

Visual Routines and Attention

by

Satyajit Rao

B.S. Mathematics, Massachusetts Institute of Technology (1990)
B.S. Brain & Cognitive Science, Massachusetts Institute of Technology (1992)
S.M. Electrical Engineering & Computer Science, Massachusetts Institute of
Technology (1991)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy


at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY


February 1998

© Massachusetts Institute of Technology 1998. All rights reserved.


Author

 Department of Electrical Engineering and Computer Science
February 6, 1998

Certified by

 Patrick Winston
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by

 Rodney Brooks
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by

 Arthur C. Smith
Chairman, Department Committee on Graduate Students

MAR 27 1998

LIBRARIES

Visual Routines and Attention

by
Satyajit Rao

Submitted to the Department of Electrical Engineering and Computer Science
on February 6, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The human visual system solves an amazing range of problems in the course of everyday activities. Without conscious effort, the human visual system finds a place on the table to put down a cup, selects the shortest checkout queue in a grocery store, looks for moving vehicles before we cross a road, and checks to see if the stoplight has turned green.

Inspired by the human visual system, I have developed a model of vision, with special emphasis on visual attention. In this thesis, I explain that model and exhibit programs based on that model that:

1. *Extract a wide variety of spatial relations on demand.*
2. *Learn visuospatial patterns of activity from experience.*

For example, one program determines what object a human is pointing to. Another learns a particular pattern of visual activity evoked whenever an object falls off a table.

The program that extracts spatial relations on demand uses sequences of primitive operations called visual routines. The primitive operations in the visual routines fall into one of three families: operations for moving the focus of attention; operations for establishing certain properties at the focus of attention; and operations for selecting locations. The three families of primitive operations constitute a powerful *language of attention*. That language supports the construction of visual routines for a wide variety of visuospatial tasks.

The program that learns visuospatial patterns of activity rests on the idea that visual routines can be viewed as repeating patterns of attentional state. I show how my language of attention enables learning by supporting the extraction, from experience, of such patterns of repeating attentional state.

Thesis Supervisor: Patrick Winston
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Rodney Brooks
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my friends Pawan Sinha, William Neveitt, Deniz Yuret, Pamela Lipson, Rahul Sarpeshkar, and the Zoo group, for many lively discussions about science and everything else under the sun.

I would like to especially thank my advisors Patrick Winston and Rodney Brooks for truly giving me the freedom to pursue my ideas.

Finally no words can express my gratitude to my parents whose emphasis on education has made this thesis possible.

Contents

1	Introduction	10
1.1	Building a humanoid robot	10
1.2	The Symbol System Approach	10
1.3	The Behavior Based Approach	12
1.4	What exactly do we want?	13
1.5	A platform that supports rich representation and inference	13
1.6	Origins of representational power	14
1.7	Visuospatial underpinnings of human cognition	15
1.8	Two Problems	16
1.9	Contributions of the thesis towards Problem #1	16
1.10	Contributions of the thesis towards Problem #2	18
1.11	Outline of the Thesis	20
1.12	Summary	20
2	Visuospatial problem solving	21
2.1	The need for a flexible spatial analysis mechanism	21
2.1.1	Can't we simply build a special purpose solution for each task? . . .	22
2.1.2	Why should we insist on a common framework?	24
2.2	The framework of a solution: Ullman's visual routines	24
2.2.1	Key issues in making the proposal work	25
2.3	Jeeves - A VRP for a blocks world domain	26
2.3.1	An overview of Jeeves	26
2.3.2	A critique of Jeeves	27
2.4	Sonja - A concrete-situated context for the application of visual routines . .	27
2.4.1	An overview of Sonja	27
2.4.2	A critique of Sonja	28

2.5	The RG system: Visual Routines for Binary Images	29
2.5.1	Primitives of the RG system	30
2.5.2	An example of a visual routine in the RG language	34
2.5.3	A critique of the RG system	36
2.6	What is a visual routine?	37
2.7	Open issues	38
2.8	Summary	38
3	An Architecture for Visual Routines	40
3.1	Visual routines for real images	40
3.2	Early Vision	41
3.2.1	Spatial derivatives	43
3.2.2	Motion	45
3.2.3	Color saliency and segmentation	45
3.2.4	Depth	46
3.2.5	Grouping of low-level features	46
3.2.6	Choice of the early visual representation	47
3.3	The visual routine architecture	47
3.4	Moving the focus of Attention	48
3.5	Establishing properties at the focus of attention	48
3.5.1	Establishing the Figure at the FOA	49
3.5.2	Establishing Figure attributes at the FOA	51
3.5.3	Establishing Local Spatial Relations between Regions at the FOA	51
3.5.4	Global spatial relations between markers	54
3.6	Selecting Locations	54
3.7	Attentional State	55
3.8	Examples of visual routines	55
3.8.1	Example 1	55
3.8.2	Example 2	60
3.9	Evaluation of the architecture	61
3.10	A review of models of Visual Attention	63
3.10.1	Models of attention as a region/object selection mechanism	64
3.10.2	Physiological evidence for a selectional role of attention	65
3.10.3	Psychophysical evidence for a selectional role of attention	65

3.10.4	Current work in modeling attentional processes	66
3.10.5	What's missing in all these models?	66
3.11	Summary	66
4	Learning Visual Routines	68
4.1	Visual routines are learned - not planned	68
4.2	Situated patterns of activity, and interactive emergence	71
4.3	Why learn emergent patterns?	72
4.3.1	Visuospatial patterns lead to expectations that impose top-down bias on behavior	72
4.3.2	Visuospatial patterns form the basis of language, and reasoning . . .	73
4.4	Attentional traces during exploration	74
4.5	Activating similar attentional traces in memory	76
4.6	Learning an expectation from activated attentional traces	80
4.7	Learning visual routines via reinforcement learning	80
4.7.1	An overview of the U-tree reinforcement learning algorithm	80
4.7.2	A critique of Reinforcement Learning	82
4.8	The role of visual routines in cognitive development	83
4.8.1	A brief overview of Piaget's theory	83
4.8.2	Perceptual grounding of object concepts	84
4.9	Summary	86
5	Contributions	87
5.1	The two motivating problems	87
5.2	Contributions of the thesis	87
5.3	Key issues of the future	89

List of Figures

1-1	The Two problems that this thesis deals with.	17
1-2	A Language of Attention	18
1-3	Expectations about an object passing behind an occluder may be in terms of changes of attentional state	18
1-4	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.	19
2-1	Examples of visuospatial problems	23
2-2	Yarbus [22] showed subjects a painting, asked them various questions about it, and recorded their eye-movements	24
2-3	The attention system and enumeration oracle	26
2-4	A visual operator from Sonja	28
2-5	Visual routines are interactions, not programs.	29
2-6	Four document-processing applications written using the same set of primitives	31
2-7	The area operator. Each location is labeled with the area of the region to which it belongs.	32
2-8	The Nearest Neighbor linking operator. Each location forms a link to the nearest occupied location.	33
2-9	Left: The location map. Right: The Voronoi diagram obtained from Nearest Neighbor links and Connected Component information	33
2-10	Only locations that have a particular value are selected.	33
2-11	Only regions colored by the seed locations are selected.	34
2-12	A summarizer function computes the summary over regions and delivers a mapping between region and value	34
2-13	The first two steps of the visual routine for extracting the smallest wedge in this pie-chart	35
2-14	The last four steps of the visual routine for extracting the smallest wedge in this pie-chart	35

2-15	The final steps of the visual routine for extracting the smallest wedge in this pie-chart	36
3-1	An overview of the architecture of the system	42
3-2	Early vision: spatial filter responses are computed across many scales . . .	44
3-3	Early Vision: Motion information	45
3-4	Bottom-up color salience	46
3-5	Framework for organizing the primitive operations	47
3-6	figure ground candidate from motion information	50
3-7	figure ground candidate from motion information: The dominant component of the filter response suggests the size and orientation of the figure at the FOA	50
3-8	Two kinds of spatial mechanisms are needed for capturing global and local relations between regions.	52
3-9	Matching local contexts: A sample local context is shown in (a), and the corresponding ratios matrix of this context is shown in (c). The image in (a) is matched to several previously acquired local contexts, the best matches are show in (d)	53
3-10	Markers on salient features are useful for monitoring their gross relative positions	54
3-11	This schematic shows the composition of the operations for detecting the object that is being pointed to	56
3-12	The system has to determine the object that the human is pointing to and look at it.	56
3-13	A “human template” is applied everywhere, the high value locations indicated in red suggest the locations at which a human might be present	57
3-14	58
3-15	The system uses the hand orientation to select an area of the room and saccade to the most salient blob there	59
3-16	An example of pointing where the system finally saccades to an object with low bottom-up saliency.	60
3-17	The system uses its expectation for a change of local context to look at the appropriate location for the re-appearance of the ball after having lost it. .	62
3-18	A schematic depiction of Treisman’s Feature Integration Theory. Treisman views attention as the ‘glue’ that binds features to objects.	65
4-1	Two patterns of activity (a-c) and (d-f), that are similar in some way, what is common? and how can we learn it?	69
4-2	$4\frac{1}{2}$ -month-old infants preferentially look longer at the “impossible event” in the bottom row, presumably because some expectation was violated. What is the representation of this expectation? and how was it acquired?.	69

4-3	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.	70
4-4	Biases about where to look and what to do at the focus of attention are necessary for exploration.	74
4-5	The attentional trace for an event generated by the bottom-up exploratory behaviors of Figure 4-4. Only the tracking behavior was active	75
4-6	The top row is the query segment of a “falling” ball, the next 10 rows show the 10 closest trajectories to the query trajectory in attentional state space.	77
4-7	The top row is the query segment of a “bouncing” ball, the next 10 rows show the 10 closest trajectories to the query trajectory in attentional state space.	78
4-8	The top row is the query segment of a ball “passing behind” an occluder, the next 10 rows show the 10 closest trajectories to the query trajectory in attentional state space. There was really only one other example in experience that was similar.	79
4-9	The top row shows the learned local spatial context for an object falling to the right. The bottom row shows a visualization of the template. Time goes from left to right in these sequences.	80
4-10	A schematic of a U-tree decision tree for organizing attentional state space. From McCallum[32]	81

Chapter 1

Introduction

Chapter Outline

This thesis is the result of a desire to build a robot with human-like abilities. This chapter

- Discusses some of the approaches to building a human-like robot.
- Outlines some promising new directions.
- Describes the two problems that this thesis deals with.
- Summarizes the contributions of the thesis.
- Gives a brief outline of the thesis.

1.1 Building a humanoid robot

Building a robot with human like capabilities has been one of the holy grails of Artificial Intelligence. After more than 40 years of trying to do this we are still far from the goal. During this period, the two main approaches that have dominated our efforts to build intelligent machines are the “Symbol Systems” and “Behavior based” approaches. A brief review of the contributions of these approaches will help us formulate some key questions that still remain to be answered.

1.2 The Symbol System Approach

The “Symbol Systems Approach” of classical Artificial Intelligence is based on the separation of reasoning from perception and action. It assumes that knowledge and the manipulation of knowledge, can and should be separated from the particular details of the physical body (the sensors and effectors). The framework in which traditional Artificial

Intelligence works is that perception delivers a symbolic¹ description of the world in terms of a fixed set of predicates like $\text{ON}(A, B)$, or $\text{LIKES}(X, Y)$. The “reasoning module” usually contains hand-crafted knowledge about the world expressed in some knowledge representation language (e.g. if-then rules, frames, scripts, or declarative statements in some form of mathematical logic). This knowledge, and the world description, are used by a problem solving mechanism to achieve some goal. The solution is passed to an execution module that performs actual actions in the world to achieve the goal. This is the basic framework.

This approach to achieving cognition arose from a conjunction of two factors - a somewhat restricted view of human intelligence, and a bias of what the available technology (digital computers and programming languages like LISP) was good at doing - symbol manipulation. The influence of our models of computers on our models of thought, or how the problem was defined to fit the tools, is well described by Brooks[8]. So we will briefly consider the other major influence, namely what the early researchers wanted when they said “human like intelligence”. Human intelligence was characterized by the ability to solve certain kinds of “difficult problems” - like solving a problem in formal logic [37], proving mathematical theorems, discovering scientific laws from data [36], and multicolumn subtraction [49]. The choice of these problems as representative of human cognition is interesting - it reflects a folk view of intelligence, most of humanity hardly ever engages in these tasks and the few who do are usually the ones regarded as “intelligent”, hence (the reasoning went) investigate these “high order intellectual processes” and you have the key to human intelligence. The tool for investigating these “high order intellectual processes” was introspection. For example a student would be given a problem in formal logic and asked to speak his/her thoughts aloud while constructing the proof. These verbalizations would be recorded and then analyzed for clues about the problem solving process [37]. After analyzing the problem solving process in the other problems too, it was found that if you strip away the particulars of the problem then you could explain problem solving in these limited class of problems as search in some space of symbols. This fact in conjunction with the bias of the available technology led to the introduction of “symbol manipulation” into a *definition* of intelligence. A physical symbol system was defined as a bunch of symbol structures (tokens, expressions) and some processes to create and modify the structures. The Physical Symbol System Hypothesis states that a physical symbol system has the *necessary* and *sufficient* means for *general intelligent action* [38]. The main problems with the traditional approach are:

- *Building representations for particular tasks instead of representations that could support a variety of tasks.* Consequently we end up with a collection of separate programs that play chess, prove theorems, or do medical diagnosis. There is nothing wrong with this if our intention is to build tools that can assist humans, however we should not fool ourselves into thinking that the collection of these programs says anything about cognition.
- *It evades knowledge acquisition.* It does not address the basic issue of how the knowledge (e.g. rules of chess, or behavior of liquids) is *acquired* in the first place. The practice of isolating domains and *hand-coding* facts about them is clearly evading the issue and one of the factors responsible for the performance of these systems not scaling.

¹There are many different notions of “symbol”. In this particular case “symbol” is used in a very narrow sense similar to the tokens and expressions of LISP.

- *The abuse of Abstraction.* Brooks [9] points out that coming up with descriptions like ON(A, B) from a scene (both the choice of that particular feature, and the detection of that feature) is the essence of intelligence and the hard part of the problem. He says

Under the current scheme the abstraction is done by the researchers leaving little for the AI programs to do but search.

- *Weak Inference.* In spite of the emphasis on problem solving, the systems exhibited very weak inference capabilities, they could hardly infer anything beyond what they were told. For example, if the system is trying to build a tower of blocks by putting B on C (described as ON(B,C)), and then A on B (described as ON(A,B)), and each block happens to be skewed a little to the right because of some error in positioning. The tower collapses and the description of the history (ON(A,B), ON(B,C)) contains no information about what went wrong. The description is in fact identical to descriptions of towers that did not collapse. The tendency of the knowledge engineer is to make the symbolic primitives more elaborate; by introducing degrees of ON-ness in the description, for example. This is just a temporary fix, and does not address the source of the problem. The weak inference stems from two sources. The first source of weak inference is that only one kind of representation is being used - the symbolic representation used in the previous example can only make certain types of information explicit (a qualitative representation of ON-ness) and will break when you attempt to use it for tasks for which it is not suited. The second source of weak inference is the artificial separation of perception from reasoning. Prematurely throwing away the perceptual representation of the image without knowing what's important in it for some task is a mistake. Certain kinds of reasoning (about why the tower collapsed) may be best performed using "perceptual" representations and processes.
- *All Knowledge may not have to be represented explicitly.* A robot can behave in a manner such that a human might attribute some knowledge and goals to it, but that doesn't mean that the robot should be built with an explicit representation of that knowledge or goals.

1.3 The Behavior Based Approach

The more recent approach to building intelligent systems is the behavior based approach pioneered by Brooks [9, 8, 7] as a reaction to the problems of the Symbol Systems Approach. The main motivation of the behavior based approach is that the ability to perceive and act evolved long before, and took much longer to evolve, than the abilities of playing chess, or proving theorems. Hence it is worthwhile to investigate simple perception action loops because they might give you some insight into more complicated forms of intelligent behavior.

In the behavior based approach one attempts to build robots that perform simple tasks in complex environments. The hope is that it is easier to start with a complex environment and some simple tasks, and then gradually start increasing the complexity of the task, rather than the other way around. The chief insights of behavior based A.I are:

- Embodiment and Situatedness - Investigate intelligence by working in the real world with real physical robots that directly sense the world. This prevents you from dealing with artifactual problems in simulated environments, and also lets you exploit constraints in the real world.
- Exploiting the complexity of the environment - There is a way to design robots that can lead to complex behavior with non-centralized representations of the world. The complexity of the behavior is a product of the dynamics of the interaction of the robot with a complex environment rather than some complicated internal representation of the world. The trick to designing robots that exploit the complexity of the environment is by designing a collection of behaviors that interact with each other via the environment. The behaviors themselves are very simple independent perception action loops tightly coupled to the environment.
- The world is it's own best model - The world is sensed continuously instead of maintaining and updating some complicated internal model.

With these insights the behavior based approach has been successful in developing a subsumption architecture methodology for incrementally building collections of behaviors. However, there is still a large gap between the cognitive abilities of the most complicated behavior based systems to-date and the human like robots that we would like to build. To be fair, it has never been claimed that the principles of subsumption architecture were alone sufficient either as an explanatory device, or as a constructive recipe for human-like intelligence. However, the question must be asked if a collection of hundreds of human-like behaviors, like looking preferentially at human faces, imitating gestures, e.t.c bring us any closer to understanding or building human-like intelligence. This brings us to the key question of what we want when we say “human-like intelligence”?

1.4 What exactly do we want?

What is it about the human brain that makes it so unique in the animal kingdom? How did human intelligence evolve? Can we find a way to understand how the differences in brains and bodies between species corresponds to the different cognitive landscapes that they inhabit? For instance, why is it that even a “genius” fly or frog could never be able to contemplate the solar system or plan for next week? In the context of the evolution of cognition these questions strike at the core of what it is to be human.

In the remainder of this section I will take a particular position on the question of what distinguishes human cognition. Doing so will help us pose some questions which must be answered if we are to build a humanoid robot. This thesis itself is a first step towards answering these questions.

1.5 A platform that supports rich representation and inference

The most striking aspect of human intelligence is the astonishingly broad range of things that we can represent and infer. For example we can have concepts like “solar system”,

or understand how a bicycle pump works, or prove a theorem in mathematics. One truly appreciates how peculiar it is that we can do these things when one observes that the representational capabilities of most animals are very tightly coupled to the environmental niche in which they evolved. Homo Sapiens on the other hand can represent things like “atom” or “striped elephant” - things that it may never have perceived, or may not even exist. The key point to remember is that *it is not our questionable expertise at any one of the tasks mentioned above, but that we can do them at all and that we can do so many of them that characterizes human intelligence*. Therefore, a fundamental question in Artificial Intelligence which we must answer if we are to build robots with human like intelligence is what is the source of this representational power? *How can we engineer powerful representational and inference capabilities in a robot?* It cannot be stressed enough that the idea is not to build specialized representations for individual task domains, but rather to build a representational platform that will allow the robot to learn to play chess, or answer a question about how toasters work without a human having to handcode any representations particular to game playing or qualitative physics. This, then is what I want when I say “human like intelligence”.

1.6 Origins of representational power

An important clue to the origin of representational and inference capabilities in humans lies in our evolutionary history. It seems unlikely that in the course of evolution we suddenly developed specialized neural substrates devoted to playing chess, or doing arithmetic. From an evolutionary standpoint it seems much more likely that such abilities are purely accidental byproducts of more basic previously evolved capabilities like representing space, manipulating objects, natural language syntax, and a representation of social relations [42], [5] - representations and processes that evolved as a response to the pressures of being embodied, and being in the company of other similarly embodied humans. The implication of this simple hypothesis is that you may be solving an integral for example², by using visuospatial operations on spatial representations of some symbols, and imputing syntactic categories and anthropomorphic roles to the components of the expression. Whether you believe this particular example or not is not important. The point is that some such explanation of our abstract cognitive abilities in terms of representations and processes meant for more mundane activities *must* exist. Clearly, ants, frogs, and monkeys have visual, motor, and communication systems too. Yet we know that even the smartest of them cannot even begin to grasp the concept of an atom - a concept that is well within the reach of an adult human. We need a deep understanding of what is special about our representational systems and their interactions that lets us do things that other species cannot.

If we somehow understood what it is about embodied representations and processes and their interactions that gives rise to the “accidental” byproducts of being able to do calculus or think about atoms, then we would be well on our way to understanding the source of human cognition.

In order to gain such an understanding, we need to

²I choose this example because doing mathematics is considered to be one of the abstract cognitive feats of the human brain.

1. Pick a small set of representations and processes that that originally evolved to facilitate certain tasks like navigating through space, or manipulating objects, or for social interaction.
2. Explain what it is about each of them that makes it flexible enough to be re-used in novel ways.
3. Explain how interactions between different representations (e.g. when a visuospatial pattern of activity gets re-expressed in a syntactic representation as an atomic symbol) creates a whole new level of complexity that we label as “higher-level” cognition.

I believe that pursuing such a research program will be very rewarding in our efforts to understand natural intelligence, as well as in building human-like intelligence. As a first step, in this thesis I look at some visuospatial representations and processes from the point of view of the program described above.

1.7 Visuospatial underpinnings of human cognition

What is vision for? The 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. Recognizing objects, and directing action are important functions of vision no doubt, however, visual processes may also 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.

The next natural question is how? What are the common underlying representations/processes in our visual system that support both the mundane tasks of getting about the environment, as well as the more “abstract” tasks like the ones mentioned above? The following three processes are likely candidates:

1. Detecting spatio-temporal regularities: The regularities or patterns that we see in the spatio-temporal behavior of objects 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[23]). Therefore, it is important to have mechanisms that detect spatio-temporal regularities.
2. Extracting spatial relations on demand: Visual processes that evolved to direct action should not be tied to extracting only particular features of the environment. While they should be capable of serving the visual needs of various behaviors such as tracking, or grasping, they should not be limited to only extracting the visual features required by those behaviors. The visual machinery needs to be more powerful, and capable of extracting a potentially unbounded number of visual features that go well beyond what the hardwired behaviors require. The spatial analysis machinery in humans has this kind of flexibility, thereby lending itself for use not only on a variety of real world spatial problems (e.g. does the equator pass through Zaire on a map? or what

is the human pointing to?), but also synthesized spatial descriptions (e.g. a spatial representation of your schedule for the day).

3. Synthesizing spatial descriptions on demand: This capability allows the application of the visual problem solving machinery to a hypothetical situation formed from the precedents and defaults that have been learned.

The following example illustrates how these three capabilities might work together. 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.

It is my view that 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.

1.8 Two Problems

The goal of this thesis is to propose and demonstrate mechanisms for the first two problems mentioned in the previous section - *extracting spatial relations on demand* and *learning spatial regularities from experience*.

The first problem that I deal with is to find a spatial analysis mechanism that is robust and versatile enough to handle a wide variety of spatial tasks, Figure 1-1(a) shows a schematic.

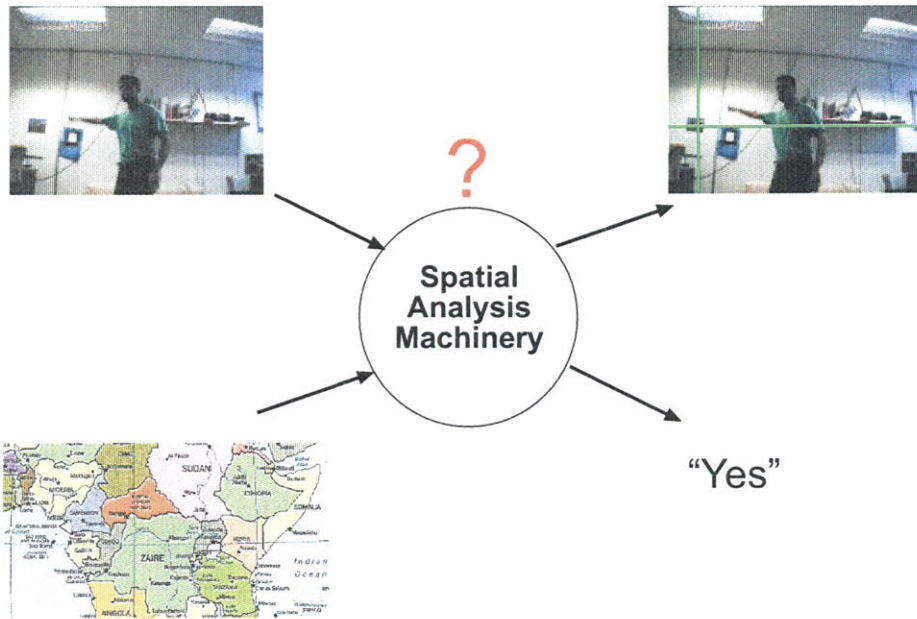
The second problem is to suggest a visuospatial representation for concepts like “fall” or “more”, *and* show how such representations might be learned from experience. Figure 1-1(b) shows a schematic. The representation of these spatial concepts will in fact drive the spatial analysis machinery that sought as part of the first problem.

1.9 Contributions of the thesis towards Problem #1

The main contributions of this thesis towards finding a robust spatial analysis machinery that can solve a wide variety of spatial problems, are as follows:

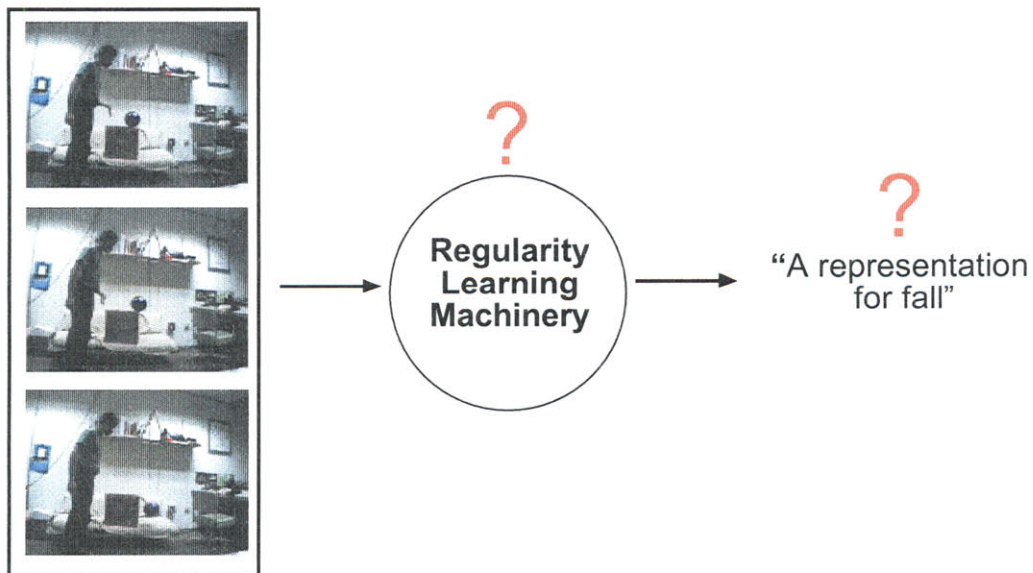
1. *A Language of Attention*. The machinery for extracting spatial relations and the mechanisms of visual attention are closely linked. I present a specific proposal for a spatial analysis architecture, the core of which is a new language of attention. Figure 1-2 shows a schematic. All procedures for extracting spatial relations are constructed by picking primitive operations from the three classes of operations. A

What is the human pointing to?



Does the equator pass through Zaire?

(a) Problem #1: Finding a versatile spatial analysis machinery.



(b) Problem #2: Finding a representation for visuospatial events, and a means of learning concepts in this representation.

Figure 1-1: The Two problems that this thesis deals with.

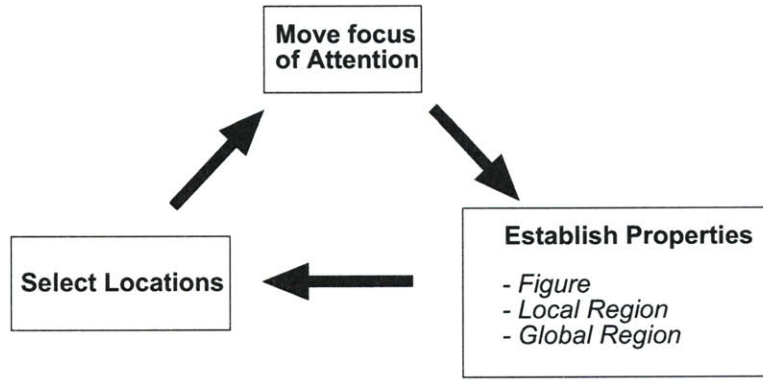


Figure 1-2: A Language of Attention

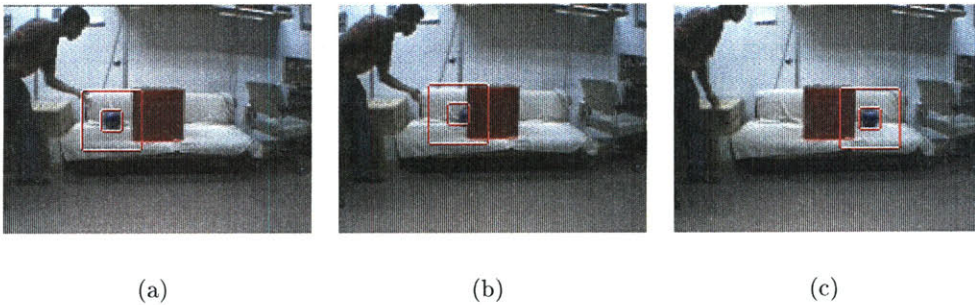


Figure 1-3: Expectations about an object passing behind an occluder may be in terms of changes of attentional state

typical procedure (called a visual routine) involves moving the focus of attention to a location, establishing certain properties and performing some operations at that location, selecting location(s) based on the properties just established, moving the focus of attention to one of those locations, and starting the cycle all over again.

2. *Attention is more than selection.* The proposed architecture extends and enriches prior models of Attention. The function of visual attention has traditionally been seen as "Selection", where one region is selected from many. In the proposed architecture (as shown in Figure 1-2), Selection is just one of the classes of operations, and is effective in spatial problem solving only in conjunction with the other classes.
3. *A versatile architecture for real-images.* As far as I are aware this is the first attempt to construct an architecture for real images with the explicit goal of being able to handle a very wide variety of spatial problems.

1.10 Contributions of the thesis towards Problem #2

The main contributions of this thesis towards finding a perceptual representation for concepts like "fall", and learning such representations is as follows:

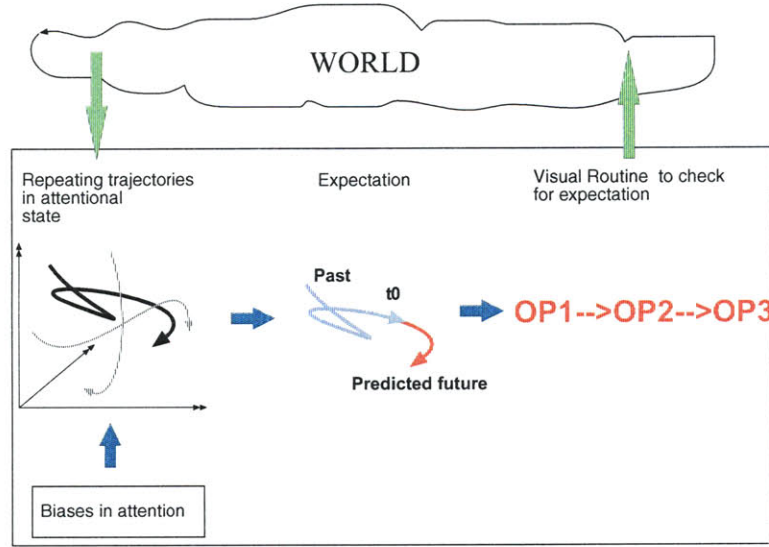


Figure 1-4: Regularities in the world and attentional biases repeatedly drive the attentional state through patterns. High frequency patterns become expectations. Expectations generate visual routines to check for the predicted part of the pattern.

1. *Patterns in Attentional State* I make a specific proposal for the perceptual representation underlying many visuospatial concepts, namely that it is in changes of “Attentional State”. While Niyogi[39] has recently made a similar proposal, and demonstrated it in a simulated world of simple geometric shapes, I present a model for real-world images, and also show how recurring patterns in attentional state may be learned. Figure 1-4 shows a schematic. Structure in the world and biases in attention lead to certain repeating trajectories in attentional state space. These patterns can be learned and constitute the perceptual representation for the event (e.g. an object falling), and also act as the expectation for future events.
2. *A novel parsimonious view of Cognitive Development.* Figure 1-3 shows a sequence where a ball passes behind an occluder and emerges on the other side. It has been shown by Spelke[52] and Baillergeon[3] that children have strong expectations for such events by the age of $4\frac{1}{2}$ months, and that a violation of these expectations (for instance if the ball were to emerge from behind another occluder which is some distance away) causes surprise. The question of interest here is what is the regularity that the child has learned and how is this represented? Spelke and Baillergeon’s explanations rely on something called “object-concepts” and knowledge of some sort of “intuitive physics”. I propose a much simpler explanation in terms of changes of attentional state, and show how such expectations can be learned.
3. *A promising candidate for perceptual grounding.* The question of “perceptual grounding”, i.e. if there is some perceptual representation that concepts like “fall”, or “more” bottom out in, has always been of great interest in Artificial Intelligence. The “patterns of attentional state” proposal made in this thesis is a concrete candidate for such a representation. I think that “patterns in attentional state” constitute a good representation because they are expressed in terms of the “language of attention”, the same language that is used to check for these events in a top-down manner. In other

words the representation of the regularity is in a form that makes it easy to check for it in the future.

1.11 Outline of the Thesis

In Chapter 2 I introduce the problem of extracting spatial relations on demand. I describe Ullman’s [58] visual routines proposal, and review subsequent efforts to implement this proposal in various domains. I concentrate on a system by Mahoney[29] for extracting spatial relations from binary images. While static binary images are a very different domain from the dynamic gray-scale/color images of real scenes, the binary image domain will serve to illustrate the compositional feature of my solution for real-world images.

In Chapter 3 I present a novel model of visual attention for generating visual routines. Visual routines are *sequences of operations for extracting spatial relations*, or more generally (as we shall see) *patterns of activity in attentional state*. This is the core chapter that lays out the families of spatial representations and operations used by my system. I compare my model of visual attention to models of visual attention in psychophysics, neuroscience, and machine vision.

In Chapter 4 I present a scheme for learning visual routines from experience. Whereas in Chapter 3 the focus is on the primitive operations of visual routines and how they can be strung together to extract different spatial relations, the focus in Chapter 4 is on how visual routines automatically come about from experience.

1.12 Summary

The long-term motivation is to build human-like cognition. Abstract cognitive feats may be accidental by-products of previously evolved representations and processes that evolved to support more mundane activities. A central challenge therefore, is to understand/model/dissect these representations to understand what makes them versatile and re-usable. I propose that the combination of 3 visuospatial mechanisms makes a very powerful and flexible system for the representation of “common-sense” spatial concepts, and spatial inference. In this thesis I will propose models for two of these mechanisms: extracting spatial relations on demand, and learning “common-sense” spatial regularities from experience.

Chapter 2

Visuospatial problem solving

Chapter Outline

In this chapter introduces the problem of extracting spatial relations on demand. I describe Ullman's [58] visual routines proposal, and review subsequent efforts to implement this proposal in various domains. concentrate on a system by Mahoney[29] for extracting spatial relations from binary images. While static binary images are a very different domain from the dynamic gray-scale/color images of real scenes, the binary image domain will serve to illustrate two important features:

- The power of composing visual routines by repeatedly choosing primitives from a small set of operations.
- The fact that visual routines work by repeatedly setting up successive frames of reference.

These two features will carry over to my visual routine architecture for real-images described in Chapter 3.

2.1 The need for a flexible spatial analysis mechanism

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 2-1 shows a small sample of problems that our visual system may be presented with during the course of a day. In Figure 2-1(a) one must determine the locations of the humans and the table and check for a particular spatial relationship between them. In Figure 2-1(f) segmenting the wedges is the hard part of the problem, because there is no clear demarcation between them. In Figure 2-1(h) 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 2-1(b) 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 [33] 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 machine-vision has been skewed towards object recognition with relatively little attention to the spatial mechanisms.

What kinds of visual mechanisms in humans make it possible to handle the spatial tasks in Figure 2-1 ? 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 ¹. Unfortunately, we do not know about the mechanisms at this level of detail. One of the problems that many different areas in the brain are involved in even a “simple” visuospatial task, and multicellular recording techniques are still in their infancy. In 1967 Yarbus [22] recorded the eye-movements of subjects after showing them images like the one in Figure 2-2 and presenting them with a specific question about the image. The sequence of foveations reveal several interesting features like the tendency to return repeatedly to a prior foveation point, however they shed no light on the kinds of computations performed at a particular point of foveation, or the kind of state maintained across foveations.

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.

This chapter and the next one deal with the problem of finding spatial analysis machinery that can account for this competence, i.e the ability to extract different kinds of spatial relations on demand.

Before exploring the problem in more detail we need to lay some issues to rest.

2.1.1 Can’t we simply build a special purpose solution for each task?

It is certainly possible - albeit after some work - to come up with algorithms for each individual task. In fact, if one is building a well defined application where a small fixed set of spatial tasks is required, and known in advance, it is practical to handcode individual solutions for each task. However, the main goal of this thesis is to make progress on the

¹This is of course a crude caricature of one pathway, there are doubtlessly other forward and back-projection pathways involved



(a) Is the table between the two people?



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



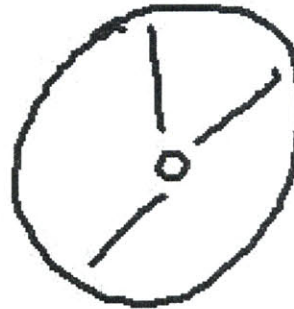
(c) Is the green rod touching the red box?



(d) How many cups are there on the green book?



(e) What is the human pointing to?



(f) Which is the smallest wedge in this pie-chart?



(g) Is the blue ball falling?

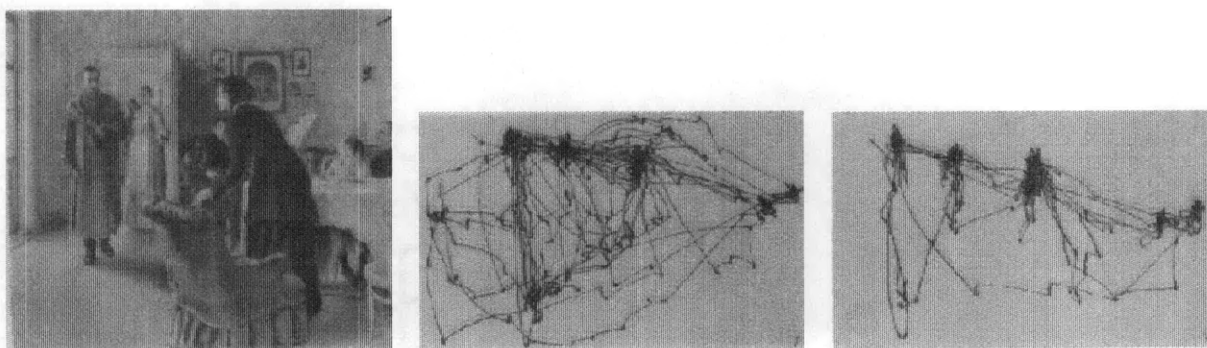


(h) Does every plate have a cup?



(i) Which person is closer to the door?

Figure 2-1: Examples of visuospatial problems



(a) "The unexpected visitor"

(b) Examine the picture

(c) Estimate the ages of the people

Figure 2-2: Yarbus [22] showed subjects a painting, asked them various questions about it, and recorded their eye-movements

issue of building a visual system for a humanoid robot or intelligent room that like the human visual system can handle an open-ended set of spatial problems. In other words, the goal is to find a common framework that solves all these problems as they arise. It is this insistence on an open-ended system that makes the problem hard.

2.1.2 Why should we insist on a common framework?

There are two good reasons why we should:

1. Scalability: A humanoid robot or intelligent room may have to deal with thousands of spatial problems in the course of a day. Writing a different program for each one is just not practical.
2. Learning: If we want to learn to detect different spatial events and relations, having a common vocabulary to express the commonalities is essential.

2.2 The framework of a solution: Ullman's visual routines

Shimon Ullman [58] 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.

2.2.1 Key issues in making the proposal work

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?

In the remainder of this chapter I will review some work on the topic of visual routines by Horswill, Chapman, and Mahoney & Rao. At the end of the chapter we will be in a position to assess their contributions in the light of the issues mentioned above, and see what are the open problems in coming up with a computational model for visuospatial problem solving. The next two chapters, which form the core of the thesis, directly address these open questions.

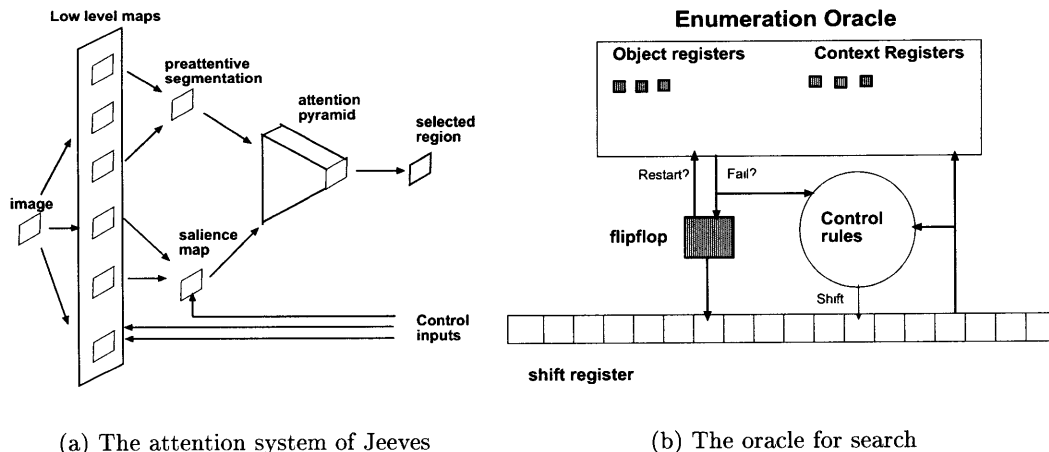


Figure 2-3: The attention system and enumeration oracle

2.3 Jeeves - A VRP for a blocks world domain

“Jeeves” is a Visual Routine Processor constructed by Ian Horswill [21] for performing “visual search” to answer simple conjunctive queries in a blocks-like world. In a typical instance the system examines a scene of colored blocks and finds a block “X” that satisfies a query like: $green(X) \wedge vertical(X) \wedge on(X, Y)$ where Y is some other block. The system also computes some simple spatial relations like “On”.

2.3.1 An overview of Jeeves

Figure 2-3 shows a schematic of the architecture of the system. The early-level vision consists of a set of color maps (R,G,B, R/I, G/I, B/I), intensity, and Laplacian edges. A preattentive segmentation stage carves the scene into regions of uniform color. A set of task-specific pixel-by-pixel control weights is used to compute a saliency map. The “attention” uses the saliency map to modulate the segmentation map, and selects the region with the highest integral of salience. The bounding box, centroid, and average low-level map values of the winning region are computed. Two important pieces of attentional state are a set of markers which hold the centroids of some regions, and the “Return Inhibition Map” which is a retinotopic map that masks out regions that should not be selected.

Given a query like $green(X) \wedge vertical(X) \wedge on(X, Y)$ The expression is preprocessed to indicate the variables that need to be enumerated and fed to an “enumeration & backtracking” automaton. The control logic of the automaton causes it to backup over the literals if a particular variable binding fails. The reader is referred to [21] for the details. The important points to note are that the logic variables are implemented by markers (which are essentially pointers to regions in the image), and that backtracking state (values of a variable that have already been tried) are maintained by the return inhibition map. After enumerating the various variable bindings, the solution to the query - if one exists - is left as a set of markers (regions) that satisfy the query.

2.3.2 A critique of Jeeves

The main focus of the Jeeves architecture is the control mechanism for variable enumeration and backtracking during visual search for regions that satisfy a certain conjunction of attributes. My main criticism of Jeeves as a viable architecture for a visual routine processor is that *the control architecture is not the hard part of most real-world spatial problems*. To illustrate this point consider the problem in Figure 2-1(e) of determining what the human is pointing at. This problem would have to be presented to Jeeves as the query $Pointing(Y, X) \wedge Human(Y)$. Jeeves would then try to enumerate over different regions in the image to find the ones that satisfy the expression, the implicit assumption here is that somehow there exist “procedures” for determining the predicates “Pointing()”, and “Human()”. *But coming up with visual operations for determining the truth of these predicates is the essence, and the hard part of the visual routines problem!* The Jeeves architecture has little to say about this matter because its focus is on the enumeration of arguments of the predicates, and not as much on the robustness/versatility of the mechanisms for determining the truth of the predicates. Finally, let us briefly focus on the spatial mechanism that Jeeves does use for determining the truth of predicates like $On(X, Y)$. Jeeves drops a ray downward from the marker for region X and checks to see it crosses a region Y, directly below. How versatile are such ray tracing methods for establishing different kinds of spatial relations? (Chapman [10] too heavily relies on ray tracing methods to establish spatial relations). The answer is that they are certainly useful for establishing a coarse spatial relationship between regions, but they are not sufficient for capturing somewhat finer local spatial relationships. For instance in Figure 2-1(c) drawing a ray between the centroids of the box and the rod is not going to help answer the question.

2.4 Sonja - A concrete-situated context for the application of visual routines

David Chapman’s “Sonja” [10] is a system that plays the video game Amazon. In this game the player controls the actions of a video icon that has to battle ghosts and demons while trying to pick up various objects in its two dimensional world.

2.4.1 An overview of Sonja

The machine’s inputs are a symbolic abstraction of a 2D world and occasional instructions from a human. The two major components of the system are a set of peripheral visual operators and a central system that controls application of the operators. At any given instant the peripheral operators compute various geometrical properties of the world. Figure 2-4 shows the inputs and outputs to a visual operator in Sonja. The central system specifies some of the operands of the visual operator and also a control signal that tells the operator to execute or not. The intermediate state that is referenced and updated by the operator is accessible to all the operators.

The central system consists of units called *proposers* and *arbiters*. Each proposer proposes the application of a particular visual operator with certain parameters. Each proposer has a condition under which its proposal is valid. There may be different proposers with different

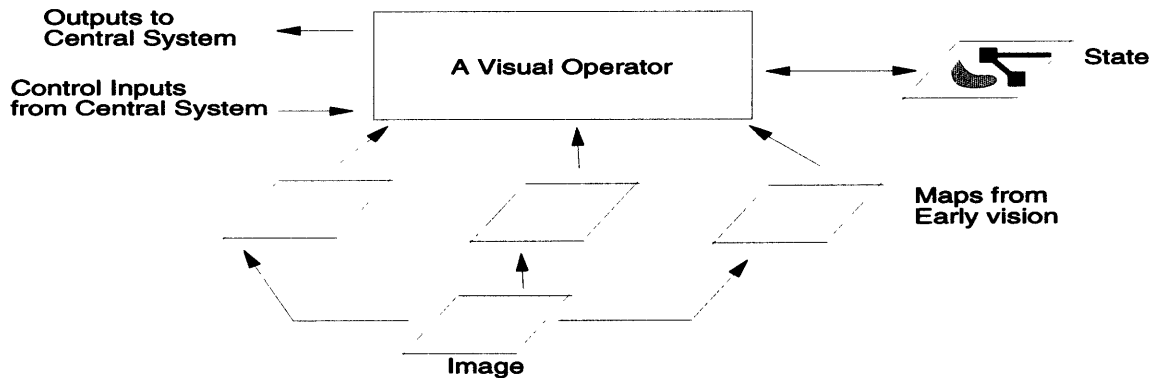


Figure 2-4: A visual operator from Sonja

suggested parameter settings for the same visual operator. If the conditions of all these proposers happen to be true then an arbiter implements some fixed override scheme on the various proposals to select one of them (a set of operands) for the visual operator. The central system is therefore a large unmodular net of proposers and arbiters whose input is from the visual operators and whose output at a particular time is the application of a visual operator with some set of arguments.

Figure 2-5 shows the interaction between the world, the visual operators, and the central system during the operation of Sonja. The x-axis multiplexes time and space. The trapezoids represent the entire central system. Only the operator that is finally chosen and its inputs and outputs is shown in the figure. The triangles represent delay elements that store information within the central system between clock ticks. The key point made here by Chapman is that *Visual Routines are interactional patterns, NOT programs*. In the figure some aspect of the world gets noticed by the bottom up visual operator OP26 which reports it to the central system. Based on this new development and the past history the central system directs another visual operator OP37 to report on some aspect of the scene. The information from OP37 and past history is used to take some action (OP99) which affects the world, and simultaneously give some instructions to OP37. Thus, Chapman points out the notion of a visual routine is an abstraction that an observer can make while looking at the interaction. In other words, nowhere in the central system is there a specification that the sequence OP26 - OP37 - OP99 - OP37 should be applied under some conditions, nor is it the case that some single unit planned this sequence of application of operators. This pattern of interaction just happened to dynamically arise given the states of the central system and the states of the world.

2.4.2 A critique of Sonja

The main purpose of Sonja is to demonstrate various aspects of the concrete-situated approach (e.g, situatedness, routine activity, dynamics of interaction) in a simple simulated world. Visual routines per-se are not the focus of the work, so it is not surprising that the mechanisms for extract spatial relations in Sonja do not scale to real-world spatial problems. The set of spatial relations between markers computed by the peripheral operators are not rich enough to handle many of the problems in Figure 2-1 for instance.

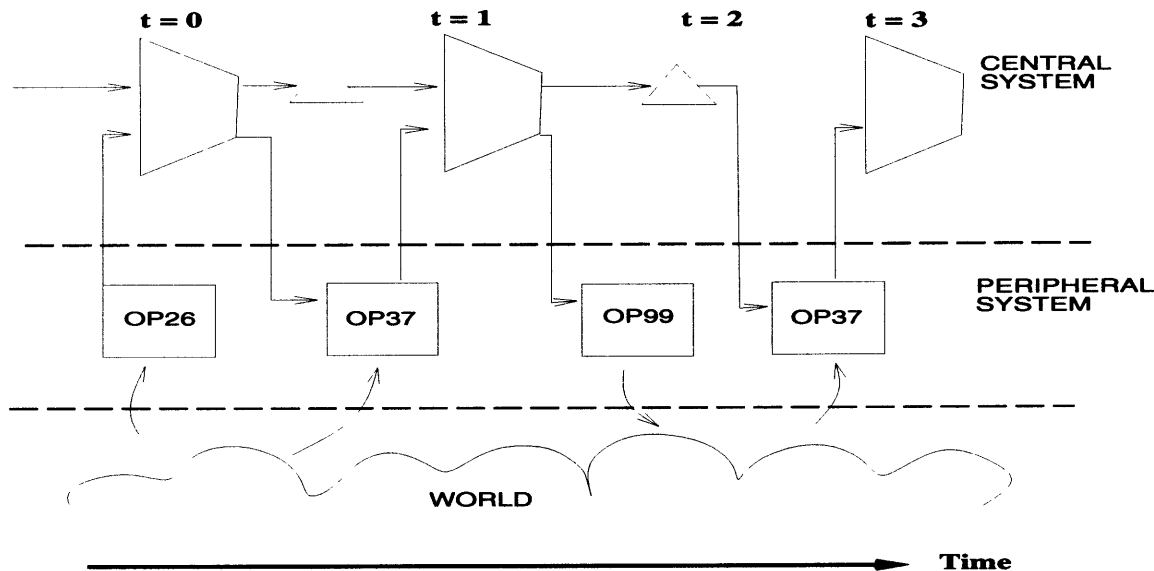


Figure 2-5: Visual routines are interactions, not programs.

However, Chapman makes two important contributions to the issue of composition of visual routines.

1. He emphasizes that there might be no central box that sequences or plans visual routines, but that the sequence of applications of various operators emerge from the interaction of the central system with the world. The very same observation was made by Brooks [6] in his criticism of traditional views about Planning. The behavior based approach that he used to build insect like robots does not have central representations of plans. Instead, it has a collection of mostly independent perception-action loops that are tightly coupled to the environment. The similarity between Chapman's insight about visual routines and Brooks's insight about planning can be seen clearly if we view each of the peripheral operators as a behavior and flatten out the central system as the various excitations and inhibitions between behaviors.
2. Chapman stresses the "routine" aspect of visual routines. Namely, the dynamics of the interaction between an agent and the environment give rise to certain routine patterns the agent falls into.

Finally I review a visual routine language developed by James Mahoney [29] for extracting spatial relations from Binary Images, and also some specific examples of visual routines written in this language implemented by this author.

2.5 The RG system: Visual Routines for Binary Images

The "Reverse Graphics" system developed by Jim Mahoney is a versatile set of spatial analysis primitives which can be combined together in different ways to implement a host

of applications. Figure 2-6 shows some of the applications implemented by Rao using a common set of primitives.

In Figure 2-6(a) the input to the system is a crude hand-drawn sketch of a pie-chart. The system analyzes the sketch to extract the relative sizes of the wedges and renders an equivalent pie-chart. Note that the system is not simply a rendering program that “cleans up” edges. It is extracting the wedges from the sketch in spite of the fact that there is no clean demarcation between the wedges.

In Figure 2-6(b) the input to the system is a crude hand-drawn sketch of a directed graph. The system analyzes the spatial relations between the elongated components and the circular ones to distinguish between edges, vertices, and edge-labels. The system distinguishes an edge label of “0” from a vertex by its relative position with respect to the edges. The system can use the graph structure extracted to control some application or simply render it as shown.

Figure 2-6(c) shows a “layout” application. The input to the system is a set of panels with a unique number of dots on each one, and a master page which contains a crude sketch of the desired layout. The system segments the master page into different regions and pairs each region with an input panel based on a count of the number of dots. Finally the panels are laid out in the corresponding regions.

Figure 2-6(d) shows an “editing” application. A transparent slide is overlayed on a document, and editing marks are made on it (e.g, delete, move, rotate). The slide and the document are scanned in, the system analyzes the marks, determines the portions of the document that are being referred to, and applies the transformations.

As mentioned before, the important point here is that each of these applications was written² using the *same set* of primitives. The following sections describe the set of primitives, and go through the complete visual routine for the pie-chart example in Figure 2-6(a).

2.5.1 Primitives of the RG system

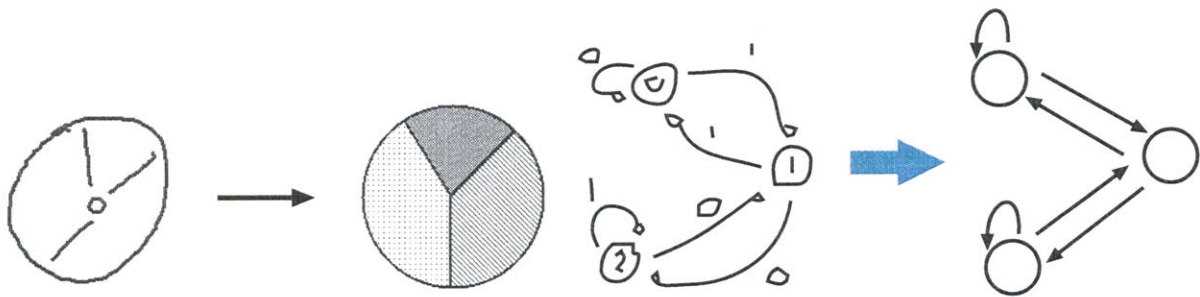
All primitive operations of the RG system are defined over three types of quantities.

The first type of quantity is called a *Property Map*. Property maps show the values that a particular property takes on at every location in the image. Orientation is one such property, for example. Property maps are defined across scales in a positionally exhaustive way. That is, moving to a coarser scale only increases the local area over which the property is computed, it does not change the number of locations at which the property is computed. Property maps are represented by heirarchichal levels of integer arrays.

The second type of quantity is a *Location Map*. Location maps show the locations at which some property assumes a certain value, for example all locations in the image that are red, or all locations in the image that are inside some region. One can perform Boolean operations over the domain of Location Maps. Location maps are represented by bitmaps.

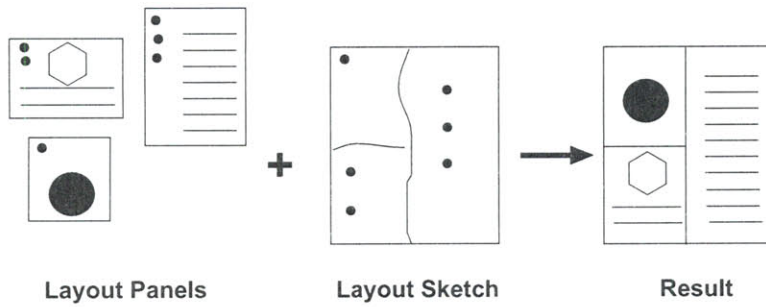
The third type of quantity is a *Summary Value*. It is the result of applying a summarizing function over particular regions of property or location maps. The perimeter is one such

²all these applications were hand-coded by the author, there were *not* automatically generated by the system

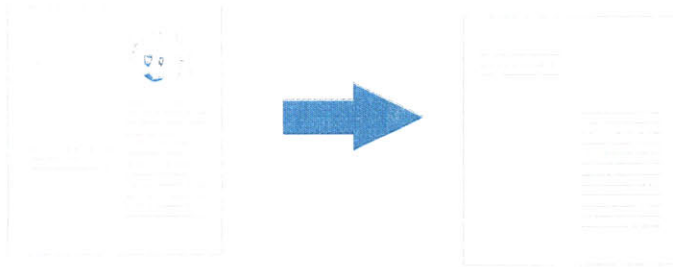


(a) visual routines extract the wedge sizes from the pie-chart sketch and hand them over to a rendering program

(b) visual routines extract the structure of the directed graph from the sketch and hand them over to a rendering program



(c) visual routines for the layout application interpret the marks on the layout sketch to determine what goes where



(d) visual routines for the editing application interpret the editing marks on the document and make the necessary changes

Figure 2-6: Four document-processing applications written using the same set of primitives

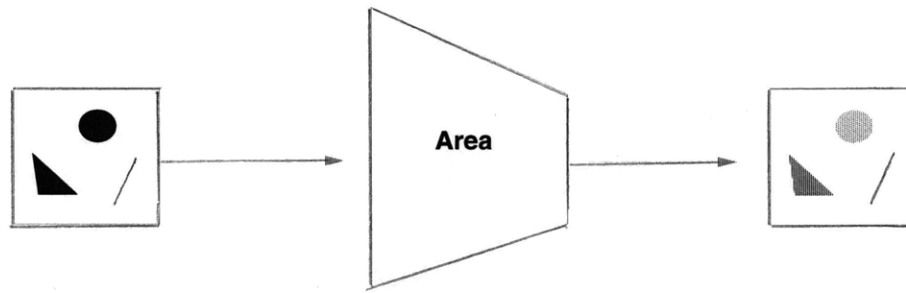


Figure 2-7: The area operator. Each location is labeled with the area of the region to which it belongs.

summarizing function. A mapping between regions and summary values is obtained as a result of applying the summarizing function.

There are three classes of elementary operations defined over the three basic data types.

1. $LocationMap \rightarrow PropertyMap$
2. $PropertyMap \rightarrow LocationMap$
3. $PropertyMap \times LocationMap \times Summarizer \rightarrow SummaryValue$

The first class of operators establishes properties at locations, the second class selects locations from Property Maps, and the third class summarizes property values over selected regions.

Establishing properties: Location map \rightarrow Property map

This class of operations takes a location map and at every location computes some property to yield a property map. There are several types of properties computed.

Region limited properties. A *region* is a set of connected locations on a location map. This class of operations compute some global property that is a function of all the locations that constitute a region, and then produces an image with each location of the region labeled with the “global” property. The *Area* operator for instance simply finds the areas of the regions in the location map. Each pixel of each region is labeled with the area of that region as shown in the schematic in Figure 2-7. Other examples are of region limited properties are elongation, connected components, and the angle of the axis of least inertia.

Diameter limited properties. At each location a property of all the locations within some constant radius around that location is computed. Orientation, size, and density are examples of such properties.

Relative Spatial properties. This class of operators computes one or more *links* at every location. A link from location A to B indicates the association of pixel A with pixel B because of some property, for example B might be the region nearest to pixel A. The *Nearest Neighbor* operator for instance takes a location map as input produces an image where each location has a link to the nearest location of the region nearest to it, as shown in as shown in the schematic of Figure 2-8. The positionally exhaustive link information is very useful

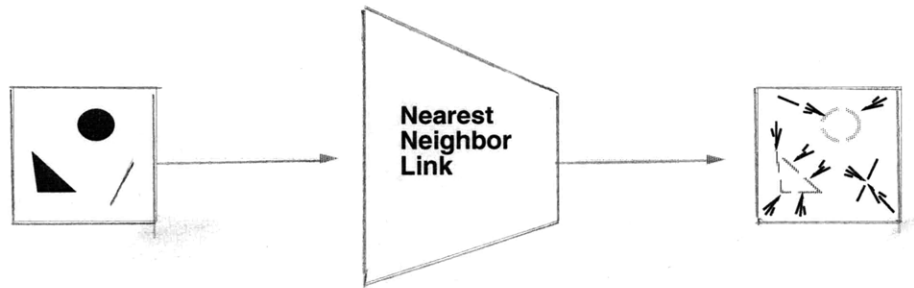


Figure 2-8: The Nearest Neighbor linking operator. Each location forms a link to the nearest occupied location.

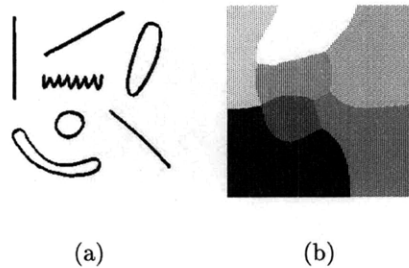


Figure 2-9: Left: The location map. Right: The Vornoi diagram obtained from Nearest Neighbor links and Connected Component information

in computing other properties like grouping. For example, the link information and the connected Components information could be used to compute a Vornoi diagram in one step in the following trivial way: Each location labels itself with the connected-component label of the location it is linked to. Figure 2-9 shows an example of a location

Selecting locations with certain properties: Property map \rightarrow Location Map

Select locations that have a particular property value This operation selects all locations that have a particular value from a property image as shown in the schematic of Figure 2-10.

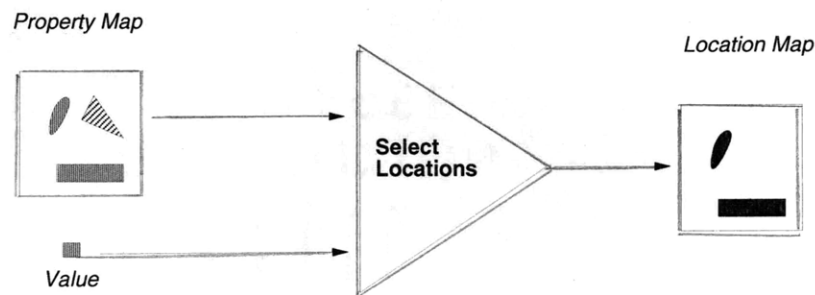


Figure 2-10: Only locations that have a particular value are selected.

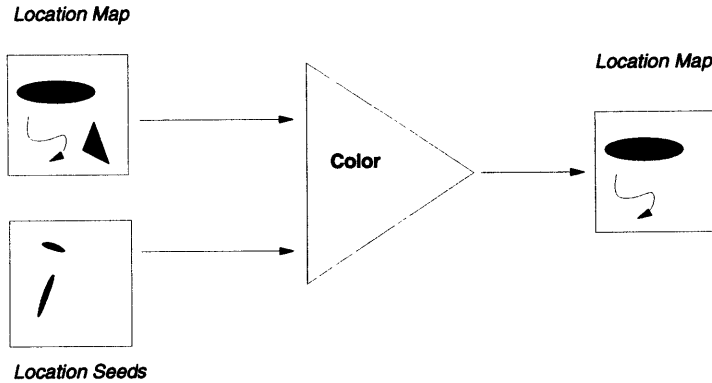


Figure 2-11: Only regions colored by the seed locations are selected.

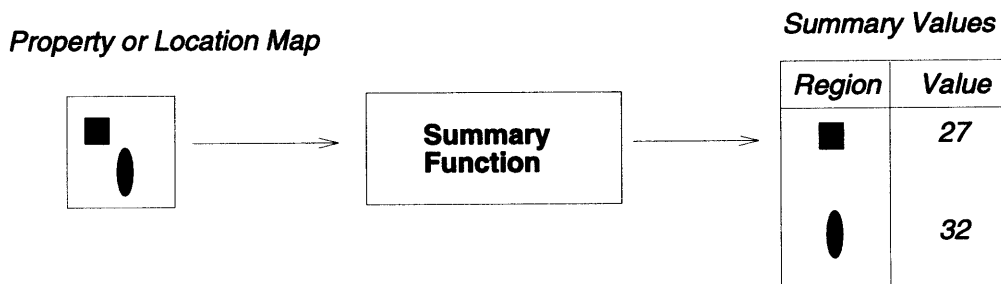


Figure 2-12: A summarizer function computes the summary over regions and delivers a mapping between region and value

Coloring This operation selects locations that are connected to a particular bunch of seed locations. This is accomplished by coloring from the seed locations and stopping at the boundaries defined by the regions. Figure 2-11 shows a schematic.

Computing the aggregate of a property over a region: Property Map \times Location Map \rightarrow Summary Values

The third class of operations summarizes information over regions to yield a mapping between regions and values. Any function that maps many values to one can be used as the summarizer function, for example area, perimeter, mean, mode, AND, OR e.t.c.

2.5.2 An example of a visual routine in the RG language

In this section I show an example of a visual routine implemented with the primitives described in the previous section. Figure 2-13(a) shows an image of a pie-chart. The problem is to “find the smallest wedge in the pie-chart”. The main goal of going through this example in detail are:

- To introduce the idea of a visual routine being a sequence of operations chosen from a small family of operations.
- To show the compositional power of the basis set of operations in the RG system.

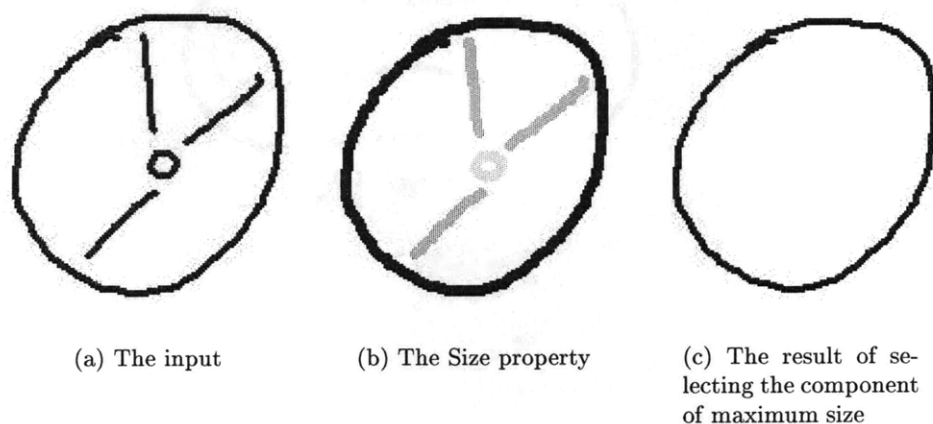


Figure 2-13: The first two steps of the visual routine for extracting the smallest wedge in this pie-chart

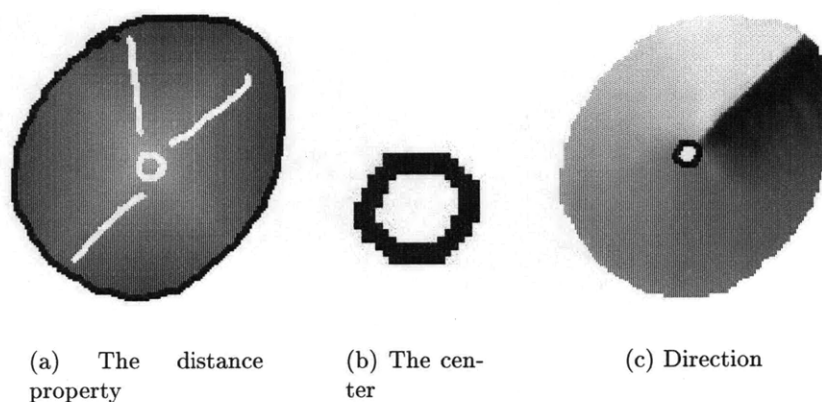


Figure 2-14: The last four steps of the visual routine for extracting the smallest wedge in this pie-chart

Areas: $\text{LocationMap} \rightarrow \text{PropertyMap}$ The first operation takes the location map and computes an “Area” property map. That is, each black pixel in the location map is labeled with the area (a number) of the connected component to which the pixel belongs, giving the result shown in Figure 2-13(b).

Select Max Size: $\text{LocationMap} \leftarrow \text{PropertyMap}$ Having computed the “size” property we now select the component that has the maximum size. This operation yields the perimeter of the pie-chart as shown in Figure 2-13(c).

Distance: $\text{LocationMap} \rightarrow \text{PropertyMap}$ Having extracted the perimeter of the pie-chart, we apply the “distance” property to the perimeter location map. The result as shown in Figure 2-14(a) is that each white pixel gets labeled with its distance from the closest black pixel.

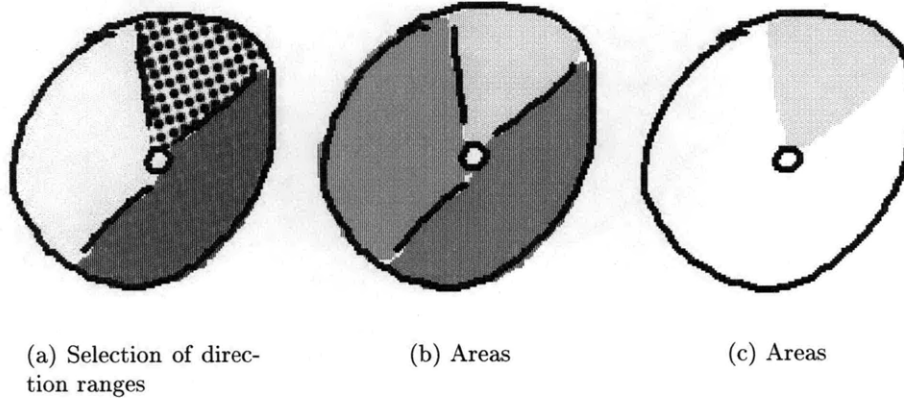


Figure 2-15: The final steps of the visual routine for extracting the smallest wedge in this pie-chart

Select max distance component: $\text{LocationMap} \leftarrow \text{PropertyMap}$ The component with the maximum values of the distance property is selected. This is done by looking at the distributions of the distance property over each connected component and selecting the one with the highest mean. The result as shown in Figure 2-14(b) is that the center of the pie-chart is selected.

Direction: $\text{LocationMap} \rightarrow \text{PropertyMap}$ Applying the direction property to the location map containing only the center component of the pie-chart yields a property map where each white pixel is labeled with its direction from the nearest center pixel. The result of this operation is shown in Figure 2-14(c).

Select ranges of direction: $\text{LocationMap} \leftarrow \text{PropertyMap}$ Each of the components now has a distribution of directions from the property map computed in the previous step. The components with unimodal distributions are selected, this yields the elongated spokes of the pie chart. Each of these spokes has a dominant mode in direction. All the pixels with directions between successive values of the modes of the spokes are selected thus yielding each individual wedge, as shown in Figure 2-15(a)

Areas: $\text{LocationMap} \rightarrow \text{PropertyMap}$ Having extracted the locations of each of the wedges, we apply the Area operator again (but this time on the wedge locations, not the locations in the original input of Figure 2-13(a)). This yields the property map shown in Figure 2-15(b) with each pixel labeled with the area of the wedge to which it belongs.

Select min Area: $\text{LocationMap} \leftarrow \text{PropertyMap}$ Finally we simply select the pixels with the minimum value from the area property map just computed, yielding the locations of the smallest wedge as shown in Figure 2-15(c).

2.5.3 A critique of the RG system

In section 2.2.1 we listed two key issues that a visual routines proposal must deal with. Mahoney squarely focuses on the first issue, namely of assembling a rich, expressive set of

primitives. How successful has he been at doing this?

As mentioned in section 2.5 the RG system has been used by this author to write visual routines for several document processing applications. The language is expressive, in that the programs for extracting spatial relations and properties are compact. For instance the program for computing the Voronoi regions of an image is less than five lines.

As demonstrated by the example in 2.5.2 the RG system derives its effectiveness from one simple but powerful idea: Spatial relations can be extracted by selecting locations, using them as a “frame of reference” to establish certain properties, then selecting other locations based on those properties and repeating this cycle over and over again. As we will see in the next chapter this idea also carries over to real images - albeit with different kinds of selection operations and properties.

2.6 What is a visual routine?

There are a several different notions of the term “visual routine” that have been used in the works described above.

The original definition of visual routines by Ullman [58] views them as a sequence of “primitive” operations drawn from a fixed basis set in order to extract some spatial relation from a scene. Mahoney and Rao’s visual routines for binary images are consistent with this definition. Chapman on the other hand views visual routines as patterns of visual activity that emerge due to interaction with a changing environment. He uses the word “routine” literally in the sense of habit, as rote or repeating activity, for example my tendency to look at the top-left corner of my screen every time I hear a beep, to check if the mailbox icon has changed color, is one of the routine visual activities that I engage in. Both Ullman’s and Chapman’s perspectives on visual routines are important and even though they may seem different from each other they are in fact very closely related and simply emphasizing *different modes of visual attention*³.

Chapman’s emphasizes the “pattern of visual activity” rather than the “extraction of spatial relations”. My visual routine for checking to see if I have email does not have much of a spatial component to it (other than looking at the top-left corner of my screen). On the other hand my visual routine for checking to see what somebody is looking at does need to have some spatial sophistication, in that I will have to locate the persons head, the eyes, get the gaze direction, and then look along that direction for some salient object. The point here is that in both of these examples of visual routines, Chapman’s emphasis is simply on the fact that they are both *patterns* of visual activity. Ullman on the other hand is concerned with fleshing out the spatial machinery of the visual routines that make it possible to locate the top-left corner of the screen or the gaze direction of the eyes.

Yet another characterization of visual routines views them as a sequence of operations to accomplish visual “search”. For example searching for conjunctions of features like a blue L among a field of green L’s and blue T’s [14]. Viewing a visual routine as a procedure for “visual search” is popular in the machine-vision literature. However, “search” is just one of the functions of visual routines. It would be hard and artificial to cast many of the spatial

³I will have more to say about this point in the next two chapters.

problems of Figure 2-1 as “search”.

2.7 Open issues

My criteria for evaluating a visual routines proposal are

- Is it a proposal for real-world images?
- How versatile is the spatial analysis machinery at extracting different kinds of spatial relations?
- Is there a scheme for automatically generating visual routines?

Among the works reviewed in the previous sections “Jeeves” is the only one that deals with real images. However, as we saw, the focus there was on the control architecture (enumeration and backtracking of logic variables) rather than on the mechanisms for extracting spatial relations. Mahoney and Rao do focus on spatial mechanisms, but their mechanisms do not carry over very well to real images because of the difficulty of figure-ground segmentation and the dynamic nature of the real world. Neither Horswill, nor Mahoney & Rao deal with the issue of how visual routines might be composed or automatically generated. Chapman has demonstrated some novel ideas about situated activity that provide an alternative approach to “composing” visual routines, namely visual routines just “emerge” from interaction with the environment. However his architecture for visual routines is very specific to the simulated world of the Amazon video game where the spatial relations that need to be extracted are known in advance.

In the following chapters:

- I present an architecture for visual routines for extracting a wide variety of spatial relations from real-world dynamic images. Another way of looking at this is a kind of “language of visual attention” that can be used to construct programs (visual routines) for extracting spatial relations.
- I also directly address the issue of how visuospatial concepts/expectations can be learned from experience and thereafter generate visual routines.

2.8 Summary

In this chapter we discussed the problem of finding a spatial analysis machinery for real-world visuospatial problems. The important requirement was that the machinery be versatile enough to deal with a variety of problems. There are several reasons for insisting on a common framework to handle all these problems: scalability, learning, and the fact that there *is* something that is common to all the tasks listed in Figure 2-1. We adopted a “visual routine” framework proposed by Ullman, and reviewed some of the work on visual routines. We saw that there are still several open issues regarding the construction of a versatile mechanism for real images. However the following ideas from the reviewed work do carry over to the solution that we discuss in the following chapters:

- The power of composing visual routines by repeatedly choosing primitives from a small set of operations.
- The fact that visual routines work by repeatedly setting up successive frames of reference.
- The fact that visual routines may emerge from interaction with the environment, rather than being planned.

Chapter 3

An Architecture for Visual Routines

Chapter Outline

In this chapter I present a *language of visual attention* for generating visual routines. Visual routines are *sequences of operations for extracting spatial relations*, or more generally *patterns of activity in attentional state*. This chapter lays out the families of spatial operations, and illustrates their use with some examples of visual routines. Finally, I compare my model of visual attention to other models of visual attention in psychophysics, and neuroscience.

3.1 Visual routines for real images

In the following sections I describe the architecture and primitives of a system for constructing visual routines, and then use this visual routine language to construct visual routines for some spatial tasks. In what sense is the proposed model a *language*? The word *language* is used because the model is generative in the same way that one generates sentences of a language by stringing together elements of syntactic categories. The only difference is that the sentences in the language of attention are *procedures* for extracting spatial relations.

A visual routine for even a “simple” task like deciding which of two objects is bigger, exercises many visual as well as non-visual faculties. In order to build an architecture that supports visual routines for a variety of spatial tasks one cannot avoid confronting many of the hard problems in vision like figure-ground segmentation, feature tracking, and shape recognition, to name only a few. One way of avoiding some of these problems is to simplify the environment by using children’s blocks, or painting the walls, or recording images from a clean desk, or a rotating turntable, or restricting the kinds of events that can take place. I have tried to avoid such simplifications. Instead, I have tried to use crude approximations to the hard problems. Another factor that has made the architecture design hard is that this is *not* a system with a specific task or goal. In fact, the whole point of this enterprise is to build a substrate that can support a variety of goals and spatial tasks like the ones in Figure 2-1.

How then should one evaluate the architecture that is going to be presented? The bottom line is how expressive are the set of primitives presented? That is, how large is the set of spatial tasks for which one could write visual routines using these primitives? We will return to this question towards the end of the chapter.

Figure 3-1 is an overview of the architecture of the system. The system can be divided into three distinct levels. At the first level (starting at the bottom) early-visual and preattentive properties are computed. The defining characteristic of this level is that the image properties are computed independent of the task at hand. The preattentive properties computed at this level are available to next one. The second level, which is the main focus of this chapter, shows the machinery for visual routines. At this level the top down goals and bottom-up information from the preattentive stage are used to fashion a specific visual routine. A visual routine in my framework, as mentioned earlier, is a sequence of operations drawn repeatedly from a basis set of operations. The third layer of the system consists of a collection of primitive behaviors each of which have specific biases about where to look next.

The following sections describe the first two levels of the architecture. The third level, which involves exploratory behaviors and mechanisms for learning patterns in visual activity, is the subject of the next chapter.

3.2 Early Vision

The early visual representation of the scene consists of the following components:

1. A 20 dimensional vector at every point in the image. This vector is obtained by applying 5 spatial filters across 4 different scales to the gray-scale component of the image.
2. A 60 dimensional vector at every point in the image. This vector is obtained by applying 5 spatial filters across 4 different scales to each of the 3 normalized color components of the image.
3. Motion information consisting of the location and direction of motion.
4. A color saliency map - indicating blobs of high color contrast

The above choice of features may be characterized as “blob-vision”, because it makes explicit blobs of various sizes and orientation, and their direction of motion. The decision to make blobs rather than regions (of arbitrary shape and size) explicit, is because blobs are much easier to compute and localize. Furthermore, segmenting an image into regions must require some top-down biases about the task. Segmenting the image into regions at an early stage may not only be hard but also ill-defined with purely bottom-up information (we return to this point in our discussion of color-segmentation later in this section).

The rest of this section describes the five components listed above and also the tradeoffs in making this particular choice of early visual features.

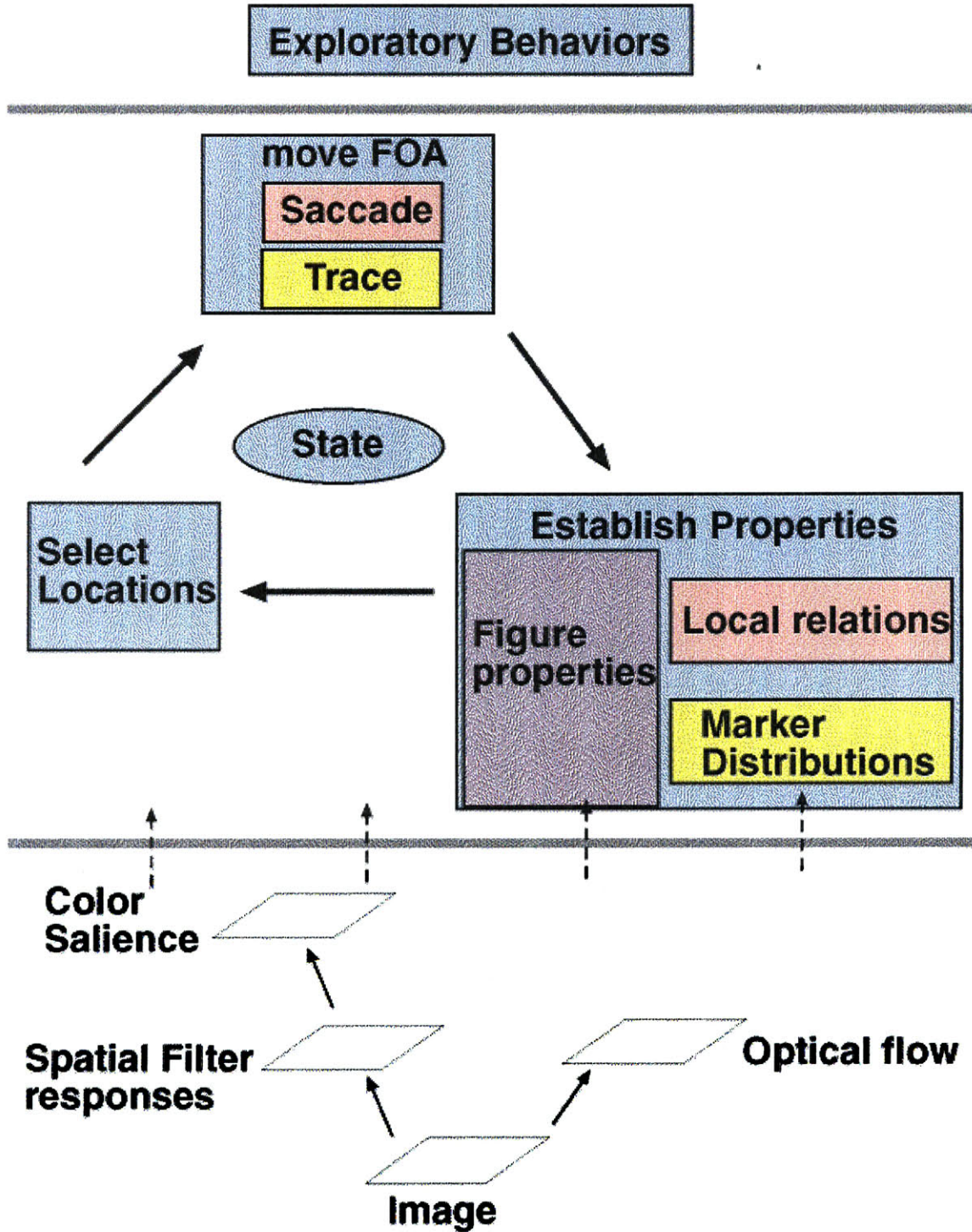


Figure 3-1: An overview of the architecture of the system

3.2.1 Spatial derivatives

For the spatial filters I use the steerable first and second order Gaussian filters as described by Freeman & Adelson [13]. The Gaussian function (shown with a sigma of 1 for convenience) is:

$$G(x, y) = e^{-(x^2+y^2)} \quad (3.1)$$

The first order derivatives in the 0 and 90 degree directions, i.e. derivatives with respect to x and y are:

$$G1_x = -2xe^{-(x^2+y^2)} \quad (3.2)$$

$$G1_y = -2ye^{-(x^2+y^2)} \quad (3.3)$$

The x and y derivatives completely span the space of all first order directional derivatives. In other words the directional first derivative in any arbitrary direction θ can be synthesized as a linear combination of the x and y derivatives:

$$G1_\theta = \cos(\theta)G1_x + \sin(\theta)G1_y \quad (3.4)$$

The second order derivatives G_{xx} , G_{xy} , and G_{yy} can be similarly computed. From these partial derivatives the directional derivatives in the 0, 60, and 120 degree directions can be calculated. The G_0 , G_{60} , and G_{120} responses span the space of all second order directional derivatives.

$$G2_\theta = k_1(\theta)G1_0 + k_2(\theta)G2_{60} + k_3(\theta)G2_{120} \quad (3.5)$$

where

$$k_j(\theta) = \frac{1}{3}[1 + 2\cos(2(\theta - \theta_j))] \quad (3.6)$$

using $\theta_1 = 0$, $\theta_2 = 60$, and $\theta_3 = 120$

The $G1_x$, $G1_y$, $G2_0$, $G2_{60}$, and $G2_{120}$ filters are applied to the image at different scales. The $G1$ and $G2$ filters correspond to edge and blob detectors respectively. Figure 3-2(b) shows the 5 filters and the result of applying them to an image at 4 different scales. Each point in the image therefore has a response vector of 20 elements. The feature vector at a point of interest in the image (shown with green cross hairs) is shown as a vector in Figure 3-2. These feature vectors are identical to the ones used by Rao & Ballard [47], except that their vectors have an additional set of four $G3$ responses at each scale.

Figure 3-2 shows the application of the spatial filters to the gray-scale component of the image. The same filters are also applied to the normalized color components of the image, that is to the R/I , G/I , and B/I images, where $I = (R + G + B)$. The 20 responses for each of these three color components is lumped together into one 60 dimensional vector.

Edge-detection. I have preferred to work with the raw spatial derivatives rather than edges, firstly because thresholding introduces unwanted parameters at an early stage, and secondly I am assuming that any regions of interest will be picked up by the blob detectors. True, this is an assumption about the environment, however the alternative of finding edges and stitching them together correctly to derive the boundary of a region is an unsolved problem, see Sha'ashua and Ullman[50], Alter and Basri[2] for such approaches. The choice that I have made does well for the minimal amount of computation it requires.

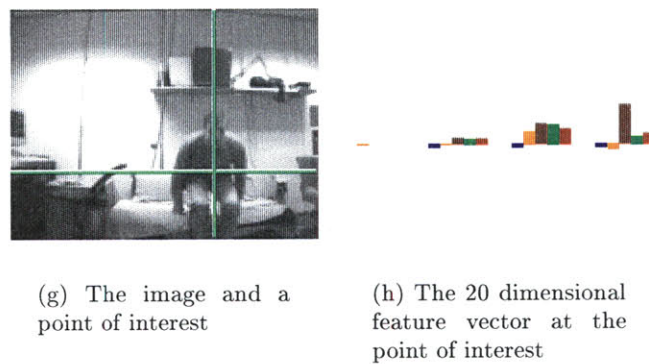
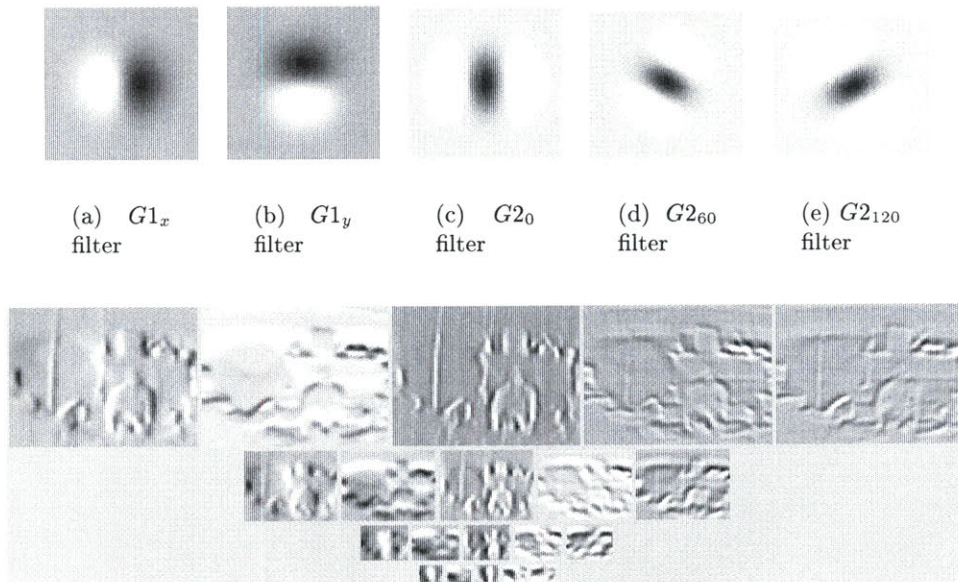


Figure 3-2: Early vision: spatial filter responses are computed across many scales

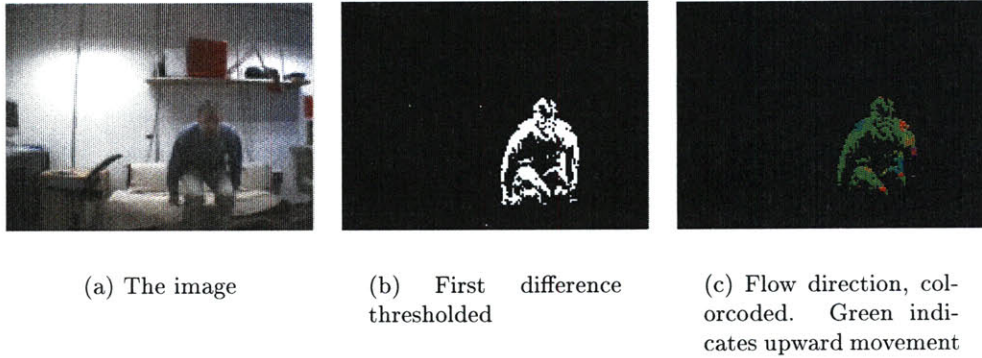


Figure 3-3: Early Vision: Motion information

3.2.2 Motion

The motion component of the early visual representation is shown in Figure 3-3. A simple correlation scheme is used to compute optical flow, as follows: after finding the first difference in brightness and thresholding it we get the locations in the image where there has been a significant change of brightness as shown in Figure 3-3(b). At each of these locations we grab a fixed sized patch centered on the first difference location in the previous image and search for that patch around that location in the current image. In other words we are establishing a correspondence between the previous image and the current one only at locations with significant changes of brightness. The correspondence gives the displacement vectors for the flow. The chief drawback of this method is that for very fast moving objects (or more precisely, when the displacement of object patches between frames is more than the search radius for the patches) the method gives totally erroneous results. One way to address this is to look at motion at multiple scales (just as we do for the spatial derivatives). I have not implemented a pyramidal scheme for motion computation but the reader is referred to Bergen and Hingorani [4] for such an approach. The multiple scale flow computation picks up large displacements at coarse scales, and propagates the flow down to finer scales. The correlation scheme that I use is only applied at a single scale but suffices for my purposes because I am interested in only the gross direction of motion of the patches on the object, not the precise magnitude of the displacement.

3.2.3 Color saliency and segmentation

Finally I compute a bottom-up measure of color saliency to highlight blobs of high color contrast. An example is shown in Figure 3-4. The color saliency is computed simply by adding together the response of the blob detectors to the color components of the image. This is clearly very simplistic, a more sophisticated measure of saliency would find the color gradient everywhere, do some form of color segmentation and take the contour integral of the gradient for each region, yielding a measure of saliency of the region. Horswill [21] uses such a method in his Visual Routine Processor architecture.

See Pal[40] for a review of color segmentation techniques. Color Segmentation - the bottom-up segmentation of the image into a bunch of uniform color regions, is something I did ex-



Figure 3-4: Bottom-up color salience

plore but discarded as being too unreliable. There are two reasons why color segmentation of the whole image at an early stage is not a good idea. Firstly we do not have a good understanding of color perception - the human visual system extracts something that is experienced as “color” which remains remarkably invariant to changes in illumination. In machine vision we are still far from extracting that invariant from the RGB signal, consequently color segmentation algorithms are still very sensitive to changes in illumination. Secondly, segmenting the image into regions *before* we know the task at hand is wasteful because the location(s) of interest for the task may not be any of the bottom-up color segments (for instance it may be where the tip of your pen meets the paper - rather than the paper or the pen by themselves). I feel that color salience is a more robust property at the early-vision level.

The decision about what kind of method should be used to compute color saliency, or whether we should do any color-segmentation clearly depends on what we want to use the results for. In the case of this system, the color saliency is computed just to suggest where next to look at. As the localization of regions is not my goal, using the responses of the color blob detectors to suggest *hot spots* where attention should be diverted, suffices for my purposes.

3.2.4 Depth

The system currently does not extract any depth information. Depth/Gradient information would certainly have been very useful for figure/ground segmentation [17], detecting occlusion, or making surfaces explicit [16]. However, given that making a visual routines architecture for 2D images is hard enough, I did not want to burden ourselves with the computation of depth/gradient information.

3.2.5 Grouping of low-level features

There has been considerable work on *perceptual grouping* [50][2][28] - the grouping of low-level features like a bunch of line terminators that are aligned, into a high-order feature. Such processes would have produced a richer early-visual representation. However I did not explore this avenue because I wanted to start with a simple set of features.

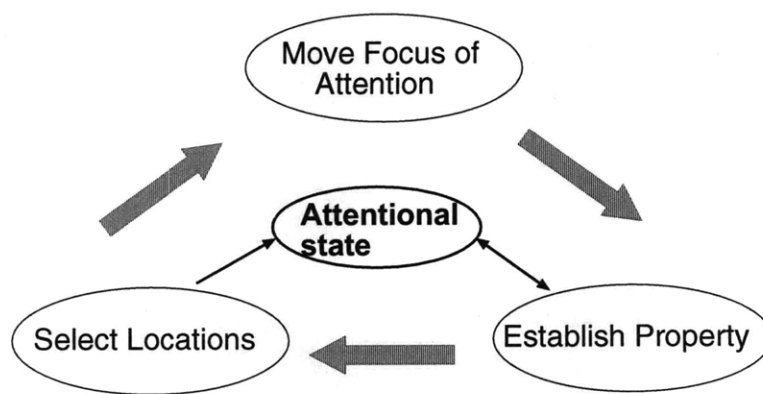


Figure 3-5: Framework for organizing the primitive operations

3.2.6 Choice of the early visual representation

So far in this section I have described the five components that make up the choice of the early visual representation that is passed on to higher stages. There is clearly a tremendous simplification here compared to the kinds of features that have been suggested (see the classic text by Marr [30] for instance) as components of a $2\frac{1}{2}$ D sketch.

There are many other well known early vision features that have been omitted from the choice that has been made. Treisman [14] and her collaborators have designed a “pop-out” paradigm for testing if a particular feature is part of the early-visual representation or not. The feature in question is presented in an array of distractors. If the feature “pops-out” i.e. clearly draws attention to itself irrespective of the size of the distractor array then it is regarded as being part of the early-visual/preattentive representation. A long list of such features is being compiled. Clearly the nature of the early visual representation is still an active area of research, and I have tried to choose as simple a set as possible without getting bogged down at this stage because the main focus of this thesis is on the next stage where task-specific visual routines are synthesized.

3.3 The visual routine architecture

In this section I describe the visual routine primitive operations and their organization. It is at this stage in the architecture that a top-down task specification and the bottom-up early visual information result in the synthesis of a visual routine from the set of primitive operations.

Before I describe the individual primitive operations, a high-level overview of the organization of the primitives is in order. Figure 3-5 shows a simple framework for organizing the primitive operations. All visual routines can be viewed as a sequence of operations chosen from three broad families of operations. One class of operations *moves the focus of attention*. Another class of operations *establishes properties at and with respect to the focus of attention*. A third class of operations *selects locations* by selecting ranges of values of the scene properties just established. The attentional state consists of scene properties that have been established during prior foveations.

A typical visual routine follows the flow of control indicated in the schematic, where we move the focus of attention to a new location, then establish one or more scene properties, then perhaps store some of these scene properties in attentional state and also if necessary compare the current scene properties to previously stored properties in the attentional state. Then we select some location(s) in the image based on the scene properties just established, move to the new location ... and the cycle continues (with not necessarily the same choices of scene properties and selection operations of course).

The preceding description of a visual routine simply stresses the sequential nature of visual routines and the three classes of operations. The following is a higher-level non-mechanistic view of what these sequences of operations are doing.

- *Visual routines establish local contexts, and use them as a local frame of reference to select the next place to look at.* For example a visual routine to locate a person's hand may first locate the person, and use the size, position of head, and torso to constrain the locations where the hand may be found.
- *Visual routines monitor changes in attentional state.* The representation of many abstract spatial concepts like "picking up something" or "falling off" is in terms of changes in certain spatial properties that are monitored.

For the next several sections you will see explanations of the individual pieces of Figure 3-5. While these descriptions are essential to understand the core of the system, it is easy to get lost in the detail and lose track of the fact that these operations *work together* to make a visual routine. Concrete examples of visual routines will be presented after the descriptions of the primitive operations.

3.4 Moving the focus of Attention

There are three ways in which the focus of attention of the system can be moved:

1. *Saccading* to a location. The focus of attention simply jumps from one point to another with an arbitrarily large displacement.
2. *Tracing* a feature. A feature like a curve is continuously traced, i.e with very short displacements in the focus of attention.
3. *Tracking*. The focus of attention continuously follows an object as it moves about.

3.5 Establishing properties at the focus of attention

While the early visual properties are computed everywhere, all the time, there are some operations that are applied only at the focus of attention (FOA). As mentioned in section 2.1, the eye-movement experiments of Yarbus tell us very little about what is going on at the focus of attention. In Figure 2-2(c) for instance when the focus of attention jumps from the central woman to the child seated at the table (or the other way around, the trace is not

Scene Properties at/w.r.t the Focus of Attention		Associated Operations	
1	Figure	1	<i>figure_ground_motion</i>
		2	<i>figure_ground_fv</i>
2	Figure Attributes	3	<i>get_size</i>
		4	<i>get_direction</i>
		5	<i>get_orientation</i>
3	Local Region Spatial Relations	6	<i>match_local_context</i>
		7	<i>match_regions_relative</i>
4	Relative Marker Distributions	8	<i>match_marker_distributions</i>

Table 3.1: Eight operations which can be applied at the focus of attention, divided into four classes by the scene properties that they are associated with.

labeled with time), what kinds of scene properties are being extracted at each location? In this particular case the size of the individuals may be extracted as clues to their age. The point here is that we have little idea about the computations being performed at the focus of attention to accomplish the task at hand. However, a specification of the vocabulary of operations at the focus of attention, that are available to the visual system, is essential for any realistic model of visual attention¹. In this subsection I describe my choice for the vocabulary of properties and operations at the FOA.

A note about my use of the terms “properties” and “operations”: properties refer to the kind of thing being extracted, for instance the “figure” at the FOA, or “the local spatial context around the figure”. Every property has one or more associated operations to extract it. Furthermore there are operations to compare properties (e.g. an operation to compare local spatial contexts). Table 3.1 shows a list of scene properties established at the focus of attention and the associated operations. The value in separating properties from operations is that even though future models of attention may propose different sets of operations, I believe that the properties with which they will be associated, will not be very different from the ones proposed here.

I now describe the operations of Table 3.1 in more detail.

3.5.1 Establishing the Figure at the FOA

Separating the figure from ground has been one of the major problems of machine vision. I take the position that there are several independent candidates for the figure at the focus of attention, and that it is up to the task to decide which one should be used. The system currently uses two independent bottom-up methods to determine the figure.

figure_ground_motion: Motion information is particularly valuable for figure/ground segmentation. The first temporal derivative in image brightness is thresholded to obtain the location of the moving points (actually only the edges), and then a connected components algorithm is run over the binary location image to segment the points into different groups based on their proximity to each other. The group/figure that is closest to the current

¹I discuss this important point at the end of the chapter, where existing models of attention are reviewed.

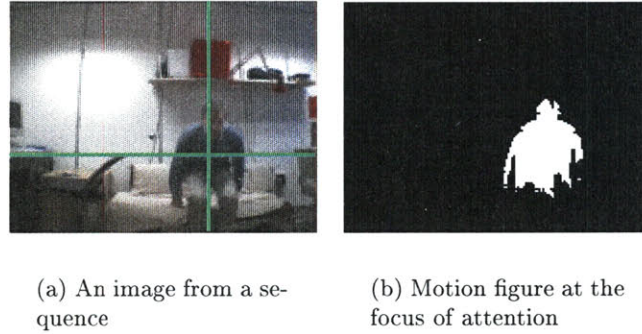


Figure 3-6: figure ground candidate from motion information

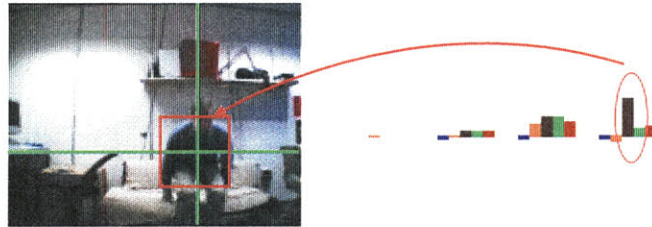


Figure 3-7: figure ground candidate from motion information: The dominant component of the filter response suggests the size and orientation of the figure at the FOA

point of foveation is chosen as the current figure. However, this group/figure is closer to the silhouette of the figure we really want because first the temporal derivative only yields information near spatial edges. An additional step is necessary to identify the points inside the moving figure. The phenomenon by which points inside a moving figure inherit the motion of the edges of the figure is known as “motion capture”. I use the heuristic of common-fate to implement motion-capture - points that are surrounded by moving points that belong to the same figure, are also labeled as part of the figure. Figure 3-6(b) shows an example of the result of the motion capture algorithm. The approximations used in the *figure_ground_motion* primitive have several limitations. First difference locations are not indicative of the motion locations for fast motion (large displacements) and the local proximity rule used to get the connected components is very sensitive to noise. It would be better to connect some higher-level feature (edge segments based on smoothness, or patches) rather than points to make it more robust. In spite of these problems the *figure_ground_motion* primitive gets the silhouette of the moving feature most of the time.

figure_ground_fv: If one views the image as a collection of blobs and edges of different sizes and orientations, then the dominant blob or edge at a particular point in the image is a crude but perfectly good candidate for the figure at that point. This blob or edge may be part of a larger figure, the arm of a human for instance. The dominant blob or edge at the point of foveation is trivial to extract; it is simply the maximum component of the 20 dimensional feature vector at the point of foveation. Figure 3-7 shows an example. I *define* whatever this returns as the figure at the focus of attention.

Clearly the two procedures for bottom-up figure-ground described above are selecting very different kinds of figures. The first one is selecting entire moving objects, the second one

is selecting salient blobs or edges, whether they are moving or not. Both the motion and spatial feature-vector candidates for the figure are available, the task will decide which one gets used.

An important bottom-up candidate for the figure which I have not implemented comes from selecting edges of the same disparity at or near the focus of attention. Grimson, Lakshmiratan et al, [17] show that the true role of stereo may not be to extract absolute depth (for which the error increases rapidly with error in angle of vergence) but to suggest which edges may belong to the same figure based on having disparity values close to the disparity value at the point of vergence.

Besides ignoring the other bottom-up candidates for figure-ground segmentation, I am also ignoring the role of top-down biases exerted by object models on figure-ground segmentation. It may not be correct to assume that figure-ground segmentation must precede object-recognition. The two processes may be closely intertwined, with bottom-up figure candidates selecting some object models which in turn insert top-down biases that “fill-in” parts of the figure. A discussion of these important issues is outside the scope of this thesis. The two bottom-up candidates of figure described above should suffice for my purposes.

Having established the figure at the focus of attention, we establish some attributes of the figure.

3.5.2 Establishing Figure attributes at the FOA

get_orientation The orientation of the figure is the dominant orientation of the spatial feature vector.

get_size The size of the figure is simply the scale at which the dominant orientation exists.

get_direction Assuming that the figure has been tracked for the past few frames, the direction of motion of the figure is the average direction over the past few frames.

Note that all the figure attributes are of the dominant blob/edge extracted by *figure_ground_fv*.

We now move on to discussing two families of operations: one for extracting spatial relations between *regions*, and the other for spatial relations between *features that can be approximated by point markers*. Figure 3-8 motivates the need for two types of operations for capturing relative spatial relations. Figure 3-8 shows two regions that are in different spatial configurations. In the case of (a) when the regions are “far” apart, the distance and direction between the centroids of the regions is sufficient for capturing the spatial relationships between the regions. However, when the regions are close together as in (b), the centroids are poor replacements for the regions, a different representation will be needed. One way of looking at the two types of operations is in terms of *local* versus *global* measures of spatial relations, where “local” is defined with respect to the size and extent of the region of interest.

3.5.3 Establishing Local Spatial Relations between Regions at the FOA

One class of operations at the FOA is devoted to establishing local spatial relations between *Regions* in the image. In particular we will consider the local spatial context - the portion of the image in the immediate vicinity of (and including) the figure. The size of the local

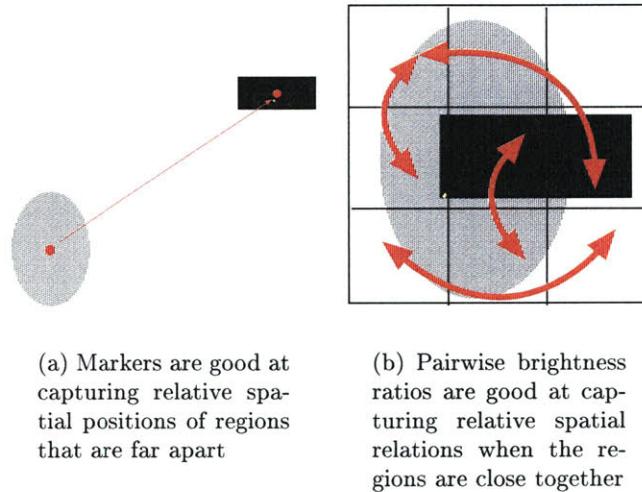


Figure 3-8: Two kinds of spatial mechanisms are needed for capturing global and local relations between regions.

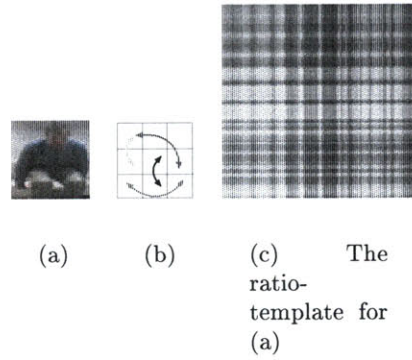
context is decided by the size of the figure. In the case of the figure from motion the bounding box of the figure is used as the local context, as shown in Figure 3-6(a). In the case of the figure returned by `figure_ground_fv`, the local context is the bounding box of the dominant spatial filter. Figure 3-9 shows several examples of local contexts around `motion_figures`.

Why is the local spatial context important? It is an empirical fact about our world that it has repeating local structure. Several spatial relations like being “on” something, or “touching” something, or “inside” something, are examples of purely local spatial relations, in the sense that the spatial extent of the receptors needed to detect these relations between two regions, does not grow with the size of the regions.

One way of capturing local structure, i.e. the distribution of “stuff” around the point of foveation, is to compute pairwise brightness ratios between patches, as used by Sinha[51]. Figure 3-9(b) shows a schematic. The image is divided into a mosaic of patches, and the ratios of the brightnesses between every pair of patches is computed, giving a cross-ratio matrix as shown in (c).

match_local_contexts: The ratio-matrix acts as a template for the patch, and can be used for matching. Figure 3-9(c) shows the top fifty matches to the image in (a). All of these local contexts were acquired autonomously by the system during scene exploration². Note that the template does a good job of retrieving humans in similar postures. The retrieval of the blue-ball in the last few images is understandable given that the silhouette of the rising person matches the silhouette of the ball. The use of ratio-templates was proposed by Sinha[51] as a representation for objects.

²I will discuss scene exploration in the next chapter



(d) The best matching local contexts

Figure 3-9: Matching local contexts: A sample local context is shown in (a), and the corresponding ratios matrix of this context is shown in (c). The image in (a) is matched to several previously acquired local contexts, the best matches are show in (d)

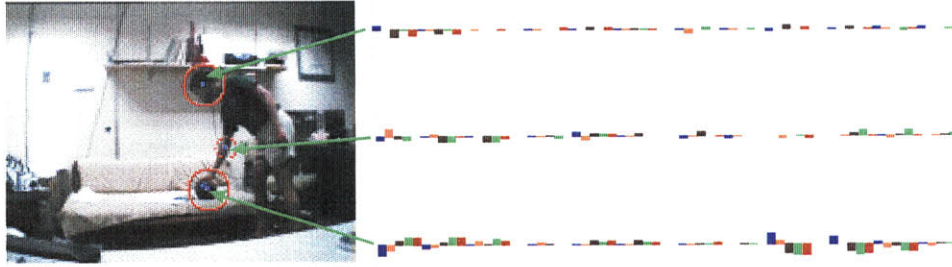


Figure 3-10: Markers on salient features are useful for monitoring their gross relative positions

	Selector	description
1	select_color	Select locations that match a particular color
2	select_fv	Select locations that match a feature response vector
3	select_blob_atscale	Select all blobs at a particular scale
4	select_motion	Select all locations moving in a particular direction
5	select_region_relative	Select locations that are in a particular spatial configuration with respect to a region
6	select_marker_relative	Select the likely locations for a marker based on its relative position to other markers

Table 3.2: Selection operations

3.5.4 Global spatial relations between markers

Another class of operations at the FOA is devoted to establishing global relative spatial relations between *widely separated regions* in the image. As mentioned earlier, sometimes regions may be simply characterized by a single “point-marker”, and the relative spatial relations by the distance and direction between point markers. Figure 3-10 shows an example where markers on the head, arm, and object, are used to characterize a “pickup” event. The relative positions of markers is useful in capturing certain invariances in scenes and events. The relationship between the markers shown in Figure 3-10 is not very different even in different instances of “picking-up” objects.

3.6 Selecting Locations

Operators that select locations in the image take a property image, and a range of values as input and simply selects locations in the property image that fall within that range. Table 3.2 shows the selection operations used by the system. Selection operations may be composed - for example one can select all blobs of a particular size within a certain distance and direction of the point of foveation.

3.7 Attentional State

Attentional state consists of:

1. The current figure, its direction of motion, and its local context.
2. Previously selected regions, their figure attributes, and their local global spatial relations with respect to the current figure.

This concludes the description of the architecture of the system. We are now finally ready to see how the families of operations described in the previous sections work together.

3.8 Examples of visual routines

In this section I present some examples of visual routines which will use the operations described in the previous section. While the tasks may appear to be different, the visual routines described in this section all use the same framework (shown in Figure 3-5) of foveating to a location, establishing some properties with respect to the point of foveation, selecting certain locations, and then moving on to the next foveation while monitoring changes in attentional state.

It should be clarified that in the following examples the sequence of operations that make up the visual routine were programmed, they were not automatically sequenced by the system. The automatic composition of visual routines by learning from experience is the subject of the next chapter. The emphasis in this chapter is on the basic attentional framework for visual routines.

3.8.1 Example 1

Figure 3-12 shows two scenes where a human is trying to direct the attention of the robot/room to a particular object. The task is to construct a visual routine that will direct the system to the object being pointed to, by stringing together operations described in the previous sections. Figure 3-11 shows a schematic of the order in which the operations are applied. The goal here is not to construct the best pointing-gesture-recognition-program, but to show the utility of the particular choice I have made for the *language of attention*. This example highlights a point made earlier in this chapter (in section 3.3) about what visual routines do: *they establish local contexts, and use them as a local frame of reference to select the next place to look at.*

Find the human At first the system must find the human in the scene. It uses a "human template" that is formed out of G1G2 kernel feature vectors. Figure 3-13(a) shows the points that make up the template (each point has a G1G2 feature vector associated with it). Such templates are very similar to ratio-templates [51] and are in a form suitable for learning from experience. Figure 3-13(c) shows the results of applying the template to all locations in the image in (b). The red locations in (c) indicate a high match at the center of the human.

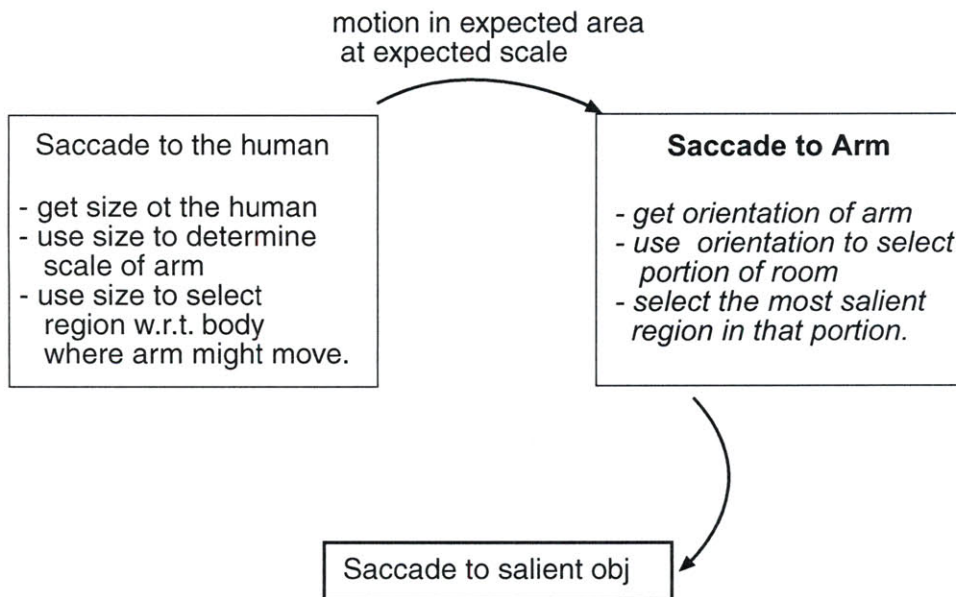


Figure 3-11: This schematic shows the composition of the operations for detecting the object that is being pointed to



Figure 3-12: The system has to determine the object that the human is pointing to and look at it.

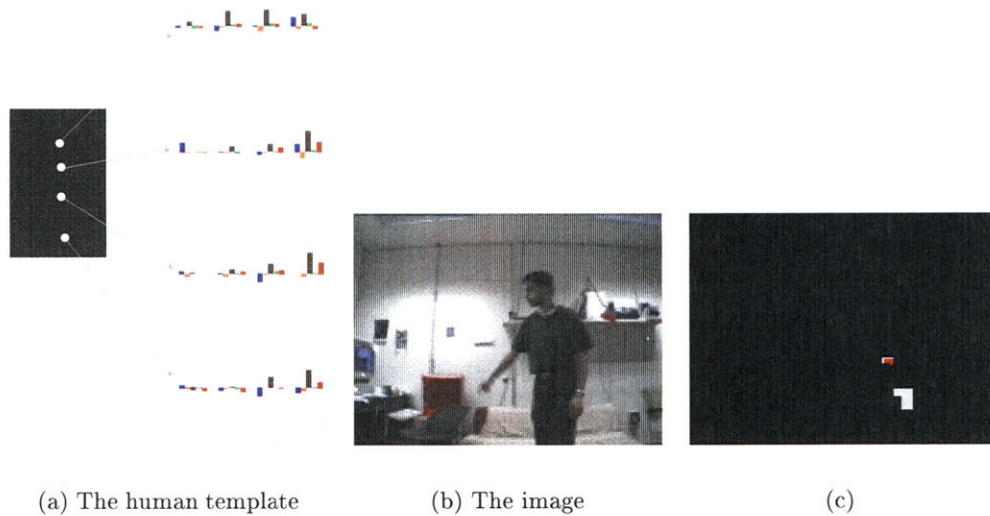


Figure 3-13: A “human template” is applied everywhere, the high value locations indicated in red suggest the locations at which a human might be present

Saccade to the human Figure 3-14(b) shows the system saccading to the location with the highest match.

Having isolated the human, the next task is for the system to isolate the hand with respect to the body, and then shift attention to the hand.

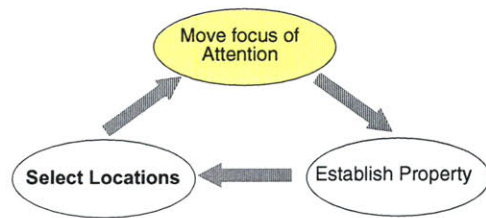
The next task is to detect if the human is in the process of pointing at something. The system attends to motion within a certain radius of the point of foveation. Furthermore, it uses the size of the human it is looking at to select the scale at which it must search for the arm.

Figure 3-14(d) shows the area around the point of foveation being monitored for motion. Figure 3-14(e) Shows the subset of monitored locations at which motion is detected.

Select “arm-scale” blobs in the motion region The size of the human determines the scale at which the arm is likely to be detected. Only blobs at this lower scale (at which the arm is likely to be detected) are attended to in the moving regions. Figure 3-14(g) shows the likely locations of the arm.

Saccade to the arm Figure 3-14(i) shows the system at its new point of foveation on the arm. Having saccaded to the arm - which is now the focus of attention - the orientation of the arm is determined.

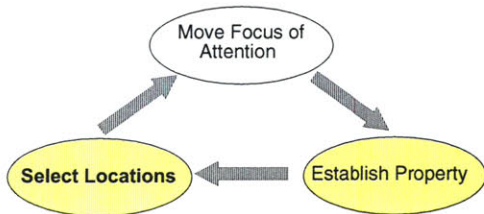
Select a region in the room Figure 3-15(d) shows the region selected in the room. The system’s attention now shifts to the selected region of space. Within this space it selects a blob that is salient in its color contrast. Figure 3-15(e) shows the bottom-up color salience everywhere in the image. Figure 3-15(g) shows the system finally saccading to the most salient location within the selected region. In this particular example the location



(a)



(b) The system saccades to the human



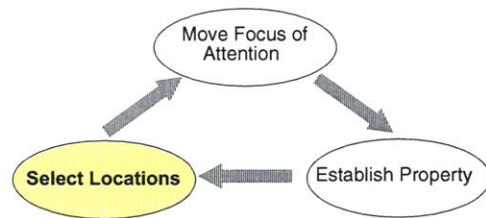
(c) Detecting the pointing action



(d) Select Motion within a certain radius of the point of foveation



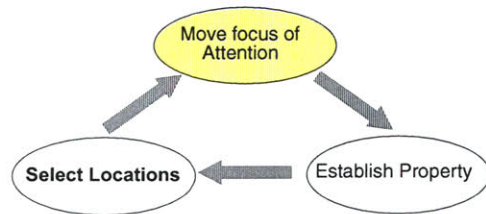
(e)



(f) Detect the arm



(g) The hand is selected among the attended blobs

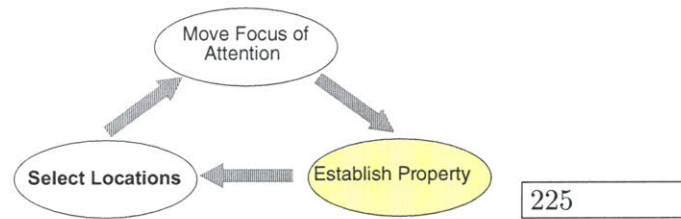


(h) Saccade to the arm



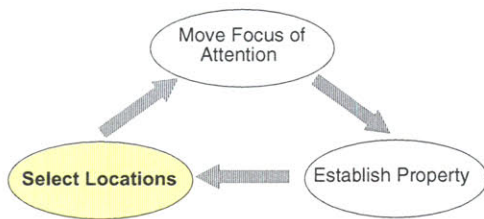
(i) The system saccades to the hand.

Figure 3-14:

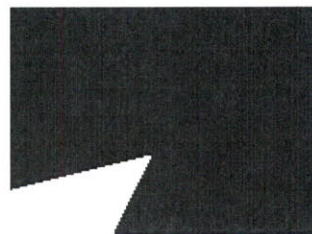


(a)

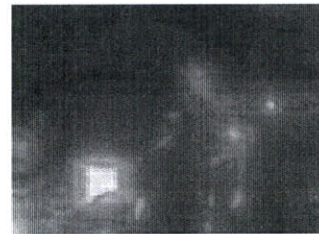
(b) Determine the orientation of the current figure - the arm



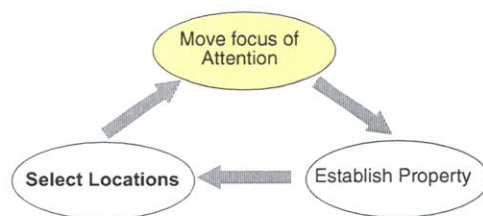
(c)



(d) Select an area of the room



(e) Find the most salient blob in the selected area



(f)



(g) Saccade to the salient blob

Figure 3-15: The system uses the hand orientation to select an area of the room and saccade to the most salient blob there

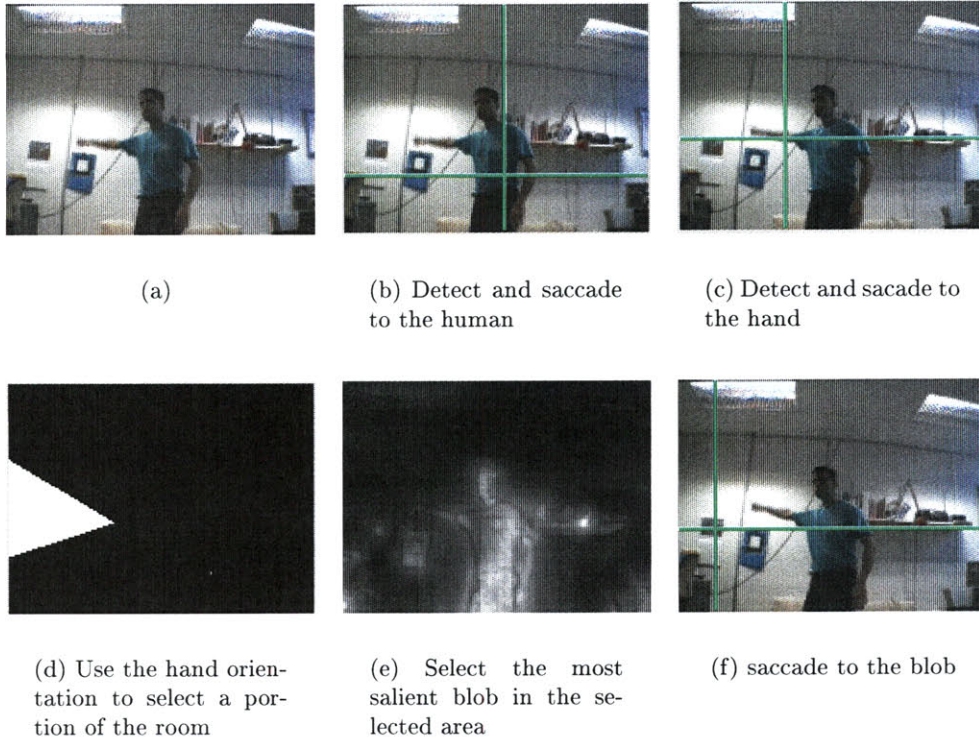


Figure 3-16: An example of pointing where the system finally saccades to an object with low bottom-up saliency.

finally saccaded to happens to be the most salient in the image, however as Figure 3-16 shows this need not be the case. Even though the bottom-up saliency of the calendar in the left of the image is low compared to other parts of the image, the selected cone shaped region can be viewed as imposing a top down modulating effect leading to the selection of the calendar as the most salient location.

The previous example of a visual routine emphasizes the framework described in Figure 3-5. The system repeatedly establishes a locus of attention, there it establishes some properties with respect to the point of foveation, selecting from the range of values of these properties enables the system to shift its focus of attention to some region of space where it may monitor a particular event or feature. At another level, we saw that the system repeatedly sets up successive frames of reference. For example it used the size and orientation of the human to determine the size and locations at which to search for the hand.

3.8.2 Example 2

In this example I describe a simple visual routine for tracking a moving object even as it passes behind an obstacle. Experiments with infants [3] show that $3\frac{1}{2}$ month old infants already have strong expectations for such events, for instance that the object is likely to emerge at the other end of the object, and are surprised when these expectations are violated. One important and unanswered question is: what is the representation of these expectations? One of the contributions of this thesis is the concrete proposal that the

representation of the expectation is in terms of changes of attentional state, where the attentional state and the associated operations are as specified in the previous sections. In other words, changes of attentional-state are sufficient to explain the behavior of the child without invoking more abstract “beliefs” about objects or requiring any “intuitive physics”. Such abstract knowledge could arise later, but the point here is that it is not *necessary* to explain the behavior of the child.

In the following example the visual routine is hand-sequenced. In the next chapter I show how expectations can be learned in an unsupervised manner and thereafter generate visual routines. For now however the point of this example is to show that

- The changes of state of the figure (the ball) and the local spatial context contain sufficient information for constructing the visual routine that checks for the ball re-emerging at the opposite edge.
- The set of operators described in previous sections (in particular the operator for matching local-spatial context described in section 3.5.3) is capable of implementing this visual routine.

In this example the task is to simply track the ball. Figure 3-17(a) shows a schematic of the visual routine that tracks even in the presence of occlusion.

Track Object Figure 3-17(b) and (c) shows a ball and the local context (shown as a large red rectangle) around it as it is tracked. At every instant the figure (the ball), its attributes (direction of motion, size, orientation) and the local context around it are being monitored.

Lose ball at the right edge The system waits until the tracker loses the ball (Figure 3-17(c)(e) and the local context matches the local context shown in (d), i.e. a dark edge to the right of center.

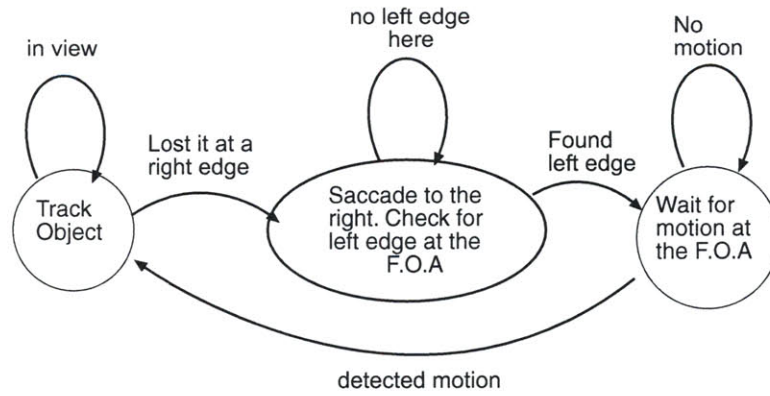
Saccade to the right searching for the opposite edge: Once it loses the ball at an edge, the system saccades to the right, while searching for a local context that matches the one shown in (g).

Wait for motion: Having found the opposite edge the system maintains its focus of attention on this edge while waiting for motion within the current local context (f). When it finally sees motion here (h) it saccades to the moving object, and starts tracking it again (i).

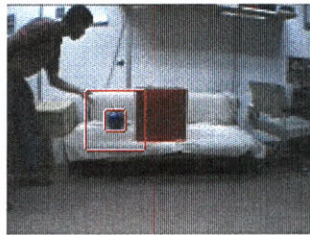
The example emphasizes how the kinds of information attended to in the local-context changes during the course of the event. In frame (b) only the moving figure is being attended to, in (c) (e) (f) (g) the spatial structure of the local context (in particular the dark edge to the right, and the dark edge to the left) is relevant. In (h) only the motion information within the local context is relevant.

3.9 Evaluation of the architecture

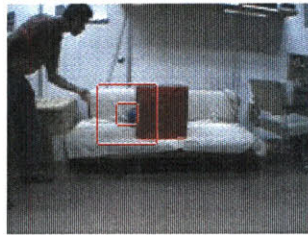
The previous sections showed some examples of visual routines implemented in terms of the proposed language of attention. An empirical way of demonstrating the versatility of the language is to implement hundreds of visual routines for everyday tasks. While extensive



(a) A schematic of transitions in attentional state



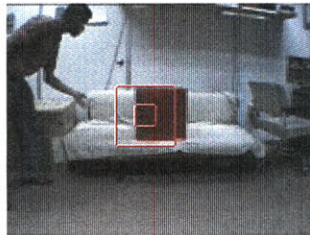
(b)



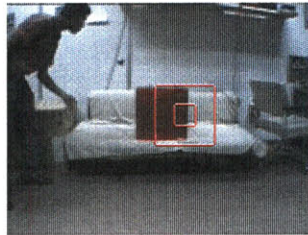
(c)



(d)



(e)



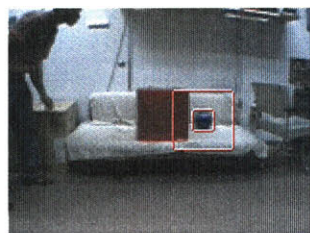
(f)



(g)



(h)



(i)

Figure 3-17: The system uses its expectation for a change of local context to look at the appropriate location for the re-appearance of the ball after having lost it.

testing is certainly a high priority, there is another way of evaluating the language, without trying to enumerate all the visual routines that one can write (which is an impossible task for a compositional grammar). Visual routine primitives and the visual routines that they produce (sequences of primitives) are at different levels of complexity, and therefore can be evaluated independently (even though the routines are constructed from the primitives). In other words, we can make some requirements of what visual routine primitives should do, and similarly make some requirements of what visual routines should do and check to see if these requirements are met.

How robust are the primitives at establishing spatial relationships between regions? There are really two questions here. First, how does one characterize the spatial relations between two regions? and secondly do the primitives I use capture these spatial relations? An exhaustive way of characterizing the relative spatial relationship between two regions (say A and B), is to make explicit the distance and direction distributions of every point in one region (say A) with respect to every point in B. A little reflection will show that any relative spatial relation (e.g. all patches in A which have a patch of B immediately to the right and below) is easily computed from the distance and direction distributions. *I claim that the primitives in my system that capture local region relations and global marker distributions capture the information in these distance and direction distributions.* As mentioned in section 3.5.3 the local region relations capture the information in the distributions when the regions are close together, and the global marker relations between the region centroids capture the information in the (unimodal) distributions when the regions are far apart. Hence the primitives completely characterize the spatial relations between two regions.

How robust is the language at constructing visual routines to capture visuospatial patterns of activity? The language that I have proposed is built around two insights about what visual routines do:

1. Visual Routines establish successive local frames of reference. For example in the pointing example the size and orientation of the human were used to setup a local frame of reference, which was used to select expectations about the location of the hand. Which in turn was used a frame of reference to select a portion of the room.
2. Visual Routines monitor changes in attentional state.

Given that the language I have proposed has been explicitly constructed with these two functions of visual routines in mind, my language is effective in capturing visuospatial patterns of activity insofar as visuospatial patterns of activity indeed perform the two functions listed above.

I now look at the proposed language of attention in the light of existing models of visual attention.

3.10 A review of models of Visual Attention

"Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form of one out of what seem several simultaneous objects or

trains of thought.” - William James (1907)

Hypotheses regarding the role of attention in visual processing have changed little since William James’ 1907 proposal. Attention is viewed primarily as a selection mechanism that allows the visual system to allocate its limited resources to specific regions in an image for further specialized processing. To achieve the goal of reducing or constraining overall input to the cognizer, selective attention may filter out some features while enhancing our perception of other features. The attentional system, in essence, is entrusted with the task of controlling the information flow that incessantly impinges on an observer.

Over the past several decades, a host of researchers have elaborated this general theme. In this section I review some of the empirical data and the most prominent proposals regarding the role of attention. For more extensive descriptions, the reader can refer to any of a number of other reviews (Posner [44]; Posner & Petersen [45] Posner & Rothbart [46]; Tomlin & Villa [55]).

The purpose of this review is to establish that the models of attention that have been developed so far have adopted a rather narrow outlook as to the role of attention in visual processing. This will serve to highlight one of the main points of this thesis, namely, that attention is not merely a stand-alone selection process, but is instead a component of the larger visual routines machinery.

3.10.1 Models of attention as a region/object selection mechanism

LaBerge[26] has been a strong proponent of the selectional role of attention. In his model, attention is merely an enhancer of the attended area, and a suppressor of the surround. The attentional machinery can achieve this contrastive effect in one of three ways:

1. Via a greater enhancement of the attended area than that of the surround, or
2. Via a lesser degradation of the attended region than that of the surround, or
3. Enhancement of the attended region and degradation of the surround. In LaBerge’s model, then, attention is what heightens our sensitivity to specific subsets of our surroundings.

To ground out his model in actual brain structures, LaBerge proposes that attentional processing is carried out by thalamocortical circuits - by the transfer of information to and from the cortex from the thalamus. The thalamus serves to selectively amplify the information coming into the visual system and this amplified information is fed back to the system. The superior colliculus, he believes, plays a role in determining what location to attend to.

A proposal that seeks to refine the general notion of attention as a selection mechanism hypothesizes a role for attention in object selection. The basic idea is that attention facilitates individual object selection by binding multiple features to objects (Treisman and Galade [14]) or by putting the features in the correct “object files” (Kahneman and Treisman, 1984). Treisman’s feature integration theory assumes that the key to the process of combining preattentive features into objects is focused attention. Such a process comes into

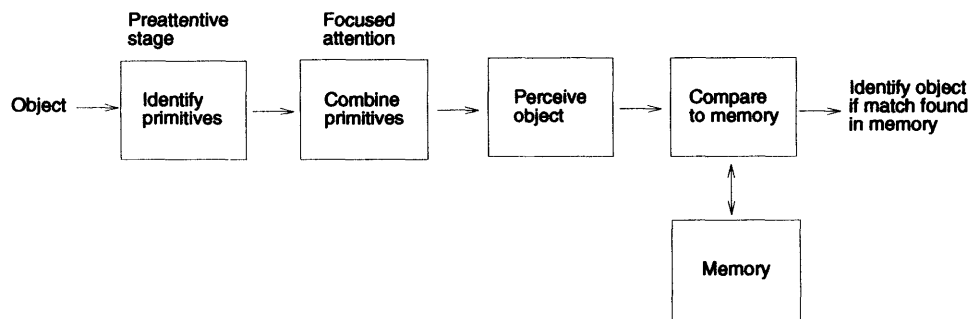


Figure 3-18: A schematic depiction of Treisman's Feature Integration Theory. Treisman views attention as the 'glue' that binds features to objects.

play in complex conjunction search situations, wherein the target object may be defined as a conjunction of multiple attributes. The flow-chart in Figure 3-18 shows how focused attention is involved in the overall process of object perception according to the Feature Integration Theory.

3.10.2 Physiological evidence for a selectional role of attention

Researchers have found attentional modulation/selection of neuronal response in several visual cortical areas. Results from area V4 are particularly striking (Moran and Desimone [35]; Haenny, Maunsell and Schiller [18]; Haenny and Schiller [19]; Spitzer[54]). In some of their early work, Moran and Desimone trained monkeys to attend to a colored bar shape in one location and to ignore a different one in the other location. When both stimuli were in the receptive field of a cell, the cell responded well if the stimulus to which the cell was responsive appeared at the attended location but was attenuated significantly if the stimulus appeared at the unattended location. In general, V4 single-cell data indicates that cells exhibit a gain in firing rate when a single object is presented and attended to. However, when two objects are presented and only one of them is attended, cells show no change in firing rate if the stimulus is a target (preferred) and exhibit attenuation if the stimulus is a distractor (non-preferred).

3.10.3 Psychophysical evidence for a selectional role of attention

Informal psychophysical evidence for a selectional role of attention comes from several everyday experiences. An example may be picking out a conversation at a party to listen to or concentrating on one instrument, say the clarinet, during a musical performance. When attention is directed to the desired voice or the chosen instrument, the key information is more easily extracted from the environment, making it available for further processing—pulling out the content of the conversation or determining the success of the clarinet player, for instance. The facilitation of some information comes at the expense of the inhibition of competing information. So, it is difficult to sustain a current conversation when a nearby one is attended to, and it is hard to keep track of other players when the clarinet is selected for listening.

There is also a host of formal experimental evidence that supports the selectional role of

attention. For instance, Burkell and Pylyshyn (1995) have built upon the work by Yantis and Jonides (1984) (who demonstrated that sudden onsets attract "attention" or "priority tags"). They have found that subjects can separate the subset of search items whose locations are precued by a late-onset marker from the other tokens in a display. Treisman and Sato (1990) have also found evidence for the selection role of attention

3.10.4 Current work in modeling attentional processes

A number of research teams are working towards modeling eye-movements and attentional fixations. Prominent efforts in this regard include (Rao and Ballard, 1995; Rao et al., 1996; Itti and Koch, 1996). The common underlying strategy in all these pieces of work is the computation of saliency maps on some pre-specified image attributes such as color or luminance contrast. Attentional fixations then wander from one image region to another in the order of their computed saliency. Rao et al., 1996, for instance, build iconic scene representations using oriented spatial filters at multiple scales. The attentional 'itenary' is constructed in a coarse-to-fine fashion with higher salience being attached to larger scale filter responses. Itti and Koch follow a similar strategy. They combine multi-scale image features into a topographic saliency map. The relative weights of the image features can be learned.

3.10.5 What's missing in all these models?

It is important to note that all these efforts are concerned only with the control strategy for determining the loci of attention over time, but say nothing about the processing that happens *at the attended locations*. Going back to the Yarbus example in Figure 2-2(c) simply selecting various parts of the scene does not accomplish the task at hand of estimating the ages of the people, there is some crucial task-specific computation at each successive focus of attention. The "attention = selection" view stems directly from viewing attention as an independent selectional mechanism whose main function is to facilitate object-recognition. In this thesis, I shall argue that this view-point needs to be expanded and that attention needs to be viewed as an integral component of the visual routines machinery, and consequently any realistic model of attention must have a proposal for the vocabulary of operations that can be performed at the focus of attention. In this more comprehensive view, selection is just one facet of what attention does; a more important role it plays is in the establishment/extraction of inter or intra region spatial properties. These properties are critical for the performance of visual routines.

3.11 Summary

In this chapter I presented the architecture of a system for constructing visual routines, and showed some examples of visual routines implemented in terms of this language. The examples highlighted two ways of looking at what visual routines do: namely, establishing successive changes of frames of reference, and monitoring changes in attentional state. The proposed language of attention represents a significant step beyond prior models of Attention. Whereas prior models viewed the role of Attention as just "Selection", I emphasized

two other families of operations which are also crucial to a realistic model of Attention. I also made some specific proposals about the operations that constitute each family of operations. While the specific operations may change in the future as we better understand visual routines, we claim that the families of operations and the types of the operations will endure.

Chapter 4

Learning Visual Routines

In this chapter I present a scheme for learning visual routines from experience. Whereas in Chapter 3 the focus was on the primitive operations of visual routines and how they can be strung together to extract different spatial relations, the focus here is on how visual routines automatically come about from experience. In this chapter I will provide an answer to the following related questions:

1. If visual routines for extracting spatial relations are not explicitly sequenced or composed how do they come about?
2. Event expectations: By the age of six months infants already have certain expectations about events in the world. Figure 4-2 shows an example of an event that surprises $4\frac{1}{2}$ -month-old infants. How do these expectations arise?
3. Event recognition: We use terms like “pick-up” or “collide” to describe certain events even though the details of the individual instances may differ widely in their details (Figure 4-1 shows two different instances of “pickup”). What is invariant in different instances of pick-up or collide that leads us to describe them the same way? And how do we learn these invariants?

The answers to these questions will hinge on the following ideas: Visual routines are learned - not planned. Event expectations can be formed by learning frequently occurring patterns in attentional state. Extracting invariances requires proactive exploration.

4.1 Visual routines are learned - not planned

In section 2.1.2 we saw that there were two important issues that any “basis-set” theory of visuospatial problem solving must deal with: The choice of a basis set of primitive operations, and their composition. The previous chapter dealt with the first issue, it described a

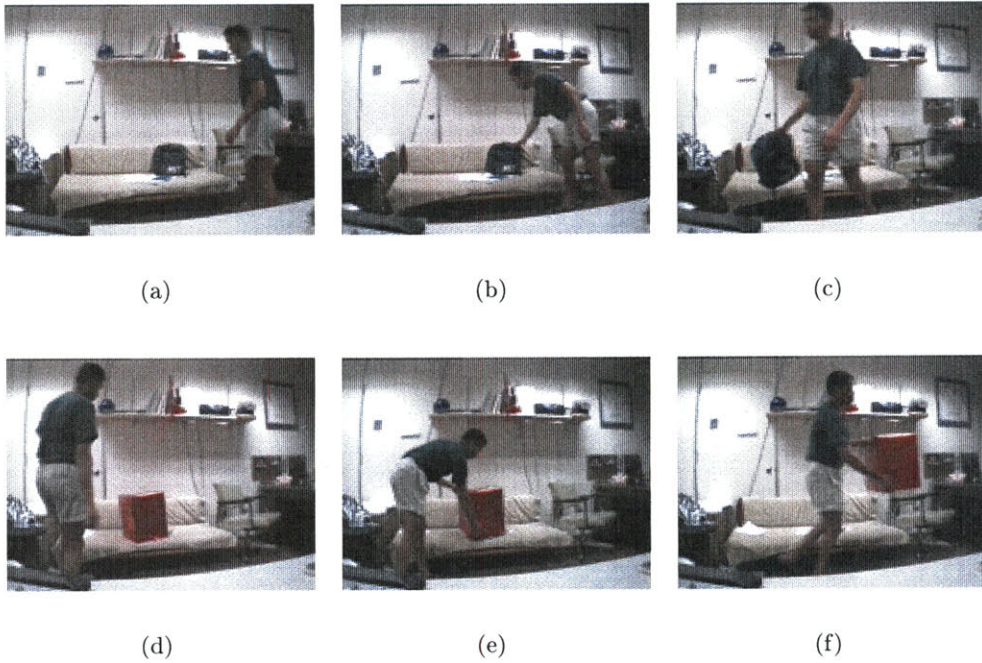


Figure 4-1: Two patterns of activity (a-c) and (d-f), that are similar in some way, what is common? and how can we learn it?

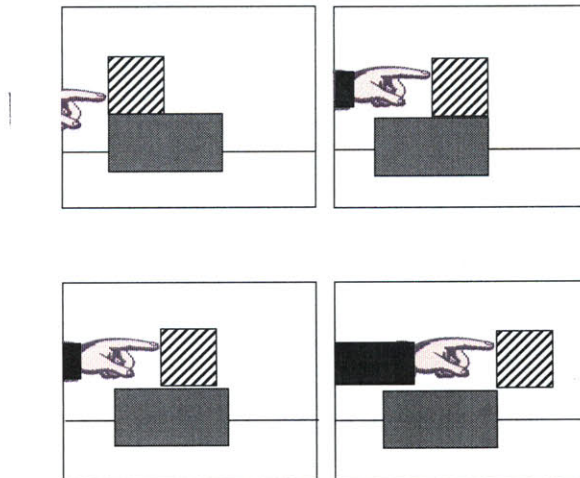


Figure 4-2: $4\frac{1}{2}$ -month-old infants preferentially look longer at the “impossible event” in the bottom row, presumably because some expectation was violated. What is the representation of this expectation? and how was it acquired?.

[3]

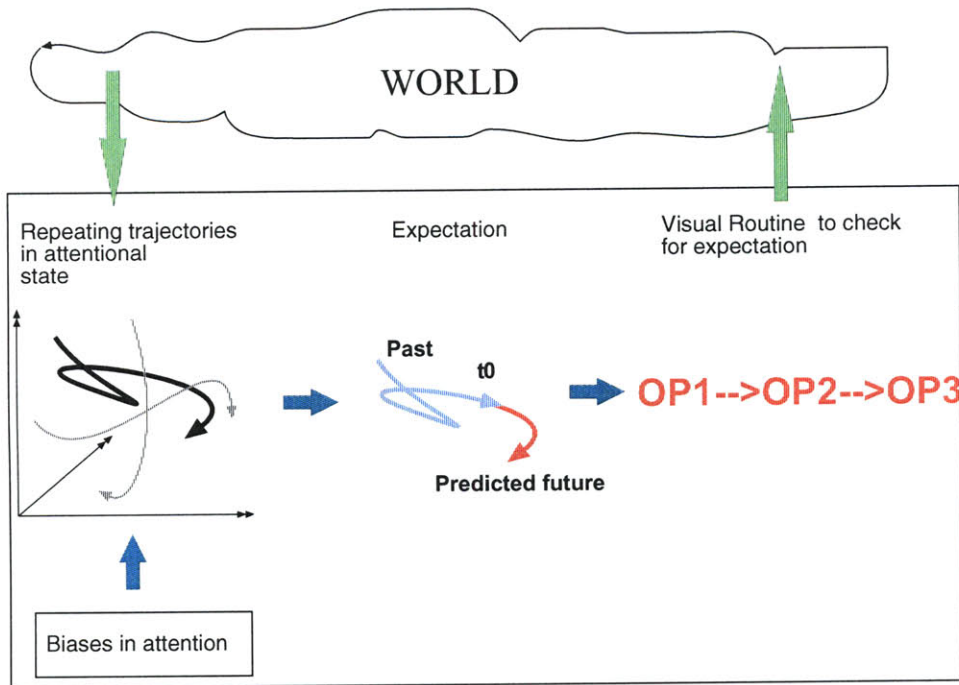


Figure 4-3: Regularities in the world and attentional biases repeatedly drive the attentional state through patterns. High frequency patterns become expectations. Expectations generate visual routines to check for the predicted part of the pattern.

small family of operations (a kind of “language of visual attention”) and showed examples of how visual routines could be constructed by repeatedly choosing operations from the family of operations. The visual routines in the examples however were composed by hand, in this chapter we deal with the remaining important issue, namely how visual routines can be automatically composed. How can a system know how to put together a visual routine to determine what a person is *pointing* at or if one object is *touching* another object? This statement of the problem implies an “explicit sequencer”, a box that when given a new visuospatial problem, somehow “knows” how to sequence the primitive operations to solve that problem. I take the position that there is no such box, no explicit sequencer for constructing visual routines.

An overview of my approach

The following points list the main features of my approach, a schematic of which is shown in Figure 4-3.

1. Patterns of visual activity emerge from interaction with the environment (as we saw in section 2.4 in “Amazon” [10]). 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 3.7.
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 as shown in Table 3.2 every 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.

In the following sections I first explain the motivation and issues surrounding each of these points before delving into the specific instantiation in my system.

I first clarify what I mean by the first point.

4.2 Situated patterns of activity, and interactive emergence

In his book *Catching Ourselves in the Act* Horst Hendriks-Jansen [20] examines the role of natural selection in the explanation of human behavior and the nature of explanatory models of human intelligence. His main message is that *patterns of emergent activity are the only basis for an explanation of natural intelligence*. A brief discussion of his work will shed light on the idea of “patterns of activity” which as we have seen is a useful way of thinking about visual routines.

Jansen does a thorough job of analyzing the implicit assumptions that underlie various computational theories/explanations/recipes for human-like intelligence including the symbol-systems approach, connectionism, and behavior-based A.I. While this is not the place to detail Jansen’s numerous well-crafted arguments, there is one point that recurs in his criticism of existing explanations of intelligence:

Deliberately designed computational models do not have explanatory power because they they do not reveal anything more about the-thing-being-modeled beyond what was put in. He says (of the deliberately designed models):

They require clearly defined parameters that delineate the similarities between the model and the *explicandum* right from the start. The model does not help to define the parameters, and it has no power to reveal previously unsuspected causal and structural relations. These need to have been defined before the model is ever built, since it is built in terms of those parameters in an attempt to prove that certain formally specified relations exist in the *explicandum* just as they do in the model.

Jansen concludes that the “natural” kinds for explaining intelligence must therefore come from some other place than the model itself. He suggests that the *patterns of activity* that a system exhibits when it interacts with the environment are a good candidate for these

“natural kinds”. An example of such a pattern of activity is the wall-following behavior that a robot exhibits when its low level reflexes of Avoiding obstacles, Aligning to some direction, Strolling, and Correcting (see [31] for the details) interact with an environment that happens to have walls. the important point that Jansen makes is that the “wall-following” activity is not a predetermined series of movements, it emerges from both, the presence of the wall and the low-level reflexes, and *has a structure that lies above the reflexes that produce it*. Furthermore, the notion of a wall does not have an internal representation inside the robot, and no formal description of the wall is necessary to produce the activity.

What do visual routines have to do with all this?

A lot. The concept of frequently occurring traces in attentional state is identical to Jansen’s “patterns of activity” (in fact I use the same phrase). I feel that I have built upon Jansen’s notion of patterns of activity in one important way. Jansen would like find out what the natural kinds are for humans and how these “natural kinds” arose via natural selection. To him that would constitute a valid *explanation* of intelligence. My goal is to use the natural kinds to *construct* intelligence. I would like to suggest that *one could build a system that can notice and grasp the patterns of activity that it goes through, and use them as a vocabulary or scaffolding for higher-level cognition*.

4.3 Why learn emergent patterns?

We now move on to the second point mentioned in the outline of section 4.1, namely that the emergent patterns of visual activity are learned. A reasonable question here is “If visual routines emerge from interactions with the environment, why bother learning them?” After all, won’t they just happen automatically when needed? It is important to understand why this argument is flawed.

4.3.1 Visuospatial patterns lead to expectations that impose top-down bias on behavior

Suppose someone is trying to get a child to look at some object in the environment. In some situations, pointing at it works because the object is very salient to begin with, and the child’s attention shifts from your hand (or gaze direction) to the object purely due to bottom-up biases (i.e. without using the direction of the hand as a cue). However, there will be other situations where the object is not very salient, and having learned the relation between hand orientation and object position from prior situations makes all the difference in being able to locate the object. In fact these are the situations where the pointing gesture really has function. Putting it another way, learning the correlation when the signal is “strong” helps you in situations when there is noise and the signal is weak, because now you have a model. Another example is the visual routine that you go through of looking both ways before crossing the road. Sometimes you may be driven through this routine by a bottom-up behavior of looking at a looming stimulus in the periphery. Noticing that this pattern happens a lot on roads, remembering the pattern, making it an expectation, and pro-actively looking left and right before you cross a road, has clear survival value. The

bottom-up behavior is still in effect, but it may be too late by the time it is triggered, or it may not be triggered at all if portions of the road are not in your peripheral vision (e.g. you're at an intersection).

The reason for learning emergent patterns discussed above is essentially arguing for the benefit of having expectations, in that *expectations about your interaction with the world constitute a kind of model of the world, which can drive behavior in a top-down manner to pro-actively extract signal from noise*. There is another different but equally important reason for learning emergent patterns in experience: *patterns in visual experience form the basis for "higher-level" concepts that can be used for communication between agents that share the same set of patterns by virtue of having the same or very similar visual routine architecture*.

4.3.2 Visuospatial patterns form the basis of language, and reasoning

Consider the following significant aspects of human cognitive development:

1. Children learn regularities in the environment well before they can speak, [52] [3].
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.

In the light of such facts, I would like to suggest that:

1. A vast number of visuospatial patterns are extracted at a very early stage in child development in an unsupervised manner.
2. Subsequent supervised learning only requires a few examples because the child almost has the concepts, and the supervised learning is simply helping label a pattern that has already been learned.
3. The visuospatial patterns are required for language development, and must necessarily be acquired before spatial words can be learned.
4. The visuospatial patterns are used for higher cognitive functions(like reasoning about quantity).

In the next few sections I describe the details of the instantiation of the approach described in 4.1 in my system, and show the visual routines generated as a result of exploration and learning.

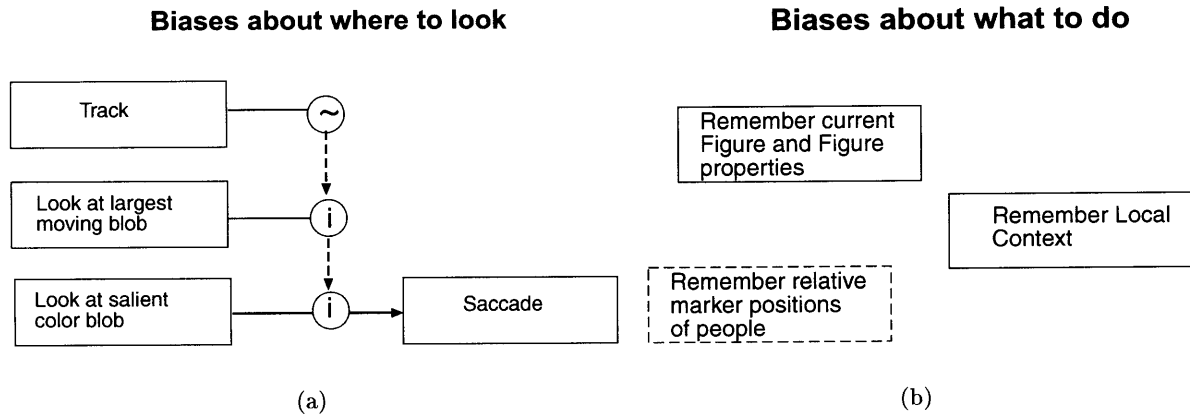


Figure 4-4: Biases about where to look and what to do at the focus of attention are necessary for exploration.

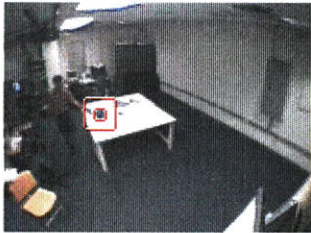
4.4 Attentional traces during exploration

In the examples of the previous chapter the hand-coded visual routine, together with the scene, decided where the system looked, and what properties got monitored across foveations. 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.

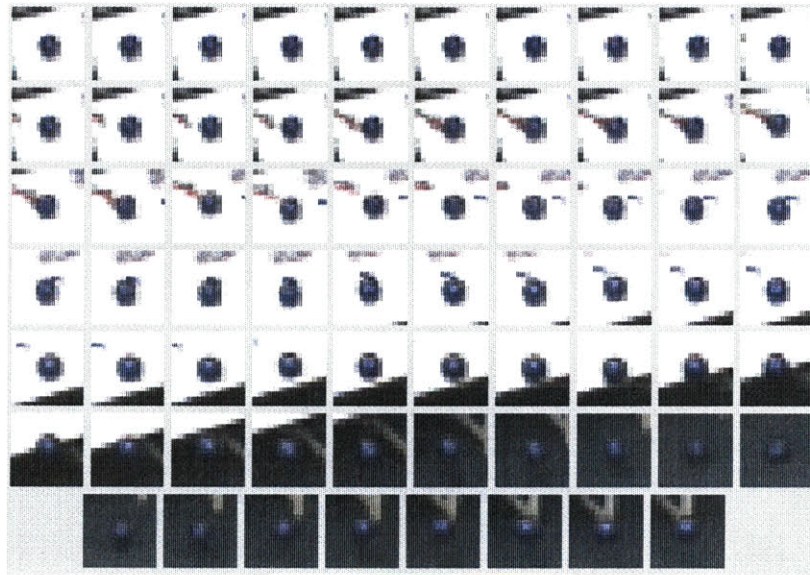
A system of simple behaviors shown in Figure 4-4(a) implements bottom-up biases about where to look. One behavior tracks moving blobs, another one looks at the largest moving blob. A third behavior looks at color-salient objects if nothing is moving. At present the behaviors have been individually tested but the subsumption structure (indicated by the dashed lines in the figure) has not.

Apart from *where to look*, there must also be biases about *what to do* at the focus of attention, and what state to maintain across foveations, during free exploration. Throughout the thesis I have emphasized that any realistic model of attention must take into account what happens at the focus of attention. A large part of the previous chapter was devoted to a description of the properties and operations at the focus of attention. These operations could be used to establish different kinds of spatial properties at the FOA depending on the task at hand. However, what does one do during exploration? There are a couple of possibilities:

- One possibility is to always extract certain properties. For instance the figure attributes, and the local spatial context around the figure may be important properties to extract and monitor across foveations regardless of the situation.
- Another possibility (not mutually exclusive) is to have other representational systems and motivations *suggest* tasks. What does this mean? So far I have only talked about the visuospatial machinery for a humanoid robot, but clearly the visuospatial machinery doesn't exist in isolation, but is a vehicle for a variety of goals and motivations that need to check for certain situations in the environment. To take just one ex-



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



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

Figure 4-5: The attentional trace for an event generated by the bottom-up exploratory behaviors of Figure 4-4. Only the tracking behavior was active

ample; A drive to look for people, and check to see who is next to whom may be a high priority to parts of the brain concerned with social interaction, so there may be a strong bias to put markers on people (as opposed to other objects) and monitor the relative spatial positions of these markers during exploration of a scene.

Given that I have been considering the visuospatial machinery in isolation, and because I want to start with a simple scheme, I've adopted the first option, of always monitoring a fixed set of features (listed below). Furthermore in all of the examples in this chapter I have used a blue ball as the object of interest, so as not to mix in the issue of object recognition and obscure what I am trying to show, namely that it is possible to learn visuospatial patterns of activity using the local and spatial representations described in the previous chapter.

Summarizing the conditions under which the learning experiments were conducted: A blue ball was used, and this fact was used to aid in the tracking. During tracking the properties monitored (the attentional state) was restricted to the following components:

1. The direction of motion of the figure.
2. The presence or absence of the figure (i.e. is the size of the figure non-zero)
3. The immediate spatial context around the figure.

In other words I will not for the moment consider the other components of attentional state described in 3.7.

Figure 4-5(a) shows some snapshots of the ball being automatically tracked as it is pushed over the edge of the table. The local context around the ball (shown as a large red square) is scaled to a fixed size of 40×40 . Figure 4-5(b) shows the full trace of the local context, from left to right in row major order, the ball is pushed, it rolls to the edge of the white table, falls over and bounces. Note that the direction of motion of the figure, and presence or absence of the figure is also part of the state but has not been shown in Figure 4-5.

4.5 Activating similar attentional traces in memory

Multiple instances of events involving the ball that may be described as “falling”, “colliding”, “bouncing”, “passing-behind”, and “picking-up” were recorded. Attentional traces for each of these events were stored as in memory, forming different trajectories in the attentional state space. Trajectories of “similar” events should be close together in the state space.

As it is hard to depict the entire state space and highlight the trajectories that cluster together (as shown in the schematic of Figure 4-3), I will adopt another way of showing the same information. I will match a segment of an attentional trace to *all* segments of attentional traces of *all* examples, and display the best matches. If the choice of the attentional state space has been a judicious one, then we should expect to see segments of attentional traces of “similar” events.

For now I use a very simple-minded method for matching trajectories; I match trajectories based solely on the absolute values of the figure attributes, (i.e. the direction of motion of

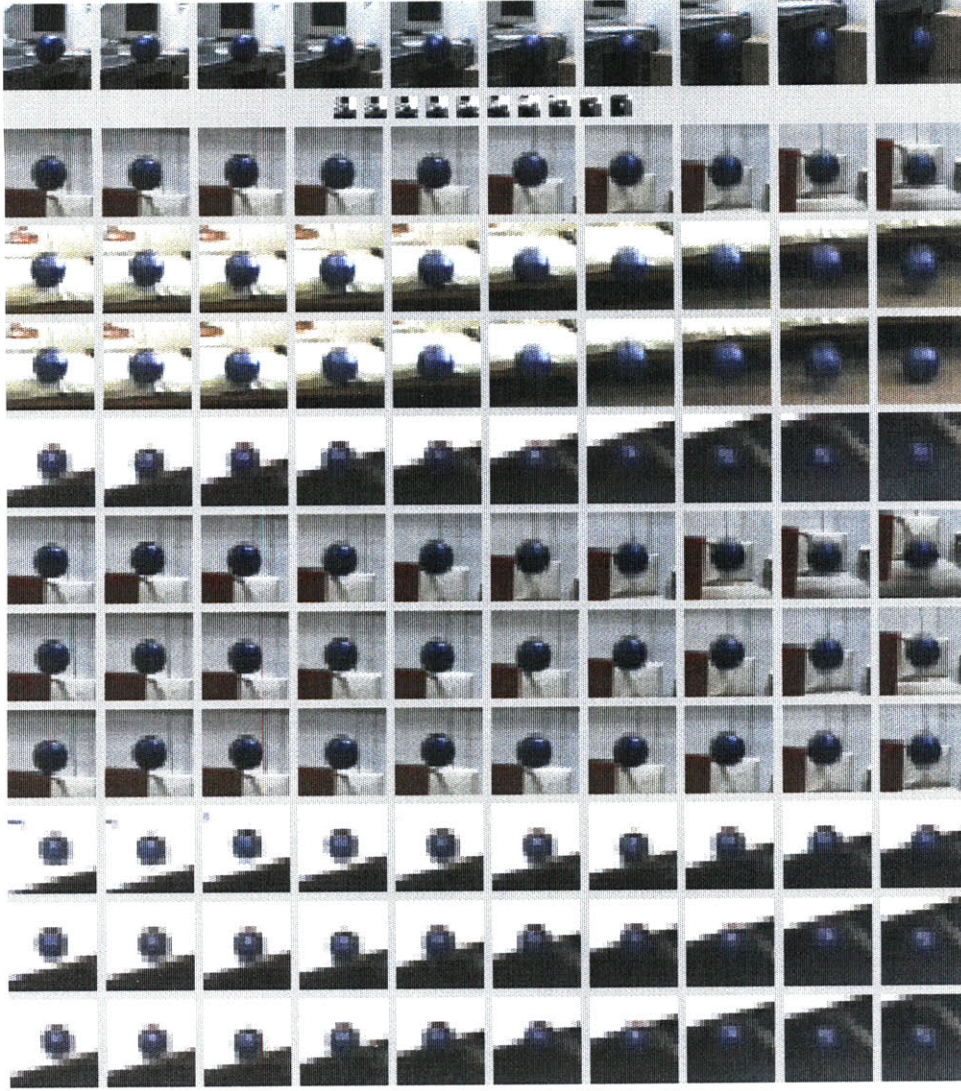


Figure 4-6: The top row is the query segment of a “falling” ball, the next 10 rows show the 10 closest trajectories to the query trajectory in attentional state space.

the ball, and whether it was being tracked or not). This can be thought of as moving a little sphere along a trajectory in state space, and considering only trajectories that fall within the volume swept out in state space. It would be more reasonable to match *differences* in attentional state, which would be much more robust. However, this is only a first pass at demonstrating the approach described in 4.1 and should not be taken as representative of the full capabilities of the system.

Figure 4-6, shows the 10 closest trajectories in the attentional state space to segments of a ball “falling to the right”. The topmost row shows the new or query sequence, the next 10 rows show the closest trajectories. Time goes from left to right. Some of the matching trajectories are almost identical to other ones except shifted in time.

Figure 4-7, and Figure 4-8 show the 10 retrieved best trajectories for queries of a ball “bouncing”, and “passing behind” an occluder.

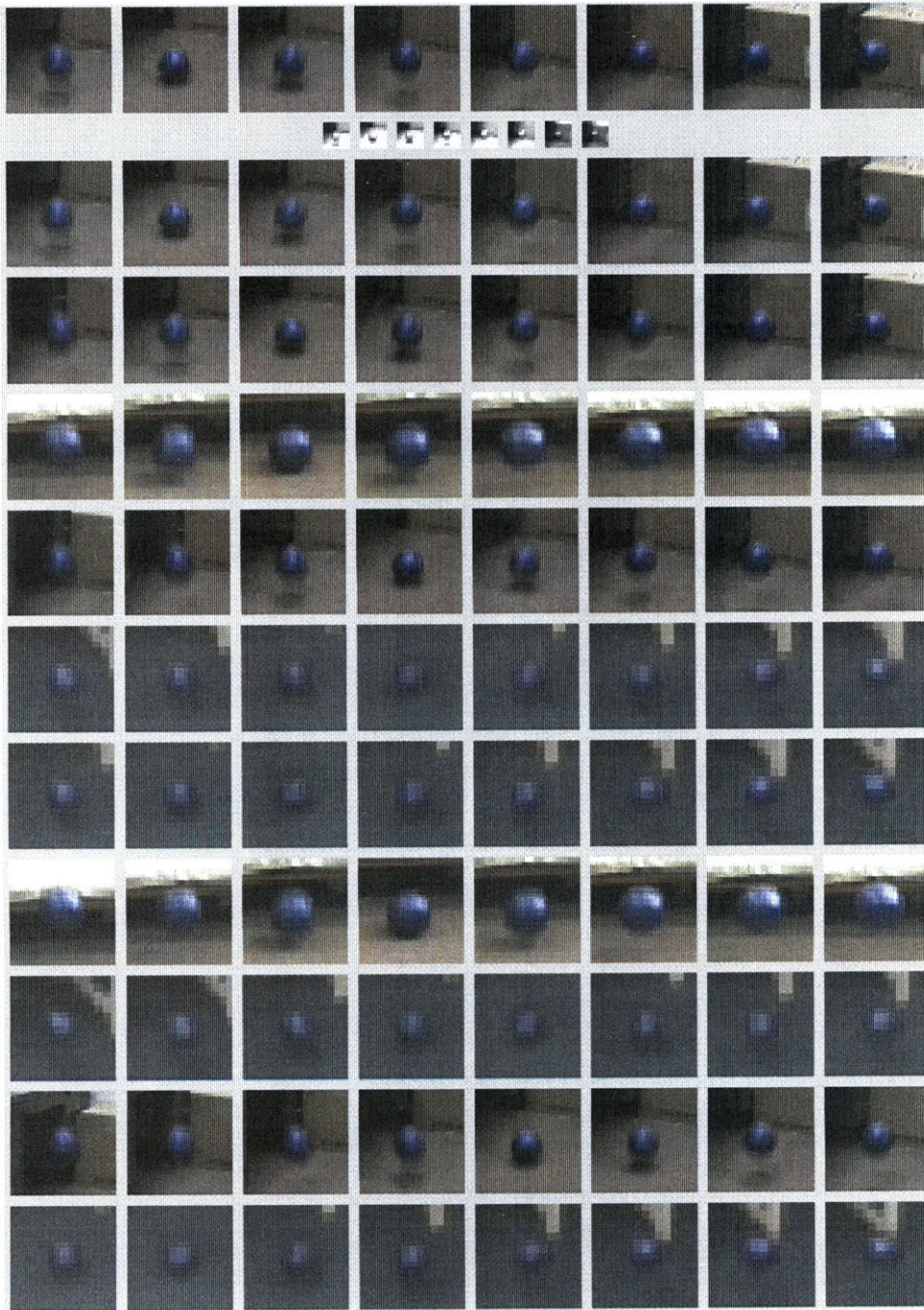


Figure 4-7: The top row is the query segment of a “bouncing” ball, the next 10 rows show the 10 closest trajectories to the query trajectory in attentional state space.

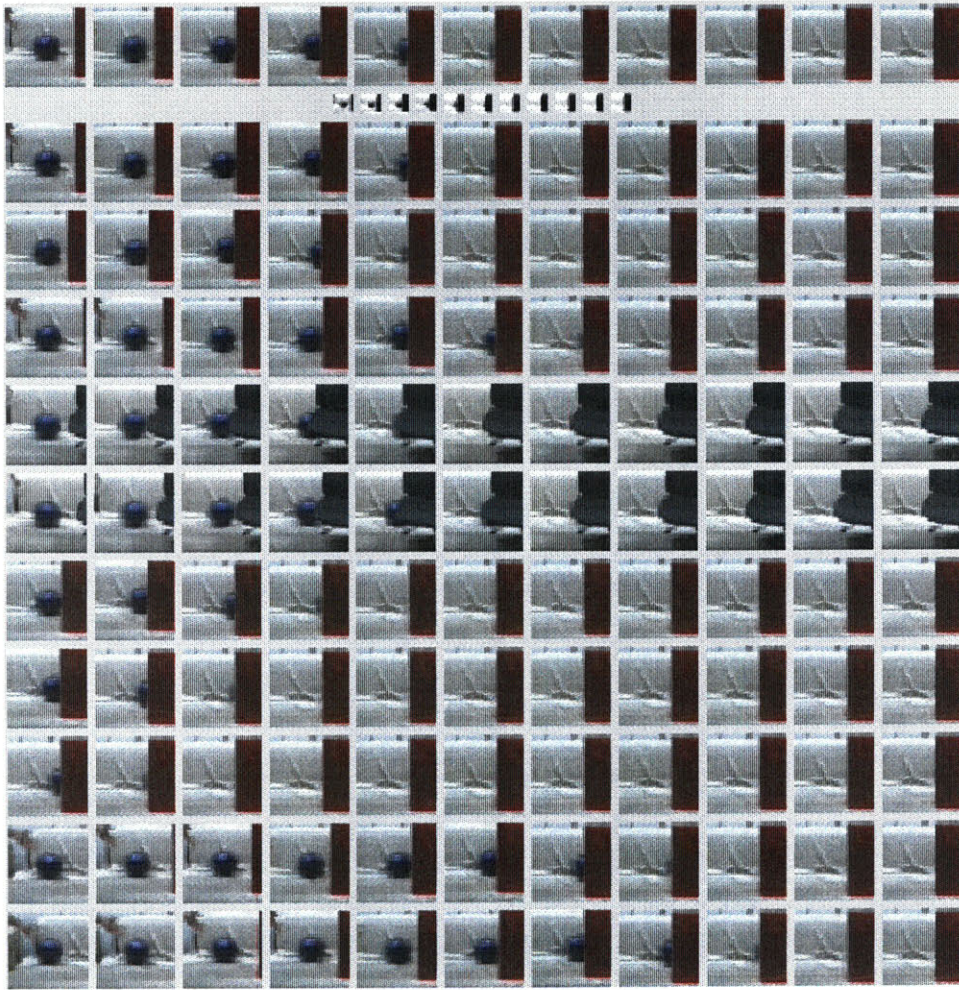
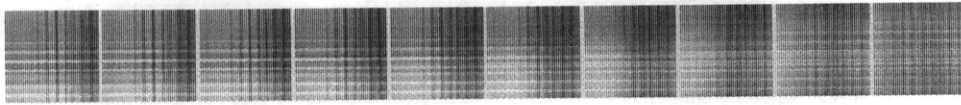
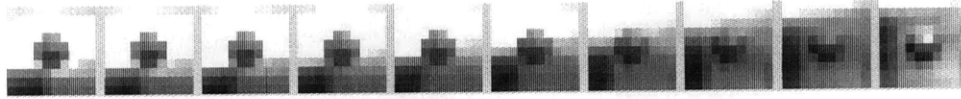


Figure 4-8: The top row is the query segment of a ball “passing behind” an occluder, the next 10 rows show the 10 closest trajectories to the query trajectory in attentional state space. There was really only one other example in experience that was similar.



(a) The averaged ratio templates of activated traces of Figure 4-6



(b) A visualization of the averaged ratio templates

Figure 4-9: The top row shows the learned local spatial context for an object falling to the right. The bottom row shows a visualization of the template. Time goes from left to right in these sequences.

4.6 Learning an expectation from activated attentional traces

Having activated/retrieved “similar” attentional traces, the recurring pattern is learned by simply averaging the local spatial contexts of the activated traces. For example, in Figure 4-6 the local context ratio-templates for the 10 traces are averaged to get an “expectation or template for falling to the right” as shown in Figure 4-9(a).

4.7 Learning visual routines via reinforcement learning

In this section I review an alternative approach for learning visual routines and compare it to the approach described in this chapter.

4.7.1 An overview of the U-tree reinforcement learning algorithm

McCallum [32] describes a *U-Tree* reinforcement algorithm for learning visual routines. Visual routines are viewed as a sequence of perceptual actions to redirect the purposefully narrow attention window to relevant parts of the environment. The agent’s task at every step is to choose where to look, where to focus covert attention, and which perceptual computations to perform on the available sensory data. McCallum concludes that learning a strategy for perceptual actions is an integral part of learning to perform the task. He then points out that Reinforcement Learning is well suited for learning of attentional shifts because Reinforcement Learning’s chief strength lies in its ability to choose sequences of actions based on the current state and some policy or reward function. The problem of hidden state (due to a limited attentional window) is addressed by having some short-term memory for states and actions in the immediate past.

McCallum presents a new reinforcement learning algorithm called *U-Tree*. The input to the algorithm at each time step is a transition triplet of action-just-taken, current-state, reward. From this time series of triplets the algorithm constructs a *decision-tree* which takes a new

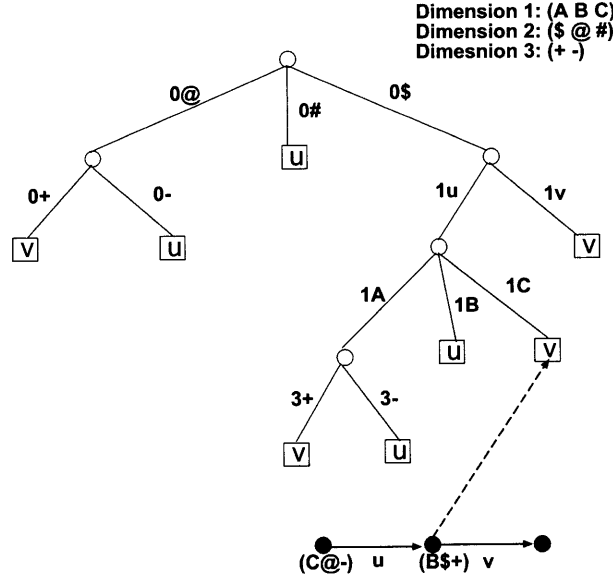


Figure 4-10: A schematic of a U-tree decision tree for organizing attentional state space. From McCallum[32]

triplet and classifies it in a bin. Figure 4-10 shows a schematic of a U-tree along with a sequence of instances at the bottom. In this U-tree there are two possible actions, \boxed{u} , and \boxed{v} , which appear at the leaves of the tree along with an estimated utility Q-value (not shown in the figure). There are three input dimensions which take on values (A, B, C), (, #, \$), and (+, -) respectively. The branches of the U-tree indicate a classification decision based on perceptual state in the recent past (e.g. 1C asks if the state C was observed 1 time-step ago). Given some new observation like (B \$ -) which was reached after taking the action \boxed{u} , the observation is classified by following the branches of the tree from the root: 0\$ because we are seeing \$ now, then 1u because the last action taken was \boxed{u} , then 1C because C was observed in the previous instance. This leads to a leaf which suggests that the next action taken be \boxed{v} along with the expected utility of this action.

The U-tree is constructed by starting with one node, and then dividing up the instances classified by that node into separate nodes under it. The partitioning is done based on testing the instances for statistically significant differences based on expected future discounted reward. The statistical test applied to check if the instances came from different distributions is the Kolmogorov-Smirnov test.

McCallum tests the system on a simulation of highway driving. Seven perceptual dimensions (e.g. hear-horn, gaze-object, gaze-side, gaze-speed, e.t.c.) each with a discrete set of values are available to the system. The system also receives one of three rewards at each time step: -10.0 for scraping past a slow truck, -1.0 when a fast truck honks at the agent, and 0.1 for making progress. Given a seven dimensional input and a reward, the agent must choose one of five actions: gaze-forward-left, gaze-forward-center, gaze-forward-right, gaze-backward, and shift-to-gaze-lane. The result of using the U-tree approach is that the system creates only 143 internal tree nodes, out of a possible state space of 2500^3 .

4.7.2 A critique of Reinforcement Learning

The criticism that follows is not directed particularly at the U-tree method but at the reinforcement learning framework in general. There is clearly a difference in paradigm between the reinforcement learning approach described above and my approach to learning as described in section 4.1. The following are some of the major differences:

Representation is key, abstracting away from it is bad. In any learning approach there are three questions of paramount importance, *what is being learned?* *In what representation is it being learned?* and *How is it being learned?* i.e. what is the learning method being used? Once we have decided what it is that we want to learn, picking a representation is key, because picking a good representation that makes important things explicit greatly simplifies the learning method used. This is the route that I have chosen. In chapter 3 I have spent some effort in choosing a simple but rich set of properties and operations, so that when it comes to picking a learning method I can use a relatively simple memory-based method of noticing high-frequency patterns. Reinforcement Learning on the other hand abstracts away from the representation and places its emphasis on a complicated learning method that requires a reward at every step. Delayed reward methods exist, but are even more complicated because they have to deal with severe credit assignment issues. Notice that in the highway driving problem above, once the problem was formulated in terms of 7 perpetual dimensions, 3 rewards, and 5 actions, as far as the learning method was concerned it could well have been about trading options in the stock market. The learning method doesn't care where the problem came from, in fact this is viewed as a "strength" of Reinforcement Learning, and makes it seductive because it can be used in any domain. However, this "generality" comes at a very costly price. The learning method now has to do significantly more work and introduce artificial constructs like "reward" to unreach the regularity from the data. There is a strong parallel here between the use of Reinforcement Learning like methods and the promotion of "Search in some space of symbols" as a "general" problem solving method in the early days of A.I. because both approaches abstract away from the *representation*, the choice of which is the crucial part of the problem.

Notice structure in the world, don't look for a reward at every step. My approach depends crucially on the fact that there is *structure* in the world which interacts with the biases of the system to produce repeating patterns in attentional state. It is artificial to posit rewards for every time step of every event, firstly because it begs the question of who is going to come up with a reward structure for every task, but more importantly because in many situations like observing a ball falling, rewards just don't make sense. *Events just happen. Being able to notice recurring ones, and predict their occurrence is reward in itself.* I am not claiming that rewards have no place at all in learning. In supervised learning, a gesture of appreciation to a child, or a tidbit of food to a dog, could help reinforce some behavior. However in most situations adopting a "reinforcement with reward" framework is not necessary to learn.

Low-level exploratory behaviors incorporate implicit biases, making explicit rewards unnecessary. Low-level exploratory behaviors like "tracking", or "look-for-human", implicitly encode biases about what to do next; the tracking behavior is "rewarded" by continued tracking, and the "look-for-human" by finding a human to look at every time step. Therefore they already incorporate very specific notions of reward particular to their goals and do not need explicit and arbitrary measures of reward. Reinforcement Learning

on the other hand by abstracting away from the problem, has no biases about what to do and has the full panoply of action choices available at every time step. This lack of bias will be claimed as a virtue because it makes the system more “general”, however as biology shows again and again, presence of biases in low-level behaviors speeds up learning and leads to the interactive emergence of patterns of activity [20].

4.8 The role of visual routines in cognitive development

In this section I will review some past and recent work on child development, and examine the exciting implications of visual routines on cognitive development.

4.8.1 A brief overview of Piaget’s theory

Any discussion of child development must begin with a summary of Jean Piaget’s monumental contributions. The following are the four stages of development as hypothesized by Piaget:

1. Sensorimotor Stage[0-2yrs]: In this stage sensorimotor schemas of increasing complexity are learned. The world is explored through some very basic reflexes like grasping, or sucking. 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 child repeats actions until he/she can reproduce them reliably. Exploration becomes more goal-directed as the child tries to do things to repeat an interesting stimulus. Finally the child moves from sensorimotor to representational intelligence as action sequences are simulated in the head rather than by active manipulation (there is evidence that such simulations are possible in certain motor and visual areas [15] of the brain). The defining feature of this stage is that the child’s concept of the world is completely in terms of its actions upon it.
2. Pre-operational stage [2-7yrs]: The child represent objects and events even in their absence. Deferred imitation, symbolic play, drawing, and spoken language mature during this stage. However, Preoperational thought is still strongly tied to perception. It is egocentric and cannot grasp conservation laws.
3. Concrete operational stage [7-11yrs]: Logical thought is applied to concrete problems. Thought finally breaks through from perceptual limitations that characterized the previous two stages, and can now manipulate classes and relations leading to what Piaget calls “mobility” of thinking.
4. Formal operational stage [11-15yrs]: Piaget believes that logic is the mirror of thought, rather than the other way around, i.e. the function of formal logic is to make explicit the mental operations that occur at the highest stage of human cognitive development [24]. At this stage the adolescent can see abstract structure in discourse, metaphors, and analogies of all kinds.

What aspects of Piaget’s [41] theory are relevant to our efforts to build a humanoid-robot? 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, whereas logical-mathematical has to be invented.
- **Bootstrapping and Subsumption:** The stages do not em 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.

Piaget’s main contribution is in the cataloging of the changes of behavior due to cognitive development. Two big issues that still remain unanswered and are particularly relevant to someone trying to *build* human-like intelligence are:

1. **What** develops? What exactly are the cognitive structures that undergo qualitative changes?
2. **How?** What are the mechanisms that bring about the changes?

Piaget’s suggests *schemas* as cognitive structures, and *assimilation*, and *accommodation* as mechanisms of specialization and generalization of schemas. Piaget’s schemas (as interpreted by Drescher [11]) are essentially state action state triplets, that denote a state transition caused by some action. While Piaget’s suggestion of a schemas and assimilation and accommodation mechanisms is a useful beginning, much remains to be done in finding specific instantiations in terms of brain structures.

I will return to the two issues mentioned above after reviewing some work on cognitive development that is more recent. One of the most actively researched and debated issues in cognitive science relates to the somewhat vague notion of *object concepts*. Specifically, what kinds of object concepts exist in the brain and how they develop from infancy through adulthood. Many studies in developmental psychology have attempted to probe these questions. I review some of the key ones with a view to pointing out a major deficiency that they all share - a lack of grounding of the cognitive notions into well-defined perceptual primitives and operations. In fact, I will suggest that the perceptual grounding may, at least in some cases, obviate the need to posit the existence of ill-defined high-level object concepts in the developing or adult brain.

4.8.2 Perceptual grounding of object concepts

Spelke and her colleagues have been very active in trying to understand the course of development of object concepts and physical knowledge [52, 53]. The major themes they have identified in their empirical results are:

- The cognitive capacities for object perception in adults trace their roots back to early infancy. Infants seem capable of perceiving, reasoning about and acting upon objects in a relatively sophisticated manner.

- The capacities to perceive and reason about objects are closely related to one another.
- Infant's perception of objects is based upon fundamental physical constraints set down in the environment, such as empirical cohesion within boundaries, continuity of motion paths and inter-object contacts. Since it intuitively seems that the principles of cohesion, continuity and contact are critical to adult's conceptions of objects, a corollary of Spelke et al's suggestion is that high-level object concepts come into being relatively soon after birth in human infants (point 1, above).

In short, the developing brain is believed to embody knowledge about physical principles in the environment. These principles guide object perception and reasoning. Furthermore, since these principles are rather complex, (for instance, the relations of connectedness are believed to operate on the underlying three-dimensional structure of the world, rather than on two-dimensional image information), the object concepts they support are quite sophisticated. Accordingly, the object perception processes are believed to occur late in the visual system.

Another prominent group that has sought to investigate physical reasoning and object perception in infancy comprises Baillargeon and her colleagues. Baillargeon took as her starting point Piaget's [1952, 1954] suggestion that the notion of object permanence during occlusion does not develop until nearly the "advanced" age of nine months. In his experiments with infants younger than 9 months, Piaget repeatedly observed that children did not search for objects that they had observed being hidden. Apparently, Piaget reasoned, the children incorrectly assume that the objects cease to exist when concealed by other objects. However, Baillargeon and others realized that infants might perform poorly on search tasks simply because of difficulties inherent in planning a means-end search sequence rather than due to incorrect beliefs about occlusion events. Using the violation-of-expectation experimental paradigm (which exploits infant's tendency to look longer at novel than at familiar stimuli [Banks, 1983; Olson and Sherman, 1983; Spelke, 1985]), Baillargeon demonstrated in a series of studies that contrary to traditional claims, even very young infants "appreciate" that objects continue to exist when occluded. For instance, the infants expressed surprise when a tall object seemingly disappeared while traveling behind a squat occluder but not when the object was shorter than the occluder's height. The lack of surprise in the latter case led Baillargeon to conclude that infants *believe* that objects continue to exist when masked by other sufficiently large objects. Baillargeon's other studies have sought to probe infants knowledge of contact relations as determinants of object stability and their ability to reason about collision phenomena. In the former, the investigators presented evidence to support the conclusion that infants acquire the abstract concept of *support* rapidly and then progressively refine it over time (to be able to determine how much support is required for stability). The latter set of experiments effectively tested for the presence of the notion of conservation of momentum in young infants. Children were surprised upon seeing an object that had just suffered a collision with a moving object stay stationary. The conclusion reached here was that children rapidly acquired the basic notion of conservation of momentum and then refined it (how far should an object move after collision) gradually.

The empirical studies have made significant contributions to our understanding of the perceptual capabilities of infants. However, important open questions remain regarding what conclusions can legitimately be drawn from these experimental results. For instance, is it valid to conclude on the basis of this data that infants do indeed possess a high-level object

concept? Stated differently, is a high-level object- concept critical to explaining these results? Also, does it suffice to label the infants knowledge as an object concept or a belief in some physical law? Is it not important to ground out this somewhat vague cognitive notion in actual computable perceptual primitives and operations?

In this thesis I have proposed a perceptually grounded representation for physical knowledge about events in terms of changes of attentional state. In comparison to the works discussed above I would like to emphasize the relative merits of my proposal:

1. A perceptually based representation that implicitly incorporates the observed physical contingencies, is inherently simpler than an ill-defined high-level construct. Occam's razor suggests that the former should be the preferred way of accounting for experimental data. Also, by obviating the need to posit high-level object concepts, my scheme becomes a viable candidate for explaining "object knowledge" in simpler animals as well.
2. A scheme that embodies a suggestion for how it may be implemented in a real system is to be preferred over one that is mentalistic and vague.
3. My scheme allows for a uniform explanation of all experimental settings rather than having to account for each data set by positing knowledge of a different physical law.
4. Perhaps the most important motivator for my perceptually based scheme is that it relies solely on the observed contingencies rather than on any pre-ordained physical laws. Experiments have proven over and over again that physics does not govern perception, the statistics of the environment do.

4.9 Summary

This chapter focuses on the question of how visual routines can be learned. The key idea is that repeating patterns in attentional state can be learned. Subsequently these learned expectations generate visual routines to check for the corresponding event. The results shown were for only a small subset of the attentional state. More work needs to be done to show learning in the full state space. This paradigm differs significantly from the reinforcement learning paradigms currently being used. The representation of the learned patterns as *changes of attentional state* has exciting implications for the perceptual grounding problem as well as for theories of cognitive development in children.

Chapter 5

Contributions

Chapter Outline

This final chapter reviews the goals of the thesis. It states what was done, and describes what remains to be done.

5.1 The two motivating problems

I set out to solve two problems:

1. To find a robust, versatile spatial analysis machinery that can be used to solve a wide variety of spatial tasks.
2. To learn visuospatial concepts like *fall* or *more*, in an unsupervised manner.

The main motivation for the first problem is that a versatile spatial analysis machinery is necessary not only to deal with a variety of perceptual tasks from moment to moment, but the same machinery may also be re-used for spatial inference on *synthesized/imagined* representations.

The motivation for the second problem is that there must be some notion of *fall* before the system can check if an object is falling. I believe that such visuospatial concepts are learned from experience in an unsupervised manner.

5.2 Contributions of the thesis

My main contributions towards the first problem are:

- *Asserting that a single versatile mechanism for visuospatial analysis is necessary to build vision system with human level capabilities.* An application-oriented approach of having a collection of disparate programs for extracting various spatial relations will not suffice.

- *Proposing a language of attention.* I reviewed the visual routines proposal of Ullman and the related work that has been done since, and took up the two key challenges that any visual routine theory must address: developing a compact and expressive basis set of primitive operations, and developing an automatic means of combining them for a spatial task. I observed that any theory of visuospatial analysis is intimately related to a theory of visual attention. I proposed a basis set of primitives categorized into three families of operations. These families of operations constitute a *language of attention* in the sense that a visual routine is constructed by stringing together primitives from each class of operations. The word *language* is used because the model is generative in the same way that one generates sentences of a language by stringing together elements of syntactic categories. The only difference is that the sentences in the language of attention are *procedures* for extracting spatial relations.

How should you evaluate my proposal? The primary requirement of any visual routine proposal is that it be very robust at handling a wide variety of spatial tasks. The most convincing way to show that a proposal meets these requirements is by demonstrating that it can handle hundreds of real world spatial problems. While this acid test has yet to be conducted, I argue that the system can be evaluated on the basis of the degree to which it does what visual routines are supposed to *do*. Namely, visual routines are supposed to establish spatial relations between regions (at the focus of attention), and at a higher level (across several shifts of the focus attention) exploit local structure in the world, by setting up *successive local frames of reference*. I showed that the primitives of the system were explicitly designed to establish *local* and *global* spatial relationships between regions, thereby completely characterizing the spatial relations between regions. Moreover at a higher level, I showed that the system does establish successive local frames of reference, as demonstrated in the pointing example where the system uses the local context of the human to find the hand, and then uses the local reference frame on the hand to select a portion of the room.

The proposed model of attention goes beyond prior models of attention which viewed the role of Attention as mere selection. While selection is a crucial function of attention, there are other crucial issues (like the task-specific computations that are carried out at the focus of attention) that must be accounted for by any realistic model of attention. My model is a more comprehensive model of visual attention with selection taking its place as one of several families of operations that work together in the solution of some spatial problem.

If visual routines are to be truly useful they should be automatically composed by the system, not handcoded by a user. My main contributions with regard to the composition issue are

- *Visual routines are learned and not planned* (this idea builds on Chapman's insight that visual routines emerge from interaction with the world). My approach is to notice recurring patterns of activity in *attentional state*, learn those patterns, and use them as expectations to generate visual routines. I showed some examples demonstrating the learning of trajectories in attentional state, but much work remains to be done (see the following section).
- *The perceptual representation of visuospatial concepts is in terms of changes of attentional state.* In other words I am suggesting a concrete representation for perceptual

grounding. This has important implications on theories of cognitive development. It makes the theories simpler and more concrete by explaining infant behavior in terms of perceptual representations rather than in terms of “object-concepts” or “intuitive physics”.

5.3 Key issues of the future

In this thesis I proposed and demonstrated the feasibility of some new ideas for the synthesis and learning of visual routines. However, it is necessary to go beyond a demonstration of feasibility and extensively test the proposed language of attention and learning scheme, on a real-time head-eye platform. There is still considerable work to be done to make this happen, however the prospect of having a system that learns thousands of visuospatial concepts/patterns-of-activity and uses them as a scaffolding for communication and reasoning, is exciting and much closer than before.

Bibliography

- [1] N. Ahuja and M. Tuceryan. Extraction of early perceptual structure in dot patterns: Integrating region, boundary, and component gestalt. *Computer Vision, Graphics, and Image Processing*, 48:pp304–356, December 1989.
- [2] T.D. Alter and R. Basri. Extracting salient contours from images. *Proceedings of IEEE Conference on Computer Vision, Pattern Recognition*, June 1996.
- [3] Renee Baillargeon. Physical reasoning in infancy. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, chapter 11. M.I.T. Press, 1996.
- [4] J.R. Bergen and R. Hingorani. Heirarchical motion-based frame rate conversion. Technical report, David Sarnoff Research Center, April 1990.
- [5] Derek Bickerton. *Language and Species*. The University of Chicago Press, first edition, 1990.
- [6] Rodney A. Brooks. Planning is just a way of avoiding figuring out what to do next. Working Paper 303, M. I. T. Artificial Intelligence Laboratory, September 1987.
- [7] Rodney A. Brooks. Elephants don't play chess. *Robotics and Autonomous Systems*, 6:pp3–15, 1990.
- [8] Rodney A. Brooks. Intelligence without reason. A. I. Memo 1293, M. I. T. Artificial Intelligence Laboratory, April 1991.
- [9] Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:pp139–159, 1991.
- [10] David Chapman. Vision, instruction and action. A. I. Technical Report 1204, M. I. T. Artificial Intelligence Laboratory, April 1990.
- [11] Gary L. Drescher. *Made up Minds - A Constructivist Approach to Artificial Intelligence*. MIT Press, 1991.
- [12] R. Eckhorn, H. J. Reitboeck, M. Arndt, , and P. Dicke. A neural network for feature linking via synchronous activity: results from cat visual cortex and from simulations. In R. M. J. Cotterill, editor, *Models of Brain Function*, pages pp255–272. Cambridge University Press, 1989.
- [13] William T. Freeman and Edward Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:pp891–906, 1991.

- [14] Treisman & Galade. A feature integration theory of attention. *Cognitive Psychology*, 12:pp97–136, 1980.
- [15] A.P. Georgopoulos, A.B. Schwartz, and R.E. Kettner. Neuronal population coding of movement direction. *Science*, 233:pp1416–1419, 1986.
- [16] W.E.L. Grimson. *From Images to Surfaces: A Computational Study of the Human Early Visual System*. M.I.T Press, first edition, 1982.
- [17] W.E.L. Grimson, A. LakshmiRatan, P.A.O. O'Donnell, and G. Klanderman. An active visual attention system to “play where’s waldo”.
- [18] P.E. Haenny, J.H.R. Maunsell, and P.H. Schiller. State dependent activity in monkey visual cortex: Ii. retinal and extra-retinal factors in v4. *Exp. Brain Res.*, 69:pp245–259, 1988.
- [19] P.E. Haenny and P.H. Schiller. State dependent activity in monkey visual cortex: I. single cell activity in v1 and v4 on visual tasks. *Exp. Brain Res.*, 69:pp225–244, 1988.
- [20] Horst Hendriks-Jansen. *Catching Ourselves in the Act*. The M.I.T. Press, first edition, 1996.
- [21] Ian Horswill. Visual routines and visual search: a real-time implementation and an autotmata theoretic analysis.
- [22] A.L. IArbus. *Eye movements and Vision*. Plenum Press, first edition, 1967.
- [23] Mark Johnson. *The Body in the Mind - The Bodily Basis of Meaning, Imagination, and Reason*. The University of Chicago Press, 1987.
- [24] John L. Phillips Jr. *The Origins of Intellect Piaget’s Theory*. W.H. Freeman and Company, second edition, 1969.
- [25] K. Koffka. *Principles of Gestalt Psychology*. Harcourt Brace, New York, 1935.
- [26] D. LaBerge. Computational and anatomical models of selective attention in object identification. In Michael Gazzaniga, editor, *The Cognitive Neurosciences*. MIT Press, 1995.
- [27] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academics, Boston, 1985.
- [28] James V. Mahoney. Image chunking: Defining spatial building blocks for scene analysis. Master’s thesis, M.I.T, 1987.
- [29] James V. Mahoney. A programming language for interpreting graphic images. Technical report, XEROX Palo Alto Research Center, April 1995.
- [30] David Marr. *Vision*. W.H. Freeman and Company, first edition, 1982.
- [31] M.J. Mataric and R.A. Brooks. Learning a distributed map representation based on navigation behaviors. In *Proceedings of 1990 USA Japan Symposium on Flexible Automation*, pages pp499–506, 1990.

- [32] Andrew Kachites McCallum. Learning visual routines with reinforcement learning. In *AAAI Fall Symposium 1996*, pages pp82–86, 1996.
- [33] M. Mishkin, L. G. Ungerleider, and K. A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neuroscience*, 6:pp740–749, 1983.
- [34] J. Moran and R. Desimone. Selective attention gates visual processing in extra- striate cortex. *Science*, 229:pp782–784, 1985.
- [35] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:pp782–784, 1985.
- [36] Allen Newell, J.C. Shaw, and Herbert Simon. Empirical explorations with the logic theory machine: A case study in heuristics. In Feigenbaum and Feldman, editors, *Computers and Thought*. 1961.
- [37] Allen Newell and Herbert Simon. Gps: A program that simulates human thought. In Feigenbaum and Feldman, editors, *Computers and Thought*. 1961.
- [38] Allen Newell and Herbert A. Simon. Computer science as empirical inquiry: Symbols and search. In John Haugeland, editor, *Mind Design*.
- [39] Sourabh A. Niyogi. Deducing visual relationships from attentional representations. In *AAAI Fall Symposium*, pages pp63–69, 1996.
- [40] N.R. Pal and S.K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):pp1277–1294, 1993.
- [41] Jean Piaget. *The Origin of Intelligence in Children*. W.W. Norton and Company, first edition, 1963.
- [42] Steven Pinker. The evolution of cognition. Notes for an I.A.P. 92 talk at M.I.T.
- [43] J. R. Pomerantz and E. A. Pristach. Emergent features, attention, and perceptual glue in visual form perception. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):pp635–649, 1989.
- [44] M. I. Posner. Attention as a cognitive and neural system. *Current Directions in Psychological Science*, 1:pp11–14, 1992.
- [45] M. I. Posner and S. E. Petersen. Attention as a cognitive and neural system. *Annual Review of Neuroscience*, 13:pp25–42, 1990.
- [46] M. I. Posner and M. K. Rothbart. Attentional mechanisms and conscious experience. In A. D. Milner & M. D. Rugg, editor, *Foundations of Neuropsychology Series*, pages pp91–112. New York: Academic Press, 1992.
- [47] Rajesh P.N. Rao and Dana Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:pp461–505, 1995.
- [48] I. Rock and S. Palmer. The legacy of gestalt psychology. *Scientific American*, 12:pp48–61, 1990.

- [49] Rosenbloom, Laird, Newell, and McCarl. A preliminary analysis of the soar architecture as a basis for general intelligence. *Artificial Intelligence*, 47:pp289–325, 1991.
- [50] 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.
- [51] Pawan Sinha. Image invariants for object recognition. *Investigative ophthalmology and visual science*, 35, 1994.
- [52] 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.
- [53] Elizabeth S. Spelke, Peter Vishton, and Claes von Hofsten. Object perception, object-directed action, physical knowledge in infancy. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, chapter 10. M.I.T. Press, 1996.
- [54] H. Spitzer, R. Desimone, , and J. Moran. Increased attention enhances both behavioral and neuronal performance. *Science*, 240:pp338–340, 1988.
- [55] V Tomlin, R. S. & Villa. Attention in cognitive science and sla. *Studies in Second Language Acquisition*, 16(2), 1994.
- [56] A. Treisman and S. Sato. Conjunction search revisited. *J. Exp. Psychology; Hum. Perception*, 16:pp459–478.
- [57] D. A. Trytten and M. Tuceryan. Segmentation and grouping of object boundaries using energy minimization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages pp730–731, Maui, Hawaii, June 1991.
- [58] Shimon Ullman. Visual routines. A. I. Memo 723, M. I. T. Artificial Intelligence Laboratory, June 1983.
- [59] M. Wertheimer. Laws of organization in perceptual forms. In W. D. Ellis, editor, *A Source Book of Gestalt Psychology*, pages pp71–88. Harcourt Brace, 1938.
- [60] Patrick Henry Winston. Learning structural descriptions from examples. In Patrick Henry Winston, editor, *Psychology of Computer Vision*. MIT Press, 1975. Based on a Ph.D Thesis, MIT 1970.
- [61] Patrick Henry Winston. *Artificial Intelligence*. Addison-Wesley Publishing Company, third edition, 1992.