Visual Indexes and Nonconceptual Reference[1]

Zenon Pylyshyn, Rutgers Center for Cognitive Science
(Notes for NYU's Language and Mind Seminar, March 2, 2004)

First a little historical background

Just as Moliere's Monsieur Jourdain discovered that he had been speaking prose all his life without realizing it, so I discovered not too long ago that what I had been doing without realizing it occupied a position in the philosophical landscape. It turns out that coming from a very different perspective I had taken a position on a set of questions that philosophers had been worrying about for much of the past 30 or more years. My clandestine involvement in philosophical issues began when a computer science colleague and I were trying to build a model of geometrical reasoning that would draw a diagram and notice things in the diagram as it drew it [1]. One problem we found we had to face was that if the system discovered a right angle it had no way to tell whether this was the intersection of certain lines it had drawn earlier, and if so *which particular* lines. Moreover, the model had no way of telling whether this particular right angle was identical to some bit of drawing it had earlier encountered and represented as, say, the base of a particular triangle. There was, in other words, no way to determined the identity of an element (I will use this neutral term until I discuss what qualifies as an element) at two different times if it was represented differently at those times. This led to some speculation about the need for what we called a "finger" that could be placed at a particular element of interest and that could be used to identify it as particular token thing (the way you might identify a particular feature on paper by labeling it). In general we needed something like a finger that would stay attached to a particular element and could be used to maintain a correspondence between the individual element that was just noticed now and one that had been represented in some fashion at an earlier time. The idea of such fingers (which came to be called "Fingers of INSTatiation" or FINSTs) then suggested some empirical studies to see if humans had anything like this capability. Thus began a series of experimental investigations of FINSTs that occupied me and my students for much of the past 20 years.

During this period I found myself thinking about why vision needed the sort of link provided by FINSTs to connect cognitive representations and the sensible world. My initial interest in FINSTs was a response to the fact that diagrams did not come into existence all of a sudden, but were constructed over time. It soon became clear that it does not matter how the figure came into existence, since the representation of the figure is itself built up over time. We clearly don't notice all there is to notice about a scene in an instant – we notice different things over a period of time as we move our eyes and our focal attention around. Consequently we may notice and represent the very same token element differently at different times. There is plenty of evidence that even without eye movements or movements of the "spotlight of attention" we construct perceptual representations incrementally over time [2-19], so we cannot escape the need to keep track of individual elements (or whatever we eventually find serve as the elements that we keep track of) *qua individuals* over time.

Around the same time as we undertook these experiments [initially reported in 20,21] another set of experiments were independently published by Daniel Kahneman [22], who introduced the concept of an *Object File*. An Object file is the *conceptual* representation of elements with which they are associated. Although this was not stressed in the Kahneman et al. report, object files are connected to individual things and keep accumulating information about the individuals as it tracks them, but it does not use that information in order to determine which individual it is associated with (i.e., it does not use the contents of the object file keep track of the individuals). In other words, object files are in fact attached to objects by what we were calling FINSTs.

As with many ideas, it took a long time to appreciate that the basic idea was actually a proposal that introduced nonconceptual representation. Somewhere around this time it began to strike me that FINSTs had to be a very special sort of connection, different from what psychologists had been studying under the term

[1] This is a rough piece. I have merged material from [31] with some recent thoughts that may place it into philosophical context. In the process I may have made a mess of the organization. If you find it so, you are welcome to read instead, the original reference which you can get from http://ruccs.rutgers.edu/ftp/pub/papers/cognition2001-reprint.pdf.

"attention" and different from the semantic connection of *satisfaction* with which philosophers have had a long-standing love-hate relationship. They differ from what psychologists call *focal attention* in that (1) there needs to be several of them – the task I sketched briefly above can't be done with only one FINST index (I will give examples why later) (2) Indexes are generally assigned by virtue of events taking place in the visual field – i.e. they are data-drive, (3) FINSTs stay connected to the same element (whatever such a thing turns out to be) as the element moves around and changes any or all of its properties, (4) Because of the last point, the indexes' attachment to the appropriate element cannot be mediated by a description of the element in question both because there is no description that uniquely and for all time characterizes a particular token element (unless one uses a historical record of past properties, and then the question still remains why that sequence of properties makes it the same individual) and also because what the index is intended for is to keep track of elements qua individuals, independent of whatever properties they may in fact have. Although these properties largely reflect empirical facts about vision that have since been discovered by experiments, they are inherent in the function which FINSTs were called upon the perform. Moreover the above 4 properties already mark FINSTs as being quite different from the sorts of mind-world (or representation-world) connections that psychologists (and AI people) had postulated in the past, because they both served to refer to the element in question and do so without representing the element as falling under a description: The relation between the representation and the thing represented is not one in which the thing *satisfies* the conceptualization. The far-reaching consequence of the FINST view are hard to overestimated. While this is a topic for later, I can give you a sense of how odd this idea is if I point out that it means, for example, that FINSTs provide a mechanism for referring to things even though, in an important sense, the referrer does not know what he is referring to just because the reference does not carry a conceptualization with it! To refer to something (say that object in the corner) without referring to it as a cat, or as some mass shaped thus-and-so, or as a patch of tawny-color, or (as Quine might put it) as a collection of undetached cat parts, is a strange notion indeed. Yet I there must be a stage in the visual process where something like this happens, otherwise we could not construct our conceptual representations on scaffold of causal connections to the world, as we must.

On the issue of whether it makes sense to postulate a nonconceptual form of reference I have in my general corner, certain AI people who speak of embodied or situated cognition, philosophers who speak of essential indexicals [23], and logicians who argue for bare demonstratives [24], which are closely related to FINSTs. In the opposite corner are those who wish to inoculate their picture against skeptical arguments by assuming that the most primitive reference must be accessible to conscious experience [John Campbell speaks of the mechanism as "conscious attention" 25], or who escape the conundrum of not knowing what one is referring to by assuming that the most basic form of reference must pick out locations (or at least space-time regions) [26,27] or, at any rate, that they are not non-conceptual [28]. At this point I simply want to alert you to the fact that much philosophical baggage hangs on how we describe what goes on in the earliest stages of visual perception (where by earliest I mean logically, neurologically and temporally early, though not necessarily ontogenetically early).

Why do we need nonconceptual reference?

The most general view of what vision does is that it computes a representation of a scene which 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 there may perhaps be a somewhat different version of this representation that may become available for the immediate control of motor actions). This form of representationa represent a situation "under a description", that is, it represents the elements of the situation as members of some category or as falling under a certain concept. This is a fundamental characteristic of cognitive or intentional theories which distinguishes them from physical theories [29]. We need this sort of representation because what determines our behavior is not the physical properties of the stimuli around us, but how we interpret or classify them — or more generally *what we take them to be*. It is not the bright spots we see in the sky that determine which way we set out when we are lost, but the fact that we see them (or represent them) in a certain way or under a certain concept (e.g., as the pointer stars in the big dipper or as the North Star). It is because we represent them *as* members of a certain category that our perception is brought into contact with our knowledge of such things as astronomy and navigation. Moreover, what we represent need not even exist, as in the case of the holy grail, in order to determine our behavior. In other words, it is the fact that we conceptualize our

perceived or imagined world that allows us to think about it. 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 can, under certain conditions, also refer to or represent some things without representing them in terms of concepts. We can refer to some things, as we say, *preconceptually* (I notice that the preferred term in philosophy appears to be *non*conceptually). 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 has. A term such as *this* or *that* is called a "demonstrative". Philosophers like Ernie Lepore [23] have argued that demonstratives are ineliminable in language and thought. The reason for the ineliminability of demonstratives also apply in the case of 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 nonconceptual way. If we could only refer to things in terms of their category membership, our concepts would always be rooted only in other concepts 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. What I propose is 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 a particular token individual. 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[2]. 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. The need to pick out, or refer to, individual things (let's start out by using the informal term "things" since characterizing what we pick out is one of the tasks of this paper) is not a something that arises under exotic circumstances – of the sort dreamt up by philosophers – but arises every time you look out and see the world. It arises for a number of very good reasons and is generally associated with what is referred to in psychology as *focal* or *selective attention*. This is not the place to analyze these reasons, but it may be useful to at least list them since they are not always recognized or appreciated. The need for *selecting* or *picking out token things* arises because of:

1. The limited capacity of the mind to process information. Because information processing is limited, some selection is required. The proper way to characterize the dimension along which the mind is limited and consequently the basis for selection are important empirical questions on which there is now interesting convergent evidence (we will consider the evidence pointing to *objecthood* as the unit of attention or the elements over which attention selects).
2. In encoding or conceptualizing a scene it is necessary to keep track of individual tokens in order to build a consistent representation. This arises in part because a representation must be constructed incrementally over time as parts of the representation that are encoded (or noticed) at different times and must be put into correspondence. This is discussed below.
3. Information about the world is "packaged" or presented in certain ways that lead to what Austen Clark [27] calls the "binding problem" [after Anne Treisman 30] or the "many properties problem", who had

[2] There is one obvious way in which we could refer to a unique individual without using a description that uniquely applies to that individual, and that is by looking directly at it or attending to it. But this way requires making an indexical reference since it involves thoughts such as: *What I am looking at now* (or *this*). Indexicals appear to be a superset of demonstratives since you can use "here" to refer to where you are (indexical use) or-- by pointing -- to a place on a map (a demonstrative use). Both arguably entail a component of nonconceptual reference. I will not be dealing with the ***locative*** type of indexicals that refer to locations since I will be arguing that Visual Indexes refer to things rather than locations (referring to something as a location is already at least partly conceptual).

introduced the term earlier]. Very early in the information-processing stream of visual analysis we must distinguish between properties present in a scene and different conjunctions of these properties present on individual objects or places (i.e. we distinguish between various scenes containing redness, greenness, squareness and roundness; for example, we distinguish between a scene containing a red square and a green circle and a scene containing a red circle and a green square). This occurs at an extremely primitive level in vision (Clark would say it occurs in sensation, but I find that notion opaque – I say it occurs in *early vision* or in the visual module) and the informational basis for this encoding must be present prior to the application of concepts like circle and square and even red and green. It must be evident in the way the perceptual world is primitively parsed – otherwise that conceptualization would not get a start.

4. Patterns that are visually discriminable cannot be detected or represented unless we can designated which things partake in that pattern. **Collinear**(x,y,z), or **Inside**(x,y) or **Above**(x,y) or even **Location**(p, axes[x,y,z]) cannot be evaluated unless some tokens in a scene are bound to the arguments x, y and z.
5. Many visual patterns can only be discriminated if a serial process operates over the elements, which requires that token elements be somehow "marked" as they are referred to by what Ullman calls "visual routines" that carried out over the marked tokens. Predicates such as "containing n items" or "is inside a closed contour" or "are on the same contour" all require the operation of a serial process over the scene and this process requires that certain things in the scene be picked out and referenced (most psychologists refer to this picking out as "marking" or "tagging" but that is very misleading way of talking since nothing is done to the distal scene nor to a representation of it – we simply pick out and assign a demonstrative reference to the things picked out).

In these notes I focus on the problem of establishing a correspondence between individual things in the world and their counterparts in the visual representation, 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, 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 than a description. And that, I suggest, is why we must have something like a visual indexing mechanism which *nonconceptually* 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 need for individuating and indexing: Empirical motivations [for more details see 31]

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 token individual* among many similar ones in the visual field (especially when its location or properties are changing), a description will not do. 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 or happens to occupy a particular location in space. 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. I appreciate that being a particular individual encumbers the individuation process with the need for conditions of individuation and real full-blooded individuals must meet this condition and therefore must be *conceptualized as* that individual. But the visual system in its ignorance appears to solve a subset or a scaled-down version of the individuation problem that is sufficient for its purposes and which more often than not does correspond to real individuals (or real objects). That is the beauty and the ugliness of the visual module – it does things expeditiously that turn out to be the right things to do in this sort of world, a world populated mostly by objects that move in certain rigid ways, in which discontinuities in lightness and in depth have arbitrarily low probability because real scene edges occupy a vanishingly small part of the universe, in which accidental alignments have vanishingly small probability of occurring, in which the light tends to come from above and casts shadows downward, and so on. It is a blissfully ignorant but superlatively successful design for our sort of world. Vision is attuned to just the right properties which it picks out without benefit of knowledge and expectations of what is likely to be in some particular scene at some particular time.

So I claim that a very important and neglected aspect of vision is the nonconceptual connection by which it picks out what I have been calling objects (but which it does not conceptualize as objects because it knows nothing of objecthood). In arguing for the insufficiency of conceptual (or descriptive) representations as the sole form of visual representation, I appeal to three empirical assumptions about early vision: The assumption that individuation of objects precedes detecting and encoding their properties, the assumption that the detection of objects is primitive and nonconceptual or preconceptual (i.e., does not itself involve the appeal to any encoded or perceived properties), the assumption that the detection of properties proceeds by the prior detection of objects that bear those properties, and the assumption that visual representations are built up incrementally.

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

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 spatio-temporal 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 Figure 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 [32]. Computing global patterns such as collinearity, or others discussed by [33], 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.

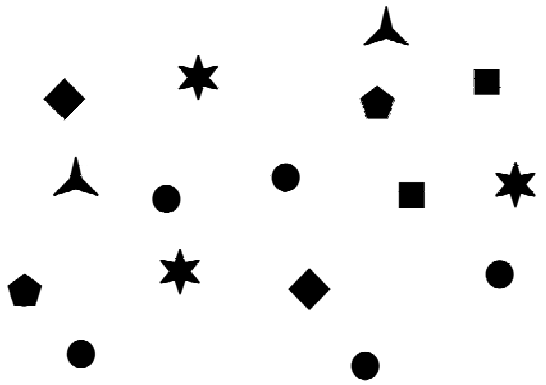

Figure 1 Find 4 or more items that are collinear. Judging collinearity requires both selecting the relevant individual objects and ignoring all their local properties.

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 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. Shimon Ullman [33], as well as a large number of other investigators [34-36] 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. James Intriligator's dissertation [described in 37] showed 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 ([38] to measure attentional resolution. Figure 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 [37] show that the process of individuating objects is separate and distinct from that of recognizing or encoding the properties of the objects.

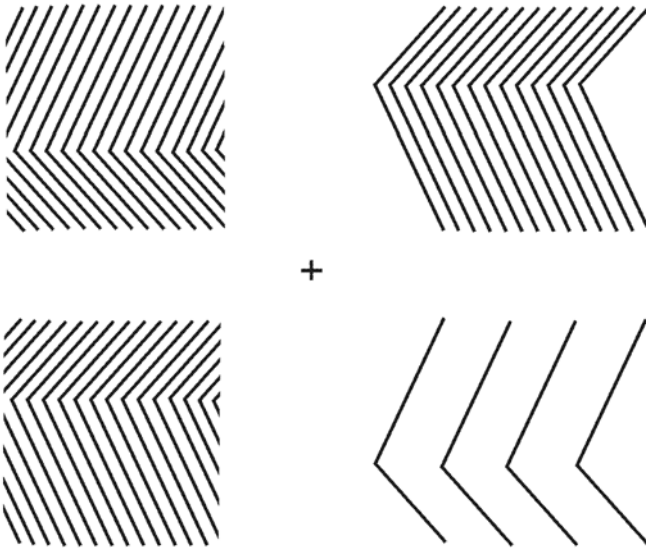

Figure 2 At a certain distance if you fixate on the cross you can easily tell which groups consist of similar-shaped lines, although you can only **individuate** lines in the bottom right group. For example, you cannot count the lines or pick out the third line from the left, etc., in the other three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

Studies of rapid enumeration (called *subitizing*) described in the work that Lana Trick and I did [39] 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 Figure 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.

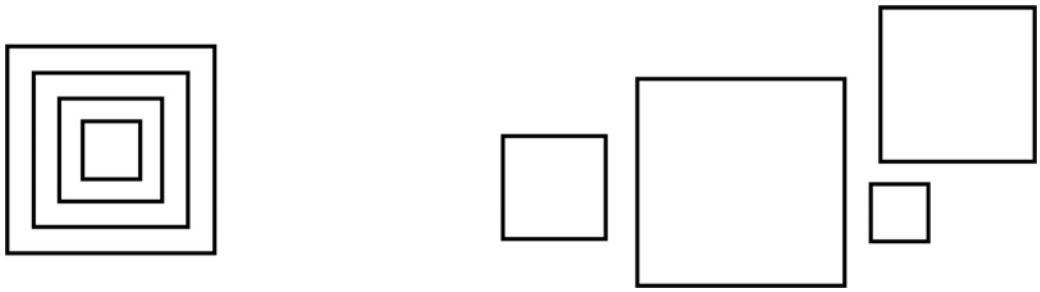

Figure 3. Squares arranged so they cannot be preattentively individuated (on the left) cannot be subitized, whereas the ones on the right are easily subitized [based on 39].

2. Detection of visual properties consists in the detection of properties-of-objects, as opposed to the detection of properties *tout court* or properties-at-locations (Clarks' 'Feature Placing')

When a property is encoded by the visual system it is encoded not just as a property existing in the visual field, but as the property of an individual perceived thing-in-the-world. The claim has frequently been made that features are detected as occurring at a location (talk of "feature placing" explicitly assumes that this is what happens, at least within the level of sensation). I claim that the visual system does not just detect the presence of redness or circularity in the visual field, or the presence of such properties at some particular

location in some frame of reference: It detects that certain individual *objects*[3] 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. Some of them are sketched next.

(a) Object-Based Attention and single-object advantage. The *first kind of evidence* comes from the observation 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 [40]. So for example, if you are asked to judge the relative heights of the two vertices in the figure below, you are faster when instructed to view the lighter portion as the object in (a) compared to (b).

Figure 4 Figures used to demonstrate single-object advantage in judging properties of a shape within one figure vs between two figures.

Other evidence supporting this conclusion comes from a variety of sources [many of which are reviewed in 41], including experiments in which objects move through space or in which they move through feature space. [More examples are discussed in my book, just published last month 42]. Also, clinical cases of hemispatial visual neglect and Balint Syndrome, implicate an object-centered frame of reference. Patients with the symptom known as simultanagnosia, who reportedly can only see one object at a time, nonetheless can report properties of two objects if they are somehow linked together. This sort of object-specificity of feature encoding is exactly what would be expected if properties are always detected as belonging to an object. Object-based attention has been widely studied in current vision science and most of the more impressive evidence comes from cases where objects move so it is possible to distinguish between objecthood and location.

(b) The Binding Problem and detecting conjunctions. Another kind of evidence foor the primacy of objectood comes from the fact that people can distinguish the occurrence of features in a scene from their co-occurrence in individual objects. This has been called the binding problem or the multiple-properties problem by Austen Clark [27]. The assumption is that in early vision (or, as some people put it, in *sensation*) people can distinguish between different displays that consist of redness, greenness, circularity and squareness. For example they can distinguish between a display consisting of a red circle and a green triangle from one consisting of a green circle and a red triangle. When people detect conjunctions of properties (in, for example, rapid-search experiments),we find that they must use focal attention to pick out the individuals that have both properties. The usual assumption among psychologists about how the binding problem is solved is that it is done in terms of the common *location* of the bound properties. This assumption is made in Treisman's Feature Integration theory [43], in Clark's theory of sentience, in Campbell's analysis of consciousness [26] and in most psychological theories [see, e.g., 44]. But this will not work in general and where it does work, it confounds location and objecthood: Two or more features can be bound either by being in the same location or by being features of the some object.

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 is one carried out by Mary-Jo Nissen [45]. 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

[3] It's time to stop beating around the bush and referring to *things* and *elements* when what I am going to be getting to is that the most primitive reference in vision is to precursors of objects – to things that usually turn out to be objects (though they may sometimes turn out to be only shadows).

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 is 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 Hal Pashler [44, p 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) *Object specific effects move with moving objects*. 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 on "object files" [22] and our own studies using multiple-object tracking (MOT) [see below, as well as 46,47]. 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.

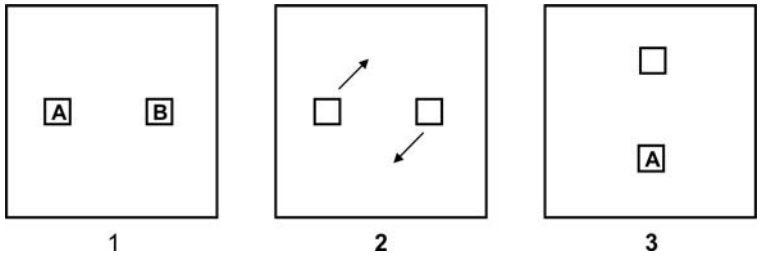

Figure 5: Studies showing facilitation of naming of a letter (the letter could be named faster) when it recurs in the same box as it was at the start of the trial, even though this was not predictive of which letter it was (since half the time it was the letter that had been in the other, equally distant, box).

Similarly, related studies by Steven Tipper [48] showed that the phenomenon known as *inhibition of return* (whereby the latency for switching attention to an object increases if the object had been attended in the past 300 ms to about 900 ms) was specific to particular objects rather than particular locations within the visual field [though later work by 49, suggests that location-specific IOR also occurs].

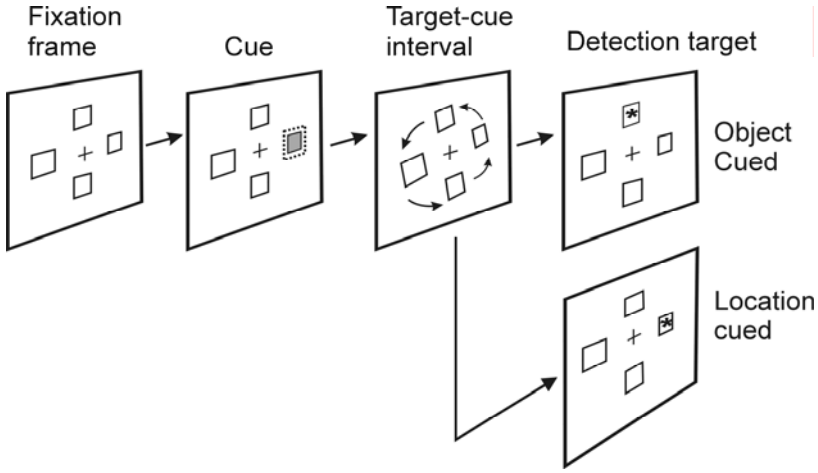

Figure 6 Inhibition of Return (IOR) is a phenomenon whereby items that are attended and then attention is removed from them become **more** difficult to re-attend during a period of from abut 300 ms to 900 ms after. It has been shown that what is inhibited in IOR is mostly the individual object that had been attended – IOR travels with the object as it moves.

While there is evidence that unitary focal attention, sometimes referred to as the "spotlight of attention," may be moved through space and appears to spread away from its central spatial locus, many other attention phenomena appear to be attached to objects with little evidence of spreading to points in between the objects. Egly et al [50] showed that attention seems to spread within closed contours, but also that it only spreads through contoured regions that are perceived to be a single object – it does not stop spreading when it encounters the edge of an occluding surface but continues on through to the rest of the same object on the other side of the occluder.

Visual representations are constructed incrementally

Another empirical finding is that our visual representation of a scene is not arrived at in one step, but rather is built up incrementally. This finding has strong theoretical support as well. A number of theoretical analyses [33,51] 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 computed serially by scanning a beam of attention over certain parts of 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., 52,53,54]. 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 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.

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 property or feature is present in the scene, you need to add this information to the current representation. How do you know which represented item is the relevant one so you can add the information to the appropriate item? 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?

You might think that 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 many of their very obvious properties without their representation being updated [55-59].

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 in adding the predicate Q(α) 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.[4] 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 arises when the objects appear across different views, as when a display is examined by moving the eyes. As mentioned earlier, 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 [1]. 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 had left the high-resolution field of view and then returned again as the eyes moved about. It is for just such reasons that FINST indexes were introduced some 25 years ago.

Some Experimental Evidence for a visual indexing mechanism

Nonconceptual selection

The following experiment by my student Jacquie Burkell [60] illustrates and provides evidence in favor of the assumption that the visual system has a mechanism for picking out and accessing individuals prior to encoding their properties. Burkell showed that sudden-onset location cues (which we assume cause 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 Figure 7) a number of placeholders (11 in the case illustrated), consisting of black X's, appeared on the screen and remained there for one second. Then an additional 3-5 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 through only 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 with the number of items in the display). As expected, the target was detected more rapidly when it was one of the subset 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

[4] 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.

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 nontargets differed from it in one feature and others differed from it in another feature).

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: 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* (whether it constituted a simple-search or a conjunction-search task) determined the relation between number of search items and 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.

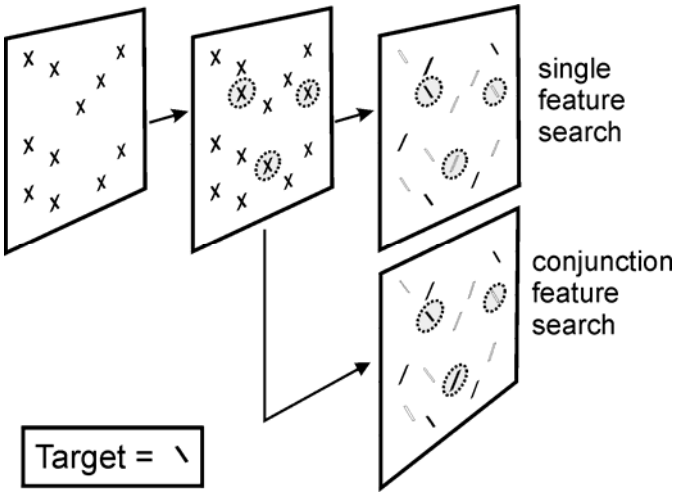

Figure 7: Sequence of events in the Burkell & Pylyshyn [60] study. In the first frame the observer sees a set of placeholder Xs for 1000 ms. In the second frame, "late onset" placeholders appear for 100 ms, signaling the items that will constitute the search subset. In the third frame, all placeholders 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 nontargets 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. More recently we carried out this experiment where an eye movement was interposed between the late onset placeholders and the search display, so that we could ask whether the indexes applied during the cues would withstand a saccade and allow items to be accessed in the search phase even though they were on different places on the retina. We found that they did.

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 nontargets 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*, without having to scan the display searching for the subset items. We have also carried out the above experiment under rather technically difficult conditions in which subjects had to move their eyes in the brief period between getting the late-onset cues and the start of the search process. Although we did not get exactly the same clean results we were able to show that indexes assigned to the cued objects survive an eye movement. This means that after the rapid saccade they were able to pick out the cued objects even though they were now in a different place on the retina. Having such a mechanism provides the beginnings of an account of how the world retains is apparent stability in the course of the 100,000 or so rapid movements each day – it does it by maintaining a cross-saccade correspondence on a few significant objects. Studies have shown that we do not need (and do not have) the ability to recall more than a few items from one fixation to another so this mechanism may be all we need.

This type of study provides a clear picture of the property of indexes that we have been emphasizing: They provide an *direct access mechanism*, rather like the access provided by pointers in a computer or a demonstrative in language. Certain primitive visual objects can be indexed without appealing to their properties (the indexing being due to such transients as 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 primitive visual object by an index.

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 any of 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 each individual object just as if each individual object had a unique label or proper name. Although (as I will argue later) I believe these two functions to be distinct, I have proposed that they are both realized by a primitive mechanism called a *visual index*, some of the details of which will be sketched later. 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 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 Figure 8.

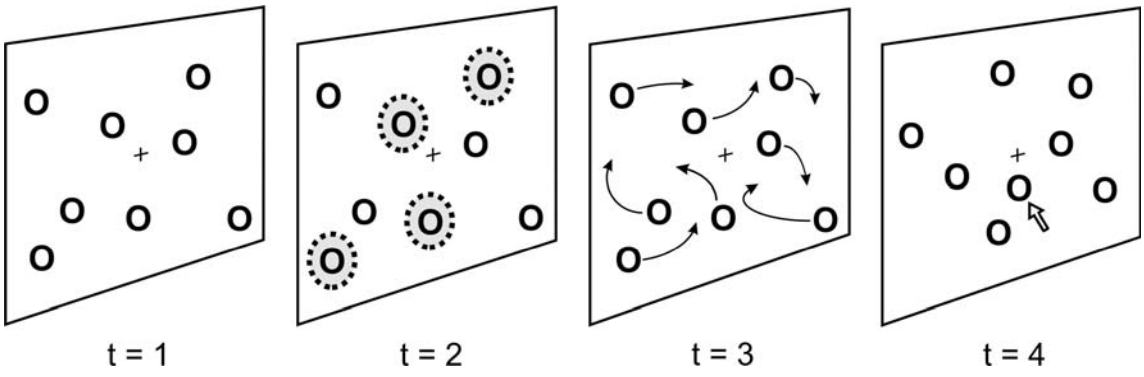

Figure 8. Illustration of a typical Multiple-Object Tracking Experiment. A number (here 8) of identical objects are shown (t=1), then a subset (the "targets") is selected by flashing them (t=2), after which the objects move in unpredictable ways (with or without self-occlusion) for about 10 seconds. 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. (Demonstrations of this experiment can be viewed at: http://ruccs.rutgers.edu/finstlab/demos.htm)

In a typical experiment, observers are shown anywhere from 8 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 seconds 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 [20], have shown that observers can indeed track up to 5 independently moving targets within a field of 10 identical items. The question we must ask is: How can this be done? What mechanism makes this possible? If it were to be done using some description of each object it would have to be a process that encodes each object's location, since location is the only property that distinguishes one object from the other. Such an process would have to use focal attention since a reasonable assumption from previous work on attention is that objects must be attended in order to encode their properties to be encoded. A possible algorithm is this the one shown in Table 1.

Table 1. Serial Scanning algorithm for tracking objects in MOT

1. **While the targets are visually distinct, scan attention to each target in turn and encode its location on a list.**
2. **When targets begin to move, check the *n*'th position in the list and go to the location encoded there: Call it Loc(*n*).**
3. **Find the closest element to Loc(*n*).**
4. **Update the actual location of the element found in #3 in position *n* in the list: this becomes the new value of Loc(*n*).**
5. **Move attention to the location encoded in the next list position, Loc(*n*+1).**
6. **Repeat from #3 until elements stop moving.**
7. **Report elements whose locations are on the list.**

This algorithm visits each object in turn and encodes its location. 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 nonconceptual tracking of a primitive perceptual individuality.[5]

Recently a large number of additional studies in our laboratory [62-67] as well as in other laboratories [38,61,68-70] have replicated these multiple object tracking results using a variety of different methods, confirming that observers can successfully track around 4 or 5 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, 71] 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 [62]. Using a different tracking methodology, [72] also failed to find any evidence of a "spread of attention" to regions between targets [see also 73]. It appears, then, that items can be tracked despite the lack of distinctive properties (and, indeed when their properties are

[5] 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 [**61** Yantis, S. (1992) Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology* 24, 295-340 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 [**62** Sears, C.R. and Pylyshyn, Z.W. (2000) Multiple object tracking and attentional processes. *Canadian Journal of Experimental Psychology* 54 (1), 1-14, 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 [**46** Pylyshyn, Z.W. (1994) Some primitive mechanisms of spatial attention. *Cognition* 50, 363-384, 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.

changing) and despite constantly changing locations and unpredictable motions.[6] Taken together these studies suggest that what [75] referred to as the *early vision system* [an essentially encapsulated system, discussed at length in 76], 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 multiple object tracking task exemplifies what is meant by "tracking" and by "maintaining the identity" of objects. It also operationalizes the notion of "primitive visual object" as whatever allows nonconceptual selection and multiple-object tracking.[7] 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 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 [67]. Indeed, as we saw in see section 2, some well-defined visually-resolvable features do not allow individuation (see Figure 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, [66] 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 [80] 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 [81] 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 [64] shows 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).

More findings concerning multiple object tracking

We have been studying this Multiple Object Tracking paradigm for many years now and have shown a number of surprising results. A sampling of some of these results is presented here because I will may occasion to refer to them when I return to consider the nature of the mechanism that makes it possible to track these elementary objects.

1. *The number of objects that can be tracked is limited to 4 or 5*. We have shown that people can track up to 5 target objects with a high degree of accuracy for periods of at least 10 seconds (4 objects is a more common limit). Even children 5 years of age manage to track 2 or 3 objects. The ability to track increases slightly when the objects are prevented from coming too close to one another and when they move very slowly – though in the latter case it seems that what is involved is a cognitive strategy or grouping and memorizing the location of the group and then tracking another group, etc. What is special about MOT is that it somehow discourages people from playing such games, and the natural way to explain that is to say that tacking demands a great deal of attention – which is consonant with the phenomenology of doing the task. But this assumes a single pool of attentional resource – an

[6] In a set of yet-unpublished studies [**74]** we showed 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. Bahrami [**69]** also showed that observers cannot detect changes in color or shape on either nontargets or targets while tracking.

[7] The concept of a "proto-object" is a general one that has been used by a number of writers [77-78], and sometimes using some other term, such as "preattentive object [**79]** 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" (inasmuch as it is treated as the same enduring individual when it changes location) and less than a recognized 3D distal object. Beyond that, the exact nature of a proto-objects depends on the theory in question.

assumption that we are beginning to question and in the process finding reasons to believe that the actual keeping-track-of-individuality aspect of MOT does not draw on a general attentional resource. This needs some empirical unpacking and I am willing to be drawn in to a discussion of this issue if the need arises!

2. *Objects can be tracked extremely well even under conditions where subjects are unlikely to have kept a record of objects' locations and updated these during the tracking.* We simulated a serial updating algorithm based on the assumptions that (a) encoding the location of objects requires focal attention, (b) focal attention is unitary and must be moved to the object whose location is encoded, and (c) measures of the speed of movement of focal attention provide a reasonable estimate of the outside limit on the speed of attention movement. This showed that with the displays used in [20], performance would not have exceeded about 30% whereas human tracking performance was consistently over 85%.
3. *There is no sign of attention enhancement between targets.* If what is attached to targets is a form of attention then it is punctate and does not spread from targets. Simple shape discrimination of a new item is faster on a target than on a nontarget, but there is no advantage when the new item occurs in between targets. If targets receive attention then the attention does not spread to points between targets [72,82].
4. *Objects can be tracked even if they disappear briefly and completely from view, providing the disappearance occurs in a manner that is compatible with objects going behind an occluding surface*. If objects disappear by shrinking to a point and then growing out from a point, they are poorly tracked. Similarly if targets disappear at an edge other than a fixed leading edge (as they would if they were moving behind an occluder) tracking is poor. (It also happens if all objects suddenly disappear from the display and remain invisible for up to half a second – I may have more to say about this later.)
5. *Under certain conditions objects can be indexed voluntarily.* We have always assumed that indexes were "grabbed" by certain events in the world, since they are our first line of causal influence by the perceptual world on our representation. But we showed that given enough time objects can be indexed voluntarily – so index assignment can be conceptually based. However this requires extra time and it requires more extra time when there are more targets. What this suggests is that voluntary indexing proceeds by moving focal attention to each target in turn, while it is visibly distinct, allowing an index to be "dropped off" along the way. This, in turn, suggests that the role played by attention (and conceptually-driven search) in this case is to facilitate the automatic grabbing of indexes by picking out regions in which the targets are locally distinct. This is in the same spirit as I have elsewhere argued that attention modulates visual processing, even though early vision is modular and informationally encapsulated, by allow automatic modular visual processes to operate at selected objects or locations [76].
6. *Tracking performance is the same whether or not the objects are pairwise distinct in appearance*. Of course you can't just make the objects different since they could then be "tracked" simply by recalling what color or shape the targets were. So we developed a different technique. We arranged that no two objects were ever the same color, but the colors were changed continuously and gradually throughout the trial by moving through the color circle (no special care was taken to ensure that the color sequence was isoluminant, so this result also suggests that changes in luminance neither enhance nor detract from tracking). Performance on this version was compared with the case where objects changed color in exactly the same manner but all changes occurred in synchrony so that at any instant all objects were the same color. We also compared this with the standard case where objects did not change color and we examined in for easy and difficult tracking tasks (i.e., one in which objects were kept apart vs allowed to overlap) and for long and short trial durations. The latter two manipulations changed performance in the expected way (long trials led to more errors and repulsion made the tracking easier) but the affected all three color-change conditions the same.
7. *Recall of object properties is very poor or even nonexistent.* We have shown [74] that subjects notice very little about the objects they are tracking. Another study from Iran [69] showed that people are poor at detecting changes in object colors if these occurred while the object is out of view (e.g., behind an occluding surface).
8. *Not all well-defined clusters of properties can be tracked*. We showed that clusters of features that are perceived as well-integrated parts of larger objects cannot be tracked as individuals [67]. For example, if a good-Gestalt object is formed by connecting a target with a nontarget (in a procedure we called "target-merging"), then such targets cannot be readily tracked. In the extreme, when objects are points and they are connected in target-nontarget pairs by a line (so that observers had to track the end of a line) they could not do it, even though they could easily track the unconnected points moving along the identical trajectories.
9. *When targets disappear briefly the location of their reappearance is not anticipated – so tracking does not appear to use prediction of future location.* Under a variety of conditions Brian Keane [83] found that if after a brief disappearance, objects reappear where they would have been had they kept moving during the disappearance interval, tracking performance is poor and much worse than if the objects reappear close to where they disappeared (which would happen only if they stopped moving the instant they disappeared). This result holds for several different modes of disappearance (suddenly or progressively going behind an occluder).

10. *Recall of the identity of targets is poor*. We did a number of studies in which subjects not only had to track the targets, but also had to report identifying labels that had been assigned to each of them at the beginning of the trial (during the identifying period when the targets were flashed). The labels were either their initial location (one of the four corners) or the numbers (1,2,3,4) that appeared inside the target circles at the start of the trial. In a variety of different experiments we consistently found that the identification label (ID) was retained much more poorly than the objects' status as a target or nontarget. This is extremely puzzling in view of the fact that in order to track a particular target it was necessary to track its identity from the time it was visibly distinct at the start of the trial.[8] The only other memory load imposed by the ID task was recalling the 4 paired associates involving the internal reference and the external label – and we showed that this was trivially easy even under conditions of simultaneous tracking and even when we looked only at trials in which all targets were correctly tracked or when targets were visually distinguished from nontargets. We showed that this is because target-target pairs are more readily confused when they come close together than are target-nontarget pairs.
11. *Moving nontargets are inhibited*. In answering the question why targets are more often confused with other targets than with nontargets, one possible hypothesis is that the nontargets are in some sense taken out of the running for being confused with targets. There is some precedent for the view that nontargets might be inhibited from receiving attention since such a phenomenon has been demonstrated in several different contexts in studies of attention (e.g., in the case of what is called "inhibition of return" (IOR), wherein objects that had been attended in the previous 300 to 900 ms are not as readily attended a second time [see the review in 84]. It has also been shown that IOR occurs on up to 5 objects at once [85] and moves with the inhibited objects when they move [48]. Exactly what this means is not so clear, especially in the light of our surprising finding that each individual nontarget, *and not the space between objects*, was inhibited. Inhibition was measured using a probe dot technique introduced for just such purposes by [35]. In the course of a tracking trial, subjects had to note whether a small dot appeared anywhere in the display. After correcting for differences in visibility at different locations (by running a probe-detection control condition with exactly the same display but without requiring subjects to track) we showed that targets and the empty space between objects had about the same level of probe detection but that detection on nontargets was lower. This indicates that nontargets are inhibited as they move around (rather than targets being enhanced). But that is rather surprising because it means that inhibition tracks the nontargets. Yet by hypothesis only things that are indexed can be tracked and there are only 4 or 5 indexes! We also have some recent preliminary evidence that places near nontargets are not inhibited – they appear to have the same attention on them as targets and places near targets or in empty space. These results have made me rethink what goes on when there are more than 4 objects in a display, and it has resulted in an amended view of how indexing works. It has suggested that what many of us have been referring to as individuation should actually be decomposed into at least two aspects or stages – a stage of clustering or segregation or circumscription, and a stage where the cluster is given a demonstrative reference. The first sort of individuation occurs in parallel and is a purely physical aggregation process that yields tokens that may (or may not) receive a reference token or index. Non-aggregated (or empty) places in the visual field cannot get reference tokens and so cannot take part on such psychological processes as apparent motion or stereo disparity computation, which require distinct tokens aggregated over space and time in order to partake in correspondence assignment.
12. *There is some evidence that the specifically tracking aspect of MOT may not draw on a single general attentional resource*. We have found that secondary tasks that are sufficiently attention-demanding to produce a decrement in common tasks generally thought to be attentive did not lead to a decrement in tracking. Moreover it appears that observers automatically track moving objects, at least for short durations, even when such tracking is not required for the task at hand and their tracking performance is not assessed. Since this particular finding is sufficiently counterintuitive we are continuing to explore the reasons for it. In fact we recently have found evidence that tracking of items that were once distinguished (say by being flashed) happens whether or not the observer intends to tack them – i.e. whether or not there is a tracking task involved. It appears that having once been distinguished in some manner – even in a totally task-irrelevant manner – carries with it the special status that can be demonstrated in such measures as the ability to detect a small faint probe dot that appears unpredictably on the objects. We have shown this special tagging property of once-distinguished objects in several different experiments.

[8] Not everyone finds this obvious, but in fact the logic of tracking requires that for each object that is correctly tracked it must be the case that the subject knows that it belonged to the set of targets in the immediately preceding instant and the only way to know that it was a member of the set of targets in the preceding instant is by knowing which object it was in that set and that requires knowing which object it was in the preceding instant and so on until one works back to the beginning of the trial when the targets were distinct. What people might be thinking is that you can be wrong as to which object it was and still know that it was a target. This indeed describes what we found; now the question is why should you be wrong about which object is was when two targets come together while not when a target and a nontarget come close together. I do present an hypothesis in the next paragraph.

A theory of visual indexing and binding: The FINST mechanism

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 – e.g., 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.

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 4 or 5) internal symbols or parts 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 the visual objects as the *same* individuals despite changes in their 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 else discontinuity in the presence of local occlusion cues such as those mentioned above in discussing the Yantis [86] and Scholl [66] 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 [74].

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 the distal scene; (2) recently activated clusters compete for a pool of 4-5 visual indexes or FINSTs; (3) index assignment is primarily stimulus-driven, although cognitive factors, such as scanning focal attention until an object is encountered that activates an index, may have a limited effect; (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 perceptual 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 in Figure 9. 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 resulting in an existing binding being lost). As new properties of the inducing element are detected 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 I have in mind here is one that appears in computer science when we speak of pointers or when variables are assigned values. To have this sense of reference is to be able to access the referents in certain ways: to interrogate them in order to determine some of their properties, to evaluate multi-place predicates over them, to move focal attention to them, and in general to *bind* cognitive arguments to them, as would have to be done in order to execute a motor command towards them. The important thing here is that the inward arrows are purely causal and are instantiated by the non-conceptual apparatus which, following the terminology suggested by [75], I have discussed under the name

early vision [76]. The indexing system latches onto certain kinds of spatiotemporal objects because it is "wired" to do so, or because it is in the nature of its functional architecture 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.

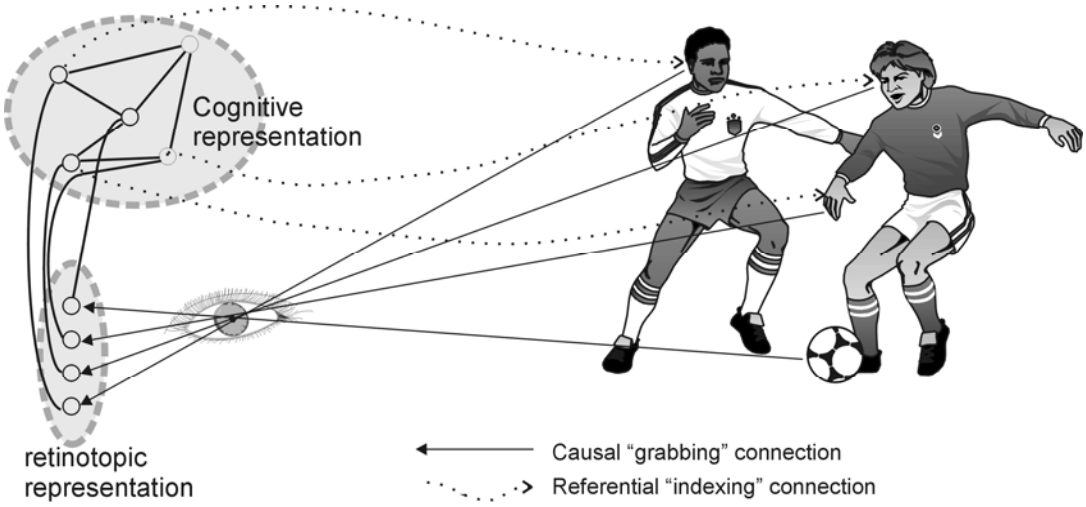

Figure 9: Sketch of the types of connections established by visual indexes between the primitive visual objects or proto-objects and parts of conceptual structures, depicted here as a network.

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 [87] 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 it 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 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, or between indexing and knowing where something is, and so on). In fact a very simple network, such as the one described by [88] can implement such a function[9] [the application of the Koch & Ullman network to Visual Index theory has been explored in 90,91]. All that is required is a winner-take-all circuit whose convergence on a certain active place on a spatiotopic map enables a signal to be sent to that place, 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 92]. The important point about such a network, which makes its pointing function essentially pre-conceptual, is that the process that sends the probe signal to a particular place *uses no*

[9] 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. Any early vision system, such as Marr's, will contain sensors and a way of clustering features [**75]**. 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 & Ullman [**88]** 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 [**89]**.

encoding of properties of that place, not even its location. Being able to probe a certain place depends only on its being the most active by some measure [such as the activation measures assumed in many theories of visual search, like those of 43, or 89]. 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 results in the probe location moving in response to the movement of the primitive object [see 88, for details – also see Note 5].

Discussion: Objects and the Mind-World Connection

What I have described is 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 [93] has made the generalized notion of object the centerpiece of a radical reformulation of metaphysics. We share 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 we share with Smith 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 that is 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 I have been proposing 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. This general conclusion is also supported by evidence from clinical neuroscience, where it has been shown that deficits such as unilateral neglect [94] or Balint syndrome [95] apply over frames of reference that are object-based, wherein what is neglected appears to be specified with respect to individual objects. From this initial idea I 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 the work by [96-98] and 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, Alan Leslie [98] describes a number of studies in which one or two objects are placed behind a screen and the screen is then lowered 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 nonconceptual 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 multiple-object tracking examples, the notion of primitive visual object I have introduced 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 defined by Liz Spelke and others as a "bounded, coherent, three-dimensional physical object that moves as a whole" [99]. Are such Spelke-objects different from what we have been calling visual objects or proto-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 [97,100], quite possibly the *concept* of an object is not involved in picking out these objects, just as no concept at all (i.e., no description) plays a role in such phenomena as multiple-object tracking. 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 that 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 at least in some of the phenomena — see [98]. 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 Wynne [101] 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 the 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 the them come into existence by being *poured from a beaker* [102] 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. Several studies have shown that the precise *manner* in which objects disappear and reappear matters to whether or not they continue to be tracked [66]. 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 multiple-object tracking paradigm. But this sort of effect of recent visual history is quite plausibly subsumed under the operation of a non-conceptual 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 [76]).

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 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 ideas I have been discussing suggest an alternative and rather intriguing possibility. It is the notion that *primitive visual object* is the primary and more primitive category of early (nonconceptual) 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 what I argued is needed for independent reasons — some 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 many philosophers have argued [103]. 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 not to succumb to 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 104]. Without such a nonconceptual 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 interdefined 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. And because you can pick out *that* individual you can move your gaze to it or you can reach for it – your motor system cannot be commanded to reach for a red thing, only to reach for a particular individual.

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 that whatever the eventual solution turns out to be, it will have to respect a collection of facts, 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.

References

1 Pylyshyn, Z.W., Elcock, E.W., Marmor, M. and Sander, P. (1978) Explorations in visual-motor spaces. In *Proceedings of the Second International Conference of the Canadian Society for Computational Studies of Intelligence*

2 Kimchi, R., Behrmann, M. and Olson, C.R., eds (2003) *Perceptual organization in vision: Behavioral and neural perspectives*, Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers. (2003). xii, 475pp.

3 Kimchi, R. (2000) The perceptual organization of visual objects: A microgenetic analysis. *Vision Research* 40 (10-12), 1333-1347

4 Nakatani, K. (1995) Microgenesis of the length perception of paired lines. *Psychological Research* 58 (2), 75-82

5 Parks, T.E. (1995) The microgenesis of illusory figures: Evidence for visual hypothesis testing. *Perception* 24 (6), 681-684

6 Sekuler, A.B. (1994) Local and global minima in visual completion: Effects of symmetry and orientation. *Perception* 23 (5), 529-545

7 Sekuler, A.B. and Palmer, S.E. (1992) Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General* 121 (1), 95-111

8 Bachmann, T. (1991) Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology* 3 (1), 87-103

9 Froufe, M. (1989) What is perceived when one "does not perceive"? Backward masking: An equivocal technique for analyzing the microgenesis of visual perception. *Estudios de Psicologia. No* 37, 105-123

10 Sergent, J. (1989) Ontogenesis and microgenesis of face perception. *Cahiers de Psychologie Cognitive* 9 (1), 123-128

11 Bachmann, T. (1989) Microgenesis as traced by the transient paired-forms paradigm. *Acta Psychologica* 70 (1), 3-17

12 Bachmann, T. (1987) Different trends in perceptual pattern microgenesis as a function of the spatial range of local brightness averaging. *Psychological Research* 49 (2-3), 107-111

13 Tucker, V. and Broota, K.D. (1985) Effect of exposure duration on perceived size. *Psychological Studies* 30 (1), 49-52

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

15 Westerlundh, B. (1983) Personal organization of the visual field: A study of ambient to focal reports of threatening stimuli. *Archiv fur Psychologie* 135 (1), 17-35

16 Reynolds, R.I. (1978) The microgenetic development of the Ponzo and Zoellner illusions. *Perception & Psychophysics* 23 (3), 231-236

17 Frohlich, W.D. and Laux, L. (1969) Sequential perception, microgenesis, integration of information and orienting reactions: I. Actual genetic model and orientation reaction. *Zeitschrift fuer Experimentelle und Angewandte Psychologie* 16 (2), 250-277

18 Nesmith, R. and Rodwan, A.S. (1967) Effect of Duration of Viewing on Form and Size Judgments. *Journal of Experimental Psychology* 74 (1), 26-30

19 Kimchi, R. (2003) Visual perceptual organization: A microgenetic analysis. In *Kimchi, Ruth (Ed); Behrmann, Marlene (Ed); et al. (2003). Perceptual organization in vision: Behavioral and neural perspectives.*, pp. 117-154, Lawrence Erlbaum Associates, Publishers

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

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

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

23 Perry, J. (1979) The problem of the essential indexical. *Noûs* 13, 3-21
24 Lepore, E. and Ludwig, K. (in press) The semantics and pragmatics of complex demonstratives. *Mind*
25 Campbell, J. (2003) Reference as attention. *Philosophical Studies*?, 265-276
26 Campbell, J. (2002) *Reference and Consciousness*, Oxford University Press
27 Clark, A. (2000) *A Theory of Sentience*, Oxford University Press
28 McDowell, J. (1994) *Mind and World*, Harvard Univ Press
29 Pylyshyn, Z.W. (1984) *Computation and cognition: Toward a foundation for cognitive science*, MIT Press
30 Treisman, A. (1995) Modularity and attention: Is the binding problem real? In *Visual selective attention* (Bundesen, C. and Shibuya, H., eds.), Lawrence Erlbaum Associates
31 Pylyshyn, Z.W. (2001) Visual indexes, preconceptual objects, and situated vision. *Cognition* 80 (1/2), 127-158
32 Navon, D. (1977) Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology* 9, 353-383
33 Ullman, S. (1984) Visual routines. *Cognition* 18, 97-159
34 Ballard, D.H., Hayhoe, M.M., Pook, P.K. and Rao, R.P.N. (1997) Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* 20 (4), 723-767
35 Watson, D.G. and 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
36 Yantis, S. and Jones, E. (1991) Mechanisms of attentional selection: temporally modulated priority tags. *Perception and Psychophysics* 50 (2), 166-178
37 He, S., Cavanagh, P. and Intriligator, J. (1997) Attentional resolution. *Trends in Cognitive Sciences* 1 (3), 115-121
38 Intriligator, J. and Cavanagh, P. (2001) The spatial resolution of attention. *Cognitive Psychology* 4 (3), 171-216
39 Trick, L.M. and Pylyshyn, Z.W. (1994) Why are small and large numbers enumerated differently? A limited capacity preattentive stage in vision. *Psychological Review* 101 (1), 80-102
40 Baylis, G.C. and Driver, J. (1993) Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Percepton and Performance* 19, 451-470
41 Scholl, B.J. (2001) Objects and attention: The state of the art. *Cognition* 80 (1/2), 1-46
42 Pylyshyn, Z.W. (2003) *Seeing and visualizing: It's not what you think*, MIT Press/Bradford Books
43 Treisman, A. and Gelade, G. (1980) A feature integration theory of attention. *Cognitive Psychology* 12, 97-136
44 Pashler, H.E. (1998) *The Psychology of Attention*, MIT Press (A Bradford Book)
45 Nissen, M.J. (1985) Accessing features and objects: Is location special? In *Attention and performance XI* (Posner, M.I. and Marin, O.S., eds.), pp. 205-219, Lawrence Erlbaum
46 Pylyshyn, Z.W. (1994) Some primitive mechanisms of spatial attention. *Cognition* 50, 363-384
47 Pylyshyn, Z.W. (1998) Visual indexes in spatial vision and imagery. In *Visual Attention* (Wright, R.D., ed.), pp. 215-231, Oxford University Press
48 Tipper, S., Driver, J. and Weaver, B. (1991) Object-centered inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology A* 43A, 289-298
49 Tipper, S.P., Weaver, B., Jerreat, L.M. and 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
50 Egly, R., Driver, J. and Rafal, R.D. (1994) Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General* 123 (2), 161-177
51 Tsotsos, J.K. (1988) How does human vision beat the computational complexity of visual perception. In *Computational Processes in Human Vision: An interdisciplinary perspective* (Pylyshyn, Z.W., ed.), pp. 286-340, Ablex Publishing
52 Calis, G.J., Sterenborg, J. and Maarse, F. (1984) Initial microgenetic steps in single-glance face recognition. *Acta Psychologica* 55 (3), 215-230.
53 Reynolds, R.I. (1978) The microgenetic development of the Ponzo and Zollner illusions. *Perception and Psychophysics* 23, 231-236
54 Sekuler, A.B. and Palmer, S.E. (1992) Visual completion of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General* 121, 95-111
55 Rensink, R.A. (2000) Visual search for change: A probe into the nature of attentional processing. *Visual Cognition* 7, 345-376
56 Rensink, R.A., O'Regan, J.K. and 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
57 Rensink, R.A., O'Regan, J.K. and Clark, J.J. (2000) On the failure to detect changes in scenes across brief interruptions. *Visual Cognition* 7, 127-145
58 Simons, D.J. (1996) In sight, out of mind: When object representations fail. *Psychological Science* 7 (5), 301-305
59 Simons, D.J. and Levin, D.T. (1997) Change blindness. *Trends in Cognitive Sciences* 1, 261-267
60 Burkell, J. and Pylyshyn, Z.W. (1997) Searching through subsets: A test of the visual indexing hypothesis. *Spatial Vision* 11 (2), 225-258
61 Yantis, S. (1992) Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology* 24, 295-340
62 Sears, C.R. and Pylyshyn, Z.W. (2000) Multiple object tracking and attentional processes. *Canadian Journal of Experimental Psychology* 54 (1), 1-14
63 Blaser, E. and Pylyshyn, Z.W. (1999) Measuring the independence of attention to multiple features (abstract). *Perception* 28, p56
64 Blaser, E., Pylyshyn, Z.W. and Holcombe, A.O. (2000) Tracking an object through feature-space. *Nature* 408 (Nov 9), 196-199
65 McKeever, P. (1991) Nontarget numerosity and identity maintenance with FINSTs: A two component account of multiple-target tracking. In *Psychology*, University of Western Ontario
66 Scholl, B.J. and Pylyshyn, Z.W. (1999) Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology* 38 (2), 259-290

67 Scholl, B.J., Pylyshyn, Z.W. and Feldman, J. (2001) What is a visual object: Evidence from target-merging in multiple-object tracking. *Cognition* 80, 159-177
68 Culham, J.C., Brandt, S.A., Cavanagh, P., Kanwisher, N.G., Dale, A.M. and Tootell, R.B.H. (1998) Cortical fMRI activation produced by attentive tracking of moving targets. *J Neurophysiology* 80 (5), 2657-2670
69 Bahrami, B. (2003) Object property encoding and change blindness in multiple object tracking. *Visual Cognition* 10 (8), 949-963
70 Viswanathan, L. and Mingolla, E. (2002) Dynamics of attention in depth: Evidence from multi-element tracking. *Perception* 31 (12), 1415-1437
71 Eriksen, C.W. and St. James, J.D. (1986) Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics* 40, 225-240
72 Intriligator, J. and Cavanagh, P. (1992) Object-specific spatial attention facilitation that does not travel to adjacent spatial locations. *Investigative Ophthalmology and Visual Science* 33, 2849 (abstract)
73 Awh, E. and Pashler, H. (2000) Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception & Performance* 26 (2), 834-846
74 Scholl, B.J., Pylyshyn, Z.W. and Franconeri, S.L. (submitted) The relationship between property-encoding and object-based attention: Evidence from multiple-object tracking.
75 Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*, W.H. Freeman
76 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
77 Rensink, R.A. (2000) The dynamic representation of scenes. *Visual Cognition* 7, 17-42
78 Di Lollo, V., Enns, J.T. and 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
79 Wolfe, J.M. and Bennett, S.C. (1997) Preattentive object files: shapeless bundles of basic features. *Vision Research* 37 (1), 25-43
80 Yantis, S. and Gibson, B.S. (1994) Object continuity in apparent motion and attention. *Canadian Journal of Experimental Psychology* 48 (2), 182-204
81 Nakayama, K., He, Z.J. and Shimojo, S. (1995) Visual surface representation: A critical link between lower-level and higher-level vision. In *Visual Cognition* (Kosslyn, S.M. and Osherson, D.N., eds.), pp. 1-70, MIT Press
82 Sears, C. (1991) Information processing at multiple locations in the visual field. In *Psychology*, University of Western Ontario
83 Keane, B. and Pylyshyn, Z.W. (2003) Does tracking disappearing objects in MOT involve predicting the locus of reappearance? In *Vision Sciences*
84 Klein, R. (2000) Inhibition of return. *Trends in Cognitive Sciences* 4 (4), 138-147
85 Snyder, J.J. and Kingstone, A. (2000) Inhibition of return and visual search: How many separate loci are inhibited? *Perception and Psychophysics* 62 (3), 452-458
86 Yantis, S. (1998) Objects, Attention, and Perceptual Experience. In *Visual Attention* (Wright, R., ed.), pp. 187-214, Oxford University Press
87 Posner, M.I., Snyder, C. and Davidson, B. (1980) Attention and the detection of signals. *Journal of Experimental Psychology: General* 109, 160-174
88 Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* 4, 219-227
89 Wolfe, J.M., Cave, K.R. and Franzel, S.L. (1989) Guided search: an alternative to the feature integration model for visual search. *J Experimental Psychology: Human Perception and Performance* 15 (3), 419-433
90 Acton, B. (1993) A Network Model of Visual Indexing and Attention. In *Electrical Engineering*, University of Western Ontario
91 Pylyshyn, Z.W. and Eagleson, R.A. (1994) Developing a network model of multiple visual indexing (abstract). *Investigative Ophthalmology and Visual Science* 35 (4), 2007-2007
92 Pylyshyn, Z.W. (2000) Situating vision in the world. *Trends in Cognitive Sciences* 4 (5), 197-207
93 Smith, B.C. (1996) *On the Origin of Objects*, MIT Press
94 Driver, J. and Halligan, P. (1991) Can visual neglect operate in object-centered coordinates? An affirmative single case study. *Cognitive Neuropsychology* 8, 475-494
95 Robertson, L., Treisman, A., Friedman-Hill, S. and Grabowecky, M. (1997) The interaction of spatial and object pathways: Evidence from Balint's syndrome. *Journal of Cognitive Neuroscience* 9 (3), 295-317
96 Xu, F. and Carey, S. (1996) Infants' Metaphysics: The case of numerical identity. *Cognitive Psychology* 30, 111-153
97 Spelke, E., Gutheil, G. and Van de Walle, G. (1995) The development of object perception. In *Visual Cognition* (Vol. 2) (Kosslyn, S.M. and Osherson, D.N., eds.), pp. 297-330, MIT Press
98 Leslie, A.M., Xu, F., Tremolet, P.D. and Scholl, B.J. (1998) Indexing and the object concept: Developing `what' and `where' systems. *Trends in Cognitive Sciences* 2 (1), 10-18
99 Spelke, E.S. (1990) Principles of object perception. *Cognitive Science* 14, 29-56
100 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
101 Chiang, W.-C. and Wynn, K. (2000) Infants' tracking of objects and collections. *Cognition* 75, 1-27
102 Carey, S. (1999) Establishing representations of new individuals: New infant results and old studies by Michotte. In *Object Cognition: Underlying mechanisms and their origins (May 20-21)*, Unpublished
103 Hirsch, E. (1982) *The Concept of Identity*, Oxford
104 Miller, G.A., E. Galanter, and K. H. Pribram. (1960) *Plans and the Structure of Behavior*, Holt, Rinehart & Winston