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. This theory is closely related to the need for indexical reference, although it has been taken to radical extremes by some authors. 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. As Andy Clark pointed out in a recent
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 Slezak2) 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 behaviors6. 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, Simon6 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. 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 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

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).
on, in which indexical reference to situations-dependent items is used. Some workers suggest that the environment is used as an extension of memory, because people do not commit everything they see (or even as much as they could) to memory as they 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 perceivereactor. 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.19

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 everything has those properties. For instance, knowing many facts about the North Star (e.g. it is stable in the northern sky, it is not easy to re-examine parts of it, etc.) would not help the robot 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.

**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 he 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 Chris 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. 6)

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 belief. 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 theories would not be able to explain their actions.

**Reference**


**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, here, before, after and so on. A demonstrative is that subset of indexicals, including or and this (and a more general notion introduced by David Kaplan and written as Deict, which select individuals through a deliberate act of demonstrating or pointing). On the other hand, the term deictic simply indicates a pointing reference and is used here interchangeably with demonstrative.
Shimon Ullman (Ref. 6) 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. 3). They involve detecting that: (a) an element lies inside a closed curve; (b) two (or more) elements lie on a single curve; (c) several elements are collinear; and (d) there are exactly three elements in a display. Trick and Pylyshyn (Ref. 8) have explored case (d) in some detail and demonstrated that rapid enumeration of small numbers of individuated objects (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 pattern 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 limited capacity pre-attentional stage in vision. Appl. Psychol. Med. 10, 1–23

Fig. 1. Patterns recognized by using visual routines. The appropriate tasks in these examples are: (a) Which dots are on the same 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? (e) How many squares are there in each group? (f) Which objects are collinear? (g) Which of these objects are collinear? (h) Which objects are collinear? (i) Which objects are collinear? (j) How many squares are there in each group?

Box 3. Visual routines

Shimon Ullman (Ref. 6) 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. 3). They involve detecting that: (a) an element lies inside a closed curve; (b) two (or more) elements lie on a single curve; (c) several elements are collinear; and (d) there are exactly three elements in a display. Trick and Pylyshyn (Ref. 8) have explored case (d) in some detail and demonstrated that rapid enumeration of small numbers of individuated objects (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 pattern 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 limited capacity pre-attentional stage in vision. Appl. Psychol. Med. 10, 1–23
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 as opposed to the other four 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’ studies, 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, 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).

Using the multiple object tracking paradigm, a great deal has been learned about the indexing mechanism. Novel findings will be presented on 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; (2) changes in their color and shape go unnoticed when they are tracked; and (3) it takes less time to find a property among targets than among non-targets.

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 “mnesiological identity”) of certain types of object. We refer to the term 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.

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. 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. One 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 two 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. 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 clicking the mouse button. (c) Eight identical circles appear on a screen and four of them flicker briefly. The observer’s task is to keep track of the four circles that initially flickered as opposed to the other four 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’ studies, 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, 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).
Opinion

Pylyshyn – Situated vision

the whole display is always a conjunction search, a difference between the two search times was

correlation search takes longer and depends on the number of non-targets in the search set. As

target can be distinguished from a non-target using only a combination of two features. The

depend of the number of non-targets. The bottom panel shows a conjunction search in which a

target differs from each non-target by a single feature). The search is rapid and does not de-

target item is red and the others green, or if it is horizontal

The target (e.g. whether it constitutes a single-feature search or

subset of a larger set of items could be selected using the visual

subset selection

One of our assumptions, which has received considerable in-

approximate position is known in advance. Both of these pre-

dictions have been confirmed[30,31]. (Fig. I (a) in Box 3 also

illustrates that concentric squares, which cannot be indi-

viduated without attentively tracing their contours, cannot

be subitized. However, the same squares arranged side by

side can easily be subitized.)

Subset selection

Many writers speak of ‘marking’ or ‘tagging’ items in a dis-

play[32,33]. Steve Yaniss[34], who was one of the first to show

that attentional priority is conferred by the sudden appear-

ance 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 case of items we discussed earlier, for example in con-

nection with the ‘correspondence problem’ for incremental

visual encoding. Placing a tag on an object in a partial or ab-

tract representation does not help to direct relationships in

the world that are not yet encoded, or to direct the atten-

tion-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 indis-

pent in relating descriptions to diagrams, such as in the

context of solving problems in geometry, because they en-

able one to refer directly to individual objects in a dia-

gram without specifying their properties. What the visual sys-

tem 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 exis-

tence 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 representa-
tion 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 net-

work that does just this (Box 4) and its application to the pur-
Box 4. A possible implementation of a visual index

Koch and Ullman (Ref. 3) 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. 1. In Fig. 1, the sensors are an array of units, the activation levels of which are mapped into a topographic buffer (or ‘mirror’).

This, in turn, leads 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 maintain 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 inferences of the focus of a topological 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. 3), this simple network shows how a pointer can be easily implemented.

References


Fig. 1. A winner-take-all network for implementing aspects of a visual index. (See text for details.)
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 observer to direct the gaze to a specific location, 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 ‘reservoir’. The movement of blocks from the reservoir 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 anything 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\(^1^3\). 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 only three indexes. This accords well with the visual index

**Fig. 3.** One of the conditions of the Kahneman et al.\(^42\) 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 (Frame 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).
workers have suggested that infants develop the rudiments of sensitivity to the numerosity of objects in their view. Some between the ages of four and ten months exhibit an apparent pothesis both contain ideas that help to explain why infants of the concept of both an object and a number at an early age. The authors argued that this might indicate a capacity to the objects as the same as ones they had seen earlier (Box 6). However, infants did not distinguish whether the balls were the correct colours as young as five months old (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 re-
moved, 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. 3). 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.

Conclusions

I have argued that the visual system (and perhaps also the cog-
nitive system) needs a special kind of direct reference mecha-
nism to refer to objects without having to encode their prop-
erties. 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 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. 1, involved copying a pattern (the ‘model’) to a ‘workspace’ by obtaining blocks from a ‘resource’ and tracking 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 retaining

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


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 re-
moved, 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. 3). 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

object tracking11 and other related work12 provided some clues. I use the term ‘grap’ 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 systems 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 that reflects its location and 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.29 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 changing environment. In any case, the preference for following pointers, rather than accessing memory, appears to be a strong one. Jeremy Wolfe and his colleagues51,52 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 provides the encoding of any sensory properties) is a contact with what have been termed visual objects or proto-objects16. 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.23,24,25 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.

Acknowledgments

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 RBNH Grant 501-MH44054 awarded to the author.

References


Current Trends 2000 Online Access
As part of a Current Trends initiative, institutional print subscribers will be able to obtain online access to individual Current Trends titles for the first time through BioMedNet for a premium. Further details and prices can be obtained from info@current-trends.com or ct.subs@rbi.co.uk

Current Trends titles are also available via ScienceDirect licensing agreement. Contact info@sciencedirect.com, telephone the New York office +1 212 462 1980 or fax +1 212 462 1985 for details. Personal subscribers continue to be eligible for FREE online access to their print journal, also through BioMedNet.