Feature-Placing and Proto-Objects

Austen Clark
Department of Philosophy
103 Manchester Hall U-2054
University of Connecticut
Storrs, CT 06269-2054

Philosophical Psychology, 17(4), December 2004, 451-477.

Abstract. This paper contrasts three different schemes of reference relevant to understanding systems of perceptual representation: a location-based system dubbed "feature-placing", a system of "visual indices" referring to things called "proto-objects", and the full sortal-based individuation allowed by a natural language. The first three sections summarize some of the key arguments (in Clark 2000) to the effect that the early, parallel, and pre-attentive registration of sensory features itself constitutes a simple system of nonconceptual mental representation. In particular, feature integration--perceiving something as being both *F* and *G*, where *F* and *G* are sensible properties registered in distinct parallel streams--requires a referential apparatus. Section V. reviews some grounds for thinking that at these earliest levels this apparatus is location-based: that it has a direct and nonconceptual means of picking out places. Feature-placing is contrasted with a somewhat more sophisticated system that can identify and track four or five "perceptual objects" or "proto-objects", independently of their location, for as long as they remain perceptible. Such a system is found in Zenon Pylyshyn's fascinating work on "visual indices", in Dana Ballard's notion of deictic codes, and in Kahneman, Treisman, and Wolfe's accounts of systems of evanescent representations they call "object files". Perceptual representation is a layered affair, and I argue that it probably includes both feature-placing and proto-objects. Finally, both nonconceptual systems are contrasted with the full-blooded individuation allowed in a natural language.

I hope to draw your attention to a tiny, murky, but conceptually fascinating portion of vertebrate sensory processing. This domain starts at the transducer surfaces and ends (roughly) at the point at which attention is focused on some object the perception of whose properties requires integration of information across several processing streams. Processing prior to that point is often called "pre-attentive" processing. If for example we are talking about vision, then all these processes lie firmly within what is called "early" vision; but it all happens *before* your attention is attracted to something, or before you become *aware* of what you are seeing. It is a murky domain because by the time attention gets slewed around to them, these processes have already finished their job and are long gone.

Even within this murk I believe we can discern the outlines of psychological processes that have representational structure; and indeed different processes within the domain have representational structures of different and increasingly sophisticated kinds. The least sophisticated kind I call *feature-placing*. The name is in honor of Sir Peter Strawson's work on feature-placing languages, but the sensory scheme is pre-linguistic and non-conceptual. I will describe the highlights of that scheme; but then go on to explain how even in pre-attentive vision we can find representational structures that are a step up or two from bare feature-placing. In particular there is an intriguing discussion underway in vision science about reference to entities that have come to be known as "proto-objects" or "pre-attentive objects". These are a step up from mere localized features, and they have some but not all of the characteristics of "objecthood" or of "objective particulars".

It has become commonplace to think of visual representation as having multiple and distinct layers. Proto-objects will be the values of variables in one layer of visual representation. That layer is above feature-placing but below the apparatus of individuation and sortal concepts found in a natural language.

I. Features

"Feature" is a lovely word because it names a point of intersection of three different traditions. It is a phenomenal property; an attribute represented in a cortical feature-map; and a Strawsonian feature. To take these in turn.

The old-fashioned and austere notion of "phenomenal property" is a property of appearance: a characteristic of *how something appears.* To use C. D. Broad's (1927) term, the first goal in the study of sensory processes is try to understand the *facts of sensible appearance*: how something looks, or feels, or smells, or tastes. To understand these facts of sensible appearance we have to understand what similarities and differences the creature senses among the stimuli that confront it. Stimulus characteristics alone do not determine these phenomenal similarities and differences. Similarities of color cannot be read off from similarities of wavelength; differences in pitch are not simply differences in frequency; thermal perception is not a matter of registering the surrounding temperature. Instead we must, in a certain sense, enter into the "point of view" of the sentient organism, so as to understand how its sensory systems are constituted to treat some classes of stimuli as relatively similar, and others as relatively different. In a story that is now relatively well known, these phenomenal similarities and differences, and relative similarities and differences, can, if conditions are auspicious, provide us with the materials to construct a quality space, in which distance between two points corresponds to the relative similarity of the qualities they present.

The quality space will have some finite number of dimensions, corresponding to the number of independent dimensions of variation that the creature can sense among the stimuli that confront it in that modality. The metric for such a space is *not* provided by stimulus characteristics such as wavelength, pitch, or temperature; but by the degree to which those various stimuli happen to be perceived, by the creature in question, *as* being similar. A "phenomenal property" is, on this reading, a determinate value in some dimension of variation in sensory appearance. A fully determinate phenomenal property will correspond to a point in quality space.

It might be possible in some modalities to determine introspectively the number of dimensions available in the quality space. But if it is possible, it is very difficult. For example, it is only if one restricts stimuli to surface colors presented in carefully controlled laboratory conditions that variations in apparent surface color can be described in just three dimensions, such as hue, saturation, and brightness. Even then, the mere dimensionality of a space does not determine which axes are used to construct it. The axes in terms of which a sensory system actually transacts its business are *not* introspectively accessible. So it is possible to be surprised about some aspects of the phenomenal structure of our own experience. In fact we have been treated to some surprises: we learn, for example, that surface colors are (in places) registered using an axis that runs from yellow through gray to blue; another from red through gray to green; and a third from white to black. If this pans out, then a particular determinate value of "degree of yellowness/blueness" will be identified as a phenomenal property of surface color.

Biophysicists and sensory physiologists call such a determinate value a sensory (or perceptual) *feature.* Once we begin to understand the particular axes of discriminability which a sensory system actually uses, we can begin to explain why appearances in that modality are, for the creature in question, structured as they are. There has been *enormous* progress on this front in the last twenty years. A theme common to many different vertebrate sensory systems is the discovery of cortical "feature maps".

Feature maps have two identifying characteristics: first, one finds cells of some identifiable kind in some cortical region that respond to some particular dimension or dimensions of variation in phenomenal appearance. (States of the cells are states registering those variations.) Second, they are called "maps" because one finds that those cells are arranged in roughly topographical (here, retinotopic) order, so that adjacency relations are more or less respected. As Francis Crick and Christof Koch put it:

> Different cortical areas respond, in general, to different features. For example, the neurons in area MT are mostly interested in motion and depth, those in area V4 in color and shape, and those in 7a in position in space relative to the head or the body. [282] ... if you are currently paying attention to a friend discussing some point with you, neurons in area MT respond to the motion of his face, neurons in V4 respond to its hue, and neurons in auditory cortex ... respond to the words coming from his face (Crick & Koch 1997, 282-284)

How common are these feature maps? They are rife. Typically there are many feature maps per modality; in vision there are certainly "pathways", if not separate maps, for colour, motion, fine form, spatial frequency, stereoscopic depth, and other visual attributes, yet to be identified. Even within 'one' dimension of variation in phenomenal appearance such as colour, one finds multiple feature maps at different levels within the neuroanatomical hierarchy: blob cells in V1, thin stripes in V2, projections to V4, and so on (Davidoff 1991, 20-5). By one count, in the macaque monkey we have now identified thirty two distinct visual areas (see Felleman and Van Essen 1991). Larry Weiskrantz has dubbed this array the "oil refinery". Where in the oil refinery does the crude oil of *information* get distilled into the elusive subjective qualities of conscious sensation? Stay tuned!

In talking about feature maps in the past I have tended to emphasize Anne Treisman's "feature integration" model of attention. This is a very influential and well-known model of visual selective attention; it has attracted lots of research, and undergone significant revisions over the years. But it is important to note that feature-placing is *not* dependent upon this particular model; it can be drawn from even simpler and more pervasive background assumptions. These are assumptions that Treisman's model shares with others that are contrary to hers. (So they provide a good place on which to focus our efforts in philosophy of psychology!) They are, roughly:

> (1) that the initial stages of visual processing proceed in largely independent and parallel streams, working in a bottom-up and automatic fashion. Each such stream constructs a representation of a large portion (perhaps all) of the visual field; and

> (2) that these initial stages are followed at some point by a second set of limited capacity processes, whose activities at any one time are limited to small portions of the visual field. A compendious view hence requires these processes to proceed sequentially, from location to location.

Ulric Neisser suggested these ideas 37 years ago; and they are now a common background assumption of many different theorists. Treisman shares this inheritance. Her model includes a collection of independent feature maps, with each indicating incidence of its own favorite features over the entire visual field. The limited capacity process is spatial selective attention. It has the job of "binding" together the features represented in different feature maps when those features are represented as characterizing the same place-time. Because selective attention has limited capacity, tasks that require such binding require a serial search, location by location.

Suppose we start to modify the feature integration theory of attention by changing assumptions in it that have proven to be empirically problematic. (Treisman herself has done this; her model is quite different now than it was in 1980!) For example, we drop the claim that all feature conjunctions require attention; conjunctions of color and stereoscopic depth can give pop-out. We drop the claim that there is a dichotomy between "serial" and "parallel" search, allowing instead that searches can vary continuously in efficiency. Those that are very efficient have minimally increasing reaction times as distractors increase; those that are very inefficient have sharply increasing reaction times as we add distractors. But in between you might have all sorts of different slopes. The result is something like Jeremy Wolfe's "guided search" model (see Wolfe 1994, 1996a). It has the same underlying architecture, however. We have multiple independent streams, extracting features in different dimensions, and representing incidence of those features over large portions of the visual field in distinct and independent "feature maps". Those multiple feature maps are then followed by a limited capacity process, which has the job of integrating results over the collection.

Even the less portentous vocabulary of distinct visual *streams* can land us in difficulties. Suppose you believe that parvocellular and magnocellular streams carry information about distinct sets of visual attributes (which they do: fine form and color vs. motion and luminance contrast). Suppose you also believe that some perceptual episodes require contributions from both streams (which some do: seeing a red fire truck in motion, for example). Then you too face some variant of the "property binding" problem.

All these research programs share the idea of "feature maps". In any such model a feature is a dimension of discriminability among appearances. A feature map indicates the incidence of values of one or more of such dimensions across the ambient optic array; and these maps are typically organized retinotopically; or at least in an observer-relative way.

II. Placing

What then is the problem that these two background assumptions allegedly lead us into? One might think that a sufficiently compendious catalog of sensory features could exhaust the content of the act of sense. This assumption is shared by anyone who thinks a mental life confined to sensation is nothing but a stream of sensory qualities, a flux of qualia. We could characterize its content using nothing more than a bunch of general terms and conjunction. At one moment the stream of experience registers "brown and wet and slippery and cold" but a happier moment is "green and dry and warm and sweet-smelling." Now I think this old picture contains a big mistake, one which is still doing damage in neuropsychology. Not only do we need more terms than just feature terms; we need a totally different *kind* of term--one which has logical and semantic properties distinct from those of any possible feature term.

Why do we need anything more than features? Why can't we treat appearances of location in just the same way that we treat appearances of hue, brightness, or texture? I argue that if we did, we would run into an insoluble problem whenever we try to represent one thing as having more than one feature--for example, represent one patch as being both red and stippled. Such a feat cannot be managed if sensory location is treated just like all the other features.

The problem of how to represent one thing as having more than one feature has (at least) three different names in the three different disciplines of philosophy, psychology, and neuroscience. In philosophy it has been most recently called (by Frank Jackson) the "Many Properties" problem, though it has earlier variants in Kantian "synthesis" or the Lockean "compounding" of simple ideas. Psychologists call it "feature integration". Neuroscientists call it "binding". (It is as if the three language communities, each stuck in their own alpine valley, have three different names for the same high peak.) Any solution to this problem, I argue, requires a distinction in kind between features and their apparent locations. Terms for features and terms for places must play fundamentally distinct and non-interchangeable roles, because otherwise one could not solve the binding problem.

Let us start with the philosophical version. In his book *Perception*, Frank Jackson (1977, 65) discussed what he called the "Many Properties" problem: the problem of discriminating between scenes that contain all the same features, but differently arranged. Consider for example the perception of combinations of colors and textures: red vs. green, cross-hatched vs. stippled. We can readily distinguish between the following two scenes:

Scene 1: cross-hatched red next to stippled green
Scene 2: stippled red next to cross-hatched green

But what sorts of sensory capacity are required in order to discriminate one scene from the other? A creature equipped merely with the capacity to discriminate cross-hatched from stippled, and red from green, will fail the test, since both scenes contain all four features. Instead the creature must divvy up the features appropriately, and somehow register that scene 1 contains something that is both red and cross-hatched, while scene 2 does not.

In neuroscience the Many Properties problem has come to be known as the "binding" problem--or at least it yields one kind of the many different kinds of binding problem. Colour, motion, fine form, stereoscopic depth, and so on, are registered in distinct "feature maps" localized in different regions of cortex. How is it that activity in the various disjoint regions manages to represent one thing as being both *F* and *G*? This is perhaps the simplest and clearest kind of binding problem. It is sometimes called "property binding" (see Treisman 1996, 171); I shall focus exclusively on it.

Suppose we are trying to choose between Frank Jackson's Many propertied scenes. One is

cross-hatched red next to stippled green

while the alternative is

stippled red next to cross-hatched green.

The fact that we can distinguish between the two shows that our visual system can solve the Many Properties problem. Both scenes involve two textures and two colours, and so the simple capacities to discriminate textures or colours separately would not render the scenes discriminable from one another. Instead one must detect the overlap or co-instantiation of features: that the first scene contains something that is both stippled and red, while the second contains something that is both cross-hatched and red. The features of texture and colour are 'integrated', are perceived *as* features of one thing; to use the neuroscience term, there is a 'binding' of cross-hatched and red in the first scene, and of stippled and red in the second. Only the particular combinations (or bindings) of features differentiate the two scenes from one another.

What makes property binding possible? Careful consideration of this question will reveal the formal differences between the qualitative characteristics of sensory experience and the spatial ones.

Suppose that fine form and colour are carried by different streams within the visual nervous system. One registers that something is stippled, and a different one registers that something is red. The conclusion we seek is that something is both stippled and red**.** Formally, listing the contributions of our feature maps as premises, we have:

1. Something is red.
2. Something is stippled.

———————————————

Something is both red and stippled.

The inference is not valid; to make it truth-preserving, we must ensure that the stippled thing is the red thing, or at least that there is some overlap between stippled and red. Formally, using *x* and *y* as names for whatever is sensed as stippled or red respectively, we need to add a third premise:

1. x is red.
2. y is stippled.
3. $x = y$.

———————————————

Something is both red and stippled.

Now we have a valid inference, but to get it some variant of the third premise is essential. Unless we can identify the subject matter of the first premise with that of the second, we cannot logically secure the conclusion.

Such identification may be partial. That is, we do not need the stippled portion and the red portion of the scene to be perfectly coincident, or identical, as long as there is some overlap. If *x* is the stippled portion, and *y* the red, then it suffices that some part of *x* is identical to a part of *y*. This variant of the third premise allows the conclusion that something is both stippled and red. But notice that there is still an underlying identity required. We need some stippled *x* such that the very same *x* is red.

Now what exactly is this something such that the same something is both stippled and red? According to Anne Treisman's feature integration model, and all the work on the binding problem which has followed from it, what makes it possible to bind stippled and red together is that they are both features *of the same place-time.* As Treisman and Gelade put it:

> stimulus locations are processed serially with focal attention. Any features which are present in the same central 'fixation' of attention are combined to form a single object. Thus focal attention provides the 'glue' which integrates the initially separable features into unitary objects. (Treisman and Gelade 1980: 98)

Features scanned are bound together because at the time of scanning they characterize the same place. Identity of place-times drives the bindings: what gets bound to the stippled feature are whatever other features are found at the same spatio-temporal location. So the system must in one way or another detect that coincidence of locations in order for binding to proceed successfully. Treisman's model includes a 'master map' of locations, whose job is to ensure that such coincidence can be detected readily.[1]

With this we can make our schema a bit more realistic. A map for fine form (or textons) registers the locations of texture features, and similarly for a colour map. The two premises might be rendered

1. Here it is red.
2. There it is stippled.

where the 'here' and 'there' are stand-ins for whatever capacities are employed to identify the place-times that are stippled or red respectively. To achieve the binding of stippled to red, to get to the perception of something that is both stippled and red, the same form of suppressed premise is necessary, namely

3. Some region here = a region there.

The regions overlap. Without one or another identity, feature integration fails. The premises would fail to show that there is something that is both stippled and red.

So we must *identify* the place of one feature with the place of another. Identifications (or the requisite identity statements) are logically interesting animals. They are not at all equivalent to conjunctions. To get to identity statements we need to add a new *kind* of term, with a distinct function. These are singular terms, names or

terms like names, that are used to identify. So if feature integration works as these models propose, then within sentience itself we find capacities that fill two distinct logical functions. One is proto-predicative: the capacity to sense red (or any other feature) both here and there. The other is proto-referential: the capacity to identify the red region as one that is also stippled. This is an informative identity, grasped sub-personally. According to these models, property binding--successful feature integration--*consists in* sub-personal grasp of such identities.

So in these models, we need two *kinds* of term: two place-holders with different roles in the schema 'appearance of quality Q at region R'. These identify a location and attribute qualities to that location. The two roles cannot be collapsed or interchanged. I claim that this 'appearance of quality Q at region R' is the form of non-conceptual representation employed in any vertebrate sensory modality that can solve the many properties problem.

Suppose we try to collapse this partition. Cast 'here' and 'there' not as spatial identifiers, but as spatial qualitative attributes: something like 'hitherness' and 'thitherness'. These attributes are to be conjoined to the other features found in the scene. Our first scheme would yield

(stippled & red & hither & green & cross-hatched & thither)

while the second presents

(cross-hatched & red & hither & stippled & green & thither).

Unfortunately, the two conjunctions are precisely equivalent. If we treat the spatial character of experience in this way, we lose the capacity to distinguish between the two scenes. As Quine (1992: 29) puts it, 'conjunction is too loose'. We must somehow focus the attribution of qualities: we need stippled and red in one place, and green and cross-hatched in another, distinct place. If places are reduced merely to dimensions of qualitative variation, this focusing becomes impossible, and feature integration disintegrates. Since we *can* readily discriminate between the two scenes, such a collapse of our partition cannot be allowed. Even in appearance we make a distinction between qualities and the places at which they appear.

Although our conclusion describes a conjunction of features, the logic that gets us there is *not* conjunction. Feature conjunction is not conjunction. It is predication: *joint* predication.

Consider the difference between the conjunction of 'Lo, a pebble' and 'Lo, blue', and the form 'Lo, a blue pebble'. As Quine says,

[1] This would be "location-based" selection, and there is plenty of evidence for it. But there is also evidence that visual selective attention can use object-based coordinate schemes as well. See section V.

> The conjunction is fulfilled so long as the stimulation shows each of the component observation sentences to be fulfilled somewhere in the scene—thus a white pebble here, a blue flower over there. On the other hand the predication focuses the two fulfillments, requiring them to coincide or amply overlap. The blue must encompass the pebble. It may also extend beyond; the construction is not symmetric. (Quine 1992: 4)

'Pebble' is a sortal, but the logic is impeccable. To get something that is both blue and a pebble, we must identify regions: find that the blue region amply overlaps the one occupied by the pebble. This cannot be done with conjunction.

Instead the work of binding is the work of identification. This map and that map map the same territory. Or: the region characterized by feature *F* is the same region characterized by feature *G*.

We might put it this way: the "master map" in Treisman's model is the part of the system that has the job of securing *identifications*, of determining if and when two or more feature maps have a *common* subject matter. I claimed that Treisman's model shares a basic architecture with many others: multiple independent streams, deriving their own feature maps in parallel; followed by some limited capacity process whose job it is somehow to coordinate and select the outputs needed from those cortical areas. Treisman uses "master map" to name whatever part of the system does the job of cross-identification. This names a function, and that same function goes by other names in other theories. Wolfe calls it the "activation map", and Crick and Koch call it a "saliency" map. Its function, I argue, is to secure identifications.

III. Sensory Feature-placing

So we need not only features, but also the placing thereof. This at last explains the third connotation for the word "feature": an older echo, from the association with Sir Peter Strawson's "feature-placing" languages (see Strawson 1954, 1963, 1974).

The referring expressions of such a language are confined to indexicals and deictical demonstratives (it, here, there, this, that). Its predicates all characterize "features", described as ways of filling space or "kinds of stuff". The only other vocabulary is a present tense copula. The prototypical sentence indicates the incidence of a feature in some demonstratively identified space-time region. Examples of feature-placing sentences:

> It's cold here; it's windy there.
> This is mud. That is ice.
> Here it is sticky. There it is slippery.

White and shiny may characterize a region in just the same way as ice, cold, slippery, and bright. The "features" characterize kinds of stuff (mud and ice) or, more generally, ways of filling space. Under the latter category we find a vast and interesting array of phenomenal properties: white, shiny, cold, slippery, sticky, bright.

The language has no proper names, no definite descriptions, no identity sign, no numerical identity, no quantifiers, and no means for "dividing reference". Feature terms are restricted to mass terms; no sortals or count nouns (puddles, patches, or poodles) are allowed. "Icy here, muddy there" is meant simply to indicate the local incidence of features; it is at some remove from "here is a *patch* of ice, there a mud *puddle*". Patches and puddles can be counted and identified; features only indicated.

Feature-placing sentences name no particulars, and often have the form of a subjectless Lichtenbergian proposition: 'It is cold. It is windy.' Interestingly, since "cold" qualifies (see Strawson 1974: 137), most sensory qualities will as well. It is startling to see how much introspective content can be cast in this form. We have not only Lichtenberg's 'It thinks', but the colloquial 'It hurts' and 'It stinks'. Wilfrid Sellars (1981) suggested that the fundamental form of color ascription is, ultimately, analogously, 'It pinks'.

In the study of sensation all three of these traditions can merge seamlessly. A sensory feature is a phenomenal property, and our predicates thereof can all live happily within the confines of a Strawsonian feature-placing language. The similarities and differences among these features can be inferred from the discriminations and relative similarities that the creature makes among the stimuli that confront it. The study of such discriminations has recently revealed that they are often implemented in the form of distributed, topographically organized, cortical feature-maps. Such maps register differences within some subset of the dimensions of variation in appearance within a given modality. But it has been surprising to discover the neural identity of the axes in terms of which the sensory nervous system actually makes its discriminations, and also to discover that the business is transacted in so many different, and widely distributed, cortical locales.

Sensory systems indicate place-times and characterize qualities that appear at those place-times. They employ a primitive variety of mental representation--probably the most primitive variety. In our schema "appearance of quality *Q* at region *R*" the first member is a sensory quality, to be analysed and understood using the previously developed account of quality space (Clark 1993). The second is a stand-in for the capacities of spatio-temporal discrimination in virtue of which the sensory quality appears *where* it appears. We might be able to discriminate *n* dimensions of qualitative variation at a region picked out by *m* dimensions of spatial discrimination. So a fully determinate specification of a sensory feature might require an n-tuple $[q_1...q_n]$, and the region at which it appears might be identified by coordinates $[r_1...r_m]$. Typically, these regions are regions of space-time, in or around the body of the sentient organism.

Specification of the content of an act of sense requires pairs of the form $([q_1...q_n], [r_1...r_m])$, and (I argue) the pairing principle is analogous to the tie between subjects and predicates (see Clark 2000, section 2.5). Mechanisms to account for the spatial character of sensory experience have a referential or proto-referential function. They function something like demonstratives. Coordinates $[r_1...r_m]$ describe the current activation of the mechanisms of spatial discrimination in virtue of which *Q* appears where it does. Mechanisms that serve to characterize the qualities that appear at those locations have a predicative or proto-predicative function. These "predicates" are just determinate values along the various dimensions of qualitative variation in the modality in question. A fully determinate specification $[q_1...q_n]$ identifies a point in quality space by specifying values for *each* of the *n* dimensions of qualitative variation discriminable at the region in question. Such features or sensory qualities can have multiple simultaneous instances: they serve as characterizing universals or general terms.

These sensory processes have attributive and referential powers that are akin to those of a simple feature-placing language. Restrict the predicates in the language to characterize what I earlier identified as sensory features. Restrict its "placing" capacity to the demonstrative adverbs "here" and "there". Then, in a given context, a sentence in such a language demonstratively indicates a region and attributes some feature or features to it. The *sensation* of a smooth, light red, concave region next a bumpy, brown, convex locale does something logically and semantically analogous; it picks out places and attributes features to them. But the latter does not require or proceed in a natural language; any creature that can discriminate regions red and concave from regions brown and convex has the requisite representational capacity.

IV. Visual Proto-objects

Now I am certainly not the only person to have noticed that vision represents (and must represent) not only visual features but also the

items that such features characterize. It has been gratifying to discover and follow a very interesting literature on what have come to be known as "visual proto-objects" or "pre-attentive objects".

One thing that is gratifying about this literature is that it provides an independent set of arguments for the conclusion I just reached via feature-placing: that we need not only features but some mechanism to pick out or select that which the features characterize. Zenon Pylyshyn starts his very interesting 2001 Cognition article by saying

> Vision suited for the control of action will have to provide something more than a system that constructs a conceptual representation from visual stimuli; it will also need to provide a special kind of direct (preconceptual, unmediated) connection between elements of a visual representation and certain elements in the world. Like natural language demonstratives (such as 'this' or 'that') this direct connection allows entities to be referred to without being categorized or conceptualized. (Pylyshyn 2001, 127)

This "direct connection" he calls "visual indexing", and it is explicitly a connection of direct *reference.* He says a visual index should be viewed "as performing a demonstrative or preconceptual reference function." (Pylyshyn 2001, 127). "FINST" is used as a synonym for "visual index"; it is an acronym for "finger of instantiation"--the pointing finger that often accompanies a demonstrative. He goes so far as to say

> What I have in mind is very like a proper name insofar as it allows reference to a particular individual. However, this reference relation is less general than a name since it ceases to exist when the referent (i.e. the visual object) is no longer in view. In that respect it functions exactly like a demonstrative...(Pylyshyn 2001, 138 n 2)

Pylyshyn gives some interesting and independent conceptual arguments for why we need this reference relation. But perhaps the most compelling grounds are demonstrations of some empirical phenomena.[2] One he calls "multiple object tracking". At the beginning of the trial, a number of the tokens on the screen flash briefly. There is a brief pause, then all the tokens start to move randomly, for roughly ten seconds. The task is to track the ones that flashed. The surprise is that most people have no trouble tracking four to six tokens simultaneously over varying intervals. Those tokens can (to some degree) change shape or color without getting lost, they can endure "virtual occlusions", and so on, though the exact degree of such

[2] Reprints and demonstrations can be found at the Visual Attention Lab website at http://ruccs.rutgers.edu/finstlab/finstsum.html; follow the "Demos" link. You may need to download Quicktime to view them.

forbearance is still under investigation.Pylyshyn proposes that the visual system operates with four to six discrete visual pointers or indexes.

> The visual systems needs a mechanism to *individuate and keep track of particular individuals in a scene* in a way that does not require appeal to their properties (including their locations). Thus, what we need is a way to realize the following two functions: (a) pick out or individuate primitive visual objects, and (b) provide a means for referring to these objects as though they had labels, or, more accurately, as though the visual system had a system of pointers. Although these two functions are distinct, I have proposed that they are both realized by a primitive mechanism called a *visual index* ... (Pylyshyn 2001, 141)

The initial flash serves to attach each index to a referent. That reference can be maintained over time even though the tokens change their location and their properties. Pylyshyn calls it "preconceptual" because it is not reference by description: the properties change, yet tracking is unperturbed.

A visual proto-object can now be defined: it is anything that can be picked out and selected by a "visual index". These are the values of the variables--the items that would satisfy--identifiers found in early vision. He says :

> The concept of a "proto-object" is a general one that has been used by a number of writers ... in reference to clusters of proximal features that serve as precursors in the detection of real physical objects. What these uses have in common is that they refer to something more than a localized property or "feature" and less than a recognized 3D distal object. Beyond that, the exact nature of a proto-object depends on the theory in question. (Pylyshyn 2001, 144n)

Prima facie there exists a big difference between visual indexing (or more generally, proto-objects) and the scheme earlier described as feature-placing. Reference to proto objects can be maintained even though they *change* their locations and their properties. Pylyshyn proposes a small number of *discrete* tokens which are much more like the demonstratives in a natural language than the demonstrative function I propose for sensory feature-placing. But is this contrast real?

V. Why regions?

The first point to make is that there is no logical incompatibility between the two schemes. We might have both. In fact, Pylyshyn's account seems to *require* that we have both. He says "In assigning

indexes, some cluster of visual features must first be segregated from the background or picked out as a unit" (Pylyshyn 2001, 145). This "picking out" is closely related to the Gestalt notion of figure-ground separation.

Consider the difficulty of recognizing an instance of a relational property, such as Collinear(x_1, x_2, ... x_n). Pylyshyn argues that visual indices are needed to attach appropriately ordered values to the various arguments in such lists. As long as *n* is no more than the number of available indices, this argument is compelling and plausible. But remember that we must have already solved the problem of segregating each such x_i from the background. Such segregation requires picking out at least some of the boundaries of x_i and perceiving them *as* boundaries of something distinct from the background. That task requires discrimination of spatial relations at a level of detail that seems impossible to convey using just four or five indices; the "local form" features of the boundary alone, for example, require a far more prolix and compendious indexing of places in patterns of spatial relations than could be crammed into five indices. Even something as simple as edge detection requires registering a "discontinuity of intensity" across spatial intervals; it too is a relation between values at places. And we have not one, but *n* such figures to segregate from the ground. How could we manage the entire task if *all* the direct reference of visual representation had to be achieved using just four or five visual indices?[3] Acknowledging that there is also an earlier and simpler kind of direct reference, found in the placing of features needed to form feature-clusters, makes this problem go away. Feature-maps have the kind of compendious direct reference to places that makes such clustering computationally tractable.

Once formed, a "cluster of visual features" remains the most primitive kind of primitive object. Pylyshyn summarizes the first two steps of his model as

> (1) early visual processes segment the visual field into feature-clusters ... and (2) recently activated clusters compete for a pool of four to five visual indexes (Pylyshyn 2001, 146)

[3] If *all* the direct reference that occurs in visual perception occurs with indices bound to feature-clusters segregated from the background, then it is difficult to explain how we manage to perceive that background itself. Whereas undifferentiated spatial extents are easy pickings for feature-placing. More generally, shape discrimination requires discrimination of spatial relational properties, and it is difficult to believe that all the discriminabilities among shapes can be managed using just four or five argument terms.

So what we are tracking might be just feature-clusters. How might this work?

> to maintain the identity of moving clusters (i.e. to implement a 'sticky' binding) all one needs is a mechanism that treats time-slices of clusters that move continuously over the retina as the same cluster. It could do so, for example, by following the rule that if the majority of the elements in the cluster ... continue to be present in a succeeding cluster then consider both clusters to be the same. (Pylyshyn 2001, 147 note 7)

There is no contradiction in supposing that early vision might use both schemes of representation, with feature-placing being the simpler, and earlier, kind. Indeed, given all the evidence earlier found for location-based selection, it is hard to avoid this conclusion. A brief scan of that evidence may help clarify the import of the suggestion.

Michael Posner's classic experiments on spatial cuing of attention were among the first to suggest that perceptual systems can pick out places (see Posner 1978; Posner, Snyder & Davidson 1980; Posner, Inhhoff, Freidrich & Cohen 1987; Posner & Petersen 1990; Posner & Rothbart 1992). Subjects fix their gaze on a mark centered on the screen, and keep it there throughout the experiment. A "cue" is briefly presented to indicate one or another side of the screen. (This might be done in various ways: a light might flash on one side; an arrow might appear near the fixation cross, pointing to one side or another; one side of the screen might brighten momentarily, and so on.) Then, after a brief interval, a "target" is presented on one or the other side of the screen. As quickly as possible, the subject must identify its location or one or another of its features. Posner found that if the cue predicts the subsequent location of the target, subjects are both faster and more accurate than if they have no cue at all; they are slower and less accurate if the cue occurs on the wrong side. Posner proposed that the cue directs attention to a location, where it remains fixed for awhile even after the cue disappears. It takes some processing, time, and work to disengage attention from a location and move it somewhere else. Attention is somewhat "sticky", to use his memorable metaphor. If the target appears in the same place as the cue, reactions are fast; if it appears elsewhere, the time and work involved in shifting attention explains why responses are slower and more error-prone. Without any cue at all attention may or may not happen to be focused nearby, so we get intermediary values. Note well that all of these results are independent of the need for eye movements. Posner called it "covert orienting"; attention moves independently of the eyes.

The critical suggestion for our purposes is that attention can be directed at a location; its deployment is controlled (sometimes) by spatial coordinates, independently of what is or is not found at the place they serve to indicate. The problem here posed for object-based accounts is to explain how the facilitation or inhibition endures in the interval after the cue stops and before the target appears. There is no obvious object one can nominate to serve this role.

Imagine that the cue has just vanished, and attention remains directed at least momentarily at the place of the cue. The way in which attention is directed at that place is a precise analogy for the way in which (I claim) feature-placing indicates places. Feature-placing can indicate or pick out a place by spatial coordinates derived from the operation of the sensory system itself; it does not need some object to glom onto--some object to which to attach its referential force, or with which to supplement or clarify its identification. Any sensory system that can manage spatial discrimination independently of object attributes must somehow register differences in location independently of attributes found at those locations; those registrations provide the "spatial coordinates" in question. An activation of an ordered *m*-tuple $[r_1 ... r_m]$ of values of such registrations (where the value of *m* is determined by the dimensionality of the spatial discriminations possible) serves to indicate or pick out a place. It does not individuate a region or identify one place-time in a way that would sustain re-identification; but nonetheless one can speak of (for example) attention "directed at" that place, or of a surface feature appearing *at* that place. Attention is centered on the place indicated by the coordinates in question; the smooth, light red, concave region appears (roughly) thereabouts.

Activation of an ordered *m*-tuple of registrations in the spatial discrimination mechanisms of a sensory system indicates a place in somewhat the way that a pointing finger does when pointing accompanies a (sortal free) demonstrative. From the mountaintop one might point to parts of the forest in view, saying "that's coniferous; that's deciduous", or (pointing to the rock underfoot on the summit) "That's lichen; this is moss". The orientation of the finger in the air is the analogy for the *m*-tuple of activated coordinates, where *m* is determined by however many coordinates we need to describe the direction of pointing; the ostended region is the place indicated by those coordinates. There is nothing here that would allow us to determine precisely which regions are *not* to be included; instead there is a high-probability central locus, grading off to less likely penumbral regions. Nor is there anything in the demonstration that would allow a precise answer to the question of whether one has or has not later succeeded in identifying the same place again. Nevertheless, pointing can serve to indicate places; it is, after all, sometimes useful in sortal-free demonstrations. Sensory spatial discrimination can do something analogous.

Posner was also one of the first to demonstrate a phenomenon called "inhibition of return": targets placed in a location from which attention has just departed are detected more slowly, and with more errors, than those placed in a spot as yet unvisited (see Posner & Cohen 1984). Other experiments also demonstrate facilitation or inhibition effects that seem to attach to locations per se, whether or not they are occupied by objects. For example, if attention is cued from one location to another, the intermediary positions are also briefly facilitated; targets presented along that track are processed more quickly and accurately, "as if a spotlight were being swept across the visual field" (Farah 2000, 177; see Shulman, Remington, & McLean 1979). Other experiments address the "width" of the spotlight, showing that locations near a cue are facilitated, while those far enough away (eg "outside the spotlight") are inhibited (Hoffman & Nielson 1981). Philosophers should audit this cottage industry, which is busily probing the properties of the spotlight: finding its speed, showing that it can move faster than the eyes, determining its width, discovering whether it can be "zoomed", and so on. Entry level accountants can be forgiven if they think that these are mere metaphors, though their cash value is still undetermined.

Beyond experiments in visual search and reaction times there are other, more theoretical grounds for thinking that feature-placing picks out places. First, space-time provides a simple and universally available principle of organization: a relatively easy way to correlate the multiple feature maps within one modality. As already mentioned, Treisman's "feature integration" model makes precisely this suggestion, and illusory conjunctions are cited as evidence for it. If spatial relations provide the organizing principles by which multiple feature maps are integrated, then disorders of spatial perception should have consequences that are more global and more profound than those in which a single feature map (or even a single processing stream) is impaired. Treisman (1998, 1300) suggests that simultanagnosia confirms this prediction; it is impairment in the "master map", a "loss of space". (See also Driver and Vuilleumier 2001.)

Feature-placing is found not only in vision, but in any vertebrate sensory modality capable of solving the many properties problem. In modalities other than vision the primacy of place comes more clearly to the fore. For example, both audition and somesthesis attribute features to locales, and those locales seem to be indicated independently of the identification of objects. Spatial discrimination in audition is surprisingly good; even humans can locate sounds with fair accuracy, in a purely auditory fashion. Unless a sound is an object, discriminating the location of a sound is not discriminating the location of an object. It does indeed make sense to speak of the organization of "auditory scenes", where the elements of the scene are not objects (in any ordinary sense of the word) but mere streams of sound. In illusions of auditory localization, one mis-locates a sound; the ringing of the telephone might sound as if it is coming from a place in which no sound sources are to be found. That place is indicated or picked out by auditory spatial discrimination in a way that is similar to the way that attention can be directed at a place or a color can appear at a place. Yet audition does not identify an object at that place; it merely attributes auditory features (pitch, loudness, timbre, etc.) to that locale. It is an "object-free" indication; the subject matter of the attribution is picked out independently of (and prior to) identification of any objects. Similarly, although we sometimes talk of aches and pains as if they are objects, perception of aches and pains and other sensations cannot be assimilated to perception of objects.

Just as location can provide a viable medium of exchange between different feature maps within a modality, it can also provide one across modalities. The philosophical point is familiar to readers of Aristotle and Kant. To perceive that red concave region as also being both smooth and cold, one must somehow attribute visual, tactile, and thermal features to the same locale, and be able to pick out their common subject matter using any of the modalities involved. The spatio-temporal region characterized by those qualities seems the only viable candidate.

Perhaps the most compelling modern evidence for the reality of something like the "common sense" is found in experiments on cross-modal cuing of selective attention. Posner showed how a visual cue could serve to direct attention to one or another side of a visual display; auditory and tactile cues have been shown to do the same thing. Subjects in experiments reviewed in Driver and Spence (1998) looked at a display with a center light (on which to fixate) and vertical pairs of lights on either side of that fixation point. In each trial, one of the upper or lower lights in one of the pairs on one or the other side of the center would light up, and the subject would report "up" or "down" as quickly as possible. Just as in Posner's experiments, a visual cue on the correct side preceding the target stimulus would both speed responses and improve accuracy. But Driver and Spence also placed loudspeakers behind both sides of the display, and found that auditory cues also served to facilitate visual responding.

It seems that a sound can attract visual selective attention to its locale. The implication is that there is some mechanism for translating auditory locations into visual ones. That is, the system uses the location of the sound to provide directions for visual selective attention, so there must be some means of translating the auditory coordinates for that location into ones that can direct visual selective attention. Otherwise, how would the sound help visual responses? Driver and Spence (1998) found similar facilitatory effects for tactile cues (a slight vibration on the left or right hand of the subject), and for cuing in reverse directions (eg a visual cue speeds a tactile discrimination, and so on). The only pairing for which cross modal facilitatory effects could not be demonstrated was visual cuing without eye movements for auditory discriminations. If subjects were allowed eye movements, facilitation did occur.

We seem then to have a mutual inter-translatability between visual, auditory, and tactile coordinates for the locations of stimuli. A cue presented in one such modality can serve appropriately to direct selective attention in another. There are no sensible features shared across all three modalities; the only attributes shared across all three are spatio-temporal locations. So these cross-modal effects provide compelling evidence that each modality has means of picking out locations, and furthermore that the means used in one modality can be translated into terms that can aid selective attention in another. It is not quite the "common sense", but it does some of the same jobs.

Driver and Spence (1998) have also demonstrated that the coordinate translation process is modifiable--as it has to be, since spatial relations between receptors in different modalities can change. This can happen on the fly, with a simple shift in posture. For example, consider what happens if the subject crosses his or her arms, so that the right hand lies closest to the left side of the visual display (and conversely). If we use a tactile cue on the right hand, will it draw visual attention to the left side of the display, or (as previously) to the right side? It would be more adaptive if the system were smart enough to implement the former alternative, and that is what Driver and Spence found. They call the coordinate translation process "remapping". It is also needed between vision and audition: if a subject shifts his gaze, but not his head, then the locations of loudspeakers relative to the visual center-point will all shift. (The loudspeaker that was to the right of one's previous fixation point might now be directly under it.) Yet Driver and Spence (1998) found cross-modal cuing remaps accurately in that experiment as well; the loudspeaker that is now to the left of the visual fixation point draws visual attention to the left, e.g. to a different visual location than it did previously.

VI. Feature-placing v. Proto-objects

So at last the question is on the docket: does preattentive vision (or preattentive perceptual processing in general) have a referential capacity that is above and beyond that provided by feature-placing?

To approach this question sensibly, one would first need to know what exactly are the resources available at the starting point. That is, what exactly *are* the visual features in terms of which early vision does its business? It bears emphasis that right now *we do not know*. We may be in for some big surprises when their identities are revealed. They may not correspond to ones that are readily accessible introspectively, or to ones that are intuitively obvious.[4] Here are two intriguing quotes on this question from Anne Treisman**.** She says "The critical question is what counts as a feature for the visual system" (Treisman 1998, 1301). The question goes back a ways; in 1988 she described it as "the question of how to decide what is and what is not a functional feature in the language of visual coding" (Treisman 1988, 203).

A second surprise about the features is that they themselves can grow into constructions that have surprisingly complicated logical properties, in what Treisman calls the "feature hierarchy". We think of color as the prototypical "punctate" feature, one which might characterize the smallest point that one could see. The simplest "visual field" would be a sum of such points, each with its own particular color. But "features" can be much more complicated. A visual texture, for example, characterizes more than one point. It is a characteristic of a "surface": of a topologically connected sum of such points, at a certain depth (or moving at a certain speed). When ascribing a texture to a region, one is not saying that every point in that region has that texture; instead various points in it stand in

[4] Even surface color cannot be treated as "a" feature, since it has at least three independent dimensions of variation. How exactly might the visual system manage chromatic discriminations? At certain points it clearly uses an opponent-process organization: we have axes for yellow-blue, for red-green, and for white-black. But it is premature to conclude that there is a chromatic feature map that uses just those three axes. Similarly, discrimination of line orientation (vertical, horizontal, tilted) could be managed in all sorts of ways. Even though it would be intuitively pleasing, we cannot assume that there is one map that registers such orientation in angular coordinates.

relations to one another consistent with that texture. Textures, then, are more complicated logically; they are higher up in the feature hierarchy.

Features that are higher in the hierarchy invoke more spatial dimensions: we go from a one dimensional point, to a two dimensional surface, to a two and half dimensional oriented surface, to full 3 d objects. Even though David Marr scorned the idea of features, the hierarchy shows a similar progression of representations with an increasingly rich spatial ontology: a point in the array of intensities, an oriented edge in the primal sketch, a surface in the 2.5 d sketch, and finally a three dimensional object. We also get a similar progression of increasingly sophisticated coordinate schemes with which one might refer to items in that ontology: retinotopic, head-centered, ego-centric, allocentric. Finally, there is an increasingly complicated logical structure for features higher in the hierarchy. If you think of these "features" as data structures, those data structures become increasingly elaborate higher up. It is relatively simple to assign a color to a point, but even something as simple as a 2d kink in a 2d contour is much more complicated. One must find the "discontinuities of intensity", as Marr puts it, and then find a contiguous series of loci of such discontinuities. That gives you, perhaps, a contour. Then the contour itself has different orientations at different places along its length; a kink is a discontinuity in those orientations.

But kinks or contours are still very simple compared to developments at the end stage. Think of "glistening" or "shimmering" as potential visual features. They require reference to developments over time, as do any characterizations of visually perceptible motion. You see a hawk glide over the shimmering lake; a cat on the shore bristles. The tableau presents us with (at least) three features each of which is four dimensional.

Jeremy Wolfe has done much to clarify the question of which features are "basic" features. What he means by a "basic" feature is defined by two characteristics: (1) it supports efficient search (reaction times that used to be called "parallel search") and (2) it can on occasion allow for "effortless texture segmentation". Local differences on the feature in question can "pop out". Using this criterion, Wolfe (1996a) produces a list of "basic features" in vision.

Five are relatively unsurprising, as their physiology is becoming better understood: color, orientation, motion, size, and stereoscopic depth. By "color" Wolfe means just differences on a two dimensional color circle, ignoring luminance differences. The latter have not been adequately tested in visual search tests to see if white and black act just like other colors, or whether luminance is perhaps a distinct basic feature. "Orientation" is at least orientation in a frontal plane (horizontal v. vertical); it might or might not include the tilt of that plane. "Motion" might or might not refer to separable dimensions of direction and speed. "Size" might just be spatial frequency.

Others of Wolfe's basic features come from very different levels in the "feature hierarchy". For example, he gives curvature, vernier offset, glossiness, and pictorial depth cues as four more basic features. Curvature and vernier offset could be features of an edge, contour, or boundary, and so not get to the logical complexity of a surface feature. But glossiness requires at least a surface oriented in depth, and so is higher up in the hierarchy. The pictorial depth cues (occlusion, perspective, etc.) and tilt (if it is on the list) are similar: these are features that don't make sense unless characterizing at least 2.5 dimensional entities. Motion and stereoscopic depth could well be at the top of the hierarchy, requiring the tracking of 3 dimensional surfaces over time.

Wolfe says that the most controversial and problematic of his basic features are the "form primitives". Part of the problem is that there is no consensus on the primitives sufficient to describe shapes. Various primitives have been shown to support pop-out and efficient search: line terminations, junctures, closure, intersections, convergence, containment, and other topological features. Shape turns out to be the crux. It is a complicated relational property, and shape terms have some but not all of the logical characteristics of sortals. As Peacocke (1992a,b) pointed out, one can count the number of rectangles that are visible, and the number might vary depending on whether one means "distinct tokens" or "distinct types". I think he was quite right to provide a special place for shapes: they are above his "situated scenarios" but below propositions, so they are "proto-propositional".

Two items might have identical sets of "local" features yet differ in their arrangement, and so have different shapes. Perception of shape seems to require perception of relational properties. One must identify the various elements of the arrangement as the *N* terms that stand in some *N*-place relationship to one another. This would be a rather complicated data structure; it is up there a bit in the feature hierarchy. If you have two shapes that differ only in their arrangement of the same *N* local features, this discrimination would require very careful placing of each of the *N* terms into the appropriate slot. If binding requires attention, this would require a lot of attention! Yet we have what we think of as a sensory capacity to discriminate shapes. Can it be preattentive? If so, how? Those I think are two of the most interesting questions alive in vision science today.

Wolfe's own answer to the question is no. Preattentive objects have a limited set of basic features, restricted to "properties that define visual surfaces and that divide one surface area from the next in texture segmentation" (Wolfe 1994, 222). These properties can include features of local form (curvature, vernier offsets, terminations, etc) but they do not include the overall shape of a thing. Tokens which share all the same local form features, but which differ in their arrangement, cannot be discriminated from one another without the application of focal attention. So preattentive objects are for Wolfe "shapeless bundles of basic features" (see Wolfe & Bennett 1997).

VII. Why not objects?

Now for some quick contrasts between feature-placing, proto-objects, and objects proper. The reference found in simplest varieties of feature-placing is to a place, while proto-objects can be tracked across places. Space-time regions do not move relative to one another, while proto-objects do. A feature map contains vast quantities of information sustaining the spatial discriminability of the features it registers, and it is the use of such information that yields the capacities we describe as "indicating" or "picking out" places. In contrast, deictic codes, object files, and visual indices are discrete and few in number; at any time there are at most five or six of them available. Once each has been assigned a value, new items can be tracked only if one of those assignments is changed. The spatial reference of a feature map cannot in this way be isolated to the references of a few discrete terms. To put it crudely, a FINST refers in something like the way a demonstrative term in natural language refers; a feature-map refers (or better, "indicates regions") in something like the way a map refers.

There are also important contrasts between proto-objects and objects proper. The most fundamental is that proto-objects need not be objects at all; instead they include whatever is *perceived as* an object. Pylyshyn says that multiple object tracking [MOT]

> operationalizes the notion of 'primitive visual object' as whatever allows preconceptual selection and MOT. Note that objecthood and object-identity are thus defined in terms of an empirically established mechanism in the human early vision system. A certain (possibly smooth) sequence of object locations will count as the movement of a single visual object if the early vision system groups it this way--i.e. if it is so perceived. (Pylyshyn 2001, 144)

This fundamental idea of defining a "visual object" as whatever the visual system treats *as* an object is shared by other proto-object

theoreticians. Dana Ballard carefully describes the computational and representational properties of deictic codes; his proto-objects are whatever might be the values of those codes. Kahneman, Treisman, and Gibbs (1992) likewise confine their discussion to characterizing the representational and cognitive characteristics of the fleeting representations they call "object files". They do not use the term "proto-object", but Wolfe does: for him a proto-object is whatever is the referent of a preattentive object file. By this criterion proto-objects include in their ranks some merely intentional objects, such as the ones that are perceived as moving objects in episodes of *apparent* motion. In such demonstrations (and demonstrations of the phi phenomenon, the "tunnel effect", and virtual occlusion[5]) nothing in fact moves, even though something appears to move and to change its properties. That non-existent thing that appears to move is a proto-object, since an object file has been opened for it, and the visual system treats it *as* an object.[6]

Under this interpretation proto-objects are intentional objects, and they may be *merely* intentional objects. Something forms a visible group and is perceived as an object, moving through space and enduring through time, but that something need not be a full-blooded object, or an "object" in any ordinary sense of the word. For this reason the terminology that refers to the representational systems instead of their putative referents--to "object files", "deictic codes", or "visual indices" instead of "proto-objects"--is somewhat less confusing, though also less alluring and provocative. Proto-objects can then be defined as the referents, whatever they are, of the deictic terms of that system of representation.[7]

[5] A "virtual occlusion" occurs if the target appears to be occluded by a figure which is not itself visible. The "tunnel effect" requires a gradual occlusion, followed later with gradual disocclusion some distance away. "Subjects report a compelling impression that a single object disappeared into a tunnel or behind a wall, traveled invisibly in the interval, and finally reappeared at the other end." (Kahneman, Treisman & Gibbs 1992, 180). Pylyshyn notes on his website that tunneling seems to block MOT.

[6] According to Kahneman, Treisman and Gibbs 1992, the main constraints on the creation of an object file are spatio-temporal: that something can be perceived as if it moves on a continuous trajectory. Such a trajectory can include occlusions and tunneling effects. If it satisfies those constraints, the putative visual object can change in kind, color, shape, size, and all other sensible properties and yet still be perceived as one thing, whose properties change.

[7] A hypothesis that fits most of the available examples: over its lifetime, an object file names an apparent spatio-temporal trajectory; a four

Other contrasts emerge if we shift attention, as recommended, to the differences between systems of reference to proto-objects and systems of reference to objects in natural languages. The former is preconceptual, direct, and derived from the casual mechanisms of perceptual processing. None of its elements are products of convention. It does not require sorting or identifying its referent as a member of a kind, nor does it require criteria of identification and re-identification sufficient to distinguish the "same one, again" from "a different one, but qualitatively identical". Finally, object files and FINSTs are short-lived. They require an on-going information link to their putative referent, and they expire shortly after that link is broken--in the visual case, shortly after the referent goes out of view. Systems of reference to objects in natural languages need not share any of these characteristics, though deictical demonstratives come closest to sharing at least some of them.

Pylyshyn is refreshingly clear on the distinctions between the reference of a FINST and the more full-blooded varieties:

> The individual items that are picked out by the visual system and tracked primitively are something less than full-blooded individual objects. Yet because they are what our visual system gives us through a brute causal mechanism (because that is its nature) and also because the proto-objects picked out in this way are typically associated with real objects in our kind of world, indexes may serve as the basis for real individuation of physical objects. While it is clear that you cannot individuate objects in the full-blooded sense without a conceptual apparatus, it is also clear that you cannot individuate them with only a conceptual apparatus. Sooner or later concepts must be grounded in a primitive causal connection between thoughts and things. (Pylyshyn 2001, 154)

The "full-blooded" individuation of objects presumably requires mastery of principles of individuation and reidentification, yielding a firm distinction between numerical and qualititative identity. FINSTs are pre-conceptual, and lack the wherewithal to satisfy this condition.

VIII. Some Tentative Conclusions

So does preattentive vision have a referential capacity that is above and beyond that provided by feature-placing? In particular, does it manage to identify objects, in some viable sense of that phrase, or is still basically just indicating the incidence of features? Now this is a big question, and this paper will have served its purpose already if it has stimulated some interest in it. But it is in order to draw some tentative conclusions about the powers and inadequacies of the schemes of reference under discussion.

First, notice that the grounds on which one calls a proto-object an "object" are themselves a bit tenuous. They come down to three: they have distinct locations; they are trackable; and they are countable. These attributes of objecthood are perceptible only after the process of figure-ground segmentation has done some work. A feature cluster that might grab a visual index must in some way stand out as a figure, segregated, distinct, or separable from an otherwise undifferentiated background. It has a "distinct" location not in the sense that all its borders are perceptible and precise, but rather in the sense that at least some parts are clearly separated from their background; some places are clearly occupied by some portion of the figure, and others are not. That it can be counted and tracked are other implications of the same segregation of figure from ground.

Visual features higher up in the feature hierarchy can themselves separately possess those three characteristics. Recall that features are not confined to punctate properties; an "edge", as a relation among points characterized by sensibly distinct features, is also well within the ambit of a sensory feature map. Furthermore, these edges--these "discontinuities of intensity", where features sensibly *change*--can be (and are) localized. They can be placed. Such localization relies on the ability to discriminate differences in features, but that ability is something sensory, and was already granted when we described the ordering of discriminations that yields a range of features.

"Edges" though are a few steps up in the feature hierarchy, and they have some interesting logical properties. Like shapes, edges can function as quasi-sortals: they have distinct locations, they are countable, and they can be tracked. The lip of an ocean wave provides a distinctive edge that can be discriminated even when the wave is far out from the beach. You can track the waves as they roll in, and count them as they crash on shore.[8] Pylyshyn says:

> Objecthood need not be specific to a particular modality, and more general notions of objecthood might turn out to be theoretically useful.

dimensional worm. At any moment during its lifetime it indicates the location of its "object" at that moment. The sensible properties of the thing at that place are not essential to this function.

[8] Waves also have discriminable shapes, with distinctive and named geometric features such as "crest" and "trough". The lip of the wave is not the only thing tracked when one tracks a wave; one might track the entire shape, as it appears to move towards the shore. This is true even though parts of the wave do not have distinct locations; the trough is not something that can be segregated from the background.

> For example, if we define objecthood in terms of trackability ...then objecthood may become a broader, and perhaps theoretically more interesting notion. For example, it appears that even when visual 'objects' are not distinguished by distinct spatio-temporal boundaries and trajectories, they still function as objects in many other respects. They may, for example, be tracked as individuals and they may exhibit such object-based attention phenomena as the single-object detection superiority.... (Pylyshyn 2001 145 n)

Now feature-placing can give us the wherewithal to locate, track, and count some putative entities, such as the waves in the ocean as they come crashing onto the beach. But waves are still quite a ways away from being individuals or "objective particulars". Consider: Two waves can fuse; one wave can split into two; "a" wave has entirely different parts at different times; and if all those parts are arranged in just the same way at a different time, it is a purely verbal question whether we have the same wave again, or a different one that is qualitatively identical. Waves are often mentioned by Buddhist thinkers who want to point to something that appears to be an object but is not. In this I think they are right.

What additional steps are needed to get us from higher-level, locatable, trackable, and countable features first to proto-objects, and then to objects proper?

Feature-placing fails in several ways to have the referential powers of visual indices. First, there is no obvious way that feature-placing could be pressed to yield a small set of discrete terms (four to six visual indices or pointers) each of which can have its reference separately assigned. Nor is there means within feature-placing to track a referent independently of its location, or to track it over time. Although some higher-order features characterize developments over four dimensions (such as the direction of motion, motion trajectories, and the earlier examples of "glistening" and "shimmering"), strictly speaking even these features do not track anything over time; they merely ascribe a fixed feature to a spatio-temporal region that is somewhat more temporally extended than most. Lacking tense, feature-placing cannot register that this feature (here and now) was that one then.

But these expressive inadequacies point to a more fundamental difference between feature-placing and proto-objects. Although features higher up in the hierarchy can separately possess the properties of being discretely locatable, trackable, and countable, there is no mechanism in feature-placing to produce a *cluster* of such features--a cluster which has those characteristics as a unit. As noted earlier, these three attributes of objecthood all arguably derive from the process of segmentation and grouping, a process resulting in a group of features being perceived *as* a group, and that group being perceived as something distinct from its background. We need somehow to form a "feature-cluster", and that feature-cluster must become something like a figure, distinct from ground. While there certainly are grouping processes involved in the perception of shape (for example), feature-placing proper seems to lack the power to treat a group of distinct features as a group, and to distinguish "it" (that feature-cluster as a whole) from its background.

Perhaps this last inference is wrong, and the grouping processes involved in shape perception give us something close to what we need. Shape perception is a critically important locus in this domain. We need to think more about it. It is important too in the step from proto-objects to objects proper. Recall that Strawson points to *shape* as potential raw material out of which one can construct sortals. He says

> ...the idea of a simply placeable feature might include--might indeed be--the idea of a characteristic shape, and in this way provide a basis for criteria of distinctness for individual instances of the feature. (Strawson 1954, p. 253).

Consider the "idea of cat", he says:

> if there is to be a general concept of the cat-feature, corresponding in the required way to the notion of an individual instance, it must already include in itself the basis for the criteria of distinctness which we apply to individual cats. (Roughly, the idea of cat, unlike that of snow, would include the idea of a characteristic shape.) (Strawson 1954, 247-48.)

Ayer too had a similar idea, with his notion of "cat patterns" (see Ayer 1973: 91-2). Shape is very interesting, because it is unarguably a sensory feature, yet it provides some, but not all, of the resources needed to make the step to sortals.

Reflection on that step yields a final observation on the powers of feature-placing. I argued that solving the binding problem--solving the problem of feature integration--requires identification. One must in some sense grasp an identity between regions indicated by separate feature maps. Yet feature-placing per se has no identity sign, and (strictly speaking) no identification. It serves only to indicate the incidence of features, not to identify particulars bearing properties. So does feature-placing secure the identification, or not?

The work of Treisman and Wolfe suggests a hard-line answer. For Treisman coincidence of features is something that no feature map can manage on its own; it requires the additional mechanism of selective attention, operating on something like a "master map". For Wolfe visual selective attention is not *always* required for feature integration, but it is still *sometimes*. In particular, perception of the overall shapes of things must still call on its services. Preattentive objects are "shapeless bundles of basic features".

If this is right then the hard-line answer is, as Strawson guessed, that feature-placing precedes the introduction of identification and identities, but it also provides the materiel which makes that introduction possible.

> These ultimate facts do not contain particulars as constituents but they provide the basis for the conceptual step to particulars. The propositions stating them are not subject-predicate propositions, but they provide the basis for the step to subject-predicate propositions. (Strawson 1963, 218).

We can sharpen the sense in which the "placing" is proto-referential. To establish that one region is both *F* and *G*, we need contributions from the two separate feature maps in which the families of *F* and *G* are found. Both maps must contribute a term, indicating a region; and those two regions must overlap. Unless each feature-placing map can contribute such an indicator, feature integration fails. But the mechanisms of feature-placing alone cannot establish that the two indicators identify the same region. In short, the placing in feature-placing provides materiel with which to make the step to full-blown reference, but it alone cannot complete that step. It is proto-reference; or, if you like, protean reference.

References

Ayer, Alfred J. (1973). *The Central Questions of Philosophy*. Harmondsworth, Middlesex: Penguin Books.

Ballard, Dana, Hayhoe, M. M., Pook, P. K. & Rao, R. P. N. (1997) Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences 20* (4): 723-67.

Broad, C. D. (1927). *Scientific Thought*. London: Routledge and Kegan Paul.

Clark, Austen (1993). *Sensory Qualities*. Oxford: Oxford University Press.

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

Crick F. and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in Neuroscience,* 2: 263-75.

Crick F. and Koch, C. (1997). Towards a neurobiological theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (eds.), *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press, 277-292.

Cussins, Adrian (1992). Content, embodiment, and objectivity: The Theory of Cognitive Trails. *Mind 101* (October): 651-88.

Davidoff, Jules (1991). *Cognition through Color*. Cambridge, Mass.: MIT Press.

Driver Jon & Vuilleumier, Patrik (2001). Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition 79*: 39-88.

Driver, Jon & Spence, Charles (1998). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society of London B 353*: 1319-1331.

Farah, Martha J. (2000). *The Cognitive Neuroscience of Vision*. Oxford: Blackwell Publishers.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex,* 1: 1-47.

Goodman, Nelson (1977). *The Structure of Appearance.* 3rd edn.

Boston: Dordrecht Reidel.

Hoffman J. E. & Nielson B. (1981). Spatial selectivity in visual search. *Perception and Psychophysics 30*: 283-290.

Jackson, Frank (1977). *Perception: A Representative Theory*. Cambridge: Cambridge University Press.

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

Peacocke, Christopher (1983). *Sense and Content: Experience, Thought, and their Relations*. Oxford: Clarendon Press.

Peacocke, Christopher (1992*a*). Scenarios, concepts, and perception. In Tim Crane (ed), *The Contents of Experience: Essays on Perception.* Cambridge: Cambridge University Press, 105-135.

Peacocke, Christopher (1992*b*). *A Study of Concepts.* Cambridge, Mass.: MIT Press.

Posner M. I, & Rothbart, M. J. (1992). Attentional mechanisms and conscious experience. In A. D. Milner and M. D. Rugg (eds.) (1992). *The Neuropsychology of Consciousness*. London: Academic Press, 91-112.

Posner M. I., Inhhoff A. W., Freidrich F. J. & Cohen A. (1987). Isolating attentional systems: a cognitive-anatomical analysis. *Psychology 15*: 107-21.

Posner, M. I & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience 13*: 25-42.

Posner, M. I. (1994). Attention: the mechanisms of consciousness. *Proceedings of the National Academy of Science USA, 91* (August): 7398-7403.

Posner, M. I. (1978). *Chronometric Explorations of Mind*. Hillsdale, NJ: Lawrence Erlbaum.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology 32*: 3-25.

Posner, M. I. & Cohen, Y. (1984). Components of visual orienting. In *Attention and Performance* (vol 10): *Control of Language Processes*. Ed. by H. Bouma & D. G. Bouwhuis. Hillsdale, NJ: Lawwrence Erlbaum, 531-556.

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

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

Pylyshyn, Zenon (2000). Situating vision in the world. *Trends in Cognitive Science 4*(5): 197-207.

Pylyshyn, Zenon (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition* 80: 127-158.

Quine, W. V. O. (1992). *Pursuit of Truth*. Revised edn. Cambridge, Massachusetts: Harvard University Press.

Sellars, Wilfrid (1981), "The Lever of Archimedes", *The Monist, 64*: 1-90. (See lecture 2, "Naturalism and process", *The Monist*, 64: 36-65, paragraph 101 ff.)

Shulman G. L., Remington R. W., & McLean J. P. (1979). Moving attention through visual space. *Journal of Experimental Psychology: Human Perception and Performance 5*: 522-526.

Strawson, P. F. (1954). Particular and general. *Proceedings of the Aristotelian Society,* 54: 233-260.

Strawson, P. F. (1963). *Individuals*. New York: Anchor Books.

Strawson, P. F. (1974). *Subject and Predicate in Logic and Grammar*. London: Methuen & Co. Ltd.

Treisman, Anne (1996) The binding problem. *Current Opinion in Neurobiology* 6: 171-78.

Treisman, Anne (1988) Features and objects: The fourteenth annual Bartlett Memorial Lecture. *Quarterly Journal of Experimental Psychology A* 40: 201-37.

Treisman, Anne (1998). Feature binding, attention and object perception. *Philosphical Transactions of the Royal Society of London B* 353: 1295-1306.

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

Vecera, Shaun P. & Farah, Martha J. (1994). Does visual attention select objects or locations? *Journal of Experimental Psychology:*

General 123(2): 146-60.

Weiskrantz, Lawrence (1997). *Consciousness Lost and Found.* Oxford: Oxford University Press.

Wolfe, J. M., Friedman-Hill S. R., Stewart M. I. and O'Connell, K. M. (1992a). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance* 18(1): 34-49.

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

Wolfe, J. M. (1996a). Visual search. In Harold Pashler (ed), *Attention*. London: Taylor & Francis (University College London Press), 13-73.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1(2): 202-238.

Wolfe, J. M. (1996b) Extending guided search: why guided search needs a preattentive "item map". In Arthur F Kramer, Michael G. H Cole, and Gordon D Logan (eds) *Converging operations in the study of visual selective attention*. Washington, DC: American Psychological Association, 247-70.