An Outline for Engineering Visual Cognition

Giulio Sandini and Satyajit Rao

September 22, 2000

1 Introduction

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 programmed into a robot. It is fundamentally a result of a developmental process that is intimately tied to the interaction of the robot with its environment, and can therefore only be engineered by a developmental process.

For a developmental program one must be be able to specify

- **What is built in apriori:** The hardware, the pre-programmed vision and motor architecture, and an initial set of pre-programmed behaviors.
- **What develops and How:** What kinds of structures get learned because of the initial set of behaviors? How do these structures help bootstrap the learning of new behaviors?

Naturally, we do not yet have the complete answers to these questions. However, in the next few sections we describe our current thinking/approach.

2 What is built in apriori

In this section we describe the hardware, the visual routine machinery which we believe is crucial for the flexibility of visual cognition, the motor routine machinery, and an initial set of pre-programmed behaviors.

2.1 Hardware

Giulio: Please fill in according to BabyBot's specs:

2.2 Vision Architecture

The architecture focuses on two problems are at the core of visual cognition: One is the ability to extract spatial relations on demand (eg if a blob is bigger or above another blob), and the second is the ability to learn spatiotemporal regularities (eg if an object is released it is likely to move downwards).

Early and Preattentive vision

Figure 1 is an overview of the proposed visual architecture. The system can be divided into four distinct levels. At the first level (starting at the bottom) early-visual are computed. The defining characteristic of

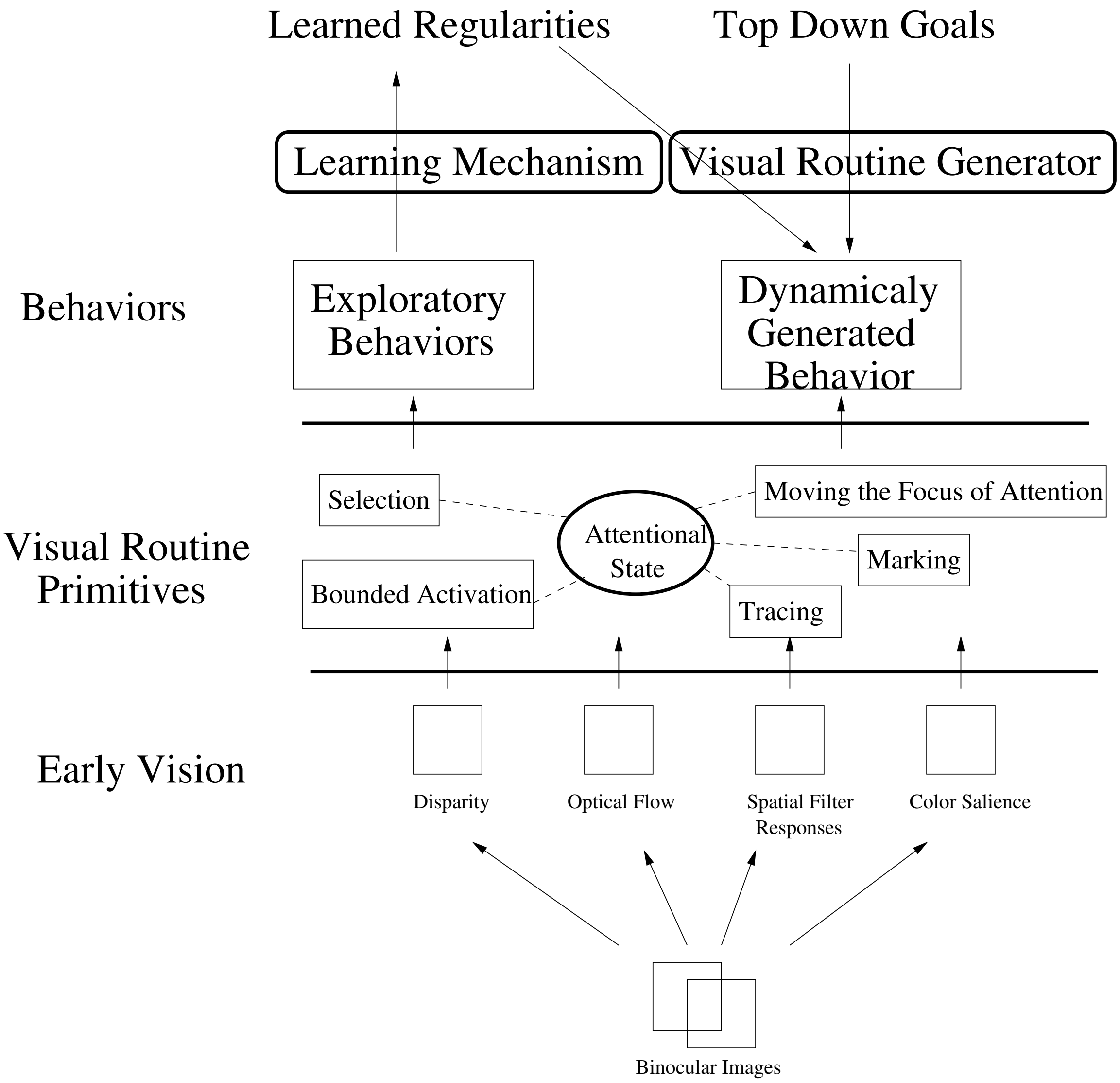

Figure 1: An overview of the architecture of the system

this level is that the image properties are computed independent of the task at hand. The sample of these properties follows

1. Binocular disparity.
2. Optical Flow
3. The responses of spatial filters as described by Freeman & Adelson [**?**]. Five spatial filters (2 first order derivatives and 3 second order derivates of the Gaussian) across 4 different scales resulting in a 20 dimensional vector at every point in the image.
4. A color saliency map - indicating regions of high color contrast.

There has been considerable work on *perceptual grouping* [3][1][2] - the grouping of low-level features like a bunch of line terminators that are aligned, into a high-order feature. One or more of such line or region based groupings (not shown in the figure) may be used to produce a richer early-visual representation.

Visual Routines and Attention

The second level, shows the machinery for visual routines. 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 [4] 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.

To Ullman's primitives we add the the very important notion of **attentional state**, i.e. a set of short-term properties (eg, size of an object, or a set of marker positions, or the relative position of one region with respect to another) that are constantly accessed and modified by the primitive operations.

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

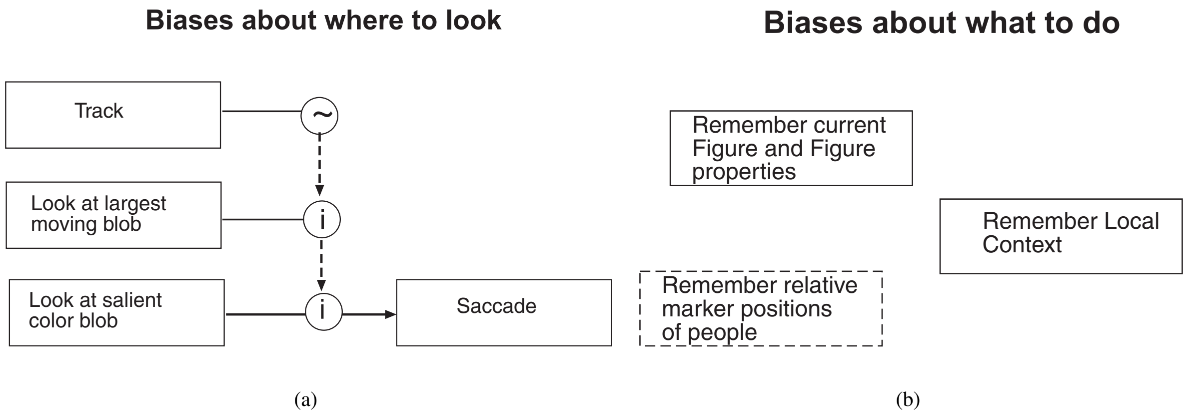

Figure 2: Biases about where to look and what to do at the focus of attention are necessary for exploration.

inside the big circle". However in order to flesh out the details of the visual routines proposal we will have to to address 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?

Exploratory behaviors

The third level in Figure 1 shows the behaviors, some of them pre-programmed and some dynamically generated. These behaviors are constructed from the visual routine primitives as well as from some motor routine primitives (not shown in the figure). Some of the behaviors we intend to implement are:

- Saccading to a sudden motion or salient color region.
- Saccading to a sudden sound.
- Saccading to a human face (the assumption here is that human face detectors are pre-built and not learned)
- Smooth pursuit tracking (the parameters are learned)
- Grasp reflex, i.e. closing the fingers around something when it comes brushes against the palm of the hand.
- Random reaching arm movements

The approach that we take is that pre-programmed behaviors cause the robot to interact with and explore the environment. During free exploration however, i.e. when the system is not trying to solve any particular problem, it *must* have biases about where to look next, *and* what to do at the focus of attention.

Learning spatial regularities and generating novel behavior

Exploration of the environment leaves a "trace", ie a time sequence of attentional states. Figure 3 shows an example of an attentional trace. Recurring spatial relationships and events thus leave repeating traces in attentional state, and these traces are a rich source for capturing an abstract description of the event. For example when watching a ball roll over the edge of a table, the pre-programmed behaviors for tracking the

(a) A ball is tracked, while its attributes and the local spatial context around it are attended to.

(b) A trace of the local context for the entire event, only this trace is remembered

Figure 3: The attentional trace for an event generated by the bottom-up exploratory behaviors of Figure 2. Only the tracking behavior was active

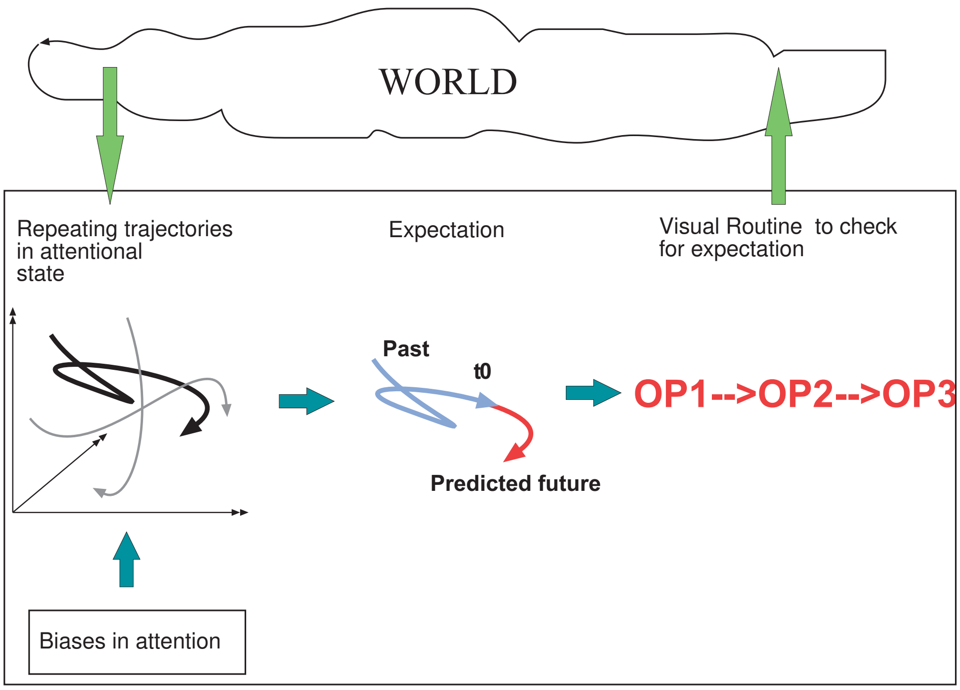

Figure 4: Regularities in the world and attentional biases repeatedly drive the attentional state through patterns. High frequency patterns become expectations. Expectations generate visual routines to check for the predicted part of the pattern.

ball are engaged. Over many separate instances of balls rolling over tables the intersection of the traces generated in attentional state is a good representation of what's spatially common to balls rolling of tables. Furthermore, because the pattern learned is a "trace in attentional state space", it can be easily matched to the description generated as it views a new instance of a ball rolling of a table. The match between the initial trajectories of the learned pattern and the ongoing pattern of activity in attentional state enable prediction of what's going to happen next (the ball falling down and hitting the floor). This expectation is enough to generate a visual routine causing the robot to look down at where the ball is going to fall. Figure 4 shows a schematic of this idea.

3 What develops and how

References

[1] T.D. Alter and R. Basri. Extracting salient contours from images. *Proceedings of IEEE Conference on Computer Vision, Pattern Recognition*, June 1996.

[2] James V. Mahoney. Image chunking: Defining spatial building blocks for scene analysis. Master's thesis, M.I.T, 1987.

[3] A. Shaashua and S. Ullman. Grouping contours by iterated pairing network. In Lippmann, Moody, and Touretzsky, editors, *Advances in Neural Information Processing Systems*, pages pp335–341. November 1990.

[4] Shimon Ullman. Visual routines. A. I. Memo 723, M. I. T. Artificial Intelligence Laboratory, June 1983.