An Outline for Engineering Visual Cognition

Giulio Sandini and Satyajit Rao

September 6, 2000

1 Introduction

In this section we provide some more details about our effort to engineer visual cognition. In previous sections we emphasized the need to understand the flexibility and adpativeness of our visual representations and processes, in particular:

- Vision-specific contributions
- Interfaces with non-visual modalities
- Developmental rules and behaviors.

For reasons that will become clear we start with the last two third issues and then consider the first one.

2 Developmental Rules and cross-modal behaviors

First we would like to state two biases. We choose to

- *Explore visual cognition through a humanoid robot*, because we believe that cognition is by definition about being in and interacting with the real world, and therefore a physical body is essential. Furthermore a human-like body besides being easier to interact with would give insights into our own cognition.
- *Explore visual cognition via a developmental program*. We believe that full blown adult human visual cognition is far too complex to be directly progammed into a robot. It is fundmantally a result of a developmental process that is intimately tied to the interaction of the robot with its enviroment, and can therefore only be engineered by a developmental process.

What aspects of Piaget's [**?**] theory are relevant to our our efforts to engineer visual cognition? There are three aspects of Piaget's theory that stand out in this regard:

- **Development = change of cognitive structures**: cognitive and intellectual change is the result of a development process. Cognitive development is a coherent process of successive qualitative changes of "cognitive structures".
- **Active exploration and construction**. Development of cognitive structures is ensured only with active exploration of the environment and social interaction. All knowledge is a construction resulting from the child's actions. Physical knowledge is constructed through discovery.
- **Bootstrapping and Subsumption**: The stages of Piagetian development do not *replace* each other in a sequence, but are layered so that the later stages grows out of and *modifies* the actions of the previous ones.

Building a human-like robot is the relatively easy part, much harder is the issue of engineering a developmental process. To this end, we forsee the following three broad stages in development.

2.1 Learning about ones own body through it's action on the world

The first stage in development is very similar to Piaget's "Sensorimotor" stage of development [0-2yrs]. In this stage visuo-motor patterns of increasing complexity are learned. The world is explored through some very basic reflexes like grasping, or flailing ones arms, or foveating on an object of high color contrast. At first these reflexes work independently, but soon they start chaining together into more complex sequences. Behaviors start distinguishing between stimuli and are subject to reinforcement. The robot repeats actions until it can reproduce the actions and their effects reliably. For instance it soon learns that executing certain motor commands in certain postures will bring an elongated object (it's own arm) into certain portions of the visual field, or that executing the grasping reflex in a particular posture produces a certain somatosensory sensation (of pinching itself). Exploration becomes more goal-directed as the robot tries to do things to repeat an interesting stimulus. Finally the robot moves from sensorimotor to representational intelligence as the consequence of action sequences that can be simulated in the head rather than by active manipulation (for instance knowing that a certain flailing action will bring the arm into view). The defining feature of this stage is that the robot's discovery of it's own efferents in terms of it's afferents via action upon the world, in other words the robot discovers it's control of it's own end effectors (eg it's arms, eyes, neck...) via it's sensors (vision, touch, sound...) via active exercising of a set of pre-programmed behaviors.

2.2 Learning about the external world through ones body

Now that the robot is not surprised, and can indeed anticipate some of the results of it's own actions (for example how the scene might change if it rotates it's head to the right), it can finally begin to explore the external world with its end effectors. Again, the only reason this becomes possible at all is that the robot can distinguish between itself and the world. So grasping an external object is perceived as such, because it doesn't produce the predictable somatosensory effects of grasping oneself. Now the robot can finally learn patterns in the external behavior of objects like tracking the object as it is released from grasp and learn the correlation of the loss of contact and the downward motion of the object. The robot learns that most objects do not move by themselves unless acted upon, that they do not pass through one another, that they usually make some kind of sound when they come into contact with oneanother. A vast repository of such patterns or regularities that are actively discovered by the robot constitute it's "commonsense" knowledge about the world.

2.3 Projecting experience onto perceptual metaphors

Once the robot has accumulated a large number of patterns about the behavior of the external world (and as we shall see, patterns in its own internal state while perceiving an external event) It certainly becomes possible to learn symbolic tags for these patterns like "fall", or "bigger" with little effort. The reason that children pick up the "semantics" of langauge so quickly with such little supervised learning may be that they already have the meaning as a result of the previous stages of development! The purpose of clustering and tagging the perceptual patterns, may to be able to understand future experience in terms of these abstract tags instead of the raw patterns.

This section so far highlights the stages that we expect to see in development of visual cognition, but does not give much details on the specific *represenations and mechanisms involved*, ie *What* develops? and *How*?. For instance what is the represenation of the pattern of something "falling" and what kinds of mechanisms extract this pattern. The following section tries to address these issue in more detail.

3 Vision specific representations and processes in cognitive development

Many creatures have some form of visual cognition, and go through stages of cognitive development to get there, so one might ask what makes human vision special? In prior sections we mentioned that the ability to to (i) extract spatial relations (ii) learn spatial regularities and (iii) synthesize visuospatial descriptions were key to the flexibility of human visual cognition. In this section we describe how we intend to explore the representations and mechanisms that support these abilities.

3.1 Extracting spatial relations on demand

The problem of extracting spatial relations on demand (eg finding the larger of two objects, or the countries that the equator passes through on a map, etc) is closely linked to visual attention and the topic of "visual routines".

Shimon Ullman [?] initially proposed the problem of finding a versatile spatial analysis mechanism and also described the framework of a solution. The essence of his solution is that there exists a set of elementary operations that when combined in different ways produce different visual routines for doing various spatial tasks. The elementary operations therefore form a kind of basis set for visual routines.

Ullman suggests that visual processing is divided into two stages. The bottom-up, spatially uniform, viewer-centered computation of the base representation (like the 2 1/2 D sketch) followed by the extraction of abstract spatial properties by visual routines. Visual routines define objects and parts, their shapes and spatial relations. The formation and application of visual routines is not determined by visual input alone but also by the specific task at hand. The elementary operations are not all of the same type, some of them operate in parallel across the entire image, others can be applied only at a single location at a time. It is suggested that these characteristics of the operators reflect constraints inherent to the computation they perform, not because of a shortage of resources. The structures computed as a result of the application of various visual routines are incrementally pieced together so that subsequent processing of the same image can benefit from the intermediate results of previous computations. For example when asked to count the number of red objects in a scene, the intermediate result of the locations of the red objects is maintained to help answer a later question about the biggest red object.

Ullman suggests the following as plausible elementary operations:

1. *Shift of Processing focus:* A process that controls where an operation is applied.
2. *Indexing:* Locations that are the odd-man-out in the base representation (e.g an island of blue in a sea of red), attract the processing focus directly. They are called indexable locations and serve as starting points for further processing.
3. *Bounded activation or Coloring:* The spreading of activation in the base representation from a location or locations. The activation is stopped by boundaries in the base representation.
4. *Boundary Tracing:* Moving the processing focus along one or more contours.
5. *Marking:* Remembering a particular location so that processing can ignore it or return to it in the future.

Composing a basis set of elemental spatial operations to solve spatial tasks is an attractive idea. Clearly this kind of explanation is far more plausible than having specialized feature detectors like "the smallest object inside the big circle". However in order to flesh out the details of the visual routines proposal one has to deal with two key issues:

1. Choice of a set of primitives: What is a good set of elementary operations?
2. Composition: Given a set of primitives how do they get strung together to perform some spatial task? Is there a need for an explicit sequencer?

3.2 Learning Spatial Regularities and Visual Routines/Behaviors

Once again a viable theory of visual attention is key to learning spatial regularities like "under", or "pick up", or the visuo-motor routine for crossing a road. The following points list the main features of our approach, a schematic of which is shown in Figure 1.

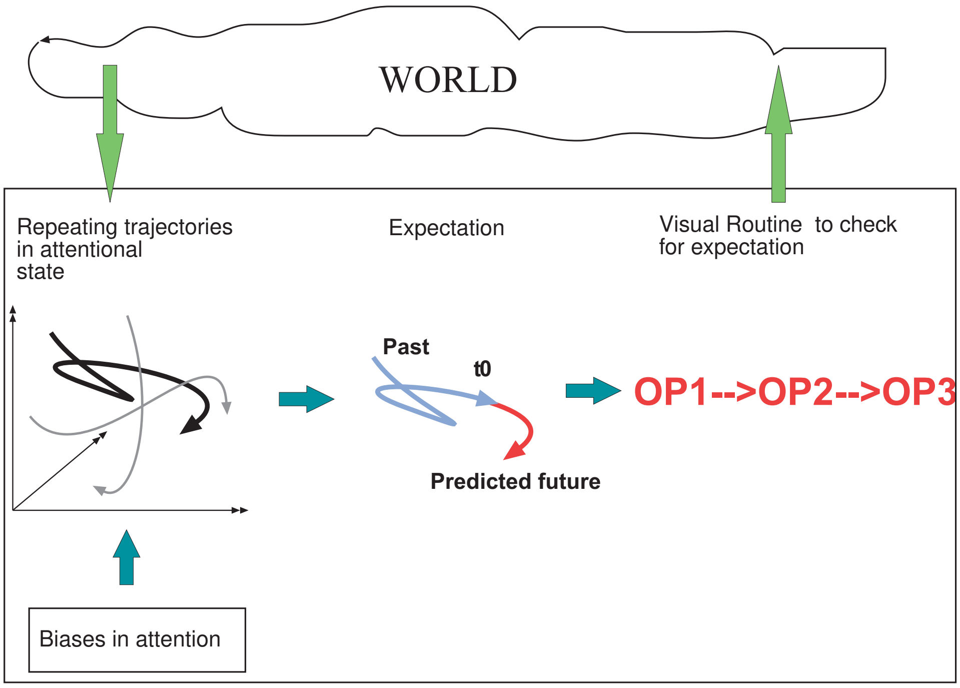

Figure 1: Regularities in the world and attentional biases repeatedly drive the attentional state through patterns. High frequency patterns become expectations. Expectations generate visual routines to check for the predicted part of the pattern.

1. Patterns of visual activity emerge from interaction with the environment . Exploration of the environment leaves a "trace" in attentional state. Regularities in the world (recurring spatial relationships and events) and biases in attention (the tendency to track moving objects for instance) cause repeating trajectories in attentional state. The components of attentional state are as described in section **??**.

2. These emergent patterns in attentional state can be learned. The repeating "patterns of activity" in one's attentional state can stand out simply because of the higher frequency with which they occur.

3. A partial match of the current trajectory (shown in blue) with the learned prior pattern leads to the prediction of the rest of the pattern (shown in red).

4. The predicted sequence of attentional states generate a corresponding "visual routine" which is used to check for those states. The transformation of predicted states into a routine is easy because every visual property has a corresponding operation to extract it.

5. Exploration must be pro-active. In order to learn certain abstract spatial invariants (like pick-up) one must actively select regions and monitor properties to notice any regularities.