The Empirical Case for Bare Demonstratives in Vision

Zenon Pylyshyn, Rutgers Center for Cognitive Science

Table of Contents

The Empirical Case for Bare Demonstratives in Vision

Zenon Pylyshyn,
Rutgers Center for Cognitive Science,
Rutgers, the State University of New Jersey, New Brunswick, NJ

1. Background: Representation in language and vision

One of the most important ideas that developed in the late 20th century, and for which Chomsky, Fodor and Newell/Simon can take much of the credit, is realism about mental representations. In the human sciences, realism about theoretical (and especially mental) entities had fallen out of fashion in the middle third of the 20th century. It seems to me that there were two things that made the difference in bringing cognitivism, back into psychological science. One was the work that began with Hilbert and was developed by Turing and Church and Markov and others who formulated the abstract notions of mechanism and of what we now call "information processing." This is the lineage that led to Cybernetics and later to Artificial Intelligence, though a very large proportion of the field would now probably dissociate itself with that "logicist" part of the family tree, just as earlier Logicists like Frege dissociated themselves with psychological pursuits. The other development that brought mentalism back was the discovery that it was possible to treat some aspects of the human capacity for language in a way that made it at least appear to be compatible with mechanism. These developments encouraged many people to hope that one day we might have an explanatory theory of some of the mechanisms of linguistic competence, not just a taxonomic description of a corpus of linguistic utterances. The specific results achieved in transformational grammar, coupled with the generative or procedural aspect of the theoretical mechanisms (which, after all, wore the formal garb of Post Production systems and of Markov Algorithms) gave us hope that we were on the track of a theory of language understanding and language production.

Well, we were wrong about a lot of things, and especially about how a theory of grammar might be incorporated into a theory of comprehension/production (recall, for example, the decisive failure of the “derivational theory of complexity”). Many of the early ideas of psycholinguistics were later abandoned. What remained, however, was the basic belief that both rules, which included “rules of grammar”, and formal structures (of sentences) would play a central role in the theory of not only the language capacity, but also of cognition more generally. Moreover, ever since those developments in the late 50's and early 60's, talk about rules and representations no longer meant we were describing a corpus of behavior; rather when we spoke of rules we were referring to an internal property of some system or mind. We now routinely spoke of rules and the structures that they generate as being “internally represented”.

What was meant by the phrase “internally represented,” however, was far from clear – even to those of us who spoke that way. And it does not get any clearer if one adopts Chomsky's way of putting it when, for example, he says that a theory of the speaker/hearer “involves rules”, or that the theory postulates a certain rule R “as a constituent element of [the speaker/hearer's] initial state” or “attributes to ...[the speaker/hearer] a mental structure ... that includes the rule R and explains his behavior in terms of this attribution” (Chomsky, 1986, p243); or when he says that a speaker is “equipped with a grammar” or “internalizes a system of rules”. Yet, despite the uncertainties, none of us doubted that what was at stake in all such claims was nothing less than an empirical hypothesis about *how things really are inside the head of a human cognizer*. We knew that we were not speaking metaphorically nor were we in some abstract way describing the form of the data.

The way the story has gone within the study of language, including psycholinguistic studies of human performance, is now familiar to cognitive scientist, at least in broad outline if not in detail. But there is another area of cognitive science, quite different from the study of language, that has also made considerable progress: That is the area of visual perception. Under the important influence of David Marr (Marr, 1982), who saw the parallels between his enterprise and Chomsky's; visual perception, like language, was seen as being essentially modular (Pylyshyn, 1999), as amenable to the sort of competence-performance distinction that made progress in linguistics possible; and as fundamentally concerned with questions of representation. There has probably been more progress in the study of visual perception –and more interaction between the evidence of psychophysics, phenomenology and neuroscience– than in any other area of cognitive science. At the same time, there has been nearly as much misunderstanding and ideological dispute in the study of visual perception as there has been over the years in the study of language. In what follows I will discuss one recent line of work in which I have been involved that concerns the nature of the representations underlying visual perception (including one major shortcoming of the received view).

In addition to the broad methodological point that we need to distinguish between competence and performance, which informs both linguistics and vision science, the two fields share other properties, both methodological and substantive.

2. Some parallels between the study of vision and language

The study of language and linguistic processes (learning, parsing, understanding and genarating) developed in parallel with the development of our understanding of what the basic goals were and how the major problems in the field were to be understood. The study of visual perception has also developed in a similar way, as we developed a clearer view of its goals and to such questions as the following.

(1) Is vision a distinct process or is it continuous with cognition? If the former, then how can we draw the boundary between vision and cognition?

(2) Are the sources of evidence used in the study of vision special in any way? Do they, for example, include the equivalent of the sort of “judgments” (of grammaticality and ambiguity) used routinely by linguistics?

(3) What is the function computed by vision? Can we characterize the inputs and the outputs of vision –i.e., the representations that vision computes– in a perspicuous way, and in a way that shows its connection with general cognition?

(4) What form of representation is computed by visual processes? Is the form of representation similar to the form of representation computed for language (e.g., Logical Form) or must it be different in fundamental ways?

In what follows I will focus primarily on the last item (4). Before I do that, however, I would like to point out some considerable similarities between vision and language processing as well as similarities of methodology faced by vision science and linguistics. On the face of it there are many similarities between language and vision. They are both productive so there is no limit on how many patterns can be generated or recognized and in both cases similarities among patterns require appeal to the structure of the stimuli (i.e., both achieve their paradigmatic structure – the similarities and differences among distinct stimuli – by virtue of differences in the syntagmatic or syntactic structure among elements within each stimulus). Another way to put this is that in both vision and language there is syntactic structure which must be expressed by structure-dependent rules. Recognition of the type of each linguistic stimulus proceeds by the reconstruction of its structure through a process called parsing. In vision recognition also proceeds by a form of parsing (as developed, for example, in the recognition-by-components theory, Biederman, 1987). Also both language and vision reveal a substantial amount of innate structure and what rule-learning there is has to deal with the poverty of the stimulus – the fact that a finite set of samples of patterns is logically insufficient for inferring the structural rules.

Moreover, determining the structure of individual patterns (parsing) must deal with missing parts: stimuli in both language and vision contain unexpressed parts that are filled-in by the observer: language structures contain gaps, deletions and traces that are not expressed in the physical signal, and vision routinely deals with partially-occluded patterns which are completed and filled in by the visual system (by a process called amodal completion, as illustrated by the many Kanizsa figures, see e.g., Kanizsa, 1979). The filling-in in both cases is done by modular processes, as opposed to being inferred from general knowledge. These general similarities suggest that the processes in both cases may be similar even though they are independent of one another.

Now consider the 4 questions set out above. The first question (#1 above), whether language and vision are distinct modules (or, as Chomsky puts it, different organs) has, I believe, been answered in the affirmative in both domains (I have argued the case for a visual module in Pylyshyn, 1999). Although there remain border skirmishes, as there always are at borders, it seems clear that vision and language both involve distinct functions and even distinct areas of the brain. The debate ultimately turns on the question of where the boundary is and that awaits the development of better and more general theories because ultimately it is the theory that tells you how to deal with the gray areas.

The same might be said of the second question (2). Linguistics has always used intuitions of native speakers regarding such phenomena as grammaticality, ambiguity and paraphrase. But these were subject to considerable argument in early years of generative grammar because one can’t just ask someone whether a sentence is grammatical or ambiguous or whether two sentences mean the same thing. The very notions of grammaticality, ambiguity and sameness of meaning are *theory laden* (to use a term from Hanson, 1958). The sentence that Chomsky used in his earliest writing to illustrate the difference between grammaticality and acceptability

("Colorless green ideas sleep furiously") was the subject of criticism and many people produced interpretations of the sentence to show it was meaningful. Intuitions of grammaticality are always problematic.[1] Yet in recent times the use of intuition in linguistics has not disappeared – it continues to play a central role in linguistic theory-building. But now it is used to answer well-posed questions derived from the theories. There is a similar problem in vision science where the appeal to "how things look" or to the contents of conscious experience is similarly problematic. I need not list the many ways that conscious experience is misleading nor the well-known cases where vision is unaccompanied by any conscious experience at all. In fact when people report on what they experienced, or "what things look like" their reports may be guided by their own folk theories and expectations. As I have recently claimed (Pylyshyn, forthcoming, Chapter 4) although we cannot stop using "what things look like" as a source of evidence, we need to use this kind of evidence in conjunction with evolving theories, just the way linguistic intuitions have been tamed by theories.

Question (3) is more problematic. In the case of language the input is either an acoustical stream or a string of linguistic units or formatives: phonemes or morphemes or lexical items, depending on whether the theory is to accommodate phonology, morphology or only syntax. It is widely held that these are independent levels of description that can be addressed separately. In the case of vision one might think of the input as consisting of an image, such as found on the retina [2]. But a case can be made that vision is an active process so the input might be better described in terms of what Gibson called the *ambient optical array* through which the organism moves and explores. There is also the question of the output of vision (assuming that vision really is an independent module). In has generally been assumed that the output of vision is much like the output of the language analyzer – logical forms. In any case few people think of the output of vision as anything but a symbolic description since without that vision would not inform the organism and lead to belief fixation (the exception being people who have advocated a theory of mental imagery that claims it uses the mechanisms of vision because in that case vision and mental images both generate displays in the brain, as opposed to logical forms).

The question can be raised of whether Logical Form, such as discussed in language, or some other essentially *descriptive* form of representation is adequate for representing visual percepts. The answer I am offering is that it is not. But I am not about to suggest that visual percepts should be thought of as *pictorial* or *analogue* or any other sort of ill-understood formats that many writers have proposed (Kosslyn, 1994). I find such proposals to be either hopelessly underspecified and metaphorical or else clearly false, although this is not the place to say why (see, however, Pylyshyn, 2002; Pylyshyn, forthcoming). What I claim is that the representations underlying visual percepts are mostly symbolic conceptual descriptions of roughly the classical sort. But I will also argue that notwithstanding the need for a logical form to allow perception to inform thoughts, this form of representation is incomplete in at least one critical respect – it lacks resources for picking out and referring to particular token individuals in the world. They lack, in other words, the special power that *demonstratives* have in language.

Demonstrative terms (and indexicals in general) differ from other terms primarily in the way that they function, the way they convey information. There they play a very important role in communication, thought and action, where they refer to token individuals. It is there that they come essentially into contact with perception; demonstratives pick out individual tokens in the perceptual field, both in communication and in thought. They are, as Perry and Kaplan have argued, indispensable in language and thought (Almog, Perry, & Wettstein, 1989; Perry, 1979). What remains controversial among philosophers of language is whether there are *bare* demonstratives or only complex demonstratives. A bare demonstrative refers to an individual without at the same time referring to it as something that falls under some conceptual category or other (as when we think “this”) whereas a complex demonstrative works like a descriptive noun phrase to pick out an individual that has the properties mentioned or implied by the referring expression (as in "this brown dog"). Ernie Lepore has been one of the defenders of the position that there are bare demonstrative, and moreover that complex demonstratives rely on the prior selection made by the bare demonstrative implied by demonstrative phrases (Lepore & Ludwig, 2000). This is exactly the position that I have taken with respect to visual demonstratives. Since one of the functions that demonstrative reference plays (either in spoken language or the language of thought) is that of grounding conceptual representations in perception, then it must be that at least some of the things that perception picks out must be picked out without regard to the conceptual category it falls under – in other words it must contain a bare demonstrative.

2.1 Augmenting the *Language of Thought* to include demonstratives

I have defended the appropriateness of what I call here the classical symbolic view of visual representation on a number of different grounds (Pylyshyn, 2003). For example I have cited such properties as the abstractness and variability in definiteness of our visual representations (the way sentences can be abstract and variable in the sorts of details they encode) and the necessity that the system of representations meet the usual requirements of productivity and systematicity that Fodor and I discussed in connection with our critique of connectionist proposals (Fodor & Pylyshyn, 1988). I believe that compositional symbolic representations are the only form of representation that even come close to having the sort of requisite expressive power for visual percepts, even though they remain incomplete in a number of ways, such as their inability to conveniently encode magnitudes and the inability to individuate and reference tokens of visual objects. It is the latter shortcoming that I will discuss in this essay. A more extensive argument, with empirical evidence to support the detailed assumptions, is presented in (Pylyshyn, 2001a, 2003, forthcoming).

Theories of visual perception attempt to give an account of how a proximal stimulus (presumably a pattern impinging on the retina) can lead to a rich representation of a distal three-dimensional world and thence to either the recognition of known objects or to the coordination of actions with visual information. Such theories typically provide an effective (i.e., computable) mapping from a 2D pattern to a representation of a 3D scene, usually in the form of a symbol structure. But such a mapping, though undoubtedly one essential purpose of a theory of vision, leaves at least one serious problem. The problem is that of connecting visual representations with the world in a certain critical way. This problem occurs for a number of reasons, but for our purposes I will emphasize just one such reason: the mapping from the world to our visual representation is not arrived at in one step, but incrementally. We know this both from empirical observations (e.g., percepts are generally built up by scanning attention and/or one's gaze) and also from theoretical analysis — e.g., Ullman, (1984) has provided good arguments for believing that some relational properties, such as the property of being inside or on the same contour, have to be encoded serially by scanning a display. But then one problem arises immediately: If the representation is built up incrementally, we need to know that a certain part of our current representation refers to a particular individual object in the world. The reason is quite simple. As we elaborate the representation by uncovering new properties of a scene that we have partially encoded we need to know where (i.e., to which part of the representation) to attach the new information. In other words we need to know when a certain token in the existing representation should be taken as corresponding to the same (real, physical, individual) object as a particular token in the new representation, so that we can append newly noticed properties to the representation of the appropriate individual objects.

A possible way in which a purely descriptive representation could pick out individuals is by using definite descriptions. It could, for example, assert things like “the object *x* that has property P” where P uniquely picks out a particular object *x*. In that case, in order to add new information, such as that this particular object also has property Q one would add the new predicate Q and also introduce an identity assertion, thus asserting something like P(*x*) $\wedge$ Q(*y*) $\wedge$ *x* = *y* (and, by the way, adding this new compound descriptor to memory so that the same object might be relocated in this way when a further new property of that object is later noticed).[3] But this is almost certainly not how the visual system adds information. This way of adding information would require adding a new predicate Q to the representation of an object that is *picked out by a certain descriptor*. To do that would require first recalling the description under which *x* was last encoded, and then conjoining to it the new descriptor and identity statement. Each new description added would require retrieving the description under which the object in question was last encoded.

The alternative to this unwieldy method is to allow the descriptive apparatus to make use of singular terms such as names or demonstratives. If we do that, then adding new information would amount to adding the predicate Q(*a*) to the representation of a particular object *a*, and so on for each newly noticed property of *a*. Empirical evidence that we will review below suggests that the visual system's Q-detector recognizes instances of the property Q *as a property of a particular visible object*, such as object *a*, this is the most natural way to view the introduction of new visual properties by the sensorium. In order to introduce new properties in that way, however, there would have to be a non-descriptive way of picking out *a,* such as a singular term or a name or a demonstrative. This is, in effect, what labeling objects in a diagram does through external means and what demonstrative terms like “this” or “that” do in natural language.[4] This alternative is *prima facie* the more plausible one since it is surely the case that when we detect a new property we detect it as applying to *that* object, rather than as applying to some object in virtue of its being the object with a certain (recalled) property.[5] Such intuitions, however, are notoriously unreliable so later in this paper I will examine empirical evidence which suggests that this view is indeed more likely to be the correct one. For example, I will describe studies involving multiple-object tracking that make it very unlikely that objects are tracked by regularly updating a description that uniquely picks out the objects. In these studies the only unique descriptor available is location, and under certain plausible assumptions the evidence shows that it is very unlikely that the coordinates of the points being tracked are being regularly updated so that tracking is based on maintaining identity by updating descriptions.

There are a number of other reasons why a visual representation needs to be able to pick out individuals the way demonstratives do (i.e., independent of their properties or locations). For example, among the properties that are extracted (and presumably encoded in some way) by the visual system are a variety of relational predicates, such as **Collinear**(X1, X2, ...Xn) or **Inside**(X1,C1) or **Part-of**(F1,F2), and so on. But these predicates apply over distinct individual objects in the scene independent of what properties these individuals have. So in order to recognize a relational property involving several objects we need to specify which objects are involved. For example, we cannot recognize the **Collinear** relation without picking out which objects are recognized as collinear. If there are many objects in a scene only some of them may be collinear so we must associate the relation with the objects in question. This is quite general since properties are predicated of things, and relational properties (like the property of being "collinear") are predicated of several things. So there must be a way, independent of the process of deciding which property obtains, of specifying which objects (in our current question-begging sense) have that property. Ullman, as well as a large number of other investigators (Ballard, Hayhoe, Pook, & Rao, 1997; Watson & Humphreys, 1997; Yantis & Jones, 1991) talk of the objects in question as being "tagged" (indeed, "tagging" is one of the basic operations in Ullman's theory of visual routines). The notion of a tag is an intuitive one since it suggests a way of *marking objects* for reference purposes. But the operation of tagging only makes sense if there is some thing on which a tag can literally be placed. It does no good to tag an internal representation (unless one assumes that it is an exact copy of the world) since the relation we wish to encode holds in the world and may not hold in the representation. But how do we tag parts of the world? What we need is what labels gave us in the previous example: A way to name or refer to individual parts of a scene *independent of their properties or their locations*.

What this means is that the representation of a visual scene must contain something more than descriptive or pictorial information in order to allow re-identification of particular individual visual elements. It must provide what natural language provides when it uses names (or labels) that uniquely pick out particular individuals, or when it embraces demonstrative terms like “this” or “that”. Such terms are used to indicate particular individuals. This assumes that we have a way to *individuate* [6] *and keep track of particular individuals in a scene* even when the individuals change their properties, including their locations. Thus what we need are two functions that are central to our concern: (a) we need to be able to pick out or individuate distinct individuals (following current practice, we will call these individuals *objects*) and (b) we need to be able to refer to these objects as though they had names or labels. Both these purposes are served by a primitive visual mechanism that I call a *visual index*. So what remains is for me to provide an empirical basis for the claim that the visual system embodies a primitive mechanism of the sort I call a *visual index* or a FINST. I begin with a description of the first of the two functions it provides, that of *individuating primitive visible objects*.

2.2 Primitive visual objects

Let me first provide a sketch of how the notion of an object has come into general use in the study of vision and visual attention. I will first describe a number of experiments that suggest that the detection of certain properties, such as color or shape or location, are perceptually separate from the detection of the individuals that bear them, and that the detection of objects likely precedes the detection of their properties. Then I will describe some experiments that further show that what the visual system detects when it is said to detect objects is not a proximal feature-cluster, but something that persists despite certain sorts of changes in its properties, including its location. By then we will see that the application of the term *object,* while still insufficient to bear the load of what is required of a real *individual*, as philosophers understand this term, begins to be much more interesting. In fact it offers a construct that I will call a *primitive visible object* that will be the building block for a story of how certain thoughts can be grounded on basic perceptual processes — i.e., how we can think about something for which we have no concept.

Evidence of independent recognition of objects and their properties in early vision. Interest in what is now referred to as object-based attention may have begun with the observation that under certain conditions there appears to be a dissociation between the perception of certain properties and the perception of which objects have those properties. In fact it seems as though attention is required in order to bind properties to their bearers. For example, Anne Treisman and her colleagues showed that when properties of items not under direct attentional scrutiny were reported from a visual display there were frequent errors in which properties were assigned to the wrong items, resulting in what are called "illusory conjunctions". For example, (Treisman & Gelade, 1980) found that if attention was distracted by a subsidiary visual task (such as naming digits at the center of the display), subjects frequently reported seeing the correct shape and color of items but in the wrong combinations resulting in erroneous *conjunctions* of color and shape (e.g., they reported that the display contained a red X and a green O when in fact it had contained a green X and a red O). The illusory conjunctions appear with a large variety of properties of objects (Treisman & Gelade, 1980; Treisman & Sato, 1988). For example, illusory conjunctions occur for shape properties so that a display with right oblique lines, L-shaped and S-shaped figures led to the misperception of triangles and dollar signs. There is also evidence that certain object properties can be detected while their locations are either misidentified or unknown.[7] Thus you might see that a display contains the letter X but fail to detect *where* it was located, or "see" it to be at the wrong location (Chastain, 1995; Treisman, 1986). There has also been considerable interest in recent years in the so-called "two visual systems" view (Ungerleider & Mishkin, 1982) which claimed that there are two streams of visual processing in the brain: A dorsal stream that encodes *where* a thing is and a ventral stream that encodes *what* it is (its identity).[8] These and related studies (including demonstrations that people can attend to large random shape embedded within other shapes which they must ignore – Rock & Gutman, 1981) suggested that attention is allocated to what are called *objects* (or individuals) rather than to particular *places, regions,* or *properties*. There is even evidence from the clinical syndrome known as unilateral neglect that what is neglected must be described in relation to perceptual objects rather than locations in space (Tipper & Behrmann, 1996).

Evidence that extracting several pieces of information from a display is easier if they are part of one object. The notion that objects are detected and then visual properties are bound to them at a very early stage in visual perception has also received support from studies showing that it is faster to find (and identify) several features or a properties if they are associated with the same object (and also features that are part of different objects interfere less in a search task). For example, (Duncan, 1984) and later (Baylis & Driver, 1993) showed that access to relational properties of two features (such as "larger than") is faster when the features in question belong to the same perceptual object than when they are parts of different objects which nonetheless are objectively in the same relative relation (e.g., the same distance apart). These studies all point to the idea that objects are selected first and *then* properties of these objects may be encoded and available for judgments.

Evidence for access to multiple objects. In order to detect such relational properties as that a number of points are collinear or that a point is inside a closed contour the visual system must have a way to refer to the individuals over which these predicates are supposed to apply. In general, to evaluate P(x,y) both x and y need to be bound to the individuals in question. Yet attention has generally been assumed to be unitary: you can devote attention to only one thing at a time (not one *place* at a time[9]). Since we can move attention from object to object there must be some way to specify which object to move it to next. We must have some pre-attentional access or variable binding mechanism. So the mechanism for binding mental variables to objects must be more primitive than and precede the allocation of focal attention. Visual Index Theory (Pylyshyn, 2001b) claims that prior to the allocation of focal unitary attention visual indexes (or FINSTs) must be “grabbed” by portions of the visual landscape. The function of these indexes is to provide a way to access objects on demand, or to bind parts of the cognitive representation to objects. How many objects? Empirically we have found the number to be around 4 or 5 over a wide variety of experimental paradigms.

Several properties of the indexing process are illustrated by a series of studies we have performed involving selecting a subset of items in a visual search task. The search task we used was adapted from one originally introduced by (Treisman & Gelade, 1980). In a series of studies, Jacquie Burkell and I (Burkell & Pylyshyn, 1997), used the sudden-onset of new objects (which we called “late-onset placeholders”) to control search. The empirical question was whether the search would be confined to the subset defined by the late-onset objects - those that we assumed had been indexed. The answer was unequivocal: Only indexed objects constituted the search set. Moreover, it made no difference how far apart the indexed objects were, showing that they did not have to be searched out before being matched against the search criteria. (For more details on these and a number of other studies see, Pylyshyn, 2003; Pylyshyn, Burkell, Fisher, Sears, Schmidt et al., 1994).

2.3 Individuating and tracking primitive visual objects: Multiple Object Tracking studies

Perhaps the clearest way to see what is being claimed when I say there is a primitive mechanism in early vision that picks out and maintains the identity of visible objects is to consider a set of experiments, carried out in my laboratory, to which the ideas of visual individuation and identity maintenance were applied. The task is called the *Multiple Object Tracking (MOT) Task*.

In a typical experiment, subjects are shown a screen containing anywhere from 12 to 24 simple identical objects (points, spheres, plus signs, figure-eight shapes) which move across the entire visual field in unpredictable ways without colliding. A subset of these objects is briefly rendered distinct (usually by flashing them on and off a few times). The subject's task is to keep track of this subset of objects (called "targets). At some later time in the experiment (say 10 seconds into the tracking trial) one of the objects is again flashed on and off. The subject must then indicate whether or not the flashed (probe) figure was one of the targets. A large number of experiments, beginning with studies by (Pylyshyn & Storm, 1988), have shown clearly that subjects can indeed track up to 5 independently moving identical. Moreover, we were able to argue that the motion and dispersion parameters of the original Pylyshyn & Storm experiment were such that tracking could not have been accomplished using a serial strategy in which attention is scanned to each figure in turn, storing its location, and returning to find the figure closest to that location on the next iteration, and so on. Based on some weak assumptions about how fast focal attention might be scanned and based on actual data on how fast the objects actually moved and how close together they had been in this study, we were able to conclude that such a serial tracking process would very frequently end up switching to the wrong objects in the course of its tracking. This means that the moving objects could not have been tracked 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 or *individuality.* This is exactly what I claim is going on — tracking by maintenance of a primitive perceptual individuality.

Recently a large number of additional studies in our laboratory have replicated these multiple object tacking results, confirming that subjects can successfully track several independently moving objects.[10] Moreover, 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 (Pylyshyn, Burkell, Fisher, Sears, Schmidt, & Trick, 1994). Using a different tracking methodology, Intriligator & Cavanagh (1992) also failed to find any evidence of a “spread of attention” to regions between targets. 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. Taken together these studies implicate a notion of primitive visible object as a category induced by the early visual system, preceding the recognition of properties and preceding the evaluation of any visual predicate.

The multiple object tracking task exemplifies what I mean by “tracking” and by “maintaining the identity” of objects. It also operationalizes the notion of “primitive visible object” — a primitive visible object is whatever attracts a FINST index and allows multiple-object tracking. Note that this is a highly mind-dependent definition of objecthood. Objecthood and object-identity are defined in terms of a causal perceptual mechanism. A certain sequence of object-locations will count as the movement a single object if the early (pre-attentive) visual system groups it this way — i.e., if it is so perceived — whether or not we can find a physical property that is invariant over this sequence and whether or not there exists a psychologically-plausible description that covers this sequence. The visual system may also count as one individual object certain kinds of disappearances and reappearances of visual objects. For example, Scholl & Pylyshyn (1998) have shown 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 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), then they are tracked as though they were continuously moving objects. Performance in the MOT task does not deteriorate if targets disappear in this fashion although it suffers dramatically if targets suddenly go out of existence and reappear, or if they slowly shrink away and then reappear by slowly growing again at exactly the same place as they had accreted in the occlusion condition.

2.4 A theory of Visual Indexing and Binding: The FINST mechanism

The basic motivation for postulating Visual Indexes is that, as we saw at the beginning of this essay, there are a number of reasons for thinking that 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* must be maintained or tracked despite changes in the individual's properties including its location in the visual field. Our proposal claims that this is done *primitively* without identifying the object through a unique descriptor. The object in question must be segregated from the background or picked out as an individual (the Gestalt notion of making a figure-ground distinction is closely related to this sort of “picking out”). Until some piece of the visual field is segregated and picked out, no visual operation can be applied to it since it does not exist as something distinct from the entire field.

In its usual sense (at least in philosophy), picking out an individual requires having criteria of individuation — i.e., requires having a sortal concept. How can we track something without re-recognizing it as the same thing at distinct periods of time, and how can we do that unless we have a concept or a description of it? My claim is that just as the separation of figure from ground (the "picking out") is a primitive function of the architecture of the visual system, so also is this special sort of preattentive tracking. What I am proposing is not a full-blooded sense of identity-maintenance, but a sense that is relativized to the basic character of the early visual system. The visual system cannot in general re-recognize objects as being the same without some descriptive apparatus, but it can track in a more primitive sense, providing certain conditions are met (several of these conditions were mentioned earlier in discussing the Yantis and the Pylyshyn & Scholl results cited above).

What this means is that our theory is concerned with a sense of *picking out* and *tracking* that are not based on top-down *conceptual* descriptions, but are given pre-conceptually by the early visual system, and in particular by the FINST indexing mechanism. Moreover, the visual system treats the object so picked-out as distinct from other individuals, independent of what properties this object might have. If two different objects are individuated in this way they remain distinct as far as the visual system is concerned. Moreover, they remain distinct despite certain changes in their properties, particularly changes in their location. Yet the visual system need not know (i.e., need not have detected or encoded) any of their properties in order to implicitly treat them as though they were distinct and enduring visual tokens. Of course there doubtless are properties, such as being in different locations or moving in different ways or flashing on and off that allow indexes to be assigned to these primitive objects in the first place. But none of these properties define the objects — they are not essential properties. What is an essential property is that it attracted an index for *any possible reason*! My claim is that to index *x, in this primitive sensory sense,* there need not be any concept, description or sortal that picks out *x*'s by type.[11]

The basic idea of the FINST indexing and binding mechanism is that a causal chain leads from certain kinds of visible events, via primitive mechanisms of the early visual system, to certain conceptual structures (which we may think of as symbol structures in Long Term Memory). This provides a mechanism of reference between a visual representation and what we have called primitive visible objects in the world. The important thing here is that the inward effects are purely causal and are instantiated by the non-conceptual apparatus of what I have called early vision (Pylyshyn, 1999). This apparatus guarantees that under certain conditions the link will maintain a certain continuity, thus resulting in its counting as the *same link*. It is tempting to say that what makes it continuous is that it keeps pointing to the same thing, but according to our view this is circular since the only thing that makes it the *same thing* is the very fact that the it the index references it. There is no other sense of “sameness” so that “primitive visible object” as we have defined it is thoroughly mind dependent.

By virtue of this causal connection, the conceptual system can *refer to* any of a small number of primitive visible objects. It can, for example, interrogate them to determine some of their properties, it can evaluate visual predicates (such as ***Collinear***) over them, it can move focal attention to them, and so on. The function that I am describing is extremely simple and only seems complicated because ordinary language fails to respect certain distinctions (such as the distinction between individuating and recognizing, indexing and knowing where something is, and so on). Elsewhere (Pylyshyn, 2003) I provide an extremely simple network, based on the Koch & Ullman (1984) winner-take-all neural net, which implements such a function.

3. What does all this have to do with connecting vision and the world?

What we have described is a mechanism for picking out, tracking and providing cognitive access to what we call an object (or, more precisely, a *primitive visible object*). The notion of an *object* is ubiquitous in cognitive science, not only is vision but much more widely. I might also note that it has been a central focus in developmental psychology where people like Susan Carey and Fei Xu have studied “a child's concept of object” (Xu, 1997), and in clinical neuroscience, where it has been argued that deficits such as unilateral neglect must be understood as a deficit of object-based attention rather than space-based attention. Space does not permit me to go into any of these fields although I am engaged in a larger project where I do examine the connections among these uses of the term “object”. But I would like to draw your attention to the fact that giving objects the sort of central role in vision that I have described suggests a rather different ontology. Just as it is natural to think that we apprehend properties such as color and shape as *properties of objects*, so it is also 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. Everyone from Kant to modern cognitive scientists take this for granted — that's (in part) why it is so natural to think of mental images as having to be embedded in real space in the brain. Yet the findings I have described in the study of visual attention (as well as other areas of psychological research which I cannot describe here, but see, Pylyshyn, 2003) suggests an alternative and rather intriguing possibility. It is the notion that *primitive visible object* is the primary and more primitive category of early (preattentive) perception, so that we perceive objecthood first and determine location the way we might determine color or shape — as a property associated with objects. If this is true then it raises some interesting possibilities concerning the nature of the mechanisms of early vision. In particular it suggests what we argued is independently needed — a mechanism for directly referring to objects in a way that does not rely on having a unique description under which that object falls. This is the function (of “demonstrating”) served by the hypothesized visual index mechanism.

Notice that when I am careful I hedge my use of the term *object* in making this claim, as I must because what I have been describing is not the notion of an object in the usual sense of a physical object or individual. *Object* or *individual* are sortal concepts whose individuation depends on assuming certain conceptual categories. But our notion does not assume any concepts. The individuals that are picked out by the visual system and tracked primitively are something less than full blooded individuals. Yet because they are what our visual system gives us through a brute causal mechanism — because that is its nature — it serves as the basis for all real individuation. As philosophers like (Wiggins, 1979) and (Hirsch, 1982) have argued, you cannot individuate objects in the full blooded sense without a conceptual apparatus — without sortal concepts. But similarly 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 faired well and has been abandoned in cognitive science. However the principle of grounding concepts in perception remains an essential operation if we are not to succumb to an infinite regress. Visual indexes provide a putative grounding for basic objects and we should be grateful because without them (or at any rate something like them) we would be lost in thought without any grounding in causal connections with the real-world objects of our thoughts. With indexes we can think about things (I am sometimes tempted to call them *FINGs* since they are interdefined with *FINSTs*) without having any concepts of them: One might say that we can have *demonstrative thoughts*. And nobody ought to be surprised by this since we know that we can do this: I can think of this here thing without *any description* under which it falls. And, perhaps even more important, because I can do that I can reach for it.

If this analysis is correct – if people do select visual objects before they represent their properties – then treating demonstrative terms as consisting of bare demonstrative (plus additional properties based on the rest of the complex), rather than complex demonstratives that pick out objects-with-specified-properties – makes sense. It makes sense for all the reasons that (Lepore & Ludwig, 2000) have given together with the empirically-motivated grounds suggested here – namely, that attentive selection (or FINST selection) at its initial and most primitive nonconceptual stage picks out visual objects before it encodes their properties. The property encoding places *conceptual* logical forms into the Object Files which were created empty after a new object came into view or was noticed.

Well I have probably waded deep enough into philosophy for the modest purposes of this essay. Needless to say there are some details to be worked out as this is a work-in-progress. But I hope I have at least made the point that there is a real problem to be solved in connecting visual representations to the world that is different in principle from the representations of sentences referred to as Logical Form. Whatever the eventual solution to the problem of visual representation turns out to be, it will have to respect a collection of facts some of which I have 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.

References

Allen, R., McGeorge, P., Pearson, D., & Milne, A. B. (2004). Attention and expertise in multiple target tracking. *Appl. Cognit. Psychol., 18*, 337-347.

Almog, J., Perry, J., & Wettstein, H. (Eds.). (1989). *Themes from Kaplan*. New York, NY: Oxford University Press.

Alvarez, G. A., Arsenio, H. C., Horowitz, T. S., & Wolfe, J. M. (2005). Are mutielement visual tracking and visual search mutually exclusive? *Journal of Experimental Psychology: Human Perception and Performance, 31*(4), 643-667.

Alvarez, G. A., & Cavanagh, P. (2005). Independent attention resources for the left and right visual hemifields. *Psychological Science, 16*(8), 637-643.

Alvarez, G. A., & Scholl, B. J. (2005). How Does Attention Select and Track Spatially Extended Objects? New Effects of Attentional Concentration and Amplification. *Journal of Experimental Psychology: General, 134*(4), 461-476.

Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition, 10*(8), 949-963.

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

Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. *Psychological Review, 94,* 115-148.

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

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

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

Cavanagh, P. (1992). Attention-based motion perception. *Science, 257*, 1563-1565.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9*(7), 349-354.

Cavanagh, P., Labianca, A. T., & Thornton, I. M. (2001). Attention-based visual routines: Sprites. *Cognition, 80*(1-2), 47-60.

Chastain, G. (1995). Location coding with letters versus unfamiliar, familiar and labeled letter-like forms. *Canadian Journal of Experimental Psychology, 49*(1), 95-112.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

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. *J Neurophysiology, 80*(5), 2657-2670.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*(4), 501-517.

Fougnie, D., & Marois, R. (2006). Distinct Capacity Limits for Attention and Working Memory. Evidence From Attentive Tracking and Visual Working Memory Paradigms. *Psychological Science, 17*(6), 526-534.

Goodale, M., & Milner, D. (2004). *Sight Unseen*. New York, NY: Oxford University Press.

Hanson, N. R. (1958). *Patterns of Discovery*. Cambridge: Cambridge University Press.

Hirsch, E. (1982). *The Concept of Identity*. Oxford, UK: Oxford.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of attention. *Cognitive Psychology, 4*(3), 171-216.

Jovicich, J., Peters, R., Koch, C., Braun, J., Chang, L., & Ernst, T. (2001). Brain areas specific for attentional load in a motion-tracking task. *Journal of Cognitive Neuroscience, 13*, 1048-1058.

Kanizsa, G. (1979). *Organization in vision: Essays on Gestalt perception*. New York: Praeger.

Keane, B. P., & Pylyshyn, Z. W. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognitive Psychology, 52*(4), 346-368.

Kosslyn, S. M. (1994). *Image and Brain: The resolution of the imagery debate*. Cambridge. MA: MIT Press.

Lepore, E., & Ludwig, K. (2000). The semantics and pragmatics of complex demonstratives. *Mind, 109*, 199-240.

Lindberg, D. C. (1976). *Theories of vision from al-Kindi to Kepler*. Chicago: University of Chicago Press.

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

Ogawa, H., & Yagi, A. (2002). The effect of information of untracked objects on multiple object tracking. *Japanese Journal of Psychonomic Science, 22*(1), 49-50.

O'Hearn, K., Landau, B., & Hoffman, J. E. (2005). Multiple Object Tracking in People with Williams Syndrome and in Normally Developing Children. *Psychological Science, 16*(11), 905-912.

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

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. (1994). Some primitive mechanisms of spatial attention. *Cognition, 50*, 363-384.

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. (2001a). Connecting vision and the world: Tracking the missing link. In J. Branquinho (Ed.), *The Foundations of Cognitive Science* (pp. 183-195). Oxford, UK: Clarendon Press.

Pylyshyn, Z. W. (2001b). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*(1/2), 127-158.

Pylyshyn, Z. W. (2002). Mental Imagery: In search of a theory. *Behavioral and Brain Sciences, 25*(2), 157-237.

Pylyshyn, Z. W. (2003). *Seeing and visualizing: It's not what you think*. Cambridge, MA: MIT Press/Bradford Books.

Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking (MOT): I. Tracking without keeping track of object identities. *Visual Cognition, 11*(7), 801-822.

Pylyshyn, Z. W. (forthcoming). *Things and Places: How the mind connects with the world (Jean Nicod Lectures Series)*. Cambridge, MA: MIT Press.

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

Rock, I., & Gutman, D. (1981). The effect of inattention on form perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 275-285.

Saiki, J. (2003). Feature binding in object-file representations of multiple moving items. *Journal of Vision, 3*(1), 6-21.

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 target-merging in multiple-object tracking. *Cognition, 80*, 159-177.

Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (1999). When are featural and spatiotemporal properties encoded as a result of attentional allocation? *Investigative Ophthalmology & Visual Science, 40*(4), 4195.

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

Slemmer, J. A., & Johson, S. P. (2002). *Object tracking in ecologially valid occulsion events.* Paper presented at the Vision Sciences 2002, Sarasota, FL.

Suganuma, M., & Yokosawa, K. (2002). *Is multiple object tracking affected by three-dimensional rigidity?* Paper presented at the Vision Sciences Society, Sarasota, FL.

Tipper, S. P., & Behrmann, M. (1996). Object-centered not scene-based visual neglect. *Journal of Experimental Psychology: Human Perception and Performance, 22*(5), 1261-1278.

Treisman, A. (1986). Properties, parts, and objects. In K. Boff & L. Kaufmann & J. Thomas (Eds.), *Handbook of Perception and Human Performance*. New York: Wiley.

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

Treisman, A., & Sato, S. (1988). Conjunction search revisited.

Trick, L. M., Perl, T., & Sethi, N. (2005). Age-Related Differences in Multiple-Object Tracking. *Journals of Gerontology: Series B: Psychological Sciences & Social Sciences, 2,* 102.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In J. Ingle & M. A. Goodale & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549-586). Cambridge, MA: MIT Press.

vanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science, 14*(4), 498-504.

Viswanathan, L., & Mingolla, E. (1998). Attention in depth: disparity and occlusion cues facilitate multi-element visual tracking (Abstract). *Investigative Ophthalmology and Visual Science, 39*(4), 634.

Viswanathan, L., & Mingolla, E. (2002). Dynamics of attention in depth: Evidence from multi-element tracking. *Perception, 31*(12), 1415-1437.

Wiggins, D. (1979). *Sameness and Substance*. London: UK: Blackwell.

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

NOTES

[1] Consider the following two sentences. Which one (if any) is grammatical? “I am having trouble deciding between/among P” where P is some numerical predicate. The choice, according to some grammars, depends on whether P yields exactly two alternatives. But that is not decidable in general. Does that mean that grammar contains undecidable rules? Clearly not: What is shows is that a rule one believed to be a rule of grammar turns out not to be part of grammar at all, something that intuition is powerless to decide.

[2] This obvious point took hundreds of years to appreciate. Only after Kepler’s seminal analysis of how an image can be focused by a lens was the role of the retinal image appreciated (Lindberg, 1976).

[3] Strictly speaking the definite description that uniquely picks out a certain object at a particular time is a quantified expression of the form: $\exists x P(x)$, where P is the unique property of the object in question. When an additional predicate Q that pertains to the same object is to be added, the unique descriptor is retrieved and the new stored expression added: $(\exists x \exists y \{P(x) \wedge Q(y) \wedge x=y\}$. If a further property R of the same object is detected at some later time, the last expression must be matched to the object at which R is discovered and its descriptor updated to the expression $\exists x \exists y \exists z \{P(x) \wedge Q(y) \wedge R(z) \wedge x=y \wedge y=z\}$. This continual updating of descriptors capable of uniquely picking out objects is clearly not a plausible mechanism for incrementally adding to a visual representation. It demands increasingly large storage and retrieval based on pattern matching.

[4] Notice that the need for demonstratives remains even if the representation were picture-like instead symbolic, so long as it was not an exact and complete copy of the world but was built up incrementally. If the picture depicts some state of affairs in the world we still have the problem of deciding when two pictorial bits are supposed to depict the same object. We still need to decide when two picture-fragments are supposed to depict the same object (even though they may look different) and when they are supposed to depict different objects. This is the same problem we faced in the case of symbolic representations. We don't know whether the thing in the picture that is depicted as having the property P is the thing to which we must now add the depiction of the newly-noticed fact that it also has property Q. Without a solution to that puzzle we don't know to which part of the picture to add newly noticed properties.

[5] There is another alternative for picking out objects that I will not discuss here because the evidence I will cite suggests that it is not the correct option for visual representations. This alternative that assumes the existence of demonstratives, as we have done, except the demonstratives in question are *place demonstratives* or *locatives,* such as “this place”. Such an apparatus would allow the unique picking out of objects based on their location and would overcome the problem with the pure descriptivist story that we have been describing. That alternative is compatible with the view presented here although, as we will argue, the idea that object individuation is mediated by location alone (or location alone) does not seem to be supported by the empirical data..

[6] As with a number of terms used in the context of early vision (such as the term “object”), the notion of *individuating* has a narrower meaning here than in the more general context where it refers not only to separating a part of the visual world from the rest of the clutter (which is what we mean by individuate here), but also providing identity criteria for recognition instances of that individual. As is the case with *objecthood* and other such notions, we are here referring primarily to primitive cases - i.e. ones provided directly by mechanisms in the early vision system (in the sense of Pylyshyn, in press) and not constructed from other perceptual functions.

[7] This claim is contentious. There have been a number of studies (reviewed in Pashler, 1998) showing that in those cases where an object is correctly identified, its location generally can be correctly reported. However, what these studies actually show is that for objects whose shapes (or in some cases color) can be correctly reported, their location can usually also be reported. From our perspective this only shows that there is a precedence ranking among the various properties of an object that are recorded and reported and that rough location may be higher on the ranking than other properties. What the experiments do not show (contrary to some claims) is that *in order to detect the presence of an object one must first detect its location*. The studies described below (dealing with multiple Indexing) suggest ways to decide whether an object has been detected in the relevant sense (i.e., individuated and indexed, though not necessarily recognized). The theoretical position to be developed here entails that one can *index* an object without encoding its location. There are, so far as I know, no data one way or another regarding this prediction.

[8] More recent studies have shown that the what-where dichotomy is not quite the right way to distinguish the two visual systems. Rather it appears that one of the systems (the ventral system)

specializes in recognition while the other (the dorsal system) specializes in visual-motor control (Goodale & Milner, 2004).

[9] Location-based attention is not ruled out by these studies. It still remains possible that a "spotlight of attention" can be scanned across a display in search of objects of interest. However, these studies do show that at least *some* forms of attention are directed to whole objects irrespective of their location in space.

[10] Some published research includes (Allen, McGeorge, Pearson, & Milne, 2004; Alvarez, Arsenio, Horowitz, & Wolfe, 2005; Alvarez & Cavanagh, 2005; Alvarez & Scholl, 2005; Bahrami, 2003; Blaser, Pylyshyn, & Domini, 1999; Blaser, Pylyshyn, & Holcombe, 2000; Cavanagh, 1992; Cavanagh & Alvarez, 2005; Cavanagh, Labianca, & Thornton, 2001; Culham, Brandt, Cavanagh, Kanwisher, Dale et al., 1998; Fougnie & Marois, 2006; Intriligator & Cavanagh, 2001; Jovicich, Peters, Koch, Braun, Chang et al., 2001; Keane & Pylyshyn, 2006; Ogawa & Yagi, 2002; O'Hearn, Landau, & Hoffman, 2005; Pylyshyn, 1989, 1994, 2004; Pylyshyn et al., 1994; Pylyshyn & Storm, 1988; Saiki, 2003; Scholl & Pylyshyn, 1999; Scholl, Pylyshyn, & Feldman, 2001; Scholl, Pylyshyn, & Franconeri, 1999; Sears & Pylyshyn, 2000; Slemmer & Johson, 2002; Suganuma & Yokosawa, 2002; Trick, Perl, & Sethi, 2005; vanMarle & Scholl, 2003; Viswanathan & Mingolla, 1998, 2002; Yantis, 1992).

[11] I am claiming that there is a mechanism in early (pre-conceptual) vision that latches onto certain entities for purely causal reasons, not because those entities meet conditions provided by a cognitive predicate - i.e., not because they constitute instances of a certain concept. In other words if P(*x*) is a primitive visual predicate of *x* then the *x* is assumed to have been independently and causally bound to what I have called a primitive visible object. Although this sort of latching or seizing by primitive visible objects is essentially a bottom-up process, this is not to say that it could not in some cases be guided by intentional processes, such as perhaps scanning one's attention until a latching event is located or an object meeting a certain description is found. For example, it is widely assumed (Posner, Snyder, & Davidson, 1980) that people can scan their 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.