Psychology 354
Seeing And Visualizing

Situating The Symbolic
Classical Theories Of Vision
Pylyshyn's Hybrid Theory

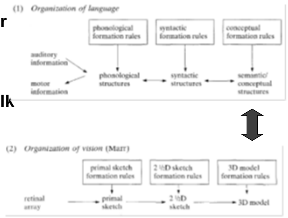
From Situation To Symbolic

- Embodied cognitive science is a strong reaction to the classical or symbolic approach
- "Some ideas are merely the perennial recycling of behaviorist ideology in psychology, which attempts to empty the organism of thought and replace it with the increasingly complex reflexes" (Pylyshyn, 2000)
- Interestingly, symbolic researchers are adopting some sort of situation or embodiment to reduce the computational load of their theories

Zenon Pylyshyn

On Beyond Zebra

- Jackendoff, has long argued for strong links between syntax, semantics, and visually-based entities (objects, events)
- "The problem of how we can talk about what we see can be understood more clearly in terms of diagrams (1) and (2). What is necessary for this task to be possible at all is a set of correspondence rules linking forms of information in the two faculties" (Jackendoff, 1987)
- Jackendoff's proposed link is given as well on the right, in the form of the double arrow


(1) Organization of language
phonological formation rules
syntactic formation rules
conceptual formation rules
auditory information
motor information
phonological structures
syntactic structures
semantic/ conceptual structures
(2) Organization of vision (Marr)
primal sketch formation rules
2 ½D sketch formation rules
3D model formation rules
retinal array
primal sketch
2 ½D sketch
3D model

Ray Jackendoff

Grounding Language In Vision

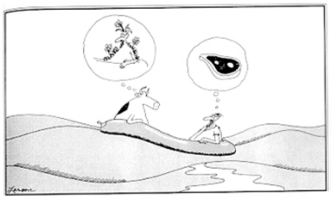
- More modern theories attempt to streamline language processing by situating it with vision
- One example is the cross-channel early lexical learning model (CELL)
- This model extracts phonetic features from recorded speech, and links these to the three-dimensional shape models derived from visual processing
- The goal is to ground semantics into situated visual entities – as demonstrated by Toco the robot in the video on the right

Deb Roy

Toco, demonstrated

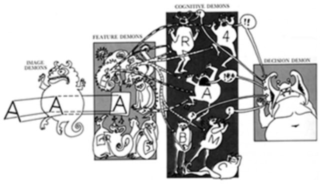
Perception

- Perception can be described – in a particularly classical sense – as constructing meaningful representations of the external world
- How might this kind of perceptual processing be mediated?
- Are hybrid theories of perception possible?



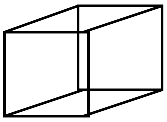
Data-Driven Processing

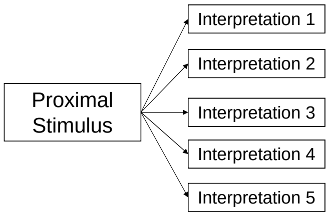
- Perception might be a data-driven process, not affected by beliefs, desires, goals
- The visual world is a set of data that completely determines visual processing
- An example is the feature accumulation theory of vision, such as Selfridge's Pandemonium theory



Perceptual Underdetermination

- One problem with pure data-driven theories is that visual perception suffers from a poverty of the stimulus
- Perception is underdetermined
- The proximal stimulus does not uniquely determine the distal stimulus that produced it





Top-Down Processing

- The poverty of the stimulus might be dealt with by theory-driven or top-down processing – cognitive processing
- Knowledge of the world creates expectations, and these expectations are used to reduce visual ambiguities.
- "Seeing is believing", as in Bruner's New Look theory
- "We do not perceive the world merely from the sensory information available at any given time, but rather we use this information to test hypotheses of what lies before us. Perception becomes a matter of suggesting and testing hypotheses" (Gregory, 1978, p. 221)

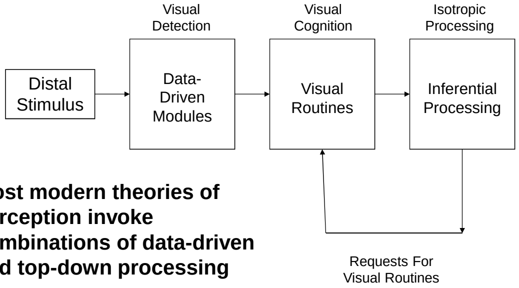
Jerome Bruner

Richard Gregory

Top-Down Problems

- These theories are dangerously incomplete
- Pure top-down perception would not be very adaptive
- Pure top-down theories also have trouble being completely scientific

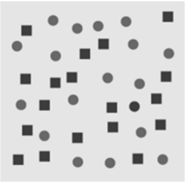
A Compromise View



- Most modern theories of perception invoke combinations of data-driven and top-down processing
- They are ripe for permitting hybrid theories in cognitive science

Treisman’s Visual Search Task

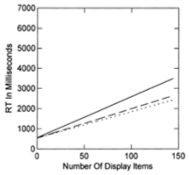
- Evidence for a compromise theory of visual perception comes from studying visual search
  - Present visual displays with different numbers of elements
  - In half the displays, all elements are distractors
  - In the other half, one of the distractors is different – the target
- Subject must decide quickly and accurately if a target is present in any given display
- Latency is dependent measure
- Independent measure is number of elements, and types of elements

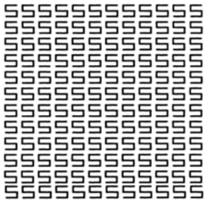


How long does it take to find the red circle?

Serial Search

- Nonprimitive targets (built from combinations of features) are found slowly, and the time to find them depends on the number of distractors
- Note increasing slope of reaction time functions

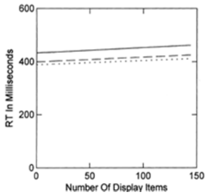




Visual search display from Dawson & Thibodeau (1998)

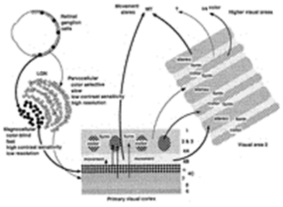
Popout

- Primitive features pop out of the display, as shown in the data on the right from Dawson and Thibodeau (1998)
- The time to find such targets is not affected by the number of distractors
- A target the pops out is defined by a primitive feature
  - Color
  - Motion
  - Orientation
  - Contrast



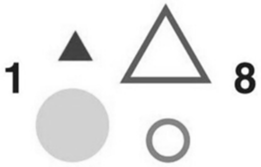
Popout And The Brain

- Livingstone and Hubel noted that the features that popout are also those for which there is evidence of detection by biological transducers
- Thus popout reflects data-driven processing
- Early vision segregates the world into maps of a small number of feature types: color, motion, orientation, contrast



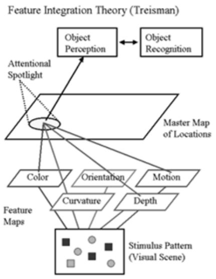
Illusory Conjunctions

- But not all of Treisman’s work reflects data-driven processing
- Cognitive processes like attention are revealed too
- Treisman and Schmidt (1982) found that when attention is divided, subjects frequently report seeing illusory combinations of features – objects not present, but created by recombining features that were present
  - Subjects were presented images like the one below
  - First asked to report black numbers, then the shape at each location
- 20% of trials subjects report object not present (small green triangle



A Compromise Of Processes

- Treisman incorporated both types of processing – data-driven and top-down – in her feature integration theory
- Early vision fills primitive feature maps
- An attentional spotlight is required to paste features from different maps together, permitting ‘object files’ to be linked to semantic memory



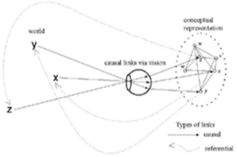
Indexing, Not Feature Integration

- Pylyshyn has proposed a theory of seeing and visualizing that departs from Treisman's in some key respects
  - Single attentional spotlight replaced with multiple indices
  - Indices do not group features, but serve literally as pointers to objects in the world that can later be inspected
- Turns feature accumulation on its head – objects are picked out before their features
- "The problem with descriptive forms of representation lies in the way in which representations are related to objects, including where an observer is situated in the world. In particular, descriptive representations failed to deal with indexical properties and relations" (Pylyshyn, 2000)

Zenon Pylyshyn

Finger Instantiations (FINSTs)

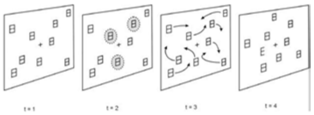
- Pylyshyn has proposed a novel intentional theory to situate representations of objects
- He proposes an attentional tag called a FINST
- A FINST is like having a finger attached to an object
- The object can be tracked, but not necessarily by an attentional spotlight
- The object is in consciousness, but does not require an explicit representational description
- Multiple FINST scan be employed at the same time



The FINST Theory

MOT Support For FINSTs

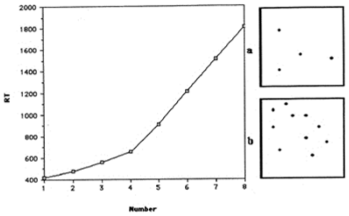
- Evidence for the FINST theory began with studies of multiple object tracking (MOT)
- A subset of targets would blink, drawing FINSTs
- Then all objects would move random way for a period of time
- Subjects had to say whether objects at the end were targets or not
- Subjects could track four or five targets in this way
- A single spotlight of attention could not account for their accuracy



Multiple object tracking demonstrations

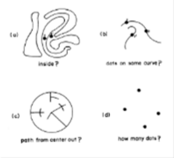
Subitizing Support For FINSTs

- Trick and Pylyshyn argue that FINSTs are also used to enumerate elements
- Their research shows that we can easily subitize 4-5 elements, but enumerating more requires longer processing; RT curves suggest a dissociation of processing types
- FINST deployment can be used to perform this task – but only for as many elements as there are FINSTs



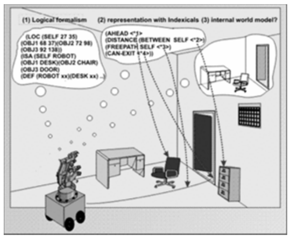
Visual Cognition And FINSTs

- Why are FINSTs necessary?
- There are an infinite number of spatial relationships between objects that can be computed
- We cannot compute them all in advance
- Objects tagged with FINSTs can have relationships computed when necessary
- Objects can have their properties (which might be continually changing)accessed via FINSTs as well
- Objects can be individuated even as their properties change



Implications

- Objects are directly referenced
- The mechanism that does this does not encode their properties
- That is, objects are initially detected without being conceptualized
- A consequence of indexing certain visual objects is that it becomes possible to bind indexed objects to arguments of cognitive representations or cognitive motor programs
- This is an example of a symbolic theory that is situated – a hybrid theory



FINSTs And Scaffolding

- The situating of cognition in Pylyshyn's theory arose from his long interest in why we find it helpful to think with diagrams
- Early work with Elcock attempted to explore this issue with a computer simulation
- But this simulation was explicitly scaffolded!
- "Since we wanted the system to be as psychologically realistic as possible we did not want all aspects of the diagram to be 'in its head' but, as in real geometry problem-solving, remain on the diagram it was drawing and examining" (Pylyshyn, 2007, p. 10)

Edward Elcock

Limited Fields Of View

- Another realistic characteristic was to limit the information taken in at a glance
- "We also did not want to assume that all properties of the entire diagram were available at once, but rather that they had to be noticed over time as the diagram was being drawn and examined. If the diagram were being inspected by moving the eyes, then the properties should be within the scope of the moving fovea" (Pylyshyn, 2007, p. 10

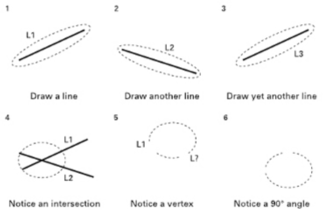
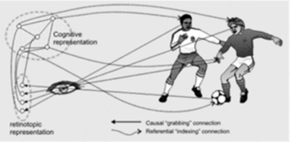


Figure 1.1

As we draw lines (which we see through a narrow foveal view shown by the ellipses) we need a way to refer to particular ones. We can do that by associating them with a description (e.g., "… is at 28° from horizontal") or by placing a label near them. What else do we need in order to re-recognize them when they recur as an intersection or a vertex, or when a second vertex is recognized, or when another property of a vertex (e.g., being 90°) is noticed?

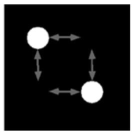
Individuating And Tracking

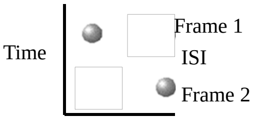
- With not all information available at a glance, elements had to be individuated from one another
- Furthermore, individuated identities must be tracked as the fovea moved, or even as objects moved (and changed appearance)
- This tracking cannot be based on sets of accumulated features: “If objects can change their properties, we don’t know under what description the object was last stored” (Pylyshyn, 2003b, p. 205)
- FINSTs are used to track identies
- “Think of demonstratives in natural language -- typically words like this or that. Such words allow us to refer to things without specifying what they are or what properties they have” (Pylyshyn, 2007, p. 18)


Cognitive representation
retinotopic representation
Causal "grabbing" connection
Referential "indexing" connection

Sticky FINSTs

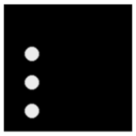
- To track individuated identities, FINSTs must stick to their indexed elements, even as they move or change appearance
- FINSTs must be ‘sticky’ by remembering what went where
- The FINST mechanism must therefore be able to solve the motion correspondence problem
- This is a problem of underdetermination that is confronted by systems that see apparent motion

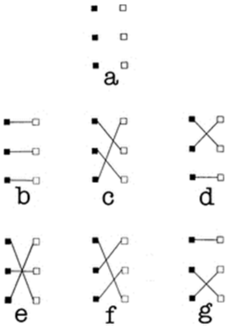




Motion Correspondence

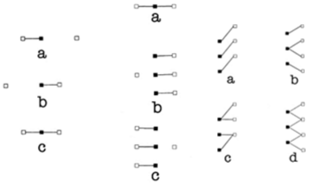
- The problem of “what went where” is called the motion correspondence problem
- With N elements in Frame 1 and Frame 2, there are N! possible interpretations.
- However, only one of these will be correct





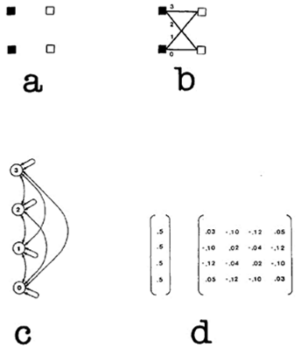
Constraining Correspondence

- Dawson (1991) argued that three natural constraints could be exploited to solve the motion correspondence problem
- Nearest Neighbour
- Relative Velocity
- Element Integrity



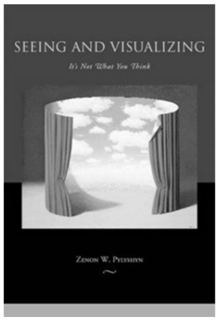
A Correspondence Network

- Dawson (1991) demonstrated how these three natural constraints could be exploited by a PDP network for motion correspondence
- Pylyshyn also appeals to connectionist networks for making FINSTs ‘sticky’



A Hybrid Theory

- Pylyshyn’s theory is clearly an example of hybrid cognitive science
- It invokes internal descriptions
- It employs indices to the world and uses them for cognitive scaffolding
- It exploits connectionist networks to permit FINSTs to track objects that move or change appearance



A Classical Twist

- How does Pylyshyn reconcile his hybrid theory with his place as a classical pioneer?
- He argues that the data-driven elements of his theory – FINSTing and tracking – are not cognitive!
- "I propose a distinction between vision and cognition in order to try to carve nature at her joints, that is, to locate components of the mind/brain that have some principled boundaries or some principled constraints in their interactions with the rest of the mind" (Pylyshyn, 2003b, p. 39)
- The key to the particular carving of the system in his theory is that early vision, which includes preattentive mechanisms for individuating and tracking objects, does not do so by using concepts, categories, descriptions, or inferences