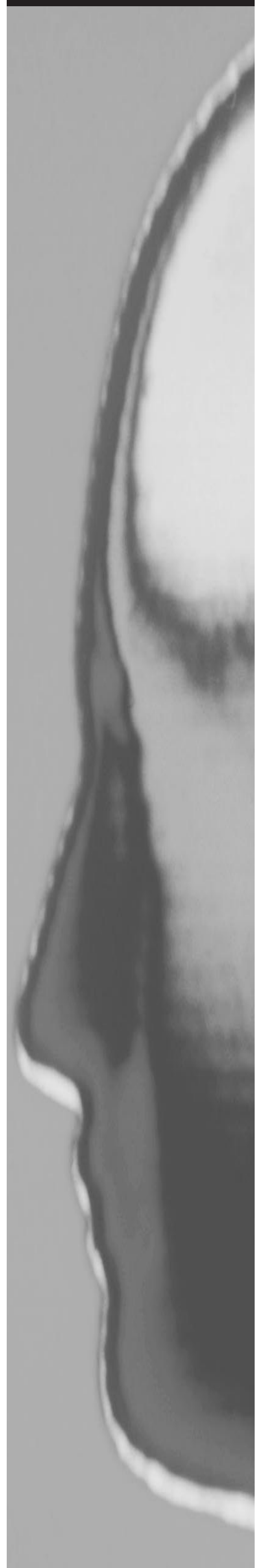


Situating vision in the world

Zenon W. Pylyshyn

Recently, there has been a great deal of interest in what has been called 'situated cognition', which has included claims that certain forms of representation are inadequate for modeling active organisms or agents such as humans and robots. In this article, I suggest that a weakness in classical theories of visual representation is the way in which representations connect with the real world, which may account for many of the concerns expressed by the situated cognition community. Specifically, I claim that what current theories lack is any provision for a certain form of direct, preconceptual connection between objects in the visual world (visual objects or proto-objects) and their representations in the visual system. This type of connection is akin to what philosophers and semanticists have referred to as an 'indexical' or 'demonstrative' reference and what some cognitive scientists have referred to as 'deictic pointers'. I explain why such a mechanism is needed and suggest that many workers have, in fact, been studying precisely this under the term 'visual index'. The visual index hypothesis is illustrated with the results of some relevant experiments, including multiple object tracking, visual routines and subset-selected visual searches. Indexing theory provides a synthesis that has profound implications for explaining a wide range of psychophysical findings, certain results in infant cognitive development and also some ancient problems in the philosophy of mind.

Representations are the basic building blocks of cognitive explanations of human behavior. It is an article of faith in cognitive science that no theory of cognition is complete without them, at least since behaviorist theories have been abandoned in explaining intelligent behavior. Representations, no matter what their ideological flavor, function in the same way as descriptions: they use the conceptual resources of the mind to encode properties of the world in much the same way as language uses words. Even neural-network models are representational (or 'intentional', in philosophers' terminology) in that they represent aspects of the world as either having certain properties, belonging to categories or falling under concepts. A hallmark of such representations is that they can be incorrect and can misrepresent the real world. For example, they can represent a shadow as a tiger or the shorter of two lines as the longer. It has been known for some time – especially in the field of artificial intelligence – that a conceptual description alone (what Bertrand Russell called a 'definite description') is inadequate for encoding certain types of knowledge that we all possess, such as how to perform certain actions (e.g. play the violin, hit a golf ball into a hole or find your way home along a familiar route). Here, I shall concentrate on one way in which conceptual representations are inadequate for the encoding of beliefs based on visual inputs and expose one particular shortcoming of descriptive representations, namely their lack of what is called **indexical** reference (see Glossary).

The problem with descriptive forms of representation lies in the way in which representations are related to objects, including where an observer is situated in the world. In particular, descriptive representations fail to deal with indexical properties and relations. These are context-dependent properties, defined in terms of their relation to an agent or actor, and are critical in determining many kinds of action. Without additional resources, descriptive representations do not connect with the world in a manner that enables actions to be determined. A viewpoint that has been gaining some currency in cognitive science, sometimes referred to as 'situated' or 'embodied' cognition, attempts to minimize the role of representations in explaining intelligence[1,2]. This theory is closely related to the need for indexical reference, although it has been taken to radical extremes by some authors[3]. The idea that we can minimize or even eliminate our reliance upon representations by appealing to the immediate environment has become popular among different research groups for quite different reasons. Some ideas are merely the perennial recycling of behaviorist ideology in psychology, which attempts to empty the organism of thought and replace it with increasingly complex reflexes. Many writers from the artificial intelligence school of cognitive science (and related philosophical positions) have proposed technical arguments for minimizing the relative importance of representations and increasing the role played by direct, unmediated interactions with the world[3–5]. As Andy Clark[1] pointed out in a recent

Z.W. Pylyshyn is at the Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ 08903, USA.

tel: +1 732 445 0635
fax: +1 732 445 6715
e-mail: zenon@ruccs.rutgers.edu

1364-6613/00/$ – see front matter PII: S1364-6613(00)01477-7

Box 1. Three possible forms of 'thoughts' of a robot

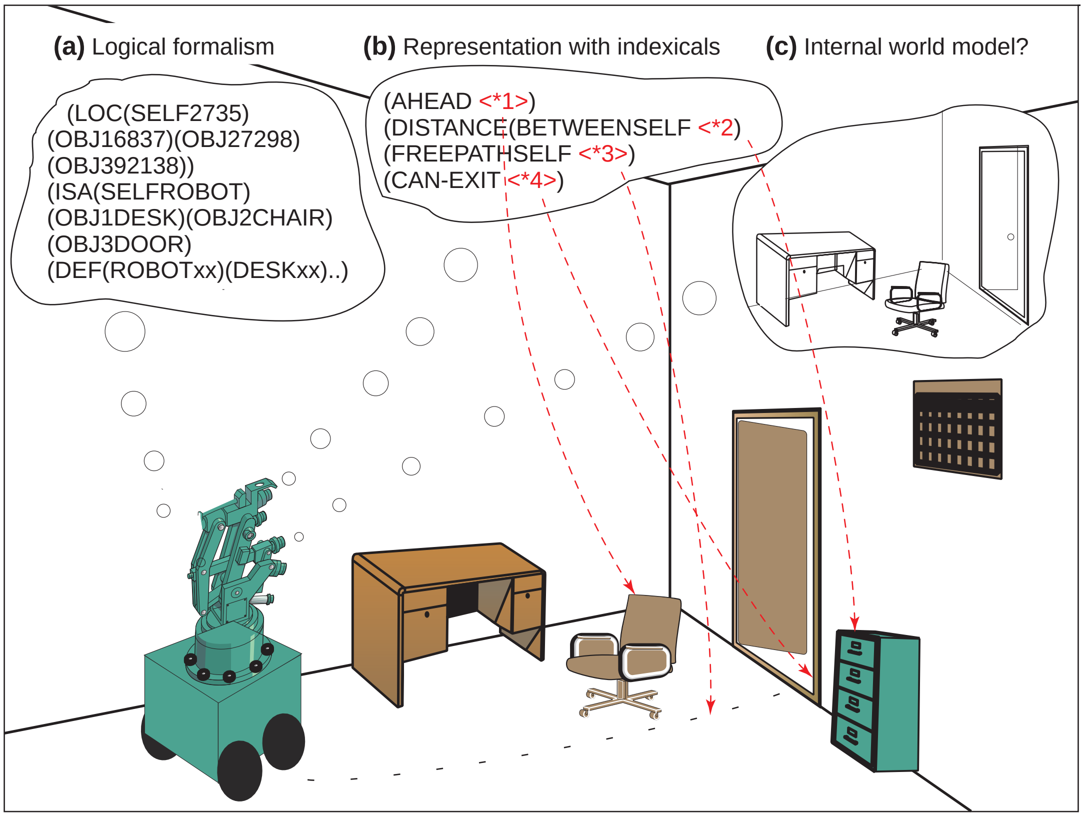
What knowledge representation does a robot require to carry out actions in the real world? All that a robot needs to navigate is a means of directing its attention to individual objects; additional information can then be encoded as needed. The first form of representation (logical formalism; Fig. Ia) is not only inefficient but also fails to allow individual items to be referred to, except via their properties. This is insufficient for the purpose of acting upon objects because an action needs to be 'told' which individual object to act upon, not what its properties are. The need for a mechanism like the one illustrated by the second alternative (Fig. Ib) has been recognized by many workers in robotics (Refs a–c) and philosophy (see Box 2). The third form of representation (internal model; Fig. Ic) has the same disadvantages as the first but requires, in addition, an inner intelligence (or 'homunculus').

References

a Agre, P.E. (1995) Computation and embodied agency. *Informatica* 19, 527–535

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

c Lespérance, Y. and Levesque, H.J. (1995) Indexical knowledge and robot action: a logical account. *Artif. Intell.* 73, 69–115



trends in Cognitive Sciences

Fig. I. Three different ways in which a robot might represent its world. Note the advantage of a representation that uses demonstrative (or deictic) pointers or indexes to select objects in the field of view **(b)**, compared with encoding all potentially useful information [**(a)** or **(c)**].

review, the suggestions put forth in these arguments range from a variety of ways to augment symbolic representations to the assertion that the problem of understanding the physical basis of intelligence needs to be radically reformulated. These arguments (and the counter-arguments of Peter Slezak[2]) are not dealt with here, except to note that the visual index hypothesis (see below) has certain affinities with the basic concerns expressed by the situated vision community. In particular, many authors appeal to the need to take into account the nature of the actual environment, as opposed to represented environments, in explaining intelligent behavior.

The idea of paying attention to the environment may, ironically, originate from a major proponent of symbolic representation theory. In a famous monograph, Herb Simon gave an intriguing example of the importance of paying attention to the environment in explaining complex behaviors[6]. The problem that Simon posed was one of explaining the path of an ant in a desert. Viewed from above, the path might appear highly complex. This might tempt one to hypothesize a complicated representation and an equally complex procedure in the brain of the ant to explain its behavior. However, Simon[6] argued that the complex path could be the result of a very simple process operating in a complex environment: the ant may simply be heading in the direction of the sun and avoiding large obstacles. Similarly, many workers interested in problem-solving and memory have noted that if people have a problem written down, they often do not represent every aspect of it in their mind. For example, the procedures that children learn for doing arithmetic assume that the addends are written down and available to be examined[7]. Consequently, the instructions for addition can refer to such things as 'the next column to the left' or 'the number at the top of the current column' and so

Box 2. The need for demonstratives in encoding beliefs

In arguing that demonstratives (and other indexicals, like the italicized words in '*I* am *here now*') are essential and cannot be replaced by descriptions, philosopher John Perry gives the following example.

> The author of the book *Hiker's Guide to the Desolation Wilderness* stands in the wilderness beside Gilmore Lake, looking at the Mt. Tallac trail as it leaves the lake and climbs the mountain. He desires to leave the wilderness. He believes that the best way out from Gilmore Lake is to follow the Mt. Tallac trail up the mountain… But he doesn't move. He is lost. He is not sure whether he is standing beside Gilmore Lake, looking at Mt. Tallac, or beside Clyde Lake, looking at the Maggie peaks. Then he begins to move along the Mt. Tallac trail. If asked, he would have to explain the crucial change in his beliefs in this way: 'I came to believe that *this* is the Mt. Tallac trail and *that* is Gilmore Lake.' (Ref. a, p. 4)

The point of this example is that to understand and explain the action of the lost author, it is essential to use demonstratives, such as the terms 'this' and 'that', in both the description and, more importantly, in stating the author's beliefs. A more elaborate description of what the mountain trail looked like might have helped to bring the author to the right beliefs, but the problem would have remained unsolved until he had the thought that the trail in front of him ('that' trail) is, in fact, the Mt Tallac trail. Without some way to directly select the referent of a descriptive term and link the perceived object to its cognitive representation, people would be unable to act upon their knowledge and theorists would not be able to explain their actions.

Reference

a Perry, J. (1979) The problem of the essential indexical. *Noûs* 13, 3–21

on, in which indexical reference to situation-dependent items is used. Some workers suggest that the environment is used as an extension of memory[8], because people do not commit everything they see (or even as much as they could) to memory as their eyes explore a scene. Rather, what they possess is a way of returning to and re-examining parts of the scene as required. There are additional examples (see Box 1), but what should be noted here is that to use the environment in this way, people have to be able to keep track of individual objects in it and use tracked objects as markers for cognitive activities. This article describes a mechanism – the visual index – that makes this possible.

Demonstrative reference

Using what I have been referring to as **demonstrative** references avoids the need to encode a scene exhaustively in terms of absolute or global properties and can instead refer to certain relations between the objects and the perceiver/actor. This simplifies certain kinds of planning by providing information in an optimal form for making decisions about actions. Box 1 illustrates three possible ways in which a robot might represent an environment through which it must navigate. It demonstrates that less computation is required if actual pointers to objects in the scene are used as part of the representation, because it allows relevant objects to be selected directly. How a robot with such a capacity could be constructed is currently the focus of serious investigation in artificial intelligence[9,10].

The ability to use indexical references is much more profound than the previous examples may have suggested. There is a crucial difference between representing the fact that there is *something* that possesses certain properties on the one hand, and on the other hand representing the fact that *this very thing* has those properties. For instance, knowing many facts about the North Star (e.g. it is stable in the northern sky, it can be located by extrapolating a line through the pointing stars of the Big Dipper, etc.) is completely different to knowing that the object currently being looked at is the North Star. Yet only the latter belief will lead one to take certain actions if, for instance, one was lost. In fact, the only way that knowing how to determine the position of the North Star is useful is if it is also known which stars are the pointing stars of the Big Dipper. Consequently, a representation is required that directly connects the token objects to which beliefs refer or upon which certain actions can be performed. John Perry gives a lucid example (Box 2) of how the decision to take a particular course of action can arise from the realization that a particular object referred to in a description, and a particular thing that one sees, are one and the same. This realization cannot be formulated without the resources of an indexical or demonstrative type of reference. In what follows, I shall explain what this particular insight tells us about what is required by the visual system to provide information upon which beliefs and actions can be based.

It is a central theme of this article that demonstrative reference is essential not only in robotics, where vision must connect with actions, but also in giving an adequate account of certain properties of vision itself. I suggest that we already possess a theory of a mechanism that serves this purpose, which was developed for quite independent reasons in the course of studying visual attention[12–16]. First, consider some examples in which demonstratives are required by the visual system. We know that a visual representation is not constructed in a single step; there are intermediate stages, which do and do not include the involvement of eye movements[17,18]. Several recent reviews of the experimental literature[19] have made it clear that much less is noticed with each glance at a scene than was initially hypothesized, especially when a scene is changing. Features of a scene are noticed

Glossary

Indexical, demonstrative, deictic: I use these terms almost interchangeably in this article. However, in the technical philosophical literature, an **indexical** is sometimes taken to be a more general concept and to include all terms with context-sensitive referents, such as *I, me, you, now, before, here* and so on. A **demonstrative** is that subset of indexicals, including *that* and *this* (and a more general notion introduced by David Kaplan and written as *Dthat*), which select individuals through a deliberate act of demonstrating or pointing[11]. On the other hand, the term **deictic** simply indicates a pointing relation and is used here interchangeably with demonstrative.

Box 3. Visual routines

Shimon Ullman (Ref. a) described several characteristic visual patterns, the detection of which might involve the construction of a procedure based on more basic operations. Four of these are illustrated below (Fig. I). They involve detecting that: (a) an element lies inside a closed curve; (b) two (or more) elements lie on a single contour; (c) several elements are collinear; and (d) there are exactly n elements in a display. Trick and Pylyshyn (Ref. b) have explored case (d) in some detail and demonstrated that rapid enumeration of small numbers of individuals (called subitizing) only occurs if, first, the individual objects can be pre-attentively individuated, as in the right (but not the left) group in (d) and, second, that the process is not altered if the observer knows in advance where the elements will be. This suggests that subitizing involves only the counting of active pointers.

References

a Ullman, S. (1984) Visual routines. *Cognition* 18, 97–159

b Trick, L.M. and Pylyshyn, Z.W. (1994) Why are small and large numbers enumerated differently? A limited capacity pre-attentive stage in vision. *Psychol. Rev.* 10, 1–23

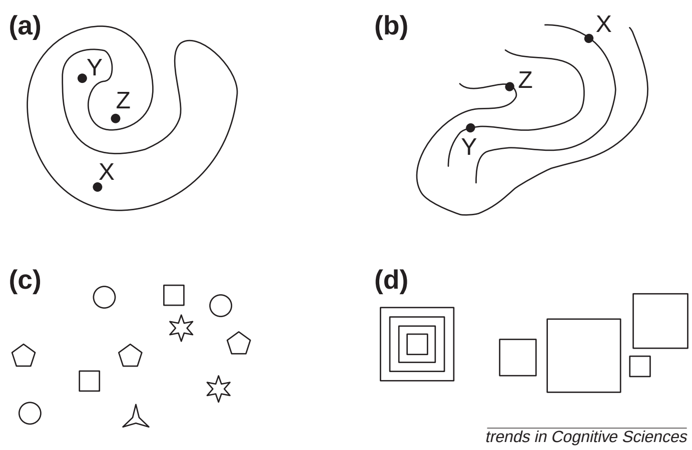


Fig. I. Patterns recognized by using visual routines. The appropriate tasks in these examples are: **(a)** Which dots are inside a closed curve?; **(b)** Which dots are on the same curve?; **(c)** Which of these objects are collinear?; **(d)** How many squares are there in each group?

over a period of time. When the visual system notices new properties of objects that it has already partially encoded, how does it determine which object in its current representation is the object to which the newly perceived property belongs? If the visual system only has a description of the scene as interpreted before the new property was noticed, the only way of deciding where to attach the new property is to first determine how the object might have been previously encoded and subsequently attach the new property to that part of the encoded scene. For example, if at one instant a particular part of the scene was represented as containing two intersecting lines and one subsequently noticed that a certain angle in the scene was a right angle, how would one know where to attach this new information? It needs to be attached to the correct vertex or intersection in the current representation. However, if there are many intersections in the scene, the visual system needs some way to determine which intersection in the current representation is the one now known to be a right angle. I shall call this the 'correspondence problem' for incremental visual encoding.

If the visual system had an accurate encoding of the 3-D coordinates of all the objects in a scene, it would be able to determine the coordinates at which any new (and old) information was stored to solve the correspondence problem. However, experimental evidence suggests that such encoding is not available. As the eyes explore a scene, very little information is retained from one fixation to the next, and none of the retained information relates to the absolute locations of objects. Changes in a scene are rarely updated unless attention is focused on an object that changes[8,20]. Clearly, some other mechanism is needed to solve the correspondence problem. One possible solution is to have a pointer from a representation of an object to an actual object (in the scene), which could act as a demonstrative reference. (Note that the pointer has to point to an object, rather than a location, in order for this to function in dynamic scenes.) Such a pointer would allow the system to map a newly perceived property onto a representation of the object that had been previously (incompletely) encoded. The visual system would need this sort of capacity to be able to represent a particular object, irrespective of how it had been encoded. This is exactly what has already been suggested and precisely what a visual index is designed to provide.

The ability to select individual objects in a scene without regard to their properties has been assumed in most theories of vision[21] and is what makes it possible to recognize that a set of individual objects forms a pattern. In an influential paper, Shimon Ullman[22] argued that the visual system needs certain basic operations from which to construct the capacity to detect more complex patterns. A particular pattern-recognition process constructed in this way is called a visual routine. Ullman showed that to detect certain arrangements of element, a visual routine needs to be serially executed (Box 3). One basic operation that Ullman assumed is known as 'marking' and allows individual items to be 'tagged' or bound to an argument of a visual routine. This is the operation with which I am concerned, although I take a different approach. Instead of implying that an inscription of some sort (e.g. a tag, mark or label) is placed on an object, I propose that a pointer, called a visual index or, for historical reasons, a FINST (for FINger of INSTantiation) is set to point to the object. Note that the properties of individual objects are not used to detect their relational pattern; in fact, the properties must be explicitly ignored. The same applies when acting on an object, such as moving the gaze to it. If one looks at a uniform, repetitive texture with no visually unique elements (e.g. a repetitive wallpaper pattern), one would have no difficulty in attentively selecting a small number of specific elements and moving one's attention from one element to another, providing they are not too close together[23]. This could be achieved even without a visual frame of reference such as walls and a ceiling, as might be the case if one was looking at a uniform pattern of tiny lights in complete darkness. Such examples suggest that the visual system has some way to select or individuate a small number of token visual elements without relying upon uniquely encoding each one. This is what the visual indexing theory proposes. A more familiar way of putting this is to say that the visual system can distinguish between object types and object tokens and assigns attention to individual tokens (e.g. in a visual search[24]). Below, I describe several experiments that were directly motivated by the visual index hypothesis and which support my assertions about the nature of visual indexes and their role as preconceptual reference pointers.

Multiple object tracking

Experimental research on visual indexes started with the following experiment (Fig. 1). Eight identical circles appear on a screen and four of them flicker briefly. Subsequently, all eight circles begin to move randomly on the screen and continue to do so for about ten seconds, after which they stop moving. The observer's task is to keep track of the four circles that initially flickered (but are now identical to the other circles) and to identify them at the end of the trial. The only special characteristic of the targets is that they were visually distinguishable at some time in their history to enable their selection as targets. In our laboratory and others[12,25–28], observers can consistently track four objects with more than 87% accuracy. The question is, how do they do it? One possible answer is that they transfer their attention from one object to another in a consistent pattern while updating the stored locations of the targets. However, in the original study of multiple object tracking[12], we argued that (given some conservative assumptions about how locations are encoded and how quickly attention can be scanned) this process would lead to much poorer tracking performance (around 35% correct) than we actually observed (which was >85% correct).

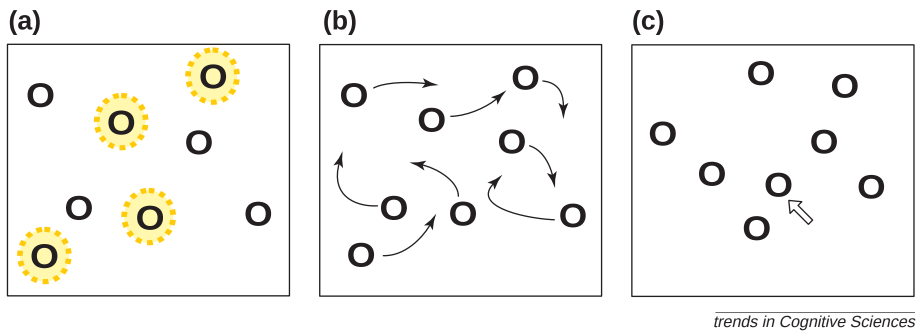


Fig. 1. A version of the multiple object tracking experiment. (a) Eight circles appear on the screen and four of them are briefly flashed to indicate that they are targets. Subsequently, all circles move randomly for ten seconds **(b)**. The observer's task is to select the four targets by placing the cursor on them and clicking the mouse button **(c)**. (For an animated illustration, see http://ruccs.rutgers.edu/finstlab/motOccl.mov)

Using the multiple object tracking paradigm, a great deal has been learned about the indexing mechanism. (Novel findings will be presented in a forthcoming special issue of the journal *Cognition*, devoted to objects and attention.) As an example, we found that certain well-defined clusters of features (e.g. dots) cannot be tracked when they are joined to non-target objects because they become the end points of lines and thus do not constitute individual visual objects from the perspective of the visual system. We also discovered that (1) tracked objects continue to be tracked successfully even though they disappear completely (though briefly), provided that the mode of disappearance is compatible with temporary occlusion behind a screen[27], (2) changes in their color and shape go unnoticed when they are tracked[29] and (3) it takes less time to find a property among targets than among non-targets[26].

In retrospect, it makes sense that the visual system should have a mechanism that can select and track a small number of objects irrespective of what their properties are. Otherwise, the only way in which an observer could determine that an object continued to be the same object would be to notice that it continued to have the same properties. However, the properties of an object can change with an object still remaining the same (e.g. 'It's a bird, it's a plane… no, it's Superman!'). We have argued that the visual system is designed to keep track of the individuality (or what some workers call the 'numerical identity') of certain types of object. We refer to the sorts of object that observers can keep track of in this way as visual objects or 'proto-objects'. This is because we have reason to suspect that although the visual system does not 'know' about physical objects, it nevertheless can track certain visual patterns that are typically associated with physical objects. Indeed, this may be why such a mechanism was incorporated into the human visual system through the process of evolution.

Other evidence of visual indexes

A basic assumption of visual indexing theory is that the visual system has a way of selecting and accessing a small number of visual objects without having to use a description. If this is true, then establishing indexes for several objects should allow an observer to select them rapidly by following pointers provided by indexes, without having to search for an object that fits a description. In addition to the multiple object tracking studies described above, two lines of evidence from my laboratory supports this contention. The first is based on one of the visual routines alluded to by Ullman[22] (see Fig. I in Box 3). This is the process of *subitizing*, in which a small set of items (usually fewer than about four) are enumerated more rapidly and reliably than a larger set. The second line of evidence is the demonstration that observers can select a subset of objects upon which to operate and ignore other items among which the subset is interspersed.

Subitizing or rapid enumeration

It has long been known that observers can enumerate up to four objects rapidly and accurately, but that greater numbers take far longer and are enumerated less accurately[30]. Reaction time increases by about 60 ms per item when there are between two and four objects and by about 100–200 ms per item when there are more than four. Our explanation of this phenomenon is that when a small number of distinct objects is suddenly displayed, each object is assigned an index from a pool of four or five available indexes. Enumeration is then carried out by counting the number of 'active' indexes. Thus, for between two and four objects, it is not necessary to count each object by consulting the display. However, when a greater number of objects is displayed, a more complex process must be adopted: the display has to be consulted, subsets have to be indexed and subitized and their totals then added to the running sum. This makes the counting process slower and more prone to errors. Two predictions of the subitizing hypothesis are, first, if objects cannot be 'individuated' without focal attention, then they cannot be subitized. Second, if subitizing occurs, it should not matter what the position of the objects is or whether their

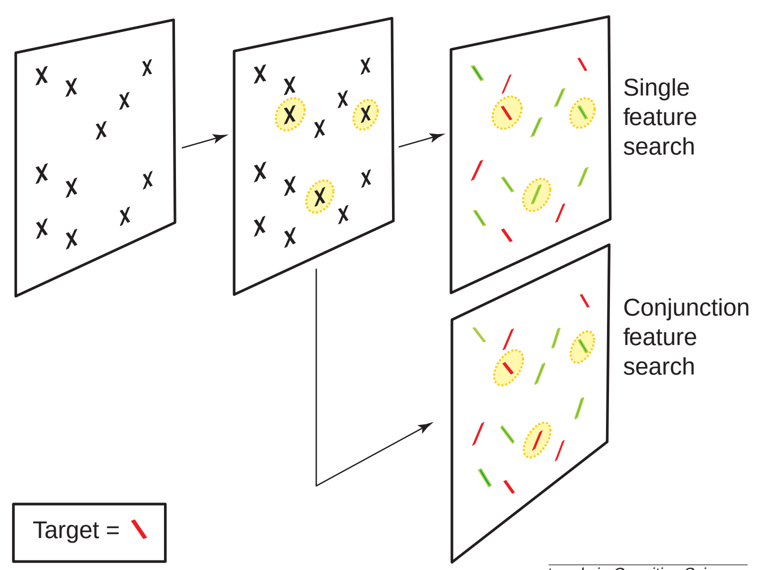


Fig. 2. Experimental evidence that late-onset items are selected and made available for searching. In the top row, the selected subset constitutes a single feature search (i.e. the target differs from each non-target by a single feature). The search is rapid and does not depend of the number of non-targets. The bottom panel shows a conjunction search in which a target can be distinguished from a non-target using only a combination of two features. The conjunction search takes longer and depends on the number of non-targets in the search set. As the whole display is always a conjunction search, a difference between the two search times was taken as evidence that the search is confined to the selected subset. (Modified from Ref. 33.)

approximate position is known in advance. Both of these predictions have been confirmed[30,31]. (Fig. I (a) in Box 3 also illustrates that concentric squares, which cannot be individuated without attentively tracing their contours, cannot be subitized. However, the same squares arranged side by side can easily be subitized.)

Subset selection

One of our assumptions, which has received considerable independent support[32], is that objects that suddenly appear in the visual field are assigned visual indexes. Another assumption is that once an object has been indexed, it can be directly accessed without having to search for it on the basis of its properties. This is the sense in which indexes are like demonstratives or pointers. We tested this hypothesis with an experiment that involved preselecting a subset of items by the sudden appearance of place markers[33]. The experimental design is illustrated in Fig. 2. It relies on the following well-known findings. When observers are asked to search for a unique item in a field of other items, they can find the target item quickly, when it differs from the others based on one particular feature, almost irrespective of how many other items there are (e.g. if the target item is red and the others green, or if it is horizontal whereas the others are vertical, it appears to 'pop out'). By contrast, if the target item shares some properties with other items, such that the target can only be identified by a combination or conjunction of features, then the search time is not only slower but depends a great deal on the number of non-target items. We were interested in whether a discontiguous subset of a larger set of items could be selected using the visual index mechanism. If so, the nature of the subset in relation to the target (e.g. whether it constitutes a single-feature search or a conjunction-feature search) should determine the speed and accuracy of the search. As the difference between these two types of search depends on the whole set, we can then test whether a complete subset could be selected. We hypothesized that if the subset could be selected in this way, the search would be slower if the selected subset constituted a conjunction search than if it constituted a single feature search set (as shown in Fig. 2). If we found that to be the case (so that the search speed depended only on the property of the selected subset), then this would provide strong evidence for the assumption that indexed items can be selected and accessed directly in the search task. Furthermore, if items are indexed and accessed directly without the necessity of scanning the display, their dispersion (or distance apart) should have no effect on the speed of the search task. We found evidence to support both of these hypotheses[33], which suggested that there are indexes that point to late-onset items (Fig. 2). Findings similar to these have been reported by Watson and Humphreys[34], who attribute the selection of item subsets to an 'inhibition' of unselected items, rather than to 'activation' of selected items. Their explanation is perfectly compatible with the visual index hypothesis because it represents one way of implementing an indexing mechanism. Indeed, a possible neural implementation that relies on the inhibition of non-targets in one stage of a neural network model exists (see Box 4).

Tagging versus pointing

Many writers speak of 'marking' or 'tagging' items in a display[22,32,34]. Steve Yantis[32], who was one of the first to show that attentional priority is conferred by the sudden appearance of new objects in a scene, suggests that such items are marked with priority tags and therefore visited first in a search. This may be an accurate description of his findings, but leaves open the question of where a tag is actually placed in the sorts of cases we discussed earlier, for example in connection with the 'correspondence problem' for incremental visual encoding. Placing a tag on an object in a partial or abstract representation does not help to detect relationships in the world that are not yet encoded, or to direct the attention-scanning or eye-movement system. On the other hand, placing a tag on something in the real world would help, but this requires that labels be affixed to objects in the real world. Labels are indeed useful and are, in fact, nearly indispensable in relating descriptions to diagrams, such as in the context of solving problems in geometry, because they enable one to refer directly to token individual objects in a diagram without specifying their properties. What the visual system requires is a way to refer to visual objects in exactly this manner. If it detects that certain items are collinear, as in Fig. Ic in Box 3, it must be able to detect not just the existence of collinearity somewhere in the world, but also which particular objects form a linear pattern, i.e. that the predicate COLLINEAR(x,y,z…) holds of the individual objects x, y and z and not others. Indexes are pointers that provide a link between visual objects and mental objects (e.g. symbols) without requiring that either be labeled or categorized.

Although it is easy to imagine how parts of a representation could be marked, some workers have wondered how the brain could possibly implement a pointer to an object. Koch and Ullman[35] have proposed a plausible neural network that does just this (Box 4) and its application to the par-

Box 4. A possible implementation of a visual index

Koch and Ullman (Ref. a) proposed a winner-take-all neural network that could serve as an implementation of a visual index, although they view it as a mechanism for scanning focal attention. The essential aspects of the network function are illustrated in Fig. I. In Fig. I, the sensors are an array of units, the activation levels of which are mapped into a topographic buffer (or 'mirror').

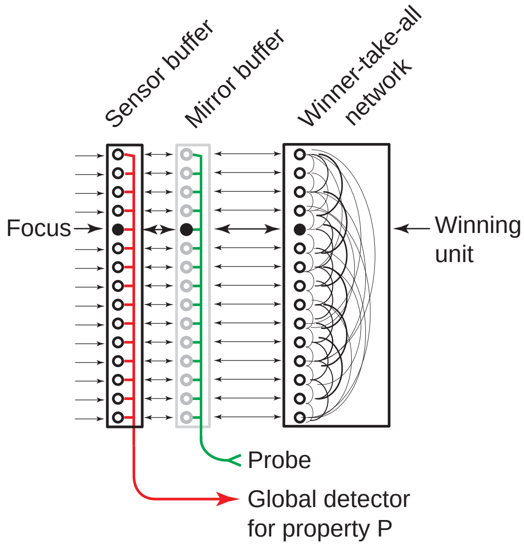


Fig. I. A winner-take-all network for implementing aspects of a visual index. (See text for details.)

This, in turn, feeds into a winner-take-all network that converges on the most active region (I shall call it the focus) and turns all other units in the buffer off. (In their paper, Koch and Ullman actually provide a design for a winner-take-all circuit that is guaranteed to converge rapidly on, and retain the value of, the most active input.) As a result of the inhibition of all but the most active unit, it is possible to send a probe signal through the buffer, which is then routed via an AND 'gate' to property detectors at the focus region. This probe can then be used to check whether certain global property detectors fire. If the property detector for some property P_i (assumed to be set just below a threshold) fires (Indicating the presence of property P_i), then we know that the focus, rather than some other region, is the site of property P_i. In this way, it is possible to make property inquiries of the focus of a topographical array. This is precisely the functionality that visual indexes are assumed to provide. Notice that it is possible to examine the properties of a focal region of the retinotopic display without knowing any of its properties (including its location) other than that it is the most active region in the visual field. Although other properties of visual indexes assumed in the visual indexing hypothesis, such as multiplicity of pointers and object tracking, require additional assumptions (Ref. b), this simple network shows how a pointer can be easily implemented.

References

a Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227

b Pylyshyn, Z.W. and Eagleson, R.A. (1994) Developing a network model of multiple visual indexing. *Invest. Ophthalmol. Vis. Sci.* 35, 2007

ticular requirements of the visual index hypothesis have been discussed[36]. It is presented here because it illustrates how a pointer system can select, and provide access to, an object without encoding where the object is physically located or what its properties are (in a similar way to what a pointer or address does in a computer).

Other related research

Object-file theory

Danny Kahneman and co-workers[37] showed in several experiments that individual objects (as opposed to their locations or other properties) provide the locus for storing and accessing various properties associated with those objects. They made use of the well-known 'priming effect', whereby the prior occurrence of a particular letter decreases the recognition time for that letter. Kahneman *et al.*[37] showed that the priming effect for a letter traveled with the box in which it had occurred (Fig. 3). From this, they concluded that when an object first appears in a scene, an object file is created for it. Thereafter, any subsequent information about the scene is filed in accordance with the associated object. Object-specific benefit can also be demonstrated in other situations where the benefit is more clearly derived from object-type information stored in an object file[38]. Object-file theory is closely related to visual index theory, the primary difference being that research into visual indexes concentrates on the pointer or link, whereas Kahneman *et al.*[37] were concerned with what information is stored at the cognitive end of a link. In the present article, the primary concern is the link between mind and object, because this is what provides the particular demonstrative or preconceptual mind–world connection that I claim is needed by the visual system.

Trans-saccadic integration

One of the most intriguing potential applications of visual indexing theory is in helping to account for how humans integrate information across successive visual fixations. As mentioned above, less information is encoded with each glance than has been previously assumed. Research by several workers[39–42] has shown that information about the properties and relative locations of a small number of objects is retained from one fixation to another and that even major changes in a scene are rarely noticed during saccades. Nevertheless, humans have the impression of a large, panoramic scene. Such a scene does indeed exist, but it is in the real world and not in the mind. Our perception of a scene may be partial and abstract, but we do not perceive a fragmented and incoherent collage of objects as we shift our gaze. How this feat is accomplished is not known, but one of the central pieces of this puzzle is the question of how humans compute the correspondence between objects seen in different fixations. Unless it is possible to decide that a particular object in one glance is the same object as in another glance, there is no way to build a coherent representation. In fact, given how little information is retained between fixations, it may be that this 'correspondence problem' is paramount and that other unknowns (such as why the world appears to remain stable despite the constantly shifting input to the eyes)

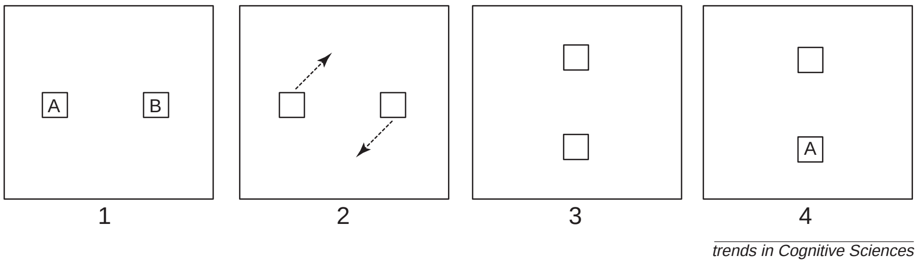


Fig. 3. One of the conditions of the Kahneman *et al.*[37] study of how individual objects provide the locus for storing and accessing properties. Boxes containing letters are shown briefly (frame 1). The empty boxes move to a new location (frames 2 and 3) and a letter that had appeared previously in one of the boxes is then displayed. When the letter is in the same box in which it first appeared, naming time is faster than when it appears in the other box (as illustrated in frame 4).

will fall into place once the correspondence problem is solved. Solutions to the correspondence problem have been proposed, the most notable of which is the saccade target theory[43,44]. It proposes that properties of the object to which an observer is about to move the eyes are encoded in detail so that it can be located again, after which it can serve as a landmark for re-computing the correspondence of other objects. This presumably requires some minimal memory of the relational pattern among objects[42]. From my perspective, this correspondence problem is the same as the correspondence problem discussed earlier in connection with the issue of incrementally constructing a visual representation. We might, therefore, expect the same mechanism to be involved in computing both trans-saccadic correspondence and correspondence within a single image, namely the direct establishment of correspondence for a small number of salient objects using visual indexing. However, the visual index hypothesis differs from the saccade target theory in two ways. First, it assumes that no detailed properties for establishing the uniqueness of any particular object need to be encoded, because it provides a means for keeping track of several objects without having to encode their properties. Second, it assumes that four or five objects can serve as landmarks, because the same number of visual indexes is available for trans-saccadic tracking. If visual indexes survive saccadic motion, they could provide a solution to the problem of trans-saccadic correspondence. Although much remains to be discovered about what happens to indexes during saccades, it has been established that tasks such as multiple object tracking are not affected by voluntary eye movements. This is based on the fact that, although some tracking studies mentioned here (e.g. Ref. 12) included controls to prevent eye movements, observers were free to move their eyes as they wished in others.

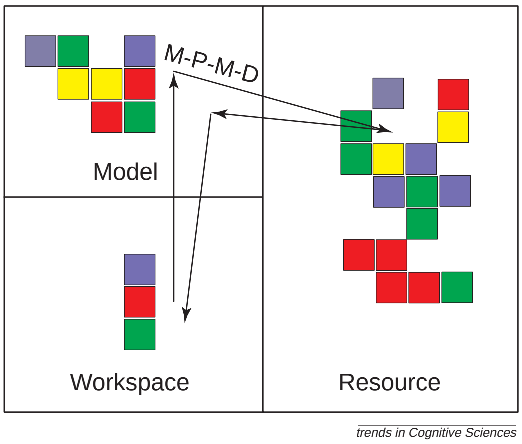
Deictic strategies for visual-motor coordination

Dana Ballard and his colleagues[45] have proposed a reference mechanism that is similar to the visual index hypothesis, although it uses the direction of gaze as the primary means of referencing. Ballard *et al.*[45] studied how direction of gaze functions in visual representations to enable the use of what they term **deictic** perceptual-motor strategies (see Glossary). They argued that the task of perceptual-motor coordination is rendered computationally far more tractable if the motor control of actions that are directed at a visual scene can be cast in terms of a local coordinate system that is based on where the eye is pointing at a particular moment in time. This allows perceptual representation to be more compact, because it can refer to objects or directions in terms of the current (i.e. momentary) gaze direction. From the perspective of visual indexing theory, gaze could serve much the same function as visual indexes; that is, it allows the object of the gaze to be referred to without having to encode its properties. However, in the visual index hypothesis, it is assumed that four or five independent indexes can be assigned simultaneously and that index assignment precedes the movement of the gaze to an object. Indeed, one of the assumptions of indexing theory is that only indexed objects can be the targets of motor commands, including the command to move the gaze to a particular object.

Ballard *et al.*[45] illustrated their deictic pointer mechanism with a copying task (Box 5). They monitored gaze direction as subjects worked on the simple task of building a copy (in a designated 'workspace') of an arrangement of colored blocks (the 'model') that they could freely examine, using a supply of blocks obtained from a 'resource'. The movement of blocks from the resource to the workspace was achieved using a pointing device (a mouse), and the path of the block and eye movements was continuously monitored. The results suggested that subjects did not memorize large parts of the pattern that they needed to copy, even if this was well within their memory span. Instead of looking at the model only four times (which is all that would be required to encode and copy patterns consisting of two blocks), subjects made 18 fixations of the model and did not memorize any more than what was needed for the next basic action of moving one block. The strategy of using the direction of gaze as the focus of memory representation illustrates the use of a deictic strategy wherein pointing (gazing) into a real scene takes precedence over memorizing (at least at the beginning of the trials). Ballard *et al.*[45] concluded that 'performance in the blocks task provides plausible evidence that subjects use fixation as a deictic pointing device to serialize the task and allow incremental access to the immediately task-relevant information'. They added, 'These results support the computational interpretation of the limitations of human working memory. Rather than being thought of as a limitation on processing capacity, it can be seen as a necessary feature of a system that makes dynamic use of deictic variables'. This conclusion is in agreement with the assertion, based on visual indexing theory, that the bottleneck in visual processing does not lie in the limited capacity of short-term memory, but rather in the number of variable bindings between objects and cognitive symbols that can be made using visual indexes[13]. Although Ballard *et al.*[45] concentrated on the importance of the direction of gaze as a deictic pointer, their scheme also used up to three additional deictic pointers. They showed that, in principle, a tower of blocks can be copied using three pointers, regardless of how complex the tower is. In other words, the block copying task can, in principle, be performed using only three indexes. This accords well with the visual index

Box 5. Using a deictic strategy to copy block patterns

Ballard and colleagues (Ref. a) studied how people encode simple patterns of blocks to construct a copy of the pattern. The task, illustrated in Fig. I, involved copying a pattern (the 'model') to a 'workspace' by obtaining blocks from a 'resource' and stacking them in appropriate relations. The strategy that most subjects used was one of moving their gaze frequently to the model and encoding only one simple aspect of the model at a time (e.g. color or location). This strategy relies on the model remaining fixed and benefits from being able to refer to only one block at a time (this being the target of the current gaze fixation). The two most common patterns are shown in Fig. I. The authors showed that eye movements to the model that preceded the pickup were likely to obtain color information, because changing the color during a saccade made very little difference to timings (and subjects did not notice the color change). Note that each of these strategies requires deictic pointers (in other words, visual indexes) to keep track of the blocks being encoded and moved.



a Ballard, D.H. *et al.* (1997) Deictic codes for the embodiment of cognition. *Behav. Brain Sci.* 20, 723–767

Fig. I. Using deictic pointers in a block copying task. Eye and block movements observed in the study of Ballard and colleagues (Ref. a). The most common pattern, adopted in 43% of the moves (here labeled M-P-M-D), is the following. The model is visited with the eye, after which it moves to the pickup point in the resource (perhaps to see which blocks are available and where they are located) and then back to the model (presumably to obtain the color or the location information). Then both the eye and the cursor move with the block to the workspace, drop off the block and repeat the cycle. The second most common strategy (adopted in 28% of moves) involved moving the eye directly to the resource for pickup, then to the model (to check where it should go), and finally to the workspace for drop-off.

hypothesis, wherein it is assumed (based on experimental data such as the multiple object tracking task) that there is a limit of about four indexes.

Indexes, object files and infant detection of numerosity

In a recent review[46], Alan Leslie argued that the object-file theory of Kahneman *et al.*[37] and the visual indexing hypothesis both contain ideas that help to explain why infants between the ages of four and ten months exhibit an apparent sensitivity to the numerosity of objects in their view. Some workers[47–49] have suggested that infants develop the rudiments of the concept of both an object and a number at an early age. Leslie and co-workers[46] showed that infants can distinguish between one and two objects at an earlier age than the age at which they use the property of these objects for recognizing the objects as the same as ones they had seen earlier (Box 6). The authors argued that this might indicate a capacity to index objects at an earlier age that that at which infants can store certain property information in associated object files.

Conclusions

I have argued that the visual system (and perhaps also the cognitive system) needs a special kind of direct reference mechanism to refer to objects without having to encode their properties. Thus, on initial contact, objects are not interpreted as belonging to a certain type or having certain properties; in other words, objects are initially detected without being conceptualized. This kind of direct reference is provided by what is referred to as a demonstrative, or more generally, an indexical. The view that I have presented assumes that certain properties of a visual scene result in indexes being assigned or 'grabbed' from a small pool of available indexes. Although I claim that objects are not indexed by virtue of an encoding of some property (or that the visual system does not search for certain properties), there is clearly some property (or set of properties) that causes indexes to be assigned, just as there is some property that causes red photoreceptive cells to fire regardless of what the visual system is looking for or expecting. Little is known about what properties cause indexes to be grabbed or to remain attached while objects move around or change their properties, although the research into multiple

Box 6. Leslie's object indexes

Alan Leslie *et al.* (Ref. a) showed 12-month-old infants two objects of different colors (a red ball and a green ball), one at a time. After the infants had seen each ball several times, the balls were placed behind a screen. When the screen was removed, infants looked for longer at a display containing only one ball than at a display containing two balls, a result that has been demonstrated by other researchers in infants as young as five months old (Ref. b). However, infants did not distinguish whether the balls were the correct colours (i.e. they looked for the same amount of time at a display with a red and a green ball as at one with two red balls). It appears that infants used color to determine that there were two objects (and therefore, according to Leslie's account, to allocate two indexes), but did not encode and store the color information in the associated object file or use it to determine expectations of what was behind the screen.

References

a Leslie, A.M. *et al.* (1998) Indexing and the object concept: developing 'what' and 'where' systems. *Trends Cognit. Sci.* 2, 10–18

b Wynn, K. (1998) Psychological foundations of number: numerical understanding in human infants. *Trends Cognit. Sci.* 2, 296–303

object tracking[13] and other related work[50] provided some clues. I use the term 'grab' because I assume (at least provisionally) that the act of index assignment is purely data-driven, as it is assumed that attention cannot be deliberately directed towards an object unless that object has already been indexed. I do not rule out the possibility that the eye-movement or attention-scanning system can be directed to locations that are specified, in some limited way, relative to other indexed objects. For example, it is quite plausible that attention can be focused in a certain direction or towards a place described in simple terms such as 'midway between point X and Y', where X and Y are already indexed. Another possibility for top-down control of index assignment is that an object can be assigned an index or an index can be left in place once focal attention has located an object of interest.

A consequence of indexing certain visual objects is that it becomes possible to bind indexed objects to arguments of cognitive representations or cognitive motor programs. This sort of binding is available as long as an indexed object remains in view, or perhaps for a short time thereafter. The availability of such a binding, or demonstrative reference, means that an object can be revisited when further information about it is needed. As noted earlier (Box 5), Ballard *et al.*[45] found that people prefer to use a deictic strategy, wherein they revisit objects frequently for small amounts of information, rather than to encode and retrieve additional information from memory. Perhaps this strategy is the most efficient one in the long-term, as the situated vision community have suggested. For example, the strategy of relying on obtaining information from the environment at the last minute, rather than retrieving it from memory, is certainly a better strategy in a rapidly changing environment. In any case, the preference for following pointers, rather than accessing memory, appears to be a strong one. Jeremy Wolfe and his colleagues[51,52] showed that in a simple speeded search task, objects were routinely revisited even if they had recently been visited and even though observers knew what was there (because they could carry out the particular task from memory).

Despite the simplicity of the visual indexing hypothesis, it represents a rather radical departure in its claim about how cognition establishes contact with the visible world. It claims, in effect, that the most primitive contact that the visual system makes with the world (the contact that precedes the encoding of any sensory properties) is a contact with what have been termed visual objects or proto-objects[27]. In other words, observers may initially detect objects that have been individuated and assigned visual indexes. Subsequently, focal attention may be deployed to objects that have been individuated and indexed by this primitive mechanism. As a result of the deployment of focal attention, it becomes possible to encode the various properties of the visual objects, including their location, color, shape and so on. Recent research into what has been called 'object-based attention' adds credence to the assertion that objects play a central role in accessing and encoding information about the visual world[24,50,53,54]. Perhaps it has been wrong to think that the first contact that humans have with the world is through sensors equipped to detect properties like red or round, oblique or edge-like. Instead, what we may be equipped to detect first (both temporally and ontogenetically) is objects or their primitive precursors, proto-objects.

Acknowledgements

I wish to thank Jerry Fodor and Brian Scholl for their contributions to the ideas contained in this article. The research reported here was supported in part by NIMH Grant 1R01-MH60924 awarded to the author.

References

1 Clark, A. (1999) An embodied cognitive science? *Trends Cognit. Sci.* 3, 345–351

2 Slezak, P. (1999) Situated cognition: empirical issue, paradigm shift or conceptual confusion? In *Perspectives on Cognitive Science* (Wiles, J. and Dartnall, T., eds), pp. 69–98, Ablex

3 Agre, P.E. (1995) Computational research on interaction and agency. *Artif. Intell.* 72, 1–52

4 Agre, P.A. and Chapman, D. (1987) Pengi: an implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artif. Intell.*, pp. 268–272, Morgan Kaufman

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

6 Simon, H.A. (1969) *The Sciences of the Artificial*, MIT Press

7 VanLehn, K. and Ball, W. (1991) Goal reconstruction: how TETON blends situated action and planned action. In *Architectures for Intelligence* (Van Lehn, K., ed.), pp. 147–188, Erlbaum

8 O'Regan, J.K. (1992) Solving the 'real' mysteries of visual perception: the world as an outside memory. *Can. J. Psych.* 46, 461–488

9 Lespérance, Y. and Levesque, H.J. (1995) Indexical knowledge and robot action: a logical account. *Artif. Intell.* 73, 69–115

10 Subramanian, D. and Woodfill, J. (1989) Making the situation calculus indexical. In *Proc. First Int. Conf. Principles of Knowledge Representation and Reasoning*, pp. 467–474, Morgan Kaufman

11 Kaplan, D. (1989) Demonstratives. In *Themes From Kaplan* (Almog, J. *et al.*, eds), pp. 481–563, Oxford University Press

12 Pylyshyn, Z.W. and Storm, R.W. (1988) Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spat. Vis.* 3, 1–19

13 Pylyshyn, Z.W. (1989) The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. *Cognition* 32, 65–97

14 Pylyshyn, Z.W. *et al.* (1994) Multiple parallel access in visual attention. *Can. J. Exp. Psychol.* 48, 260–283

15 Pylyshyn, Z.W. (1994) Some primitive mechanisms of spatial attention. *Cognition* 50, 363–384

16 Pylyshyn, Z.W. (1998) Visual indexes in spatial vision and imagery. In *Visual Attention* (Wright, R.D., ed.), pp. 215–231, Oxford University Press

17 Reynolds, R. (1981) Perception of an illusory contour as a function of processing time. *Perception* 10, 107–115

18 Sekuler, A.B. and Palmer, S.E. (1992) Visual completion of partly occluded objects: a microgenetic analysis. *J. Exp. Psychol. Gen.* 121, 95–111

19 Simons, D.J. (1996) In sight, out of mind: when object representations fail. *Psychol. Sci.* 7, 301–305

20 Rensink, R. (2000) The dynamic representation of scenes. *Visual Cognit.* 7, 17–42

21 Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman

22 Ullman, S. (1984) Visual routines. *Cognition* 18, 97–159

23 He, S. *et al.* (1997) Attentional resolution. *Trends Cognit. Sci.* 1, 115–121

24 Kanwisher, N. and Driver, J. (1992) Objects, attributes, and visual attention: which, what, and where. *Curr. Dir. Psychol. Sci.* 1, 26–31

25 Yantis, S. (1992) Multielement visual tracking: attention and perceptual organization. *Cognit. Psychol.* 24, 295–340

26 Sears, C.R. and Pylyshyn, Z.W. (2000) Multiple object tracking and attentional processes. *Can. J. Exp. Psychol.* 54, 1–14

27 Scholl, B.J. and Pylyshyn, Z.W. (1999) Tracking multiple items through occlusion: clues to visual objecthood. *Cognit. Psychol.* 38, 259–290

28 Blaser, E. *et al.* (1999) Measuring attention during 3-D multi-element tracking. *Invest. Ophthalmol. Vis. Sci.* 40, 552

29 Scholl, B.J. *et al.* (1999) When are featural and spatiotemporal properties encoded as a result of attention allocation? *Invest. Ophthalmol. Vis. Sci.* 39, S872

30 Trick, L.M. and Pylyshyn, Z.W. (1994) Why are small and large numbers enumerated differently? A limited capacity pre-attentive stage in vision. *Psychol. Rev.* 10, 1–23

31 Trick, L. and Pylyshyn, Z.W. (1993) What enumeration studies tell us about spatial attention: evidence for limited capacity pre-attentive processing. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 331–351

32 Yantis, S. and Johnson, D.N. (1990) Mechanisms of attentional priority. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 812–825

33 Burkell, J. and Pylyshyn, Z.W. (1997) Searching through subsets: a test of the visual indexing hypothesis. *Spat. Vis.* 11, 225–258

34 Watson, D.G. and Humphreys, G.W. (1997) Visual marking: prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychol. Rev.* 104, 90–122

35 Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227

36 Pylyshyn, Z.W. and Eagleson, R.A. (1994) Developing a network model of multiple visual indexing. *Invest. Ophthalmol. Vis. Sci.* 35, 2007

37 Kahneman, D. *et al.* (1992) The reviewing of object files: object-specific integration of information. *Cognit. Psychol.* 24, 175–219

38 Henderson, J.M. and Anes, M.D. (1994) Roles of object-file review and type priming in visual identification within and across eye fixations. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 826–839

39 Henderson, J.M. (1997) Trans-saccadic memory and integration during real-world object perception. *Psychol. Sci.* 8, 51–55

40 Irwin, D.E. (1991) Information integration across saccadic eye movements. *Cognit. Psychol.* 23, 420–456

41 Carlson-Radvansky, L.A. and Irwin, D.E. (1995) Memory for structural information across eye movements. *J. Exp. Psychol. Learn. Mem. Cognit.* 21, 1441–1458

42 Carlson-Radvansky, L.A. (1999) Memory for relational information across eye movements. *Percept. Psychophys.* 61, 919–934

43 McConkie, G.M. and Currie, C.B. (1996) Visual stability across saccades while viewing complex pictures. *J. Exp. Psychol. Hum. Percept. Perform.* 22, 563–581

44 Currie, C.B. *et al.* The role of the saccade target object in the perception of a visually stable world. *Percept. Psychophys.* (in press)

45 Ballard, D.H. *et al.* (1997) Deictic codes for the embodiment of cognition. *Behav. Brain Sci.* 20, 723–767

46 Leslie, A.M. *et al.* (1998) Indexing and the object concept: developing 'what' and 'where' systems. *Trends Cognit. Sci.* 2, 10–18

47 Wynn, K. (1998) Psychological foundations of number: numerical understanding in human infants. *Trends Cognit. Sci.* 2, 296–303

48 Xu, F. and Carey, S. (1996) Infants' metaphysics: the case of numerical identity. *Cognit. Psychol.* 30, 111–153

49 Spelke, E. *et al.* (1995) The development of object perception. In *Visual Cognition* (Kosslyn, S.M. and Osherson, D.N., eds), pp. 297–330, MIT Press

50 Yantis, S. (1998) Objects, attention, and perceptual experience. In *Visual Attention* (Wright, R., ed.), pp. 187–214, Oxford University Press

51 Horowitz, T.S. and Wolfe, J.M. (1998) Visual search has no memory. *Nature* 394, 575–577

52 Wolfe, J.M. *et al.* Post-attentive vision. *J. Exp. Psychol. Hum. Percept. Perform.* (in press)

53 Baylis, G.C. and Driver, J. (1993) Visual attention and objects: evidence for hierarchical coding of location. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 451–470

54 Umiltà, C. *et al.* (1995) Object-centered orienting of attention. In *Visual Selective Attention* (Bundesen, C. and Shibuya, H., eds), pp. 165–188, Erlbaum