Situating Vision in the World

Zenon W. Pylyshyn (2000)

Demonstratives in Vision

Abstract

- **Classical theories lack a connection between visual representations and the real world.**
- **They need a direct pre-conceptual connection between (proto-)objects in the visible world and visual representations.**
- **FINSTs (fingers of instantiation, aka Visual Indexes) do the needed work.**

REPRESENTATIONS

- **They encode properties of the world in the same way that words do.**
- **They can be incorrect. (We can mis-represent a wolf as a dog).**
- **But conceptual (descriptive) representations lack *indexical* reference.**
- **Indexical reference: representations whose reference depends on what is being pointed at by the speaker.**

REPRESENTATIONS

- **Situated Vision tries to address the problem of connecting to the world by *eliminating* representations altogether. (Agre takes it to this extreme.) Behaviorism –agents are just a bunch of complex reflexes.**

- **Visual Index theories and Situated Vision theories agree that there is a need to take into account the nature of the *actual* environment, not only the represented one, in explaining intelligent behavior**.

Extended Minds

- **We use the world as an extension of our minds. We can search through a visual scene the way we search through our own memories.**
- **We store object pointers so we can "look up" information about them in the world.**
- **Herb Simon (a major proponent of representations) has noted that, e.g., ants may seem to exhibit complex behaviors, but they likely in fact follow simple rules, like *move in the direction of the sun and avoid large obstacles*.**

Extended Minds

- **In learning to add, we can learn a rule in terms of "the next column on the left".**
- **To use the environment in this way, people need to keep track of individual objects, and use tracked objects as markers for cognitive activities. The visual index (FINST) is a *mechanism* for making this possible.**

- **DEMONSTRATIVE REFERENCE**
  - **Using demonstrative reference (deictic pointers) avoids the need to encode the scene in terms of global properties, and allows the encoding of relations between the objects and the perceiver/actor.**
  - **Relevant objects can be selected directly.**
  - **Contrast: There is *something* that is the North Star vs. *This very thing* is the North Star. The latter allows action.**

CONSTRUCTION OF THE IMAGE

- **Object representations need to be posited anyway, to explain how the visual system *constructs* the image.**
- **The representation that is the output of the visual system is constructed from the retinal input in a series of stages. Some of the construction involves movement of the eyes (saccades).**
- **Look at a page, it is all clear, but fix your eyes and only a small area is clear.**
- **In scanning the system has to match points across inputs. This is the correspondence problem for incremental visual encoding.**

CONSTRUCTION OF THE IMAGE

- **Solving correspondence would be easy if the agent had an accurate 3D representation of all objects in the coordinates. But experiments show that little information is carried from one fixation to the next, and there aren't representations of absolute locations of objects. Changes in a scene are rarely noticed, unless attention is on the changing object.**
- **Another mechanism is needed to solve correspondence. Demonstrative pointers would do the needed work.**
- **Must point to *objects*, not locations, to work in dynamic scenes.**
- **The mechanism can't use descriptions of objects because the properties change**

CONSTRUCTION OF THE IMAGE

- **Also, for pattern recognition, the system has to execute serial "visual routines". The routines often involve the marking/tagging of objects. For example, a routine is needed for the counting of embedded squares.**
- **Instead of tags, Pylyshyn suggests pointers.**
- **FINSTs are pre-conceptual reference pointers in the sense that the objects are identified and represented without appeal to their properties (i.e., the concepts they fall under).**
- **The FINST might latch onto an object because of the object's properties, but at the introspectible output image, the object representations are not descriptions.**

MULTIPLE OBJECT TRACKING

- **First Exp: 8 dots on a screen, 4 of them flicker, and then they all move around the screen for about 10 seconds. When they stop, the subjects' task is to identify the ones that had flickered.**
- **Subjects consistently track the 4 objects (87%)**
- **How? Not by encoding locations and updating information about locations –too much attention and scanning is required.**
- **Other tracking experiments:**
  - **Dots can't be tracked when connected to non-target dots with a bar (the visual system seems to treat the barbells as objects).**
  - **Dots can be tracked thru occluders and thru changes in color and shape. It takes less time to find a property among targets than non-targets.**

• OTHER EVIDENCE FOR INDEXES

- **Ullman's *subitizing*, or *rapid enumeration* (for < 4 items).**
- **RT increases by 60ms/item from 2 to 4 items; by about 100ms/item when more than 4 items.**
- **FINST explanation: there are 4 or 5 available pointers. Enumeration of active pointers does not require visual scanning of the display.**
- **Predictions:**
  - **If objects aren't individuated with focal attention, they can't be subitized (evidence: concentric squares).**
  - **Objects that suddenly appear in the environment get indexed and once indexed, it can be accessed without searching for it by its properties. (evidence: pop-out)**
- **Pointers vs. Priority Tags: Steve Yantis' tags don't explain correspondence or directed eye movements. Tags in the real world might help, if we also had matching tags in the representation. That's, in effect, what pointers do.**

IMPLEMENTATION: But how???

Koch and Ullman's neural net: It finds the most active unit and shuts off the rest. Then a detector for property P is sent out. If the detector fires, then we know that the focus region has property P.

[But what stays active over time???]

RELATED RESEARCH

- **Object file theory: Kahneman's lexical priming experiments show that actual objects rather than their locations provide the locus for storing/accessing properties of the objects.**
- **Priming: Prior occurrence of a letter decreases RTs to the letter. The priming effect travels with the box in which the letter occurred.**
    - **Explanation: when an object first appears, an object file is created and properties are stored.**

RELATED RESEARCH

- **Diexis in eye-body coordination (Ballard):Viewer-centered representation of the coordinates makes the task more computationally tractable. Gaze allows objects to be referenced without appeal to their properties (with fewer objects in the domain, the indexicals aren't ambiguous). Ballard monitored eye gaze in task of copying an arrangement of blocks.**
- **Subjects seem to use gaze as a pointing device to serialize the task.**
- **Index theory assumes that only indexed objects can be the targets of motor commands, including direction of gaze.**
- **Infants are able to distinguish between one and two objects earlier than they are able to use the objects' properties for recognizing an object seen before.**

Conclusions

The visual system needs some kind of direct reference mechanism to represent objects.

Object representations are non-conceptual -- they do not encode properties.

It was wrong to think that humans are first equipped to detect properties in the world; we seem to detect (proto-)objects first.

Final Thought

Is a real-world connection *really* required? FINSTs aren't literally fingers.

How might FINSTs be implemented? Perhaps it helps to consider how they might be implemented in a brain in a vat.