Perception, Representation and the World: The FINST that binds[1]

Zenon Pylyshyn, Rutgers Center for Cognitive Science

1. Some historical background

I recently discovered that work I was doing in the laboratory and in theoretical writings was implicitly taking 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 (Pylyshyn, Elcock, Marmor, & Sander, 1978). 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 while constructing a certain figure, and if so *which particular* lines they were. 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 use the term "element" when referring to a graphical unit such as used in experiments. Otherwise when speaking informally I use the term "thing" on the grounds that nobody would mistake that term for a technical theoretical construct. Eventually I end up calling them "Visual Objects" to conform to usage in psychology) 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

[1] To appear in Pylyshyn, Z. W. (in press). Perception, Representation and the World: The FINST that binds. In D. Dedrick & L. M.

came to be called “**FIN**gers of **INST**atiation” or **FINST**s) 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 25 years.

The idea of FINSTs as constituents of perceptual representations is a departure from the view of perceptual representation I had taken in “Computation and Cognition” some 23 years ago (Pylyshyn, 1984) because it postulated a mental symbol that was not connected to the world by the semantic relation of satisfaction but by a causal or informational link. In the course of the work since that book 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 we construct perceptual representations incrementally over time (Calis, Sterenborg, & Maarse, 1984; Frohlich & Laux, 1969; Kimchi, 2000; Nakatani, 1995; Nesmith & Rodwan, 1967; Parks, 1995; Reynolds, 1978a; Sekuler & Palmer, 1992; Tucker & Broota, 1985), so we cannot escape the need to keep track of individual objects *qua individuals* over time.

Around the same time as we undertook these experiments (initially reported in Pylyshyn, 1989; Pylyshyn & Storm, 1988) another set of experiments were independently published by Daniel Kahneman (Kahneman, Treisman, & Gibbs, 1992), who introduced the concept of an *Object File*. An Object file contains the *conceptual* representation of a (visual) object with which it is associated. Although this was not stressed in the Kahneman et al. report, object files are connected to individual

Trick (Eds.), *Computation, Cognition and Pylyshyn*. Cambridge, MA: MIT Press.

visual objects and keep accumulating information about the individuals as it tracks them. Our view is similar to that of Kahneman with two notable exceptions: (1) We were concerned primarily with the question of how an object file is associated with its appropriate object (answer: through the primitive index mechanism we call FINSTs) and (2) We assumed that the FINST index does not itself use the contents of Object Files in order to track the individual object token with which it is associated

As with many ideas, it took a long time to appreciate that the basic idea was actually a proposal that introduced nonconceptual representation. Eventually 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 "attention" and different from the semantic connection of *satisfaction* with which philosophers have had a long-standing but perplexing relationship. FINSTs differ from what psychologists call *focal attention* in several respects: (1) there are a small number of them (2) they are generally data-driven – i.e., assigned by events taking place in the visual field (3) they pick out individual things as opposed to regions, (4) they adhere to (stay connected to) the same individual thing (whatever exactly that turns out to be) as the thing moves around and changes any or all of its properties, (5) their attachment is not mediated by a description of (i.e., an encoding of properties of) the thing in question. There are two theoretical reasons why these indexes function without an encoding of objects' properties. One is that there generally is no fixed (tenseless) description that uniquely characterizes a particular token thing. Another is that one of main function purposes of FINSTs is to keep track of things qua individuals, independent of what properties they may in fact have. Although these assumptions largely reflect empirical facts about vision that have since been supported by experiments, they are inherent in the function which FINSTs were called upon the perform in our initial analysis (which I will explore using several different examples in this essay). The above 5 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 not only serve to refer to *token* things but they do so without representing the thing as falling under a concept or a description: The relation between the representation and the thing (or visual object) represented is not one in which the object *satisfies* the description. Rather it is purely causal.

The FINST, according to this story, is an instrument of reference by which one can pick out and refer to things. The reference is nonconceptual since it does not refer to things that have certain properties of that fall under certain conceptual categories. Thus it is very similar to a demonstrative (such as *this* or *that*), the only exception being that in the case of actual words, the referent is conditioned by the intentions of the speaker as well as other contextual factors, such as pointing or gazing at the referent. FINSTs may be thought of as demonstrative terms in the language of thought which allow a person to think about something in the world that was selected in perception (especially vision) because something drew attention to itself or, as we prefer to say, something grabs a FINST index. Once a FINST reference is established, it can be used to bind arguments of mental predicates, or conceptual information about the referent can be entered into the associated Object File. Although the FINST idea may seem simple enough, it has surprising consequences. To give you a sense of how far-reaching this idea is I point out that we have assumed that FINSTs provide a mechanism for referring to visual objects without appealing to their conceptual properties, which means that, in an important sense, the referrer does not know what he or she is referring to! To refer to something (say that object in the corner of my room) without referring to it as a cat, or as some mass with a particular shape, or as a patch of tawny-color, or (as Quine might put it) as a collection of undetached cat parts, is a strange notion. Yet there *must* be a stage in the visual process where something like this happens, otherwise we could not construct our conceptual representations on a foundation of causal connections to the world, as we must to avoid solipsism.

The question whether we need to postulate a nonconceptual form of reference has been much debated in philosophy. Among those whose views are consistent with the idea of nonconceptual representation (even though they may shun the term “representation”) are certain AI people It is generally assumed in cognitive science that behavior (including what we see, how we think, what we decide and what actions we take) can only be explained if we appeal to how people conceptualize the world around them. To conceptualize is to represent in terms of categories or properties. When we see something we see it as falling under a particular concept; we see it *as* a "person" or "car" or whatever. Seeing something as a P is a paradigm case of conceptual representation, where the predicate P is the relevant concept. But, as we will see in the next section, there is a problem in connecting this sort of conceptual representation with the brute physical world. As a result many people have talked about a level of nonconceptual representation as mediating between world and mind. The issue of whether it makes sense to postulate a nonconceptual form of reference has been much debated in philosophy and elsewhere in cognitive science. Among those who support the idea of nonconceptual representations are certain AI people (Brooks, 1991) or philosophers (Clark, 1999) who speak of embodied or situated cognition (and in fact some of these writers shun the use of the term "representation" entirely, although I believe that their view leads naturally to a form of nonconceptual representation). My position is closer to that of philosophers who speak of essential indexicals (Perry, 1979), and logicians who argue for bare demonstratives (Lepore & Ludwig, 2000), which are closely related to FINSTs. Many philosophers who write about the mind-world interface wish to ward off skeptical arguments by claiming that the most primitive reference must be accessible to conscious experience. John Campbell (Campbell, 2003) uses the phrase "conscious attention" to emphasize his claim about the essential conscious character of attention-based reference. Many writers also assume that the most basic form of reference must pick out locations or at least regions, believing that a mental grip on a region is the more acceptable form of contact between mind and world since it is possible to imagine regions being picked out by a "spotlight of attention". Still other philosophers deny that the mind-world link requires a non-conceptual representation at all (McDowell, 1994). 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 early in an organism's development)). I will return to these questions later but will begin by setting the stage for the view I have been defending in recent years.

2. 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 think about it – 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 representation represents a visual scene "under a description", that is, it represents the visual objects 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 (Pylyshyn, 1984). We need this sort of representation because what determines our behavior is not the physical properties of the things around us, but how we interpret or classify them – or more generally *what we take them to be* – that matters. 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 perceive or conceptualize it in certain ways that allows us to think about it. This is common ground for virtually all contemporary theories of cognition.

Although I have emphasized the representation-governed nature of cognition, this is not the whole story, even if augmented with sensory transducers (as I assumed in, Pylyshyn, 1984). It turns out that the sort of description-building view of perception is missing a critical piece; how the descriptors connect with what they describe. 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 *preconceptually* (the preferred term in philosophy appears to be *non*conceptually). For example, in the presence of a visual stimulus, we can think thoughts that involve individual things by using a term such as "*that*" and thinking "*that* is a pen" where the term (in mentalese) "*that*" refers to something we have picked out in our field of view without reference to what conceptual category it falls under or what properties it has. A term such as *this* or *that* is called a "demonstrative". Demonstratives in natural language work slightly differently than FINST do because, as a tool for communication, they are tied to the intention of the speaker and may even require pointing or some other directional gesture (such as direction of gaze), none of which concerns FINSTs.

Philosophers like John Perry (Perry, 1979) have argued that demonstratives are ineliminable in language and thought. The reason for the ineliminability of demonstratives also applies 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, how would the category be specified? It would presumably be defined in terms of other conceptual properties, and so on. In that case 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.[1] 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 FINST 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. 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 visual objects in a scene causally invoke certain visual objects in a representation, and in which the visual objects in the representation can in turn be used to refer to particular individuals in the world. The need to pick out and refer to individual things is not something that arises under arcane circumstances, but happens 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 why focal attention is essential for organisms like us (but see Pylyshyn, forthcoming), but it may be useful to at least list them since they are not always recognized or appreciated.

2.1 Some reasons why we need a mechanism for *selecting* or *picking out* token things:

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 things over which attention selects).

2. *Incremental construction of representations*. 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.

3. *Solving the binding problem* Information about the world is “packaged” or presented in certain ways that lead to what Austen Clark (Clark, 2000) calls the “binding problem” (after Treisman, 1995, who introduced the term) or the “many properties problem” (after Jackson, 1997). Very early in the visual information-processing stream we must distinguish between properties present in a scene and conjunctions of these properties present on individual objects (i.e., 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 at the level of sentience, but I prefer to 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 information would be fused and unrecoverable. I return to this topic in Section 3.2(b).

4. *Detection of patterns defined in terms of parts*. Visually discriminable patterns that are made up of parts cannot be represented unless we can specify which things partake in that pattern. The predicates **Collinear**(x,y,z), or **Inside**(x,y) or **Above**(x,y) or even **Location**(x,y,z) cannot be evaluated unless the arguments x, y, z are instantiated by objects in the scene (i.e., unless the variables are *bound*, in the computer science sense of that term, where this means bound to the values of its argument rather than bound by a quantifier).

5. *Tagging of individuals in a scene to mark them during visual processing*. Many visual patterns can only be discriminated if a serial process operates over the visual objects, which requires that

> token visual objects be somehow “marked” so they may be referred to by what Ullman calls “visual routines”. 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 – the visual system simply picks out and refers to certain token things).

In this essay I focus on the problem of establishing a correspondence between individual things in the world and their counterparts in a visual representation, since this is where the notion of a FINST index or FINST played its first theoretical role in our work. Before I describe how FINSTs are relevant to this connection, I offer a few remarks about what these things might be and also offer a few illustrations of how this sort of direct reference is missing from the usual representations that visual theories provide. Although I am concerned with the initial steps of the process that begin with nonconceptual connections between mind and world and eventually encodes a visual scene in terms of some conceptual structure. In that context we see FINSTs as a mechanism for connecting the mind with real physical objects in the world. But a FINST as a nonconceptual connection cannot, by its very nature, be guaranteed to pick out all and only individual physical objects because *physical object* is a conceptual category. Something is an individual physical object (or any other sort of individual) if it meets certain conditions (see any dictionary for a largely inadequate attempt to lay out such conditions). In particular it has to meet what Clark (Clark, 2000) has called *Strawsonian strictures*: it has be meet conditions of *individuation* and *identity*. To decide whether something is an individual physical object one must bring to bear criteria of identity (see the discussion of this point in, Strawson, 1963). What FINST indexes do is pick out a class of things that *in our kind of world* are very often coextensive with physical objects, yet which can be picked out without criteria of identity. The visual system very often yields a fast and automatic parsing of the world which provides a starting point for conceptual categories – even categories like *cause*, which can be nonconceptually recognized in certain circumstances (and the nonconceptual category can be distinguished from the conceptual one Schlottman & Shanks, 1992). Because the FINST indexes serve the function, in the overall operation of the visual system, to connect minds with physical objects (even though they may fail to do so sometimes). This is why I often speak of FINST indexes as referring to visual objects or even just "objects". They do, however, sometimes fail to select a physical object (e.g. if it is too small or too big, if the lighting is poor, or if it is an illusion, such as provided by holograms). What one does about such errors is a question that faces every theorist, since even with Strawsonian strictures there will inevitably be illusions and other sources of error and failures of re-identification. We simply recognize that there may be P-detectors even if they do not always detect all and only P's.

Before moving on to an explication of the theory and the experiments I would like to provide some additional background by way of motivation for the principles of selection and nonconceptual indexing listed above. Theories of visual perception universally attempt to provide an effective (i.e., computable) mapping from dynamic 2D patterns of proximal (retinal) stimulation to a representation of a 3D scene. Both the world and its visual representation contain certain individuals. 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 possess the properties that are 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 works more like a demonstrative reference than a description. And that, I suggest, is why we must have something like a FINST 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).

3. The need for individuating and indexing: Empirical motivations[2]

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 is 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 a something 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 encapsulated 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) in our kind of world or in our ecological niche. That is the beauty and the ingenuity 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 precise but accidental alignments have a very low probability of occurring, in which the light tends to come from above and casts shadows downward, and so on. 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. It is blissfully ignorant but superlatively successful in our sort of world.

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 visual objects. 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: (1) The assumption that individuation of object tokens is primitive and nonconceptual and precedes the detection of properties, (2) the assumption that detection of visual properties is the detection of properties-of-objects, as opposed to the detection either of properties *tout court* or properties-at-locations, and (3) the assumption that visual representations are generally constructed incrementally over time.

3.1 Assumption 1: Individuation of object tokens is primitive and precedes the detection of properties

(a) Evaluating Visual Predicates

The process of individuating visual 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 visual objects in Figure 1 are arranged linearly, especially if the visual objects in the figure were gradually changing their properties or if the figure as a whole was moving while maintaining the collinear arrangement. In general, featural properties of visual objects tend to be factored out when computing global patterns, regardless of the size and complexity of the global pattern (Navon, 1977). Computing global patterns such as collinearity, or others discussed by (Ullman, 1984), requires that visual objects be registered as individuals while their local properties are ignored. Whatever the particular algorithm used to detect collinearity among visual objects, it is clear that specifying *which* points form a collinear pattern is a necessary part of the computation.

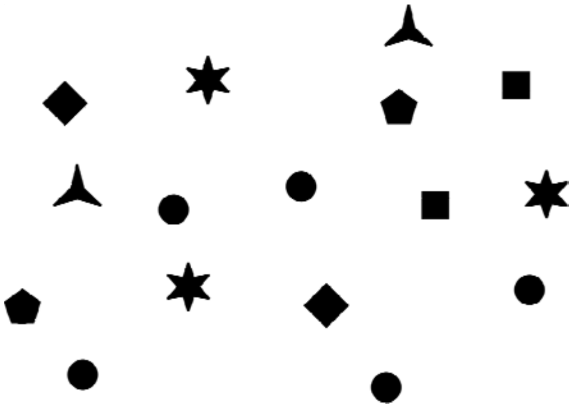

Figure 1. Find 4 or more items that are collinear. Judging collinearity requires selecting the relevant individual objects and ignoring all their intrinsic (local) properties.

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

(b) Visual individuation is different from visual discrimination

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 Intriligator & Cavanagh, 2001) 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 and Cavanagh 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 (He, Cavanagh, & Intriligator, 1997) show that the process of individuating objects is separate and distinct from that of recognizing or encoding the properties of the objects.

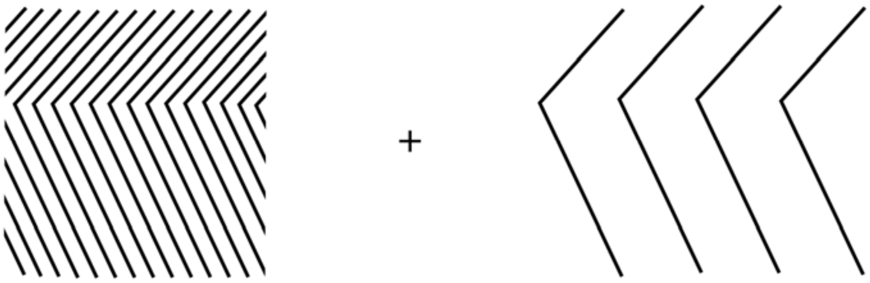

Figure 2 .At a certain distance if you fixate on the cross you can easily tell which groups consist of similar-shaped lines, although you can only **individuate** lines in the group on the right. For example, while holding the page at arms length and fixating the central cross you cannot count the lines or pick out the third line from the left, etc., in the panel on the left. (Based on Intriligator & Cavanagh, 2001).

(c) Rapid enumeration requires automatic individuation

Studies of rapid enumeration (called *subitizing*), described by Lana Trick (Trick & Pylyshyn, 1994), also show that individuating is distinct from (and prior to) computing the cardinality of a small set of objects. Trick and Pylyshyn showed that items arranged so they cannot be preattentively individuated (or items that require focal attention in order to individuate them – as in the case of items lying on a particular curve or specified in terms of conjunctions of features) cannot be subitized, even when there are only a few of them (i.e., the signature break in the function relating reaction time to number of items is not observed in those cases). 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.

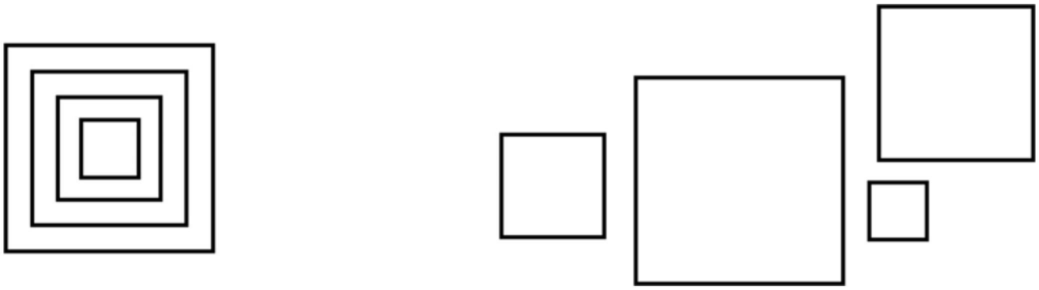

Figure 3. Squares arranged so they cannot be preattentively individuated (on the left) cannot be subitized, whereas the ones on the right are easily subitized (based on Trick & Pylyshyn, 1994).

(d) Subset selection

The following experiment by Jacquie Burkell (Burkell & Pylyshyn, 1997) 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 4) 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 to 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 produces longer search times that increase as the number of items in the display increases). 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. The *first* depends 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 only 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.

The *second* finding, of particular relevance to the present discussion 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. This is precisely what one would expect if the cued items are indexed and indexes are used to access the items *directly*, without having to scan the display. We 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. We were able to show that indexes assigned to the cued objects survive an eye movement so long as the saccade is generated in certain ways (e.g., if the eye is moved to view one of the target objects, but not if it is forced to move to the edge of the screen or to some secondary fixation point (Currie & Pylyshyn, 2003). 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 saccades each day – it does it by maintaining a cross-saccade correspondence on a few significant objects. Studies have shown that we cannot recall more than a few items from one fixation to another so this mechanism may be all we need (Irwin, 1992).

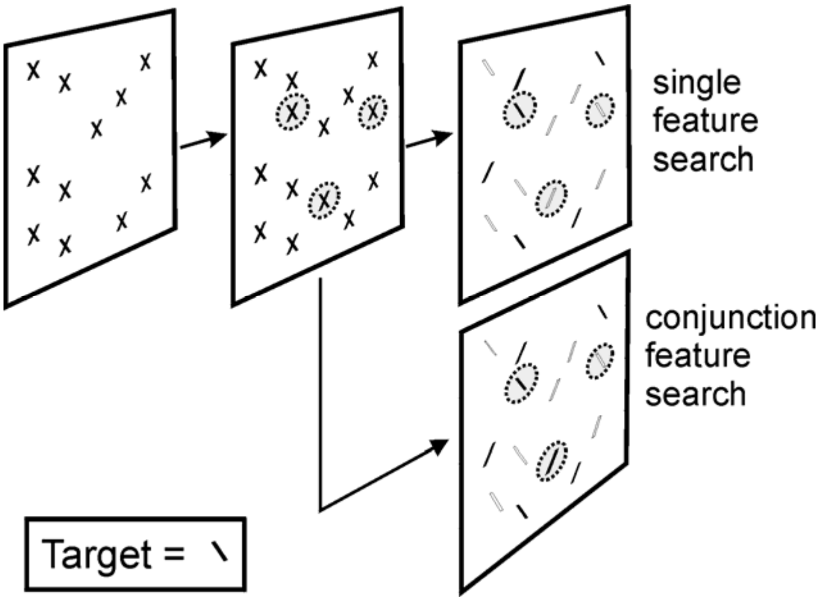

Figure 4: Sequence of events in the (Burkell & Pylyshyn, 1997) study. The observer sees a set of placeholder Xs, then 3-5 “late onset” placeholders appear briefly, signaling the items that will constitute the search items. Then all Xs change to search items (left or right oblique red or green line segments, shown here with circles around them for expository purposes) and the subject must try to find the specified target in one of two condition. In the top display the target differs from all the nontargets by one feature whereas in the bottom display, a combination of two features is required to distinguish the target.

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

3.2 Assumption 2: Detection of visual properties is the detection of properties-of-objects

When a property is first 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). 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* are red or circular or are arranged linearly. This, in turn, requires that the individuals are selected first. 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 (Baylis & Driver, 1993). 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).

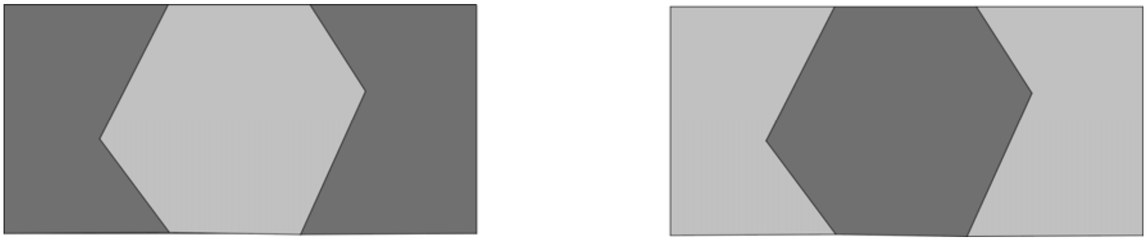

Figure 5. Figures used to demonstrate single-object advantage in judging properties of a shape within one figure vs between two figures. Based on (Baylis & Driver, 1993).

Other evidence supporting this conclusion comes from a variety of sources (many of which are reviewed in Scholl, 2001), including experiments in which objects move through space or in which they move through feature space. (More examples are discussed in Pylyshyn, 2003). 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 of properties

Another kind of evidence for the primacy of objecthood comes from the fact that we can distinguish the co-occurrence of features on an individual object from their mere occurrence somewhere in a scene. This has been called the binding problem or the multiple-properties problem. 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. 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 (Treisman & Gelade, 1980), in Clark's theory of sentience, in Campbell's analysis of consciousness (Campbell, 2002) and in most psychological theories (see, e.g., Pashler, 1998). But this will not work in general and where it does work, it confounds location and objecthood.

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 (Nissen, 1985). She argued that in reporting the conjunction of two features, observers must first locate the *place* in the visual field that has both features. In Nissen's studies this conclusion comes from a comparison of the probability of reporting a stimulus property (e.g., shape or color or location) or a pair of such properties, given one of the other properties as cue. Nissen found that accuracy for reporting shape and color were statistically independent, but accuracy for reporting shape and location, or for reporting color and location, were *not* statistically independent. More importantly, the conditional probabilities conformed to what would be expected if the way observers judged both color and shape 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 (Pashler, 1998, 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.

In contrast, it is clear that detection of objects must precede solving the binding problem because the location that would be required cannot be punctate – one must specify a region that contains both features. But which region? Try specifying the regions that share the dual (conjoined) properties in a figure such as the one in Figure 6 below. You can tell these two figures apart even though they contain the same figures and textures and can only be distinguished by which shape has which texture. The rectangular bounding region is the same so the only way to distinguish these two is to refer the particular texture to the region marked out as the outline of the figure with that texture. But you can only specify this sort of region by having selected the object and used its boundary as the region. Neither texture nor shape has a location apart from the object that has those properties. In addition, empty regions by themselves do not have causal properties and so are incapable to grabbing a FINST index.

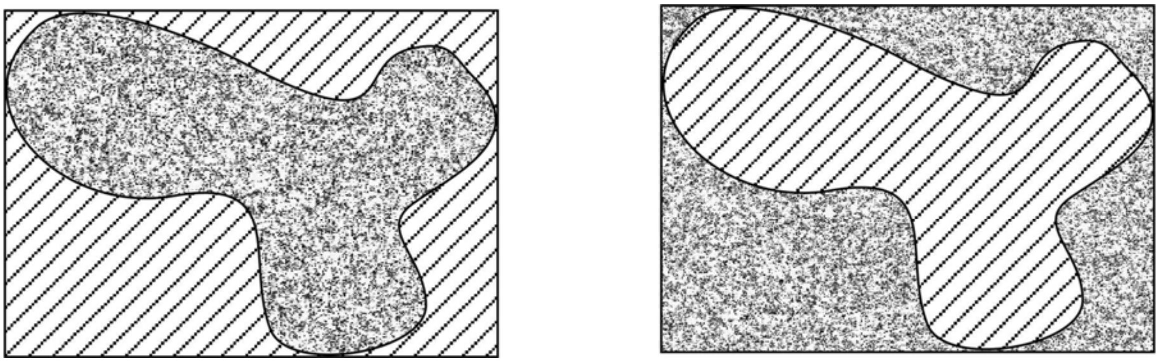

Figure 6. To distinguish these two figures you can't simply encode texture, shapes and their location, as done in Feature Maps, since they both have the same features and the same centroid (and the same bounding rectangle) location. You have to associate the texture with the region it occupies, and you can't specify that unless you have first picked out the object whose bounds constitute the relevant region.

(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” (Kahneman, Treisman, & Gibbs, 1992) and our own studies using multiple-object tracking (MOT) (see below, as well as Pylyshyn, 1994, 1998). 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.

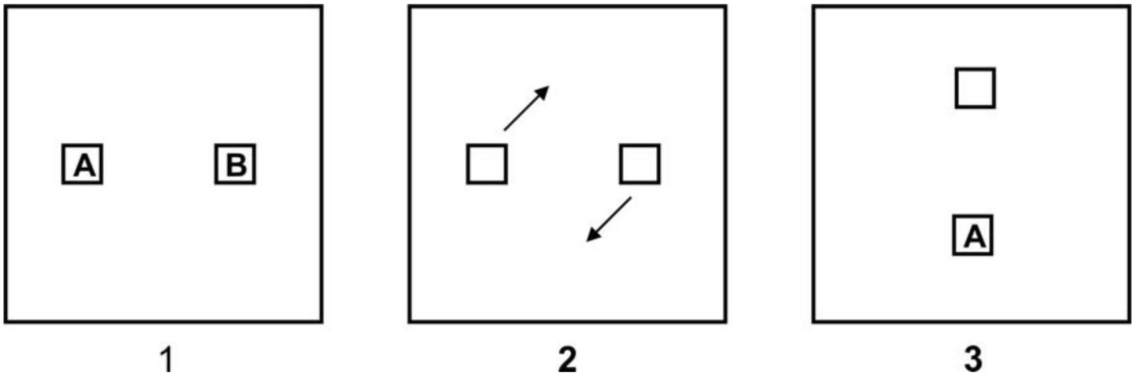

Figure 7: Studies showing facilitation of naming of a letter (the letter is named faster) when it recurs in the same box as it was in at the start of the trial, even though this was not predictive of which letter it was (since half the time it was the letter that had been in the other, equally distant, box). (Based on, Kahneman, Treisman, & Gibbs, 1992)

Similarly, related studies by Steven Tipper (Tipper, Driver, & Weaver, 1991) 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 Tipper, Weaver, Jerreat, & Burak, 1994, suggests that location-specific IOR also occurs).

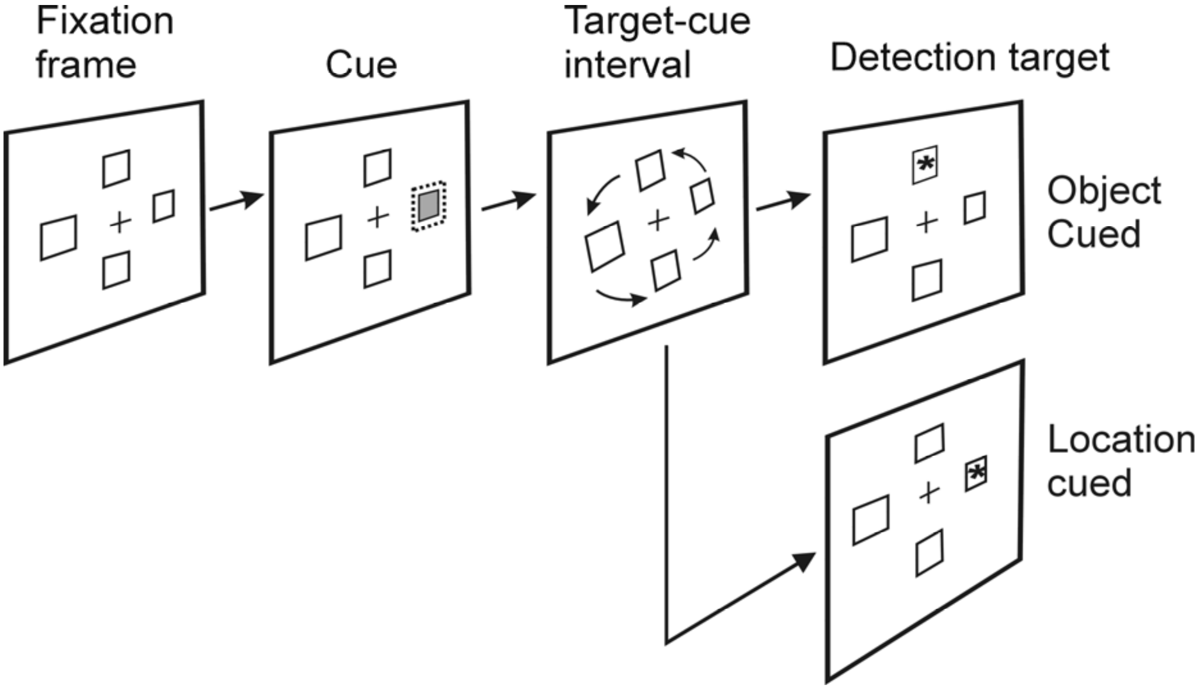

Figure 8 Inhibition of Return (IOR) is a phenomenon whereby items that are attended and then attention is removed from them become more difficult to re-attend during a period of from abut 300 ms to 900 ms after. It has been shown that what is inhibited in IOR is mostly the individual object that had 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 (but see, Sperling & Weichselgarter, 1995, for an alternative explanation of the apparent "attention movement" phenomena) 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 outside the objects in question. For example (Egly, Driver, & Rafal, 1994) showed that attention seems to spread throughout regions defined by contours, but only if those contours are perceived to be the contours of a single object.

3.3 Assumption 3: 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 (Tsotsos, 1988; Ullman, 1984) have provided good reasons for believing that some relational properties that hold between visual elements, such as the property of being inside or on the same contour, must be 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., Calis, Sterenborg, & Maarse, 1984; Reynolds, 1978b; Sekuler & Palmer, 1992). If that is so then the following problem immediately arises. If the representation is built up incrementally, we need a mechanism for determining the correspondence between representations of individual objects 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* objects 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? Or how do you know whether a particular object is a new objects or one you have seen and represented before – if you don't solve this correspondence problem correctly you will end up with a cacophony of duplicated objects in the representation of a scene. 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 previously unique 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 (Rensink, 2000; Rensink, O'Regan, & Clark, 1997; Rensink, O'Regan, & Clark, 2000; Simons, 1996; Simons & Levin, 1997). The alternative to this unwieldy method for locating a representation of a particular individual is to allow the descriptive apparatus to make use of a name or *demonstrative* reference. 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.[3] 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

4. Multiple object tracking (MOT)

I 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 *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 *FINST,* 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)* and is illustrated in Figure 9.

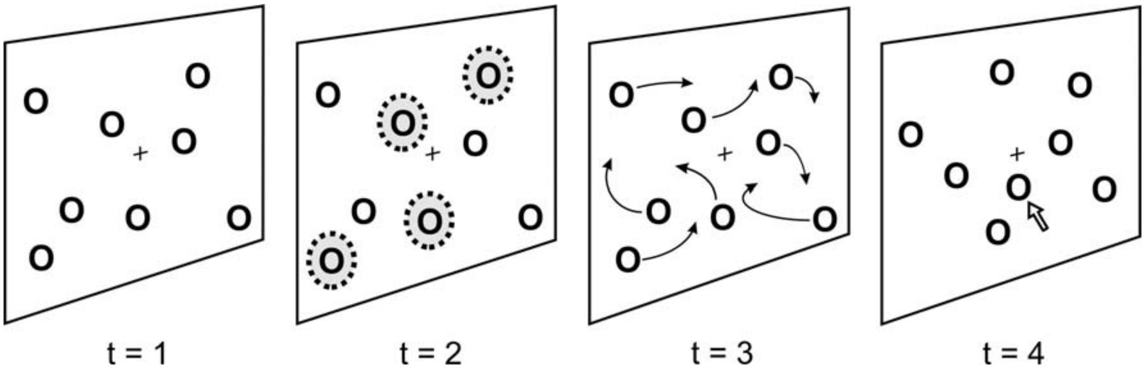

Figure 9. Illustration of a typical Multiple-Object Tracking Experiment. A number of identical objects are shown, then a subset (the "targets") is selected by blinking them, after which the objects move in unpredictable ways (with or without self-occlusion) for about 10 seconds. At the end of the trial the observer has to pick out all the targets using a pointing device. (Demonstrations of this and other MOT displays can be viewed at: http://ruccs.rutgers.edu/finstlab/demos.htm)

In a typical experiment, observers are shown anywhere from 8 to 12 simple identical objects (points, squares, circles, figure-eight shapes). A subset of these objects is briefly rendered distinct (usually by blinking 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"). After 10 or so seconds of tracking the objects stop moving and the observer must then indicate which of the now-identical objects were the targets by clicking on them using a computer mouse. A large number of experiments, beginning with the studies described in (Pylyshyn & Storm, 1988), have shown that observers can indeed track up to 5 independently moving targets within a field of 10 identical items.[4] 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 at a particular point in time. 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. So a possible tracking strategy would be to keep a record of objects' locations and visit them serially to update their location with each iteration until the end of the trial. We have simulated that algorithm on the actual displays we used and showed that with very conservative assumptions about location-encoding requiring focal attention which moves at a fine speed the best performance we could expect is about 30% which is very much less than the observed 87% . This means that the moving objects could not have been tracked by a using focal attention to update *the unique stored description of each figure* (i.e., their location). These studies suggest that the *early vision system* (an essentially encapsulated system, discussed at length in Pylyshyn, 1999), is able to individuate and keep track of about five visual objects and does so without using an encoding of any of their visual properties.

The multiple object tracking task exemplifies what is meant by "tracking" and by "maintaining the identity" of objects. It also operationalizes the notion of "visual object" as whatever allows nonconceptual selection and multiple-object tracking (since these *things* are interdefined with *FINSTs* I have sometimes called them *FINGs*). 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, such as the endpoints of lines, do not (Scholl, Pylyshyn, & Feldman, 2001). 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, (Scholl & Pylyshyn, 1999) showed that if the objects being tracked in the MOT paradigm disappear and reappear in certain ways, they are tracked as though they had a continuous existence. If, for example, they disappear and reappear by deletion and accretion along a fixed contour, the way they would have if they were moving behind an occluding surface (even if the edges of the occluder are not invisible), they are successfully tracked. However, performance in the MOT task degrades significantly in the control conditions where objects suddenly go out of existence and reappear at the appropriate matching time and place, or if they slowly shrink away to a point and then reappear by slowly growing again at exactly the same relative time and place as they had accreted in the occlusion condition. Beyond that, what qualifies as a primitive (potentially indexable) object remains an open empirical question. In fact, more recent evidence (Blaser, Pylyshyn, & Holcombe, 2000) 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).

4.1 How FINSTs are used in Multiple Object Tracking

From the point of view of FINST theory, the way MOT proceeds may be summarized as follows. When a subset of the objects blink on and off, each individual "target" captures a FINST (so long as there are not more than about 4 or 5 such blinking objects). Since objects are visually identical, the only current property that distinguishes one object from another is its location on the screen. What distinguishes targets from nontargets is that targets are the visual objects that earlier had been visually distinct in some way (in this case by their blinking) – i.e., by their past history. So in order to identify targets as distinct from nontargets it is necessary either to identify them by their location or to trace their provenance or their identity back to the start of the trial, and thereby to ascertain their origin status as target. In (Pylyshyn & Storm, 1988) we argued that it is unlikely that observers track the targets by cyclically updating a record of their locations as they move about, and then using the list of target location at the end of the trial to specify targets. On the basis of that argument we concluded that indexing does not use location information to track objects. Indexes simply attach to objects and when the objects move they carry the indexes with them (providing that the motion is within certain spatiotemporal bounds). When they stop moving, subjects can use the indexes to move their focal attention and then their gaze to each of the targets in turn. While foveating an indexed object observers can move the mouse and click on it, then move their gaze to the next indexed object and repeat.

A slight variation is needed if target objects are indicated by a property that does not automatically draw indexes (e.g., if the targets are vertical lines while nontargets are horizontal lines). We have evidence that in that case a spotlight-of-attention has to visit each of these cued targets in turn and "drop off" an index (Pylyshyn & Annan, in press). There are other findings that tell us more about the nature of these FINST indexes. For example, people do not notice when targets (or nontargets) change color or shape, they have a great deal of trouble recalling which target was which when they are identified with names or numbers at the start of the trial, and they are able to track targets even when thee targets disappear briefly but completely behind occluding barriers. When observers make errors these tend to consist in swapping the identity of one object with that of another object that is close to it, and the chances of such swaps is higher between pairs of targets than between targets and nontargets. There is also evidence that the reason for this asymmetry in swapping errors between target-target pairs and target-nontarget pairs is that nontargets are inhibited during a tracking trial (Pylyshyn, 2004, 2006).

The story of how basic MOT is carried out in terms of FINST theory is extremely simple, partly because the MOT task was designed to reflect the FINST hypothesis in a fairly direct way. But there are other findings that are not accounted for without some finer-grained assumptions about how FINSTs work. Moreover, there is more to FINST indexing than is revealed in the above story. We assume that FINSTs constitute a very general mechanism that is used not only for tracking simple elements moving on a screen, but that it also functions to allow people to keep track of things in the world. The ability to *track* things has long been recognized as an essential ingredient in identifying individual things and so the question of what our visual system treats as a thing (an individual or an object in some sense) is extremely important. Thus some of the assumptions we have made about

FINSTs have extremely far-reaching implications for how our visual system deals with individuals, properties and other aspects of the contact between mind and world. What I have found over the last several years of trying to explain to psychologists and philosophers what I think is going on, is that finding the right way to describe the empirical phenomena and explaining what they mean in a more general framework is far from an easy task. What I will do very briefly in the next section is give you a version of the story that suggests what the FINST idea might mean for the connection between mind and world. Because quite a few pieces of this puzzle are still missing I will have to go out on a limb now and then and fill in with some speculation.

5. Viewing FINSTs as nonconceptual links between mind and world

The basic motivation for postulating indexes is that, as we saw at the beginning of this essay, there are both empirical and theoretical reasons for assuming that a small 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 FINST 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. In addition what is needed is 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 objects in the world by a mechanism that binds them to individual clusters. Moreover, the clusters must be such that the representation can continue to refer to the 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 (Yantis, 1998) and Scholl (Scholl & Pylyshyn, 1999) results. They also include the requirement that the object being tracked be a perceptual whole as opposed to some arbitrary, but well defined, set of features (Scholl, Pylyshyn, & Feldman, 2001).

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 FINST indexes; (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 perceptually the same objects); and (5) only indexed objects can enter into subsequent cognitive processes, such as recognizing their individual or relational properties, or moving focal attention or gaze or making other motor gestures to them.

The basic idea of the FINST indexing and binding mechanism is illustrated in Figure 10. 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 object are detected they are associated with the index that points to that object. This, in effect, provides a mechanism for connecting objects of an evolving representation with objects in the world (stored temporarily in the Object Files mentioned earlier). 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. What is important to note here is that the inward arrows are purely causal and are instantiated by the non-conceptual apparatus which, following the terminology suggested by (Marr, 1982), I refer to as *early vision* (Pylyshyn, 1999). 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 a visual 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.

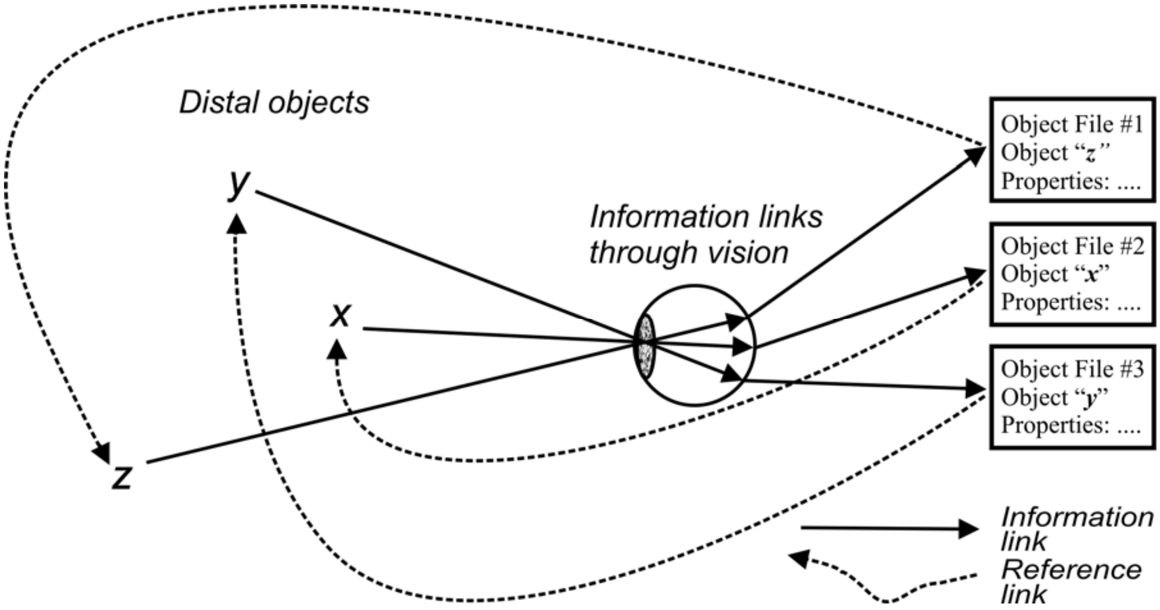

Figure 10: Sketch of the types of connections established by FINST indexes between the visual objects and parts of conceptual structures, depicted here as Object Files. Such a mechanism would clearly have applicability to everyday tasks such as monitoring players in team sports, as illustrated in the color insert (using hockey as an example).

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 (Koch & Ullman, 1985) can implement such a function (the application of the Koch & Ullman network to FINST Index theory has been explored in Acton, 1993; Pylyshyn & Eagleson, 1994). Another implementation uses an oscillatory neural network (and uses separate layers for each object, see Kazanovich & Borisyuk, 2006). All that is required is some form of 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 Appendix 5A of Pylyshyn, 2003). What is important about such a network, which makes its indexing function essentially pre-conceptual, is that the process that sends the probe signal to a particular place *uses no* *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 Treisman & Gelade, 1980; or Wolfe, Cave, & Franzel, 1989). What makes this system object-based, rather than location-based, is certain provisions in the network that ensures that a smoothly-moving object is tracked as the *same object* (e.g., this can be done by lowering the threshold of the units in the immediate neighboring in the selected unit) which results in the FINST moving with the movement of the selected visual object (for details, see Koch & Ullman, 1985; Pylyshyn, 2003, Chapter 5).

What I have described is a mechanism for picking out, tracking and providing *cognitive access* to what I have been calling *visual objects*. The notion of an *object* is ubiquitous in cognitive science, not only in vision but much more widely. It is also a foundational concern in metaphysics. But for present purposes I will take for granted that the world consists of physical objects. The view I have been proposing assumes that the visual system (or at least that part of it that we refer to as *early vision*) is encapsulated, that it is a module that works autonomously and independently of cognition. The view also relies on 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. The latter conclusion is also supported by evidence from clinical neuroscience, where it has been shown that deficits such as unilateral neglect (Driver & Halligan, 1991) or Balint syndrome (Robertson, Treisman, Friedman-Hill, & Grabowecky, 1997) 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.

My focus has been on *visual objects* – objects that are selected by the visual system without benefit of concepts and knowledge. Although I have mentioned psychophysical experiments, including Multiple Object Tracking, there are a number of findings in cognitive development that are relevant to our notion of object and index. For example, the notion of object has played an important role in the work by (Leslie, Xu, Tremolet, & Scholl, 1998; Spelke, Gutheil, & Van de Walle, 1995; Xu & Carey, 1996) 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 FINST 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 (Leslie, Xu, Tremolet, & Scholl, 1998) 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 been taken to suggest that registering the individuality of objects developmentally precedes recognizing objects by 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 is unclear whether all these uses of the term "object" is psychology mean the same thing. My present use of the term is inextricably connected with the theoretical mechanism of FINST 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 Elizabeth Spelke and others as a “bounded, coherent, three-dimensional physical object that moves as a whole” (Spelke, 1990). Are such Spelke-objects different from what we have been calling primitive visual objects?

My provisional answer to the question of the relation between these two notions of object is that primitive visual objects are in fact a subset of real physical objects (which subsume most 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 criteria for being an object (or perhaps for being a Spelke object, which is a subset of physical objects). If that is the case, then it suggests that, contrary to claims made by developmental psychologists (Spelke, Gutheil, & Van de Walle, 1995; Xu, 1997), the *concept* of an object is not involved in picking out these visual objects, just as no concept (i.e., no description) plays a role in 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 a direct demonstrative reference or *index* is involved at least in some of the phenomena — see (Leslie, Xu, Tremolet, & Scholl, 1998). However, there also appear to be cases in which clusters of features that one would expect would be perfectly good objects from the perspective of their visual properties, may nonetheless fail to be tracked as objects by 8 month old infants. Chiang and Wynne (Chiang & Wynn, 2000) have argued that *if the infants are given evidence that the things that look like individual objects are actually collections of objects* then they do not keep track of them in the studies involving placing objects behind a screen, despite the fact that the do track the visually-identical collections when this evidence is not provided. For example if infants see the apparent objects being disassembled and reassembled , or if they see the them come into existence by being *poured from a beaker* (Carey, 1999) 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), or 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 that ability to track things in psychophysical experiments is also sensitive to the way they appear and disappear, as well as the pattern by which they move. Several studies have shown that the precise *manner* in which objects disappear and reappear matters to whether or not they continue to be tracked (Scholl & Pylyshyn, 1999). In particular if their disappearance is by a pattern of accretion such as occurs when the object goes behind an occluding surface, and reappears in a complementary manner (by disocclusion) then it continues to be tracked in a multiple-object tracking paradigm. But the of effect of recent visual history is quite plausibly subsumed under the operation of a non-conceptual mechanism of the early vision system. This is consistent with the story I have been telling about how objects are selected and tracked since I have not said what the time-frame or temporal window is within which object-properties are effective either in index-grabbing or in tracking, so the immediate history (of being put down or being poured) may well be part of what determines whether the thing qualifies as a visual object (for other examples of what appear on the surface as knowledge-based phenomena but which can be understood as the consequence of a nonconceptual mechanism, see Pylyshyn, 1999).

The central role that objects play in vision has another, perhaps deeper, consequence worth noting. The primacy of objects as the focus through which properties are encoded suggests a rather different way to view the role of objects in visual perception and cognition. Just as it is natural to think that we apprehend properties such as color and shape as *properties of objects,* so has it also been natural to think that we recognize and encode 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 which that object satisfies. Perhaps this function can be served in part by the mechanism I referred to as a FINST index or a visual demonstrative (or a FINST).

Notice that what I have been describing is not the full concept 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 (Hirsch, 1982). 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 what are picked out in this way are typically 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. FINST indexes provide the needed grounding for basic objects – the individuals to which perceptual predicates apply, and hence about which cognitive judgments and plans of action are made. 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 the *THINGS* selected by *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 (of course the motor system eventually need coordinates, but that function is established further downstream, rather than being part of the command issued by the cognitive system).

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.

An alternative explanation of Multiple Object Tracking

In his chapter in this volume Brian Scholl raises an objection to my account of the multiple object tracking experiment. The alterative proposal is that tracking utilizes split attention so no visual indexes or FINSTs are needed. Because this is a perennial objection that I hear almost every time that I present this work I thought it might be worthwhile briefly to address this alternative proposal

Tracking objects and tracking sets. It is of course obvious that you don't have to remember a target's history going back to the beginning of a trial in order to track it. But even though we need not encode the history, the decision to call a particular object token a target connects to its role in the preceding instant and that, in turn, connects through a chain of individuals and sets to the initial state. Having inferred that a particular object token is a target we can then "flush" the basis for that inference and move on to the next instant in time, just as Brian Scholl says.

But that leaves a puzzle : how do you know whether a particular object had been a target in the immediately preceding instant without tracking it? According to the alternative account this is done by determining whether it was a member of the *set of targets*. Here everyone seems to assume that we can keep track of a set without keeping track of its individual members. But how can we do that? Sometimes there are ways to do this because the set has properties that the individual members do not have, either because they are aggregate properties or they are relational properties. For example we might be able to identify the targets by identifying the set to which they belonged if all the targets were in the top right quadrant of the screen, or they traveled in a rigid configuration. The most popular proposal of this sort is due to Steve Yantis (Yantis, 1992) who proposed that we could treat the set as a whole by imagining the targets being connected by an elastic band that forms a polygon – then we could track a single *distorting polygon* rather than the individual targets that form its vertices.

The trouble with polygon-tracking and related methods is that they only work if at each instant you *already know* (i.e., have some way to distinguish) which objects are the targets and therefore constitute the vertices of the polygon. The imagined elastic does not automatically wrap around the targets as it would if it was a real elastic attached to real objects; it only does so if you know which objects are the targets and wrap them accordingly. Since the objects in MOT move in unpredictable independent trajectories, then in order to keep the elastic wrapped around the targets rather than be taken over by the identical moving nontargets, we would have to first distinguish the individual targets from the nontargets. While Brian may not wish to subscribe to that particular model of MOT, he does require a similar sort of mechanism that only keeps track of the objects as a set, rather than tracking the individual objects that constitute the set. It is this desideratum that leads him to propose that tracking is purely a phenomenon of divided attention. You place an attention beam on each target so each target is tracked individually. But if you now add the novel (and gratuitous) assumption that attention beams are indistinguishable, you get tracking-by-sets without access to individual targets. (It is not clear, by the way, why one couldn't add the same assumption to the FINST version, but it's not one that has an independent motivation).

Failing to recall a name associated with a target. As Brian points out, the set-tracking hypothesis fits well with our own data (Pylyshyn, 2004) showing that recalling a particular identifier (e.g., a number or name) that had been associated with a target is much harder than simply recalling that it had been a target. In examining this finding we found evidence that the chance of attributing a particular target identifier to the wrong target was significantly higher than attributing it to a nontarget. We postulated that this asymmetry was due to the inhibition of nontargets – a hypothesis for which we subsequently found independent evidence. I now believe that there is very likely more going on in this surprising phenomenon than just index-switching. But as Brian points out, such ID errors seem natural on the account of MOT that assumes that we track sets through split (and unmarked beans of) attention and thus fail to distinguish among members of the set. However, we pay a heavy price for this naturalness since any set-tracking option not only fails to distinguish among the targets, it is missing the notion of individual entirely and so cannot account for the wide range of empirical phenomena I discussed in my chapter (as well as in chapters 4 and 5 of Pylyshyn, 2003). In addition, since one of the main functions of focal attention is to allocate resources in order to facilitate property detection, one would not expect the tracking task to be so insensitive to object properties (as reported in Bahrami, 2003; Scholl, Pylyshyn, & Franconeri, 1999).

Recall the many purposes for which FINST indexes were postulated – including distinguishing parts in recognizing patterns (using “visual routines”) and solving the “binding problem” (i.e., determining when several visual features are features of the same object). If you cannot distinguish the different attention beams you cannot associate a property with a particular object (as in the study of object-specific priming, see Noles, Scholl, & Mitroff, 2005). Such faceless attention beams appear to be little more than FINSTs without token distinctiveness or the pointer function. If you allow them to have these functions then you have FINSTs by another name – a name that, unfortunately, merges them with focal attention and so misses the special feature of FINSTs, such as their failure to encode object properties and their important non-conceptual nature. While many psychologists may not care about the latter, it is an issue that has been preoccupying me more in recent years (and which I address in Pylyshyn, 2007). It’s also the sort of issue that Cognitive Science, as an interdisciplinary pursuit, was intended to address.

References

Acton, B. (1993). *A Network Model of Visual Indexing and Attention.* Unpublished MSc Thesis, University of Western Ontario, London, Canada.

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

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

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

Annan, V., & Pylyshyn, Z. W. (2002). Can indexes be voluntarily assigned in multiple object tracking? *Journal of Vision, 2*(7), 243a.

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

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

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

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

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139-159.

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

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

Campbell, J. (2002). *Reference and Consciousness*. New York: Oxford University Press.

Campbell, J. (2003). Reference as attention. *Philosophical Studies, ?*, 265-276.

Carey, S. (1999). *Establishing representations of new individuals: New infant results and old studies by Michotte.* Paper presented at the Object Cognition: Underlying mechanisms and their origins (May 20-21), Rutgers University, New Brunswick, NJ.

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

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

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

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences, 3*(9), 345-351.

Clark, A. (2000). *A Theory of Sentience*. New York: Oxford University Press.

Currie, C. B., & Pylyshyn, Z. W. (2003). *Maintenance of FINSTs across eye movements*. Unpublished ms available athttp://ruccs.rutgers.edu/~zenon/ccurrie/TitlePage.html

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

Egly, R., Driver, J., & 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.

Frohlich, W. D., & 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.

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

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

Hochberg, J. (1968). In the mind's eye. In R. N. Haber (Ed.), *Contemporary theory and research in visual perception* (pp. 309-331). New York: Holt, Rinehart & Winston.

Horowitz, T. S., Birnkrant, R. S., Fencsik, D. E., Tran, L., & Wolfe, J. M. (in press). How do we track invisible objects? *Psychonomic Bulletin & Review*.

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

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*(2), 307-317.

Jackson, F. (1997). *Perception: A representative theory*. Cambridge, UK: Cambridge University Press.

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

Kazanovich, Y., & Borisyuk, R. (2006). An oscillatory neural model of multiple object tracking. *Neural Computation, 18*(6), 1413-1440.

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

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

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

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

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

Liu, G., Austen, E. L., Booth, K. S., Fisher, B. D., Argue, R., Rempel, M. I., & Enns, J. T. (2005). Multiple-Object Tracking Is Based on Scene, Not Retinal, Coordinates. *Journal of Experimental Psychology: Human Perception and Performance, 31*(2), 235-247.

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

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard Univ Press.

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

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

Noles, N. S., Scholl, B. J., & Mitroff, S. R. (2005). The persistence of object file representations. Perception & Psychophysics, 67(2), 324-334.

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

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

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

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

Oksama, L., & Hyona, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition, 11*(5), 631-671.

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

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

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

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press (Also available through CogNet: http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=5602).

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

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

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

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

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

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

Pylyshyn, Z. W. (2006). Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving nontargets. *Visual Cognition, 14*(2), 175-198.

Pylyshyn, Z. W. (2007). *Things and Places: How the mind connects with the world (Jean Nicod Lectures Series)*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W., & Annan, V. J. (in press). Dynamics of target selection in multiple object tracking (MOT). *Spatial Vision*.

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

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

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

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

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

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

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

Reynolds, R. I. (1978b). The microgenetic development of the Ponzo and Zollner illusions. *Perception and Psychophysics, 23*, 231-236.

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

Schlottman, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *Quarterly Journal of Experimental Psychology A, 2*, 321-342.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1/2), 1-46.

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

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object: Evidence from target-merging in multiple-object tracking. *Cognition, 80*, 159-177.

Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (1999). *The relationship between property-encoding and object-based attention: Evidence from multiple-object tracking*. Unpublished manuscript.

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

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

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

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

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

Sperling, G., & Weichselgarter, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychological Review, 102*(3), 503-532.

Strawson, P. F. (1963). *Individuals: An essay in descriptive metaphysics*. New York: Anchor Books.

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

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

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

Treisman, A. (1995). Modularity and attention: Is the binding problem real? In C. Bundesen & H. Shibuya (Eds.), *Visual selective attention*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Trick, L. M., & 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.

Tsotsos, J. K. (1988). How does human vision beat the computational complexity of visual perception. In Z. W. Pylyshyn (Ed.), *Computational Processes in Human Vision: An interdisciplinary perspective* (pp. 286-340). Norwood, NJ: Ablex Publishing.

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

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

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

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

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

Wolfe, J. M., Cave, K. R., & 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.

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

Xu, F., & Carey, S. (1996). Infants' Metaphysics: The case of numerical identity. *Cognitive Psychology, 30*, 111-153.

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

Yantis, S. (1998). Objects, Attention, and Perceptual Experience. In R. Wright (Ed.), *Visual Attention* (pp. 187-214). Oxford,GB: Oxford University Press.

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

Notes

[1] Even visual concepts, like perceived shape, cannot be specified in terms of transducer outputs (see Pylyshyn, 2003, Chapter 1). Julian Hochberg spent years searching for the geometrical basis for pattern complexity but gave up on the grounds that it was the form of the representation and not the form of the objective stimulus that mattered (Hochberg, 1968).

[2] For details see Pylyshyn (2003).) and the experimental reports cited there or in more recent reports such as: (Pylyshyn, 2004, 2006; Pylyshyn & Annan, in press).

[3] 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 is no longer in view. In that respect it functions like a demonstrative, which is why I continue to call it that, even as I use examples involving names like α.

[4] There have been well over a hundred studies in our laboratory (Annan & Pylyshyn, 2002; Blaser, Pylyshyn, & Holcombe, 2000; Keane & Pylyshyn, 2006; Pylyshyn, 2004, 2006; Pylyshyn & Annan, in press; Scholl, Pylyshyn, & Feldman, 2001) as well as in other laboratories(Allen, McGeorge, Pearson, & Milne, 2004; Alvarez, Arsenio, Horowitz, & Wolfe, 2005; Alvarez & Scholl, 2005; Bahrami, 2003; Cavanagh, 1992; Cavanagh & Alvarez, 2005; Chiang & Wynn, 2000; Horowitz, Birnkrant, Fencsik, Tran, & Wolfe, in press; Liu, Austen, Booth, Fisher, Argue et al., 2005; Ogawa & Yagi, 2002; O'Hearn, Landau, & Hoffman, 2005; Oksama & Hyona, 2004; Suganuma & Yokosawa, 2002; Trick, Perl, & Sethi, 2005; vanMarle & Scholl, 2003; Viswanathan & Mingolla, 2002; Yantis, 1992) that 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. In a set of unpublished studies (Scholl, Pylyshyn, & Franconeri, 1999) 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. This was confirmed independently by (Bahrami, 2003) who showed that observers cannot detect changes in color or shape on either nontargets or targets while tracking.