Key issues in Visual Cognition

Satyajit Rao

August 1, 2000

1 What role does vision play in human cognition?

What is vision for? and what role does it play in human cognition? In answer to these questions the typical kinds of tasks that leap to mind are recognizing people, places, locating objects, deciding where to put down your coffee cup, driving, ...etc. So many everyday tasks involve vision that it is harder to find tasks for which one does not need any vision. However, this is still a very task-oriented view of the role of vision and leads to the building of static vision systems that perform a finite set of pre-programmed tasks (maybe with some learning thrown in to improve performance with time). However there is mounting evidence that vision is much more dynamic and plays a much more central role in our cognition.

2 Vision is not just a perceptual input but an integral part of the representation of abstract concepts and reasoning

Recognizing objects, and directing action are important functions of vision no doubt, however, visual processes may also be play an important part in understanding language, understanding how a device works, doing arithmetic, or planning a trip. In other words visual processes may be deeply intertwined with the kinds of things that we would characterize as "high-level" cognition. Evidence drawn independently from Behavioral experiments, Brain-Imaging studies, and Neuroanatomy support this view:

- *Behavioral experiments show visuospatial bias for "abstract" concept representations*: Children below a certain age routinely make the "Conservation error". A few coins when spread out on a table are deemed to be "more" in number than when the same coins are bunched together in plain view of the child. There are several variants of such experiments (done with different props) that strongly suggest that initially a child's concept of "more" has a distinct perceptual bias, in that it is tied to the physical dimensions and appearance of a substance. The abstract concept of "more" or "number" as being independent of "size" emerges later but the initial confusion betrays "perceptual representations" at work. According to some theories of development the initial perceptual representations never really go away but, are simply layered over or superseded by more abstract representations (possible linguistic?) in the course of development.

- *Brain Imaging studies reveal activation of visual areas in abstract cognitive tasks*: The recent slew of studies imaging blood flow in the brain during performance of various cognitive tasks overwhelmingly show that visual areas of the brain repeatedly light up even for "non-visual" tasks like word comprehension. A recent investigation by Dehaene et al [3] investigating

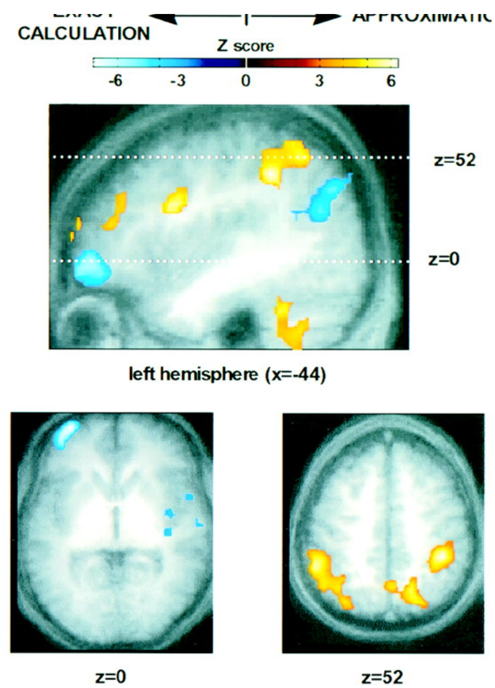


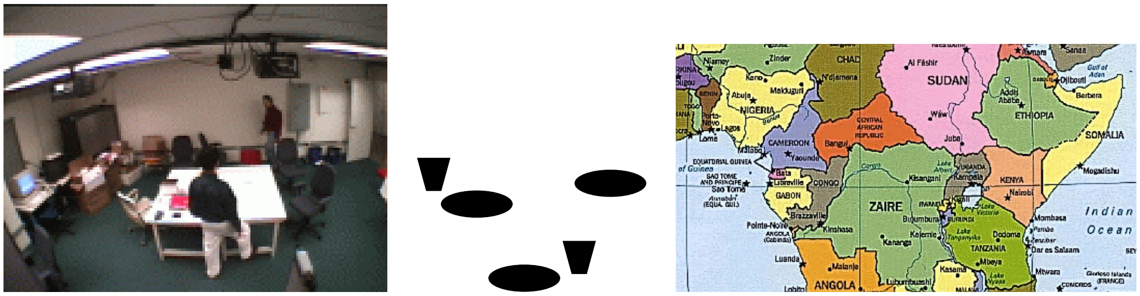
Figure 1: From Dehaene, Spelke, Pinel, Stanescu, and Tsivkin [3]: exact arithmetic calculations recruit word-association areas (in blue) while approximate-arithmetic calculations recruit visuo-spatial areas (in yellow)

the brain loci of "mathematical" thinking made the provocative finding that exact-arithmetic calculations were being supported by language specific areas and recruit areas involved in word association processes, while approximate-arithmetic problems recruited bilateral areas of the parietal lobes involved with *visuo-spatial processing*.

- *Neuroanatomy shows that almost half the brain is occupied by visual areas* Not only is the relative proportion of the visual areas large, but they are extensive feed-forward/feedback couplings to other brain areas. Such extensive hardware connectivity suggests that vision is not just a peripheral perceptual "input" to "higher-level" reasoning centers but is part of the reasoning itself.

Given the evidence from these different sources we are led to the simple but remarkable observation that

> The very **same** set of underlying perceptual/sensori-motor representations and processes in our visual system that support the everyday tasks of getting about the environment and dealing with the real world (locating and manipulating objects, extracting visuospatial relations, learning visuospatial regularities..etc) also get recruited to support other "abstract" tasks like understanding the meanings of words or doing arithmetic.



(a) Which person is closer to the door on the right?

(b) Does every plate have a cup?

(c) How many countries does the equator pass through in this map?

Figure 2: Examples of everyday visuospatial problems

3 Understanding/building the flexibility and adaptiveness that characterizes visual cognition

How is it that the same machinery can be responsible for supporting such a wide range of cognitive tasks? The key to understanding and building visual cognition is to get to the source of the tremendous flexibility and adaptiveness displayed by our visual representations and processes. In particular we need to understand

1. **Vision-specific contributions:** Identify a set of vision-specific representations and processes and show how individually or through their interaction, they provide the flexibility necessary to (i) extract spatial relations (ii) learn spatial regularities and (iii) synthesize visuospatial descriptions.

2. **Interfaces with non-visual modalities:** Understand the interaction of vision with other non-visual and multi-modal representations, both at the sensory level (eg, visuomotor representations of action and movement) and at a more abstract level. eg how a linguistic sentence interfaces with and drives visual representations and visa-versa.

3. **Developmental rules/behaviors:** Initially pre-programmed behaviors whose sole purpose is to engage the robot in activity that results in the maturation/development of visuospatial representations - which in turn result in the development of novel behaviors to replace/supersede the pre-programmed ones.

The following sections discuss these issues in more detail.

4 Vision-specific sources of adaptiveness

To understand the source of adaptiveness of our visual cognition we need to examine three kinds of abilities or competences that our visual system displays.

4.1 The ability to extract spatial relations on demand

The human visual system is remarkably adept at solving spatial problems that arise in the course of everyday activity. Whether it is finding place on the table to put down a cup, or selecting the

shortest checkout queue in a grocery store, visuospatial problems are constantly being solved in order to guide our next action. Figure 2 shows a small sample of problems that our visual system may be presented with during the course of a day. In Figure 2(a) one must determine the locations of the humans and the table and check for a particular spatial relationship between them. In Figure 2(b) one has to not only locate the plates and cups, but also keep track of which one goes with which, based on spatial proximity. In Figure 2(c) one has to somehow count only the regions that the equator passes through, and the counting process has to keep track of the regions already counted.
Even though the problems may look very different the common thread running through all of them is that one must be able to extract some regions in the image, and establish certain spatial properties between them. The selection of the regions themselves may be because of their spatial relationship with respect to other regions (e.g. when you want to find some chalk, you may look in the vicinity of a blackboard which is easier to find). In all of these examples "Object recognition" is not the crux of the problem, any hard-to-recognize object in these pictures can always be replaced by a blob without changing the essence of the problem. It is the establishment of spatial relations that is the focus here. There is widely cited evidence from Ungerleider and Mishkin [7] about two distinct pathways in the primate visual cortex - one devoted to object identity, and the other to object location, and spatial relations. The majority of the literature in computer-vision has been skewed towards object recognition with relatively less attention to the spatial mechanisms.
What kinds of visual mechanisms in humans make it possible to handle the spatial tasks in Figure 2 ? Ideally one would like to take a very simple problem (e.g which of two blobs is larger) and be able to trace the mechanisms all the way from early visual representations in V1, through the spatial selection mechanisms in V4, spatial transformations in the parietal cortex, behavioral short term memory in the prefrontal lobe, to the final motor intention of pointing to the bigger blob. This is of course a crude caricature of one pathway, there are doubtlessly other forward and back-projection pathways involved. The point is that there are multiple brain areas interacting to solve even the apparently simple problem of pointing to the larger blob.
Presently we do not have a reasonable computational model for visuospatial problem solving, but it is imperative to come up with one because these tasks are ubiquitous in everyday life, these are typical of the variety of tasks that we would expect a humanoid robot visual system to deal with from moment to moment and use the results to guide its actions.

4.2 The ability to learn spatio-temporal regularities

The regularities or patterns that we see in the spatio-temporal behavior of objects (falling, colliding, bouncing ..etc) form our model of the physical world. A vast repository of such patterns constitutes what we call *common-sense* knowledge about space. These patterns also form the metaphors into which we fit future experience (Johnson[6]). In fact it would be not be exaggeration to say that visuo-spatial patterns form the basis of language, and reasoning. Therefore, it is critical to have mechanisms that learn spatio-temporal regularities.
Consider the following significant aspects of human cognitive development:

1. Children learn regularities in the environment well before they can speak, [8] [1].
2. Children are rarely supervised in their learning of visuospatial concepts.
3. When they are supervised, for example when an adult describes something that's happening (e.g. "look at that ball falling"), children learn the concept with surprisingly few examples, compared to the large number of examples that current supervised learning programs need.

4. Children make systematic errors in reasoning that indicate a visuospatial basis for concepts and reasoning. In one experiment a child's concept of "more" is tested. The child thinks that there is "more" water in a tall narrow glass than when the same water is poured into a short broad glass. The conservation experiment is one of many experiments which indicate that initially "abstract" concepts and reasoning have a strong perceptual bias.

How are visuo-spatial-temporal patterns learned over time? What kinds of mechanisms pick up the invariants from seemingly different events? What is the representation of the learned patterns? Do the learning mechanisms share anything in common with the attentional mechanisms involved in extracting spatial relations (4.1?
These are some of the issue that need to be addressed if we are to build flexible learning systems.

4.3 The ability to synthesize/imagine visuospatial descriptions on demand

The ability to recall previously constructed spatial representations or synthesize entirely new ones from spatiotemporal defaults is the third component of the visuospatial representational system. Humans can count the windows in their house without being there, or solve problems like "An ant walks 1m North and 1m East on a plane, which is the shortest path back to the starting point?". Farah [5] cites two sources of evidence as supporting the hypothesis that the human brain uses the same machinery for imagery and perception. Behavioral data indicates that humans who have damage to right hemiparietal areas cannot even imagine the left side of previously experienced spatial scenes [2]. Medical imaging data indicates [4] that the prestriate occipital, parietal, and temporal cortices get activated during an imagery task - the same areas that get activated during perception. Having the same representation and hardware for perception and imagery is a parsimonious solution because it means that the entire visual routine machinery can be used on imagined representations just as easily as it is can during perception. The ability to create new spatial descriptions and extract various spatial relations from the imagined scene leads to a powerful problem solving capability.
For example When given the problem:

```
John is taller than Mary, and Mary taller than Susan. Is John taller
than Susan?
```

A child might *synthesize* a visual description of John, Susan, and Mary standing next to each other, and then use her *learned regularity* of what it means for one object to be taller than the next to drive a visual routine - a specific sequence of operations that are applied to the image to *extract the spatial relations* of interest.
The three capabilities of learning spatial regularities from experience, extracting spatial relations on demand, and synthesizing spatial descriptions, work together to create a very powerful and flexible mechanism that plays a major role in representation and inference in humans.

5 Interfaces with non-visual modalities

To be sure, cognition is the result of the interaction of several different representational systems, and talking about just vision is like talking about just the engine of a plane without mentioning the interaction of the engines with the wings.
At the sensory level we know that that there are several areas of the brain containing multi-modal receptive fields and self body-maps. The tactile/visual receptive fields in the pre-motor area (Graziano, Hu, Gross, 1997) where the visual receptive field is anchored to the arm and moves with it - are some

of the more interesting ones. What role do these multi-modal maps play in cognition? they are obviously geared towards the execution of action, and possibly to create a simple coordinate system for learning actions. It is certainly possible any synthesis/imagination of the interaction of the self with the world could be happening in terms of these multi-modal coordinate frames.
At a more abstract level how do linguistic representations interact with visuospatial representations? We know that a sentence like " Which is the bigger dog?" somehow gets translated to the appropriate visual routine, and the answer in turn gets transformed back into a sentence or an action (like pointing). How do these transformations and mappings work?

6 Developmental Rules/Behaviors

A finite system of pre-programmed behaviors seems to be a fairly adequate explanatory model of insect behavior. In humans at least the behaviors seem to perform a very different function. It seems like the function of early behaviors is to get the child into actively learning about it's own body and the interaction of the body with the environment (e.g Piaget's sensorimotor stage of development). Once those representations get populated and don't experience any "surprises" it seems like a new set of behaviors comes into play for learning about the external world through ones body..in other words new sets of behaviors kick in (some preprogrammed, some learned) **to actively drive the child through various stages of cognitive development**. No architecture for visual cognition would be viable without designing at least a minimal set of pre-programmed behaviors to get the cognitive development started.

References

[1] Renee Baillargeon. Physical reasoning in infancy. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, chapter 11. M.I.T. Press, 1996.

[2] Bisiach and Luzzatti. Unilateral neglect of representational space. *Cortex*, 14:pp129–33, 1978.

[3] Dehaene, Spelke, Pinel, Stanescu, and Tsivkin. Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284:pp970–974, May 1999.

[4] Martha J. Farah. The neural basis of mental imagery. *TINS*, 12(10):pp395–99, 1989.

[5] Martha J. Farah. The neurological basis of mental imagery: A componental analysis. In Kosslyn and Andersen, editors, *Frontiers in Cognitive Neuroscience*, chapter 44, pages pp559–73. MIT Press, 1992.

[6] Mark Johnson. *The Body in the Mind - The Bodily Basis of Meaning, Imagination, and Reason*. The University of Chicago Press, 1987.

[7] M. Mishkin, L. G. Ungerleider, and K. A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neuroscience*, 6:pp740–749, 1983.

[8] Elizabeth S. Spelke. Physical knowledge in infancy: Reflections on piaget's theory. In Susan Carey and Rochel Gelman, editors, *The Epigenesis of Mind: Essays on Biology and Cognition*, chapter 5. Lawrence Erlbaum Associates, 1991.