CONNECTING VISION WITH THE WORLD: TRACKING THE MISSING LINK*

Zenon Pylyshyn

Rutgers Center for Cognitive Science

Rutgers University, New Brunswick, NJ

You might reasonably surmise from the title of this paper that I will be discussing a theory of vision. After all, what is a theory of vision but a theory of how the world is connected to our visual representations? Theories of visual perception universally attempt to give an account of how a proximal stimulus (presumably a pattern impinging on the retina) can lead to a rich representation of a three dimensional world and thence to either the recognition of known objects or to the coordination of actions with visual information. Such theories typically provide an effective (i.e., computable) mapping from a 2D pattern to a representation of a 3D scene, usually in the form of a symbol structure. But such a mapping, though undoubtedly the essential purpose of a theory of vision, leaves at least one serious problem that I intend to discuss here. It is this problem, rather than a theory of vision itself, that is the subject of this talk.

The problem is that of connecting visual representations to the world in a certain critical way. This problem occurs for a number of reasons, but for our purposes I will emphasize just one such reason: the mapping from the world to our visual representation is not arrived at in one step, but rather it is built up incrementally. We know this both from empirical observations (e.g., percepts are generally built up by scanning attention and/or one's gaze) and also from theoretical analysis (e.g., Ullman, 1984) has provided good arguments for believing that some relational properties, such as the property of being inside or on the same contour) have to be acquired serially by scanning a display. Now here is one problem that arises immediately. If the representation is built up incrementally, we need to know that a certain part of our current representation refers to a particular individual object in the world. The reason is quite simple. As we elaborate the representation by uncovering new properties of a scene that we have partially encoded we need to know where (i.e., to which part of the representation) to attach the new information. In other words we need to know when a certain token in the existing representation should be taken as corresponding to the *same individual object* as a particular token in the new representation, so

that we can attribute newly noticed properties to the representation of the appropriate individual objects.

Take a concrete example. Suppose the representation of a scene takes the form of a conceptual structure whose content corresponds roughly to that of the English sentence fragment, "… four lines forming a parallelogram, with another similar parallelogram directly below it, arranged so that each vertex of the top parallelogram is connected by a straight line to the corresponding vertex of the parallelogram below it." Although we can infer that there are 12 lines in this figure, we don't have a way to refer to them individually. We can't say which lines are referred to in the first part of the description ("…four lines forming a parallelogram"), which lines are the ones that connect the two parallelogram, and so on. Without identifying particular lines we could not add further information to elaborate the representation. If, for example, on further examination, we discover that certain of the lines are longer than others, some are colored differently, some vertices form different angles than others, and so on, we would need to connect this new information to representations of particular objects in the interim representation. Conjunctions of properties (e.g., red, right-angled, lower, etc) are defined with respect to particular objects, so individual objects must be identified in order to determine whether there are property conjunctions. The question is, how can a representation identify particular objects?

Let's look at this example more closely. The content of the descriptive sentence in the above example might refer to the Necker cube shown on the left in Figure 1 (where the parallelograms in question are the figures EFGH and ABCD). Now suppose that at some point you notice, as most people do sooner or later, that face labeled FGBC is a square that appears to lie in front of (i.e., is closer to the viewer than) the square EHAD. How would we add that information to a representation whose content is like that of the sentence quoted earlier? In order to add information to a representation we need to relate the information to representations of particular elements in the figure. That's why in this example, as using diagrams in general, we label lines or vertices. Whatever form the visual representation takes it must allow the recovery of particular individual objects or parts referred to in that representation much as though they were labeled. What constraints does that impose on a representation? Can a purely descriptive representation (i.e., a description with quantifiers but no names or singular terms) do? This is a question that gets into much deeper issues than ones I can address in any detail here. Yet it needs to be

addressed at least briefly insofar as I will argue that visual representations need something like demonstratives or names in order to allow incremental elaboration (and for other reasons as well).

Common forms of representations of a simple figure such as a Necker Cube are shown in Figure 1. In order to be able to augment the description over time it would be necessary to pick out particular token objects (lines or vertices) that appear in the representation. Assuming that we have not labeled every relevant point in the figure (after all, the world does not come conveniently labeled), a possible way in which a purely descriptive representation could pick out individuals is by using definite descriptions. It could, for example, assert things like "the object *x* that has property P" where P uniquely picks out a particular object. In that case, in order to add new information, such as that this particular object also has property Q one would add the new predicate Q and also introduce an identity assertion, thus asserting that $P(x) \wedge Q(y)$ and $x \equiv y$ (and, by the way, adding this new compound descriptor to memory so that the same object might be relocated in this way when a further new property of that object is later noticed).[1] But this is almost certainly not how the visual system adds information. This way of adding information would require adding a new predicate Q to the representation of an object that is *picked out by a certain descriptor*. To do that would require first recalling the description under which *x* was last encoded, and then conjoining to it the new descriptor and identity statement. Each new description added would require retrieving the description under which the object in question was last encoded.

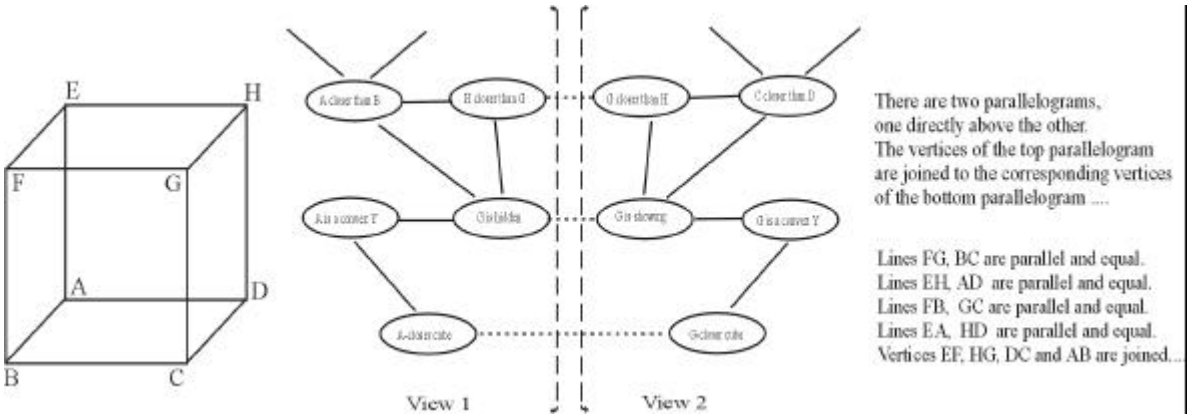

Figure 1. Some common forms of representation of a figure such as the reversing cube on the left, assuming that vertices are labeled. How could you represent the figure if it was not labeled?

The alternative to this unwieldy method is to allow the descriptive apparatus to make use of singular terms or names or demonstratives.[2] If we do that, then adding new information would amount to adding the predicate Q(*a*) to the representation of a particular object *a*, and so on for each newly noticed property of *a*. Empirical evidence for object-based attention (see the discussion in the last section of this paper and, e.g., Baylis & Driver, 1993) suggests that the visual system's property detectors (e.g., Q-Detectors) recognize instances of the property Q *as a property of a particular visible object*, such as object *a*, so that this is the most natural way to view the introduction of new visual properties by the sensorium. In order to introduce new properties in that way, however, there would have to be a non-descriptive way of picking out *a*, such as a singular term or a name or a demonstrative. This is, in effect, what labeling objects in a diagram does through external means and what demonstrative terms like "this" or "that" do in natural language.[3] This alternative is *prima facie* the more plausible one since it is surely the case that when we detect a new property we detect it as applying to *that* object, rather than as applying to some object in virtue of its being the object with a certain (recalled) property.[4] Such intuitions, however, are notoriously unreliable so later in this paper I will examine empirical evidence which suggests that this view is indeed more likely to be the correct one. For example, I will describe studies involving multiple-object tracking that make it very unlikely that objects are tracked by regularly updating a description that uniquely picks out the objects. In these studies the only unique descriptor available is location, and under certain plausible assumptions the evidence shows that it is very unlikely that the coordinates of the points being tracked are being regularly updated so that tracking is based on maintaining identity by updating descriptions.

There are a number of other reasons why a visual representation needs to be able to pick out individuals the way demonstratives do (i.e., independent of their particular properties). For example, among the properties that are extracted (and presumably encoded in some way) by the visual system are a variety of relational predicates, such as **Collinear**(X_1, X_2, …X_n) or **Inside**(X_1,C_1) or **Part-of**(F_1,F_2), and so on. But these predicates apply over distinct individual objects in the scene independent of what properties these individuals have. So in order to recognize a relational property involving several objects we need to specify which objects are involved. For example, we cannot recognize the **Collinear** relation without 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 associate the relation with the objects in question. This is quite general since properties are predicated of things, and relational properties (like the property of being "collinear") are predicated of several things. So there must be a way, independent of the process of deciding which property obtains, of specifying which objects (in our current question-begging sense) have that property. Ullman, as well as a large number of other investigators (Ballard, Hayhoe, Pook, & Rao, 1997; Watson & Humphreys, 1997; Yantis, 1998; Yantis & Johnson, 1990; Yantis & Jones, 1991) talk of the objects in question as being "tagged" (indeed, "tagging" is one of the basic operations in Ullman's theory of visual routines). The notion of a tag is an intuitive one since it suggests a way of *marking objects* for reference purposes. But the operation of tagging only makes sense if there is 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 hold in the representation. But how do we tag parts of the world? What we need is what labels gave us in the previous example: A way to name or refer to individual parts of a scene *independent of their properties or their locations*.

What this means is that the representation of a visual scene must contain something more than descriptive (or pictorial — see Note 3) information in order to allow re-identification of particular individual visual elements. It must provide what natural language provides when it uses names (or labels) that uniquely pick out particular individuals, or when it embraces demonstrative terms like "this" or "that". Such terms are used to indicate particular individuals. This assumes that we have a way to *individuate* [5] *and keep track of particular individuals in a scene* even when the individuals change their properties, including their locations. Thus what we need are two functions that are central to our concern today: (a) we need to be able to pick out or individuate distinct individuals (following current practice, we will call these individuals *objects*) and (b) we need to be able to refer to these objects as though they had names or labels. Both these purposes are served by a primitive visual mechanism that I call a *visual index*. So what remains is for me to provide an empirical basis for the claim that the visual system embodies a primitive mechanism of the sort I call a *visual index*.

Individuating and tracking primitive visible objects: Multiple Object Tracking studies

Perhaps the clearest way to see what I mean when I claim that there is a primitive mechanism in early vision that picks out and maintains the identity of visible objects, is to consider a set of experiments, carried out in my laboratory, to which the ideas of visual individuation and identity maintenance were applied. The task is called the *Multiple Object Tracking (MOT) Task* and is illustrated in Figure 2.

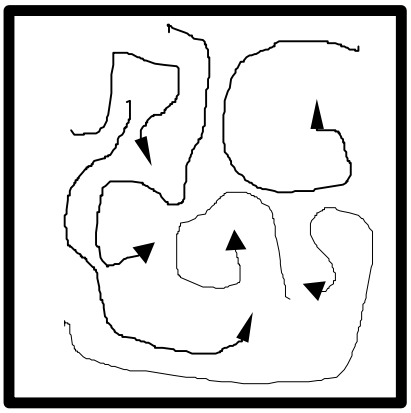

Figure 2: Typical trajectories of objects in the multiple-object tracking experiments (showing 6 objects in total).

In a typical experiment, subjects are shown a screen containing anywhere from 12 to 24 simple identical objects (points, plus signs, figure-eight shapes) which move across the entire visual field in unpredictable ways without colliding (a simplified version of which is illustrated in Figure 2). A subset of these objects is briefly rendered distinct (usually by flashing them on and off a few times). The subject's task is to keep track of this subset of objects (called "targets). At some later time in the experiment (say 10 seconds into the tracking trial) one of the objects is again flashed on and off. The subject must then indicate whether or not the flashed (probe) figure was one of the targets. A large number of experiments, beginning with studies by (Pylyshyn & Storm, 1988), have shown clearly that subjects can indeed track up to 5 independently moving identical. Moreover, we were able to argue that the motion and dispersion parameters of the original Pylyshyn & Storm experiment were such that tracking could not have been accomplished using a serial strategy in which attention is scanned to each figure in turn, storing its location, and returning to find the figure closest to that location on the next iteration, and so on. Based on some weak assumptions about how fast focal attention might be scanned and based on actual data

on how fast the objects actually moved and how close together they had been in this study, we were able to conclude that such a serial tracking process would very frequently end up switching to the wrong objects in the course of its tracking. This means that the moving objects *could not have been tracked using a unique stored description of each figure*, inasmuch as the only possible descriptor that was unique to each figure at any particular instant in time was its location. If we are correct in arguing from the nature of the tracking parameters that stored locations cannot be used as the basis for tracking then all that is left is the figure's identity or *individuality*. This is exactly what I claim is going on — tracking by maintenance of a primitive perceptual individuality.

Recently a large number of additional studies in our laboratory (Pylyshyn, 1998; Sears 1991, McKeever 1991; Scholl & Pylyshyn, in press) and elsewhere (Intriligator & Cavanagh 1992, (Yantis, 1992), and others) have replicated these multiple object tacking results using a variety of different methods, confirming that subjects can successfully track around 4 or 5 independently moving objects. The results also showed that merely widening one's breadth of attention (as assumed in the so-called zoom-lens model of attention spreading, Eriksen & St. James, 1986) would not account for the data. Performance in detecting changes to elements located inside the convex hull outline of the set of targets was no better than performance on elements outside this region, contrary to what would be expected if the area of attention were simply widened or shaped to conform to an appropriate outline (Pylyshyn et al., 1994). Using a different tracking methodology, (Intriligator & Cavanagh, 1992) also failed to find any evidence of a "spread of attention" to regions between targets. It appears, then, that items can be tracked despite the lack of distinctive properties (and, indeed when their properties are changing) and despite constantly changing locations and unpredictable motions. Taken together these studies implicate a notion of primitive visible object as a category induced by the early visual system, preceding the recognition of properties and preceding the evaluation of any visual predicate.

The multiple object tracking task exemplifies what I mean by "tracking" and by "maintaining the identity" of objects. It also operationalizes the notion of "primitive visible object" — a primitive visible object is whatever attracts a FINST index and allows multiple-object tracking. Note that this is a highly mind-dependent definition of objecthood. Objecthood and object-identity are defined in terms of a causal perceptual mechanism. A certain sequence of object-locations will count as the movement a single object if the early (pre-attentive) visual system groups it this way

— i.e., if it is so perceived — whether or not we can find a physical property that is invariant over this sequence and whether or not there exists a psychologically-plausible description that covers this sequence. The visual system may also count as one individual object certain kinds of disappearances and reappearances of visual objects. For example, Yantis has shown that when an object disappears either for a very short time *or* under conditions where it is seen to have been occluded by an opaque surface, the visual system treats the object as though it continued to exist. Similarly, Scholl & Pylyshyn (in press) have shown that if the objects being tracked in the MOT paradigm disappear and reappear in certain ways they are tracked as though they had a continuous existence and a smooth trajectory. If they disappear and reappear by deletion and accretion along a fixed contour, the way they would have if they were moving behind an occluding surface (even if the edges of the occluder are not invisible), then they are tracked as though they were continuously moving objects. Performance in the MOT task does not deteriorate if targets disappear in this fashion although it suffers dramatically if targets suddenly go out of existence and reappear, or if they slowly shrink away and then reappear by slowly growing again at exactly the same place as they had accreted in the occlusion condition.

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

I now take a few moments to review the theory of the Indexing mechanism for I intend it to serve a major function — that of providing the missing link alluded to in my title. The basic motivation for postulating indexes is that, as we saw at the beginning of this essay, there are a number of reasons for thinking that individual objects in the field of view must first be *picked out* from the rest of the visual field and the identity of these objects *qua individuals* must be maintained or tracked despite changes in the individual's properties, including its location in the visual field. Our proposal claims that this is done *primitively* without identifying the object through a unique descriptor. The object in question must be segregated from the background or picked out as an individual (the Gestalt notion of making a figure-ground distinction is closely related to this sort of "picking out"). Until some piece of the visual field is segregated and picked out, no visual operation can be applied to it since it does not exist as something distinct from the entire field.

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

What this means is that our theory is concerned with a sense of *picking out* and *tracking* that are not based on top-down *conceptual* descriptions, but are given pre-conceptually by the early visual system, and in particular by the FINST indexing mechanism. Moreover, the visual system treats the object so picked-out as distinct from other individuals, independent of what properties this object might have or whether the properties are changing in unpredictable ways. If two different objects are individuated in this way they remain distinct as far as the visual system is concerned. Moreover, they remain distinct despite certain changes in their properties, particularly changes in their location. Yet the visual system need not know (i.e., need not have detected or encoded) any of their properties in order to implicitly treat them as though they were distinct and enduring visual tokens. Of course there doubtless are properties, such as being in different locations or moving in different ways or flashing on and off that allow indexes to be assigned to these primitive objects in the first place. But none of these properties define the objects — they are not *essential properties*. What is an essential property is that, given the structure of the early visual system, the object attracted and maintained an index. My claim is that to index *x, in this primitive sensory sense*, there need not be any concept, description or sortal that picks out *x*'s by type[6]. The individuals picked out in this way by the early visual system (by a mechanism that I will describe below) are what I am referring to here as *primitive visible objects*. I use this technical terminology to distinguish these primitive visible objects from the more general sense of object, which might include invisible things, abstract things (like ideas) and other more usual notions of

object, such as tables and chairs and people — which writers like Hirsh (1982) and Wiggins (1979) and others have argued, *does* require sortal concepts to establish criteria of identity. My concern here will be with objects that are in the first instance defined in terms of the individuation (or clustering) and indexing mechanism of the early visual system, although this sort of individuation, I claim, must form the basis for full fledged individuation. The latter cannot be conceptual "all the way down" on pain of infinite regress. My claim, then, is that certain mechanisms of the early visual system lead to the automatic individuation of a small number of primitive visible objects and to the tracking of such individuals over certain sorts of changes of time and space.

The basic idea of the FINST indexing and binding mechanism is illustrated in the Figure 3 below. A series of proximal causes leads from certain kinds of visible events, via primitive mechanisms of the early visual system, to certain conceptual structures (which we may think of as symbol structures in Long Term Memory). This provides a mechanism of reference between a visual representation and what we have called primitive visible objects in the world. The important thing here is that the inward arrows are purely causal and are instantiated by the non-conceptual apparatus of what I have called *early vision* (Pylyshyn, in press). Under certain conditions this mechanism results in a link that exhibits a certain continuity or persistence, thus resulting in its counting as the *same link*. It is tempting to say that what makes it continuous is that it keeps pointing to the same thing, but according to our view this is circular since the only thing that makes it the *same thing* is the very fact that the it the index references it. There is no other sense of "sameness" so that "primitive visible object" as we have defined it is thoroughly mind dependent.

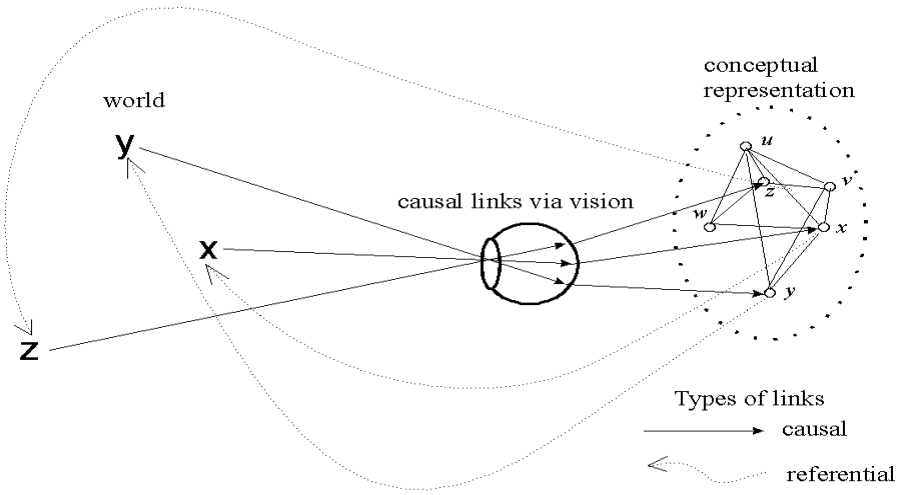

Figure 3: Sketch of the types of connections established by FINST indexes between primitive visible objects **x, y, z** *and parts of conceptual (descriptive) structures, depicted here as a network.*

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

What does all this do for connecting vision and the world?

What we have described is a mechanism for picking out, tracking and providing cognitive access to what we call an object (or, more precisely, a *primitive visible object*). The notion of an *object* is ubiquitous in cognitive science, not only is vision but much more widely. I might also note that it has been a central focus in developmental psychology where people like Susan Carey (this volume), Fei Xu (1997), Alan Leslie (Leslie, Xu, Tremolet, & Scholl, 1998) have studied "A child's concept of object". Similarly, many studies have shown that attention is allocated primarily by individual visual object, rather than in terms of regions (Baylis & Driver, 1993), a finding that is also supported by evidence from clinical neuroscience, where it has been argued that deficits such as unilateral neglect must be understood as a deficit of object-based attention rather than space-

based attention (Driver & Halligan, 1991). Time does not permit me to go into any of these fields although I am engaged in a larger project where I do examine the connections among these uses of the term "object".) But I would like to draw your attention to the fact that giving objects the sort of central role in vision that I have described suggests a rather different ontology. Just as it is natural to think that we apprehend properties such as color and shape as *properties of objects*, so it is also natural to think that we apprehend objects as a kind of property that particular *places* have. In other words we usually think of the matrix of space-time as being primary and of objects as being occupants of places and times. Everyone from Kant to modern cognitive scientist tacitly take this for granted — that's (in part) why it is so natural to think of mental images as having to be embedded in real space in the brain. Yet the findings I have described in the study of visual attention (as well as other areas of psychological research to which we will allude later) suggests an alternative and rather intriguing possibility. It is the notion that *primitive visible object* is the primary and more primitive category of early (preattentive) perception, so that we perceive objecthood first and determine location the way we might determine color or shape — as a property associated with objects.[7] 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 visible objects without using a unique description under which that object falls. This is the mechanism I referred to as a visual index or a visual demonstrative (or FINST).

Notice that when I am careful I hedge my use of the term *object* in making this claim, as I must because what I have been describing is not the notion of an object in the usual sense of an individual. An individual, as you all know, is a sortal concept whose individuation depends on assuming certain conceptual categories. But our notion does not assume the use of any concepts. The individuals that are picked out by the visual system and tracked primitively are something less than full blooded individuals. Yet because they are what our visual system gives us through a brute causal mechanism — because that is its nature — it serves as the basis for all real individuation. As philosophers like Wiggins (1979) and Hirsh(1982) have argued, you cannot individuate objects in the full blooded sense without a conceptual apparatus — without sortal concepts. But similarly you cannot individuate them with *only* a conceptual apparatus. Sooner or later concepts must be grounded in a primitive causal connection between thoughts and things.

The project of grounding concepts in sense data has not faired will and has been abandoned in cognitive science. However the principle of grounding concepts in perception remains an essential operation if we are not to succumb to an infinite regress. Visual indexes provide a putative grounding for basic objects and we should be grateful because without them (or at any rate something like them) we would be lost in thought without any grounding in causal connections with the real-world objects of our thoughts. With indexes we can think about things (I am sometimes tempted to call them *FINGs* since they are interdefined with *FINSTs*) without having any concepts of them: One might say that we can have *demonstrative thoughts*. And nobody ought to be surprised by this since we know that we can do this: I can think of this here thing without *any description* under which it falls. And, perhaps even more important, because I can do that I can reach for it.

Needless to say there are some details to be worked out so this is a work-in-progress. But I hope I have at least convinced you that there is a real problem to be solved in connecting visual representations to the world and that whatever the eventual solution turns out to be, it will have to respect a collection of facts some of which I have sketched for you today. Moreover any visual or attentional mechanism that might be hypothesized for this purpose will have far reaching implications, not only for theories of situated vision, but also for grounding the content of visual representations and perhaps for grounding perceptual concepts in general.

References

Acton, B. (1993). A Network Model of Visual Indexing and Attention, M.Sc. Thesis, *Dept of Electrical Engineering*. London, Canada: University of Western Ontario.

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**, 723-767.

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

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

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

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

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

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 Science,* **2**(1), 10-18.

Koch, C., & Ullman, S. (1984). Selecting one among the many: A simple network implementing shifts in selective visual attention. Cambridge, MA: MIT Technology Artificial Intelligence Laboratory.

McKeever, P. (1991). Nontarget numerosity and identity maintenance with FINSTs: A two component account of multiple object tracking. MA dissertation, University of Western Ontario, London, Ontario, Canada.

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

Pylyshyn, Z.W. (1998). Visual indexes in spatial vision and imagery. In R. Wright (Ed.), *Visual Attention* (pp. 187-214). Oxford, GB: Oxford University Press.

Pylyshyn, Z.W. (forthcoming). *Seeing: It's not what you think.* Book Ms in progress.

Pylyshyn, Z.W. (in press). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences.*

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

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

Scholl. B. & Pylyshyn, Z.W. (in press). Tracking multiple items through occlusions: Clues to visual objecthood. *Cognitive Psychology*, in press

Sears, C. (1991). Information processing at multiple locations in the visual field. MA Thesis, Dept of Psychology, University of Western Ontario, London, Ontario, Canada

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

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,** 90-122.

Wiggens, D. (1979). *Sameness and substance.* Oxford: Blackwell.

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.

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

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

Notes

[*] This paper is based on work-in-progress being conducted jointly with Jerry Fodor. I wish to acknowledge his contribution, as well as his critical reading of an earlier draft. The errors in this paper are not only mine, but are probably due to my refusing to accept his advice on certain points.

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

[2] Chris Peacocke has pointed out to me that calling this index-binding mechanism a *name* is misleading since names are used primarily to allow us to think about objects in their absence. The exact terminology that should be used in order to avoid misunderstanding is unclear. The term "demonstrative" implies a natural language context and an intention on the part of a speaker to demonstrate *that* object, which is not the case in the mental index I have in mind. It seems that the term "visual demonstrative" has been used with precisely the sense I have in mind, so I will henceforth confine myself to this terminology (along with the more technical phrase "index binding" which invokes the theoretical mechanism I have proposed and discussed elsewhere)..

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

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

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

[6] We are here claiming that there is a mechanism in the early (pre-conceptual) visual system that latches onto certain entities (perhaps I should say "events") for purely causal reasons, not because those entities meet conditions provided by a cognitive predicate — i.e., not because they constitute instances of a certain concept. In other words if P(x) is a primitive visual predicate of x then the x is assumed to have been independently and causally bound to what we have called a primitive visible object. Although this sort of latching or seizing by primitive visible objects is essentially a bottom-up process, this is not to say that it could not in some cases (perhaps in *most* cases) be guided by intentional processes, such as perhaps scanning one's attention until a latching event is located or an object meeting a certain description is found. For example, it is widely assumed (Posner, Snyder, & Davidson, 1980) that people can scan their attention along some path (by simply moving it continuously through space like a spotlight beam) and thereby locate certain sorts of objects. A possible consequence of such scanning is that an index may get assigned to some primitive objects encountered along the way.

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