not limited by memory) takes about 330 ms/item; if grouping is taken into account, this corresponds to a "raw" rate of about 330 x 1.4 = 462 ms/item, a value close to the attentional dwell time (Ward, Duncan, & Shapiro, 1996). This suggests the involvement of a process that can handle only one item at a time[11], with the

[10]Note that both these theses are independent of each other. The singularity thesis concerns the number of "final destinations" for the pooled information, or equivalently, the number of mid-level representations that can be said to be genuinely independent. This could be one, but it could also—at least in theory—be two, three, or some other positive integer. Meanwhile, the aggregation thesis concerns the degree and type of pooling done via the links. In principle it could well be, for example, that no pooling is done (i.e., only one link is used for each nexus), but that more than one nexus exists.

[11]It is possible to attend to more than one *item* in the image, at least if simple operations are involved (e.g., tracking the location of each item). However, the set of attended items feeds into the same nexus; it is not possible to treat them as completely separate *objects*. For example, when threading a

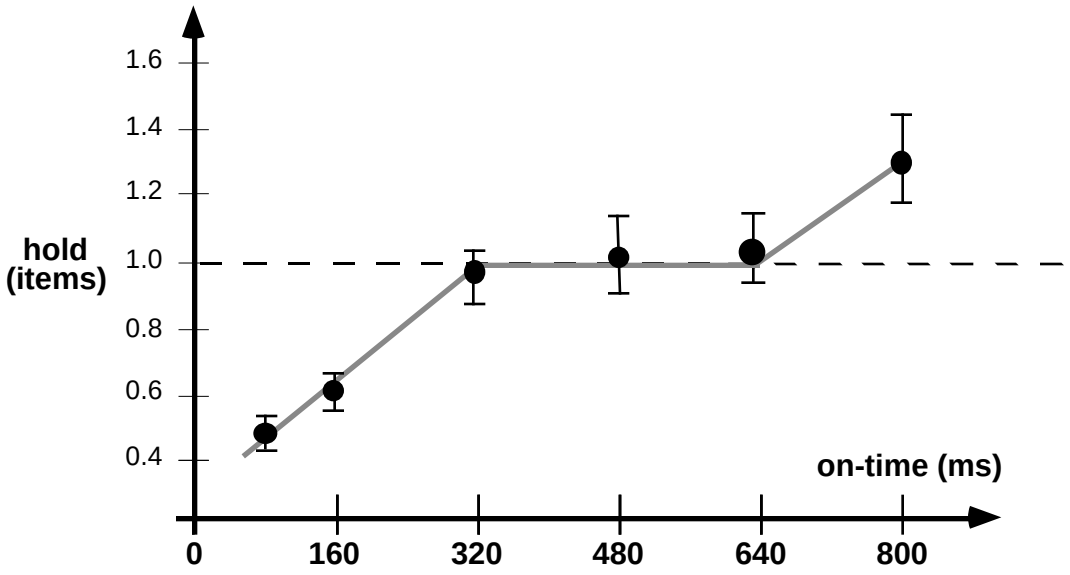

FIGURE 9.8. Attentional search for conjunction of change (orientation and contrast polarity). Data from Rensink (1999); off-time for these conditions is 120 ms. Hold increases linearly with increasing on-time up to about 300 ms. (Search speed in this realm is approximately 450 ms/item.) The asymptote here is close to just 1 item.

processing of the first item to be completed before the next can begin.

To see how the aggregation thesis can be justified, consider how information about attended items might be represented. One way (*weak aggregation*) is for each of the *n* links to maintain its own memory, with the link signalling to the nexus whenever a change occurs to the corresponding proto-object; pooling is done only over these change signals (Figure 9.9). When searching for the presence of a change, the nexus signal will be 1 for target present and 0 for target absent, even when all links are pooled. When searching for the absence of change, however, the signal would be *n-1* for target present and *n* for target absent. If *n* is 2 or more, this would lead to a weaker signal. A more effective strategy would then be to monitor only one link at a time, which would yield a nexus difference of 1 vs. 0. Note that this argument is similar to that used to explain why in standard search the presence of a low-level feature is detected more quickly than its absence (Treisman & Gormican, 1988).

Another way that attended information might be represented is via the pooling of selected properties of the proto-objects themselves (*strong aggregation*), with the nexus containing an aggregate description, such as a distribution of pooled values (e.g., three of type "a" and two of type "b"; see Figure 9.10). Search for the presence of change would involve detecting a change in this distribution, something that could be done if the comparison operation were sufficiently sensitive. (The limit on this sensitivity may be related to the limit on the number of links.) But search for the absence of change is far less tractable. For

---

needle both the needle and the thread must be attended. However, they are not independent objects, but parts of a single needle-thread system. Note that this is not always counterproductive. In the case of the needle and thread, the fact that their locations are sent to the nexus allows accurate calculation of the relative distance between them, thereby facilitating the threading of the needle.

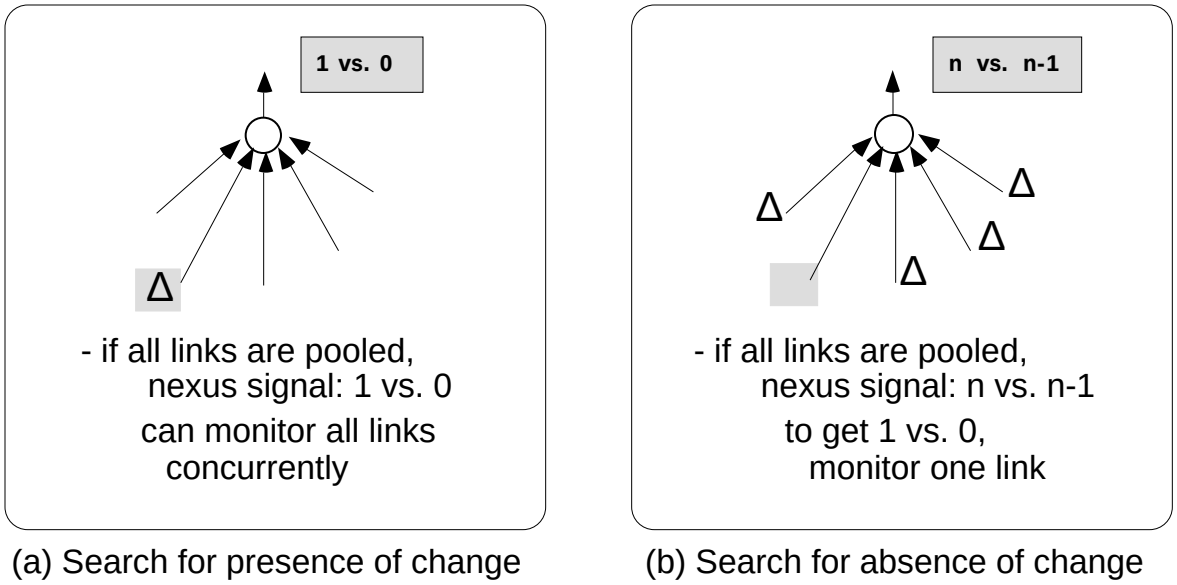

FIGURE 9.9. Explanation of search asymmetry (weak aggregation - pooling of change signals only). (a) When searching for the presence of change, the pooled signal will either be 1 (target present) or 0 (target absent). As such, there is a relatively strong signal at the nexus, even when information from several links is collected. (b) When searching for the absence of change, the pooled nexus signal will either be n-1 (target present) or n (target absent). If n is more than 2 or 3, this signal would be quite weak. To obtain a strong signal, the nexus must collect information from only one link at a time.

example, if one attended item is the target (i.e., constant) and the number of attended distractors is even the distribution may or may not change, depending on the particular items attended (Figure 9.10). A better strategy might therefore be to monitor only one item at a time. In this case, the aggregate description is exactly that of the attended item, and so can be directly used to determine whether or not that item is changing.

The finding that one item at a time is attended when detecting a conjunction of change leads to a further result: either the nexus gathers all the information from its links in parallel (thereby preventing the testing of individual links), or else each link fails to bind to it the different properties it contains. If neither of these were true—i.e., if all the properties of each link were bound to it, and if each attended link could be tested in turn—it would be possible to examine 5-6 items at each alternation, given sufficient time. The finding that only one item at a time is examined, however, means that at least one of these assumptions is incorrect. At the moment it is not known which of the two it is. It is worth noting, however, that the explanation based on lack of binding echoes the proposal for an absence of binding between the properties of each unattended low-level item, requiring an item-by-item search for a feature conjunction (Treisman & Gormican, 1988).

Although the results described above provide strong support for aggregation, they are not sufficient to determine if it is weak or strong. Settling this issue is an important task for future work. If aggregation is strong, it implies that focused attention acts by forming a partition between attended and unattended items, with the information from the attended items pooled into a single description. Note that such a scheme would solve the binding

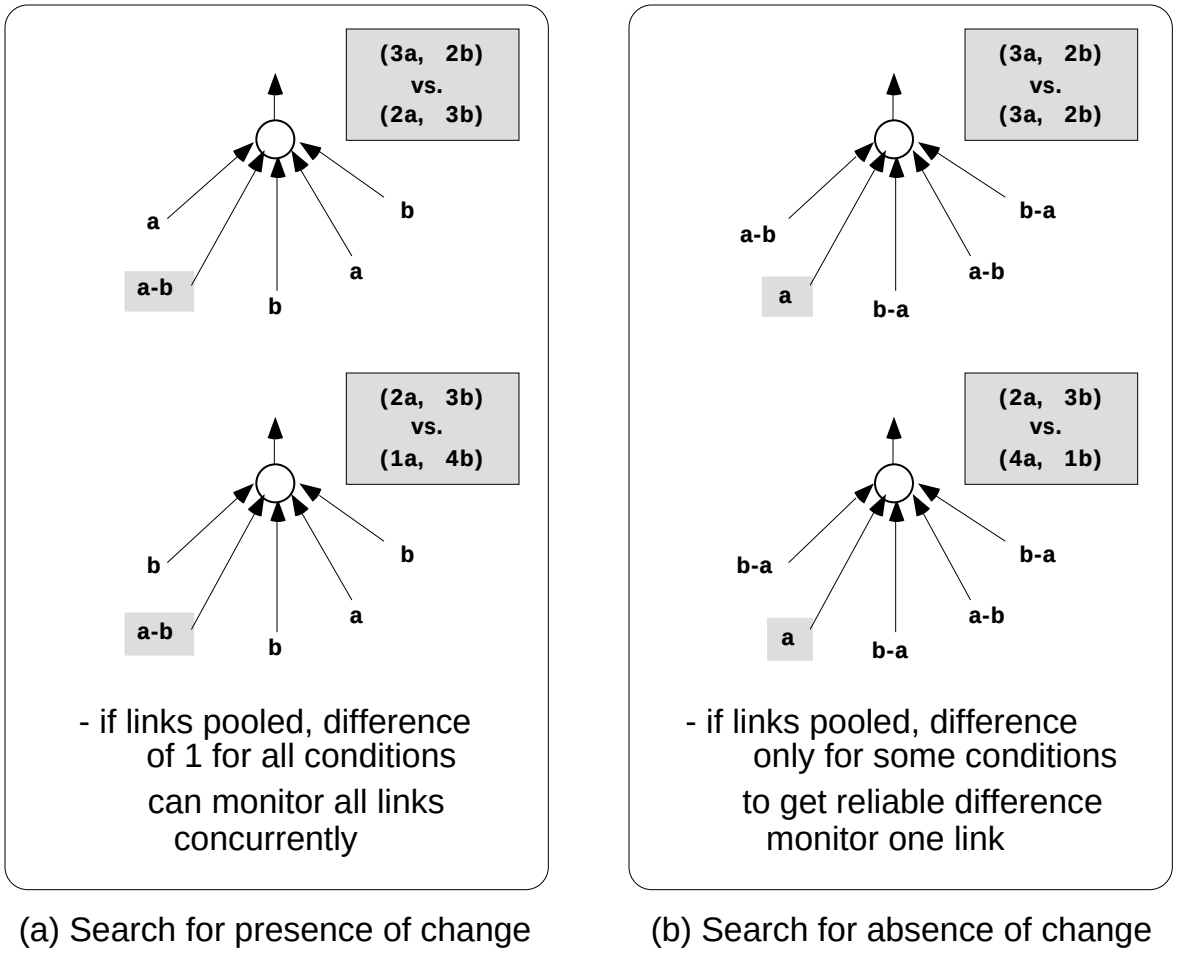

FIGURE 9.10. Explanation of search asymmetry (strong aggregation - pooling of proto-object properties). In these examples, a dash indicates change in successive displays; the letter to the left of the dash represents the first value, and the letter to the right the second. (a) Searching for the presence of change. For this situation, a change in one of the attended items always results in a change in the pooled signal (see upper and lower figures for a few examples). Thus, if the comparison process at the nexus is sensitive enough, a change in one of the links (i.e., that of the target) will always be detectable. (b) Searching for the absence of change. Here, aggregation will do one of several things, depending on the particular items linked. For example, for the values in the upper figure, the aggregate description stays the same; for the values in the lower figure, it changes considerably. Thus, monitoring the aggregate description will not always indicate when a change has occurred. To get a reliable signal of "no change" under all conditions, the nexus must collect information from only one link at a time.

problem (i.e., the question of how to assign different properties to different attended structures) simply by having no more than one discernible structure (i.e., the nexus description) in play at any time. Assigning properties to different parts of an object could be done by switching between individual parts as needed. If so, this the perception of objects would occur via a virtual representation similar to that used for perceiving scenes.

9.3. Concluding Remarks

It has been argued here that the phenomenon of change blindness has considerable potential to cast light on the way that low-level descriptions link up with high-level knowledge to result in our perception of the world. This argument centered around the proposal that focused attention is needed to perceive change. Given this, it becomes possible to use change blindness to explore attention itself, both in regards to the role it plays in the perception of our surroundings, and to the mechanisms that underlie it.

In regards to the involvement of attention in scene perception, change-blindness studies show that we do not build up a detailed picturelike representation of the scene; rather, attention provides a coherent representation of only one object at a time. To account for the fact that we subjectively experience a large number of coherent objects simultaneously, it was suggested that scene perception involves a *virtual representation* which provides a limited amount of detailed, coherent structure whenever required, making it seem as if all the detailed, coherent structure is present simultaneously. In this view, vision is an inherently dynamic process, a "just in time" system whose detailed representations are in constant flux. And focused attention is an important part of its operation, providing coherent descriptions whenever requested.

It was also shown how the original flicker paradigm could be modified to explore various aspects of attentional processing. Initial results showed that focused attention involves the collection of information into a single aggregate description. They also showed that the information contained in this description is obtained via the pooling of information from a small number of links (no more than 5 or 6). It also appears that the properties obtained from the links may not be bound together, although further experiments are needed to establish this latter point conclusively.

Taken together, these results lead to a marked shift in our understanding of mid-level vision. Rather than being a nebulous domain defined mostly in terms of what it is not,[12] it now becomes possible to map out several of its characteristics, and to connect it with other systems known to be involved with scene perception. One of the main messages emerging from this new view is that mid-level vision is heterogeneous and highly dynamic. Consequently, there is little point in looking for a complete coherent representation of a scene or an object. Instead, it is likely that only a small amount of information is ever put into coherent form at a time, with this amount being exactly that needed for the purpose at hand. (See also Ballard, 1991.)

In this view, then, mid-level vision is more concerned with the management of information than with its integration or storage. This leads to questions such as how information is retrieved when requested, how much information is maintained in the representations that guide retrieval, and how operations can be co-ordinated to minimize collisions between responses to different requests. The answers to these are, of course, unknown at the present time. But at least we have begun to ask these kinds of questions. Finding the answers to them may provide us with new insights into the mystery of how we turn the stream of photons entering our eyes into a vivid experience of our surroundings.

---

[12]The earlier characterization of mid-level vision tended to come perilously close to Gertrude Stein's famous quip about Oakland, California: "There is no there, there."

Acknowledgements

I would like to thank Ian Thornton and Carol Yin for their comments on an earlier draft of this chapter. I would also like to thank Carol Yin for her help in preparing QuickTime examples of the flicker paradigm for the accompanying CD-ROM. (Examples are also available at `www.cs.ubc.ca/~rensink/flicker`.)

References

Ballard, D.H. (1991). Animate vision. *Artificial Intelligence*, 48, 57-86.

Bridgeman, B., Hendry, D., & Stark, L. (1975). Failure to detect displacement of the visual world during saccadic eye movements. *Vision Research*, 15, 719-722.

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

Henderson, J.M. (1992). Object identification in context: The visual processing of natural scenes. *Canadian Journal of Psychology*, 46, 319-341.

Henderson, J.M. & Hollingworth, A. (1999) High-level scene perception. *Annual Review of Psychology*, 50, 243-271.

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

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.

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

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

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.

O'Regan, J.K, Deubel, H., Clark, J.J., & Rensink, R.A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7, 191-211.

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

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

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

Rensink, R.A. (1999). The Magical Number One, Plus or Minus Zero. *Investigative Ophthalmology & Visual Science*, 40, 52.

Rensink, R.A. (2000a). The dynamic representation of scenes. *Visual Cognition*, 7, 17-42.

Rensink, R.A. (2000b). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition*, 7, 345-376.

Rensink, R.A. (2000c). Seeing, Sensing, and Scrutinizing. *Vision Research*, 40, 1469-1487.

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. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition*, 7, 127-145.

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

Simons, D.J. , & Levin, D.T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 5, 644-649.

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.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95, 15-48.

Ullman, S. (1996). *High-level Vision* (pp. xi-xii). Cambridge, MA: MIT Press.

Ward, R., Duncan, J., & Shapiro, K. (1996). The slow time-course of visual attention. *Cognitive Psychology*, 30, 79-109.

Wilken, P., Mattingley, J.B., Korb, K.B., Webster, W.R. & Conway, D. (1999). Capacity limits for detection versus reportability of change in visual scenes. Paper presented at the 26th Annual Australian Experimental Psychology Conference, April, 1999.

Wolfe, J.M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting Memories*. (pp. 71-94). Cambridge, MA: MIT Press.

Woodhouse, J.M., & Barlow, H.B. (1982). Spatial and temporal resolution and analysis. In H.B. Barlow and J.D. Mollon (Eds.), *The Senses*. (pp. 133-164) Cambridge, England: Cambridge University Press.