ELSEVIER

COGNITION

Cognition 80 (2001) 127–158

www.elsevier.com/locate/cognit

Visual indexes, preconceptual objects, and situated vision

Zenon W. Pylyshyn*

Rutgers Center for Cognitive Science, Rutgers University, Psychology Building, New Wing, Busch Campus, New Brunswick, NJ 08903, USA

Received 6 November 1999; accepted 17 November 2000

Abstract

This paper argues that a theory of *situated vision*, suited for the dual purposes of object recognition and the control of action, will have to provide something more than a system that constructs a conceptual representation from visual stimuli: it will also need to provide a special kind of direct (preconceptual, unmediated) connection between elements of a visual representation and certain elements in the world. Like natural language demonstratives (such as 'this' or 'that') this direct connection allows entities to be referred to without being categorized or conceptualized. Several reasons are given for why we need such a preconceptual mechanism which individuates and keeps track of several individual objects in the world. One is that early vision must pick out and compute the relation among several individual objects while ignoring their properties. Another is that incrementally computing and updating representations of a dynamic scene requires keeping track of token individuals despite changes in their properties or locations. It is then noted that a mechanism meeting these requirements has already been proposed in order to account for a number of disparate empirical phenomena, including subitizing, search-subset selection and multiple object tracking (Pylyshyn et al., *Canadian Journal of Experimental Psychology 48*(2) (1994) 260). This mechanism, called a *visual index* or FINST, is briefly discussed and it is argued that viewing it as performing a demonstrative or preconceptual reference function has far-reaching implications not only for a theory of situated vision, but also for suggesting a new way to look at why the primitive individuation of visual objects, or proto-objects, is so central in computing visual representations. Indexing visual objects is also, according to this view, the primary means for grounding visual concepts and is a potentially fruitful way to look at the problem of visual integration across time and across saccades, as well as to explain how infants' numerical capacity might arise.

* Fax: +1-908-445-0634.
E-mail address: zenon@ruccs.rutgers.edu (Z. Pylyshyn).

0010-0277/01/$ - see front matter
PII: S0010-0277(00)00156-6

Keywords: Early vision; Visual attention; Visual indexing; Multiple object tracking; Object-based attention; Visual representation; Indexicals; Demonstrative reference; Deictics; Situated vision

1. Background: what is missing in a purely conceptual representation

In this paper I argue that a theory of how visual information is analyzed and represented for the dual purposes of recognition and control of action will have to provide a system that does more than construct a conceptual representation from proximal stimulation. Such a theory, which we might call *a theory of situated vision*, will also have to provide a special kind of direct connection between elements of a visual representation and certain token elements in the visual field, a connection that is unmediated by an encoding of properties of the elements in question. In the first part of this paper I will give a number of empirical arguments for why such a function is needed. Then I will suggest that the *visual indexing* mechanism that we have been studying experimentally for some time (Pylyshyn, 1989) provides just this sort of function.

The most general view of what vision does is that it computes a representation of a scene that then becomes available to cognition so that we can draw inferences from it or decide what it is or what to do with it (and perhaps a somewhat different version of which may also become available for the immediate control of motor actions). When we represent something, whether in vision or in thought or even in a natural language, we typically represent a situation 'under a description', that is, we represent the elements of the situation as members of some category or as falling under a certain concept. So, for example, when we say or think 'the cat is on the mat', we refer to the elements under the categories 'cat' and 'mat'. This is, in fact, a fundamental characteristic of cognitive or intentional theories which distinguishes them from physical theories (see Pylyshyn, 1984). That is because what determines our behavior is not the physical properties of the stimuli around us, but how we interpret or classify them – what we *take them to be*. It is not the bright spot in the sky that determines which way we set out when we are lost, but the fact that we see it (or represent it) as the North Star. It is because we represent it *as* the North Star that our perception is brought into contact with our knowledge of such things as astronomy and navigation.[1] This is common ground for virtually all contemporary theories of cognition.

But this is not the whole story. Although it is not often recognized we do, under

[1] A useful heuristic in determining whether something is perceived or represented 'under a description' or 'conceptually' is to ask whether it could be *misperceived*. A preconceptual (purely causal) connection has no room for misperception. We can misperceive something only when we perceive it *as* something it is not. But when we react (mechanically, neurally or biochemically) to a physical stimulus the reaction does not depend on the category under which we perceive or conceptualize it. Consequently, we cannot react to it in this way in error (or 'mis-react' to it).

certain conditions, also represent some things without representing them in terms of concepts. We can refer to some things, as I will say, *preconceptually*. For example, in the presence of a visual stimulus, we can think thoughts such as '*that* is red' where the term '*that*' refers to something we have picked out in our field of view without reference to what category it falls under or what properties it may have. A term such as *this* or *that* is called a 'demonstrative'. Philosophers like Perry (1979) have argued that demonstratives are ineliminable in language and thought. The reasons for the ineliminability of demonstratives in language and thought also apply to visual representations. Not only can we represent visual scenes in which parts are not classified according to some category, but there are good reasons why at least some things *must* be referenced in this preconceptual way. If we could only refer to things in terms of their category membership, our concepts would always be related only to other concepts (the concepts for categories) and would never be grounded in experience. Sooner or later the regress of specifying concepts in terms of other concepts has to bottom out. Traditionally, the 'bottoming out' was assumed to occur at sensory properties, but this 'sense data' view of concepts has never been able to account for the grounding of anything more than simple sensory concepts and has been largely abandoned. The present proposal is that the grounding begins at the point where something is picked out directly by a mechanism that works like a demonstrative. We will later propose that visual indexes do the picking out and the things that they pick out in the case of vision are what many people have been calling *visual objects* or proto-objects.

A second closely related problem with the view that representations consist solely of concepts or descriptions arises when we need to pick out particular token individuals. If our visual representations encoded a scene solely in terms of concepts or categories, then we would have no way to pick out or to refer to particular individuals in a scene except through concepts or descriptions involving other concepts, and so on. In what follows I will suggest a number of ways in which such a recursion is inadequate, especially if our theory of vision is to be situated, in the sense of making bidirectional contact with the world – i.e. contact in which individual elements in a scene causally invoke certain elements in a representation, and in which the elements in the representation can in turn be used to refer to particular individuals in the world.

It is this second problem – that of establishing a correspondence between individual things in the world and their counterparts in the visual representation – that I will focus on in this paper, since this is where the notion of a visual index played its first theoretical role in our work. Before I describe how a visual index is relevant to this connection function, I offer a few illustrations of how this function is missing from the sorts of representations that visual theories typically provide. Theories of visual perception universally attempt to provide an effective (i.e. computable) mapping from dynamic 2D patterns of proximal stimulation to a representation of a 3D scene. Both the world and its visual representation contain certain individuals or elements. The world contains objects, or whatever your ontology takes to be the relevant *individuals*, while the representation contains symbols or symbol structures (or codes, nodes, geons, logogens, engrams, etc. as the theory specifies). The

problem of keeping *tokens* of the representing elements in correspondence with *tokens* of individual things in the world turns out to be rather more difficult than one might have expected.

With the typical sort of conceptual representation, there is no way to pick out an individual in the world other than by finding the tokens in a scene that fall under a particular concept, or satisfy a particular description, or that have the properties encoded in the representation. What I will try to show is that this cannot be what goes on in general; it can't be the case that the visual system can only pick out things in the scene by finding instances that satisfy its conceptual representation. There are phenomena that suggest that the visual system must be able to pick out individuals in a more direct manner, without using encoded properties or categories. If this claim is correct then the visual system needs a mechanism for selecting and keeping track of individual visual objects that is more like a demonstrative reference (the sort of reference we make in language when we use demonstrative terms like *this* or *that*) than a description. And that, I suggest, is why we must have something like a visual indexing mechanism which *preconceptually* picks out a small number of individuals, keeps track of them, and provides a means by which the cognitive system can further examine them in order to encode their properties, to move focal attention to them or to carry out a motor command in relation to them (e.g. to point to them).

The idea that we need to have a means of direct reference is not new. In the study of robotic control, researchers like Lespérance and Levesque (1995) recognized that in order for a robot to function in a real environment, its system of representations must be able to deal with indexicals, or agent-centered ways of representing the world. In such representations, instead of referring expressions like 'a large red round object located at $<x,y>$' we might have 'the object in line with the direction I am heading and located between me and what I am looking at right now…'). Our notion of an index is a special case of such an indexical (other cases include locatives, such as 'here', or 'now' and personal deictic references such as 'I' or 'you'). Other artificial intelligence writers, such as Agre (1997) and Brooks (1999), have gone even further and suggested that focussing on indexical information changes the computational problems associated with planning and executing actions so radically that symbolic representations will play little or no role in these problems. While such issues are beyond the scope of this paper, I wish merely to point out that, one way or another, indexicals are playing a larger role in current theories of cognition, especially where cognition eventuates in action.

What I intend to do in this paper is first lay out some general empirical motivations for hypothesizing the existence of primitive indexing mechanisms (sometimes called FINSTs for purely historical reasons, going back to Pylyshyn, Elcock, Marmor, & Sander, 1978, where indexes were referred to as 'FINgers of INSTantiation') that *individuate* and *index*, or keep track of about four or five individual objects in the visual field. I will then present some experimental evidence showing that something like an index must be available inasmuch as people can select and keep track of four or five individual objects in controlled experiments. I will briefly review and defend the so-called FINST theory of visual indexing. Then I will discuss the relation of these ideas to other work, including work on *deictic strategies*, on

object files and on infants' sensitivity to numerosity. Finally, I will very briefly relate these ideas to some classical problems in psychology, including the problem of transsaccadic integration and the problem of grounding concepts in experience.

2. The need for individuating and indexing: empirical motivations

There are two general problems raised by the description view of visual representations, i.e. the view that we pick out and refer to objects solely in terms of their categories or their encoded properties. One problem is that there are always an unlimited number of things in the world that can satisfy any particular category or description, so that if it is necessary to refer to a unique individual object among many similar ones in the visual field (especially when its location or properties are changing), a description will often be either too complex or inadequate. A second problem is deeper. The visual system needs to be able to pick out a particular individual *regardless* of what properties the individual happens to have at any instant of time. It is often necessary to pick out an element in the visual field *as a particular enduring individual*, rather than as whatever happens to have a certain set of properties. An individual remains the same individual when it moves about or when it changes any (or even all) of its visible properties. Yet *being the same individual* is something that the visual system often needs to compute, as we shall see in the examples below.

In arguing for the insufficiency of conceptual (or descriptive) representations as the sole form of visual representation, I will appeal to three empirical assumptions about early vision: the assumption that the detection of properties proceeds by the prior detection of objects that bear those properties, the assumption that the detection of objects is primitive and preconceptual (i.e. does not itself involve the appeal to any properties), and the assumption that visual representations are built up incrementally.

2.1. Detection of visual properties is the detection of properties of objects

The *first assumption* is that when a property is detected and encoded by the visual system it is typically detected not just as a property existing in the visual field, but as the property of an individual perceived object. I will assume that the visual system does not just detect the presence of redness or circularity or collinearity in the visual field: it detects that certain individual objects are red or circular or are arranged linearly. There are a number of sources of evidence supporting this assumption, most of which were collected in connection with asking somewhat different questions.

(a) There is a great deal of evidence showing that several properties are most easily extracted from a display when they occur within a single visual object, and therefore that focal attention (which is assumed to be required for encoding conjunctions of properties) is object-based (Baylis & Driver, 1993). Evidence supporting this conclusion comes from a variety of sources (many of which are reviewed in Scholl, 2001), including clinical cases of hemispatial visual neglect and Balint syndrome, which implicate an object-centered frame of reference. This sort of

object-specificity of feature encoding is exactly what would be expected if properties are always detected as belonging to an object.

(b) Evidence often cited in support of the assumption that properties are detected in terms of their *location* is compatible with the view that it is the object with which the property is associated, rather than its location, that is primary. A good example of a study that was explicitly directed at the question of whether location was central was one carried out by Nissen (1985). She argued that in reporting the conjunction of two features, observers must first locate the *place* in the visual field that has both features. In Nissen's studies this conclusion comes from a comparison of the probability of reporting a stimulus property (e.g. shape or color or location) or a pair of such properties, given one of the other properties as cue. Nissen found that accuracy for reporting shape and color were statistically independent, but accuracy for reporting shape and location, or for reporting color and location, were *not* statistically independent. More importantly, the conditional probabilities conformed to what would be expected if the way observers judged both color and shape was by using the detected (or cued) color to determine a location for that color and then using that location to access the shape. For example, the probability of correctly reporting both the location and the shape of a target, given its color as cue, was equal (within statistical sampling error) to the product of the probability of reporting its location, given its color, and of reporting its shape, given its location. From this, Nissen concluded that detection of location underlies the detection of either the color or shape feature given the other as cue. Similarly, Pashler (1998, pp. 97–99) reviewed a number of relevant studies and argued that location is special and is the means by which other information is selected. Note, however, that since the objects in all these studies had fixed locations, these results are equally compatible with the conclusion that detection of properties is mediated by the prior detection of the individuals that bear these properties, rather than of their location. If the individuals had been moving in the course of a trial it might have been possible to disentangle these two alternatives and to ascertain whether detection of properties is associated with the instantaneous location of the properties or with the individuals that had those properties.

(c) A number of experimental paradigms have used moving objects to explore the question of whether the encoding of properties is associated with individual objects, as opposed to locations. These include the studies of Kahneman, Treisman, and Gibbs (1992) on 'object files' and our own studies using multiple object tracking (MOT) (see Section 3.2 below, as well as Pylyshyn, 1998; Pylyshyn et al., 1994). Kahneman et al. showed that the priming effect of letters presented briefly in a moving box remains attached to the box in which the letter had appeared, rather than to its location at the time it was presented. Similarly, related studies by Tipper, Driver, and Weaver (1991) showed that the phenomenon known as i*nhibition of return* (whereby the latency for switching attention to an object increases if the object has been attended in the past 300 ms to about 1000 ms) was specific to particular objects rather than particular locations within the visual field (though later work by Tipper, Weaver, Jerreat, & Burak, 1994, suggests that location-specific IOR also occurs).

While there is evidence that unitary focal attention, sometimes referred to as the 'spotlight of attention', is often location-based, and appears to spread away from its central spatial locus, the sort of attention-like phenomena that were investigated in connection with object files and IOR (and other studies to be sketched in Section 3) appear to be far more attached to objects with little evidence of spreading to points in between the objects. Using the MOT paradigm, we found that in a shape discrimination task using MOT, changes are more readily discriminated when they are associated with objects that are being tracked, with little spread to inter-object locations (Sears & Pylyshyn, 2000; see also Intriligator & Cavanagh, 1992). In all these cases what appears most relevant to the detection of properties is not their instantaneous location, but the continuing individuality – or some writers say, the continuing numerical identity – of the objects that bear those properties.

2.2. Individuation of object tokens is primitive and precedes the detection of properties

The *second assumption* is that the process of individuating object tokens is distinct from the process of recognizing and encoding the objects' types or their properties. Clearly, the visual system can distinguish two or more distinct token individuals regardless of the type to which each belongs, or to put it slightly differently, we can tell visually that there are several distinct individuals independent of the particular properties that each has; we can distinguish distinct objects (and count them) even if their visible properties are identical. What is usually diagnostic of (though not essential to) there being several token individuals is that they have different spatiotemporal properties (or locations). Without a mechanism for individuating objects independent of encoding their properties it is hard to see how one could judge that the six elements in Fig. 1 are arranged linearly, especially if the elements in the figure were gradually changing their properties or if the figure as a whole was moving while maintaining the collinear arrangement of elements. In general, featural properties of elements tend to be factored out when computing global patterns, regardless of the size and complexity of the global pattern (Navon, 1977). Computing global patterns such as collinearity, or others discussed by Ullman (1984), requires that elements be registered as individuals while their local properties are ignored. This 'ignoring' might make use of whatever selectional mechanisms may be available, perhaps including, in the collinearity example, focusing attention on lower spatial frequencies or functionally replacing the objects with points. Whatever the particular algorithm used to detect collinearity among elements, it is clear that specifying *which* points form a collinear pattern is a necessary part of the computation.

Here is another way to think of the process of computing relational properties among a set of objects. In order to recognize a relational property, such as **Collinear**(X_1,X_2,…X_n) or **Inside**(X_1,C_1) or **Part-of**(F_1,F_2), which apply over a number of particular individual objects, there must be some way to specify which objects are the ones referred to in the relationship. For example, we cannot recognize the **Collinear** relation without somehow picking out *which* objects are collinear. If

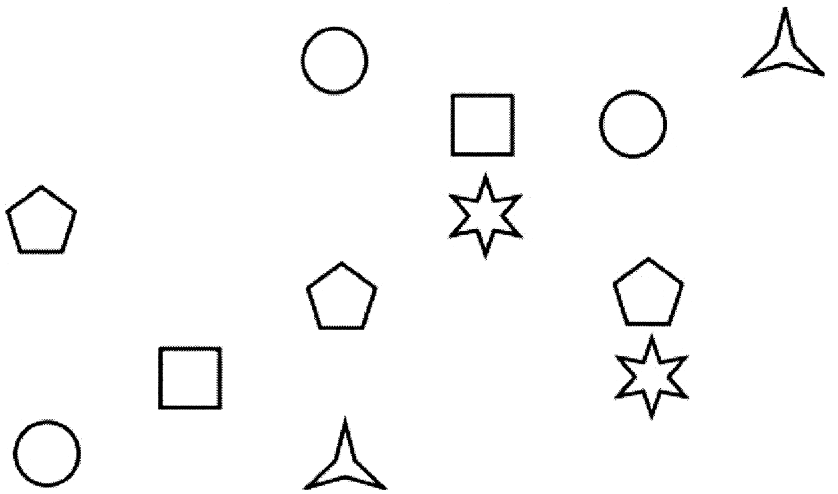

Fig. 1. Judging collinearity requires both selecting the relevant individual objects and ignoring all their local properties.

there are many objects in a scene only some of them may be collinear, so we must *bind* the objects in question to argument positions in the relational predicate. Ullman (1984) as well as a large number of other investigators (Ballard, Hayhoe, Pook, & Rao, 1997; Watson & Humphreys, 1997; Yantis, 1998; Yantis & Johnson, 1990; Yantis & Jones, 1991) refer to the objects in such examples as being 'marked' or 'tagged'. The notion of a tag is an intuitively appealing one since it suggests a way of labeling objects to allow us to subsequently refer to them. Yet the operation of tagging only makes sense if there is something on which a tag literally can be placed. It does no good to tag an internal representation since the relation we wish to encode holds in the world and may not yet be encoded in the representation. So we need a way of 'tagging' that enables us to get back to tagged objects in the world to update our representation of them. But how do we tag parts of the world? It appears that what we need is what labels give us in diagrams: a way to name or refer to individual parts of a scene *independent of their properties or their locations*. This label-like function that goes along with object individuation is an essential aspect of the indexing mechanism that will be described in greater detail below.

There are a number of other sources of evidence suggesting that individuation is distinct from discrimination and recognition. For example, individuation has its own psychophysical discriminability function. He et al. (1997) have shown that even at separations where objects can be visually resolved they may nonetheless fail to be *individuated* or attentionally resolved, preventing the individual objects from being picked out from among the others. Without such individuation one could not count the objects or carry out a sequence of commands that require moving attention from one to another. Given a 2D array of points lying closer than their threshold of attentional resolution, one could not successfully follow such instructions as:

'move up one, right one, right one, down one,…' and so on. Such instructions were used by Intriligator (1997) to measure attentional resolution. Fig. 2 illustrates another difference between individuating and recognizing. It shows that you may be able to recognize the shape of objects and distinguish between a group of objects and a single (larger) object, and yet not be able to focus attention on an individual object within the group (in order to, say, pick out the third object from the left). Studies reported in He et al. (1997) show that the process of individuating objects is separate and distinct from that of recognizing or encoding the properties of the objects.

Studies of rapid enumeration (called *subitizing*) described in Trick and Pylyshyn (1994) also show that individuating is distinct from (and prior to) computing the cardinality of a small set of objects. Trick and Pylyshyn showed that items arranged so they cannot be preattentively individuated (or items that require focal attention in order to individuate them – as in the case of items lying on a particular curve or specified in terms of conjunctions of features) cannot be subitized, even when there are only a few of them (i.e. there was no break in the function relating reaction time to number of items). For example, in Fig. 3, when the squares are arranged concentrically (as on the left) they cannot be subitized, whereas the same squares arranged side by side can easily be subitized. According to our explanation of the subitizing phenomenon, small sets are enumerated faster than large sets when items are preattentively individuated because in that case each item attracts an index, so observers only need to count the number of active indexes without having to first search for the items. Thus, we also predicted that precuing the location of preattentively individuated items would not affect the speed at which they were subitized, though it would affect counting larger numbers of items – a prediction borne out by our experiments (Trick & Pylyshyn, 1994).

2.3. Visual representations are constructed incrementally

The *third assumption* is that our visual representation of a scene is not arrived at in one step, but rather is built up incrementally. This assumption has strong support. Theoretical analyses (e.g. Tsotsos, 1988; Ullman, 1984) have provided good reasons for believing that some relational properties that hold between visual elements, such as the property of being inside or on the same contour, must be acquired serially by scanning a display. We also know from empirical studies that percepts are generally built up by scanning attention and/or one's gaze. Even when attention may not be scanned there is evidence that the achievement of simple percepts occurs in stages over a period of time (e.g. Calis, Sterenborg, & Maarse, 1984; Reynolds, 1981; Schulz, 1991; Sekuler & Palmer, 1992). If that is so then the following problem immediately arises. If the representation is built up incrementally, we need a mechanism for determining the correspondence between representations of individual elements across different stages of construction of the representation or across different periods of time. As we elaborate the representation by uncovering new properties of a dynamic scene, we need to know which individual objects in the current representation should be associated with the new information. In other

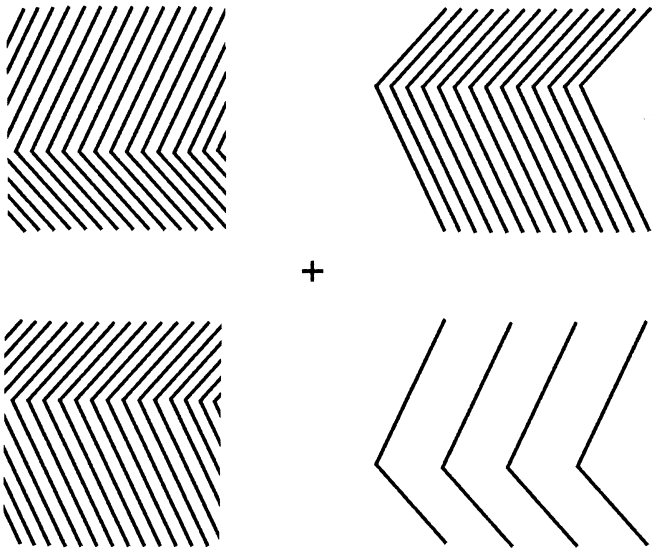

Fig. 2. At a certain distance if you fixate on the cross you can easily tell which groups consist of similar-shaped lines, although you can only *individuate* lines in the bottom right group. For example, you cannot count the lines or pick out the third line from the left, etc., in the other three groups (based on He, Cavanagh, & Intriligator, 1997).

words, we need to know when a certain token in the existing representation should be taken as corresponding to the same individual object as a particular token in the new representation. We need that so that we can attribute newly noticed properties to the representation of the appropriate individual objects. This problem remains even if the scene changes gradually so that updating can occur continuously – indeed the problem arises even if the scene remains fixed while the representation is incrementally computed (or when the percept is bistable, as in Fig. 4).

Suppose we have a representation of a scene such as the one shown on the right of Fig. 4 (a possible form of representation, which was used by Feldman & Ballard, 1982, is shown on the right). From the representation one might be able to infer that there are 12 lines in the figure. But we don't have a way to refer to the lines

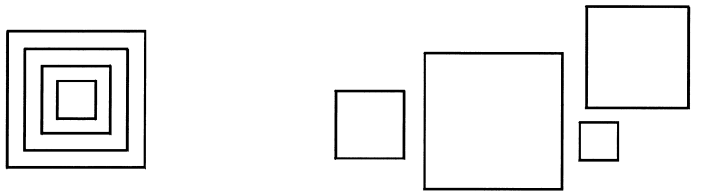

Fig. 3. Squares that are arranged so they cannot be preattentively individuated (on the left) cannot be subitized, whereas the ones on the right are easily subitized (based on Trick & Pylyshyn, 1994).

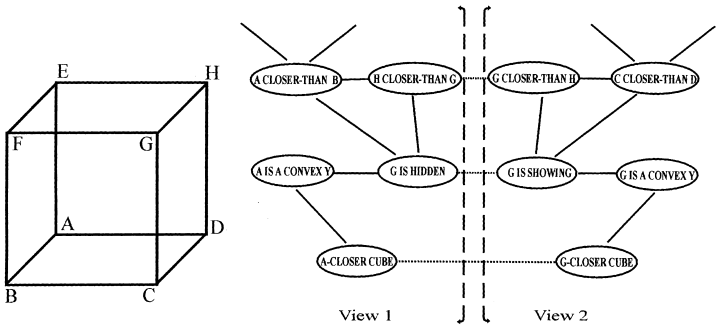

Fig. 4. One possible form of representation of a figure such as the reversing cube on the left. The figure on the right shows a connectionist network (as described by Feldman & Ballard, 1982), with solid lines corresponding to activating links and dotted lines corresponding to inhibitory links. This example assumes that vertices in the diagram are labeled. How could you represent the figure if they were not labeled?

individually. Yet without identifying particular lines we could not add further information to elaborate the representation. If, for example, on further examination, we discover that some lines are longer than others, some vertices form certain recognized angles, and so on, we would not be able to connect this new information to the representation of particular individual objects in the current representation. Because conjunctions of properties (e.g. red AND right-angled AND smaller-than, etc.) are defined with respect to particular objects, individual objects must be identified in order to determine whether there are property conjunctions.

A general requirement for adding information to a representation is that we be able to relate the newly discovered properties to *particular* elements in the existing representation of the figure. If you notice, say, that a certain angle is a right angle, you need to add this information to the representation of a particular vertex. How do you know which represented vertex it is so you can add the information to the appropriate item? In this example, as in using diagrams in general, we label lines or vertices. But of course the world does not come with every object conveniently labeled. What constraints does the need to pick out individual objects impose on the form and content of an adequate representation?

In principle it is possible to pick out an individual object by using an encoded description of its properties. All you need is a description that is unique to the individual in question, say 'the object α with property P' where P happens to uniquely pick out a particular object. But consider how this would have to work. If you want to add to a representation the newly noticed property Q (which, by assumption, is a property of a particular object, say object α), you must first locate the representation of object α in the current representation. Assuming that individuals are represented as expressions or individual nodes in some conceptual network, you might detect that the object that you just noticed as having property Q also had property P which uniquely identifies it. You might then assume that it had been previously stored as an object with property P. So you find an object in the current representation that is described as having P and conjoin the property Q to it

(or use an identity statement to assert that the object with property P is identical to the object with property Q). There are many ways to accomplish this, depending on exactly what form the representation takes. But whatever the details of such an augmentation process, it must be able to locate the representation of a *particular individual* in order to update the representation properly. Yet this may well be too much to ask of a general procedure for updating representations. It requires working backward from a particular individual in the scene to its previous representation. There is no reason to think that locating a previous representation of an individual is even a well-defined function since representations are highly partial and schematic (and indeed, the representation of a particular object may not even exist in the current representation) and an individual object may change any of its properties over time while continuing to be the same object. In fact the rapidly-growing literature on change blindness would suggest that unless objects are attended they may change their properties without their representation being updated (Rensink, 2000a,b; Rensink, O'Regan, & Clark, 1997, 2000; Simons, 1996; Simons & Levin, 1997).

The basic problem can be stated as follows: in order to properly update a representation upon noticing a new property Q, what you need to find in the current representation is not a representation of an individual with certain properties, but rather the representation of the *very individual* on which the new property Q has been detected, and you have to do that independent of what properties of the display you have already encoded at that point in time. The alternative to the unwieldy method described in the past paragraph for locating a representation of a particular individual is to allow the descriptive apparatus to make use of some functional equivalent of *demonstrative* reference (such as the type of reference corresponding to the natural language words *this* or *that*). If we had such a mechanism, then adding newly noticed information would consist of adding the predicate $Q(\alpha)$ to the representation of a particular object α, where α is the object directly picked out by this demonstrative indexing mechanism. Since, by hypothesis, the visual system's Q detectors recognize instances of the property Q *as a property of a particular visual object* (in this case of α), being able to refer to α provides the most natural way to view the introduction of new visual properties by the sensorium.[2] In order to introduce new properties into a representation in that way, however, there would have to be a non-descriptive way of picking out the unique object in question. In the following section I examine experimental evidence suggesting that such a mechanism is needed for independent reasons – and in fact was proposed some time ago in order to account for certain empirical findings.

Note that although the above discussion has been concerned mainly with reidentifying individual objects within the foveal field of view, a very similar problem

[2] The reader will have noticed that this way of putting it makes the reference mechanism appear to be a *name* (in fact the name 'α'). What I have in mind is very like a proper name insofar as it allows reference to a particular individual. However, this reference relation is less general than a name since it ceases to exist when the referent (i.e. the visual object) is no longer in view. In that respect it functions exactly like a demonstrative, which is why I continue to call it that, even as I use examples involving names like α.

arises when the objects appear across different views, as when a display is examined by moving the eyes. Interestingly, the problem that led to the postulation of a visual index mechanism in the first instance arose in connection with the attempt to model the process of reasoning with the aid of a diagram (Pylyshyn et al., 1978). The problem there is rather similar to the updating problem discussed above. But since relevant objects might have moved off the fovea into the parafovea in the course of drawing the figure, a new dimension is added to the problem of updating the representation: we need to be able to pick out individual objects that have left the high-resolution field of view and then returned again as the eyes moved about. This problem will be raised again in discussing the relation of visual indexing theory to theories of saccadic integration in Section 4.2.

3. Experimental evidence for a visual index mechanism

3.1. Preconceptual selection

The following experiment by Burkell and Pylyshyn (1997) illustrates and provides evidence in favor of the assumption that the visual system has a mechanism for picking out and accessing individual objects prior to encoding their properties. Burkell and Pylyshyn showed that sudden-onset location cues (which we assumed caused the assignment of indexes) could be used to control search so that only the locations precued in this way are visited in the course of the search. This is what we would expect if the onset of such cues draws indexes and indexes can be used to determine where to direct focal attention.

In these studies (illustrated in Fig. 5) a number of placeholders (11 in this example), consisting of black Xs, appeared on the screen and remained there for 1 s. Then an additional three to five placeholders (which we refer to as the 'late-onset cues') were displayed. After 100 ms one of the segments of each X disappeared and the remaining segment changed color, producing a display of right-oblique and left-oblique lines in either green or red. The subject had to search the cued subset for a line segment with a particular color and orientation (say a left-oblique green line). Since the entire display had exemplars of all four combinations of color and orientation, search through the entire display was always what is known as a conjunction-search task (which is known to produce slow searches in which the time it takes to locate a target increases rapidly with increasing numbers of items in the display). As expected, the target was detected more rapidly when it was one of the subsets that had been precued by a late-onset cue, suggesting that subjects could directly access those items and ignore the rest. There were, however, two additional findings that are even more relevant to the present discussion. These depend on the fact that we manipulated the nature of the precued subset to be either a single-feature search task (i.e. in which the target differed from all other items in the search set by no more than one feature) or a conjunction-search task (in which only a combination of two features could identify the target because some of the non-targets in the search set differed from it in one feature and others differed from it in another feature).

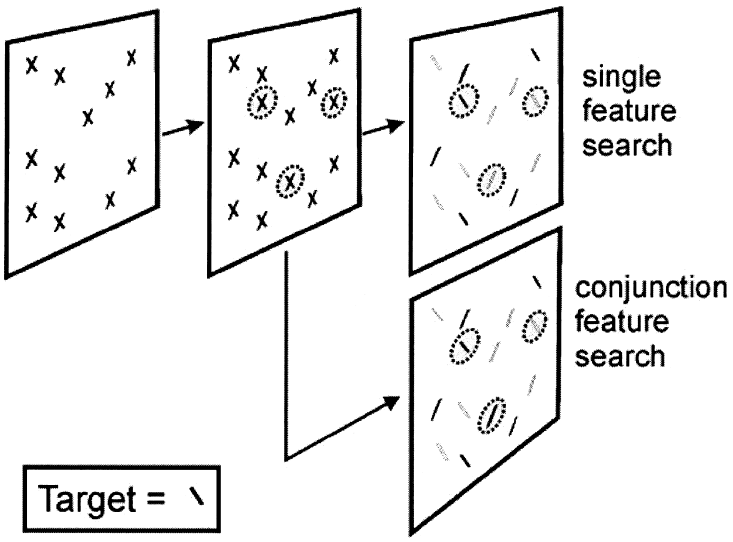

Fig. 5. Sequence of events in the Burkell and Pylyshyn (1997) study. In the first frame the observer sees a set of placeholder Xs for 1000 ms. In the second frame, 'late-onset' placeholders appear for 100 ms, signaling the items that will constitute the search subset. In the third frame, all placeholders change to search items and the subject must try to find the specified target in one of two conditions. In the top display the target differs from all the non-targets by one feature (color), whereas in the bottom display, a combination of two features is required to distinguish the target. In the experiment the bars were either red or green and the faint circles did not appear – they are only for expository purposes.

Although a search through the entire display would always constitute a conjunction-feature search, the subset that was precued by late-onset cues could be either a simple or a conjunction-feature subset. So the critical question is this: is it the property of the entire display or the property of only the subset that determines the observed search behavior? We found clear evidence that only the property of the *subset* (i.e. whether it constituted a simple-search or a conjunction-search task) determined the relation between the number of search items and the reaction time. This provides strong evidence that only the cued subset is being selected as the search set. Notice that the distinction between a single-feature and a conjunction-feature search is a distinction that depends on the entire search set, so it must be the case that the entire precued subset is being treated as the search set: the subset effect could not be the result of the items in the subset being visited or otherwise processed one by one.

Of particular relevance to the present thesis was the additional finding that when we systematically increased the distance between precued items there was *no* increase in search time per item, contrary to what one would expect if subset items were being spatially searched for. It seems that increasing the spatial dispersion of the items does not increase the time it takes to examine them, even when the examination appears to be serial (e.g. the time increases linearly as the number of non-targets in the subset increases). This is precisely what one would expect if, as we

predict, the cued items are indexed and indexes can be used to access the items *directly* (although serially), without having to scan over the display for the subset items.

This type of study provides a clear picture of the property of indexes that we have been emphasizing: they provide a *direct access mechanism*, rather like the random access provided by addresses or pointers in a computer. Certain primitive visual objects can be indexed without appealing to their properties (the indexing being due to their sudden appearance on the scene) and once indexed, they can be individually examined either in series or in parallel. In other words, one can ask 'Is *x* red?' so long as *x* is bound to some visual object by an index.

It should be noted that Watson and Humphreys (1997) independently reported a set of very similar studies and found very similar results to those of Burkell and Pylyshyn (1997). They presented a set of search items in two successive displays and showed that as long as the temporal gap between early and late items was more than about 400 ms and as long as there was no luminance change in the early items at the time the late items appeared, the late-onset items behaved as though they were the only items displayed. However, these authors argued that the underlying priority-assignment process involved 'marking' the early items for inhibition in the subsequent selection task. While we are not persuaded that the Watson and Humphreys results imply that selectional priority of late-onset items is due to the inhibition of the old items, their explanation of the selectional effect is compatible with the visual indexing theory since visual indexes could in principle be implemented by activation or inhibition of object representations or by some combination of the two (in fact an earlier implementation, following the work of Koch & Ullman, 1985, does use both activation and inhibition; see Box 4 of Pylyshyn, 2000; Pylyshyn & Eagleson, 1994). The point we make is that once 'selected', the objects can be accessed directly without using an encoding of their properties and without further scanning of the display – i.e. we assume that the mechanism of selection provides an access path or binding between objects and the cognitive processes that need to refer to them (e.g. the comparison or test operation in the search process). This is why it is significant that we found that the spatial dispersion of the objects did not affect search time.

3.2. Multiple object tracking (MOT)

We have argued that the visual system needs a mechanism to *individuate and keep track of particular individuals in a scene* in a way that does not require appeal to their properties (including their locations). Thus, what we need is a way to realize the following two functions: (a) pick out or individuate *primitive visual objects*, and (b) provide a means for referring to these objects as though they had labels or, more accurately, as though the visual system had a system of pointers. Although these two functions are distinct, I have proposed that they are both realized by a primitive mechanism called a *visual index*, the details of which are sketched in Section 4. In this section I illustrate the claim that there is a primitive mechanism that picks out and maintains the identity of visual objects, by describing an experimental paradigm

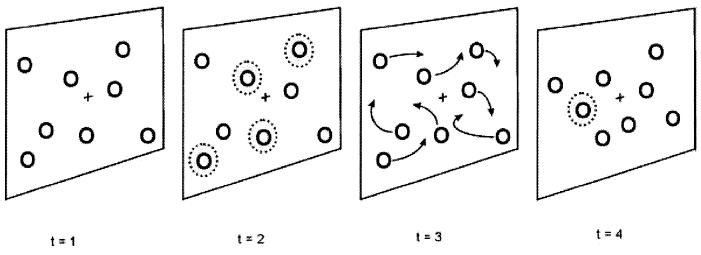

Fig. 6. Illustration of a typical MOT experiment. A number (here eight) of identical objects are shown (at $t = 1$), and a subset (the 'targets') is selected by, say, flashing them (at $t = 2$), after which the objects move in unpredictable ways (with or without self-occlusion) for about 10 s. At the end of the trial the observer has to either pick out all the targets using a pointing device or judge whether one that is selected by the experimenter (e.g. by flashing it, as shown at $t = 4$) is a target.

we have been using to explore the nature of such a mechanism. It is called the *Multiple Object Tracking (MOT) Task* and is illustrated in Fig. 6.

In a typical experiment, observers are shown anywhere from eight to 24 simple identical objects (points, plus signs, circles, figure-eight shapes). A subset of these objects is briefly rendered distinct (usually by flashing them on and off a few times). Then all the identical objects move about in the display in unpredictable ways. The subject's task is to keep track of this subset of objects (called 'targets'). At some later time in the experiment (say 10 s into the tracking trial) one of the objects is probed by flashing it on and off. The observer must then indicate whether the probed object was one of the targets. (In other studies the subject had to indicate *all* the targets using a mouse.) A large number of experiments, beginning with the studies described in Pylyshyn and Storm (1988), have shown that observers can indeed track up to five independently moving targets within a field of ten identical items. In the original Pylyshyn and Storm study we showed that the motion and dispersion parameters of that experiment were such that tracking could not have been accomplished using a serial strategy consisting of scanning focal attention to each figure in turn, encoding and storing its location, and then on the next iteration, returning to the figure closest to that location, updating that location, and so on. Based on some conservative assumptions about how fast focal attention might be scanned and using the actual trajectories of the objects of the experiments we simulated this strategy as it would apply to our experimental materials. From this we were able to conclude that such a serial tracking process would frequently end up switching to the wrong objects in the course of its tracking and would result in a performance that was very much worse than the performance we actually observed in our experiments (over 85% correct). This means that the moving objects could not have been tracked by a unitary beam of attention *using a unique stored description of each figure*, inasmuch as the only possible descriptor that was unique to each figure at any particular instant in time was its location. If we are correct in arguing from the nature of the tracking parameters

that stored locations cannot be used as the basis for tracking then all that is left is the figure's identity over time, or its persisting *individuality*. This is exactly what I claim – viz., that we have a mechanism that allows preconceptual tracking of a primitive perceptual individuality.[3]

Recently, a large number of additional studies in our laboratory (Blaser, Pylyshyn, & Domini, 1999; Blaser, Pylyshyn, & Holcombe, 2000; McKeever, 1991; Pylyshyn, 1998; Scholl & Pylyshyn, 1999; Scholl, Pylyshyn, & Feldman, 2001; Scholl, Pylyshyn, & Franconeri, 2001; Sears & Pylyshyn, 2000) as well as in other laboratories (Culham et al., 1998; Intriligator & Cavanagh, 1992; Viswanathan & Mingolla, in press; Yantis, 1992, and others) have replicated these MOT results using a variety of different methods, confirming that observers can successfully track around four or five independently moving objects. The results also showed that merely widening one's breadth of attention (as assumed in the so-called zoom lens model of attention spreading, Eriksen & St. James, 1986) would not account for the data. Performance in detecting changes to elements located inside the convex hull outline of the set of targets was no better than performance on elements outside this region, contrary to what would be expected if the area of attention were simply widened or shaped to conform to an appropriate outline (Sears & Pylyshyn, 2000). Using a different tracking methodology, Intriligator and Cavanagh (1992) also failed to find any evidence of a 'spread of attention' to regions between targets (see also Awh & Pashler, 2000). It appears, then, that items can be tracked despite the lack of distinctive properties (and, indeed when their properties are changing) and despite constantly changing locations and unpredictable motions.[4] Taken together, these studies suggest that what Marr (1982) referred to as the *early vision system* (an essentially encapsulated system, discussed at length in Pylyshyn, 1999) is able to individuate and keep track of about five visual objects and does so without using an encoding of any of their visual properties.

The MOT task exemplifies what is meant by 'tracking' and by 'maintaining the

[3] As usual one can't exclude all logically possible alternative processes for achieving these results. For example, we cannot exclude the possibility that location encoding occurs in parallel at each tracked object and then serially allocated focal attention is used for tracking, or that four parallel 'beams of attention' independently track the four targets. Another alternative that has been proposed (e.g. Yantis, 1992) is that the objects are tracked by imagining that they are vertices of a deforming polygon and tracking the polygon as a whole. This 'polygon tracking' view may describe a useful strategy for chunking the tracking objects and thus improve one's memory for where they are (which is useful for recovering from errors, as noted in Sears & Pylyshyn, 2000), but it does not supplant the need to track the individual objects since the statistically independent movement of these objects continues to define the vertices of the imagined distorting polygon. One logically cannot track the polygon without *somehow* tracking the independently-moving individual targets. Moreover, observers can track the targets perfectly well whether or not they maintain a convex polygon and whether or not they use this strategy. The strongest case for the indexing mechanism comes from the convergence of a variety of different studies (described in Pylyshyn et al., 1994, and elsewhere), no one of which is definitive, but the pattern of which supports the view that there is a distinct mechanism for individuating and keeping track of token visual objects.

[4] In a set of yet-unpublished studies (Scholl, Pylyshyn, & Franconeri, 2001) we have even shown that observers do not notice and cannot report changes of color or shape of objects they are tracking when the change occurs while they are behind an occluder or during a short period of blank screen, thus lending credence to the view that properties are ignored during tracking.

identity' of objects. It also operationalizes the notion of 'primitive visual object' as whatever allows preconceptual selection and MOT.[5] Note that objecthood and object-identity are thus defined in terms of an empirically established mechanism in the human early vision system. A certain (possibly smooth) sequence of object locations will count as the movement of a single visual object if the early vision system groups it this way – i.e. if it is so perceived. Of course it is of interest to discover what sorts of events will in fact count as visual objects from this perspective. We are just beginning to investigate this question. We know from MOT studies that simple figures count as objects and also that certain well-defined clusters of features do not (Scholl, Pylyshyn, & Feldman, 2001). Indeed, as we saw in Section 2, some well-defined visually-resolvable features do not allow individuation (see Figs. 2 and 3). We also know that the visual system may count as a single persisting individual, certain cases where clusters of features disappear and reappear. For example, Scholl and Pylyshyn (1999) showed that if the objects being tracked in the MOT paradigm disappear and reappear in certain ways, they are tracked as though they had a continuous existence. If, for example, they disappear and reappear by deletion and accretion along a fixed contour, the way they would have if they were moving behind an occluding surface (even if the edges of the occluder are not invisible), they are successfully tracked. However, performance in the MOT task degrades significantly in the control conditions where objects suddenly go out of existence and reappear at the appropriate matching time and place, or if they slowly shrink away to a point and then reappear by slowly growing again at exactly the same relative time and place as they had accreted in the occlusion condition. The persistence of objecthood despite certain kinds of disappearances was also shown in a different context by Yantis (1998) who found that when an object disappears either for a very short time or under conditions where it is seen to have been occluded by an opaque surface, the visual system treats the two exposures of the object as a single persisting object. These findings are compatible with the thesis (Nakayama, He, & Shimojo, 1995) that occlusion plays an important role in early vision. Beyond that, what qualifies as a primitive (potentially indexable) object remains an open empirical question. In fact, recent evidence (Blaser et al., 2000) has even shown that objects can be tracked even though they are not specified by unique spatiotemporal coordinates (e.g. when they share a common spatial locus and move through 'feature space' rather than real space).

[5] The concept of a 'proto-object' is a general one that has been used by a number of writers (sometimes using the same term, Di Lollo, Enns, & Rensink, 2000; Rensink, 2000a, and sometimes using some other term, such as 'preattentive object', Wolfe & Bennett, 1997) in reference to clusters of proximal features that serve as precursors in the detection of real physical objects. What these uses have in common is that they refer to something more than a localized property or 'feature' and less than a recognized 3D distal object. Beyond that, the exact nature of a proto-object depends on the theory in question.

4. A theory of visual indexing and binding: the FINST mechanism

4.1. Background motivation and assumptions of the theory

The basic motivation for postulating indexes is that, as we saw at the beginning of this essay, there are a number of reasons for thinking that a certain number of individual objects in the field of view must first be *picked out* from the rest of the visual field and the identity of these objects *qua individuals* (sometimes called their *numerical identity*) must be maintained or tracked despite changes in the individuals' properties, including their location in the visual field. The visual index hypothesis claims that this is done *primitively* by the FINST mechanism of the early vision system, without identifying the object through a unique descriptor. In other words, it is done without cognitive or conceptual intervention. In assigning indexes, some cluster of visual features must first be segregated from the background or picked out as a unit (the Gestalt notion of making a figure-ground distinction is closely related to this sort of 'picking out', although it carries with it other implications that we do not need to assume in the present context – for example that bounding contours are designated as belonging to one of the possible resulting figures). Until some part of the visual field is segregated in this way, no visual operation can be applied to it since it does not exist as something distinct from the entire field.[6]

But segregating a region of visual space is not the only thing that is required. The second part of the individuation process is that of providing a way for the cognitive system to refer to that particular individual or visual object, as distinct from other individuals. It must be possible to bind one of a small number (perhaps four or five) of internal symbols or elements of a visual representation to individual clusters or visual proto-objects. Moreover, the binding must be such that the representation can continue to refer to a visual object, treating it as the *same* individual despite changes in its location or any other property (subject to certain constraints which need to be empirically determined). The existence of such a capacity would make it possible, under certain conditions, to pick out a small number if individual visual objects and also to keep track of them as individuals over time. We are beginning to map out some of the conditions under which such individuation and tracking can occur; for example, they include spatiotemporal continuity of motion, or discontinuity in the presence of local occlusion

[6] Only *visual* objects have been considered in this paper. However, objecthood need not be specific to a particular modality, and more general notions of objecthood might turn out to be theoretically useful. For example, if we define objecthood in terms of trackability (as I do when I half-seriously introduce the term 'FINGs' at the end of this paper, or as we do in Scholl, Pylyshyn, & Feldman, 2001), then objecthood may become a broader, and perhaps theoretically more interesting notion. For example, it appears that even when visual 'objects' are not distinguished by distinct spatiotemporal boundaries and trajectories, they still function as objects in many other respects. They may, for example, be tracked as individuals and they may exhibit such object-based attention phenomena as the single-object detection superiority (Blaser et al., 2000). The auditory domain offers additional possibilities: auditory objects can be tracked either when a moving auditory source is distinguished and followed or when patterns or 'streams' are followed over time (Bregman, 1990; Kubovy & Van Valkenburg, 2001). Whether these various distinct phenomena involve the same sort of indexing mechanism remains an open research question.

cues such as those mentioned above in discussing the Scholl and Pylyshyn (1999) and the Yantis (1998) results. They also include the requirement that the element being tracked be a perceptual whole as opposed to some arbitrary, but well-defined, set of features (see Scholl, Pylyshyn, & Feldman, 2001).

Visual index or FINST theory is described in several publications cited earlier and will not be described in detail here beyond the sketch given above. The essential assumptions may be summarized as follows: (1) early visual processes segment the visual field into feature-clusters which tend to be reliable proximal counterparts of distinct individual objects in a distal scene; (2) recently activated clusters compete for a pool of four to five visual indexes or FINSTs; (3) index assignment is primarily stimulus-driven, although some restricted cognitively mediated processes, such as scanning focal attention until an object is encountered that elicits an index, may also result in the assignment of an index; (4) indexes keep being bound to the same individual visual objects as the latter change their properties and locations, within certain as-yet-unknown constraints (which is what makes them perceptually the same objects); and (5) only indexed objects can enter into subsequent cognitive processes, such as recognizing their individual or relational properties, or moving focal attention or gaze or making other motor gestures to them.

The basic idea of the visual indexing and binding mechanism is illustrated schematically in Fig. 7. Certain proximal events (e.g. the appearance of a new visual object) cause an index to be *grabbed* (since there is only a small pool of such indexes this may sometimes result in an existing binding being lost). As new properties of this object are noticed and encoded they are associated with the index that points to that object. This, in effect, provides a mechanism for connecting elements of an evolving representation with elements (i.e. objects) in the world. By virtue of this causal connection, the cognitive system can *refer to* any of a small number of primitive visual objects. The sense of reference that is relevant here is one that appears in computer science when we speak of pointers or when variables are assigned values. In this sense, when we speak of having a reference we mean that

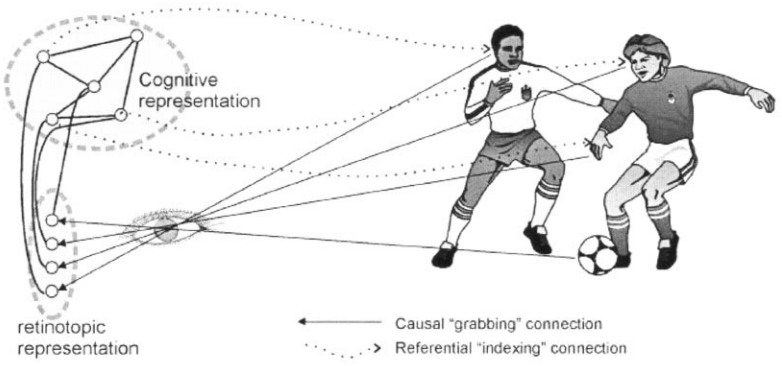

Fig. 7. Sketch of the types of connections established by visual indexes between the primitive visual objects or proto-objects and parts of conceptual structures, depicted here as a network.

we are able to access the things being referred to (the referents) in certain ways: to interrogate them in order to determine some of their properties, to evaluate multiple-argument (polyadic) predicates over them, to move focal attention to them, and in general to *bind* cognitive arguments to them. The important thing here is that the inward arrows are purely causal and are instantiated by the preconceptual apparatus which, following the terminology suggested by Marr (1982), we call *early vision* (Pylyshyn, 1999). The indexing system latches on to certain kinds of spatiotemporal objects because it is 'wired' to do so, or because it is in the nature of its functional architecture to do so, not because those entities satisfy a certain cognitive predicate – i.e. not because they fall under a certain concept. This sort of causal connection between a perceptual system and an object in a scene is quite different from a representational or intentional or conceptual connection. For one thing there can be no question of the object being *mis*represented since it is not represented *as* something (see also Footnote 1).

Although this sort of seizing of indexes by primitive visual objects is essentially a bottom-up process, it could in some cases be guided in an indirect way by intentional processes. For example, it is known (Posner, Snyder, & Davidson, 1980) that people can scan their focal attention along some path (by simply moving it continuously through space like a spotlight beam) and thereby locate certain sorts of objects. A possible consequence of such scanning is that an index may get assigned to some primitive objects encountered along the way. This is no different from the sort of indirect influence that cognition has over vision when one chooses to direct one's gaze or focal attention or the sort of indirect influence we have over other automatic functions (including such autonomic functions as heart rate) when we choose to carry out a voluntary action that leads to a change in the automatic function.

The indexing notion being hypothesized is extremely simple and only seems complicated because ordinary language fails to respect certain distinctions (such as the distinction between individuating and recognizing, or between indexing and knowing where something is, and so on). In fact a very simple network, such as the one described by Koch and Ullman (1985) can implement such a function[7] (the application of the Koch and Ullman network to visual index theory has been

[7] Although we do not address the question of how such a mechanism might be implemented in the nervous system or otherwise, alternatives are not difficult to imagine in the case of vision. Any early vision system will contain sensors and a way of clustering features (e.g. Marr, 1982). In order to maintain the identity of moving clusters (i.e. to implement a 'sticky' binding) all one needs is a mechanism that treats time-slices of clusters that move continuously over the retina as the same cluster. It could do so, for example, by following the rule that if the majority of the elements in a cluster (represented, for example, in a 'list of contributing points') continue to be present in a succeeding cluster then consider both clusters to be the same. Or alternatively, one could simply spread the activation arising from a cluster of elements to neighboring elements, thereby favoring the activation of nearby regions and so favoring continuously moving clusters. This is essentially the technique suggested by Koch and Ullman (1985) in their proposal for a neural implementation of attentional scanning. The point is that there is no in-principle puzzle about how one could implement the notion that indexes are assigned by a bottom-up causal mechanism so that once assigned the indexes are maintained as the clusters move about. Once we have such a clustering mechanism, assigning pointers to the most active of the ensuing clusters is a trivial matter and common ground to most theories of attention (e.g. the guided search theory of Wolfe et al., 1989).

explored in Acton, 1993; Pylyshyn & Eagleson, 1994). All that is required is a winner-take-all circuit whose convergence on a certain active region (or node) on a spatiotopic map enables a signal to be sent to that region, thus allowing it to be probed for the presence of specific properties (a simple sketch of such a system is given in Box 4 of Pylyshyn, 2000). The important point about such a network, which makes its pointing function essentially preconceptual, is that the process that sends the probe signal to a particular object *uses no encoding of properties of that object, not even its location*. Being able to probe a certain object depends only on its instantaneous location (say in some feature map) being the most active by some measure (such as the activation measures assumed in many theories of visual search, like those of Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989). What makes this system object-based, rather than location-based, is certain provisions in the network (i.e. enhancing of the immediate neighboring places) which result in the probe location moving in response to the movement of the primitive object (see Koch & Ullman, 1985).

4.1.1. A note on 'property-dependence' of indexing and on the attentive nature of tracking

There has been some misunderstanding about the role of object properties and of attention in tasks such as MOT. To claim that the indexing process is preconceptual is not to claim that the assignment and maintenance of indexes does not depend on properties of the objects. Clearly indexes get assigned and maintained because the objects in question possess certain properties rather than other properties. The issue is whether an *encoding* of these properties is used by the cognitive system to assign and track an index. It is also clear that objects are being tracked in the MOT paradigm because observers in these studies intend to solve a certain problem and in doing so they must pay attention to the task – indeed they often find the task very attention-demanding. What makes indexing preconceptual is not that it does not depend on properties of objects or that it is not connected with the cognitive system – it clearly is. Like other mechanisms of early (non-cognitive) vision, such as edge detectors, these mechanisms are both modular (i.e. operate on a manner that is independent of cognitive processes) and at the same time are deployed by cognitive processes in pursuit of some task. Red detectors are presumably preconceptual too, they work the way they do because of their physical–chemical properties, but they can nonetheless be deployed in a cognitive task, such as to search for a red car in a parking lot. None of these facts mean that *indexing* per se is conceptual or cognitive or even attentive.

Cognitive factors clearly enter into many aspects of performance in tasks such as MOT. Even if tracking is data-driven, as we have claimed, observers can also decide to move their attention from target to target (that they can do so is an explicit claim of indexing theory). And when they do so they can encode the relative location of targets, even if tracking does not make use of that information. In Sears and Pylyshyn (2000) we argue that one purpose for encoding this location information might be to help recover from errors. If indexes are occasionally lost due to visual distractions, a 'shadow model' of the display can be used to aid in their recovery. Sears and

Pylyshyn argued that such a strategy could account for several aspects of their data, and also might explain why constraining the targets to maintaining a more memorable configuration, such as the convex hull shape investigated by Yantis (1992), helps to improve tracking performance. In any case it is clear that more is going on in MOT experiments than just tracking based on data-driven index maintenance. There is always an overlay of cognitive activity: observers may choose to keep track of designated targets or to switch their attention to other indexed objects, or they may choose to scan their attention around until new objects capture indexes. As we suggested earlier, the cognitive system may be able to influence the indexing process in such indirect ways, and in so doing make it possible for different objects to be indexed (thus allowing objects other than flickered ones to be designated as targets). Observers are free agents, and they don't have to use the indexing mechanism provided by early vision to track targets – they may choose to attend to the sound in the next room or leave the experiment. The *total task* in these (and all other human subject) studies is clearly under cognitive control and typically requires considerable concentration.

In the past I have referred to indexing as 'preattentive' because the theory hypothesizes that indexes are not mediated by a conceptual description. But as just noted, even if the indexing and tracking mechanism is preattentive in this sense, the *task* of tracking multiple objects may require a great deal of effort and attention. In fact the theory predicts that indexes would be readily grabbed by any new object that appears in the field of view, so 'attention' may be involved in orienting the system to the relevant part of the visual field (i.e. attention may be required to control eye movements and to provide some selection of the inputs to the visual indexing system). The notion of an automatic data-driven mechanism is also compatible with the possibility that index binding decays over time and therefore requires periodic reactivation. Since many investigators take susceptibility to disruption and the need for effort as an indication that the process is 'attentive', I now avoid referring to indexing as preattentive. I continue to assume, however, that indexing and tracking are realized by an automatic and preconceptual mechanism, despite the fact that it may require that certain additional conditions be satisfied (e.g. objects must have certain properties in order to capture an index) in order for MOT to occur. This is in part because I believe the bulk of the evidence favors this view and in part because it is the alternative with the more far-reaching consequences and therefore the more interesting hypothesis to pursue, pending evidence to the contrary.

4.2. Relation to other theories

Visual index theory is closely related to a number of other theoretical frameworks. As mentioned earlier, it is very close in spirit to the object file theory of Kahneman et al. (1992) although the latter has been applied in the context of memory retrieval and has consequently emphasized the memory content of the information associated with the objects in memory. Kahneman et al. are correct when they suggest that, "We might think of [a visual index] as the initial spatiotemporal label that is entered

in the object file and that is used to address it… [A] FINST might be the initial phase of a simple object file before any features have been attached to it" (p. 216). Because of this difference in focus, research on visual indexes has been more concerned with the nature of the mechanism that allows cognition to refer to and track objects, whereas object file theory has been concerned with the question of which features of the objects are (eventually) encoded and associated with the object in memory. Thus, other investigators who appeal to the object file idea have typically asked what object-related information is encoded (Wolfe & Bennett, 1997), whereas we have looked at conditions under which only the individuality of objects is relevant to the task.

There is also a close connection between the proposed visual index mechanism and the notion of a *deictic code* discussed by Ballard et al. (1997), although their term 'deictic *code*' misleadingly suggests that the pointer actually encodes some property of the scene, as opposed to providing access to the object's properties. In the Ballard et al. discussion, the authors also point out the importance of having some way to refer to objects in a scene, without using some unique encoded properties to pick out such objects. They introduce the need for such reference both on the grounds of computational complexity and on experimental grounds, because it makes it possible to use what they call a *deictic strategy* which, in effect, allows the perceiver to minimize storage by encoding visual information only when it is needed. In their experiments they found that when observers examine a scene for the purpose of such tasks as copying a pattern of blocks, they encode very little (only the color or location of one block) on each fixation, preferring instead to revisit the scene for each new piece of information. From this Ballard et al. conclude that the object being fixated serves as a deictic reference point in building a minimal description. This is very similar to the view taken here, except that according to the present theory, such deictic references need not involve eye fixations (though they may frequently do so) and they can be directed at more than one object at a time. Indeed, we know from the work of Ullman (1984), as well as from our own work discussed earlier, that they must involve several objects.

There is also a close connection, noted earlier, between updating a representation as new properties of a scene are noticed, and updating a representation in the course of moving one's eyes about, the problem of saccadic integration. If visual indexes were able to keep track of a small number of objects as the same persisting individuals over the course of saccadic eye movements we would have a mechanism that might help to solve one of the central problems of saccadic integration: the problem of determining correspondences between objects in successive views. It had been widely believed that we maintain a large panoramic display and superimpose successive views in registration with eye movements (the registration being accomplished by using a 'corollary discharge' signal). This view has now been thoroughly discredited (Irwin, 1996; O'Regan, 1992), leaving the saccadic integration problem as an open problem in vision research.

The correspondence problem has always been at the heart of the saccadic integration problem, as well as for theories of apparent motion (Dawson & Pylyshyn, 1988), stereo vision (Marr, 1982), visual-auditory integration of location informa-

tion, visual-motor adaptation, and many other psychophysical phenomena that require a perceptual correspondence to be established between individuals. Various mechanisms have been proposed for dealing with this problem. The possible mechanisms for solving this problem rely on one or another of the two distinct ways of picking out and keeping track of individual objects that were discussed earlier: ones that appeal to a unique description of a individual object and ones that do not. The first type includes an interesting proposal called the saccade-target theory (Currie, McConkie, & Carlson-Radvansky, 2000; Irwin, McConkie, Carlson-Radvansky, & Currie, 1994; McConkie & Currie, 1996), which postulates that unique properties of *one* object (the one that serves as the target of the saccade) are encoded and searched for in the second fixation in order to establish a cross-fixation correspondence. Since Irwin (1996) has found that subjects can retain the locations of about four objects across a saccade, the other three objects would have to be located by recalling their (encoded) locations relative to the saccade target. The second option is exemplified by the visual index theory. In contrast to the saccade-target theory, visual index theory assumes that a small number of objects can be recovered from the second fixation as a side effect of their having been indexed and tracked across the saccade, without the benefit of an encoding of their properties. Of course this assumes that indexes survive the very rapid saccadic motion. Some informal indication that this may be the case comes from our own observations that MOT occurs equally well when saccades are freely permitted as when they are prevented (as in the original studies of Pylyshyn & Storm, 1988). Moreover, Henderson and Anes (1994) showed that object files were retained during saccades since object-specific priming occurred across eye fixations.

5. Discussion: objects and the mind–world connection

Visual indexing (FINST) theory hypothesizes a mechanism for picking out, tracking and providing *cognitive access* to *visual objects* or *proto-objects*. The notion of an *object* is ubiquitous in cognitive science, not only in vision but much more widely. Indeed, in a recent ambitious work inspired by ideas from computer science, Brian Cantwell Smith has made the generalized notion of object the centerpiece of a radical reformulation of metaphysics (Smith, 1996). The present work shares with Smith an interest in the question of how a connection can be established between a concept and an object (or in Smith's terms, how the world can be 'registered'), and it also shares the view that the phenomenon of tracking is central to understanding this notion. But our concern in this essay has not been to construct a notion of object free of metaphysical assumptions about the world (a sort of Cartesian skepticism), but with the notion of object beginning with some basic facts about the nature of our early vision system. We take for granted that the world consists of physical objects. The view being proposed takes its initial inspiration from the many studies that have shown that attention (and hence information access to the visual world) is allocated primarily, though not exclusively, to individual visual objects rather than to properties or to unfilled locations (Baylis & Driver, 1993). This general conclusion is also

supported by evidence from clinical neuroscience, where it has been argued that deficits such as unilateral neglect (Driver & Halligan, 1991) or Balint syndrome (Robertson, Treisman, Friedman-Hill, & Grabowecky, 1997) apply over frames of reference that include ones that are object-based, where deficits appear to be specified with respect to individual objects. From this initial idea we have sought to analyze the process of attention into distinct stages. One of these involves the detection and tracking of primitive visual objects. This stage allows attention and other more cognitive processes to access and to operate on these primitive visual objects.

Although our focus has been on *visual* objects there are a number of findings in cognitive development that appear to be relevant to our notion of object and index. For example, the notion of object has played an important role in several works (Carey & Xu, 2001; Leslie, Xu, Tremoulet, & Scholl, 1998; Spelke, Gutheil, & Van de Walle, 1995; Xu & Carey, 1996). Leslie et al. have explicitly recognized the close relation between this notion of object and the one that is involved in our theory of visual indexes. Typical experiments show that in certain situations, 8-month-old infants are sensitive to the cardinality of a set of (one or two) objects even before they use the properties of the individual objects in predicting what will happen in certain situations where objects are placed behind a screen and then the screen is removed. For example, Leslie et al. (1998) describe a number of studies in which one or two objects are placed behind a screen and the screen is then removed to reveal two or one objects. Infants exhibit longer looking times (relative to a baseline) when the *number* of objects revealed is different from the number that the infant sees being placed behind the screen, but not when the objects have different visual properties. This has widely been taken to suggest that registering the individuality of objects ontologically precedes the encoding of their properties in tasks involving objects' disappearance and reappearance.

While it is tempting to identify these empirical phenomena with the same notion of 'object', it remains an open question whether all these uses of the term refer to the same thing. My present use of the term is inextricably connected with the theoretical mechanism of visual indexing, and therefore to the phenomena of individuation and tracking, and assumes that such objects are picked out in a preconceptual manner. If the sense of 'object' that is needed in other contexts entails that individuating and tracking must appeal to a conceptual category, defined in terms of how the observer represents it or what the observer takes it to be, then it will not help us to ground our concepts nor will it help with the problem of keeping track of individuals during incremental construction of a percept. In the case of the MOT examples, the notion of primitive visual object introduced here does fill these functions. But of course this leaves open the question of what the connection is between the primitive visual object so-defined and the more usual notion of physical object, and in particular with the notion of object often appealed to in the infant studies. In those studies, an object is sometimes defined as a "bounded, coherent, three-dimensional physical object that moves as a whole" (Spelke, 1990). Are such Spelke objects different from what we have been calling visual objects?

The speculative answer to the question of the relation between these two notions

of object is that primitive visual objects are *typically* the proximal counterparts of real physical objects (which include Spelke objects). According to this view, the visual system is so structured that it detects visual patterns which *in our kind of world* tend to be reliably associated with entities that meet the Spelke criteria. If that is the case, then it suggests that, contrary to claims made by developmental psychologists like Spelke et al. (1995) and Xu (1997), quite possibly the *concept* of an object is not involved in picking out these objects, just as no concept at all of the individual objects (i.e. no description) plays a role in such phenomena as MOT. Despite this speculative suggestion, it is less clear whether a concept is involved in all the cases discussed in the developmental literature. From the sorts of considerations raised here, it seems likely that something more than just concepts may be involved in at least some cases of infants' picking out objects. It seems likely that a direct demonstrative reference or *indexing* is involved in at least some of the phenomena (see Leslie et al., 1998). However, there also appear to be cases in which clusters of features that one would expect would be perfectly good objects from the perspective of their visual properties may nonetheless fail to be tracked as objects by 8-month-old infants. Chiang and Wynn (2000) have argued that *if the infants are given evidence that the things that look like individual objects are actually collections of objects* then they do not keep track of them in the studies involving placing objects behind a screen, despite the fact that they do track the visually-identical collections when this evidence is not provided. For example, if infants see the putative objects being disassembled and reassembled, or if they see them come into existence by being *poured from a beaker* (Carey & Xu, 2001; Huntley-Fenner, Carey, & Salimando, 2001), they fail to track them as individual objects. This *could* mean that whether or not something is treated as an object depends on prior knowledge (which would make them conceptual in this case). On the other hand it may just mean that certain aspects of the recent visual history of the objects affects whether or not the visual system treats them as individual objects. What makes the latter at least a possibility is that something like this appears to be the case with other cases of the disappearance and reappearance of visual objects. As mentioned earlier, it has been shown that the precise *manner* in which objects disappear and reappear matters to whether or not they continue to be tracked (Scholl & Pylyshyn, 1999). In particular, if their disappearance is by a pattern of accretion such as occurs when the object goes behind an occluding surface and reappears in a complementary manner (by disocclusion), then it continues to be tracked in a MOT paradigm. But this sort of effect of recent visual history is quite plausibly subsumed under the operation of a preconceptual mechanism of the early vision system (for other examples of what appear on the surface as knowledge-based phenomena but which can be understood as the consequence of a non-cognitive mechanism, see Pylyshyn, 1999).

The central role that objects play in vision has another, perhaps deeper, consequence worth noting. The primacy of objects as the focus through which properties are encoded suggests a rather different way to view the role of objects in visual perception and cognition. Just as it is natural to think that we apprehend properties such as color and shape as properties of *objects*, so it has also been natural to think that we apprehend objects as a kind of property that particular *places* have. In other

words, we usually think of the matrix of space-time as being primary and of objects as being occupants of places and times. Yet the present proposal suggests an alternative and rather intriguing possibility. It is the notion that *primitive visual object* is the primary and more primitive category of early (preconceptual) vision. It may be that we detect *objecthood* first and determine location the way we might determine color or shape – as a property associated with the detected objects. If this is true then it raises some interesting possibilities concerning the nature of the mechanisms of early vision. In particular, it adds further credence to the claim that we must have a way of referring directly to primitive visual objects without using a unique description under which that object falls. Perhaps this function can be served in part by the mechanism I referred to as a visual index or a visual demonstrative (or a FINST).

Notice that what I have been describing is not the notion of an individual physical object. The usual notion of a *physical* object, such as a particular table or chair or a particular individual person, *does* require concepts (in particular it requires what are called *sortal* concepts) in order to establish criteria of identity, as philosophers like Hirsch (1982) and others have argued. The individual items that are picked out by the visual system and tracked primitively are something less than full-blooded individual objects. Yet because they are what our visual system gives us through a brute causal mechanism (because that is its nature), and also because the proto-objects picked out in this way are typically associated with real objects in our kind of world, indexes may serve as the basis for real individuation of physical objects. While it is clear that you cannot individuate objects in the full-blooded sense without a conceptual apparatus, it is also clear that you cannot individuate them with *only* a conceptual apparatus. Sooner or later concepts must be grounded in a primitive causal connection between thoughts and things. The project of grounding concepts in sense data has not fared well and has been abandoned in cognitive science. However, the principle of grounding concepts in perception remains an essential requirement if we are to avoid an infinite regress. Visual indexes provide a putative grounding for basic objects – the individuals to which perceptual predicates apply, and hence about which cognitive judgments and plans of action are made (see the interesting discussion of the latter in Miller, Galanter, & Pribram, 1960). Without such a preconceptual grounding, our percepts and our thoughts would be disconnected from causal links to the real-world objects of those thoughts. With indexes we can think about things (I am sometimes tempted to call them *FINGs* since they are inter-defined with *FINSTs*) without having any concepts of them: one might say that we can have *demonstrative thoughts*. We can think thoughts about *this* without *any description* under which the object of that thought falls: you can pick out one speck among countless identical specks on a beach. What's even more important is that at some stage we *must* be able to make judgments about things for which we do not have a description. For if all we had was descriptions, we would not be able to tell whether a particular description D was satisfied by some particular thing in the world, since we would have no independent way to select or refer to the thing that satisfied D. Without preconceptual reference we would not be able to decide that a particular description D was satisfied by a particular individual (i.e. by *that* individual) and thus we could not make judgments about nor decide to act upon a parti-

cular individual. It is because you can pick out a particular individual that you can move your gaze to it or you can reach for it – your motor system cannot be commanded to reach for something that is red, only to reach for a particular individual object.

Needless to say there are some details to be worked out so this is a work-in-progress. But there are real problems to be solved in connecting visual representations to the world in the right way, and whatever the eventual solution turns out to be, it will have to respect a set of both logical and empirical considerations, some of which are sketched here. Moreover, any visual or attentional mechanism that might be hypothesized for this purpose will have far-reaching implications, not only for theories of situated vision, but also for grounding the content of visual representations and perhaps for grounding perceptual concepts in general.

Acknowledgements

I wish to thank Jerry Fodor for his considerable help with this paper, particularly with regard to the many conceptual and philosophical questions raised by this work, and Brian Scholl for his careful reading of several drafts of this manuscript and for our many useful discussions and arguments over the interpretation of our results. This research was supported in part by NIHM grant 1R01-MH60924.

References

Acton, B. (1993). *A network model of visual indexing and attention*. Unpublished master's thesis, University of Western Ontario, London, Canada.

Agre, P. E. (1997). *Computation and human experience*. Cambridge: Cambridge University Press.

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance*, *26* (2), 834–846.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20* (4), 723–767.

Baylis, G. C., & Driver, J. (1993). Visual attention and objects: evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 451–470.

Blaser, E., Pylyshyn, Z. W., & Domini, F. (1999). Measuring attention during 3D multielement tracking (abstract). *Investigative Ophthalmology and Visual Science*, *40* (4), 552.

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature-space. *Nature*, *408* (Nov 9), 196–199.

Bregman, A. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge, MA: MIT Press.

Brooks, R. A. (1999). *Cambrian intelligence*. Cambridge, MA: MIT Press (A Bradford Book).

Burkell, J., & Pylyshyn, Z. W. (1997). Searching through subsets: a test of the visual indexing hypothesis. *Spatial Vision*, *11* (2), 225–258.

Calis, G. J., Sterenborg, J., & Maarse, F. (1984). Initial microgenetic steps in single-glance face recognition. *Acta Psychologica*, *55* (3), 215–230.

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: beyond object files and object tracking. *Cognition*, this issue, *80*, 179–213.

Chiang, W. -C., & Wynn, K. (2000). Infants' representation and tracking of multiple objects. *Cognition*, *75*, 1–27.

Culham, J. C., Brandt, S. A., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. H. (1998).

Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*, *80* (5), 2657–2670.

Currie, C. B., McConkie, G. W., & Carlson-Radvansky, L. A. (2000). The role of the saccade target object in the perception of a visually stable world. *Perception & Psychophysics*, *62*, 673–683.

Dawson, M., & Pylyshyn, Z. W. (1988). Natural constraints in apparent motion. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision: an interdisciplinary perspective* (pp. 99–120). Stamford, CT: Ablex.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, *129* (4), 481–507.

Driver, J., & Halligan, P. (1991). Can visual neglect operate in object-centered coordinates? An affirmative single case study. *Cognitive Neuropsychology*, *8*, 475–494.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: a zoom lens model. *Perception & Psychophysics*, *40*, 225–240.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254.

He, S., Cavanagh, P., & Intriligator, J. (1997). Attentional resolution. *Trends in Cognitive Sciences*, *1* (3), 115–121.

Henderson, J. M., & Anes, M. D. (1994). Roles of object-file review and type priming in visual identification within and across eye fixations. *Journal of Experimental Psychology: Human Perception and Performance*, *20* (4), 826–839.

Hirsch, E. (1982). *The concept of identity*. Oxford: Oxford University Press.

Huntley-Fenner, G., Carey, S., & Salimando, A. (2001). Sand does not count: infant individuation of objects and non-solid substances. Manuscript submitted for publication.

Intriligator, J. M. (1997). *The spatial resolution of attention*. Unpublished Ph.D., Harvard University, Cambridge, MA.

Intriligator, J., & Cavanagh, P. (1992). Object-specific spatial attention facilitation that does not travel to adjacent spatial locations (abstract). *Investigative Ophthalmology and Visual Science*, *33*, 2849.

Irwin, D. E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, *5* (3), 94–100.

Irwin, D. E., McConkie, G. W., Carlson-Radvansky, L. A., & Currie, C. (1994). A localist evaluation solution for visual stability across saccades. *Behavioral and Brain Sciences*, *17*, 265–266.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, *24* (2), 175–219.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, this issue, *80*, 97–126.

Leslie, A. M., Xu, F., Tremoulet, P. D., & Scholl, B. J. (1998). Indexing and the object concept: developing 'what' and 'where' systems. *Trends in Cognitive Sciences*, *2* (1), 10–18.

Lespérance, Y., & Levesque, H. J. (1995). Indexical knowledge and robot action – a logical account. *Artificial Intelligence*, *73*, 69–115.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

McConkie, G. M., & Currie, C. B. (1996). Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *22* (3), 563–581.

McKeever, P. (1991). *Nontarget numerosity and identity maintenance with FINSTs: a two component account of multiple-target tracking*. Unpublished master's thesis, University of Western Ontario, London, Canada.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: a critical link between lower-level and higher-level vision. In S. M. Kosslyn, & D. N. Osherson (Eds.), *Visual cognition* (pp. 1–70). Cambridge, MA: MIT Press.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

Nissen, M. J. (1985). Accessing features and objects: is location special? In M. I. Posner, & O. S. Marin (Eds.), *Attention and performance* (Vol. XI, pp. 205–219). Hillsdale, NJ: Lawrence Erlbaum.

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

Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press (A Bradford Book).

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

Posner, M. I., Snyder, C., & Davidson, B. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160–174.

Pylyshyn, Z. W. (1984). *Computation and cognition: toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

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

Pylyshyn, Z. W. (1998). Visual indexes in spatial vision and imagery. In R. D. Wright (Ed.), *Visual attention* (pp. 215–231). New York: Oxford University Press.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22* (3), 341–423.

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, *4* (5), 197–207.

Pylyshyn, Z. W., Burkell, J., Fisher, B., Sears, C., Schmidt, W., & Trick, L. (1994). Multiple parallel access in visual attention. *Canadian Journal of Experimental Psychology*, *48* (2), 260–283.

Pylyshyn, Z. W., & Eagleson, R. A. (1994). Developing a network model of multiple visual indexing (abstract). *Investigative Ophthalmology and Visual Science*, *35* (4), 2007.

Pylyshyn, Z. W., Elcock, E. W., Marmor, M., & Sander, P. (1978). *Explorations in visual-motor spaces*. Paper presented at the Proceedings of the Second International Conference of the Canadian Society for Computational Studies of Intelligence, University of Toronto, Toronto, Canada.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3* (3), 1–19.

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

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

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, *8* (5), 368–373.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition*, *7*, 127–145.

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

Robertson, L., Treisman, A., Friedman-Hill, S., & Grabowecky, M. (1997). The interaction of spatial and object pathways: evidence from Balint's syndrome. *Journal of Cognitive Neuroscience*, *9* (3), 295–317.

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition*, this issue, *80*, 1–46.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, *38* (2), 259–290.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object: evidence from multiple-object tracking. *Cognition*, this issue, *80*, 159–177.

Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (2001). The relationship between property-encoding and object-based attention: evidence from multiple-object tracking, submitted for publication.

Schulz, T. (1991). A microgenetic study of the Mueller-Lyer illusion. *Perception*, *20* (4), 501–512.

Sears, C. R., & Pylyshyn, Z. W. (2000). Multiple object tracking and attentional processes. *Canadian Journal of Experimental Psychology*, *54* (1), 1–14.

Sekuler, A. B., & Palmer, S. E. (1992). Visual completion of partly occluded objects: a microgenetic analysis. *Journal of Experimental Psychology: General*, *121*, 95–111.

Simons, D. J. (1996). In sight, out of mind: when object representations fail. *Psychological Science*, *7* (5), 301–305.

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

Smith, B. C. (1996). *On the origin of objects*. Cambridge, MA: MIT Press.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Spelke, E. S., Gutheil, G., & Van de Walle, G. (1995). The development of object perception. In S. M. Kosslyn, & D. N. Osherson (Eds.), *Visual cognition* (Vol. 2, pp. 297–330). Cambridge, MA: MIT Press.

Tipper, S., Driver, J., & Weaver, B. (1991). Object-centered inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology*, *43A*, 289–298.

Tipper, S. P., Weaver, B., Jerreat, L. M., & Burak, A. L. (1994). Object-based and environment-based inhibition of return of selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 478–499.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited capacity preattentive stage in vision. *Psychological Review*, *10* (1), 1–23.

Tsotsos, J. K. (1988). How does human vision beat the computational complexity of visual perception. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision: an interdisciplinary perspective* (pp. 286–340). Stamford, CT: Ablex.

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

Viswanathan, L., & Mingolla, E. (in press). Dynamics of attention in depth: evidence from multi-element tracking. *Perception*.

Watson, D. G., & Humphreys, G. W. (1997). Visual marking: prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological Review*, *104* (1), 90–122.

Wolfe, J. M., & Bennett, S. C. (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research*, *37* (1), 25–43.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15* (3), 419–433.

Xu, F. (1997). From Lot's wife to a pillar of salt: evidence that *physical object* is a sortal concept. *Mind and Language*, *12*, 365–392.

Xu, F., & Carey, S. (1996). Infants' metaphysics: the case of numerical identity. *Cognitive Psychology*, *30*, 111–153.

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

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

Yantis, S., & Johnson, D. N. (1990). Mechanisms of attentional priority. *Journal of Experimental Psychology: Human Perception and Performance*, *16* (4), 812–825.

Yantis, S., & Jones, E. (1991). Mechanisms of attentional selection: temporally modulated priority tags. *Perception & Psychophysics*, *50* (2), 166–178.