A Visuospatial Representational System

Satyajit Rao

September 1, 1995

Abstract

This paper describes the issues involved in the construction of a Visuospatial Representational System, the first step in my long term program for building a representational platform for cognitive tasks (Rao 1995). The main intuition in this paper is that there are certain visuospatial representations and processes that can be used not only for grasping an object or navigating around an obstacle but are also fundamental to more abstract tasks. Such as, understanding sentences about spatial events like "The ball fell into the bucket", or supporting concepts like "solar system", or proving a theorem in mathematics. It is important to look at the representational and inference capabilities of these processes because they highlight the contribution of Vision to Cognition.

In particular, there are three types of visuospatial processes that through their interactions with other representational systems (syntactic, motor–somatosensory, and social-thematic) support cognitive tasks:

1. Detecting spatio-temporal regularities.
2. Extracting spatial relations on demand.
3. Synthesizing spatial descriptions on demand.

The goal of this thesis is to construct a visuospatial representational system with the above three components, and show how they interact with other Representational Systems to support tasks like common-sense reasoning, making obvious cognitive judgments, and understanding natural language.

1 What do we want?

Building a robot with human like capabilities has been the holy grail of Artificial Intelligence. After more than 40 years of trying to do this we are still very far from that goal. There is a large gap between the cognitive capabilities of the ant like robots that we can build today and the human like robots that we would like to build. In this section I describe the outline of a program for bridging this gap. The thesis proposal in the following sections is the first step in this program.

First we must decide on what we really want when we say "human like intelligence" or "cognition". Implicit notions about the capabilities that seem to lie at the crux of human intelligence have directed (and often misdirected) entire research approaches.

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; having concepts like "solar system", or understanding how a bicycle pump works, or proving theorems 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 we want when we say "human like intelligence".

The Problem

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 (Pinker n.d.) (Bickerton 1990) - 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 example [1]) by using visuospatial operations on spatial representations of some symbols, imputing anthropomorphic roles to the components of the expression, and using inheritance properties of linguistic categories). Whether you believe this particular example or not is not the point. The point is that some such explanation of our abstract cognitive abilities in terms of representations and processes ment 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.**

[1] I choose this example because doing mathematics is considered to be one of the abstract cognitive feats of the human brain

One key idea is that there are several independent representational systems, the dynamics of whose interactions leads to the “accidental byproducts”. The thesis proposal described in the rest of this paper outlines the plan for the construction of one of these representational systems, namely a visuospatial representational system.

2 Vision and Cognition

Forty years ago the primary motivation for working on machine vision was that it would enable the machine to recognize objects in the scene. The symbolic “scene description” could then be used by other modules to “reason with”. Gradually, it was realized that this was an unreasonable requirement because there is no single “scene description”. The more recent task based view of Vision, takes a step forward by giving the visual processes specific goals depending on the task at hand. Such as looking for someone, or picking up an object. However, even this view of the function of vision puts it only on the periphery of all the representational and reasoning phenomena that we call Cognition. Recognizing objects, and directing action are important functions of vision, but there is more to it than that. *We rarely think of visual processes as useful for inference, induction, problem solving, understanding natural language, understanding how a device works, doing mathematics, or planning a trip, - some of the skills that distinguish human cognition*. This thesis is about Vision’s contribution to Cognition. It is based on the premise that vision is not simply some peripheral process that supplies information about the world to other parts of the brain that do the “reasoning”, but that it is an integral part of the “reasoning” itself. It is very important to identify and study the visual processes that support abstract representation and inference because of two reasons. Firstly, from a scientific point of view we get an insight into the origins of human cognition. We see how some visual processes which originally evolved to direct action, interact with other representational systems to support the kinds of cognitive skills that we display. Secondly, from an engineering point of view we get a clue about how to build vision systems for robots such that the visual processes not only support the routine tasks of directing actions such as reaching, grasping, and navigation, but the very same processes also help understand language, play chess, and prove theorems.

As described in section 1, by *cognition* I mean the capacity to support rich representation and inference. The emphasis is *not* on any individual task like playing chess, or solving logic problems, but on the *ability* to support many such tasks without having specialized representations tailored for game playing or theorem proving built in a priori. From now on I will use *cognition* in this sense.

Under what constraints will visual processes that evolved to direct action also support cognitive tasks like understanding language or planning a vacation?

The first constraint is on the nature of the task. The task should have a meaningful spatial analogy. That is, there should be some mapping from the entities, predicates, relations of the task domain to a spatial representation (say a 2 dimensional topographic map). Inherently spatial tasks like laying out the furniture in the room will obviously benefit from such a mapping. However, many tasks that do not appear to have anything to do with space, like planning your schedule or evaluating data, may benefit from a mapping to a spatial representation (a spatial analogy). It is no accident that we use so many spatial props such as charts, graphs, maps, and that there is an abundance of spatial metaphors in natural language. We actively map problems into a spatial representation in order to be able to draw inferences efficiently with our spatial problem solving machinery.

The second constraint is on the types of visual processes. If we somehow understood the workings of all the visual systems that nature has evolved, then what aspects of the human visual system would be point to as the key abilities that support human cognition? The following three processes are likely candidates:

- 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 1987). Therefore, it is important to have mechanisms that detect spatio-temporal regularities.
- Extracting spatial relations on demand: Visual processes that evolved to direct action should not be tied to extracting only particular aspects 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 infinite number of visual features that go well beyond what the hardwired behaviors require.
- 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 third constraint is the presence of other representational systems which are very different ways of representing events in the world. These Representational Systems help the visual system by providing a top-down bias for figure-ground segmentation, directing the focus of attention, and sometimes deciding the sequence of application of visual operators.

The goal

The goal of this thesis is to first construct a visuospatial representational system that detects spatial regularities, extracts spatial relations on demand, and synthesizes spatial representations on demand, and then show how these capabilities interact with a syntactic Representational System to support cognitive tasks such as making obvious cognitive judgments about space, and understanding simple sentences in natural language.

3 The Micro-Cog Visual System

The micro-cog visual system consists of two, color, wide-angle, 110 degree field of view, Chinon cameras each with 2 degrees of freedom. They are both mounted on a head which has three degrees of freedom. Before discussing mechanisms that extract spatial relations and synthesize spatial descriptions, I will need to make some choices about what gets computed in early and preattentive vision, the choice of spatial representations, the attentional mechanisms, and event memory. Most of these areas are far from being understood and are worthy of PhD thesis in themselves. However, my first priority is to implement all of these stages and have a complete system. Therefore, I have tried to choose very simple representations and mechanisms that will serve my purpose in each of these areas. These choices are described below.

Early Vision: The goal of early vision is to use local processes to make local properties explicit. As Figure 1 shows, the properties that I am concerned with are color, motion, and depth. The result of early vision processes is to specify these properties in a pointwise manner. The optical flow is simply computed by correlation of a patch of fixed size within a fixed displacement radius. Depth will be computed by a stereo correspondence of color patches. At the moment color is left as is, in an HSV format. There is no attempt as yet to map the HSV values to perceptual color categories as done by Syeda-Mahmood (1992).

Preattentive Mechanisms: The goal of this stage is to do crude "data driven" (bottom-up) figure–ground segmentation by using local spatially uniform processes to find large scale structure. As Figure 1 shows, salient *regions* are derived independently by local processes from each source of information - color, motion, and depth. The segmentation into regions is done by grouping pixels with common properties. Color regions are derived by linking together patches of size S x S (for some scale size S) that have similar color histograms. There are two choices for extracting motion regions; either grouping motion locations by proximity or by direction of the optical flow vector. At the moment the first method is used. Relative depth will be computed from the stereo pair of images after fixating on a region of interest. Note that at the end of the preattentive stage there are three data-driven bottom-up segmentations based on three different sources of information and a measure of saliency for each source. A single figure ground interpretation of the scene is not available until the constraint from the behavior or task currently in control is taken into account. The issue of doing data driven figure ground segmentation with local processes is a very interesting and open issue. However, I have chosen to focus on the attentional processes that operate on the results of the preattentive stage.

Spatial Representations: The salient regions from preattentive processes have so far been described only in a topographic *retinotopic* representation of space. The positions of the regions will change every time the eyes move. In order to capture the constancy in the positions of stationary objects, an *egocentric* representation of space will be used. My assumption is that this egocentric representation only describes the positions of regions that have been attended to in the recent past. In other words, only regions that have been attended to can appear in this representation. The retinotopic position of the attended

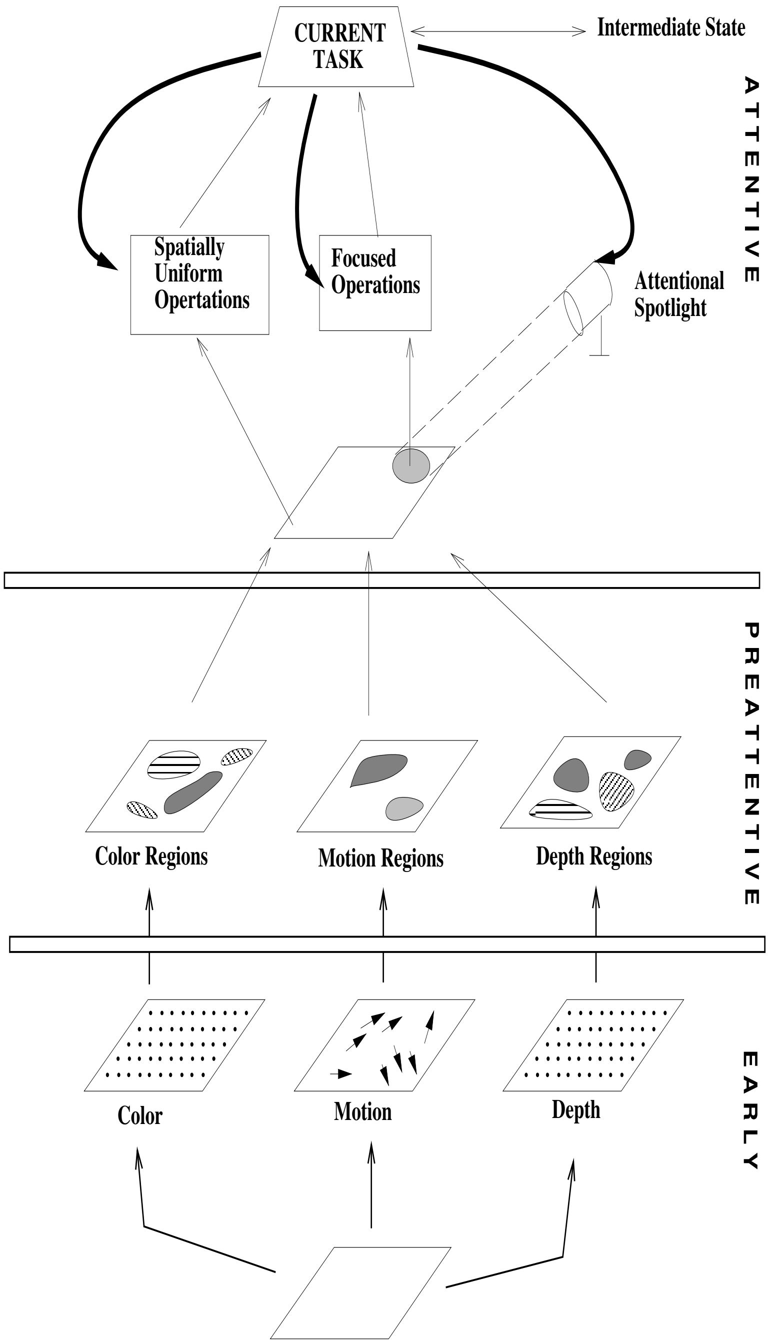

Figure 1: Early, Preattentive, and Attentive Stages of the Micro-Cog visual system

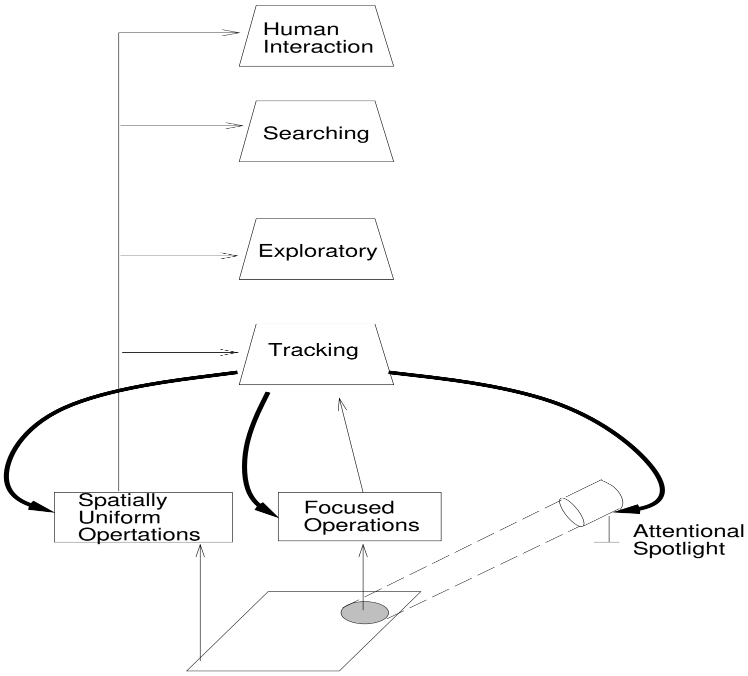

Figure 2: Attention and Behavior

region is combined with eye-position, and head position, to give the position of the region in a body centered coordinate frame.

Attention: From now on visual processing ceases to be purely bottom up, or data driven. Figure 1 shows the model of attention that will be used. The control system for a task has three resources available to it; a set of spatially uniform operations, a set of focused operations, and control on the location of the attentional spotlight. The spatially uniform operations are divided into two sets - those that are always applied irrespective of the task (Universal routines in Ullman's terminology), and those that are task specific and are only applied when that particular task is in control. Typically the three resources will be applied in the following way. First the spatially uniform operations is applied to the image. The results of these operations along with the task history determine where the attentional spotlight will be shifted. Some focused operations are applied in the attended region, the results of this with the task history determine the next cycle. The task history consists of the sequence of attended regions, as well as the integration of information from the spotlight across attentional shifts. The location of the attentional spotlight is decided not only by saliency measures from the bottom up processes of the preattentive stage but also by the task at hand. The task that is currently in control, such as looking for an object, or counting objects, decides all of the following: which operations are applied uniformly across the retinotopic maps, which operations are applied only at the attentional focus, what to do with the results of the operations at the attentional focus, where the attentional focus should be shifted to, and finally the integration of information from the focus across the shifts. These issues will be discussed in the section on Visual Routines.

Behaviors: Figure 2 shows some of the canned behaviors that will be implemented. A behavior when activated controls all the attentional resources (in the figure the Tracking behavior is shown to be in control). No two behaviors that share the same motor pathway or attempt to use the attentional spotlight can be active at the same time.

The *Exploratory* behavior: The head-eye looks around the room, fixating at regions that catch its attention. It binds the identity and locations of those regions in egocentric coordinates. This exploratory behavior will generate expectations/defaults about the absolute positions of regions in egocentric coordinates. The bottom up candidates for attention can be odd-man out locations, and the topdown candidates can be regions of "surprise". That is, regions where some expectation about the position or identity of a region has been violated. Another kind of regularity learned will be the results of one's motor actions on the environment. Gary Drescher's (Drescher 1991) Schema Mechanism constructs schemas that are simple pieces of procedural knowledge that state the effects of actions on the environment, they form a procedural model of the environment. For the head-eye system the only actions that exist are eye and head motions. The exploratory behavior helps the system build up a mapping from position of a region and an action to a new position of the region, thus building a motor representation of space.

During *Searching* The head-eye looks for an object that it has seen before. Object identity is simply determined by color histogram, and size attributes. Learned knowledge about the expected locations of objects i s used to look at some part of the room. Parallel searches for the color and size attributes of the object on the results of preattentive processes provide candidate regions. A sequential shift of attention through these regions helps bind the properties at those locations and enables a better comparison with the object model.

Another set of behaviors that involve human interaction with the head eye, such as paying attention to human gestures, take precedence over the exploratory behavior,

4 Detecting Spatiotemporal Regularities

The evolution of human intelligence is correlated with a rapid increase in size of the neocortex. One theory of the function of the neocortex views it as a device dedicated to *discovering suspicious coincidences* of events (Barlow 1985). In the visuospatial domain this means detecting coincidences and recurring patterns in the position and motion of objects. The detection of these coincidences leads to a richer model of the environment. Research with human infants (Spelke 1991) shows that babies are surprised by object motions that violate certain spatiotemporal constraints like continuity of motion and solidity, and by objects that appear to be suspended in air without support. Thus, at a very young age humans already have a very rich model of the physical behavior of the world around them. A vast repository of such physical knowledge (typically called "common-sense") is necessary for not only reasoning about the world immediately in front of us, but also for abstract cognitive tasks, some of which may have little to do with real space. For example, making sense of the sentence "Prices are going up". Johnson (1987) has described a theory of metaphor in which *recurring patterns of bodily interactions* with the world are noticed, and captured as *schemas*. Thereafter, we make sense of all our experience buy interpreting it in terms of schemata. It is the applicability of these schemas to domains far removed from the situation in which they were learned, that makes them very useful.

Hand-coding such schemas into a robot is not a viable option because of several reasons: there are too many schemas, they may depend fundamentally on the particular embodiment of the robot, we may not be able to make many schemas verbally explicit, and our verbal descriptions may be inappropriate representations. There is no substitute for mechanisms that notice recurring patterns in the environment, and also in the bodily interaction with it.

For the purpose of this thesis I will build mechanisms that learn recurring patterns in the relative spatial positions of objects, the motion profile of objects, and in eye-head movements. Examples of defaults that I expect the system to learn are: the layout of the furniture in the scene, the prefered motion directions of humans (in the horizontal direction), the mapping between a motor command and the expected effect on a scene.

I do not have details about the regularity detection mechanisms because I haven't started working on them yet.

5 Extracting spatial relations on demand

Apart from the ability to recognize objects and specify their locations (the what and where), our visual system can extract a variety of spatial problems on demand, as the following example shows. During the course of a birthday party the party organizer may have any of the following questions in mind:

1. Are the balloons still on the wall?
2. Does each kid have a cup to drink from?
3. Is there enough cider to refill a cup?

The human visual system delivers the answers to any of these questions with astonishing speed and accuracy. While all of these questions deal with the spatial location and extent of various objects in the image, the computational tasks posed by these questions are different from each other. Question 1 requires locating an object (the balloon) within the confines of another object (the wall). Question 2 not only involves locating the cups and kids but also making sure their spatial distribution is such that each kid is within reach of a unique cup. Question 3 involves at the very least, reasoning about the size of the paper cup, the size of the cider container, and the level of cider in the container. Different tasks require a visual system to extract different kinds of spatial information, and the visual system must have the machinery to do so. The problem is to find visuospatial machinery that is capable of extracting a variety of spatial relations on demand. First it will be useful to review the work done on this problem.

The framework of a solution: Visual Routines Shimon Ullman initially proposed the problem stated above (Ullman 1983) and also described the framework of a solution. The

Figure 3: A birthday party

essence of his solution is as follows: Ullman suggests that the overall processing of visual processing is divided into two stages. The bottom-up, spatially uniform, viewer-centered computation of the base representation (like the $2\frac{1}{2}$D sketch) followed by the extraction of abstract spatial properties by *Visual Routines*. It is the visual routines that define objects and parts, their shapes and spatial relations. He suggests that there exist a set of *elementary operations* that when combined in different ways produce different visual routines. The elementary operations therefore form a kind of basis set for visual routines. 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 may 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.

I find this to be an attractive framework for a solution primarily because of the core idea of composing a basis set of elemental spatial operations on demand. Clearly this kind of explanation is far more plausible than having specialized detectors for features like "the smallest object inside the big circle". The details of the solution are however far from clear. The paper is only intended to expose the problem and suggest a direction. Questions about how visual routines are synthesized from the basic operations, and what initiates them are not addressed.

A programming language for visual routines

James Mahoney at Xerox PARC (Mahoney 1992*a*) (Mahoney 1992*b*) has developed some of the elementary operations proposed by Ullman and extended the set with some new ones. He has implemented them to form a simple but powerful visual programming language where the user combines elementary operations to form visual routines that take images as arguments and return images as results. Mahoney's basic operations 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 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.

$$LocationMap \rightarrow PropertyMap$$

$$PropertyMap \rightarrow LocationMap$$

$$PropertyMap \times LocationMap \times Summarizer \rightarrow SummaryValue$$

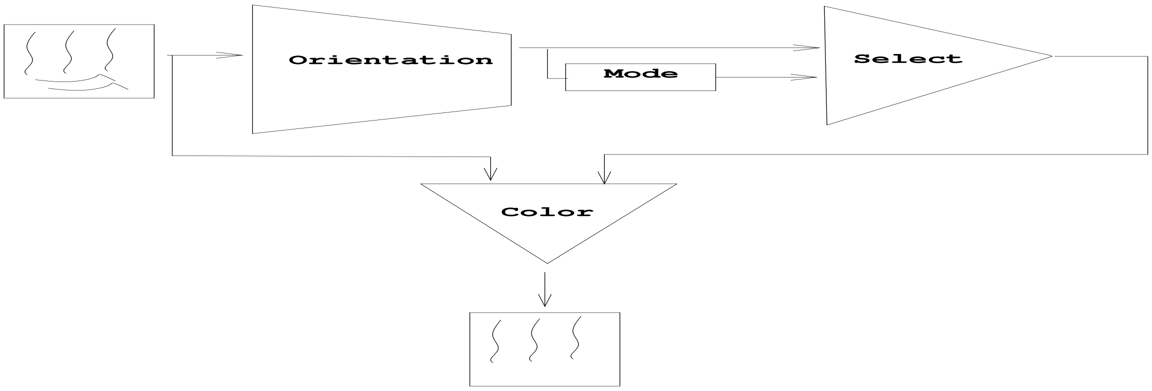

Figure 4: An example of a visual routine in Mahoney's language

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. Appendix A describes these operations in more detail.

Figure 4 shows an example of how some of the basic operations can be combined to select the items in the image with the dominant mode of orientation. First, the mode of the orientation is selected for from the image. However, the resulting regions only have the vertical pieces of the desired curves. The curves can be selected by using the first mode regions as seeds to color the image, thus giving the vertical curves.

I have found Mahoney's representation and operations to be very useful in capturing spatial relations because of the following reasons:

- *Preservation of spatial information:* The simple 2d topographic representation preserves spatial information at all stages.
- *Positionally exhaustive:* At every level of processing the information in the image is positionally exhaustive, properties are defined at every location. This enables simple *local computations* to capture abstract geometric properties.
- *Flexibility in combining operators:* A uniform representation enables operations to combine in powerful ways. Much like analog circuits whose very simple currency of voltages and currents places very little constraints on which circuit elements can be connected together, Mahoney's operators deal with only three kinds of currencies and thereby get their compositional power.

However, as of now it is the user who must compose some operators in order to extract a particular spatial relation. Mahoney has not yet addressed the problem of how the dynamics of attention can impose some constraints on the composition of these operators.

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

David Chapman's "Sonja" (Chapman 1990) 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.

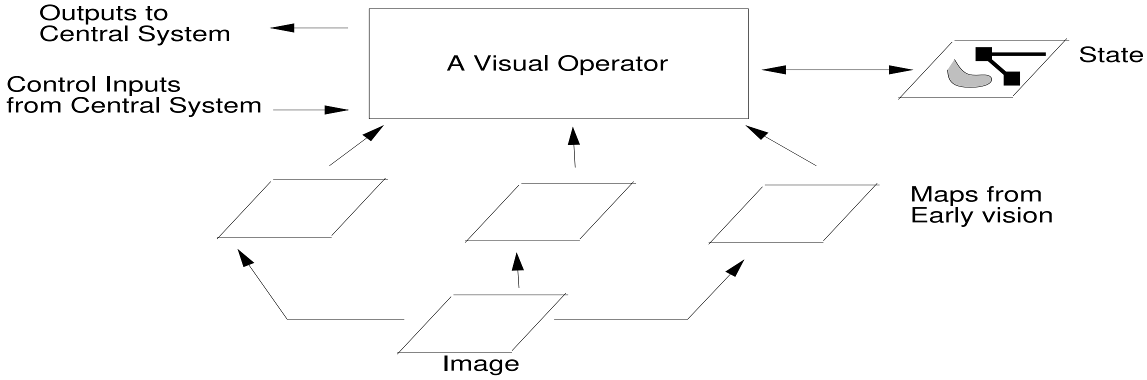

Figure 5: A visual operator from Sonja

The machine's only inputs are the visual image of the 2D world and occasional instructions from a human. For this discussion I will ignore the linguistic input and concentrate on the system's use of visual routines to determine its course of action. The two major components of the system are the visual operators and the central system that controls application of the operators.

Figure 5 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 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 6 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*

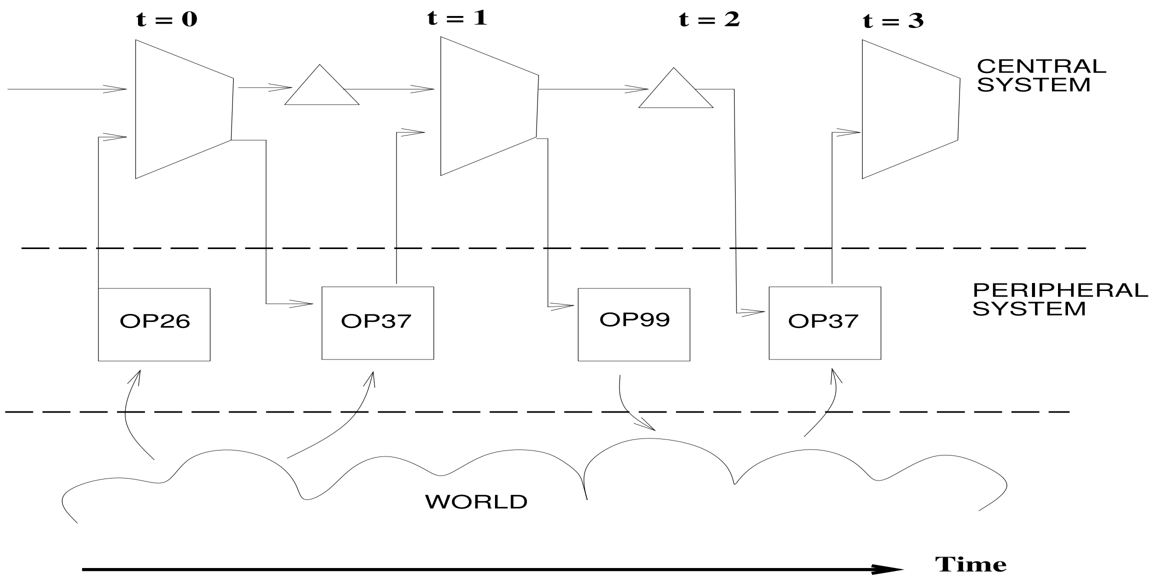

Figure 6: Visual routines are interactions, not programs.

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. The very same observation was made by Brooks (1987) 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.

In the context of the attentional framework shown in figure 1 you see that Mahoney's work has contributed to specifying a set of spatially uniform operations. Ullman's suggestions for the elementary operations are mostly about the focused operations and the control of the attentional spotlight. Chapman's contribution is in clarifying the structure of the controller Visual Routine Processor when the spatial relations being computed are fixed and known in advance.

Open issues

Testing in the real world: The effectiveness of the elementary operations suggested by Mahoney and Ullman are completely untested on real images from environments. This needs to be done urgently.

Representing new spatial concepts/defaults and having these initiate visual routines. One of the requirements of Ullman in his Visual Routines proposal was that the Visual routines processor be open-ended, that is it should be capable of extracting a large variety of spatial relations including newly defined ones. Chapman's work does not address this important issue. The Amazon world has a finite number of known objects, a small number of well defined spatial relations that are extracted. The wiring of the central system implicitly encodes the visual operators that deal with these objects and spatial relations, and the context in which the visual operators should be applied. There is no discussion of how new spatial relations can be learned and how we form visual routines to extract these relations.

Visual routine synthesis: Rather than have a synthesis unit, the dynamics of the interaction of attentional biases with other Representational Systems and the external world may be sufficient to produce an interaction that looks like a visual routine.

Initiating visual routines from other Representational Systems: How does a linguistic input like "How many chairs are near that table?" eventually lead to a visual routine?

These are the issues that I intend to deal with in the next few months.

6 Synthesizing spatial representations

The ability to recall previously constructed spatial representations or synthesize entirely new ones from spatiotemporal defaults is the third component of the visuospatial representational system. Humans can count the windows in their house without being there, or solve problems like "An ant walks 1m North and 1m East on a plane, which is the shortest path back to the starting point?". Farah (Farah 1992) cites two sources of evidence as supporting the hypothesis that the human brain uses the same machinery for imagery and perception. Behavioral data indicates that humans who have damage to right hemiparietal areas cannot even imagine the left side of previously experienced spatial scenes (Bisiach & Luzzatti 1978). Medical imaging data indicates (Farah 1989) that the prestriate occipital, parietal, and temporal cortices get activated during an imagery task - the same areas that get activated during perception. Having the same representation and hardware for perception and imagery is a parsimonious solution because it means that the entire visual routine machinery can be used on imagined representations just as easily as it is can during perception. The ability to create new spatial descriptions and extract various details leads to a powerful problem solving capability. For example, given a problem like "Tower A is taller than B but shorter than C, is Tower B shorter than C?" I would expect the spatial representational system to be able to synthesize a spatial description of the problem from the defaults that it has acquired (Figure 7 shows a possible description) and then generate a visual routine to answer the question.

The traditional propositional inference techniques require an axiom that contains transitivity of "taller than" in order to derive the fact that B is shorter than C. However, as it is a spatial problem, I would expect a spatial representational system to solve the problem with its inference processes, namely visual routines. Note that the transitivity of "taller than" is implicit in the spatial representation of figure 7, the system does not have to possess the concept beforehand. In fact it could represent the concept explicitly if necessary in another Representational System after solving many such problems.

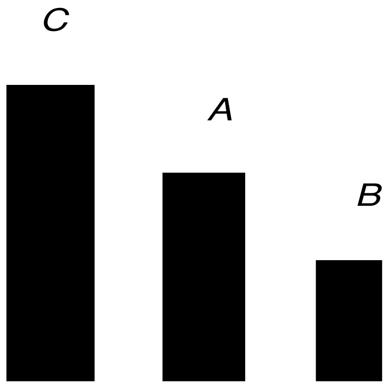

Figure 7: An imagined scene constructed from previously learned defaults about towers and a notion of "taller-than".

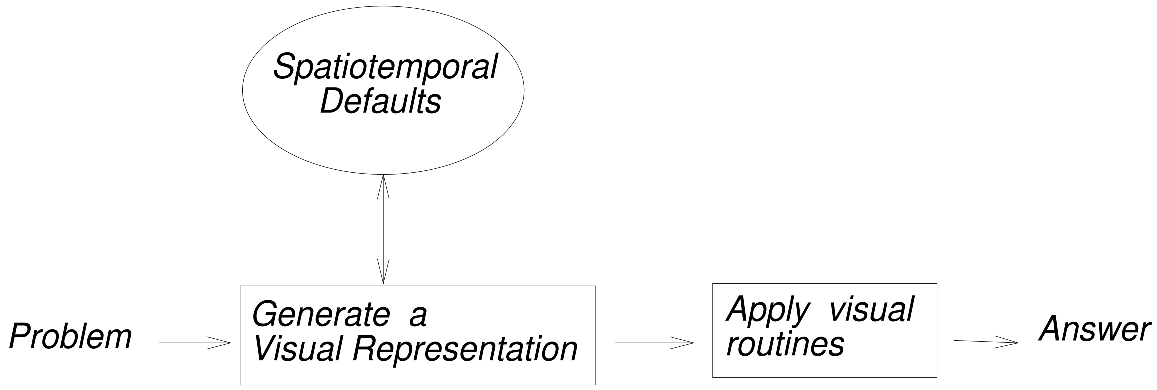

Figure 8: Problem Solving with imagery and visual routines.

The issue here is how is the system going to synthesize an image, or some other intermediate spatial representation? Figure 8 shows the generation stage drawing upon previously acquired defaults and precedents, it will also involve a feedback loop involving visual routines that keep checking if the description being generated meets the problem specifications (Tower A should be taller than B ...etc), if not some adjustments are made to the image to reduce the error. There are some problems that any image synthesis scheme will encounter. In Figure 7 the towers had their bases at the same level and the shape of the default towers was fortunately irrelevant to the problem. While synthesizing an image from the spatial constraints a lot of specific factors arise like the placement of objects, the shape, emergent spatial relations between components. Some of these are crucial to the solution of the problem (the leveled bases of the towers for instance) and some of them are irrelevant and even misleading (the shape of the towers for instance). This is a problem that humans often encounter. As a bit of anecdotal evidence, grade school students when asked to prove some theorem about an arbitrary triangle, would make the mistake of first drawing an isoceles triangle as their default triangle and then basing their proof on that diagram. As with humans, the only way to fix this is to wait for the mistake to happen and then learn from it.

This concludes the discussion of the issues involved in the construction of a visuospatial Representational System.

7 Testing the Visuospatial Representational System

Once the components of the representational system described in previous sections have been implemented, I expect to take the system through the following stages where each stage builds on the capabilities of the previous stages.

Unsupervised, behavior driven learning of spatial regularities: When both the visual and syntactic Representational Systems are in place the first step is to notice recurring patterns in the behavior driven interaction with the environment. For example the egocentric positions of certain stationery objects, or the recurring consequences of certain motor actions, or the pairing of certain word utterances with certain spatial changes in the scene. As mentioned earlier these regularities are noticed within *and* across Representational Systems. At this stage it is the behaviors that specify the visual tasks and initiate the visual routines.

Human supervised learning of certain spatial relations, and word meanings: A human teacher uses the attentional biases of the system to focus its visual attention on certain spatial relations in the scene, while simultaneously focusing linguistic attention on certain new words in the sentence uttered, thus causing a binding between the word and the spatial relation. The temporal associative memory across Representational Systems does the binding.

Answering simple queries about spatial aspects of the current scene: Queries about the current scene couched in terms of the words already learned, are put to the system. For example "Is the lamp closer to the table than the chair?")

Answering queries about hypothetical spatial events: Natural language Semantics has proved to be a very difficult problem because of the vast amounts of expertise in various domains that seem to be needed to understand even simple sentences. I address one of these domains of expertise, namely spatial relations, I want to show how the spatial expertise of the visual reasoning system enables the robot to understand simple English descriptions of spatial events by synthesizing spatial descriptions and then extract spatial relations, For example "The ball fell into the bucket" and thereafter be able to answer questions like "Is the bucket open?" and "which way is the opening facing?". This task will allow the system to demonstrate the full power of the visuospatial system because it demonstrates the working of all three components of the system.

The visuospatial system is intended to be one among many Representational Systems. A very simple version of a Syntactic Representational System is being constructed in conjunction with William Neveitt. Providing a simple language interface will permit humans to communicate their queries, while the machine can inform the human about its spatial concepts in a form that is understandable to the human.

So far the Syntactic Representational System has been presented mostly as an interface to the Visuospatial Representational System, and as providing an easy channel of communication to humans. However, it should be emphasized that this is not the role of the Syntactic system in the long-term. It is a Representational System in its own right. It captures patterns in syntactic structures, has categories, has its own attentional mechanisms, and does inference (see McAllester (1991) for one possible set of such inference rules). The syntactic categories of objects guides visual attention and problem solving, and similarly the spatial attention guides the learning of meanings of words and phrases. The dynamics of this interaction is something that I would very much like to exploit and demonstrate if time permits.

8 Schedule

The visuospatial representational system will be implemented on the head-eye system described in section 3. The early and preattentive vision will be done on C30 or C40 DSP boards, all other aspects of the system will be implemented on a workstation such as a Sparc 5 or a Silicon Graphics machine. The implementation will follow the sequence in which the issues have been discussed in this paper.

Table 1 Shows the Schedule. The first nine months will be spent implementing the various components of the system, after which the next four months will be used for testing applications.

Year	*Month*	*Topic*
1995	August	Preattentive processes
	September	Visual Routine Control
	October	Regularity Detecting Mechanisms
	November	Event Indexing and Retrieval
	December	Imagery
1996	January	Testing with other represenational systems
	February	
	March	

Table 1: Schedule for the thesis.

A The Spatially uniform operations

This appendix describes the Spatially uniform operations developed mostly by James V. Mahoney of Xerox PARC.

A.1 Establishing properties: Location map → 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. Some examples are Area, Elongation, Connected components, and the angle of the axis of least inertia. Examples of the computation of some of these properties follow.

Area This operator simply finds the areas of the regions in the location map. Each pixel of each region is labeled with the area of that region.

Region orientation

This operator computes the angle made by the axis of least inertia. Each location of a region is labeled with the angle of orientation of that region.

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 some properties.

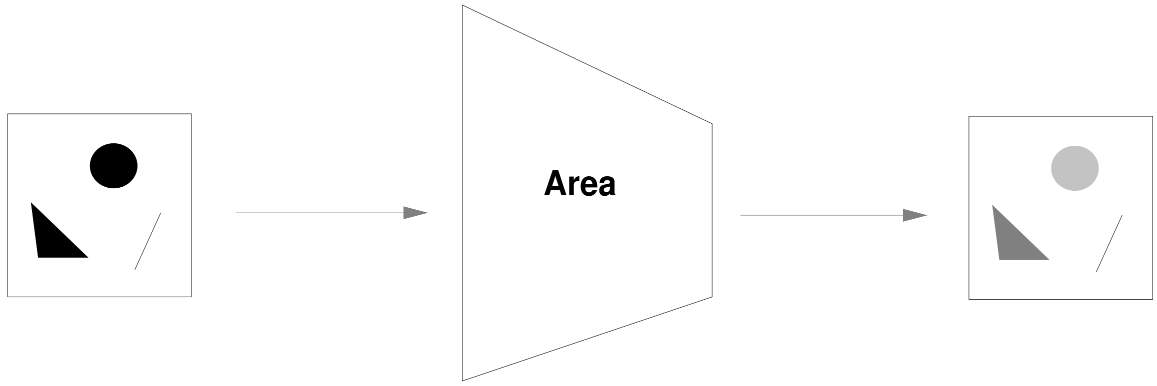

Figure 9: The area operator. Each location is labeled with the area of the region to which it belongs.

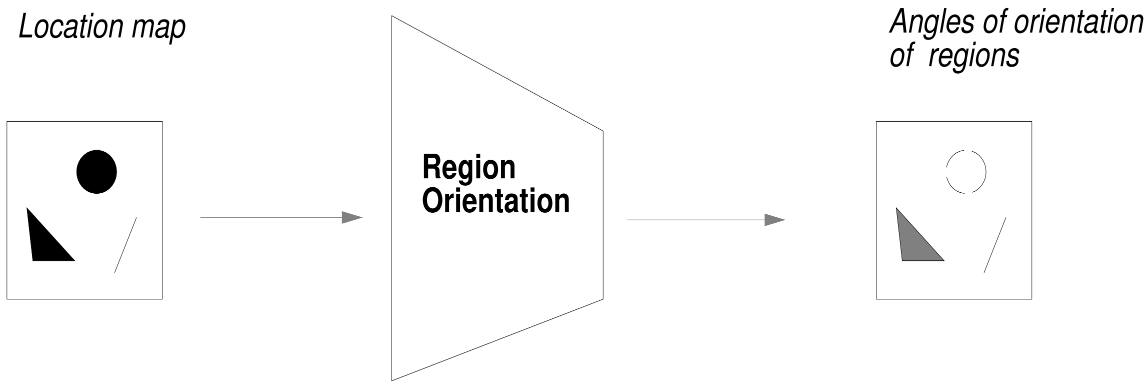

Figure 10: Each location is labeled with the orientation of the region to which it belongs.

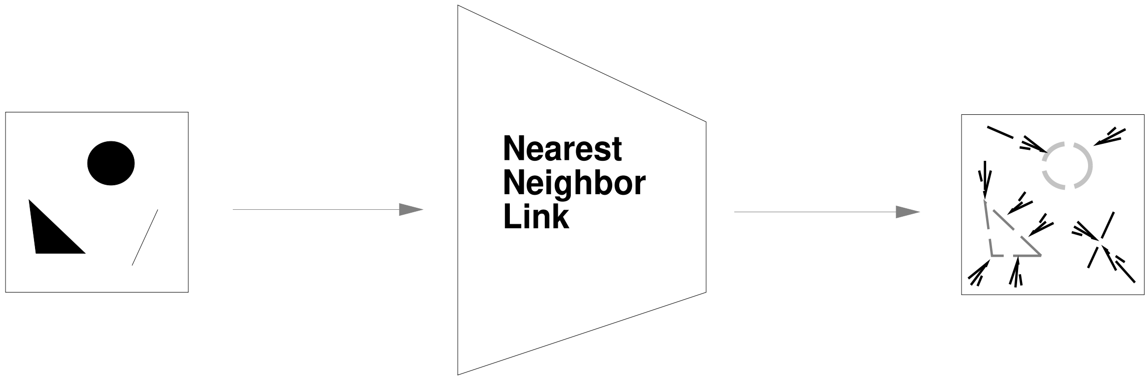

Figure 11: The Nearest Neighbor linking operator. Each location forms a link to the nearest occupied location.

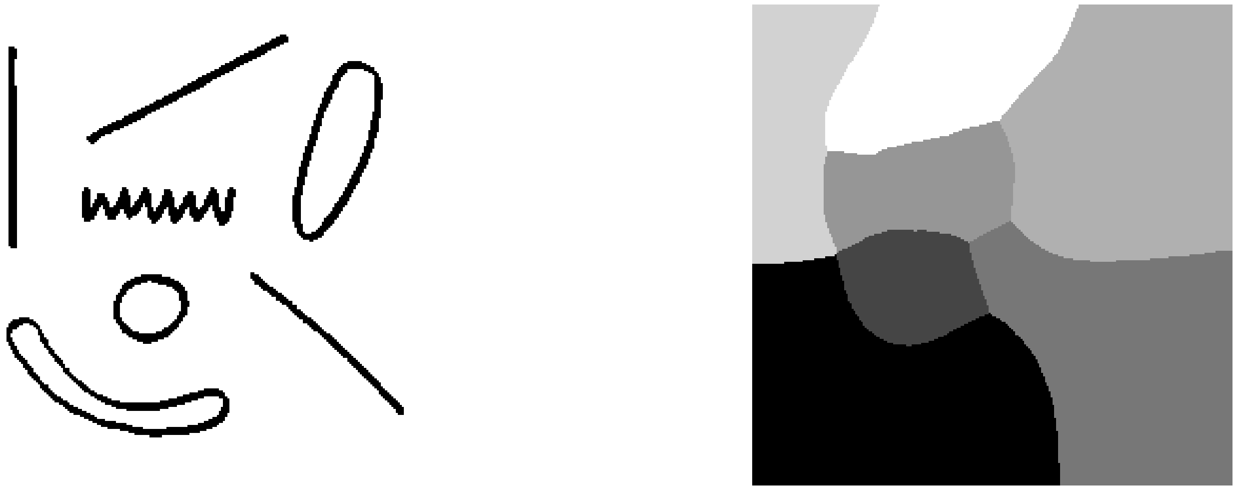

Figure 12: Left: The location map. Right: The Vornoi diagram obtained from Nearest Neighbor links and Connected Component information

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.

Nearest Neighbor: The input is a location map and the output is nearest neighbor link information. Each location forms a link to the nearest location of the region nearest to it. The positionally exhaustive link information is very useful 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.

Directional Nearest Neighbor: Each location forms links in a finite number of directions (North East West South) to the nearest region in those directions. One way of thinking

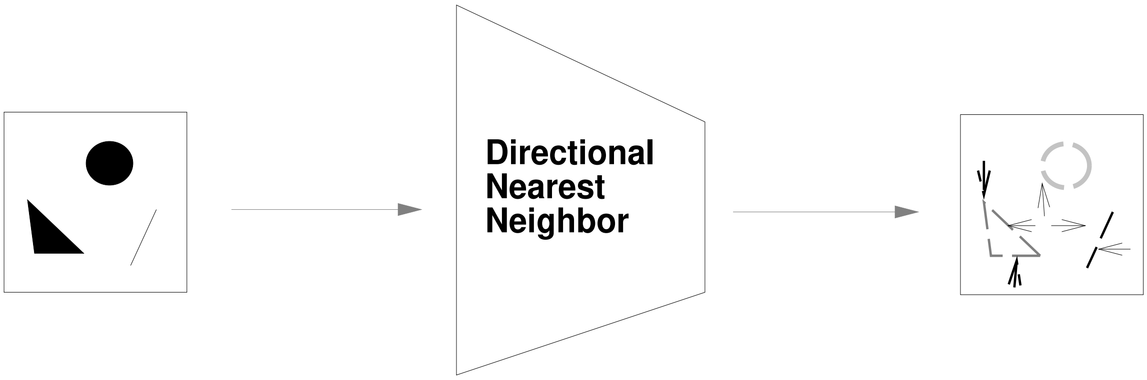

Figure 13: Relative Position links.

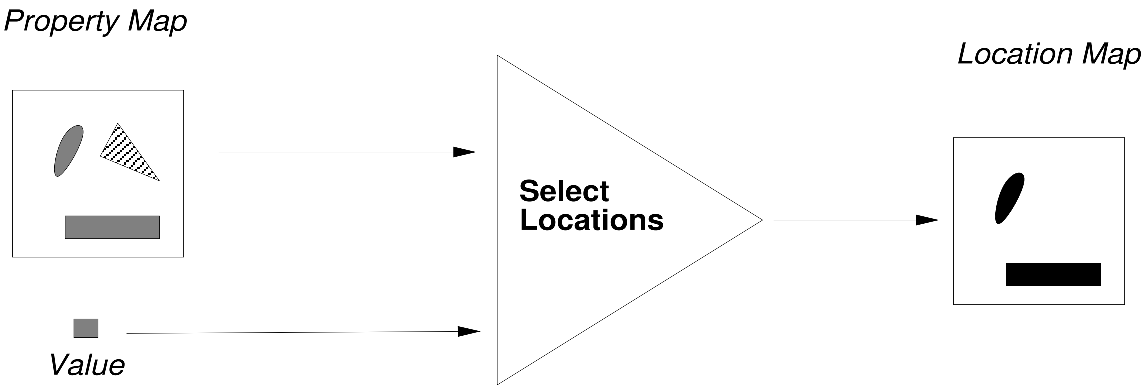

Figure 14: Only locations that have a particular value are selected.

about this is that each location independently describes the relative positions of the salient regions from its point of view.

A.2 Selecting locations with certain properties: Property map → Location Map

Select locations that have a particular property value This operation selects all locations that have a particular value from a property image.

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.

A.3 Property Map × Location Map → 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 etc.

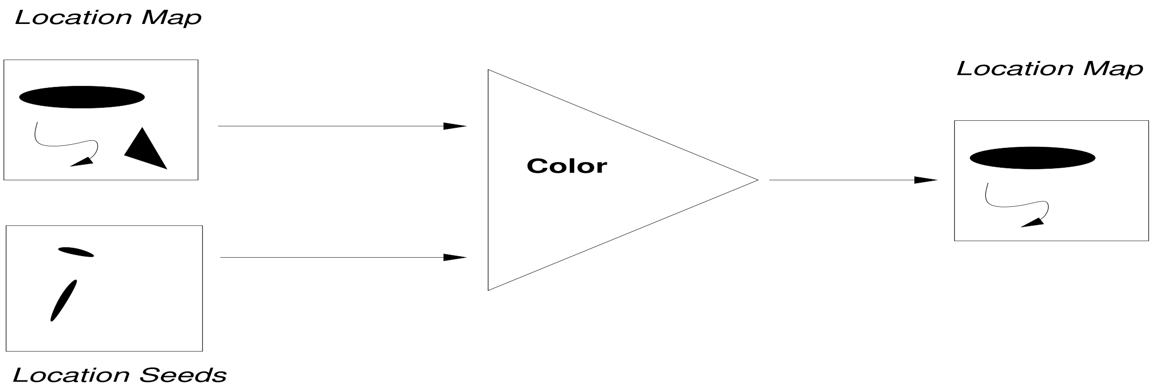

Figure 15: Only regions colored by the seed locations are selected.

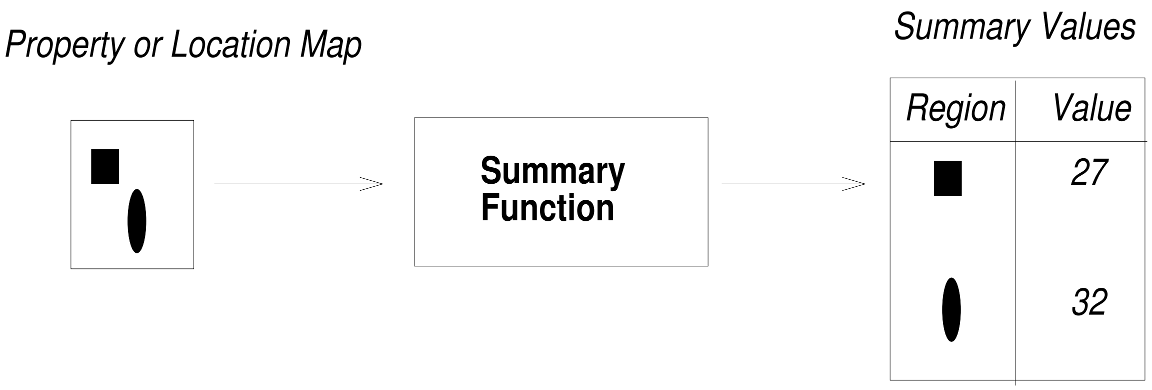

Figure 16: A summarizer function computes the summary over regions and delivers a mapping between region and value

What spatial relations do these operations make explicit?

By using the (Location Map → Property Map) operations in conjunction with the (Property Map → Location Map) operations, a variety of spatial relations can be computed. Intuitively the class of spatial relations computable by the operations described above, is the following. If the computation of a spatial relation is viewed as selecting locations where that spatial relation is true, then the question becomes what kinds of regions can you select with the above operations? By combining the two capabilities of establishing properties and then selecting locations. That is, establish some properties, select some locations, and then establish some properties of the selected locations etc) one can select all regions that are indexable wrt some other set of regions by the properties listed above (area , distance etc) where the other set of regions are themselves indexable wrt some other set of regions and so on, it seems like the operations can compute the transitive closure of indexable regions. Adding more properties improves the ability to select and hence increases the size of this closure.

References

Barlow, H. B. (1985), Cerebral Cortex as a Model Builder, *in* D. Rose & V. G. Dobson, eds, 'Models of the Visual Cortex', John Wiley and Sons Ltd, chapter 4.

Bickerton, D. (1990), *Language and Species*, first edn, The University of Chicago Press.

Bisiach & Luzzatti (1978), 'Unilateral neglect of representational space', *Cortex* **14**, pp129–33.

Brooks, R. A. (1987), Planning Is Just A Way Of Avoiding Figuring Out What To Do Next, Working Paper 303, M. I. T. Artificial Intelligence Laboratory.

Chapman, D. (1990), Vision, Instruction and Action, A. I. Technical Report 1204, M. I. T. Artificial Intelligence Laboratory.

Drescher, G. L. (1991), *Made up Minds - A Constructivist Approach to Artificial Intelligence*, MIT Press.

Farah, M. J. (1989), 'The neural basis of mental imagery', *TINS* **12**(10), pp395–99.

Farah, M. J. (1992), The neurological basis of mental imagery: A componental analysis, *in* Kosslyn & Andersen, eds, 'Frontiers in Cognitive Neuroscience', MIT Press, chapter 44, pp. pp559–73.

Johnson, M. (1987), *The Body in the Mind - The Bodily Basis of Meaning, Imagination, and Reason*, The University of Chicago Press.

Mahoney, J. V. (1992*a*), Image Chunks and their Applications, E. D. L. Memo EDL-92-3, XEROX Palo Alto Research Center.

Mahoney, J. V. (1992*b*), Image Links and their Applications, E. D. L. Memo EDL-92-2, XEROX Palo Alto Research Center.

McAllester, D. (1991), Observations on Cognitive Judgments, A. I. Memo 1340, M. I. T. Artificial Intelligence Laboratory.

Mesulam, M. M. (1981), 'A cortical network for directed attention and unilateral neglect', *Annals of Neurology* **10**, pp309–325.

Moran, J. & Desimone, R. (1985), 'Selective attention gates visual processing in the extrastriate cortex', *Science* **229**, pp782–784.

Pinker, S. (n.d.), The Evolution of Cognition, Notes for an IAP 92 talk at M.I.T.

Rao, S. (1995), Multiple Representational Systems, unpublished.

Spelke, E. S. (1991), Physical Knowledge in Infancy: Reflections on Piaget's Theory, *in* S. Carey & R. Gelman, eds, 'The Epigenesis of Mind: Essays on Biology and Cognition', Lawrence Erlbaum Associates, chapter 5.

Syeda-Mahmood, T. F. (1992), Data and Model-Driven Selection using Color-Regions, A. I. Memo 1270, M. I. T. Artificial Intelligence Laboratory.

Ullman, S. (1983), Visual Routines, A. I. Memo 723, M. I. T. Artificial Intelligence Laboratory.

Zipser, D. & Andersen, R. A. (1988), 'A backpropagation network that simulates response properties of a subset of Posterior Parietal Cortex neurons', *Nature* **331**, pp679–684.