From Imagery to Salience: Locative Expressions in Context

Alicia Abella

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Science

Columbia University
1995.

ABSTRACT

From Imagery to Salience: Locative Expressions in Context

Alicia Abella

This thesis gives a conceptual framework for representing, manipulating, measuring, and communicating ideas about topological (non-metric) spatial locations, object spatial contexts, and user expectations of relationships. This thesis articulates a theory of spatial relations, how they are represented as fuzzy predicates internally, and how they can be appropriately augmented or filtered using prior knowledge in order to produce natural language statements about location and space. This work quantifies the notions of context and vagueness, so that all spatial relations are measurably accurate, provably efficient, and matched to users' expectations.

The system combines variable aspects of computer science and linguistics in such a way so as to be extensible to many environments. The system is demonstrated both in a landmark navigation task and in a medical task, two very separate domains.

Contents

List of Figures

List of Notations

$\mathcal{O}$	set of objects in an image
$a, b, c, d, \ldots$	objects in an image
r, f	reference object and figure object
$\mathcal{P}$	set of prepositions
p_j	a preposition from $\mathcal{P}$
$\mathcal{S}$	set of superlatives
s	superlative from $\mathcal{S}$
$\mathcal{A}$	set of object features
w, h	an object's elongation in the x and y extent, respectively
$\mathcal{U}_p$	a set that contains all those objects that satisfy preposition p
A	an object's area
x, y	an object's center
W, H	an object's enlarged elongation in the x and y extent, respectively
f_{ij}	properties associated with preposition p
σ	the fuzzifying agent
p	conjoined fuzzy predicate
$\bar{p}$	thresholded conjoined fuzzy predicate
l	locative expression
$\bar{l}$	minimal locative expression
$s\bar{l}$	a super-locative expression
$T(p_1, p_2, p_3)$	a ternary relation between prepositions p_1, p_2 and p_3 that is used to perform inference network minimization
π	a set of objects that constitutes a path
$V(\pi)$	the vagueness of a path π

$$\frac{\text{Para mis padres}}{\text{To my parents}}$$

Foreword

Before beginning this exposition, we would like the reader to ponder the following questions: "How do we, as humans, use vision and natural language to convey spatial information? How do we encode all the information that we see? How do we translate this information into a form that we can use to communicate?"

Unfortunately the answers to these questions are not known for certain. Psychologists, linguists, biologists and others have each contributed their answers to these and similar questions. Jackendoff [Jackendoff, 1992] realizes that there is quite a difference in how we represent those things that we see and how we represent where those things are.

> ... what sets humans apart from other species is the ability to use these representations to express our spatial experience: talking about what things are, where they are, and how we might get to them.

He explains how there are so many ways to describe *what* an object is but there are not very many ways to describe *where* an object is. He has established two hypothesis, one that examines this disparity from a language perspective and one from a more cognitive perspective.

Jackendoff's language hypothesis suggests that there is a certain amount of filtering of information when we go from a cognitive representation of a spatial scene to a linguistic representation of what we've observed. It is perhaps this that accounts for many spatial relationships going unexpressed. This is not only true for describing where an object is but also for describing what an object is. We can recognize very complex object contours and textures but we have some difficulty in describing them using linguistic terms. In summary, language does not provide a way of describing all the things that we recognize for either object shape or object location.

We can, however, describe what an object is in complex geometric terms, but when it comes to describing where an object is, we speak in less complex geometric terms. Prepositions, which are used almost exclusively to describe the location of objects, are few in number when compared to other language constructs such as nouns. This is believed by some to be due to organizational facts about the human brain. The brain is believed to consist of submodules, one of which is concerned with object shape identification and one of which is concerned with object location. The expressive power of the submodule for localizing objects is believed to not be as strong as that for object shape identification. A more detailed analysis of this idea from Jackendoff's perspective can be found in [Jackendoff, 1987a]. These two submodules is what Ungerleider and Mishkin in [Ungerleider and Mishkin, 1982] refers to as the *what* system and the *where* system. In this thesis we will see that I will be residing primarily in the *where*. The *what* system is responsible for encoding the object's detailed shape. The *where* system only needs to store a little bit of object description and then it communicates with the *what* system through a link. What Jackendoff postulates is a correlation between the grammar we use and the nonlinguistic part of the brain.

In [Farah *et al.*, 1988] they, describe a case where a person sustained damage to the part of the brain responsible for performing the *what* tasks, and what they noted was that while the person was able to perform object location they could not perform object shape recognition very well. Levin in [Levine *et al.*, 1985] has demonstrated that a similar phenomena can happen in the imagery system. In their experiments a person could image what an object looked like but could not image the spatial relationship of objects. On the other hand they had a person who could image the spatial relationship of objects but could not image their shape.

Now that we have provided a brief look at some of the postulates regarding how we as humans encode and discuss spatial information, we will proceed by giving a brief look at what this thesis postulates.

Chapter 1 introduces the reader to the system by describing what the individual components are and how they interact. We illustrate an example of how path descriptions and density descriptions are generated in the landmark navigation domain and the radiography domain, respectively.

Chapter 2 provides a literature review. It includes a linguistic, psychological and artificial intelligence perspective on the various issues related to this work. The psychological text covers issues revolving around how we as humans treat and talk about the space around us. It also contains some references to how infants learn spatial relations. The linguistic text examines the meaning of prepositions – when are they used, how does their meaning depend on context and the various meanings that prepositions can adopt. The artificial intelligence text discusses some of the efforts being made by others with regards to natural language generation, image processing and robot navigation.

Chapter 3 consists of the semantic representation of the spatial prepositions we use in this thesis. It also discusses a process we call fuzzification, that captures some of the inherent vagueness associated with prepositions. We also define some object properties and superlatives that we use to enhance the description of the objects in the environment.

Chapter 4 describes what a locative expression is, how it is used to represent the spatial relationship of objects in an image and how they are manipulated to yield the simplest yet most informative spatial relation. Involved in this manipulation process is the notion of the vagueness of a locative expression. We will see in this chapter that the vagueness of a locative expression encodes the user decision making process and user error. A method for yielding the simplest locative expression will be explained. The methods of this chapter will be applied to the landmark navigation task.

Chapter 5 describes another method for generating the simplest locative expression. This method is based on an inference network and will be applied to the medical task.

Chapter 6 applies the concepts described in chapters 3 and 4 to the landmark navigation domain. We discuss how the final English sentence is generated.

Chapter 7 applies the concepts described in chapter 5 to the radiography domain. We will describe how the system generates medically sound sentences.

Chapter 8 provides the concluding remarks and makes suggestions for future work.

Acknowledgements

I would like to thank my first mentor, Peter Selfridge, who gave me the opportunity to spend a summer at AT&T Bell Laboratories. That experience convinced me to pursue my graduate studies. I am also greatly appreciative to all the members of the National Physical Science Consortium; without their fellowship support I would not have been able to pursue these studies. I thank Larry Tarbel, my mentor at NSA who showed confidence in my ability to successfully complete my PhD and who was very supportive during my two summers with him. I would like to thank Dr. Justin Starren who provided me with his medical expertise and answered my many novice questions regarding radiography. Special thanks also go to Dr. Jeff Newhouse who provided me with many of the radiographs that went into this work. Thanks go naturally to my advisor, John R. Kender who gave me plenty of room to pursue my own ideas. I also enjoyed and gained a great deal from his depository of "did you know ..." questions.

Special thanks go to my cousin Jorge who is a constant source of inspiration and encouraging words. I am forever thankful to my husband, Aleksandar (my unofficial advisor), who more times than I can remember sat with me while I debugged code, studied for exams, and wrote this thesis. And last, but certainly not least, I want to thank my parents, Blanca and Orlando, without whose love and support these times would have been much more difficult.

Chapter 1

Introduction: The Integration of Image and Natural Language Processing

The integration of image and natural language processing has recently emerged as an area of crucial importance to tasks that involve communicating visual information. Advances in the technology used to gather images and research results in image and natural language processing have made this integration viable. Technology has given us the means of gathering a large amount of data in the form of images, and at the same time it has developed the need to extract what is being conveyed in these images. Image processing has given us the ability to extract what is in these images and natural language processing has given us the ability to convey what is extracted.

Figure 1.1 illustrates the current state of affairs. There is a significant amount of theoretical results and systems built in the areas of image and natural language processing. The area formed by their overlap, however, contains comparably few theoretical results, and even fewer successful implementations that utilize image and natural language. The aim of this dissertation is to create a computational understanding of spatial prepositions that integrates visual and linguistic ideas and to use that understanding to generate descriptions of images from two domains

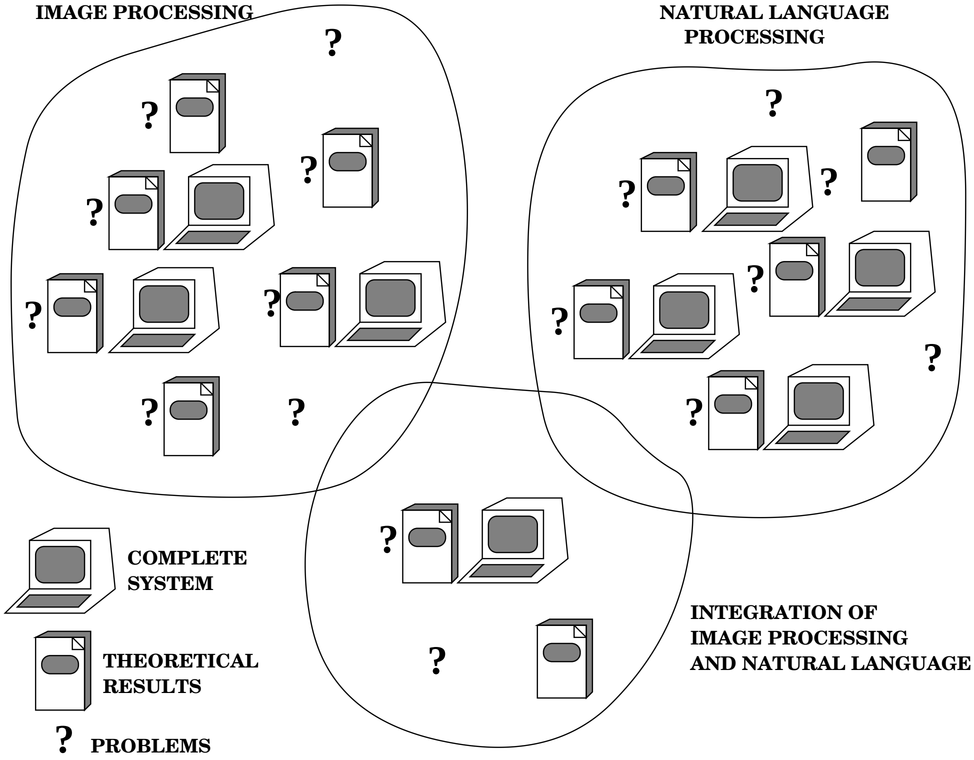


IMAGE PROCESSING
NATURAL LANGUAGE PROCESSING
INTEGRATION OF IMAGE PROCESSING AND NATURAL LANGUAGE
COMPLETE SYSTEM
THEORETICAL RESULTS
? PROBLEMS

Figure 1.1: The current state of affairs in image, natural language processing and their integration

using natural language [Abella and Kender, 1994a]. The two domains are landmark navigation and description generation of abnormal densities from radiographs of the urinary system. The text generated should resemble descriptions normally found in the domains by using knowledge that has been built into the system. The system is constructed from two types of "building blocks" – generic and domain-specific. The following section will describe in brief these "building blocks".

1.1 System Architecture

Figure 1.2 illustrates the system architecture. It consists of two types of modules and data, **generic** and **domain-specific**.

In the figure the generic modules are denoted □, generic data are ☁, the domain-specific modules are ■ and domain-specific data are ●. The

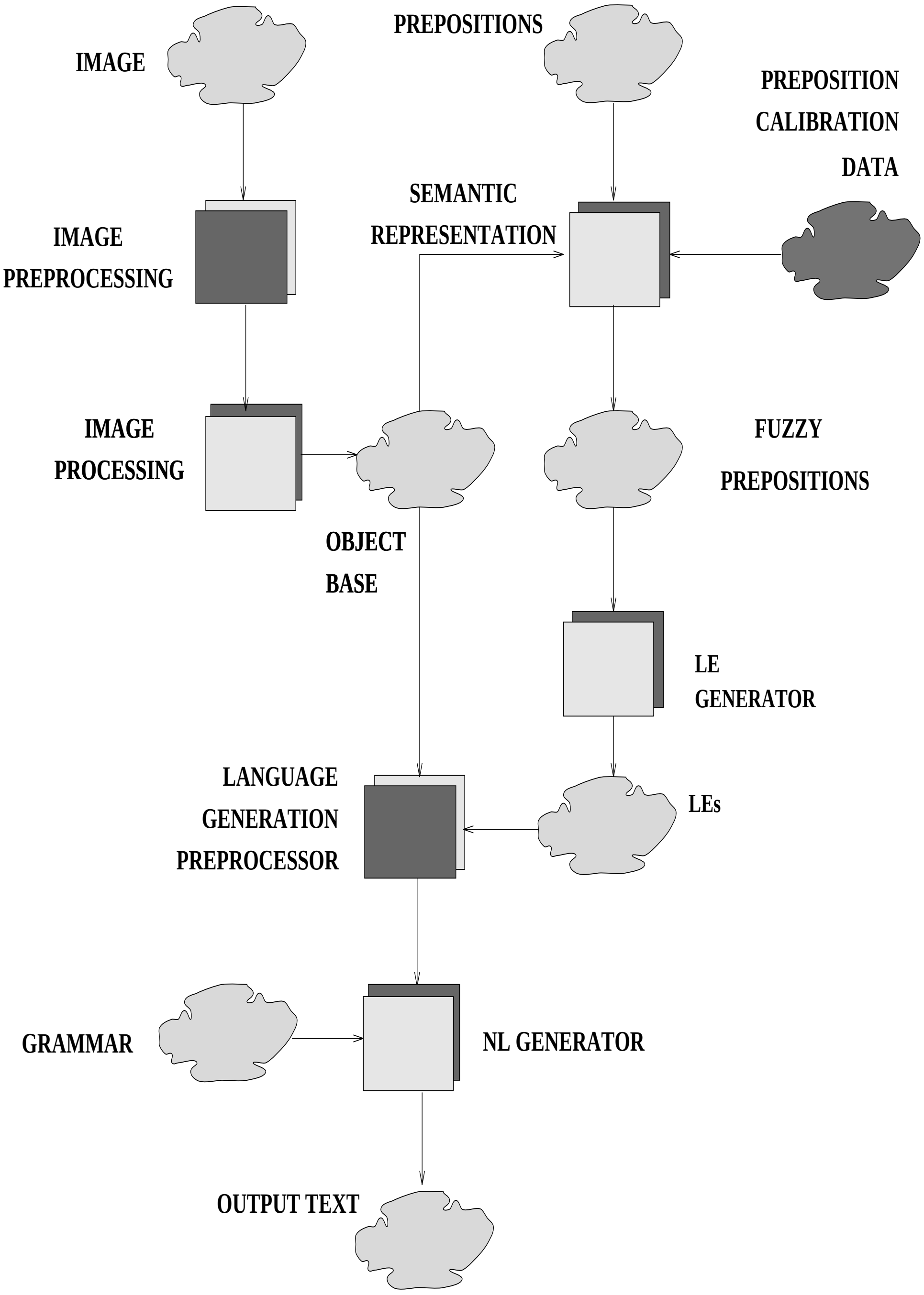


Figure 1.2: System architecture: generic and domain-specific modules and data

```
Input: gray-scale image
Output: object base
        Convert the gray-scale image to a binary image
        Locate objects in the binary image
        Compute each object's area, center, and elongation properties
```

Figure 1.3: The input, output, and steps performed by the image processing module (generic module)

generic modules form the foundation of the system. The majority of the theoretical contributions of this work are incorporated into these modules. System generality is attained through the use of the interchangeable **domain-specific** modules. The **domain-specific** modules contain knowledge about the particular domain; the knowledge necessary to produce meaningful interpretations of the spatial relationship of the object pairs.

The image processing portion of the system takes as input a grey-scale image and locates all the objects and extracts the relevant object properties (see figure 1.3). The properties extracted are an object's area, center, and elongation properties. We base this selection on the fact that in the case of spatial prepositions objects may be treated as *blobs*. We do not need to know what color shirt someone is wearing to determine if he/she is standing *near* a car. Prepositions like *in* or *inside* can regard an object as a blob as long as the blob has the capacity to surround. Likewise, *near* and *at* only require that the blob have some spatial extent. *Along* requires that an object be fairly linear and horizontal with respect to another. An object's size, however, is important. We would not normally say, "The house is *near* the bicycle.", but rather "The bicycle is *near* the house." It is normally the case that a reference object, in this case the house, is larger than the figure object. We will see, however, in the landmark navigation task that the reference object may be smaller than the figure object because the reference object is always the object that a person is looking at. The elongation property gives us a way of representing an object as a blob, and will prove pivotal in the semantic

Input: prepositions, preposition calibration data, object base
Output: fuzzy prepositions
Fuzzify each preposition
Assign to each object pair its representative preposition(s).

Figure 1.4: The input, output, and steps performed by the semantic representation module (generic module)

representation of prepositions to be covered in chapter 3.

The semantic representation module assigns a value to all the prepositions that describe all the object pairs in the image. The value is a measure of how well a preposition describes the spatial relationship of an object pair. The semantic representation of a preposition is an algebraic formula that integrates visual and linguistic information. The visual information is based on the object properties extracted by the image processing module. The linguistic information is based on data collected by asking people to select which prepositions described certain object pairs in an image. The results yielded the preposition calibration data shown in figure 1.2. The preposition calibration data is then fitted to the parameters of the semantic representation. This calibration data allows the system to capture some of the vagueness associated with prepositions by a process we call fuzzification. We commonly use phrases like *somewhat near* that express a vague belief in the spatial relationship of a particular object pair. Fuzzification captures this vagueness and yields fuzzy prepositions.

It is the responsibility of the locative expression generator to limit the number of prepositions ultimately used to describe an object pair. The semantic representation module gives the locative expression generator information about how all the prepositions are satisfied by all the object pairs. The locative expression generator must then make sense of all the information, deciding which prepositions are important and which are not. The locative expression generator then produces

Input: fuzzy prepositions
Output: Minimized locative expression
Minimize the number of prepositions that describe an object pair
while maintaining its descriptiveness

Figure 1.5: Input, output, and steps performed by the locative expression generator (generic module)

Input: object base and locative expressions
Output: semantic description of the text
Use domain specific knowledge to generate appropriate
input to the natural language generator.

Figure 1.6: The input, output, and steps executed by the language generation preprocessor (domain-specific module)

a locative expression that represents the minimal number of prepositions necessary to describe the spatial relationship of an object pair. After all a locative expression of the form "*near* and not *far* and *aligned* and not *above* and *below* and not *inside* and *next*" is not all that descriptive.

The language generation preprocessor is responsible for creating the input that the language generation module uses to generate the final English sentence. The input is a semantic description of the text to be generated. The text should convey the meaning appropriate for each particular domain. To insure this the language generation preprocessor uses knowledge that has been built into it in order to generate a meaningful semantic description. The language generation preprocessor can be looked upon as an embryo of a rule-based system. Figure 1.6 illustrates this module.

The input created by the language generation preprocessor is then supplied to the language generator along with a grammar for generating coherent sentences. The language generator and the grammar were developed by the Natural Language

Input: grammar, semantic description of text
Output: English sentence
Generate a syntactic description of the text from the input and grammar.
Interpret the syntactic description and produce an English sentence

Figure 1.7: The input, output, and steps executed by the language generator (generic module)

processing group at Columbia University [Elhadad, 1993]. The grammar is very extensive and capable of handling many types of sentences. The language generator produces the final description of the spatial relationship of the object pair in the form of an English sentence (see figure 1.7).

1.2 Landmark Navigation

Task: Given a reference object r and a figure object f generate a path description.

- *The spatial relationship between objects are expressed using spatial prepositions and superlatives. Intrinsic object properties, such as size, are sometimes chosen if they aid in generating an unambiguous description.*

- *In order to describe one object relative to another we group prepositions into locative expressions.*

- *In order to get to the desired figure object we may need to go through a set of intermediate objects.*

The goal of the landmark navigation task is to find the "best" path description for getting from a start location to a goal location. The path may not simply go from the start location to the goal location, it may need to traverse several intermediate objects along the way. The decision as to whether or not to go to an intermediate object is made by assigning to each object pair description a value that denotes

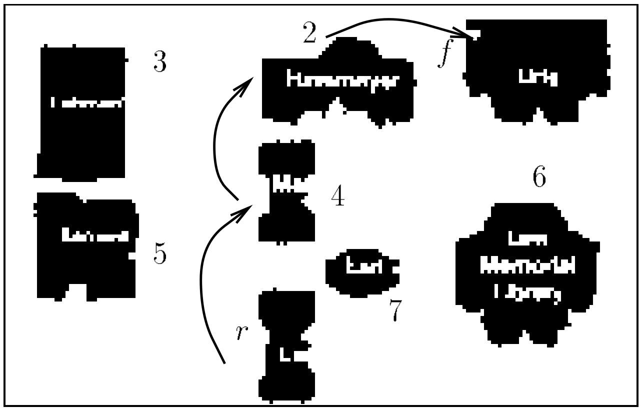


Figure 1.8: Partial campus map of Columbia University. The arrows indicate the path that the system described for the indicated r and f.

how good that description is. The goodness of a description is measured in terms of its ability to describe the spatial relationship of an object pair as unambiguously as possible. For example, if r is reference object in figure 1.8 and f is the figure object and the description given is "Locate the building to your right" there is a 33% chance that the person will find the figure object, since the building labeled 6 and building 7 are also to the right of r. In essence the person can get lost and arrive at an incorrect destination. In this example there is no clear way to describe the location of f with respect to r, therefore the system chooses to first describe the object labeled 4 and then the object labeled 2 and finally the figure object. The descriptions for object pair (4,r), (2,4), and (f,2) were better than the description for (f, r), because there was a smaller chance of a person getting lost if he/she went through the intermediate objects. The sentences generated by the system were:

(4,r) `First locate the Math building which is the nearest one above you.`

(2,4) `Then find Havemeyer which is the one above you.`

(f,2) `Uris is the one on your right.`

An interesting open issue in this task is the generation of "safe" paths. A "safe" path is one that leads to the goal with high probability regardless of the fact that

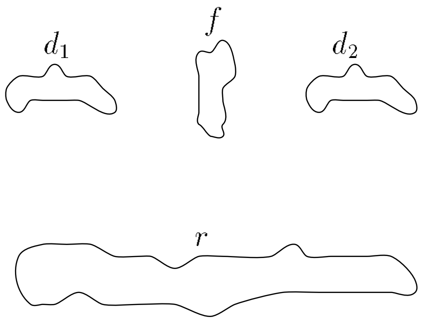


Figure 1.9: Incorrect intermediate object: The sentence "Locate the object that is aligned with you. Then find the object nearest you" may cause a user to choose either d_1 or d_2 but will, in the end, cause him/her to find the correct figure object

some of the intermediate objects along the way may have had a low probability of being recognized. An example is illustrated in figure 1.9. A feasible description of a path between f and r is "Go to the object that is aligned with you." Then go to the object nearest you." While the intermediate object may be either d_1 or d_2, the final destination in both cases is f.

1.3 Description Generation of Abnormal Densities from Radiographs

Task: Given a radiograph generate a medically sound description of where a kidney stone is located.

- *Locate the stone using the domain-specific image preprocessor.*
- *Use the semantic representation module to compute the spatial relation between the stone and the objects in a model of the urinary system.*
- *Create the appropriate semantic input based on the spatial relationship of the stone and the model parts using the language generation preprocessor.*
- *Produce the final description of the stone by combining the results produced by the language preprocessor and the domain-specific grammar.*

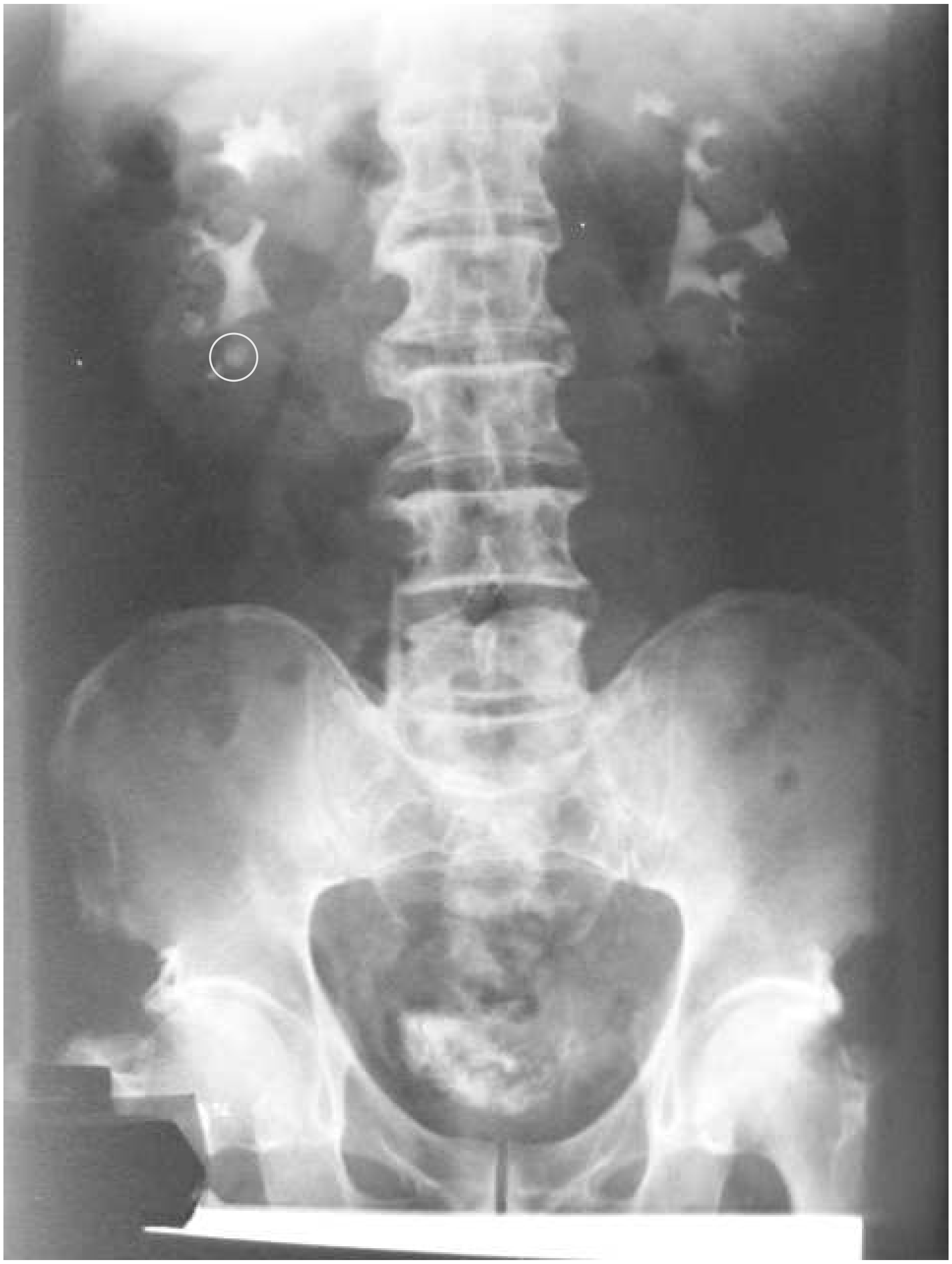


Figure 1.10: An x-ray image of the urinary system depicting a kidney stone in the right kidney (note that the right kidney appears to be on the left in the image). The stone is highlighted by the circle.

The goal of this task is to describe the location of kidney stones from X-ray images of the urinary system.[1] The system uses domain-specific knowledge in generating descriptions that most closely resemble descriptions that radiologists would generate. Radiologists describe the location of kidney stones by expressing their location with respect to an organ, a part of an organ, or the nearest bony structure, like the spinal cord. Figure 1.10 shows an X-ray image of the urinary system with a stone located in the right kidney. In this example, the actual pathology report contained the following description:
A 1 cm calyceal stone is seen in the lower pole of the right kidney.
The system generated the following description:
A calyceal stone was found in the lower pole of the right kidney.

This task is slightly different from the landmark navigation task because now the location of a stone is not simply the spatial relationship of the stone to a single reference object, but rather it may consist of the spatial relationship of the stone to several reference objects. For example, to categorize a stone as being calyceal it must be *inside* the kidney and *near* a calyx, which appear as white globular entities in the kidney in figure 1.10. The task of labeling the stone a calyceal stone is performed by the language generation preprocessor. It uses information provided to it by the locative expression generator and rules about the meaning of certain locative expressions in this context. For example in the case where the stone is *inside* the kidney and *near* the calyx the language generation preprocessor yields the semantic description that will cause the natural language processor to generate a sentence that will describe the stone as a *calyceal stone*.

1.4 Contributions

The contributions of this work are the following:

[1] This work is done in collaboration with Columbia Presbyterian Hospital's Radiology Department.

- This thesis presents a system that takes images and generates English sentences that describe the location of objects found in the image.

- The system captures spatial "understanding" via a semantic representation of spatial prepositions. This representation makes quantitative the qualitative nature of the prepositions.

- We also defined a semantic representation of intrinsic object properties. These include an object's size.

- To further enhance the description of an object's location we also defined a semantic representation of superlatives.

- We define a fuzzification technique to handle the inherent vagueness associated with prepositions. The fuzzification technique introduces flexibility to the semantic representation.

- We developed two methods for optimizing the description of an object's location. These optimization methods are necessary if we want to generate coherent descriptions.

- To assist in the selection of coherent descriptions we defined a model of user behavior and user inexactness. The system uses this model to select those descriptions that are concise yet precise.

- Based on this error model we define a measure that reveals the vagueness of a description. Little vagueness means that a user will select the intended object.

- The optimization techniques, besides finding an optimal description, also handles context. In selecting an appropriate description the optimization techniques examine how objects in the environment interact.

- The optimization techniques can be applied to three types of users. One of the optimization techniques generates descriptions for the novice user. The novice user is one that is not familiar with the environment depicted in the image. This optimization technique is based on Boolean formula minimization.

- The second optimization technique generates descriptions for the expert user. The expert is one that is very familiar with the environment depicted in the image. This technique is based on an inference network.

- Combining these two techniques we satisfy the third type of user. The third user is one that may be familiar with parts of his/her environment but not all.

- The definitions and methods aforementioned are applied to the landmark navigation task. This task consists of identifying objects in maps of theme parks. It may require the description of several objects before being able to describe the intended object.

- Also presented in this thesis is the description generation of abnormal densities found in radiographs of the urinary system.

Chapter 2

Literature Review

The underlying motivation behind this work is the integration of computer vision and natural language. There has been a substantial amount of research in both of these areas but a bridge between the two is far from being completed. Psychologists, linguists, and AI researchers have all performed their respective research on the topics of spatial reasoning, prepositional meaning, and the knowledge representation required for describing the spatial arrangement of objects in a scene. The work presented in this thesis will use results from each of these areas to create a system that will meet some of the large demands imposed by all three research areas.

2.1 A Psychological Perspective

Psychologists seem to be interested in how we simplify spatial information, how we represent it, how we manipulate it, and how we communicate it. [Miller and Johnson-Laird, 1976] gives some examples of how people communicate it, and how it differs among different languages. As for its representation, [Kosslyn, 1980] and [Pinker, 1984] speak of our long-term memory as containing a descriptive encoding of visual information and a working memory as containing a depictive representation of visual information.

The work found in [Rueckl *et al.*, 1989] defines a computational approach to the question of why there seems to exist separate cortical pathways by which we perform object identification and object localization. In [Rueckl *et al.*, 1989] they constructed a connectionist model to explore the "two-system" design of "what" and "where". The investigation was to measure the computational demands on a single system that performed both object identification and object localization versus a two-system that would process information separately. Their results showed that a two-system is more efficient than a single system. The single system would use much more of the available information about an object to represent its location than was really necessary. These results are consistent with the hypothesis that the visual system encodes object identity and location via a two-system and the reason it does so could be that it is computationally superior to do so.

Psychologists are not just interested in how we observe and talk about our physical surroundings, they are also concerned about our mental images, and how we use them, manipulate them, and "look at" them. The book [Kosslyn, 1980] gives a fairly extensive look at imagery. The big debate in the study of imagery is in whether or not we actually think and represent ideas using mental images. This controversy in known as the imageless thought controversy. In it, some experimental psychologists claim that some people do not experience mental images when solving a problem, while other psychologists claim that people do use mental images when forming ideas. An interesting experiment related to the discussion in the previous few paragraphs, is an experiment conducted by Lea found in [Lea, 1975]. In this experiment, he presented subjects with a scene that contained different objects all arranged in a circle. He then asked the subjects to memorize the scene and to associate specified words with the location of each object in the scene. Subsequently he asked them questions regarding the name of the objects and the location of the objects. In the first part of the experiment the experimenter asked to name the object n locations away from a given starting point. In the second part of the experiment, the experimenter asked for the name of the object at a given location. What was discovered was that it took longer for a person to give the name of an object associated with a location than it was to just name the location. Perhaps the time difference in naming the object as opposed to just naming the location can be attributed to the link that must be traversed in going from the “what” system to the “where” system.

In the book, [Miller and Johnson-Laird, 1976], we can find an extensive look at language and perception. In it they note how we have tried to abstract a formal concept of space from our experiences of it, and this has led to two theories, relative and absolute. Relative space is defined by the spatial relations among things. Absolute space is defined by a coordinate system independent of anything that space might contain. Miller and Johnson assert that the simplest spatial attribute of an object is its place and he quotes Aristotle’s definition of place as

1. place is what contains that of which it is the place
2. place is not part of the thing
3. the immediate place of a thing is neither less nor greater than the thing
4. place can be left behind by the thing and is separable

In summary, what Aristotle was suggesting is that place is a container coincident with what it contains. His definition does not say how a particular place is to be identified. Clarifications were made later by other philosophers. In order to take account of spatial relations our perceptual processes must register place, and relations between place. To perceive a place means perceiving a spatial region that contains that space. The region is like a shadow cast around an object and object regions may overlap. In this thesis we consider this region of space as a bounding box, and it will be used much the same way we use it to determine, for example, if two objects are near each other. If the object regions overlap then the objects are considered near. Or equivalently if their bounding boxes overlap then the two objects are near.

Miller and Johnson, stipulate that our concept of space and how objects in it are related to each other depends on our motility in it. As Urban in [Urban, 1939] notes

> our intellect is primarily fitted to deal with space and moves most easily in this medium. Thus language itself becomes spatialized, and in so far as reality is represented by language, reality tends to be spatialized.

Languages of the world differ widely in their treatment of space. Miller and Johnson note that architects have probably collected more data about the ways that people talk about space than psychologists and linguists. One such person is Lynch in [Lynch, 1960]. Lynch developed a spatial and geographical representation for what people told him about cities that they knew. Lynch noted that what he needed to represent were landmarks, paths, edges, districts, and nodes. As Miller and Johnson point out the purpose of a locative expression is to narrow the domain of search for an object, such as a landmark. Locative expressions are useful because they can be used for object identification since we use them to narrow the search domain to a point where only one object of the type described is found. This point will become very relevant when it comes time, in this research, to show the implementation of the system capable of identifying objects using locative expressions.

2.1.1 Spatial Understanding in Infants

There have been quite a bit of experiments conducted on infants and young children to investigate the beginnings and evolution of spatial understanding. The experiments vary from conducting experiments on days old infants to test how much spatial understanding they possess, to testing how children relate words to the spatial relationship of objects in their environment. The following is a glimpse at some of the experiments and beliefs behind how infants and children relate to their world spatially.

Landau and Stecket in [Landau and Stecker, 1990] performed experiments on young children to investigate how the geometric properties that they attach to an object differ from that of an object's location. In these experiments the children were introduced to a new noun by being told "This is a corp" and a corp referred to a novel object. That object was then placed on a certain location on the box and they were told "This is acrop my box". The first statement refers to the noun condition and the second statement to a preposition condition. New objects and different locations were then introduced and they were asked which of the new objects were instances of corp and which new locations were acrop. The results suggested that for the prepositions case the children either completely disregarded the shape of the object or preserved just the object's principal linear axis and accepted positions preserving the object's orientation relative to the box. In the noun condition the children were much more particular when it came to preserving the objects shape. Only objects that were similar in shape to the one that they had been told was a corp was actually recognized as one regardless of the position on the box.

The main interest in these experiments was in gaining some insights into how children come to know what geometric properties are relevant for the different expression - noun or preposition. The question raised by Landau and Stecker is how and when do children learn to generalize objects to points, volumes and boundaries when using prepositions.

Clark in [Clark, 1977] performed various experiments on infants to test their understanding of the spatial relation conveyed by certain prepositions. When performing these experiments Clark keeps in mind that children are aware of objects either possessing a flat supporting surface or an interior that cause them to serve as containers. Children seem to possess a certain fascination for containers. If a child of about two is given an overturned glass and told to place another object on the glass, in most cases the child turned the glass over and placed the object inside the glass. At about twelve months infants can place small objects on top of larger ones, and know that they can not do the reverse. The child also knows about normal orientation - which way is up for instance. The first prepositions that children seem to use is *in* and *on*, between the ages of two, and two and a half months. Clark performed experiments using the preposition pairs, up-down, over-under, top-bottom, and above-below. What Clark found was that of all the pairs up-down was the easiest for the children to grasp. Although they seem to use up-down, as well as over-under, to indicate direction rather than position in space. The next easiest pair was top-bottom, over-under, and then above-below. In summary, Clark believes that the child's nonlinguistic organizational preferences appear to play an important role in predetermining what words are more easily acquired by the child.

We find that in [J.Piaget and Inhelder, 1956] the analysis of children's development of spatial representations is a topological one. In it young children begin the task of representing space in topological terms, then advance to an understanding of Euclidean and projective information and then finally rely on the use of coordinate axes and metric information. Mandler in, [Mandler, 1988], on the other hand, feels that there are two alternatives to this hypothesis that are more appropriate for explaining a child's development of spatial representations. The first is that we are sensitive to topological and Euclidean information from birth but switch to reliance on Euclidean information as we get older. The second is that we rely primarily on topological information at all ages. He tends to support the second of these hypothesis more strongly. There seems to be more evidence that we as adults use little metric information when performing spatial thought. In other words, a number of non-Euclidean properties seem to be the most essential aspect of spatial knowledge at all ages. As already mentioned, we encode objects as inside/near/above/etc regions, landmarks, or other objects, without processing metric information because it is simply not needed and as Mandler states "... when we do not attend, we do not represent".

There has been a considerable amount of experiments conducted to try and answer the question of how infants code changes in the position of objects in an environment. It seems that there are two schools of thought aimed at answering this question. One asserts that infants relate to objects in their environment in an egocentric manner (meaning that they view themselves in relation to an object rather than recognizing the interaction between objects). The other school believes that infants can indeed relate objects to each other spatially. In [Wishart and Bower, 1982] they advocate the egocentric school of thought. They conclude that infants have an enormous difficulty in understanding the spatial relations of objects up until at least two years of age. Wishart and Bower feel that spatial understanding matures when an infant learns to overcome the tendency to use egocentric rather than geographic referents for spatial position.

In the paper [Bremner, 1978], the belief that infants use visual cues to code object positions, like the color of the background that the object is placed on, is addressed. In work conducted in [Acredolo, 1978], however, it was shown that infants under eleven months of age make no use whatever of visual referents to position. Experiments found in [Acredolo, 1978] support the claim that infants between six and eleven months rely on past accommodations to an object rather than its relationship to other objects. Other experiments have shown that if an object previously identified is moved, the infant will start searching for it where it last saw it. This leads to experimenters again to believe in egocentrism. Understanding spatial relations is a problem for a young infant in that it is difficult for them to understand the interrelations of positions in space.

In [Keating and McKenzie, 1986] the importance of landmarks for object localization by infants is investigated. The general consensus among those who have conducted experiments in this area, is that infants up to about a year in age use geometrical landmarks to localize objects after a change in position or orientation. [Acredolo and Evans, 1980] showed that the more prominent the landmarks the more effective they were as references in searching for a target object. Without a landmark the infants' behavior is primarily egocentric, localized with reference to the subject's own body. It is not until about fourteen to eighteen months that localization occurs without environmental landmarks [Rieser and Heiman, 1982].

Like the psychologists studying spatial reasoning, this study will also be concerned with how information about an environment is simplified, represented, manipulated and communicated. Now that a feeling has been gained for the psychological aspects involved in spatial reasoning, the next area of interest is in the linguistic aspects of spatial reasoning, or more specifically communication of the results of our spatial reasoning.

2.2 A Linguistic Perspective

Language imposes a strict structure on how we describe virtually every scene we observe. It has been the work of linguists, who have studied how language structures space, to detail the geometric and dimensional aspects of language. This section will be an overview on various research works in this area.

> A fundamental character of the way that space is represented at language's fine-structural level is that it is schematic. That is only particular selections of all aspects present in spatial scenes are actually referred to by linguistic elements, while all the other aspects are disregarded.

This "schematization" that Talmy, in [Talmy, 1983], refers to plays a crucial role in linguistic space descriptions. He also discusses the cognitive side of "schematization" in communication. This cognitive side refers to what a person conveying information says and how he/she specifies it and how the listener envisions what is being described. Talmy also suggests that prepositions, although limited in number, are perhaps the framework by which other linguistic elements are structured.

A scene can not be represented by a fundamental preposition, since a scene can often be very complex. Instead the scene must be partitioned. The first focus of attention, or partitioning, is around that object that we wish to describe. Then we examine a second partition or possibly more before describing what we want. Then to finally describe it we need to know the object's spatial disposition. By disposition Talmy means the object's location and orientation when stationary and its path and orientation when moving.

The spatial disposition of an object is chosen in terms of another object or group of objects. Using Talmy's definition, I will call the object of attention the figure and the other object or group of objects the reference object. In general spatial prepositions require that more geometric information be known regarding the reference object than the figure. This would agree with our cognitive process of localizing an object by considering it as a point while dividing the space around it into regions. For example, in *"The house on the cliff"* the cliff defines the space that the house is occupying and is thus considered the reference object. The preposition *across* is an example of an exception to the rule because we need to know the figure's location and orientation with respect to the reference object. For example, *"Jim laid across the bed"*, requires that we know Jim's location as well as his length perpendicular to the width of the bed. Another example is *over*, where we attribute a planar geometry to the figure with regards to the reference. For example, *"The tarp lay over the infield"*, suggests the tarp needs to be touching the whole infield, and thus must assume the shape of the infield. There is also the case when the reference object is considered a point as in *"The picture frame next to the lamp"*, where the lamp is the reference object. Talmy indicates that reference objects are larger and stationary and that spatial relations are non-symmetrical.

As an example, he uses *"The bike is near the house"*. We would not normally say, *"The house is near the bike"*. We would also not say, *"The bike and house are near each other"*. An elaboration on the geometric aspects of the prepositions used in this thesis will be given in the next subsection.

It is interesting to note how little detail we need to know about the figure or reference object when using most prepositions. For example, to answer the question *"Is Jim standing near Mary?"* we do not need to know what color shirt Jim is wearing to make our spatial judgment. Likewise to use the preposition *in* we need only know that the reference object has an interior, as in *"The fruit in the bowl"*. Objects are not heavily characterized by metric features, but rather by more qualitative and topological features such as boundary conditions (where objects touch), symmetry, and number of relevant dimensions. There are of course some expressions that require that we know part of a reference object in order to localize an object. This Talmy calls, biased geometry, and they include such terms as *in front of*, *on top of*, *on the left*, etc. Our bodies and the earth serve as biased geometry. The earth provides for the use of *north*, *south*, *east* and *west*. Our bodies provide for *in front of*, *in back of*, etc. We also find that prepositions convey temporal information, not just spatial information. For example, *"I knew that all along"*.

One last major component regarding Talmy's work is what he calls, the four imaging systems. These four systems in language attempt to capture the different kinds of relationships among objects within space or time. The temporal dimension is often integrated with the spatial domain and this yields complications. The following will be a summary of the topic. The first imaging system specifies geometries and the geometric relationship of objects to each other within different reference frames. Talmy illustrates various examples. One such example is *"The ball rolled across the rug/through the tube in 10 seconds"*. This examples indicates a point move in a bounded extent, in a bounded extent of time. The second imaging system specifies a "perspective point", when one focuses our "mental eyes" to characterize the location and distance of an object. For example, *"There is a house every now and then through the valley"*. The third imaging system specifies a particular "distribution of attention" to be given to a scene from a "perspective point". For example, *"There are freckles on the boy's face"*. The fourth imaging system indicates "force dynamics" - how objects behave due to outside forces, resistance and barriers, or lack thereof. For example, *"The ball kept rolling along the green"*. This expression makes us think that there is no obstruction and there is some force acting on the ball causing it to move.

In [Herskovits, 1986] her primary inquiry involves what she calls decoding and encoding. Decoding asks the question of whether a locative expression will convey its meaning predictably. Encoding asks the question of whether given two objects can an appropriate expression be found to express their spatial relation. These two questions are also covered in this thesis and a model for answering these questions is demonstrated.

In [Herskovits, 1986] there are four components associated with decoding and encoding They are : use types, geometric description function, tolerance shift and sense shift. A use type corresponds to various classes of uses, for a preposition, distinguished by different conventions. The geometric description function represents how we view a particular object as a geometric entity (e.g. a tree as a point) and as an indirect reference to a part of an object, or an adjacent region of space, or a projection. The tolerance is a measure of how far from the ideal meaning an expression is allowed. For example in saying *"The book is on the table*", even though there is a tablecloth between them, is an example of a tolerance shift. A sense shift tries to capture a discontinuous shift to another close relation. For example, the ideal meaning of *on* in *"The freckles on his face"* is not true. The ideal notion related to the spatial domain is that the world is made up of points, lines, surfaces, and of definitive relations of inclusion, contact, and intersection, to name a few. Of course since we know that this is not the case in the real world some shift from this ideal must be allowed.

The process that Herskovits performs for decoding begins with assuming a normal state of affairs, and if this is not the case, then to create a normal situation type by exploiting context. The normal situation type requires that conditions for the use of an expression be “normal”. By normal she means that objects obey the laws of physics, that they are where they belong, and that they interact normally with each other. The problem then becomes one of choosing a use type, a geometric description, and a tolerance consistent with what is known about the normal behavior of the objects talked about. We see use types, geometric description and tolerance shifts in Herskovits' encoding scheme. She believes that it is not enough to just satisfy a predicate in determining the applicability of a preposition. The problem here is in finding all prepositions that have one or several use types whose constraints are satisfied in the situation at hand. The constraints consist of constraints on the scene, such as allowed spatial relations between objects, and constraints on context. Herskovits breaks down the encoding process as one analogous to the decoding process. As with decoding Herskovits utilizes use types, and object knowledge to hypothesize geometric descriptions and tolerance.

In summary, Herskovits has attempted to account for some of what she calls the unruliness of language through the use of ideal meanings and shifts, use types, and interpretations that do not fit into any use type. She realizes the difficulty that context brings to the process of decoding and encoding. She also observes that some of this unruliness disappears if we accept the notion that we may approximate object shape and the environment in which they occur when using spatial expressions.

Other Work

There are various other studies related in some form or other to English prepositions and the following few paragraphs will touch upon some of them. [Bennett, 1968] discusses the question of synonymy within the framework of a stratificational theory of language taking as examples prepositions. A stratificational view of language is seen as having two ends: a meaning end and a sound end, and at each end there are a set of elements making up the meaning component and a set making up the sound component. Bennett outlines when one preposition may be substituted for another without changing the meaning of the sentence. The meaning that must not change is the cognitive meaning, for example *because of* and *on account of* have the same cognitive meaning although they differ in stylistic meaning. Bennett studies the synonymous nature of prepositions by looking at language from this stratificational point of view.

Another piece of work by Bennett, [Bennett, 1972], is concerned with what Bennett terms locative-directional distinction. Bennett presents a four-way classification of spatial expressions according to the complexity of their internal structure. Bennett refers to them as:

1. locative expressions
2. directional expressions
3. directional-locative expressions
4. directional-locative-directional expressions

The locative expression is the easiest, it just requires that the place of interest be an object or a part of an object, such as its interior. e.g. *"He is at the grocery store"* or *"He is in the grocery store"*. As an example of a directional expression we may say *"Jim ran behind the desk"*. The third case specifies a location by indicating the path that could be taken to reach it. e.g. *"Harry lives over the hill"*. An example of the fourth case is *"A car appeared from over the hill"*. The directional expression is *from over the hill* with *over the hill* being a directional-locative expression, hence the categorization directional-locative-directional.

Yet another study in prepositions by [Becker and Arms, 1969] argues that prepositions share many features with verbs and may be represented at an abstract level of grammar as predicates. Also related to the study of prepositions is the study of preposition stranding. Takami in [ichi Takami, 1992] discusses preposition stranding and gives as some examples, *"Who do you talk about the problem with?"*, and *"Last night was difficult to sleep through"*. The prepositions *with* and *through* are the stranded prepositions. It is generally believed that stranded prepositions are colloquial in style. Takami believes that preposition stranding is a phenomenon that happens primarily in discourse and that it is not a syntactic one.

These few examples already bring to light the difficulties of describing spatial relations. Language can be quite ambiguous and to chose an appropriate description for any particular spatial relation can be quite challenging, if not at times impossible. Jackendoff in [Jackendoff, 1992] cites several primary complications associated with using English prepositions, all of which are acknowledged in the previous discussion about this research area. They involve

1. how ambiguous spatial relations are forced to be expressed with our limited language

2. how the meaning of certain prepositions extend beyond the spatial realm into the time and possession realm

3. how prepositions are used as grammatical markers, such as *"I need to finish this report by June."*

How then can we ever develop a system that is to perform such spatial reasoning tasks, if we ourselves have difficulties? Fortunately, the language of spatial prepositions does afford us with a little luxury - the luxury of being able to disregard a large portion of detail from a scene that might otherwise place us in a situation for which we can not describe an object's spatial relation. The dangers here of course are that we abstract away too much information and we are left in the same indescribable bind. The hopes in this thesis is to develop a system that can ultimately interact with a user, so that if it does find itself in a bind we can supply it with additional information in the hopes of clarifying the situation. The difficulty lies in the simple fact that language is quite informal and as the builders of a computer system, meant to perform some level of language understanding and generation, we are required to convert the informal into the formal.

2.2.1 Preposition Overview

This section is an exposition in the various uses of the prepositions used in this thesis. The prepositions are explained in some of their most common contexts. The information and example are taken primarily from [Lindvist, 1976].

IN

In describing the preposition *in* Lindvist defines it as involving the idea of interiority and inclusion, especially when a static location is referred to. The idea of interiority refers to the fact that the locality of an object is in an enclosed space or specified area which contains or admits the object. An *in-phrase* according to Lindvist refers to an object that is either at rest or in motion and contained in a body or area.

The first examples of the use of *in* refers to its occurrence being effected by a three-dimensional body. In this case an object is situated somewhere in the interior of a three-dimensional body, where it may be at rest, or in motion or action. The following two examples are generalizations of this usage:

- The teapot is in the cupboard.
- The car is in the garage.

We also find uses of *in* when referring to buildings, parts of buildings, vehicles, resting-places, clothes, the body, atmospheric or physico-geographical phenomena, and collectives:

- The labs in the new building are bigger than in the old building.
- There are two fireplaces in the living room.
- The airplane, in which we traveled from New York to San Diego, was a 747.
- He was sick in bed all week. [1]
- I'm going to the dance in my blue suede shoes.
- I got a tickle in my throat just as I was about to speak.
- Battle was fought on land and in the air.
- The moon shone brightly in the sky.
- I saw him in the crowd.

Another example of the use of *in* is as an enclosure effected by a surface, expanse, or area. In general this means that an object is situated somewhere within the limits of a thing with a two-dimensional extension, such as a surface, or area, where it may be at rest, or in motion or action. An example of the general

[1] This last example is a bit misleading because when used in this expression *in* does not mean that we are enclosed in the bed, but rather we are lying *on* the bed, yet we make use of the preposition *in*.

usage in this category would be *There is a beautiful tree in the field.* We also find uses of *in* when referring to the surface of a physico-geographical expanse of a certain kind, a country, an open space or area such as a marketplace or square, a road or path etc. for example:

- Have you ever seen the face of a man in the moon?
- Have you ever been in Italy?
- While attending NYU I would often have lunch in Washington Square Park.
- There was a fork in the road.

NEAR and NEXT

The prepositions *near* and *next* will be dealt with together since they both share the idea of general proximity. *Near* and *next* occur in expressions which denote proximity in that an object is not far from another object or group of objects but rather it is situated either in its neighborhood or adjacent to it. In the case for *near* Lindvist realizes two distinct usages:

1. "in the vicinity of", a usage occurring when a phrase refers to a position of geographical features and phenomena in a neighboring area
2. "close to" when the phrase is used to describe immediacy in space

More specifically *near* may be used to refer to something that is to be found in a neighboring area, and a state or instance of closeness or approach as in something that is placed within reach, for example,

- The house is near a lake.
- We sat near the fire to keep warm.

Next to occurs in cases of proximity when an object holds a position of immediacy in space, either at the side of another object, or with reference to a succession. e.g. *"Her car was parked next to mine."*, and *"She stood at the end of the line with Jim standing next to her.* In chapter 3 we define *next* and *near* to have the aforementioned meaning since we can capture this meaning geometrically. Lindvist is clear to point out that *next* is not limited to indication of positions referring to rows only. e.g. *"She held the baby next to her breast"*. *Next* can also refer to motion as in *"The man passed next to the desk"*.

ABOVE and BELOW

What remains to be examined is *above* and *below*. *Above* occurs in cases when one object becomes situated higher in the vertical plane than another object. For *below* this definition is reversed - one object becomes situated lower in the vertical plane than another object. There exist more specific cases of the usage of *above* and *below* for example, when two objects are *above* or *below* each other in the same vertical line, when two objects are *above* or *below* each other and they are not in the same vertical line, and when a certain amount of vertical distance is specified. e.g. *"The traffic light hung above a manhole."*, *"The desk stood below the blackboard."*, *"The painting hangs above the mantlepiece"*, *"The temperature may go above* 100° *F."* *"The crew were below deck."* In chapter 3 we define two meanings of *above* and *below* the one where objects are situated higher/lower in the same vertical plane and the one where two objects are not in the same vertical plane.

Above and *below* may also be used in cases when one object is situated in relation to another in the horizontal plane. When using *below* it is implied that one object differs from another by being, as Lindvist puts it, further down a slope, nearer the sea, further south, nearer the front of a stage, etc. *Above* is the reverse of what was just defined for *below*. e.g. *"The ship was sighted 90 nautical miles above the Cape Verde Islands."*, *"I bought several acres of land below Mrs. Jones'* *property."*

More references on the topic of spatial language, including spatial prepositions, linguistic theory and semantic descriptions of English can be found in [Retz-Schmidt, Summer 1988], [Leech, 1969], [Fillmore, 1968] and [Kautz, 1985].

2.3 An AI Perspective

AI - the study of how to make computers behave intelligently. AI spans many areas that we consider to require a considerable amount of intelligence, such as language understanding and generation, perception, navigation, and planning, just to name a few. The inundation of papers in these areas is proof that there is a desire to not only better understand the underlying principles of intelligence, but to try and exhibit them through the use of computers. This inundation is further proof that we have not quite reached our goal and the need to learn more is ongoing.

An early attempt at defining a model for language understanding and developing a system to test it is illustrated in [Winograd, 1972] and [Winograd, 1973]. His focus is on the ability to write a computer program that can perform some semblance of natural language understanding and that in the process it may help clarify what language is, how we use it to communicate, and more importantly how it may bring us a step closer to understanding human intelligence. He explores in his papers the interconnections between the different types of knowledge required for language understanding. He does so by using a program that "understands" language in a limited domain. The realm of his program is a mini-world that contains blocks and pyramids together with a "robot" that can understand commands regarding moving and locating the objects in its workspace and respond to such commands in a conversational manner. The robot is not a real robot, but rather a simulated one - Winograd's concern was not in the robot itself. Winograd realizes that we can not supply a program with all the knowledge that a person brings into a conversation. We can, however, at least venture into this small world where the domain is limited and introduce enough knowledge to be able to communicate without too much worry about the ambiguities associated with language. The program that Winograd wrote consists of three parts: a syntactic parser which deals with a large-scale grammar of English, a collection of semantic routines for encompassing knowledge it might need to interpret meaning of words, and a cognitive deductive component for exploring facts and answering questions. A fourth component is a set of routines for generating English sentences in response to questions and commands.

The program written by Winograd, serves as a model for language understanding. Winograd does not, however, propose that his model is a good model of psychological processes, but rather he has shown that a model can indeed be built. The aim of his program was to address the issues of how language is used in a framework of physical objects, events and discourse. Winograd realizes that in order to devise a model for language understanding that is more complete and psychologically more intuitive requires that we learn how complex systems are organized.

We find a much less philosophical view of AI in [Glasgow and Papadias, 1992] than in [Winograd, 1972]. They define AI as a research area concerned predominantly with the discovery of tools for solving hard problems. In [Glasgow and Papadias, 1992] the main focus is in a model for representing the space occupied by an image using occupancy arrays. They use the occupancy array to retrieve information such as shape, relative distance, and relative size.

We can find the study and use of spatial expressions in the area of machine translation. In [Dorr and Voss, 1993] their primary interest is in the translation of natural language sentences where the sentence contains a spatial relation, in particular when the meaning being conveyed relates to the location or path of an object in the real world. They use real world knowledge to help remove any ambiguity associated with a sentence. In other words their knowledge representation system should help filter out incorrect translations. They define a spatial predicate in order to better identify the parts of a spatial expression. The predicate is a structure that exists at the interlingua or semantic level of representation. They then define five components within spatial predicates:

1. The type of spatial predicate, (PLACE or PATH)
2. The item being located
3. One or more reference objects
4. A spatial operator that includes a LOCATION or ORIENTATION
5. A perspective locative

As an illustration, of the usage of the above components we will use their example, *"the cup on the table"*. In this example 1 is the location on the table where the cup is, 2 is the cup, since the cup is the object being located, 3 is the table, since it is the reference object, 4 is the spatial operator *on*, and 5 has no value since the spatial relation is independent of the viewer's perspective. In summary [Dorr and Voss, 1993] have used spatial expression as an example to illustrate their method for defining the relation between the interlingua, and the knowledge representation components of a machine translation system.

Related more to the issues in this thesis is work found in [Srihari, 1991b], and [Srihari, 1991a]. Srihari developed a system called PICTION, to extract faces from photographs through the parsing of newspaper captions. This system uses spatial constraints to reduce the number of possible labels attached to each face. The use of a rule-based system helps reduce the choices even further until a unique labeling is found. The rules use spatial heuristics as well as characteristics of face (male vs female). The system does not require that a face model for all the faces that we wish to recognize be stored, as is common in vision techniques. She handles certain spatial relations that are common to newspaper captions, such as left-of, right-of and behind. The system consists of three major modules: the vision module, the natural language processing module and the interpretation module.

The system progresses through various phases before arriving at a labeling for each face found. They are:

1. Parse the caption to produce a conceptualized graph, which contains information regarding the objects hypothesized to be in the picture, such as their physical appearance, and the spatial relationship between them.
2. The interpretation module calls the vision module to generate face candidates.
3. The interpretation module then uses information from the conceptualized graph to make some initial guesses about the name-face correspondences.
4. The interpretation module then calls upon the rule-based system to discriminate among the hypotheses until a unique solution is found.

The integration of vision and language is also addressed in [Zernik and Vivier, 1988]. In this paper the authors point out the lack of a linguistic system capable of handling words such as *near* and *far* as something other than part of an abstract database. They view language processing as an objective oriented task in which lexical semantics facilitates task performance. Unlike linguists who may use a word to disambiguate, Zernik and Vivier use words as a directive for identifying an object.

Zernick and Vivier devised a computational model that receives textual and visual input and guides an "agent" by linguistic text to an object. Zernik and Vivier assert that cognitive models must receive input from a visual and language domain. They realize that this requirement poses two main challenges; language is vague and vision algorithms are lacking. Their example of the vagueness problem associated with language is the sentence *"Chicago is between L.A. and Albany"*. Although this sentence is true for a person flying between L.A. and Albany and stopping in Chicago, it is not precisely between L.A. and Albany since Chicago does not lie on a straight line between L.A. and Albany. One of the problems that Zernik and Vivier point out with image understanding systems is that they require models of objects in order to identify them in a scene, and this type of problem may also lead to searching for a plane on a railroad track, even though we know that planes are not found on railroad tracks. It is this problem that leads Zernik and Vivier to assert that world knowledge must be incorporated into a vision/language system. Zernik and Vivier support an object-oriented approach for interpreting locative expressions, rather than a geometrical oriented approach. They do so because they believe that to understand the meaning of some sentences requires that we know something about the objects in question. They give the example of *"The mouse is in the trap."* In this sentence we know that the mouse is trapped but it is not necessarily contained *in* the trap. Likewise for a sentence like *"Jim is in the pool"*.

In summary, Zernik and Vivier assert that locatives do not posses precise geometrical interpretation, and because of this fact, world knowledge should be incorporated into a vision/language system. They also assert that locatives are not just pointers to objects but rather that they guide navigation through a scene. The work in this thesis will show the effects of using locatives to navigate through a scene, but will nevertheless use a geometrical approach to do so. The preciseness that Zernik and Vivier claim cause problems with the geometrical approach will be relaxed in order to gain more flexibility in using the locatives.

We can again find reference to the need for a connection between vision and language in [Jackendoff, 1987b]. In this book Jackendoff states that one of the fundamental problems for a theory of natural language is in how we talk about what we see. In order to accomplish this some connection must be found between language and vision. In [Macnamara, 1978], Macnamara points out that in order to talk about what we see, information provided by the visual system must be translated into a form compatible with the information used by the language system. Jackendoff realizes that much of spatial thinking depends on aspects of both.

In [Mukerjee and Joe, 1990] we again find reference to spatial relations as they are used for creating spatial maps of an environment. The representation they use capture more of the functional relevancy of the spatial relations. For example they ask the question "Are the given objects aligned at some face, line or point?". If not, then "Where is one object with respect to another?" They perform alignment-based reasoning in two and higher dimensions with objects at angles. Their claim is that relations involving tangency are "more important" than others in the categorization of spatial relations. The usefulness of their system is in its ability to model uncertainties in situations where the spatial knowledge is imprecise due to language constraints or data inadequacy.

Robot navigation, can surely benefit from an integration of vision and language. To do so, however, robot navigation needs to be approached from a more qualitative angle. Robot navigation has traditionally been quantitative, but the following pieces of work approach navigation from a more qualitative point of view. In [Kuipers and Levitt, 1988], and [Kuipers, 1978] we read about three models for robot navigation built upon the knowledge that map learning and navigation consists of a sensorimotor, procedural, topological and metrical assemblage. The sensorimotor agent is used by a traveler to sense his/her/its surroundings and the ability to move through the environment based on the sensory input. The procedural agent allow the traveler to traverse previously visited routes of the environment. The topological agent is used to represent features from the environment as well as the relationships of these objects. The metric agent is needed to describe the magnitude of actions, such as the distance of travel. [Park, 1993] also approaches navigation from a qualitative point of view, and uses the notion of a sensorimotor and procedural agent. The focus is in navigating in a large unstructured environment where the focus of attention is on the navigator and map-maker. The map-maker observes the environment and supplies a starting and goal position and designs a custom map that describes how to get from the starting position to the goal position. The responsibility of the navigator is to interpret the custom map and guide the robot to the goal accurately and efficiently. The environment is qualitatively described using sensory descriptions of how the objects appear to the robot and how these objects are related topologically. Related work in this area can be found in [Kender and Leff, 1989] and [Kender *et al.*, 1990].

Instructible robots are examined by Suppes in [Suppes, 1992]. The central issue in this work is not so much robotic in nature, but rather linguistic. The purpose of the work is to approach problems of understanding grammar and semantics in English by implementing programs that instruct robots to execute various commands. Suppes establishes the importance of interaction with an instructible robot, the understanding of the users intent, and the knowledge that the robot and user share about the environment.

Also interested in communicative purposes is Nagel [Nagel, 1988]. Nagel's interest is in deriving three-dimensional descriptions of moving objects as well as stationary scene components without reliance on domain-specific knowledge. The purpose of not relying on domain-specific knowledge was to base the descriptions on general assumptions about motion. Nagel's work focuses on deriving a conceptual description of activities using verbs of motion, to perform image sequence understanding. His work combines image analysis, knowledge representation, logic, and natural language processing. He acknowledges the enormous potential for interdisciplinary research that such a problem provides.

Related to Nagel's work is research by Neumann and coworkers [Neumann, 1984], [Marburger *et al.*, 1981], [Neumann and Novak, 1983]. They have investigated a system for generating natural language descriptions of traffic scenes. Verbs of locomotion are used to express the activity present in a sequence of images. Rather than just selecting a motion verb, Neumann and coworkers have created event models that represent *a priori* knowledge about relevant motions in a scene. A motion verb like *overtake* would require that car$_1$ be ahead of car$_2$ at a certain time interval, and that in the next time interval that they be beside each other, and in the next time interval that car$_2$ be ahead of car$_1$. In this way they create a hierarchy of events some more complex than others depending on the complexity of the event they are modeling.

Wahlster and coworkers [Wahlster *et al.*, 1983] have developed a system called CityTour which can answer question in German about a discourse world related to locations and traffic activity. Examples, of some of the text it generates is

```
The car drove past the bus stop.
The tram went along the third street.
```

It can also answer questions such as

```
Did the car start?  Yes, it started a short while ago.
```

This section has given a brief overview of some of the work being pursued in the AI community with regards to natural language generation and understanding, computer vision, and robot navigation. The work in this thesis will illustrate the possibility of creating a system that relies on ideas brought forth from each of the aforementioned areas of AI research.

Chapter 3

The Semantic Representation of Spatial Prepositions

At the heart of the computational understanding of spatial prepositions is their semantic representation. A representation is necessary if the system is to be able to assign a preposition to an object pair. The semantic representation, to be discussed in this chapter, captures the vital properties sufficient for a succinct use of the chosen prepositions [Abella and Kender, 1993]. The spatial prepositions can be encoded fairly concisely because objects are treated as "blobs" and because most of the properties characterized by these prepositions can be encoded using geometric properties such as alignment and distance. The following sections will provide the details of the notations and encoding.

3.1 Notations and Definitions

We begin this exposition with some of the notation and definitions necessary to understand the semantic representation. The set of spatial relations we have chosen, usually covered in language by prepositions ($\mathcal{P}$) is

$$\{near, far, above, below, aligned,$$
$$next, inside, left, right, between\}$$

We have also defined a computational model for intrinsic object properties such as shape or color; usually covered in language by adjectives ($\mathcal{A}$). We use only:

$$\{small, medium, big\}.$$

Additionally we use superlatives ($\mathcal{S}$) both relational and intrinsic:

$$\{nearest, farthest, leftmost, rightmost, topmost,$$
$$bottommost, biggest, smallest\}$$

Let $\mathcal{O}$ be a set of objects in an image. We have defined a preposition as a predicate that maps k objects to true (1) or false (0); true if the k objects meet the requirements imposed by the preposition and false otherwise.

$$\bar{p} : \mathcal{O}^k \longrightarrow \{0, 1\}$$

where $\bar{p}$ is a preposition from $\mathcal{P}$ and $\mathcal{O}^k$ is a k-tuple of objects. Except for the preposition *between* (where k=3), k=2.

Object parameterization Now that we have defined a preposition we need to define an object. Formally, each object is represented by a six element vector that depend on an object's area A, center (x, y), and second order central moments of inertia I_{xx}, I_{xy} and I_{yy}.

The pair of objects (a, b) is represented by a 12-component vector $(a, b) \in \mathcal{R}^{12}$. It is this scaled vector that we will be using in our future calculations.

Bounding Box The parameterization of objects presented above leads to the concept of a bounding shape. A bounding shape encloses an object using certain criteria. There are various ways to compute a bounding shape for an object, one of which may be to find the maximum and minimum x and y values belonging to the

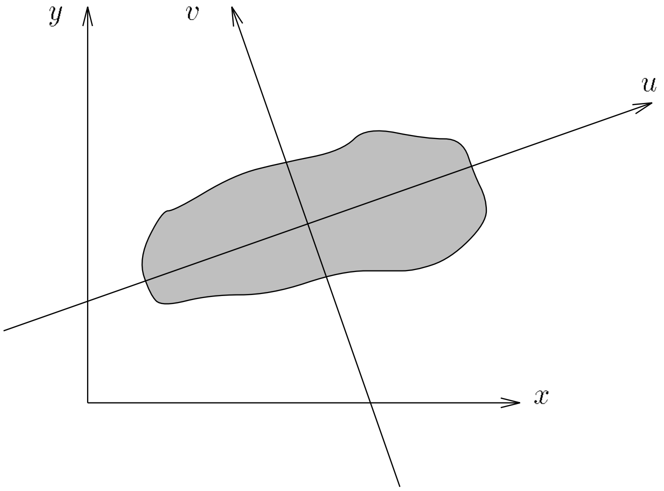


Figure 3.1: Axes of maximal(v) and minimal(u) moments of inertia

object. This, however, would not be a good bounding shape for objects that have "tentacles". The unshaded rectangle in figure 3.4 represents this bounding shape. The one that we have chosen is a bounding box and is defined through the values of w and h (defined below); it offers a measure of how much an object stretches in the x and y direction.

We define w and h in terms of the maximal and minimal moments of inertia. The maximal moment of inertia is given by the formula:

$$I_{\max} = \int\int_A u^2 du dv$$

where v and u are axes of maximal and minimal moments of inertia (see figure 3.1), respectively and A is an object's area. According to the Mean Value Theorem there exists a point, call it $(\bar{u}, \bar{v})$ such that

$$I_{\max} = \bar{u}^2 A$$

The quantity $\bar{u}$ is a measure of how much the object stretches along the u axis. We will set the half-width of our bounding box, w, to be $w = k\bar{u}$, where k is a constant to be determined, such that the bounding box of an object that is a

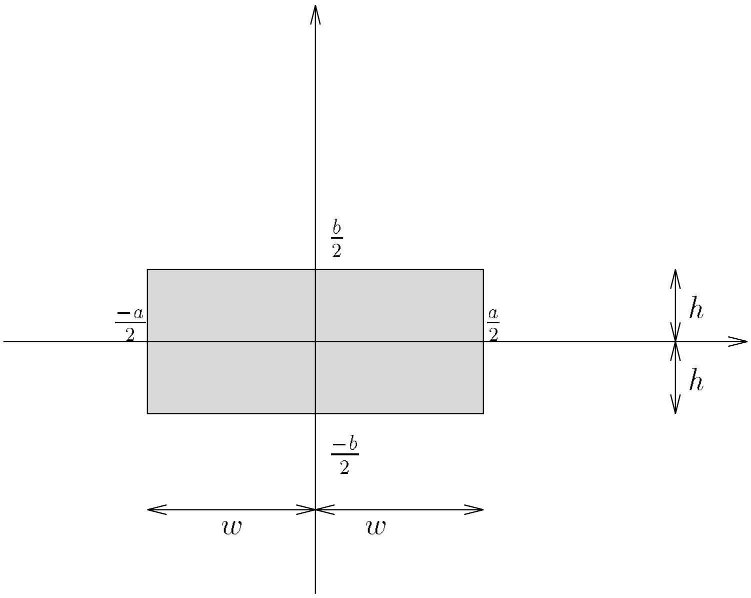


Figure 3.2: Rectangle we use to compute the width w and height h of the bounding box

rectangle should be the rectangle itself. In order to find k we compute the maximal moment of inertia [1] for a rectangle oriented in the x and y direction with width a and height b as shown in figure 3.2.

$$\begin{aligned}
I_{\max} &= \int\int_A x^2 dx dy \\
&= \int_{-a/2}^{a/2} x^2 dx \int_{-b/2}^{b/2} dy \\
&= \frac{x^3}{3}|_{-a/2}^{a/2} b \\
&= \frac{1}{3}(a^3/8 - (-a^3/8))b \\
&= 1/12a^3 b \\
&= 1/12a^2 A
\end{aligned} \tag{3.1}$$

Since we are using $\bar{u} = \sqrt{I_{\max}/A}$ as a measure of the elongation of an object

[1] We may also have chosen to compute the minimal moment of inertia, which would have given us the height of the rectangle

and we want the bounding box of a rectangle to be the rectangle itself, we set $w = a/2 = k\bar{u}$ and solve for k. Since we know that for a rectangle $I_{\max} == \frac{1}{12}a^2A$ we get that

$$\begin{aligned} w &= k\sqrt{\frac{\frac{1}{12}a^2A}{A}} \\ a/2 &= k\sqrt{\frac{\frac{1}{12}a^2A}{A}} \\ k &= 1/2\sqrt{12} \\ k &= \sqrt{3} \end{aligned} \tag{3.2}$$

In this case we were able to set w explicitly to $k\bar{u}$ because we assumed that the rectangle is oriented in the x and y direction, however, in general we need to project w and likewise h onto the x and y axis. To do so we use the following formula:

$$w = \sqrt{3}\max\{\sqrt{\frac{I_{\max}}{A}}|\cos\theta|, \sqrt{\frac{I_{\min}}{A}}|\sin\theta|\}$$

$$h = \sqrt{3}\max\{\sqrt{\frac{I_{\max}}{A}}|\sin\theta|, \sqrt{\frac{I_{\min}}{A}}|\cos\theta|\}$$

A graphical interpretation is shown in figure 3.3.

The bounding box we have just defined is illustrated by the shaded rectangle in figure 3.4. This choice of a bounding box that is aligned with the x and y coordinate (image frame) is based on human preference for alignment. In the radiography domain we find such objects as the spine which is oriented vertically and it is the axis of symmetry. In the landmark domain there is a "virtual north" preferred. For example, buildings in Manhattan are aligned with each other. In section 3.2 we will illustrate the need to define a related bounding box for the preposition *aligned* and that no bounding box was used for the preposition *between*. We do not claim that our bounding box is the best shape approximation but it yields good results, as will be illustrated in chapter 6. We may also have chosen

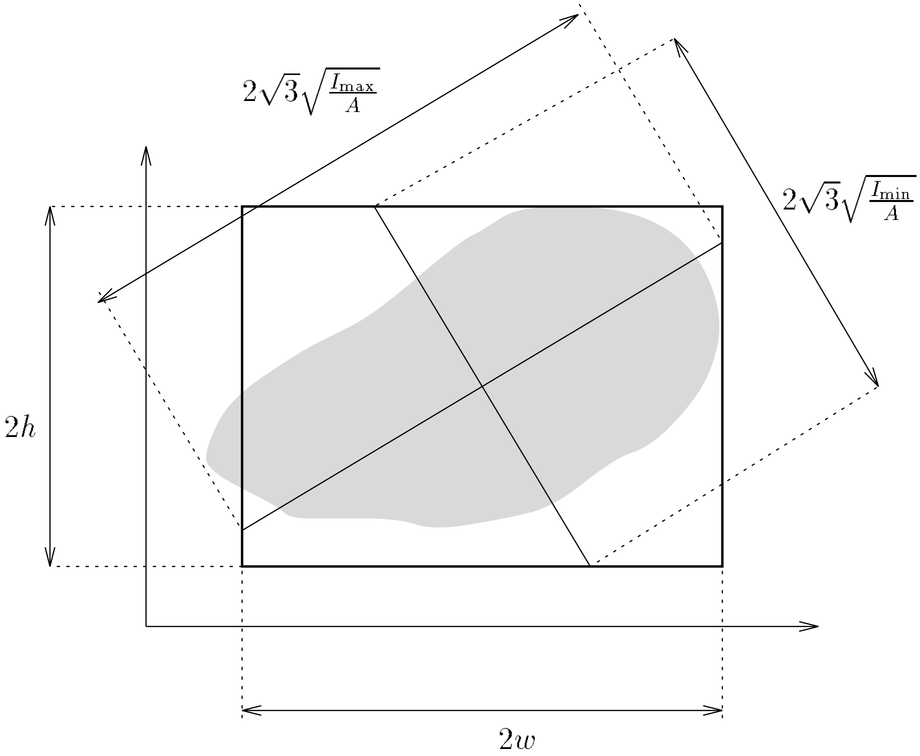

$2\sqrt{3}\sqrt{\frac{I_{\max}}{A}}$
$2\sqrt{3}\sqrt{\frac{I_{\min}}{A}}$
$2h$
$2w$

Figure 3.3: Projection of bounding box unto the x and y axis

to use a bounding circle, or an encompassing ellipse, all of which may have yielded satisfactory results.

Fuzzification of prepositions Two objects define a point in 12D space. A preposition $\bar{p}$ defines a set of points $U_{\bar{p}} \in \mathcal{R}^{12}$ such that $U_{\bar{p}} = \{(a,b)|\bar{p}(a,b) = 1\}$. The volume in this 12D space may be able to reveal some of the inherent properties associated with prepositions. In other words, examination of the space occupied by the various sets $U_{\bar{p}}$ may tell us something about the spatial prepositions. Vacancies in this 12D space may reveal why we do not have a word to describe certain spatial relationships among objects. The intersection and distances of volumes occupied by various spatial prepositions may reveal a correlation between various prepositions.

We say that objects a and b are related by preposition $\bar{p}_i$ if $(a,b) \in U_{\bar{p}}$. This "ideal" set is made up of pairs of object vectors that satisfy the constraints imposed

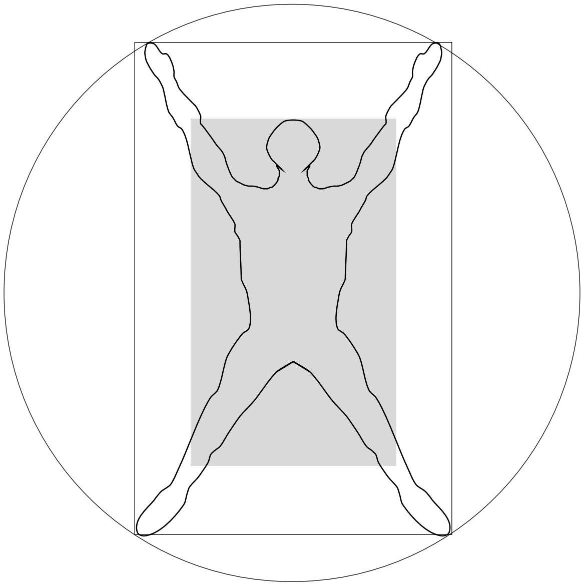


Figure 3.4: Three possible bounding shapes that could be used to represent an object's overall shape

by the preposition $\bar{p}_i$. As we well know, prepositions are inherently vague in their descriptions, and their interpretation may vary from person to person. Because of this, it is important to add some *fuzzifying* agent to our ideal set. The fuzzifying technique is as defined through fuzzy set theory [Klir and Folger, 1988]. The theory of fuzzy sets is used to represent uncertainty, contrary to the theory of classical sets[2] which represents certainty. A classical set divides objects in the world into two categories: those that certainly belong to a set and those that certainly do not belong to a set. A fuzzy set, on the other hand, divides the world much more loosely, by introducing vagueness into the categorization process. This means that members of a set belong to that set to a greater or lesser degree than other members of the set. Mathematically, members of the set are assigned a membership grade value that indicates to what degree they belong to the set. This membership grade is usually a real number in the closed interval between 0 and 1. Therefore a member that has a membership grade closer to 1 belongs to the set to a greater degree than a member with a lower membership grade. Because of its properties fuzzy set theory can find application in fields that study how we assimilate information, recognize

[2] Referred to as "crisp" sets in fuzzy set theory.

patterns [Abella, January 1992], and simplify complex tasks. In our notation the fuzzified ideal set is defined through a membership function

$$p_i(a,b) \in [0,1].$$

As a special case we define

$$\bar{p}_i(a,b) \equiv p_i(a,b) = 0 \text{ or } 1$$

3.2 The Semantic Representation

The semantic representation of prepositions entails representing objects through certain physical properties that can then serve as a basis for expressing prepositions. The physical properties we've chosen include object area, centers of mass, and elongation properties. These properties are calculated through the use of the zeroth, first, and second order moments. The basis for this choice of attributes is simplicity and familiarity. What ensues is a description of the semantic representation of language constructs from $\mathcal{P} \cup \mathcal{A} \cup \mathcal{S}$. Each preposition is defined through a set of inequalities. This results in sets $U_{\bar{p}}$ having nonzero measure (i.e. full dimensionality) in $\mathcal{R}^{12}$ which is necessary for the fuzzification procedure described in section 3.3.

3.2.1 Object Properties

Object properties, such as size, are useful for locating objects. An object's size can be combined with prepositions to enhance the overall descriptiveness of a locative expression.

We have encoded the properties *small, medium,* and *big.* In order for any of these properties to be valid at least one object must be different in size. If all the objects are the same size then obviously there can be no discrimination between them. An object must exist that is *small* and another object must exist that is *big*. Therefore, the following condition, S, must hold:

$$S \iff (\exists a) small(a) \wedge (\exists b) big(b)$$

An object a is *small* if S is true and its area A_a is not bigger than $s_{\text{small}} A_{\text{min}}$, where $s_{\text{small}} > 1$ is a constant and A_{min} is the area of the smallest object in the image. An object is *big* if S is true and if its area is greater than $(1/s_{\text{big}})A_{\text{max}}$, $s_{\text{big}} > 1$. Parameters s_{small} and s_{big} are calibrated beforehand for the particular domain and/or image. In the ensuing text we will consider $s_{\text{small}} = s_{\text{big}} = s$. An object is *medium* if S is true and it is not *small* and not *big*.

Condition S can be made true if there is any significant size difference between the objects, or equivalently, if

$$sA_{\text{min}} \leq (1/s)A_{\text{max}}$$

which translates to the condition

$$s^2 \leq A_{\text{max}}/A_{\text{min}}$$

This last condition is the algebraic representation of S. Now we can write *small, medium, big* as follows:

$$\begin{aligned} small(a) &= A_a \leq sA_{\text{min}} \wedge \neg(A_a \geq (1/s)A_{\text{max}}) \\ big(a) &= A_a \geq (1/s)A_{\text{max}} \wedge \neg(A_a \leq sA_{\text{min}}) \\ medium(a) &= A_a < (1/s)A_{\text{max}} \wedge A_a > sA_{\text{min}} \end{aligned}$$

3.2.2 Binary Prepositions

In this section we will define those prepositions from $\mathcal{P}$ that involve a reference object r and a figure object f.

NEAR

We have defined *near* so that the enlarged bounding boxes in figure 3.5 have a non-empty intersection. Mathematically this is :

$$\begin{aligned} d_x &\leq (W_f + W_r) \\ \text{and} & \\ d_y &\leq (H_f + H_r) \end{aligned}$$

where $d_x = |x_f - x_r|, d_y = |y_f - y_r|$ and

$$\begin{aligned} W_f &= w_f + \rho h_f \\ W_r &= w_r + \rho h_r \\ H_f &= h_f + \rho w_f \\ H_r &= h_r + \rho w_r \end{aligned}$$

and ρ is a statistically determined parameter determined from human psychology studies that captures how much the bounding boxes should extend. The value of ρ was determined to be 0.6.

The enlarged bounding box is necessary because the inner bounding box in figure 3.5 would not be able to handle the situation depicted in figure 3.6. If *near* depended on the bounding box shown in figure 3.6, that has sides of length $2w$ and $2h$, then the bounding boxes would not intersect and these two objects would not be considered *near*. However, we tend to consider these objects to be *near*. This

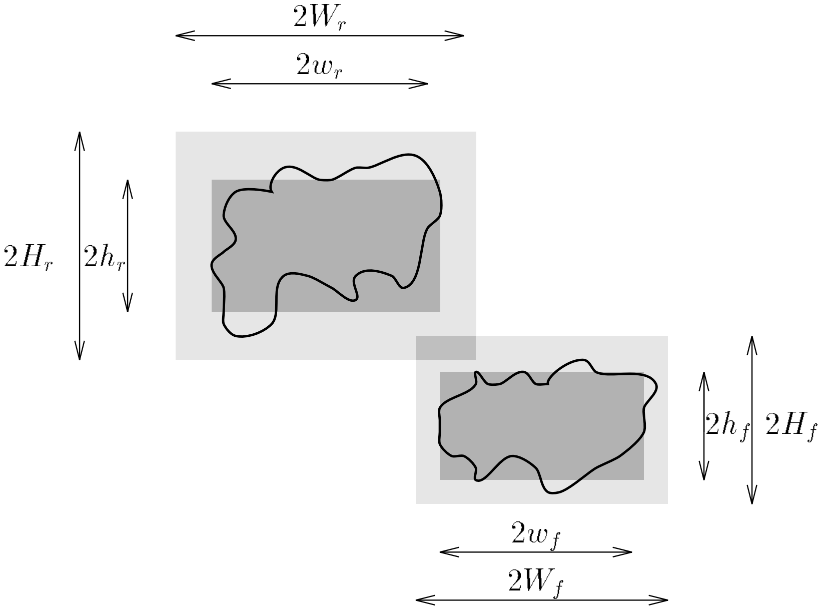


Figure 3.5: Two objects that are *near* each other

is why the enlarged bounding box is used. The enlarged bounding box combines the elongation properties of the objects in both the x and y extent. Therefore for the objects in figure 3.6 the enlarged bounding box would intersect as illustrated in figure 3.7.

FAR

Far is not the complement of *near* as one may initially suspect. We may be faced with a case where an object is neither *near* nor *far* from another object, but rather it is *somewhat near* or *somewhat far*. This notion of *somewhat* will be explained more fully when we discuss the concept of *fuzzification*. For now it suffices to say that *far* is defined so that the distance between the two enlarged bounding boxes in either the x extent or the y extent is larger than the maximum length of the two objects in that same x or y extent (see figure 3.8). Mathematically this is:

$$d_x \geq W_f + W_r + \max(W_f, W_r)$$

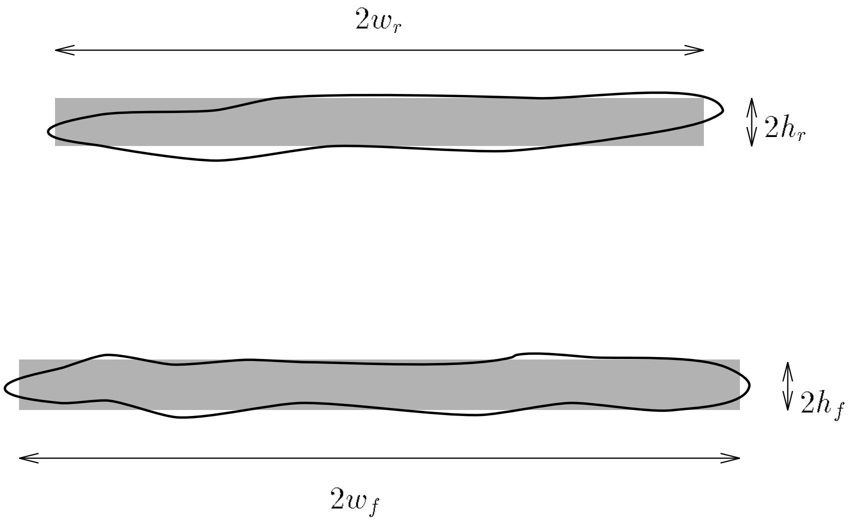


Figure 3.6: Two objects that are *near* yet whose bounding boxes do not intersect

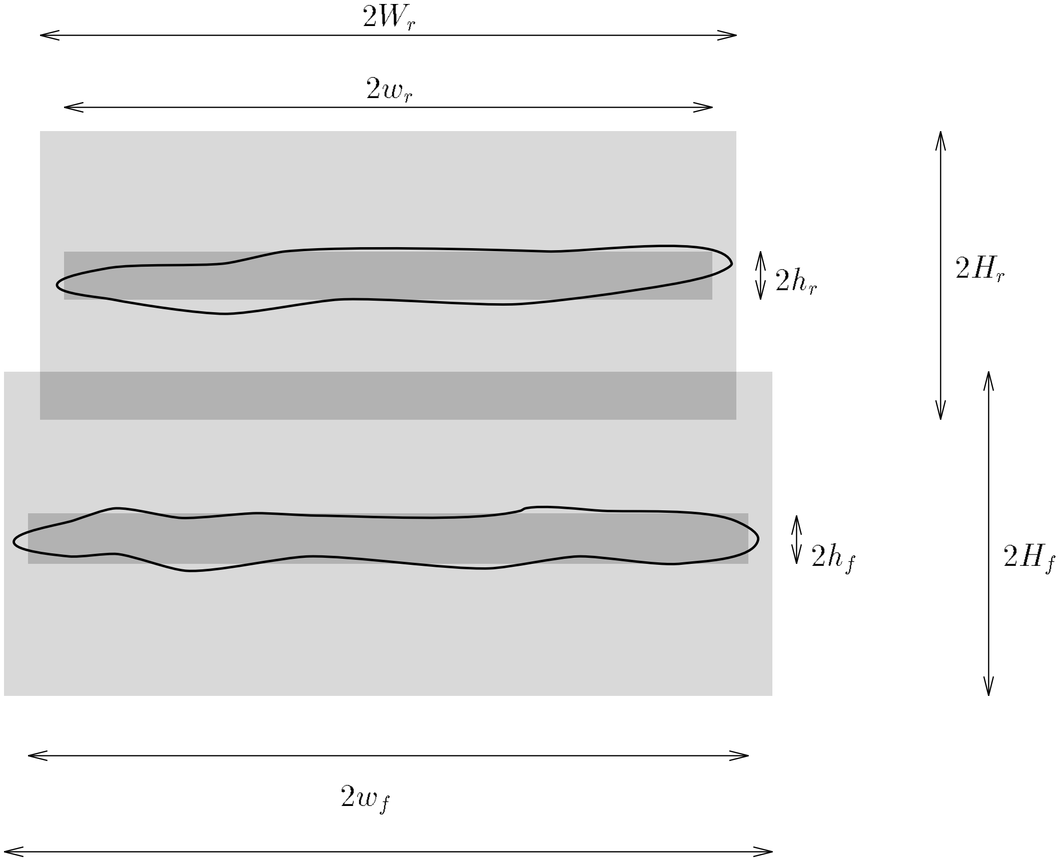


Figure 3.7: The enlarged bounding box necessary to determine that the two elongated objects are in fact *near* each other

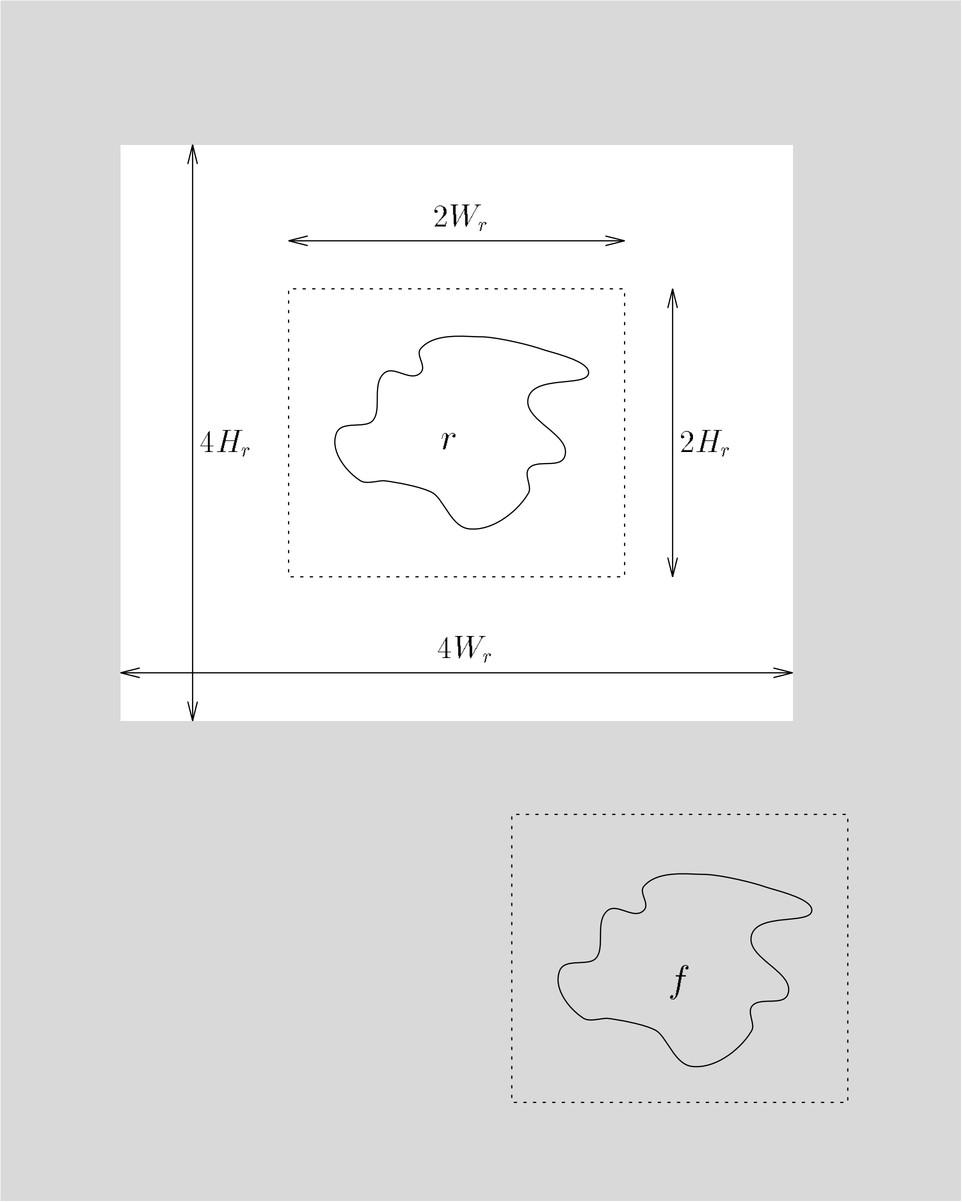


Figure 3.8: Two objects that are *far* from each other

or

$$d_y \quad \geq H_f + H_r + \max(H_f, H_r)$$

INSIDE

Inside requires that the bounding box of one object be completely embedded

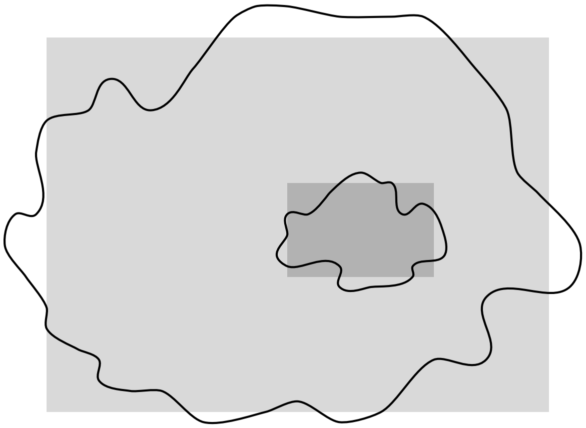


Figure 3.9: Two objects that satisfy the condition for *inside* because the bounding box of the smaller object is completely within the bounding box of the bigger object

within the bounding box of another. Formally,

$$\begin{aligned} d_x &\leq (w_r - w_f) \\ \text{and} & \\ d_y &\leq (h_r - h_f) \end{aligned}$$

ABOVE

Above requires that the projection on the y axis of the bounding box of the figure object f be *above* the projection of the bounding box of the reference object r. The mathematical relationship is

$$h_f + h_r \leq d_y$$

Note that *above* is not commutative. We define *below* similarly (see below). As with *near* and *far*, *above* and *below* are mutually exclusive prepositions. However, not-*above* does not strictly imply *below*. Two objects maybe side by side and hence are neither *above* nor *below*.

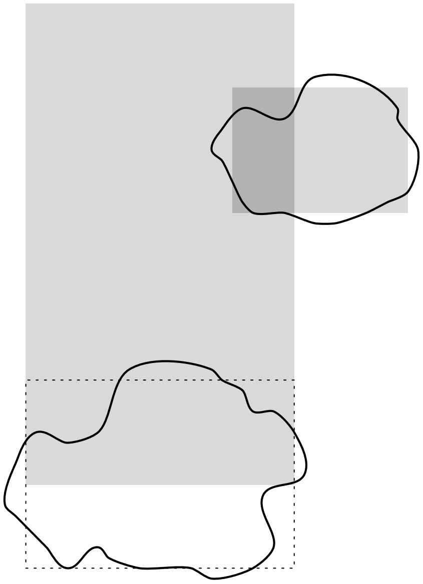


Figure 3.10: Definition of *above*: The dark shading indicates that an object is *above* another

We will find it useful to define *above*, *below*, *left* and *right* as *strictly above*, *strictly below*, *strictly left* and *strictly right* in the radiography domain. The reason for these distinctions are rooted in the different interpretations of prepositions in different domains. As we will examine in section 3.4, people interpreting maps use these prepositions differently than radiologists analyzing radiographs.

Strictly, in the case of *above* and *below* means that the projections of the bounding boxes on the x axis must intersect (see figure 3.10) The algebraic representation of this condition is

$$d_x \leq (w_f + w_r) \tag{3.3}$$

We will also find it useful to define *restricted above*, *restricted below*, *restricted left* and *restricted right*. *Restricted*, in the case of *above* and *below* means that the reference object entirely covers the figure object's extent in the x direction. The algebraic representation of this condition for *above* and *below* is

$$d_x \leq w_r - w_f$$

BELOW

Below requires that the projection on the y axis of the bounding box of the figure object f be *below* the projection of the bounding box of the reference object r. The algebraic representation is:

$$h_f + h_r \leq d_y$$

Strictly below requires the additional condition 3.3.

ALIGNED

The alignment[3] property is angular in nature, therefore its quantification involves inequalities between angles, rather than lengths as the previous prepositions had required. For this purpose we define a different type of bounding box that is centered at the object's center of mass and oriented along the object's principal inertia axes with dimensions proportional to the object's maximum and minimum moments of inertia. Angles θ, θ and Θ are as shown in figure 3.11. With this in mind, the preposition *aligned* is defined as:

$$\begin{array}{rcl} \max(\theta_f, \theta_r) < & \alpha_f & < \min(\Theta_f, \Theta_r) \\ & \text{and} & \\ \max(\theta_f, \theta_r) < & \alpha_r & < \min(\Theta_f, \Theta_r) \end{array}$$

NEXT

We've defined *next* as a combination of the prepositions *near* and *aligned.* Therefore the definition for *next* is:

[3] Although not a preposition from a language perspective we've adopted it as a spatial preposition.

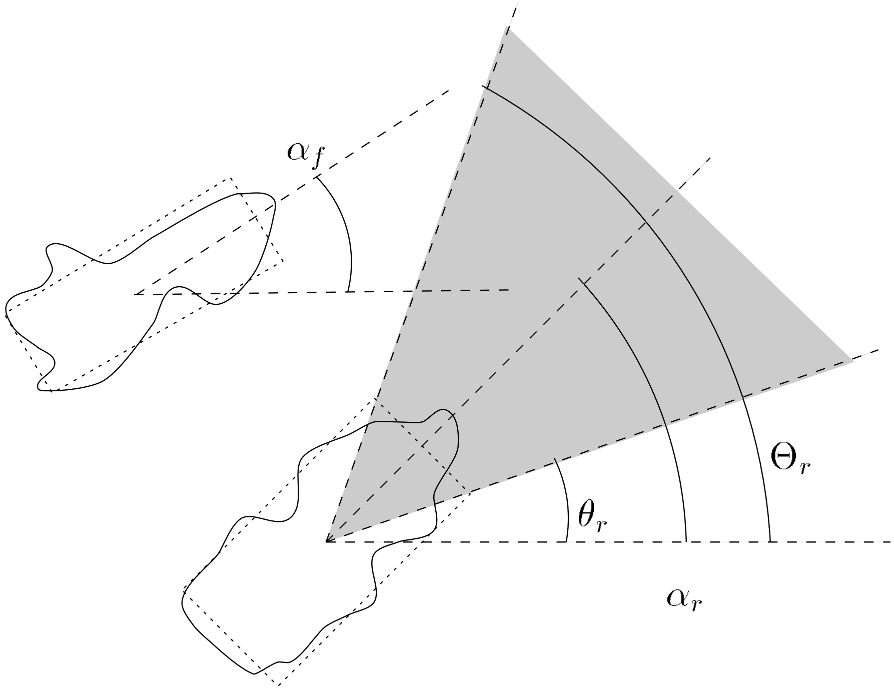


Figure 3.11: Definition of relevant angles for *aligned*

$next = near \wedge aligned$

The preposition *next* is an example of a spatial preposition that is a combination of more elementary prepositions. This hints at the possibility of a natural hierarchy of spatial prepositions.

LEFT

The preposition *left*, as we define it for the landmark navigation task, requires that the projection on the x axis of the bounding boxes of the figure object f and reference object r do not intersect and that the projection of f be to the *left* of the projection of r:

$$(w_f + w_r) \leq d_x$$

Strictly left, used in the radiography domain, requires the additional condition that the projection of the bounding boxes on the y axis intersect:

$$d_y \leq (h_f + h_r) \tag{3.4}$$

Restricted left and *restricted right*, require a condition similar to the condition defined for *restricted above* and *restricted below*. This condition is:

$$d_y \leq h_r - h_f$$

RIGHT

Like *left*, *right* in the landmark navigation domain only requires that the projection on the x axis of the bounding box of f be to the *right* of the projection of the bounding box of r:

$$(w_f + w_r) \leq d_x$$

In the radiography domain we use *strictly right* that requires the additional condition 3.4.

3.2.3 Ternary Prepositions

BETWEEN

Between, unlike all the prepositions we have covered until now, is not a binary preposition, but rather a ternary one. We have defined *between* so that it handles situations where the figure object f is *between* the reference object r and a third object a. Some examples are illustrated in figures 3.12. These figures illustrate the ideal configuration of objects for *between* to be satisfied. It is ideal because the center of all three objects are aligned. The representation, however, needs to capture slight variations from this ideal.

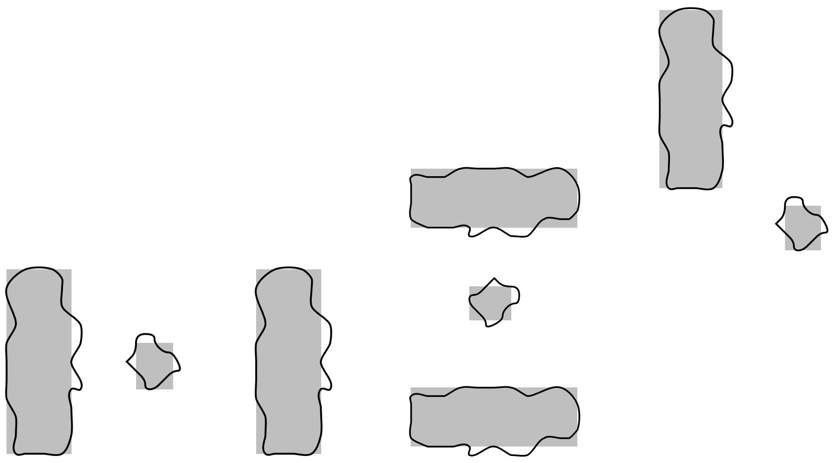


Figure 3.12: Examples of *between*

To do this we formed a triangle whose vertices are the centers of the three objects. We use the height from a vertex that is the center of the figure object of this triangle as a measure of how much the figure object is *between* the two surrounding objects. Intuitively, the height should be small. If the projections t_x and t_y (see figure 3.13) of this height are smaller than the elongation of the figure object and the elongation of the smaller of the two surrounding objects in the x and y directions respectively, then we consider the figure object to be *between* the two surrounding objects.

Analytic geometry yields the following algebraic representation for this condition:

$$t_x \leq \max(w_f, \min(w_r, w_a)) \tag{3.5}$$

$$t_y \leq \max(h_f, \min(h_r, h_a)) \tag{3.6}$$

where

$$t_x = |z(y_a - y_r)|$$

$$t_y = |z(x_a - x_r)|$$

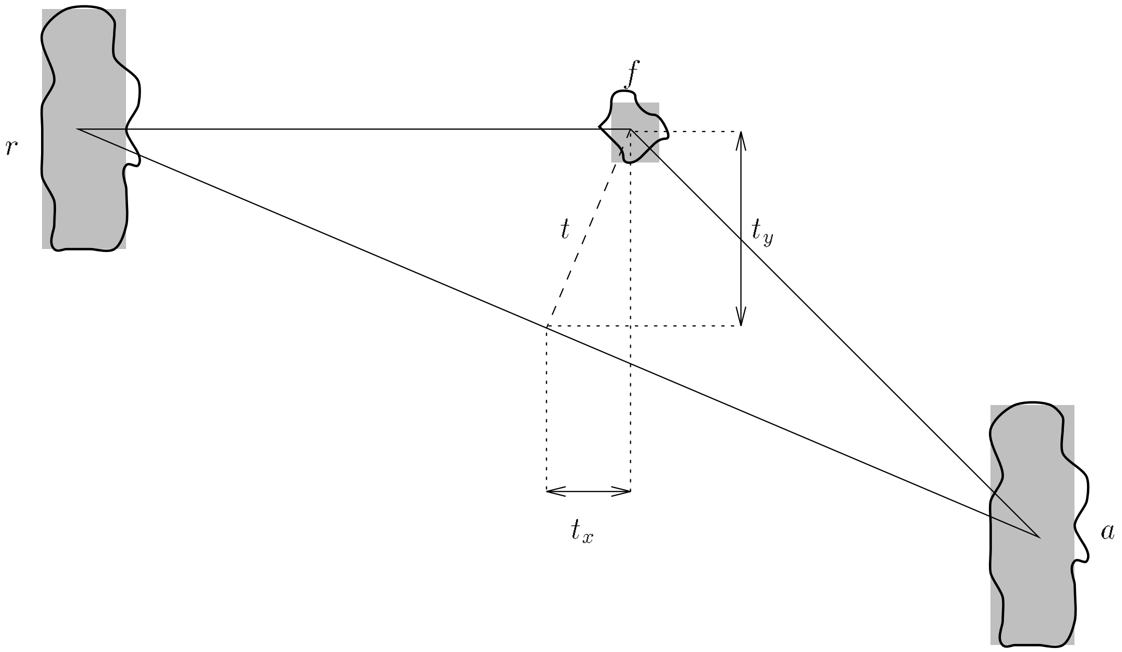


Figure 3.13: Definitions of t_x and t_y

and

$$z = \frac{x_r(y_f - y_a) + x_f(y_a - y_r) + x_a(y_r - y_f)}{(x_a - x_r)^2 + (y_a - y_r)^2}$$

The last condition in the representation ensures that the angles in the vertices that correspond to the centers of the objects r and a are acute. We do not want an obtuse angle in these vertices because we may have a small height but a situation such as the one depicted in figure 3.14. Without this last condition the figure object would be considered *between* because of the small height. Using the Pythagorean theorem we get

$$|d^2(r, f) - d^2(f, a)| \leq d^2(a, r) \tag{3.7}$$

where,

$$d^2(a, b) = d_x^2 + d_y^2.$$

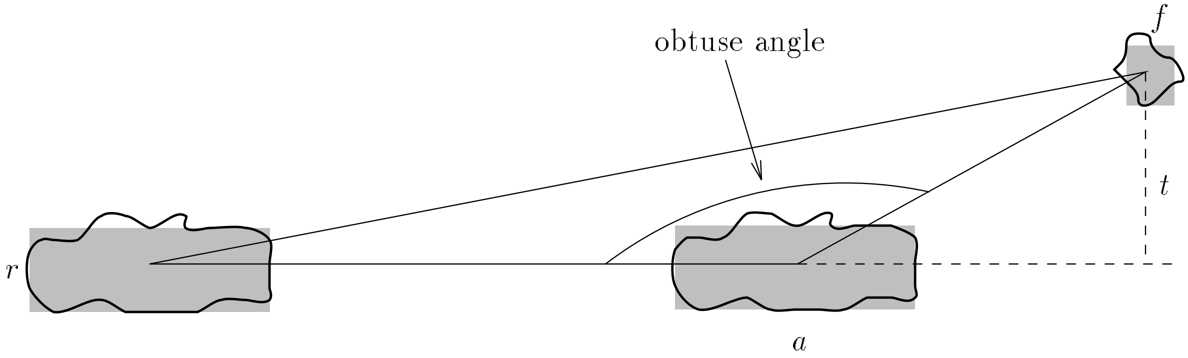


Figure 3.14: An obtuse angle in an object other than f may cause the height t to be small yet f is clearly not *between* r and a

For example, $d^2(f,a)$ is the squared distance between the center of f and a.

The conjunction of conditions 3.5, 3.6 and 3.7 defines the computational model for *between*.

3.2.4 Unary Superlatives

SMALLEST

The figure object f is the smallest object in the image if its area is smaller than the area of all the other objects in the image. We represent *smallest* as :

$$A_f \leq \min_{a \in \mathcal{O}} A_a$$

BIGGEST

The figure object f is the biggest object in the image if its area is bigger than the area of all the other objects in the image. We represent *biggest* as:

$$\max_{a \in \mathcal{O}} A_a \leq A_f$$

3.2.5 Binary Superlatives

NEAREST

The figure object f is the *nearest* one to r if there are no other objects whose distance to r is shorter. We use the Manhattan distance between bounding boxes:

$$d(a,b) = \max(d_x - (w_a + w_b), 0) + \max(d_y - (h_a + h_b), 0)$$

Object f is the *nearest* to r if

$$d(f,r) \leq \min_{a \in \mathcal{O}} d(a,r)$$

FARTHEST

The figure object f is the *farthest* one from r if there are no other objects whose distance to r is longer. The distance is defined the same way that it is defined for *nearest*. The condition for *farthest* is

$$\max_{a \in \mathcal{O}} d(a,r) \leq d(f,r)$$

LEFTMOST

The figure object f is the *leftmost* if it is the object furthest to the left. We represent this as:

$$(x_f - w_f) \leq \min_{a \in \mathcal{O}} (x_a - w_a)$$

RIGHTMOST

The figure object f is the *rightmost* one if it is the object furthest to the right. We represent this as:

$$\max_{a \in \mathcal{O}} (x_a + w_a) \leq (x_f + w_f)$$

TOPMOST

The figure object f is the *topmost* one if it is the object furthest along the y axis. We represent this as:

$$(y_f - h_f) \leq \min_{a \in \mathcal{O}} (y_a - h_a)$$

BOTTOMMOST

The figure object f is the *bottommost* one if it is the object lowest on the y axis. We represent this as:

$$\max_{a \in \mathcal{O}} (y_a + h_a) \leq (y_f + h_f)$$

3.3 The Fuzzification of Spatial Prepositions

This section describes why and how spatial prepositions are fuzzified. We will explain the fuzzification of binary prepositions; other language constructs presented in the previous sections are fuzzified in a similar way. Spatial prepositions need to be fuzzified because they are vague by their very nature; they depend on context and depend on an individual's perception of them with respect to an environment. For these reasons some leeway must be allowed when deciding if two objects are related through a given preposition.

As was mentioned in section 3.1, the preposition $\bar{p}$ is a logic predicate. It is defined as the disjunction of conjunctions of inequalities $\bar{p}_{ij}$. Formally,

$$\bar{p}(a,b) = \vee_i \wedge_j \bar{p}_{ij}(a,b),$$

where (a, b) is the 12D vector of object attributes for objects a and b, and

$$\bar{p}_{ij}(a,b) = \begin{cases} g_{ij}(a,b) \leq 0 \\ \neg(g_{ij}(a,b) \leq 0) \end{cases}$$

Functions g_{ij} define properties associated with preposition $\bar{p}$. For example, for *near* there are two inequalities that must be satisfied:

$$g_{11}(a,b) = d_x - (W_a + W_b)$$

$$g_{12}(\mu) = d_y - (H_a + H_b)$$

Hence,

$$\bar{p}_{11} = (d_x - (W_a + W_b) \leq 0)$$

$$\bar{p}_{12} = (d_y - (W_a + W_b) \leq 0)$$

The set $U_{\bar{p}}$ is defined by

$$U_{\bar{p}} = \{(a,b) | \bar{p}(a,b) = 1\}$$

i.e. object pair (a, b) belongs to the set $U_{\bar{p}}$ if it satisfies preposition $\bar{p}$.

What needs to be done now is to fuzzify the ideal set $U_{\bar{p}}$. To fuzzify $U_{\bar{p}}$ entails fuzzifying the conjunctions, disjunctions, and inequalities. The fuzzy membership function p is defined as follows:

$$p(a,b) = \vee_i^{\text{fuzzy}} \wedge_j^{\text{fuzzy}} p_{ij}(a,b)$$

where

$$p_{ij}(a,b) = \begin{cases} (g_{ij}(a,b) \leq 0)^{\text{fuzzy}} \\ \neg^{\text{fuzzy}}(g_{ij}(a,b) \leq 0)^{\text{fuzzy}} \end{cases}$$

and

$$\begin{aligned} x \vee^{\text{fuzzy}} y = & \quad x + y - xy \\ x \wedge^{\text{fuzzy}} y = & \quad xy \\ \neg^{\text{fuzzy}} x = & \quad 1 - x \end{aligned}$$

for $x, y \in [0,1]$. Function p_{ij} should satisfy the following three criteria:

- $$\lim_{g_{ij}(a,b) \to -\infty} p_{ij}(a,b) = 1$$
- $$\lim_{g_{ij}(a,b) \to \infty} p_{ij}(a,b) = 0$$
- p_{ij} should be monotonically not increasing.

For values of $g_{ij}(a,b)$ very small the value of $p_{ij}(a,b)$ should be 1, since this would mean that object pair (a,b) satisfies preposition $\bar{p}$. For values of $g_{ij}(a,b)$ very big the value of $p_{ij}(a,b)$ should be 0 since the object pair (a,b) certainly does not satisfy preposition $\bar{p}$. However, for values of $g_{ij}(a,b)$ that are not very big, but are nonetheless greater than 0, we would like the value of $p_{ij}(a,b)$ to decrease gradually as $g_{ij}(a,b)$ increases. Figure 3.15 shows the membership function that we chose:

$$p_{ij}(a,b) = (g_{ij}(a,b) \leq 0)^{\text{fuzzy}} = \begin{cases} e^{-g_{ij}(a,b)/\sigma}, & g_{ij}(a,b) > 0 \\ 1, & g_{ij}(a,b) \leq 0 \end{cases}$$

If $g_{ij}(a,b)$ is indeed less than 0 then there are no doubts that the relationship of the objects in question are satisfied by the given preposition. If however, $g_{ij}(a,b)$ is greater than 0 then the relationship is not ideally satisfied. Nonetheless we want a value that will indicate how far the relationship is from ideal. The value of $e^{-g_{ij}(a,b)/\sigma}$ is the measure used to determine how much the relationship of the objects deviates from ideal. The fuzzifying agent σ plays a crucial role in this determination. For example, if σ is 0 and $g_{ij}(a,b) > 0$ then the value of $p_{ij}(a,b)$ is

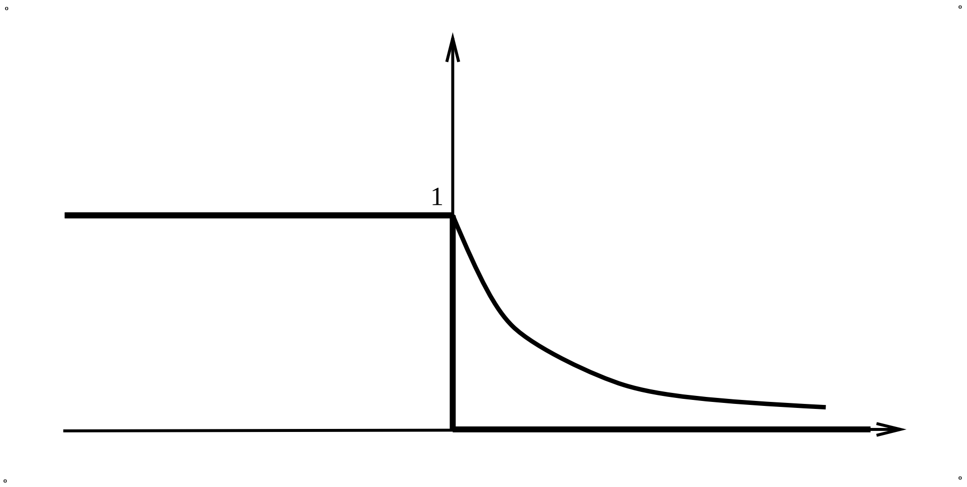


Figure 3.15: The fuzzification function

0, which is precisely what is to be expected since setting σ to 0 means that we are not allowing for any fuzzification. As σ is increased the fuzzification is increased, meaning that there is a larger allowed tolerance when determining whether or not two objects are in the relationship described by preposition $\bar{p}$. The parameter σ is set by experiments conducted on people.

3.4 Validation and Calibration

Validation and calibration of the semantic representations of section 3.2 are an important part of generating results that are as intuitive as possible. In order to test the validity of the semantic representations we designed a survey to examine how people interpret prepositions in the context of maps. We gave the survey of figure 3.17 to twenty one people and asked them to check off all the prepositions that they believed the object pairs from figure 3.16 satisfied.

We discovered from the survey that people do not think of the prepositions *above*, *below*, *left* and *right* as *strictly above*, *strictly below*, *strictly left* and *strictly right*. For example, everybody agreed that the object numbered 9 in figure 3.16 is

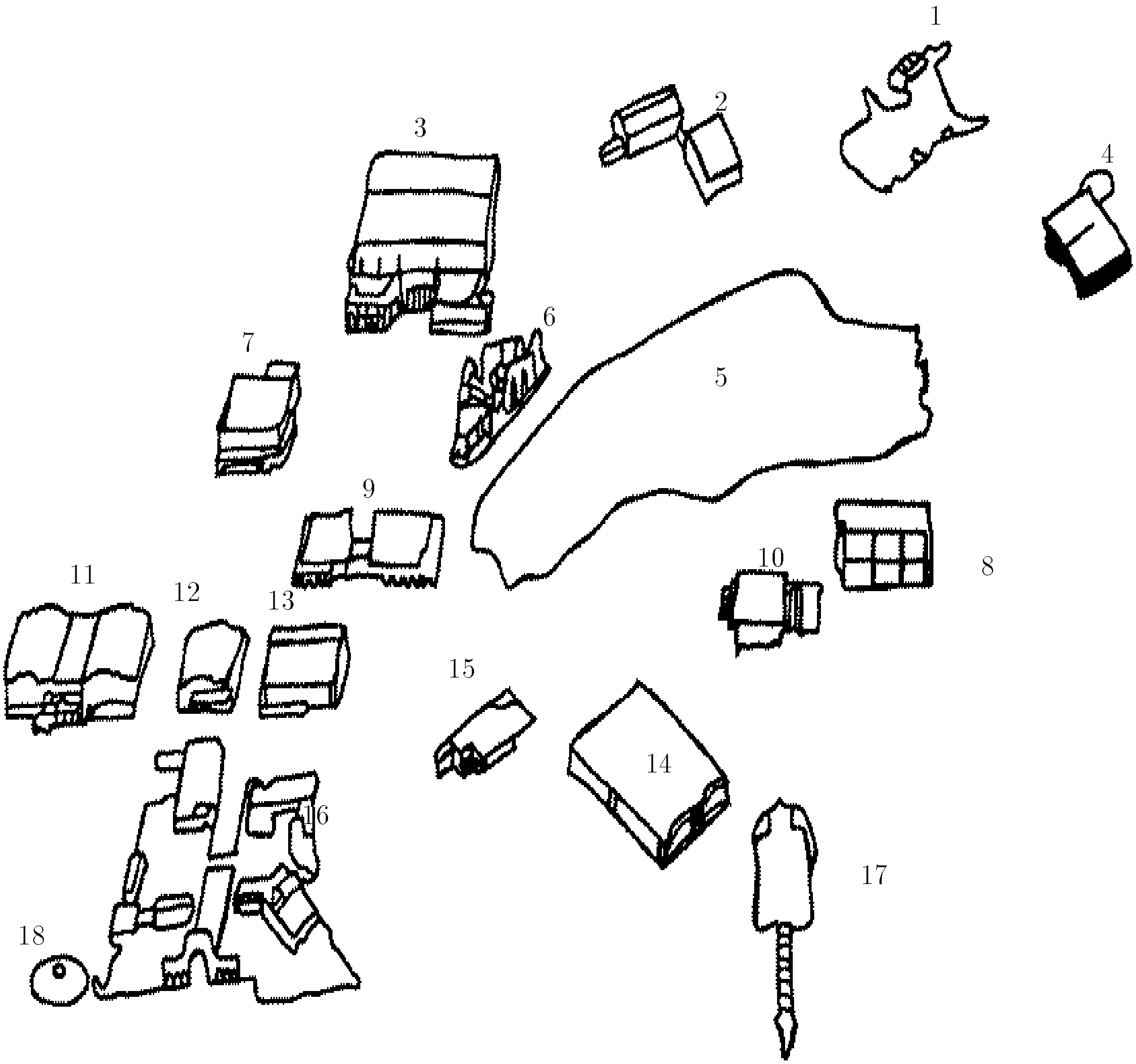


Figure 3.16: Theme Park Map of Universal Studios used in the survey

Place a check mark in the boxes you feel describe the pair of objects in the columns. You may think of this as a direction-giving task, where you are being asked to describe an object you would like to get to. The first object in each pair represents the object you want to describe. The second object represents the object you are standing by. The first column is an example. In it the object that I want to describe is the object numbered 12 in the picture and I am at the object numbered 11. Therefore in my opinion object 12 is to the right of 11, near it, and aligned with it. You may have chosen to describe object 12 differently; there is no right or wrong answer, simple use your own judgment.

	(12, 11)	(4,1)	(11,13)	(13,8)	(7,15)	(2,5)	(11,16)	(12,4)	(9,1)	(11,4)	(17,6)
near	X										
far											
above											
below											
aligned	X										
left											
right	X										
small											
medium											
big											
leftmost											
rightmost											
bottommost											

Figure 3.17: Survey used to gather calibration data and to perform validation of the semantic representations

below the object numbered 1; object 9 is certainly not *strictly below* 1. Although this is the case for the landmark navigation domain this is not the case in the radiography domain. Radiologists generate descriptions that are very local to the place where they believe a stone to be. Therefore, it is more appropriate to use the models that encodes the condition of *strictly* for these prepositions.

The survey confirmed the models from section 3.2. The most interesting case is *bottommost*. Although the center of object 18 is lower than the center of object 17 nobody considered object 18 to be the *bottommost* one. Like in our model they considered object 17 to be the *bottommost* one because it had the lowest point of all the objects on the map.

There was some confusion as to what was *small*, *medium* and *big*. From this survey we were not able to extract any understanding as to what objects on the map people considered to be *small*, *medium* or *big*. Perhaps it was simply the case that they felt that it was not necessary to describe the object's intrinsic properties if they could use a set of prepositions instead. Another possibility is that people would consider an object *small*, *medium* or *big* with respect to the reference object irrespective of the other objects. For example, quite a number of people believed that object 2 is *small* when the reference object is 5, even though object 13, for example, is about the same size and it was not considered *small* by as many people. The difference was that it was paired with object 8 which is not as big as object 5. This could explain the difference in the values for *small*, *medium* and *big* for those objects paired with figure object 11. One would expect those numbers to be the same but they were not, and this is probably due to what object object 11 was paired with; thus *small* is relative not absolute.

The survey also gives us a way of calibrating the parameters of the semantic representation so that the results we get are as close as possible to what people would expect. We discovered that people are in agreement when a preposition is definitely satisfied or not satisfied by an object pair. However, there is some discrepancy when the object pair neither satisfies nor does not satisfy the preposition exactly. This is where calibration is most vital. What we did was to find the parameters σ (and ρ in the case of *near* and *far*) that minimized the sum of the squared differences between the values for the prepositions found via the survey (see figure 3.18) and the values that the system produced for these same prepositions and object pairs. Figure 3.19 shows the quality of the calibration for *near*. Notice that the points lie fairly close to the diagonal which indicates that the system values and measured values were similar. The following table shows the values for the parameters that we calibrated:

	(4,1)	(11,13)	(13,8)	(7,15)	(2,5)	(11,16)	(12,4)	(9,1)	(11,4)	(17,6)
near	0.86	.62	0	0.05	0.81	0.9	0	0	0	0
far	0	0	.95	0.57	0	0	1.0	1.0	0.95	0.95
above	0	0	0	0.90	0.90	0.95	0	0	0	0
below	0.90	0	0.81	0	0	0	0.81	1.0	0.86	0.90
aligned	0	0.86	0	0	0.43	0.05	0	0	0	0
left	0	0.81	0.86	0.90	0.05	0.95	0.90	1.0	0.95	0
right	1.0	0.14	0	0	0	0	0	0	0	0.86
small	0.47	0.2	0.47	0.27	0.93	0.40	0.40	0.30	0	0.13
medium	0.2	0.40	0.33	0.40	0	0.53	0.13	0.40	0.53	0.27
big	0	0.2	0	0.13	0	0	0	0	0.2	0
leftmost	0	0.83	0	0	0	0	0	0	0	0
rightmost	1.0	0	0	0	0	0	0	0	0	0
bottommost	0	0	0	0	0	0	0	0	0	0.83

Figure 3.18: The probabilities that the people, sampled for the survey, believe that the prepositions are satisfied by the object pairs in the columns

preposition	parameters
near	$\sigma = .1,\ \rho = 0.6$
far	$\sigma = 0,\ \rho = 0.3$
below, above	$\sigma = 0$
aligned	$\sigma = 0.2$
left, right	$\sigma = 0.1$
small	$\rho = 1.2$

Note that the value of σ for *above*, *below* and *far* was 0 because there were not enough cases when their values were between 0 and 1.

Another method of performing validation and calibration is to compare the system against situations where we may encounter images along with corresponding text, as in guidebooks. We can use the same images found in the guidebooks and have our system generate text and compare it to the text found in the guidebooks. The map in figure 3.20 was accompanied by text that described the position of the campsites found in a national reserve in Kenya [4]. Several examples included,

Keekorok Lodge is *between* Talek River and Sand River gates.

[4] Camping Guide to Kenya by David Else, Hunter Publishing, 1989

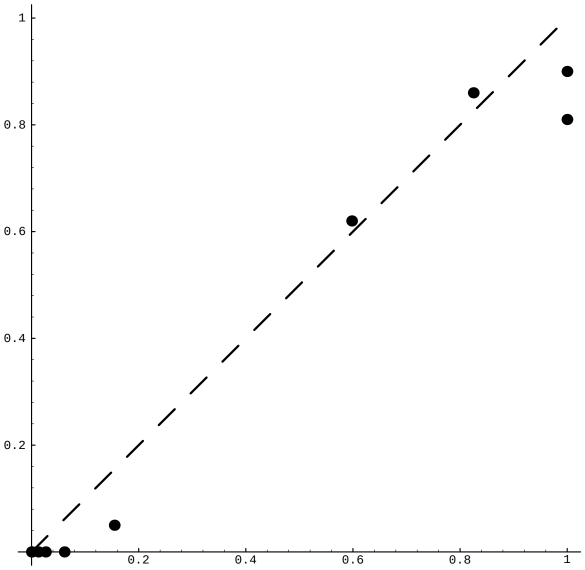


Figure 3.19: The quality of calibration for the preposition *near*. The horizontal axis represents the system's responses and the vertical axis represents the measured values.

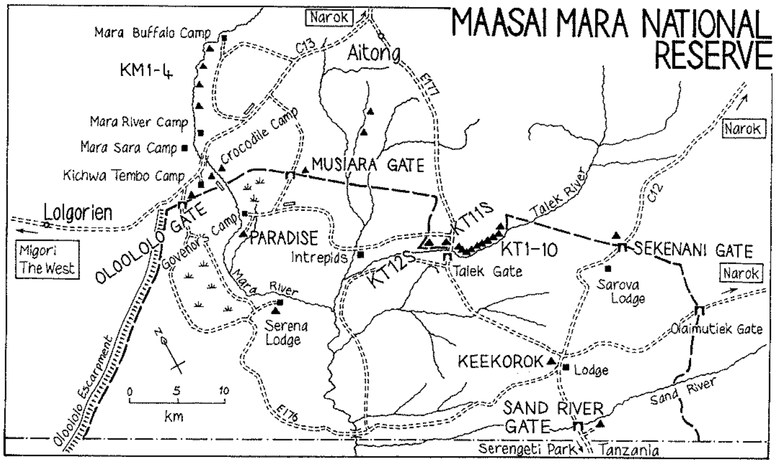


Figure 3.20: Guidebook maps can assist in validating and calibrating the semantic representations

Sekenami Gate Campsite is right *next* to Sekenani gate.

KM1-4 these four campsites are *along* the Mara River.

However, the map of figure 3.20 differs from the map of figure 1.8 and the maps that we will see in chapter 6. The map of Kenya is littered with rivers and roads that serve as landmarks as well as objects that are represented by points (or objects with very small area), which means that the descriptions the system will generate may not be appropriate for this particular map. The bounding box which is the basis of our computational model has the weakness that it does not approximate the shape of long narrow objects, such as rivers very well. In order to be able to generate descriptions that include rivers and roads we would need to devise a different way of approximating their shape that would entail a more complex computational model for the prepositions.

3.5 Conclusion

This chapter covered the methodologies encoded in the semantic representation module. The image processing module provides the information needed to determine the relationship between the figure object and the reference object, as well as a figure object's intrinsic properties, which in our case is size. Each preposition, superlative, and object property is encoded as a function within the semantic representation module. Except for the functions that determine object properties, they all take a reference object, figure object and a list of parameters. The object property functions only take the figure object and the list of parameters. The value that is returned by these functions is a measure of how well the objects satisfy the corresponding relation or property.

These computational models provide a way of quantifying the semantics of prepositions, superlatives, and features. The fuzzifying agent, that factors into the models, assists this quantification and aids in the capture of some of the vagueness associated with prepositions. For example, if our task is direction-giving and we know that the person we are giving directions to is traveling by car then we may incorporate this fact into the preposition parameterization.

We designed a survey that asked people to give their interpretation of what prepositions satisfied given object pairs from a theme park map. The results confirmed our preposition models and revealed some interesting results regarding object properties.

Since each of the relationships is a separate entity it is easy to add new language constructs; all it would entail is an extra function for each additional feature. For example, if we wanted to add color as an object property, all we would need to do is write a function that determines color based on some additional information provided by the image processing module and we would be able to proceed as before. We could apply the technique of fuzzification to the color functions as well.

The results produced by the semantic representation module consists of all the relationships between figure object and reference object for all the objects in the image. It will be the task of the locative expression generator to decide which of the relationships computed by the semantic representation module is actually necessary for an appropriate discussion of the environment.

Chapter 4

Locative Expression, Vagueness, and Minimization

In this chapter we will describe a method for determining the necessary and sufficient set of language constructs for describing objects in an image. The language constructs are from the sets $\mathcal{P}$, $\mathcal{S}$ and $\mathcal{A}$. The method is based on the Quine-McCluskey Boolean formula minimization algorithm. This method assumes that the user is a novice i.e. not familiar with the environment and hence needs to establish referent links that connect the environment in an informative way. The method uses *locative expressions* as a means of representing the spatial relationship of a reference object and a figure object. These locative expressions are manipulated to yield the best description for an object pair. The choice of description hinges on a method that defines how vague the description is. The less vague the better the description and the higher the probability that the user will select the intended figure object. We will define the vagueness measure through the use of a model of user behavior and inexactness.

4.1 The Locative Expression

A locative expression is a fuzzy predicate that represents the spatial relationship of an object pair. It is denoted $l(f, r)$. Recall that an object pair consists of a figure object f and a reference object r. For example, the locative expression $l(f, r)$ can be defined as $near(f, r) \wedge above(f, r) \wedge \neg aligned(f, r)$. It is readily convertible to its English prepositional phrase analogue. Depending on the fuzzy values of its component basic predicates, this can be translated as "f is rather near and above and not very aligned with r", or "f is not at all near, and somewhat above and not at all aligned with r". Formally a locative expression is the conjunctive normal form of fuzzy atomic formulae, each one of which corresponds to spatial relationships and object properties from a set of predicates for which computational models have been established to evaluate their truth value for a given object pair. These computational models are part of the semantic representation module, described in chapter 3. The spatial relationships correspond to prepositions, superlatives, and intrinsic object properties.

In deriving a locative expression that best describes a spatial relationship from the very many that are possible, we want it to be defined with only those predicates that are absolutely necessary for the description. We want a minimized locative expression that is a conjunction of some subset of all possible existing spatial relations. A locative expression of the form "f is *very near* and not *far* and *rather above* and not *very below* and not *aligned* and not *too inside* and to the *immediate left* and not to the *right* of r" is not very comprehensible.

4.2 Locative Expression Minimization

The first step in determining the best locative expression is the evaluation of all the possible spatial relationships involving all objects paired with a given reference object. That is, computational models compute and record the values of $p_j(f_i, r)$ for all fuzzy predicates p_j, and for all figure objects, f_i. Conceptually, this is a large two-dimensional matrix with $|\mathcal{P}|$ rows and $|\mathcal{O}| - 1$ columns, where $|\mathcal{P}|$ is the number of spatial predicates that can return fuzzy predicate values, and $|\mathcal{O}|$ is the number of objects in the image. Each cell in this two-dimensional array is filled with a value in the interval [0,1], where 0 represents complete falsehood, and 1 represents complete truth. For example, $p_7(f_3, r) = .5$ encodes that object f_3 (say, the Japanese Pavilion at Epcot Center) and object r (say, the International Gateway Pavilion) are only weakly related spatially via predicate p_7 (say, the relation "above"): in English, "The Japanese Pavilion is not really that much above the Canadian Pavilion". Given any specific object, for example, object f, its full spatial relationship to the reference object is given by the fuzzy predicate $p_1(f, r) \wedge p_2(f, r) \wedge \ldots \wedge p_{|\mathcal{P}|}(f, r)$, which has a long and nonsensical English translation. Nevertheless, since it is important, we will refer to this conjoined fuzzy predicate as[1] $p(f, r)$. Note that for any object, the form of the conjoined fuzzy predicate p is determined solely by the number and meaning of the component predicates; for any two objects a and b, $p(a, b) = p_1(a, b) \wedge p_2(a, b) \wedge \ldots \wedge p_{|\mathcal{P}|}(a, b)$. However, depending on the values returned by the computational models, the meaning of $p(a, b)$ will vary.

The second step is the minimization of this conjoined fuzzy predicate, given the context of all the other object descriptions which are themselves long conjoined fuzzy predicates. That is, we will convert p for object pair (f, r) to $\bar{l}$, where the change of letter and the addition of the bar indicate two fundamental transformation. The change letter indicates that some, even all, of the predicates p_i will be eliminated from the resulting predicate, and the bar indicates that all the fuzziness has been "crispened" to full truth values. For example, the incomprehensible example can be crispened to "f *near* and not *far* and *above* and not *below* and

[1] Note that in chapter 3 notation p referred to a single preposition but in this chapter it will refer to a conjunction of fuzzy prepositions.

not *aligned* and not *inside* and to the *left* and not to the *right* of r" and further minimized to "f *near* and not *below* r".

The minimization is accomplished by applying the Quine-McCluskey Boolean formula minimization algorithm on a modification of the two-dimensional array.

First, the matrix of fuzzy values is thresholded into values of 0, 1, and X, where X represents "don't care". In this work, we used the thresholds of 1/3 and 2/3 to partition the fuzzy values: if $p_j(f_i, r) < 1/3$, it is thresholded to 0, or pure false; if $p_j(f_i, r) > 2/3$, it is thresholded to 1, or pure true, and otherwise it is assigned the special symbol X, or "don't care". (Future research will determine how sensitive this algorithm is to these particular threshold values, and their relation to the σ of preposition calibration.) To distinguish this thresholded matrix from the matrix of fuzzy-valued predicates, we call these thresholded predicates $\bar{p}_j$, and the thresholded conjoined fuzzy predicates $\bar{p}$. Next, $\bar{p}$ for object pair (f, r) is minimized with respect to the context of all of the other $\bar{p}$ for object pairs (d_i, r), where d_i are the other objects in the image that we call distractors. We will drop the subscript on f that was being used previously and talk instead about the distractors d_i and the figure object f.

Our procedure applies the Quine-McCluskey Boolean formula minimization on $\bar{p}$ for object pair (f, r) to yield $\bar{l}$. We illustrate first by means of an example. When minimizing $\bar{p}$, we want to delete as many component predicates $\bar{p}_j$ as possible, while simultaneously ensuring that we maintain the resulting conjoined predicate's descriptive ability. We do not want the result of the minimization, to describe more object pairs than $\bar{p}$ did. It is possible for a second object, say d_1, to also be described using the same $\bar{l}$. For example, if object f has $\bar{p}(f, r) = \mathit{far}(f, r) \wedge \mathit{left}(f, r)$ and object d_1 has $\bar{p}(d_1, r) = \mathit{far}(d_1, r) \wedge \mathit{right}(d_1, r)$, then a valid minimization would have $\bar{l}(f, r) = \mathit{left}(f, r)$. But, $\bar{l}(f, r) = \mathit{far}(f, r)$ would be ambiguous, since both f and d_1 satisfy this predicate.

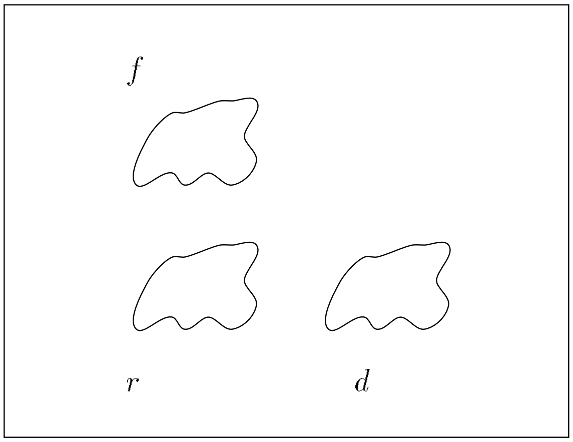


Figure 4.1: An example for the minimization process: $\bar{p}(f,r)$ is $near(f,r) \wedge above(f,r)$. $\bar{p}(d,r)$ is $near(d,r) \wedge above(d,r)$. Therefore when minimizing $\bar{p}(f,r)$ we want to eliminate the preposition *near*, since what differentiates the two object pairs is the preposition *above*. Formally, $\bar{l}(f,r) = above(f,r) = 1.0$ and $\bar{l}(d,r) = above(d,r) = 0.0$.

To illustrate, observe figure 4.1. Let's say $\bar{p}_1 = near$, and $\bar{p}_2 = above$. The conjoined predicate expressing the spatial relationship between figure object f and reference object r is $\bar{p}(f,r) = near(f,r) \wedge above(f,r)$; further, we assume that the computational models have returned the values $near(f,r) = 1.0$ and $above(f,r) = 1.0$. Likewise, the conjoined fuzzy predicate expressing the spatial relationship between object d and reference object r is again $\bar{p}(d,r) = near(d,r) \wedge above(d,r)$; however, we assume that the computational models have returned the values $near(d,r) = 1.0$ and $above(d,r) = 0.0$. Note again that in the two cases $\bar{p}$ is identical in form, but different in value.

If *above* is removed from $\bar{p}$, so that $\bar{l} = near$, this is not a good selection because now both $\bar{l}(f,r)$ and $\bar{l}(d_1,r)$ evaluate to the same value, since $near(f,r) = near(d_1,r) = 1.0$. However, if we make $\bar{l} = above$, we find that $\bar{l}(f,r) = above(f,r) = 1.0$ and $\bar{l}(d_1,r) = above(d_1,r) = 0.0$, which are readily distinguishable. Therefore the correct minimization is to remove *near*.

In the general case, the minimization procedure selects and removes predicates as long as the resulting $\bar{l}$ continues to be satisfied by the same number of objects that satisfy $\bar{p}$. Since their is some choice as to what predicate to delete, the algorithm may result in several $\bar{l}$ candidates, each one which describes the spatial relationship of figure object f to reference object r. If this is the case, then we select one of the $\bar{l}$s with the minimum vagueness (described in the next section). If two or more $\bar{l}$s have the same vagueness, we select the one with the minimal number of component $\bar{p}_j$s, preferring the $\bar{l}$ with the largest number of positive components. If there is a tie (equal vagueness, equal number of components, equal number of positive components) we select $\bar{l}$ randomly.

To compute these $\bar{l}$ candidates, we proceed with the following three steps. First, we form the set of all the descriptions $\bar{p}_j$s for the objects that we do not want to describe. Second, we find the complement of this set. Third, we apply the Quine-McCluskey Boolean formula minimization on this complement. The first step collects together those descriptions we wish to avoid in describing the figure object. The second step is all the descriptions that remain; each one is by definition not confusable with any of the objects we wish to avoid. The third step compresses these descriptions into their most economic form; as part of the minimization, the figure object is "covered" by this minimization, to yield the most compact, non-confusable, other object-avoiding description. Note that the minimization proceeds on the thresholded predicates, as the diagram in figure 4.2 indicates.

It is possible to also illustrate this avoid-and-minimize procedure by means of a Karnaugh map; we will use the example of figure 4.1. We place a 0 in the cell entry representing "*near* and not *above*" because this cell represents $\bar{p}$ for object pair (d, r) and we do not want $\bar{l}$ for object pair (f, r) to satisfy more object pairs than $\bar{p}(f, r)$ does.

A 1 is placed in the Karnaugh map shown in figure 4.3 for the cell entry representing "*near* and *above*" since this is the $\bar{p}$ for object pair (f, r) and we are looking to minimize this locative expression. The minimal expression, $\bar{l}(f, r)$ in

$$\begin{tikzcd}
p \arrow[r] \arrow[d, dashed, "\text{Direct}"'] & \bar{p} \arrow[d, "\text{Quine-McCluskey}"] \\
l & \bar{l} \arrow[l, "\text{Adverbs}"]
\end{tikzcd}$$

Figure 4.2: Quine-McCluskey Boolean formula minimization is performed on the thresholded $p(f, r)$, namely $\bar{p}(f, r)$ resulting in $\bar{l}(f, r)$. Adverbs may be added as part of a post-processing step and results in the fuzzy predicate $l(f, r)$.

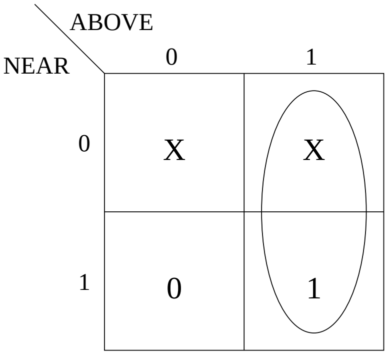


Figure 4.3: $\bar{l}(f, r)$ of figure 4.1 is *above*

Karnaugh map terms is the largest covering set of prime implicants that contain the 1 but avoid the 0s.

An X is placed in the remaining cell entries because these cell entries represent locative expressions that describe no objects from the image. Minimization is illustrated by the oval in figure 4.3. Minimization yields the $\bar{l}(f, r)$ aforementioned, namely *above*. This locative expression contains fewer predicates than $\bar{p}(f, r)$ and it does not describe other object pairs that $\bar{p}(f, r)$ didn't previously describe.

Adverbs Adverbs are used to recapture the fuzzification of the prepositions and object properties. The adverbs are selected as part of a post-processing stage. The fuzzy values $p_j(f, r)$ are retrieved for those prepositions that survived the

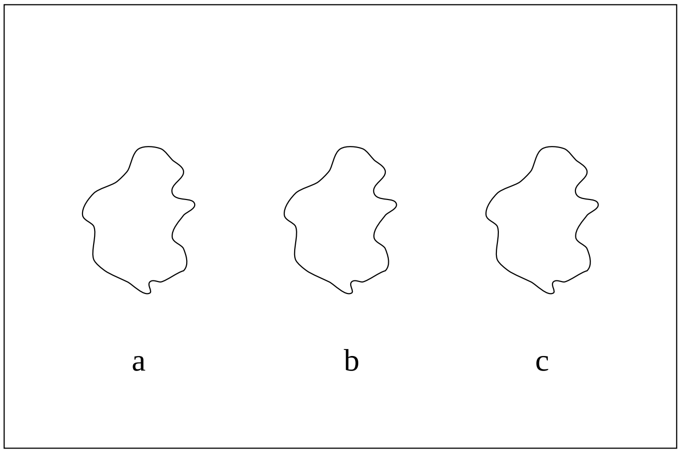


Figure 4.4: Description vagueness: the recipient of the description *near* has 50% probability of selecting the intended figure object, c, if the reference object is b

minimization procedure. The set of adverbs $\mathcal{A} = \{somewhat,\ very\}$. If $p_j(f, r)$ is greater than 0.9 then *very* is selected and a possible sentence would be "The Wonders of Life is very near the DNA statue". A $p_j(f, r)$ value between 0.67 and 0.75 would yield "The Chinese Pavilion is somewhat to the right of the Norwegian Pavilion." The interval of values assigned to each adverb, like the threshold value of 0.33 and 0.67, is the subject of future investigation.

4.3 Vagueness of a Locative Expression

Once we have found the minimal locative expression $l(f, r)$ we want to know how *good* it is, where *good* is a measure of how likely the recipient is going to locate the intended figure object. For example, in figure 4.4, if the intended figure object is a and the reference object is b and $l(a, b)$ is *near* then the recipient has a 50% probability of choosing the intended figure object.

The probability associated with the example in figure 4.4 is an obvious one, but there will be circumstances when the interpretation will vary from user to user. If the value of $l(f, r)$ is between 0 and 1 (e.g. $l(f, r) = 0.5$) then some leeway must be allowed for user differences. The survey experiments of section 3.4 revealed that situations when the objects did not satisfy a preposition exactly lead to more user disagreement. For this reason we introduced an error model.

Error model The error model provides a method for encoding user decision making and user error. The error model is flexible but we have chosen one that results exactly in the decision to define positives as more than 0.67 and negatives as less than 0.33. It assigns an interval $[l_{\min}, l_{\max}]$ for each possible value of $l(f, r)$, where $l_{\min}$ is the lowest possible value that a user would assign for a particular $l(f, r)$ and $l_{\max}$ is the largest possible value. It is expected that all users will assign values within this interval. Figure 4.5 graphically illustrates the error model that we chose. Other error models are possible. In general a "good" error model defines a small interval $[l_{\min}, l_{\max}]$ for $l(f, r)$ close to 0 and 1 and a wide interval for $l(f, r) = 0.5$, since this is where people disagree the most.

Handling distractors Distractors may lead to a misinterpretation of $l(f, r)$ resulting in the intended figure object not being found. This may happen if a user considers the value of $l(d, r)$ to be higher than $l(f, r)$ where d is a distractor. In general we consider the object with the highest value will be chosen.

Through the error model we define a measure of how likely the chosen object is f. The error model defines intervals $[l_{\min}, l_{\max}]$ for the intended figure object f and all the distractors $d_1, \ldots, d_n$. These intervals define an $(n+1)$-dimensional box $B \in [0, 1]^{n+1}$. Each point in B has $n+1$ coordinates, one which corresponds to the figure object and the remaining n correspond to the distractors. We consider that object f will be chosen over $d_1, \ldots, d_n$ with a likelihood $volume(B')/volume(B)$ where B' is a set of points from B that have a property that the coordinate that corresponds to the figure object is greater than the coordinate that correspond to all distractors.

We demonstrate this on an example with one distractor, $n = 1$. The box B is a

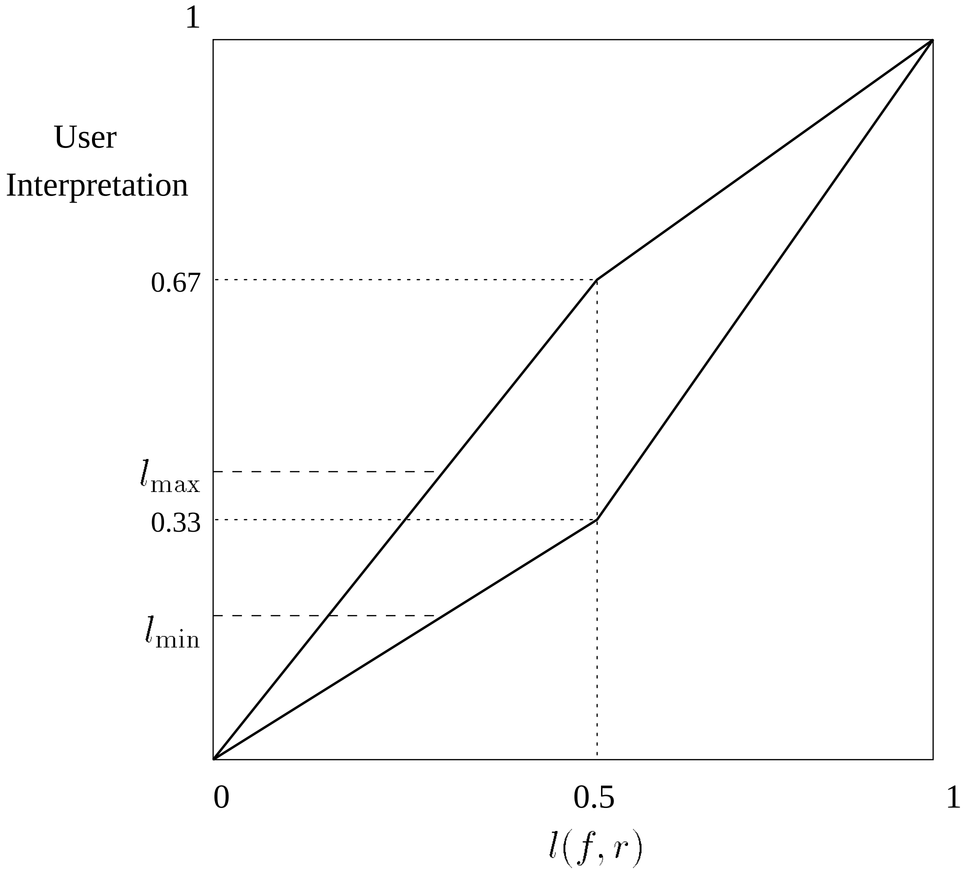


Figure 4.5: Error model

rectangle. The width of B is the interval defined by the error model for $l(f,r)$ and the height is the interval defined by the error model for $l(d,r)$. The set B' is the area of B below the diagonal (see figure 4.6). The likelihood that f will be chosen over d is $area(B')/area(B)$. Only in simple cases can $volume(B')/volume(B)$ be computed in closed form. The set B' can be complicated, an example with two distractors is given in figure 4.7. This is why we have resorted to estimating $volume(B')/volume(B)$ using a Monte-Carlo simulation.

Computing $volume(B')/volume(B)$ The value of $volume(B')/volume(B)$ is estimated using the following algorithm. We generate N $(n+1)$-dimensional random vectors uniformly distributed in B. We count how many of these vectors have the property that the coordinate that corresponds to the figure object is

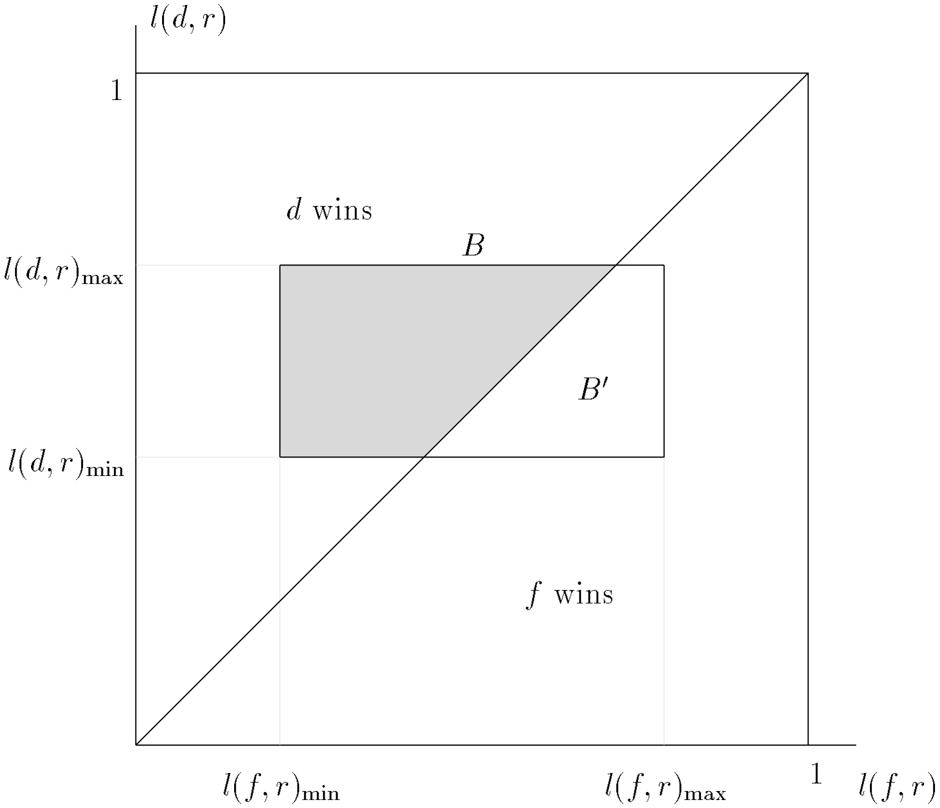


Figure 4.6: An example for determining the likelihood that f is chosen: 1 distractor

greater than all other coordinates (that correspond to distractors) of that vector. Let this number be N'. The likelihood that f will be chosen is approximately $volume(B')/volume(B) \approx N'/N$.

It may be the case that the coordinate that corresponds to f equals the coordinate of k distractors, where $0 < k \leq n$ and is greater than the coordinates of the other $n - k$ distractors. In the example of figure 4.6 this case corresponds to a point on the diagonal. In this case the count N' is incremented by $1/(1 + k)$.

Vagueness The quality of a locative expression is conveyed by a vagueness measure defined by

$$V(l, f, r) = -\log_2 volume(B')/volume(B)$$

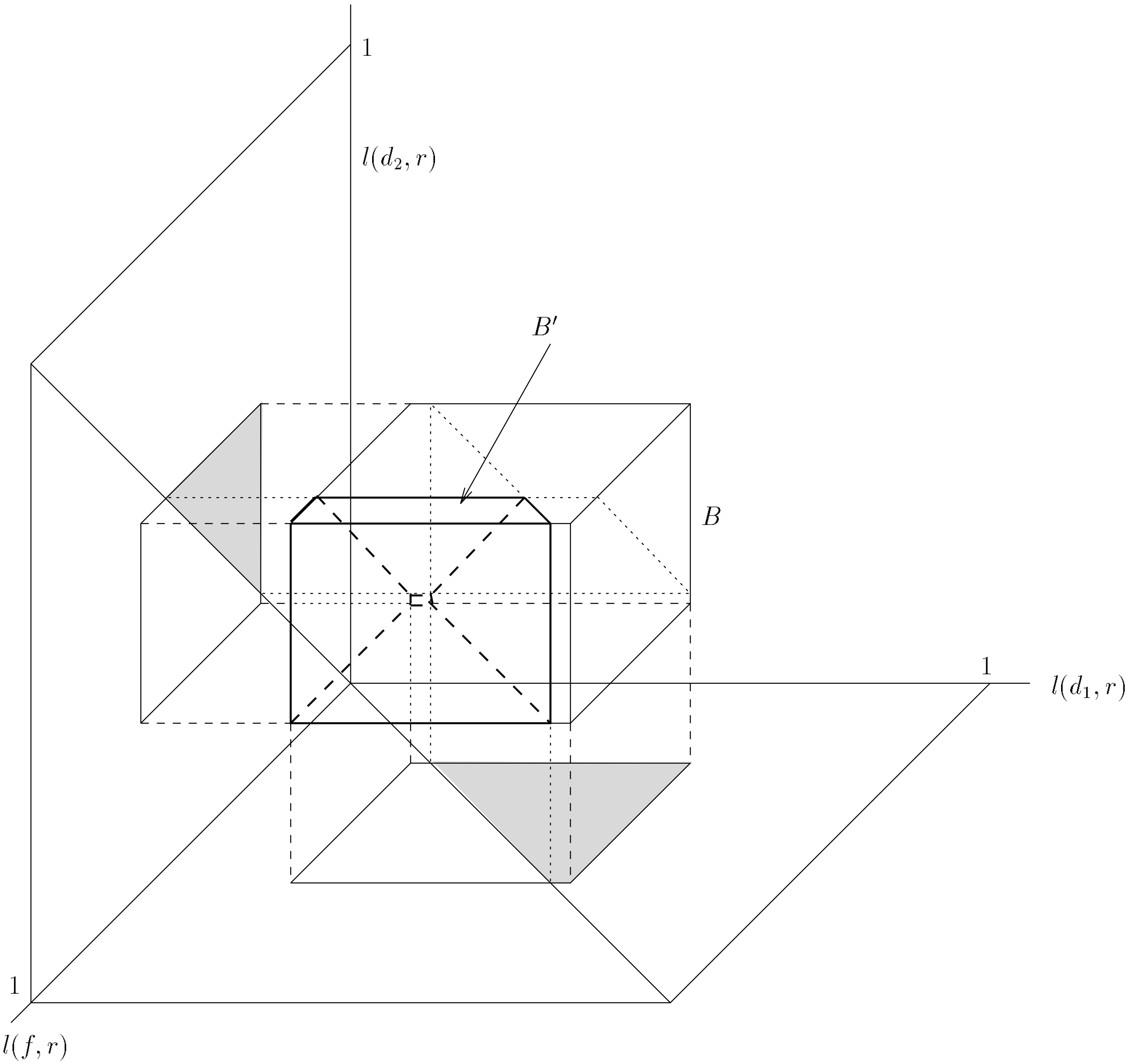


Figure 4.7: An example for determining the likelihood that f is chosen: 2 distractors

where B and B' are the sets of points described in the preceding paragraphs. "Good" locative expressions have low vagueness. The $-\log$ is useful in the landmark navigation task for computing vagueness along a path where we use Dijkstra's algorithm to search for the best path.

4.4 Superlatives

Like object size, superlatives can be used to increase the descriptive strength of a locative expression.

Global and local superlatives. Super-locative expressions We consider two types of superlatives – global and local. A global superlative is one that stands alone; it is true for an object pair in the image irrespective of the locative expressions that also may describe the object pair. For example in figure 4.8 if we wanted to describe object pair (c, r) it would suffice to say *nearest* because of all the objects in the image, c is the nearest to r. Note that for some superlatives no reference object is necessary, for example, an object that is *rightmost* is *rightmost* irrespective of other objects. *Nearest*, however, requires a reference object. A local superlative is true locally for a given locative expression. As an example, observe again figure 4.8. If we wanted to describe object b with respect to r we can not say it is *nearest* because this is simply not true, but we would like to say more than just that it is *to the left* of r since object a is also *to the left* of r. Examining the superlative list we see that we can say that b is *nearest to the left*, because of all the objects that are *to the left* it is *nearest*. Therefore we consider object b to be locally *nearest* to r.

In general we will call a combination of a superlative with a locative expression a *super-locative expression* denoted $s\bar{l}$. Observe that a global superlative is a super-locative expression with an "empty" locative expression. Also, a locative expression

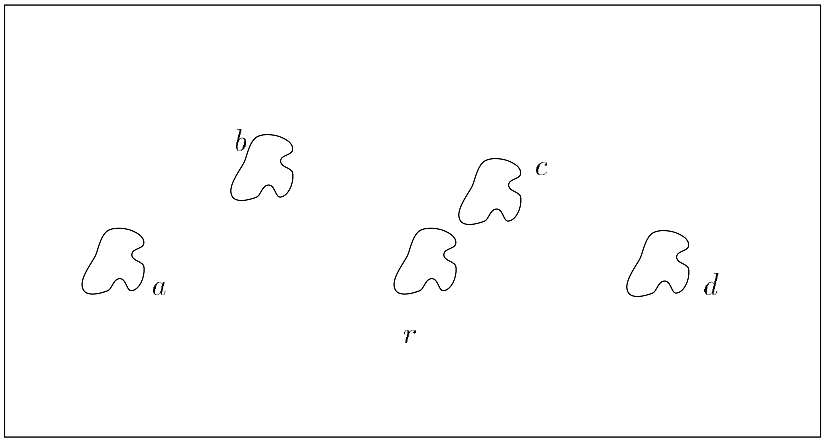


Figure 4.8: Global and local superlatives: object pair (c, r) is globally *nearest*, and object pair (b, r) is locally *nearest*, meaning that it must be combined with a locative expression to be valid. In this case it is *nearest to the left*

without a superlative is a super-locative expression that is a combination of a "null" superlative and a locative expression. A "null" superlative is by definition true for any pair of objects.

Two-step system The truth value of a preposition for object pair (f, r) is independent of the truth value of any other prepositions for object pair (f, r): for example the fact that two object are *near* does not influence the fact that they are *above*. This obvious property allowed us to combine prepositions into locative expressions: if f is *near* r and f is *above* r then we can say f *near and above* r. Conversely, if f is *near and above* r we know that f is both *near* r and f is *above* r. This is the foundation of locative expression minimization described in section 4.2.

Super-locative expressions, however, do not allow for the direct application of this minimization. This is because the superlative constituent of the super-locative expression is not necessarily true without the locative expression. For example, if f is the *nearest* object *above* r, while it is true that f is *above* r it is not necessarily the *nearest* (globally) object to r. This means that superlatives cannot be treated independently from prepositions.

This is why we separate the search for a super-locative expression into two steps. The first step is to look for the best locative expression for each superlative thus forming a set of super-locative expressions $\mathcal{L}$. The complexity of this step is $O(|\mathcal{S}|2^{|\mathcal{P}|})$. The second step is to look for the best subset of $\mathcal{L}$ using the Boolean formula minimization technique. The complexity of this step is $O(2^{|\mathcal{S}|})$. Since we have $|\mathcal{S}| < |\mathcal{P}|$ the complexity of this two-step method is $O(|\mathcal{S}|2^{|\mathcal{P}|})$. The problem of finding the minimum number of language constructs that describe f while describing as few or no distractors is NP-complete. This problem can be reduced to the Set Cover problem of finding the minimum number of subsets from a given collection, whose union is the whole set.

First step: Generating a set of super-locative expressions $\mathcal{L}$ To include superlatives the minimization technique described in section 4.2 needs modification. Recall that $\bar{p}$ for object pair (f, r) is minimized with respect to the context of all the other $\bar{p}$ for object pairs (d_i, r) where d_i are all the other objects in the image. When searching for a super-locative expression we minimize only with respect to the context of those objects d_i that satisfy a superlative better than f. If there are no other objects that satisfy the superlative better than f then the superlative alone is sufficient to describe the figure object. We will illustrate the algorithm on an example simple enough to be presented on a Karnaugh map.

Let us fill the Karnaugh map for the image depicted in figure 4.8 for two scenarios. In the first scenario let us use object c as the figure object and use a subset of prepositions and superlatives, $\mathcal{P} = \{\textit{left}, \textit{right}\}$ and $\mathcal{S} = \{\textit{nearest}, \textit{farthest}\}$. The conjoined predicate $\bar{p}$ for object pair (f, r) is "not *left* and *right*". Therefore, a 1 is placed in the corresponding cell entry in the map of figure 4.9. Next we examine which other objects are also true for this $\bar{p}$. In this example it is object d (e.g. it is *right* but not *left*) We then look through the set of superlatives checking which one is satisfied by (c, r) with respect to d. In other words, when checking to see if (c, r) satisfies s we compare c with only those objects that satisfy $\bar{p}$. We

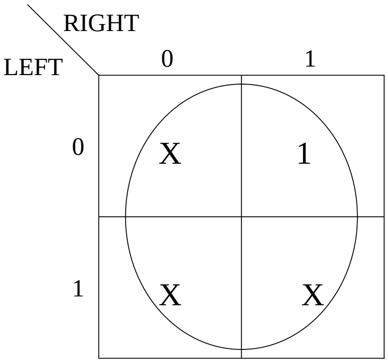
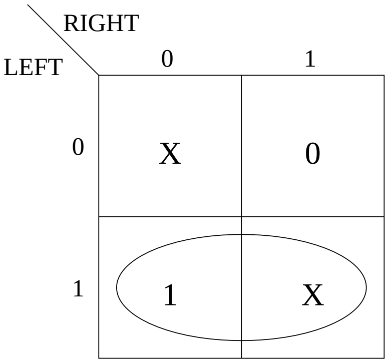


Figure 4.9: A global superlative: $\bar{l}$ = (X X), therefore (c, r) of figure 4.8 is best described using the superlative *nearest*.

do not need to look at those objects in the image that do not satisfy $\bar{p}$ since they do not influence the choice of the superlative. For our example, the superlatives that are satisfied by (c, r) with respect to $\{c, d\}$ are *nearest* and the null superlative ($\emptyset$) which by definition is satisfied by all object pairs. The next step is to fill the Karnaugh map with 0's and X's. If a distractor d_i satisfies s with respect to $\{d_i, f\}$ then a 0 is placed in the Karnaugh map that corresponds to $\bar{p}$ for object pair (d_i, r). This implies that the superlative s is not a global superlative for (f, r) If d_i does not satisfy s with respect to $\{d_i, f\}$ then an X is placed in the cell entry that corresponds to $\bar{p}$ for object pair (d_i, r). For the running example, there are no 0s in the Karnaugh Map because objects a, b and d do not satisfy *nearest* better than object c.

Minimization, for our example, eliminated all the prepositions. This result means that *nearest* is a global superlative for (c, r) and that is all that is needed to describe their spatial relationship.

In the second scenario we will step through the minimization algorithm using object b from figure 4.8 as the figure object. The conjoined predicate $\bar{p}$ for object pair (b, r) is "*left* and not right" . Object a is the only object besides b that is true for $\bar{p}$. Of the superlatives in $\mathcal{S}$, *nearest* is the only one that is satisfied by (b, r) with respect to $\{a, b\}$. Looking through all the objects in the image we find that



LEFT \ RIGHT	0	1
0	X	0
1	1	X

Figure 4.10: A local superlative: $\bar{l}(f, r) = (\mathrm{X}\ 1)$ since f is the *nearest* of all the objects to the *left* of r $\bar{l}(f, r)$ translates to *nearest to the left*.

c satisfies the superlative *nearest* with respect to $\{b, c\}$, therefore a 0 is placed in the Karnaugh map for the cell entry that corresponds to the locative expression $\bar{p}$ for object pair (c, r). The rest of the cell entries contain the "don't care" symbol. The minimization algorithm contains a heuristic that minimizes in such a way so as to yield results with the fewest negations. In the example, the algorithm returns $\bar{l}$ = *left* rather than returning *not right*. It is then added to $\mathcal{L}$ together with the superlative *nearest*.

Super-locative expressions for all superlatives are accumulated in the set $\mathcal{L}$. This set is passed to the next step described in the following paragraph.

Second step: Extracting the best element from $\mathcal{L}$ Each super-locative expression describes a set of objects. That set always contains the figure object, since the super-locative expression is constructed in a way that it is always true for the pair (f, r). However, it is possible that a super-locative expression is satisfied by some of the surrounding objects (distractors) as well. This is undesirable because the recipient of the description may misinterpret it and not choose the figure object. The goal of the second step, described in this paragraph, is to determine the best element from the set of super-locative expressions $\mathcal{L}$, that describes as few distractors as possible.

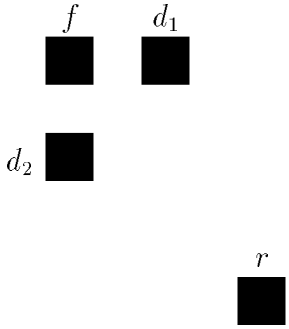


Figure 4.11: Combined superlatives: figure object f needs to be described using two superlatives: *topmost and leftmost*

We are tempted to assume that if the figure object is described using a superlative (e.g. "nearest object above you"), no confusion is possible since it would seem that only one object can satisfy any given superlative. This is however not always true, as shown in the example of figure 4.11. Here, the figure object f, whether described as *topmost* or *leftmost*, can be confused with one of the distractors, d_1 or d_2. A more precise description, *leftmost and topmost*, would eliminate the possibility of confusion. This also exemplifies the case of two superlatives in a single description.

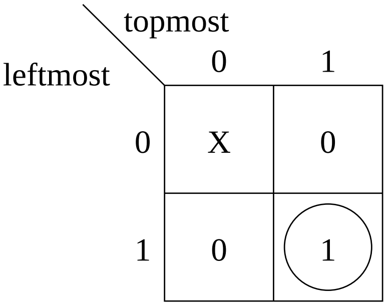
Examining the sets of objects that satisfy each of these two super-locative expressions, *leftmost* and *topmost*, reveals why it is necessary to combine them. The set that satisfies *topmost* is $\{f, d_1\}$ and the set that satisfies *leftmost* is $\{f, d_2\}$. The intersection of these two sets contains only f. It is possible to embed rules into the language generation preprocessor that could translate *leftmost topmost* into something like "upper left corner" since this may be more esthetically pleasing than combining two superlatives. However, it is rare to find sentences that contain more than one superlative. An informal survey conducted on the configuration of figure 4.11 revealed that people preferred to describe f with phrases like "upper left corner" rather than to combine two superlatives.

A super-locative expression may benefit from a combination with another super-locative expression only if the set of objects described by this super-locative expression contains other objects besides f, otherwise it is enough to use one superlative. This leads to partitioning of the set $\mathcal{L}$ into two subsets: $\mathcal{L}_1$, that contains all super-locative expressions that are satisfied only by f, and $\mathcal{L}_2$, that contains super-locative expressions that are satisfied by at least one object other than f. Either one of $\mathcal{L}_1$ and $\mathcal{L}_2$ could be empty. If $\mathcal{L}_2$ is empty and $\mathcal{L}_1$ contains more than one element then this two-step algorithm proceeds to the tie-breaker which is described in the next paragraph. If this is not the case, the strategy is to (1) extract the best combination of super-locative expressions from $\mathcal{L}_2$ (e.g. *topmost leftmost* from figure 4.11) and (2) apply a tie-breaking algorithm that selects the final description from the combination of super-locative expressions obtained in part (1) and set $\mathcal{L}_1$. The rest of this paragraph describes part (1). The next paragraph describes part (2).

In general, the algorithm for extracting the "best" subset of super-locative expressions is analogous to the algorithm for generating minimal locative expressions, with set $\mathcal{P}$ replaced with the set $\mathcal{L}_2$. Note that the "1" in the corresponding Karnaugh map will always be written in the cell that corresponds to all the elements of the super-locative expression being true. This is because we did not include the case of negated superlatives ("f is *not* the *nearest* object to r").

The Karnaugh map for the example in figure 4.11 is shown in figure 4.12. As mentioned above, the cell entry corresponding to "*topmost* and *leftmost*" contains 1 since f satisfies both super-locative expressions considered (*topmost* and *leftmost*). The cell that corresponds to "*not leftmost* and *topmost*" contains 0 because object d_1 is not *leftmost* but it is *topmost*, and cell "*leftmost* and *not topmost*" contains 0 because object d_2 is *leftmost* but not *topmost*. The remaining cell contains a "don't care" symbol. The minimization of this map results in the selection of both *topmost* and *leftmost*.

It is important to note that, due to the way it is constructed, the resulting combined super-locative expression will not contain expressions that involve negations



leftmost \ topmost	0	1
0	X	0
1	0	1

Figure 4.12: Karnaugh map for the example in figure 4.11

of superlatives (e.g. *not nearest*).

Let $S\bar{L} = (s_1\bar{l}_1,\ s_2\bar{l}_2, \ldots, s_k\bar{l}_k)$ be the combined super-locative expression obtained by the minimization process, and let $\mathcal{L}_1' = \mathcal{L}_1 \cup \{S\bar{L}\}$. We will call elements of this set simply *descriptions*. If the set $\mathcal{L}_1'$ has more than one element, we need to select one of them as our final description. This is described in the following paragraph.

Tie breaker: Vagueness and complexity In order to single out an element from the set of description candidates $\mathcal{L}_1'$ or $\mathcal{L}_1$ in the case where there are no elements in $\mathcal{L}_2$, we resort to a two level tie-breaking procedure. First, we look for those descriptions that have minimal vagueness. Second, among those that have minimal vagueness we select one that has minimal complexity (number of prepositions and superlatives involved). If there is more than one description with minimal vagueness and minimal complexity, we select one of them randomly.

The vagueness measure, defined in section 4.3, can be applied to super-locative expression directly once we define the way the truth values for super-locative expressions are computed. The truth value of a super-locative expression $s\bar{l}$ is equivalent to the truth value of the following statement: if an object d_i satisfies $\bar{l}$, then f satisfies superlative s with respect to all distractors d_i. This is because we need to examine the superlative only with respect to the set of objects that satisfy $\bar{l}$. The

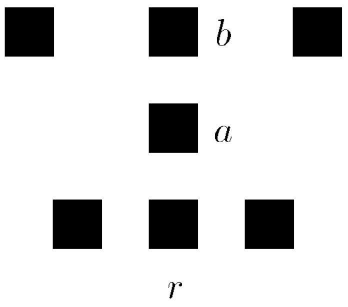


Figure 4.13: Example of a description that is not of the form $s\bar{l} \wedge s\bar{l} \ldots s\bar{l}$

"if … then" is an implication that can be replaced with "not" and "or" operators: $x \Rightarrow y \ \Leftrightarrow \ \neg x \vee y$. The final form for the truth value of a super-locative expression is (note that we dropped the bar from $\bar{l}$, to indicate the fuzzy truth value):

$$sl(f, r) = \bigwedge_{d \in \mathcal{O}, d \neq r, d \neq f} \neg l(d, r) \vee s(d, r, \{d, f\})$$

where notation $s(d, r, \{d, f\})$ stands for the value of a superlative s for object pair (d, r) with respect to the set $\{d, f\}$.

The fuzzy value of a combination of super-locative expression is the conjunction of fuzzy values of the constituents.

The complexity of a description is the total number of prepositions and superlatives in all super-locative expressions that constitute the description.

Limitations of the proposed method One limitation of the proposed "two-step" method is that once locative expressions are chosen in the first step they can no longer be simplified further as a result of what we learn in the second step. For example, combining two superlatives may render the locative expressions unnecessary.

This method does not allow for descriptions that are not of the form $s\bar{l} \wedge s\bar{l} \ldots \wedge s\bar{l}$. For example, we would not be able to generate a description of the form "the *nearest topmost* one", as would be needed in figure 4.13 to describe the object labeled b. The proposed algorithm would not be able to extract this combination of superlatives because f is not the *nearest* one to r and it is not the *topmost* one. The proposed method only combines superlatives when they satisfy the object pair (f, r). For this example, we do not know ahead of time that combining *nearest* and *topmost* would be able to distinguish b from its distractors even though b does not satisfy either of the two superlatives separately. Likewise it would not be able to describe the object labeled a as "the *topmost nearest* one". An enhancement to the proposed method would be needed and is the work of future research. It is possible that it is not practical to combine superlatives in such a way and that we would always find a different way of describing such a scenario that would not entail combining two or more superlatives.

4.5 Conclusion

The complexity of using all the language constructs that we have available led to a method based on Boolean formula minimization that selects only those language constructs that are necessary to describe the intended figure object. Determining which language constructs are necessary involves examining the surrounding context. A description needs to be chosen that has the highest probability of being chosen by the user, or, in other words, is not confusable with other objects in the image. The minimization technique together with the vagueness measure makes this possible. The addition of superlatives was made to enhance the overall description and in the process help increase the probability that the intended figure object is chosen.

We have assumed that the user of the system is a novice and the goal is to establish an understanding of the environment. This method assumes that the user knows nothing about the environment and hence needs a description that is as discriminating as possible. We will describe in chapter 5 a method for description

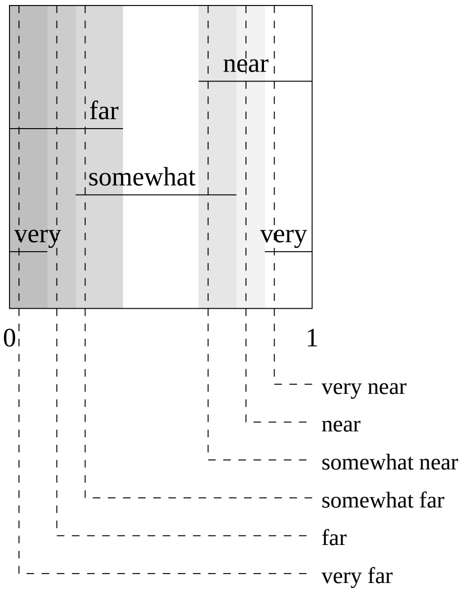


Figure 4.14: Levels of "nearness"

generation that assumes that the user is an expert, i.e. knows the location of most objects in an image. We will see how and why the methods differ.

Other Possibilities In this thesis we followed the path delineated in figure 4.2 by the solid line. Another path we could have taken is delineated by the dashed path in figure 4.2 which goes directly from p to l by a method we will call *direct fuzzy minimization.*

The idea behind direct fuzzy minimization is to create a predicate for each adverb and other language constructs and set intervals for which these adverbs and language constructs would apply. An example is given in figure 4.14. Here, we define the spectrum of "nearness" measures as *far*, *somewhat far*, *somewhat near*, *near* and *very near.*

This approach would, however, cause a significant increase in the number of prepositions which may cause the execution of the Boolean formula minimization algorithm to become impractically lengthy. A possible solution for this would be to consider an alternative minimization technique that would gather all prepositions (possibly “compound prepositions” such as *nearest above* as well) and apply a gradual buildup of the “best” description, starting from all descriptions that consist of a single preposition and examining their vagueness. If any of these descriptions has vagueness below a certain prescribed threshold, the procedure stops and returns that description. If, however, all “single-preposition” descriptions have unacceptably high vagueness, the combinations of two prepositions are examined next, and so forth. Note that in worst case this method has the same complexity as Boolean formula minimization, but in the average case the complexity is possibly lower, since descriptions tend to be short.

Another alternative to Boolean formula minimization algorithm would be to perform the fuzzy minimization described in [Kandel and Lee, 1979] for fuzzy logic circuits.

Chapter 5

Inference Network Minimization

In chapter 4 we were living in a novice world, where the goal was to establish a better understanding of the environment by generating descriptions of as many object pairs as necessary. In this chapter we will be living in an expert world where the goal is to generate as few descriptions as possible. An expert is very familiar with the environment and does not need all the descriptions he/she can gather. Instead, an expert requires filtering of all possible descriptions. Precisely how this filtering is done is the topic of this chapter. In this chapter the environment is an abdominal radiograph and the expert is a radiologist. The task is generating descriptions of stones found in radiographs. We want to insure that the descriptions are medically sound, which means that a radiologists reading it would not be able to distinguish it from descriptions generated by other radiologists.

5.1 Referent Refinement

Like the Boolean formula minimization method the first step in generating a description in the expert world is to compute the spatial relationship of the figure object to all the reference objects. The next step is to eliminate all unnecessary relationships between the figure object and reference objects. This is where the

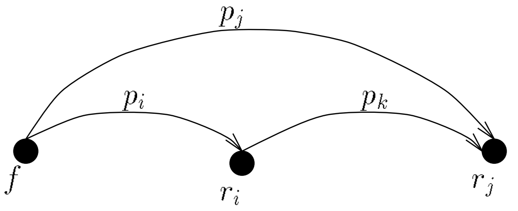


Figure 5.1: The inference rule: if a preposition p_k can be found that together with p_i implies preposition p_j then the relationship between reference object r_j and the figure object f may be eliminated from the final description.

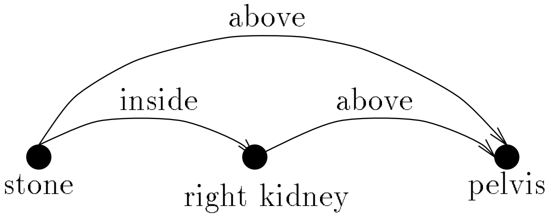


Figure 5.2: Graphical illustration of the inference rule: the two reference objects are the pelvis, and the right kidney, and the respective prepositions are *above* and *inside*

two methods differ. When eliminating these prepositions we need to look at a pair of reference objects r_i and r_j and the prepositions p_i and p_j that are true for (f, r_i) and (f, r_j). If we can find a preposition p_k that relates r_i and r_j such that the following condition holds

$$(\forall r_i, f, r_j)(p_k(r_i, r_j) \wedge p_i(f, r_i) \Longrightarrow p_j(f, r_j)) \tag{5.1}$$

then preposition p_j is redundant and we may eliminate it from the final description. Figure 5.1 graphically illustrates this condition. For example, suppose r_i is the right kidney, r_j is the pelvis, and the stone is *inside* the right kidney as illustrated in figure 5.2. In this case, p_i is *inside* and p_j is *above*. Choosing $p_k =$ *above* allows us to eliminate the fact that the stone is *above* the pelvis because the expert knows that the right kidney is *above* the pelvis and it is already known that the stone is *inside* the right kidney. Thus, it is not necessary to say that the stone is also *above* the pelvis.

The graph in figure 5.3 illustrates the inferences we use to eliminate as many relationships as possible. Note that these inferences are independent of the domain; they are provable based on the computational models from chapter 3. The proofs of these inference will be given in section 5.3. Each node represents a preposition that relates a reference object and a figure object. An edge between two nodes is labeled by the relationship between two reference objects that needs to hold in order to eliminate the preposition this edge is pointing to. The prepositions that we use in this chapter are *near, inside, above, below, left, right* because the other prepositions are not used in usual radiograph descriptions. Note that unlike the landmark navigation domain, the preposition *inside* will be used frequently. We can see from the graph that *inside* is an influential preposition – it is connected to all other prepositions. The explanation for this lies in the fact that *inside* is very localizing, the reference objects and figure object have to be in proximity. Also recall that in this domain the definitions of *above*, *below*, *left*, and *right* are *strictly above*, *strictly below*, *strictly left*, and *strictly right.*

In determining if we can eliminate a spatial relationship we follow the edges in the graph in the following manner. We begin at the node p_i that describes the relationship between figure object f and reference object r_i. If there exists an edge out of p_i that describes the relationship between r_i and r_j we follow that link to the next node p_j. If p_j describes the spatial relationship between f and r_j then we may eliminate it from the final description, because p_i and p_k imply p_j. For example let us follow the edge from *near* back to itself. We begin at the node labeled *near* if we know that f is *near* r_i. We follow the edge out of the node if r_i is *inside* r_j. If it is we may eliminate the fact that f is *near* r_j.

The dashed links in figure 5.3 represent weak links. Weak links do not satisfy condition 5.1, but a somewhat weaker condition which is that the *complement* of the preposition p_j is *not* satisfied:

$$(\forall r_i, f, r_j)(p_k(r_i, r_j) \wedge p_i(f, r_i) \Longrightarrow \neg - p_j(f, r_j))$$

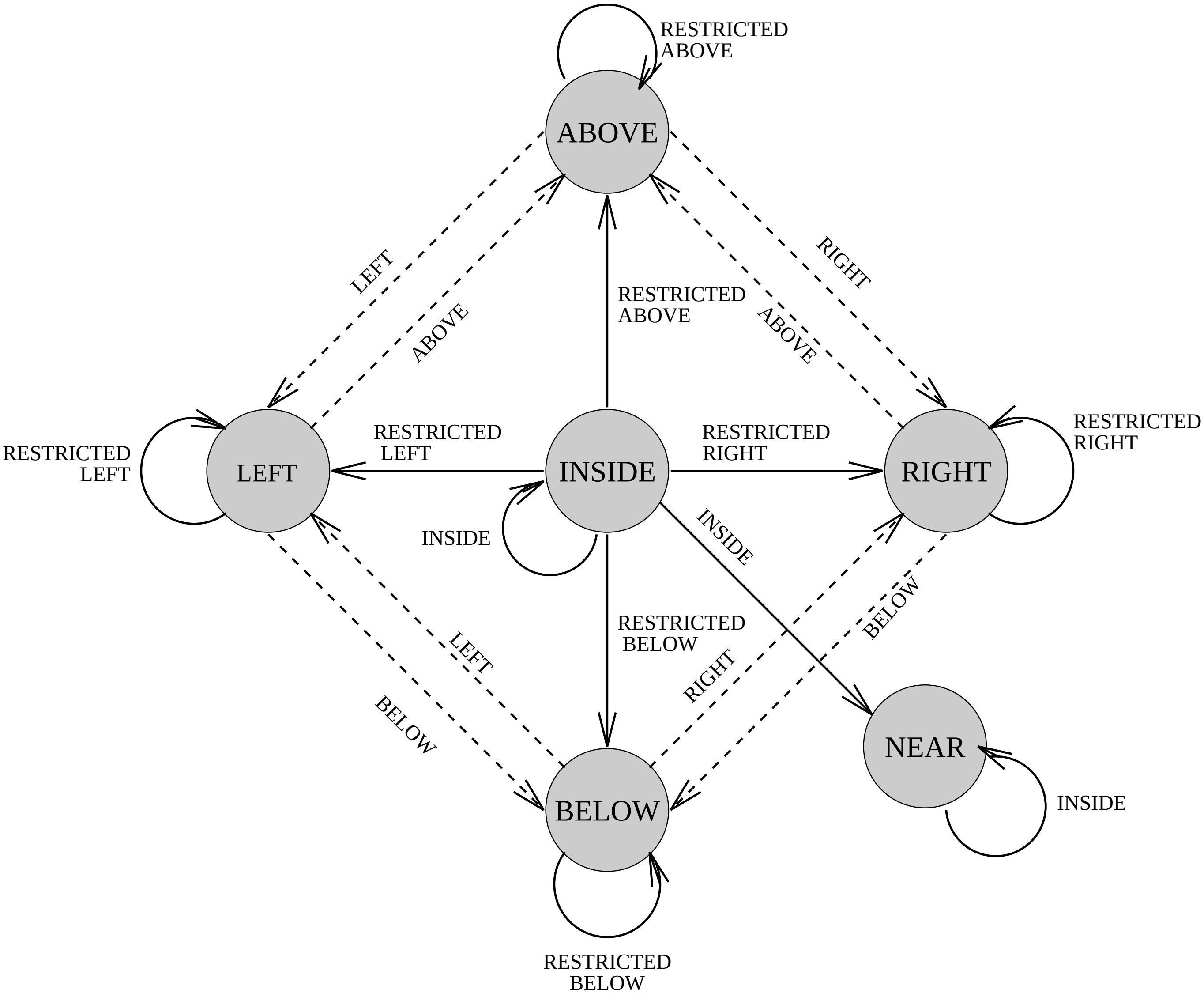


Figure 5.3: The inferences used to eliminate relationships from the final description. Nodes represent prepositions p for object pairs (f, r) and edge labels represent prepositions p for object pairs (r_i, r_j).

where $-p_j$ stands for the complement of p_j. Not all prepositions have complements. Obvious complement pairs are (*left*, *right*) and (*above*, *below*). An example is shown in figure 5.4 and figure 5.5. In figure 5.4 we know that f is to the *left* of r_1 therefore we begin at the node labeled *left* and since r_1 is *below* r_2 we follow the edge to the node labeled *below* and since f is also *below* r_2 we may eliminate this description since we know that it is certainly not *above* r_2, otherwise it could not be *left* of r_1.

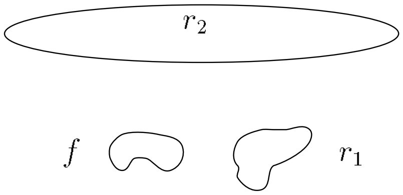


Figure 5.4: Weak inference: in this arrangement we may eliminate the description associated with reference object r_2 because we know that r_1 is *below* it and we know that f is to the *left* of r_1, therefore we know that f is not *above* r_2 (weak inference)

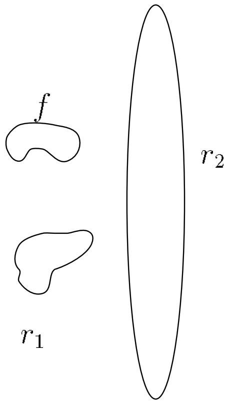


Figure 5.5: Weak inference: the edge from *above* to *left* is *left* because, according to the definitions of *above* and *left*, if r_1 is to the *left* of r_2 and f is *above* r_1 then if f is to the *left* of r_2 and we may eliminate the description associated with r_2 (weak inference)

5.2 Encoding inferences uses spanning trees

This section describes how to extract the minimal set of necessary descriptions based on the graph of figure 5.3. We begin with a set of all possible descriptions of a figure object with respect to all the reference objects. These descriptions can be thought of as a pair (p, r) where p is a preposition and r is a reference object, such that $p(f, r) = 1$. Not all of these descriptions are required to describe the figure object. The graph of figure 5.3 will enable us to detect the redundant descriptions. This graph defines a ternary relation $T(p_1, p_2, p_3)$. We say that prepositions p_1, p_2, p_3 are in relation T if the graph in figure 5.3 contains an edge pointing from p_1 to p_2 and labeled p_3. For example, $T(\mathit{inside}, \mathit{near}, \mathit{inside}) = 1$. Using this relation we will define a directed graph G whose nodes are all possible descriptions (p, r) of a figure object. Descriptions (p_1, r_1) and (p_2, r_2) are connected by an edge from (p_1, r_1) to (p_2, r_2) (denoted $(p_1, r_1) \rightarrow (p_2, r_2)$) if a preposition p_3 exists such that r_1 and r_2 are related by p_3 and p_1, p_2, and p_3 are related by T:

$$(p_1, r_1) \rightarrow (p_2, r_2) \Leftrightarrow (\exists p_3)(p_3(r_1, r_2) \wedge T(p_1, p_2, p_3))$$

For example if f is *inside* r_1 and *near* r_2 and if r_1 is *inside* r_2 then $(\textit{inside}, r_1) \rightarrow (\textit{near}, r_2)$ since $T(\textit{inside}, \textit{near}, \textit{inside})$.

If a node has an ancestor in G then the description conveyed by that ancestor would render the description conveyed by the node redundant. Therefore the minimal description consists of the roots of the spanning forest of G. In an acyclic graph the roots of a spanning forest cannot be inferred from any other nodes, however all other nodes can be inferred by them. If the graph contains cycles then the spanning forest is not unique and we must apply an ordering on the nodes. The ordering is based on the importance of the node. The importance of a node in the radiograph domain is determined by the expert. For example, a node that describes a relationship that involves the kidney is more important than one that involves the pelvis, therefore the descriptions that involve the kidney would be considered as candidates for the root of a spanning forest before the descriptions that contain the pelvis. According to the radiologists we collaborated with, the organs, ordered in decreasing order of importance, are the following:

- kidneys
- bladder
- ureters
- bony structures

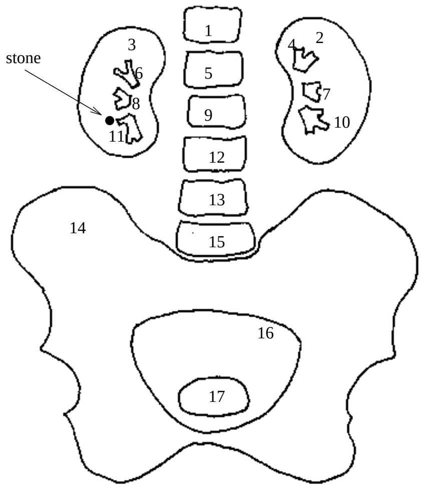


Figure 5.6: The model of the urinary system with a superimposed stone

An efficient algorithm ($O(\max(e, n))$, where e is the number edges and n is the number of nodes in G) exists for determining the roots of the spanning forest of a directed graph [Aho *et al.*, 1974].

Example We will present an example of the inference network minimization using the image in figure 5.6. This image represents a model of the urinary system with a stone found in the right[1] kidney. Each object in the model (both organs and bony structure, e.g. kidney and spinal cord) are numbered to simplify the presentation. The first step in retrieving the minimal set of descriptions is to compute all the spatial relation of the stone to all the objects in the model. The result is a list of stone descriptions with respect to the different objects. This list is `((below 8) (below 6) (left 11) (near 11) (left 10) (left 9) (left 12) (inside 3) (near 3) (left 2) (above 16) (above 14))`

From these descriptions and the graph of figure 5.3 we form the implication graph shown in figure 5.7. For example, an edge from the node labeled `below 8`

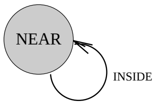
[1] The right kidney appears left in the image.

to the node labeled `below 6` exists because the stone is *below* 8 and *below* 6 and according to the graph this edge may be drawn if 8 is *below* 6 which it is. In other words, the fact that the stone is *below* 8 and that 8 is *below* 6 makes saying that the stone is *below* 6 unnecessary. The other edges are generated in this same manner. This figure also shows the spanning forest of this graph whose roots are the minimal description for the stone: `((below 8) (near 11) (inside 3))`. We will see in chapter 7 that this minimal description translate to the sentence "The right lower quadrant contains a density which may represent a stone in the lower pole calyx."

Note that in a case of a cyclic graph, it is necessary to compute the spanning forest rather than just search for nodes that do not have any ancestors.

5.3 Proofs of Inferences

In this section we will algebraically prove the correctness of the graph in figure 5.3, based on the definitions of the prepositions found in chapter 3. It is important to note that we will be using *strictly above*, *strictly below*, *strictly left* and *strictly right*, that require the conditions 3.3 and 3.4.

Proof of inference 

We will prove that

$$a \ near \ b \wedge b \ inside \ c \Longrightarrow a \ near \ c$$

Figure 5.8 illustrates this scenario. The definition of *near* (page 50) is

$$|x_a - x_b| - (W_a + W_b) \leq 0 \tag{5.2}$$

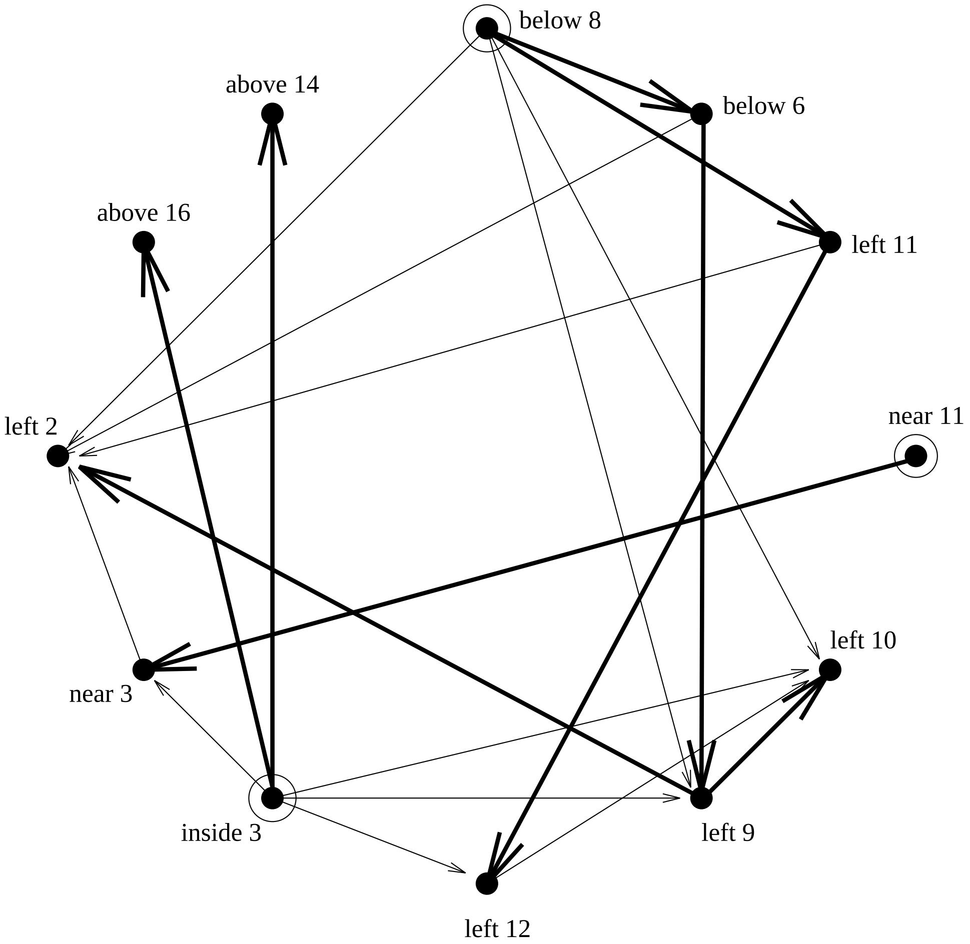


Figure 5.7: The graph G and its spanning forest (thick edges). The roots of the spanning forest are circled.

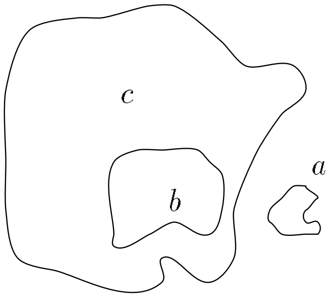


Figure 5.8: An example that b *inside* c and a *near* b implies a *near* c

and

$$|y_a - y_b| - (H_a + H_b) \leq 0$$

The definition of *inside* (page 53) is

$$|x_b - x_c| - (w_c - w_b) \leq 0 \tag{5.3}$$

and

$$|y_b - y_c| - (h_c - h_b) \leq 0$$

We will show that the implication is true on just the x component of the definition since the proof is analogous for the y component. Since $W = w + \rho h$, relation 5.2 becomes

$$|x_a - x_b| - (w_a + \rho h_a + w_b + \rho h_b) \leq 0 \tag{5.4}$$

Adding 5.4 and 5.3 yields

$$|x_a - x_b| + |x_b - x_c| - (w_a + w_c + \rho(h_a + h_b)) \leq 0 \tag{5.5}$$

By the triangle inequality $|x_a - x_c| \leq |x_a - x_b| + |x_b - x_c|$, so that 5.5 simplifies to

$$|x_a - x_c| - (w_a + w_c + \rho(h_a + h_b)) \leq 0 \tag{5.6}$$

Now we observe from 5.3 that $w_c \geq w_b$. Inserting this into 5.6 results in

$$|x_a - x_c| - (w_a + w_c + \rho(h_a + h_c)) \leq 0$$

or

$$|x_a - x_c| - (W_a + W_c) \leq 0$$

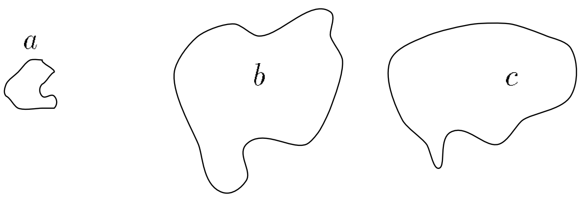


Figure 5.9: *Near* is not transitive

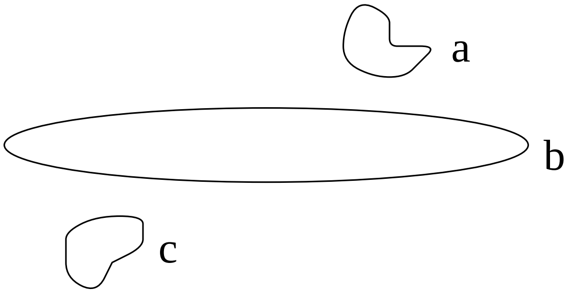
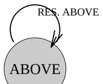


Figure 5.10: *Above* is not transitive: a is *above* b, b is *above* c, but a is not *above* c

This is precisely the definition of *near* (in the x direction) for objects a and c. Although we have just shown that a *near* b and b *inside* c implies a *near* c, figure 5.9 graphically illustrates why the edge between *near* and itself is not *near*, as we may have been tempted to think. Simply because a is *near* b and b is *near* c does not imply that a is also *near* c.

Proof of inference 

The preposition *above* as defined on page 55 is not transitive, meaning that if object a is *above* b and b is *above* c this does not imply that a is *above* c (see figure 5.10). This is the case because of the constraint imposed on *above* in the x direction (namely that the two objects overlap in the x direction). However, we can use *restricted above* which if satisfied by b and c, together with a *above* b, implies that a is *above* c:

$$a\ above\ b \wedge b\ restricted\ above\ c \Longrightarrow a\ above\ c \tag{5.7}$$

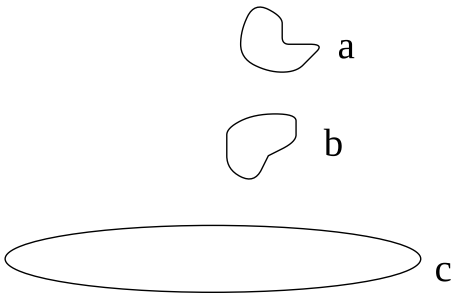


Figure 5.11: An example of *restricted above*: c entirely covers b's extent in the x direction therefore we may say that b is *restricted above* c

Recall that we consider b to be *restricted above* c if b is *above* c and c entirely covers b's extent in the x direction (see figure 5.11). Algebraically b *restricted above* c can be expressed by

$$|x_b - x_c| + w_b - w_c \leq 0 \tag{5.8}$$

and

$$h_b + h_c - (y_b - y_c) \leq 0 \tag{5.9}$$

We proceed by proving the implication 5.7. From the definition of *above* (page 55) we have that a *above* b is

$$|x_a - x_b| - (w_a + w_b) \leq 0 \tag{5.10}$$

$$h_a + h_b - (y_a - y_b) \leq 0 \tag{5.11}$$

Adding the inequalities 5.10 and 5.8 yields

$$|x_a - x_b| + |x_b - x_c| - (w_a + w_c) \leq 0 \tag{5.12}$$

The triangle inequality reduces 5.12 to

$$|x_a - x_c| - (w_a + w_c) \leq 0 \tag{5.13}$$

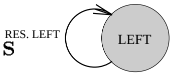
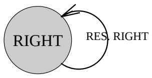
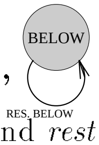
Adding 5.11 and 5.9 results in

$$h_a + 2h_b + h_c - (y_a - y_c) \leq 0 \tag{5.14}$$

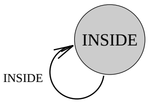
Observing that $h_b \geq 0$ inequality 5.14 becomes

$$h_a + h_c - (y_a - y_c) \leq 0 \tag{5.15}$$

Inequalities 5.13 and 5.15 represent the definition of *a above c*. Thus we have proven the implication 5.7.

Proof of inferences , , 

Prepositions *restricted left*, *restricted right* and *restricted below* are defined analogously to *restricted above*. The corresponding implications are proven the same way as the implication 5.7.

Proof of inference 

We want to prove that prepositions *inside* is transitive: if object a is *inside* b and b is *inside* c then that implies that a is *inside* c. We will show that proof only for the x coordinate since the proof for the y coordinate is the same. From the definition of *inside* page 53 we have that

$$|x_a - x_b| - (w_b - w_a) \leq 0 \tag{5.16}$$

and

$$|x_b - x_c| - (w_c - w_b) \leq 0 \tag{5.17}$$

Adding 5.16 and 5.17 we get

$$|x_a - x_b| + |x_b - x_c| - (w_c - w_a) \leq 0$$

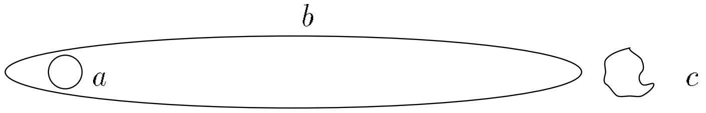


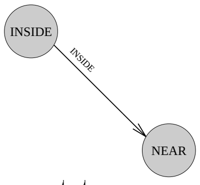
Figure 5.12: The edge from node *inside* to node *near* is not *near*

or using the triangle inequality,

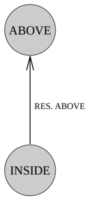
$$|x_a - x_c| - (w_c - w_a) \leq 0$$

This last inequality is the x coordinate part of a *inside* c.

Proof of inference



In this paragraph we want to prove that a *inside* b and b *inside* c implies a *near* c. For two objects to be *near* it is required that their bounding boxes intersect. For one object to be *inside* another requires that the bounding box of one object be embedded in the bounding box of the other. Therefore, if object a is *inside* object c then it must also be *near* object c. Based on the proof that *inside* is transitive we conclude that a is *inside* c and consequently *near* c. This means that a *inside* b and b *inside* c implies a *inside* c. Figure 5.12 graphically illustrates why the link from node *inside* to node *near* is not *near*. Saying that the a is *inside* b and b is *near* c does not say anything about the relationship between a and c. Reference object c may be anywhere in the vicinity of b without being in a single relationship with a. In figure 5.12 c is *far* from a. But another possible scenario would have c *above* a while still being *near* r_1.



Proof of inference

In this paragraph we prove that

$$a \ \mathit{inside} \ b \wedge b \ \mathit{restricted} \ \ \mathit{above} \ c \Longrightarrow a \ \mathit{above} \ c \tag{5.18}$$

We have *a inside b* is defined by

$$|x_a - x_b| - (w_b - w_a) \leq 0 \tag{5.19}$$

$$|y_a - y_b| - (h_b - h_a) \leq 0 \tag{5.20}$$

and *b restricted above c* is defined by

$$|x_b - x_c| + w_b - w_c \leq 0 \tag{5.21}$$

$$h_b + h_c - (y_b - y_c) \leq 0 \tag{5.22}$$

Adding 5.19 and 5.21 yields

$$|x_a - x_b| + |x_b - x_c| + w_a - w_c \leq 0$$

which together with the triangle inequality produces

$$|x_a - x_c| + w_a - w_c \leq 0 \tag{5.23}$$

Next, we add 5.20 and 5.22 and get

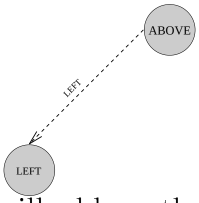
$$|y_a - y_b| - (y_b - y_c) + h_a + h_c \leq 0 \tag{5.24}$$

Since $|y_a - y_b| - y_b \geq -y_a$ we have that 5.24 implies

$$h_a + h_c - (y_a - y_c) \leq 0 \tag{5.25}$$

Inequalities 5.23 and 5.25 represent the definition of *a restricted above c*. Since *a restricted above c* $\Longrightarrow$ *a above c* the implication 5.18 is true.

Proof of inference



In this paragraph we will address the case of weak inferences denoted by the dashed edges in the graph in figure 5.3. A weak inference is categorized by the fact that we cannot prove the preposition that the weak inference is pointing to but we can prove that the complement of the preposition is not satisfied. (e.g. the complement of *left* is *right*). We interpret this inference as

$$a \ above \ b \wedge b \ left \ c \Longrightarrow \neg(a \ right \ c) \tag{5.26}$$

We will proceed with proving the implication 5.26. The condition that a is *above* b in the x direction is

$$|x_a - x_b| - (w_a + w_b) \leq 0 \tag{5.27}$$

The condition that b is to the *left* of c in the x direction is

$$w_b + w_c - (x_c - x_b) \leq 0 \tag{5.28}$$

Adding 5.27 and 5.28 yields

$$|x_a - x_b| - w_a + w_c - x_c + x_b \leq 0$$

This last inequality is equivalent to

$$|x_a - x_b| - (x_a - x_b) + 2w_c \leq w_c + w_a - (x_a - x_c) \tag{5.29}$$

where we add $w_c + w_a - (x_a - x_c)$ to both sides of the inequality. The left side of 5.29 is positive because $w_c > 0$ and $|x_a - x_b| \geq x_a - x_b$. This means that the right side in 5.29 is positive,

$$w_c + w_a - (x_a - x_c) > 0$$

which is the negation of a is to the *right* of c in the x direction. Thus a is not to the *right* of c.

The remaining weak inference are proven in an analogous way.

5.4 Conclusion

We have just described what is required to generate the minimal set of descriptions if we live in an expert world. For the set of descriptions to be useful to an expert the minimization method, that selects the smallest set of descriptions, must exploit the expert's knowledge about the environment. For the radiography domain this knowledge included the spatial relationship of the organs and bony structure. It is not necessary to say that the stone is *above* the pelvis if we know that the stone is superimposed on one of the lumbar vertebra since the expert knows that the lumbar vertebra are *above* the pelvis. It is presumed that the radiologist knows where the significant parts of the radiograph are.

It is possible that we may have an expert in the landmark navigation domain, such as a person who works say at EPCOT Center. Describing a new attraction to this person would be equivalent to describing a stone to a radiologist. We would use the inference network minimization method in this case. We may also find that we are faced with a novice in the radiography domain – like a beginning student[2]. Such a person would not know much about the environment and they would benefit most by the result that the Boolean formula minimization method would yield. In summary, both methods may be used in both domains, the choice as to which one to choose depends on the intended recipient of the descriptions, whether he/she is a novice or an expert.

The two worlds can be described as not only differing on the novice/expert scale, but also in terms of the goal. The novice world has as its goal to establish referent links (via the use of Dijkstra's algorithm on the complete graph created using the vagueness measure) and the expert world's goal is to filter referent links via the spatial implications graph created using the inference network. The two worlds establish two ends of a continuum from zero referents (e.g. the use of *topmost* does not require a referent) to many referents, with middle ground possible (i.e. a few referents). This middle ground is analogous to people indicating on a map things they already know (Spaceship Earth and the lagoon; spine and pelvis). The processing under these circumstances combines the two methods: using known referents and filtering them (objects are described only in terms of the most *salient* known object, that is, the one with the most implications) and creating as needed intermediate objects.

A lot of we have discussed in terms of selecting the minimal set of descriptions could be precompiled. However, in the middle ground, we have to resort to on-line computation. In the novice world, each figure object can be anticipated in the context of a single referent and its description precompiled without user interaction. In the expert world, the inference network can be precompiled and all the equivalence classes of locations of the figure object can be anticipated too; the X-ray can be broken into regions having the same description, so that when a stone is found at (x, y), the appropriate description can be looked up in the table.

[2] Or a person like myself when I took on this assignment

The middle ground is tricky since the user must indicate what subset of objects he/she considers to be referents; this is a 2^n subset problem, and although theoretically possible, computationally infeasible. Thus, in this middle ground, description must be truly interactive, since the computational complexity of anticipating user knowledge is infeasible. At either extreme no questions need to be asked of the user. But in the middle a possible session would go: “What do you want to describe? What do you already know?”, and the description must be created on the fly. It must be created on the fly because it is impossible to generate a proper response if the descriptions will be based on the user's knowledge context as well as the spatial context. Although this thesis does not address this “middle ground” scenario, it develops methods on each end and provides the basis for their integration.

The middle ground is not the only place where descriptions must be created on the fly. Descriptions must be created on the fly whenever an image is not known in advance. This is true for example in weather forecasting, where weather patterns are changing frequently and information is needed at the moment of receiving the image. This is also another possible application field of our system.

Chapter 6

Landmark Navigation

In this chapter we will illustrate how the concepts of chapters 3 and 4 are utilized for the task of landmark navigation. Recall that the goal of the landmark navigation task is to generate path descriptions for getting from a reference object (start location) to a figure object (goal location). In generating these path descriptions we have access to the entire map. The path descriptions are not generated as if a person were standing at a location and moving within the map, but rather looking down on the map. The path should be the "best" possible path, according to some criterion. We will describe what it means for a path to be the "best". We have chosen to illustrate the landmark navigation task on maps of theme parks.

6.1 Path Vagueness

The concept of the vagueness of a locative expression plays a crucial role in the landmark navigation task. We want to insure that the description that is generated is as good as possible, meaning that it has the highest probability of being interpreted correctly. In other words, if a person were to use that description he/she would arrive at the goal location without having deviated off course. In terms of the concepts we have defined earlier this means that we want to generate a description that yields the smallest possible vagueness. This may require that we describe an *intermediate* object or set of objects. A set of intermediate objects, together with the reference object r and the figure object f, defines a path. If the total vagueness along the path is less than the vagueness associated with describing f directly then the system chooses to describe the objects along the path.

We define the vagueness of a path using the vagueness of a single locative expression. A path π is a series of objects, $\pi = (a_0, \ldots, a_n)$ where $a_0 = r$ and $a_n = f$. The vagueness of a path is defined as

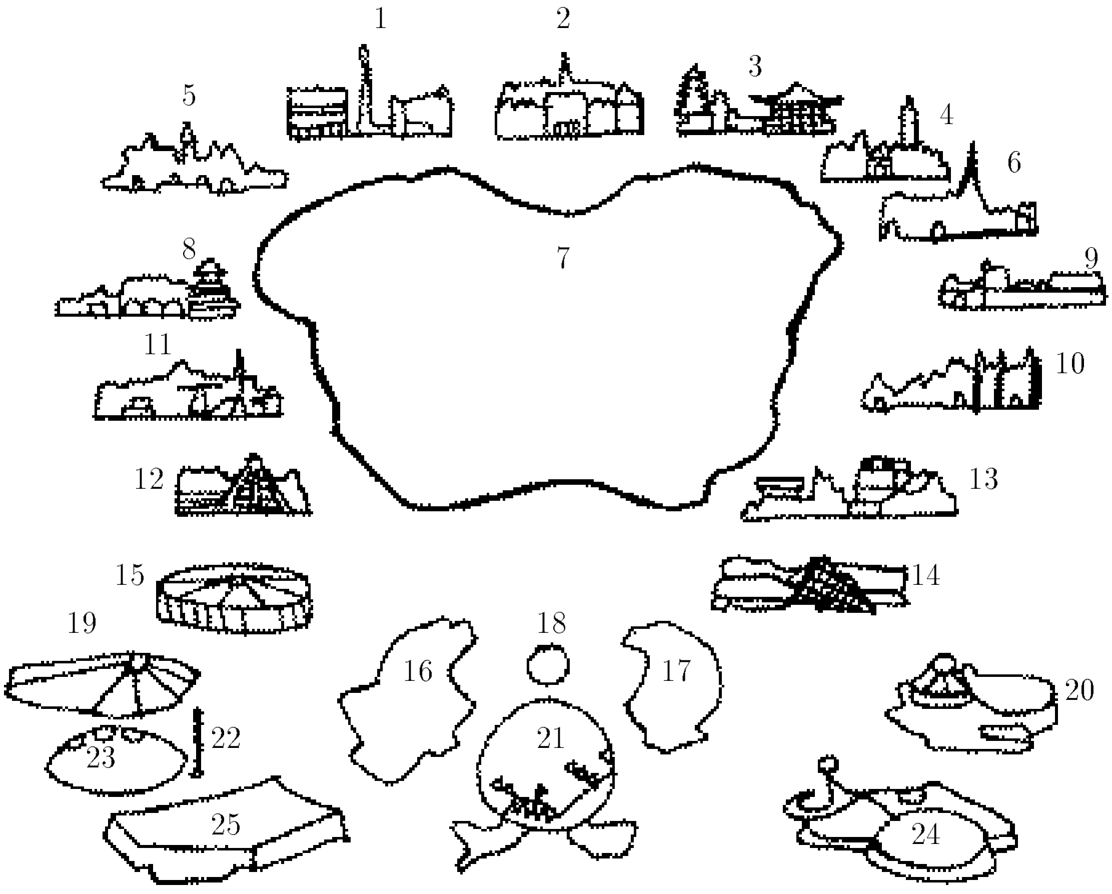
$$V(\pi) = \sum_{j=1}^{n} V(l_{j-1,j}, a_{j-1}, a_j)$$

where $l_{j-1,j}$ is the locative expression for the pair (a_{j-1}, a_j).

The vagueness $V(\pi)$ is needed in order to choose a chain of locative expressions that will more assuredly get us from r to f. The problem of finding a path with minimal vagueness is in the category of single-source shortest path problems and is efficiently solved by Dijkstra's algorithm [Aho *et al.*, 1974].

As an example of how $V(\pi)$ affects the outcome of the path description generated, observe figure 6.1. The intent in this example is to describe the goal location, $f = a_{12}$ (the subscript corresponds to the numbered objects in the figure), from the start location, $r = a_5$. The vagueness computed for (f, r) is $V(a_5, a_{12}) = 1.61$. This value is large enough to warrant the search for a better path – one with lower vagueness. The system finds that a better path is $\pi = (a_5, a_8, a_{11}, a_{12})$. Figure 6.3 illustrates the vagueness values associated with each object pair along the path from a_5 to a_{12}. Notice that a possible path may have been $\pi = (a_5, a_{11}, a_{12})$, but the combined vagueness value along this path was larger than the combined vagueness value of $\pi = (a_5, a_8, a_{11}, a_{12})$.

Table 6.2 contains l and V for each object pair along the path. From the graph



1. Italian Pavilion
2. American Adventure Pavilion
3. Japanese Pavilion
4. Moroccon Pavilion
5. German Pavilion
6. French Pavilion
7. lake
8. Chinese Pavilion
9. International Gateway Pavilion
10. United Kingdom Pavilion
11. Norwegian Pavilion
12. Mexican Pavilion
13. Canadian Pavilion
14. Journey Into Imagination
15. World Of Motion
16. Communicore East
17. Communicore West
18. fountain
19. Horizons
20. The Land
21. Spaceship Earth
22. DNA statue
23. Wonders Of Life
24. The Living Seas
25. Universe Of Energy

Figure 6.1: Theme park map of Walt Disney World's EPCOT center.

a_i	a_j	superlative	V	probability	l
5	8	leftmost	0.15	90%	near
8	11	leftmost	0.15	90%	next
11	12	topmost	0.22	86%	below
5	12		1.61	33%	medium, far, below, between

Figure 6.2: The superlative, vagueness value, probability of getting to f and l for each of the object pairs discussed in the image of figure 6.1

of figure 6.3 and the table of figure 6.2 we can see that $V(a_5, a_8, a_{11}, a_{12}) = 0.51$ which is significantly lower than $V(a_5, a_{12})$. The locative expressions for each of the object pairs along the path are also simpler than l for (a_5, a_{12}). This helps to explain why $V(a_5, a_{12})$ is so large. There are various objects in figure 6.1 that also satisfy this description.

As another example say that $r = a_1$ and the intended f is a_2. The system searches for a path with minimum vagueness and finds that the best path is the one that goes directly to f from r. The vagueness value is $V(f, r) = 0.15$ and l is *nearest and right.*

As a final example, suppose $r = a_{25}$ and $f = a_{12}$. Again the system finds that a direct path from r to f is not very good since $V(f, r) = 0.86$. Instead it generates a path that describes five intermediate objects,
$\pi = (a_{25}, a_{23}, a_{22}, a_5, a_8, a_{11}, a_{12})$ with a combined vagueness value of
$V(a_{25}, a_{23}, a_{22}, a_5, a_8, a_{11}, a_{12}) = 0.71$, which is lower than $V(a_{25}, a_{12})$ but not by much. A problem similar to that of reducing the number of prepositions in order to reduce ambiguity also holds for reducing the number of intermediate objects. Too many intermediate objects are just as likely to cause confusion as too many prepositions. The overall goodness of a description may be computed by means of a criterion that will combine certain measures deemed important. Several such measures may include the number of intermediate objects, number of prepositions in a locative expression and the vagueness of a description.

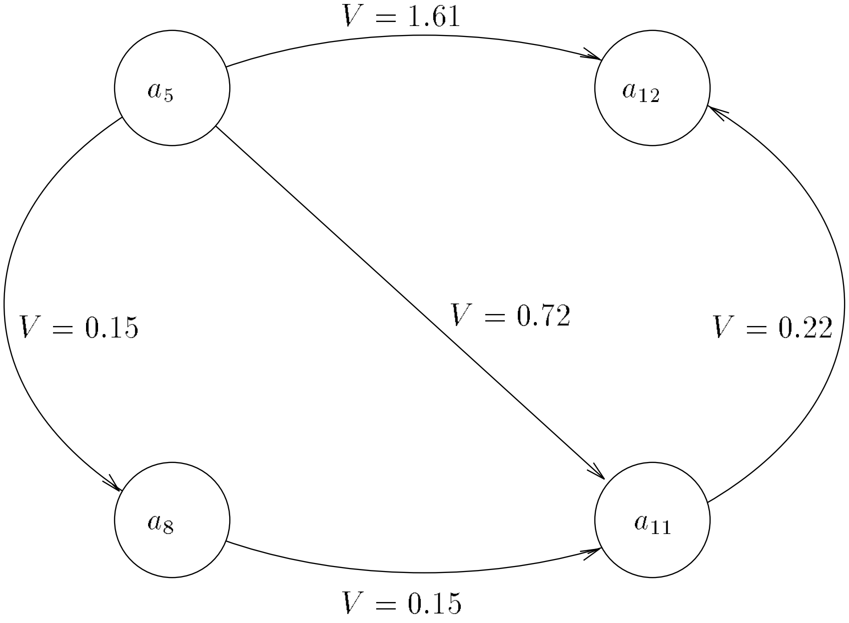


Figure 6.3: The minimal vagueness path: the nodes in the graph represent objects from figure 6.1 and the arcs represent the vagueness value of the objects on the arc. The best path for describing object 12 with respect to object 5 is $\pi = (a_5, a_8, a_{11}, a_{12})$.

Figure 6.4 illustrates, qualitatively, the influence that the choice of a criterion has on the choice of a description. The criterion here depends on vagueness (V) and path length ($|\pi|$). If the criterion function f is chosen to be $f = V$ then the optimal description is represented by point A that has minimal vagueness V_{min}. If the criterion function is $f = |\pi|$ then the optimal description is represented by point C that has minimal complexity. If, however, the criterion function is a combination of these two measures, such as $f = V|\pi|$, then the optimal description is represented by point B.

We have chosen to use the following criterion:

$$J = w \log_2 |\pi| + V(\pi)$$

In this criterion $|\pi|$ is the number of objects in the path π, $V(\pi)$ is vagueness along the path, and w is a weight parameter. The parameter w determines where the emphasis in path generation is placed. If w is 0 then no emphasis is placed

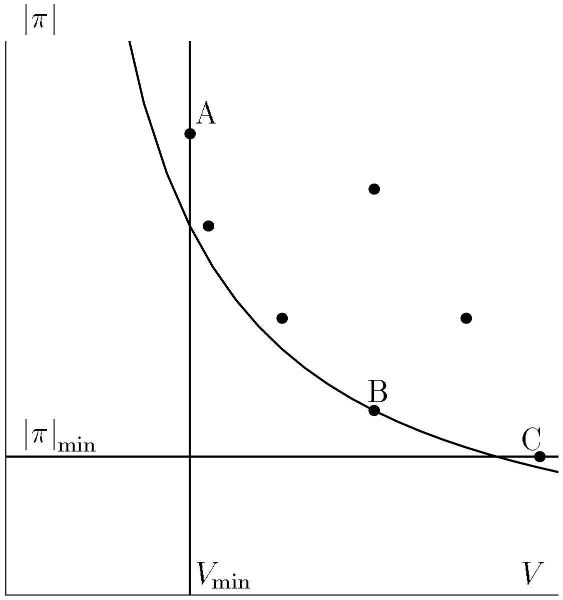


Figure 6.4: The influence of various criteria on the selection of an optimal path description

on reducing the number of intermediate objects that are chosen. If w is 1 then emphasis is placed on reducing the number of intermediate objects that are chosen. A middle of the road value for w, like 0.5, will yield some intermediate emphasis for the number of intermediate objects that are chosen. As with most uses of a criterion, x is usually sacrificed for y. In this situation, setting w higher sacrifices clarity of description for lower vagueness. This is the situation where there are too many intermediate objects. Setting w arbitrarily high sacrifices low vagueness for less intermediate objects. In summary, it is a question of generating a "quick" description, one with few intermediate objects, or a "safe" description, one with potentially many intermediate objects.

6.2 Path Description Generation

Up until this point in the chapter we have discussed only how paths are chosen but not how they are described. The language generation preprocessor is responsible for determining what the path description should look like. The NL generator is responsible for generating the desired description. The next several subsections will not only illustrate some of the path descriptions generated by the system, but also how they are influenced by the parameter w from criteria J. We will illustrate descriptions for value of $w = 0$, $w = 0.5$ and $w = 1$. We may have chosen values greater than 1 but these three values were enough to illustrate the point of "quick" descriptions versus "safe" descriptions.

6.2.1 Example 1

We will first use figure 6.1 and the examples of section 6.1 to illustrate the sentences generated by the NL generator. Table 6.5 contains the relevant information for each of the object pairs described in the examples of this section. Let us say that we are at the Italian Pavilion(a_1) and we wish to know where the American Adventure Pavilion(a_2) is. The system accepts the Italian Pavilion as the reference object and the American Adventure Pavilion as the figure object. If the system uses a value of 0 for w it finds that it is "best" to describe the American Adventure Pavilion directly from the Italian Pavilion because the vagueness is small enough, $V(a_1, a_2) = 0.15$. The NL generator generates the following sentence:

```
The American Adventure Pavilion is the nearest one
to the right of the Italian Pavilion.
```

If the system uses a value of $w > 0$ it finds the same path, since there are no intermediate objects to be removed.

Suppose next we would like to know where the Mexican Pavilion(a_{12}) is in re-

r	f	superlative	V	probability	l
1	2	nearest	0.15	90%	right
5	12		1.5	35%	medium,far below,aligned,between
5	8	leftmost	.15	90%	next
8	11	topmost	.19	88%	below
11	12	topmost	.21	86%	below
19	10		2.5	18%	far, aligned, right
19	9	rightmost	.23	85%	
9	10	nearest	.25	84%	below
19	22		.57	67%	small,below
22	17		0	100%	aligned
9	11	biggest	2.4	19%	far, aligned,left
9	24	bottommost	0.03	98%	below
24	5	farthest	.38	77%	
5	8	leftmost	.15	90%	next
8	11	topmost	.19	88%	below
9	19	leftmost	.25	84%	
19	8	rightmost	.23	85%	
19	11	bottommost	.92	53%	far, above
20	14	bottommost	.27	83%	above
14	16	rightmost	.73	60%	medium, far left between
20	16	leftmost	1.89	27%	medium, between
19	17	rightmost	1.75	30%	far, between

Figure 6.5: The superlative, vagueness value, probability of getting to f and l for each of the object pairs discussed in the image of figure 6.1

lation to the German Pavilion(a_5). Using $w = 0$, the system finds that the "best" path describes two intermediate objects, the Chinese Pavilion(a_8) and the Norwegian Pavilion(a_{11}). The vagueness $V(a_5, a_{12}) = 1.51$ but the vagueness of the path is $V(a_5, a_8, a_{11}, a_{12}) = 0.55$. The sentences generated are:

```
First, find the Chinese Pavilion which is the leftmost one
near the German Pavilion.
Then, find the Norwegian Pavilion which is the leftmost one
next to the Chinese Pavilion.
The Mexican Pavilion is the topmost one below the Norwegian Pavilion.
```

For paths that contain several intermediate objects like this one the language generation preprocessor can select time adverbials such as *first*, *then* or *afterwards* to enhance the descriptions. When the system uses $w = 0.5$ it finds the same path as it did for $w = 0$. For $w = 1$ the system finds a different description. Instead of describing two intermediate objects the system chooses to describe only one intermediate object. The vagueness of this path is $V(a_5, a_{11}, a_{12}) = 0.92$. In this example, we sacrificed a lower vagueness for less intermediate objects. The sentences generated are:

```
First, identify the Norwegian Pavilion which is the topmost one
below and far from the German Pavilion.
The Mexican Pavilion is the topmost one below
the Norwegian Pavilion.
```

Now let us say we are at Horizons(a_{19}) and we want to know where the United Kingdom is. When we give this request to the system, with $w = 0$, it finds that the path that goes directly to the United Kingdom Pavilion(a_{10}) has a fairly high vagueness value $V(a_{19}, a_{10}) = 2.5$. Instead it chooses to first describe how to locate the International Gateway Pavilion(a_9) and from there describe how to find the United Kingdom Pavilion since this yields a vagueness value $V(a_{19}, a_9, a_{10}) = 0.48$. The following are the sentences the NL generator generates:

```
First, find the International Gateway Pavilion which is
rightmost on the map.
The United Kingdom Pavilion is the nearest one below the
International Gateway Pavilion.
```

When we try this example with $w = 0.5$ and $w = 1$ the system finds the same description since a path with less intermediate objects would have to go directly to the figure object, and the vagueness of the direct path is too high.

Suppose next we would like to know where the Communicore West(a_{17}) is in relation to the Horizons(a_{19}). With a value of $w = 0, w = 0.5$ and $w = 1$ the system finds the same path description, it is the following:

```
First, locate the DNA statue which is the small one below
the Horizons.
The Communicore West is aligned with the DNA statue
```

The vagueness value $V(a_{19}, a_{22}, a_{17}) = 0.57$ as opposed to $V(a_{19}, a_{17}) = 1.7$.

As another example suppose we are at the International Gateway Pavilion(a_9) and want to know where the Norwegian Pavilion is(a_{11}). For a value of $w = 0$ the system finds a path through three intermediate objects. The sentences generated are the following:

```
First, find the Living Seas which is the bottommost one below
the International Gateway Pavilion.
Then, note the German Pavilion which is the farthest one.
Afterwards, locate the Chinese Pavilion which is the leftmost one
next to the German Pavilion.
The Norwegian Pavilion is the topmost one below the Chinese Pavilion.
```

The vagueness value $V(a_9, a_{24}, a_5, a_8, a_{11}) = 2.$ as opposed to $V(a_9, a_{11}) = 2.4$ with a description of

```
The Norwegian Pavilion is the biggest one far from,
aligned with, and to the left of the International Gateway Pavilion.
```

In this example, although there is not a significant decrease in vagueness by going through several intermediate objects, the individual descriptions along the path are clearer than the single description describing a_9 and a_{11}. When w is set to 0.5 the path description changes, it becomes the following:

```
First, locate the Horizons which is the leftmost one.
Afterwards, find, the Chinese Pavilion which is the
leftmost one far from the Horizons.
The Norwegian Pavilion is the topmost one below the Chinese Pavilion.
```

The vagueness $V(a_9, a_{19}, a_8, a_{11}) = 0.92$. If $w = 1$ we get yet another description:

```
First, note the Horizons which is the leftmost one.
The Norwegian Pavilion is the bottommost one above
and far from the Horizons.
```

The vagueness $V(a_9, a_{19}, a_{11}) = 1.1$. This example again demonstrates what we mentioned in section 6.1, and that is that it may be the case, as it is here, that clarity of description is sacrificed for less intermediate objects.

As a final example, let us take a look at a description that illustrates a figure object whose description reveals that it is a difficult to describe object. If we are at the Land(a_{20}) and we wish to know where the Communicore East(a_{16}) is the system generates the following sentences:

```
First, locate the Journey Into Imagination which is the bottommost
one above the Land.
The Communicore East is the medium-sized rightmost one far from
and to the left of the Journey Into Imagination.
It also lies between the Journey Into Imagination and another building.
```

The vagueness is $V(a_{20}, a_{14}, a_{16}) = 1.1$, as opposed to $V(a_{20}, a_{16}) = 1.8$. What makes this description difficult to follow is the use of the preposition *far* and *between*. There are several objects that could be *far* from the Journey into Imagination(a_{14}) and choosing which one is the *rightmost* one of those that are *far* depends on the perception of what is *far* and is not *far*. There are also several objects to the left of Journey into Imagination that lie between it and another building. A way to avert a situation like this one is to decrease the fuzzifying agent σ for *far*. If we know that *far* is a preposition that is not very descriptive for the current scenario, we may not want to include it in the description unless two objects are clearly and unmistakably *far*. By decreasing σ we can guarantee this.

Landmarks

There are certain objects from figure 6.1 that appear most often in path descriptions as intermediate objects. These objects are landmarks. The graph in figure 6.6 was created by generating all possible paths from each object to every other object in the image of figure 6.1 for three different values of $w, w = 0, w = 0.5$ and $w = 1$. For the image of figure 6.1 the International Gateway Pavilion(a_9) is

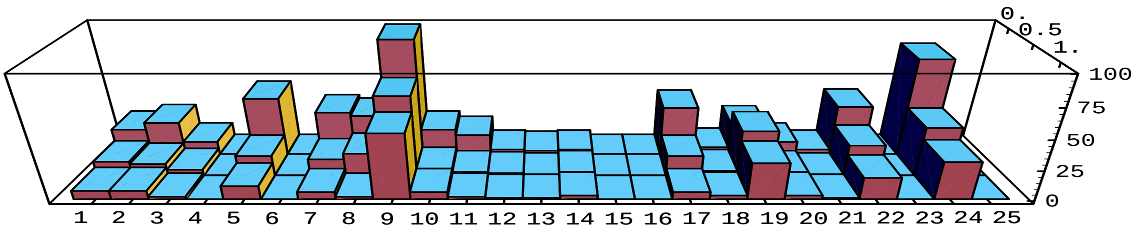


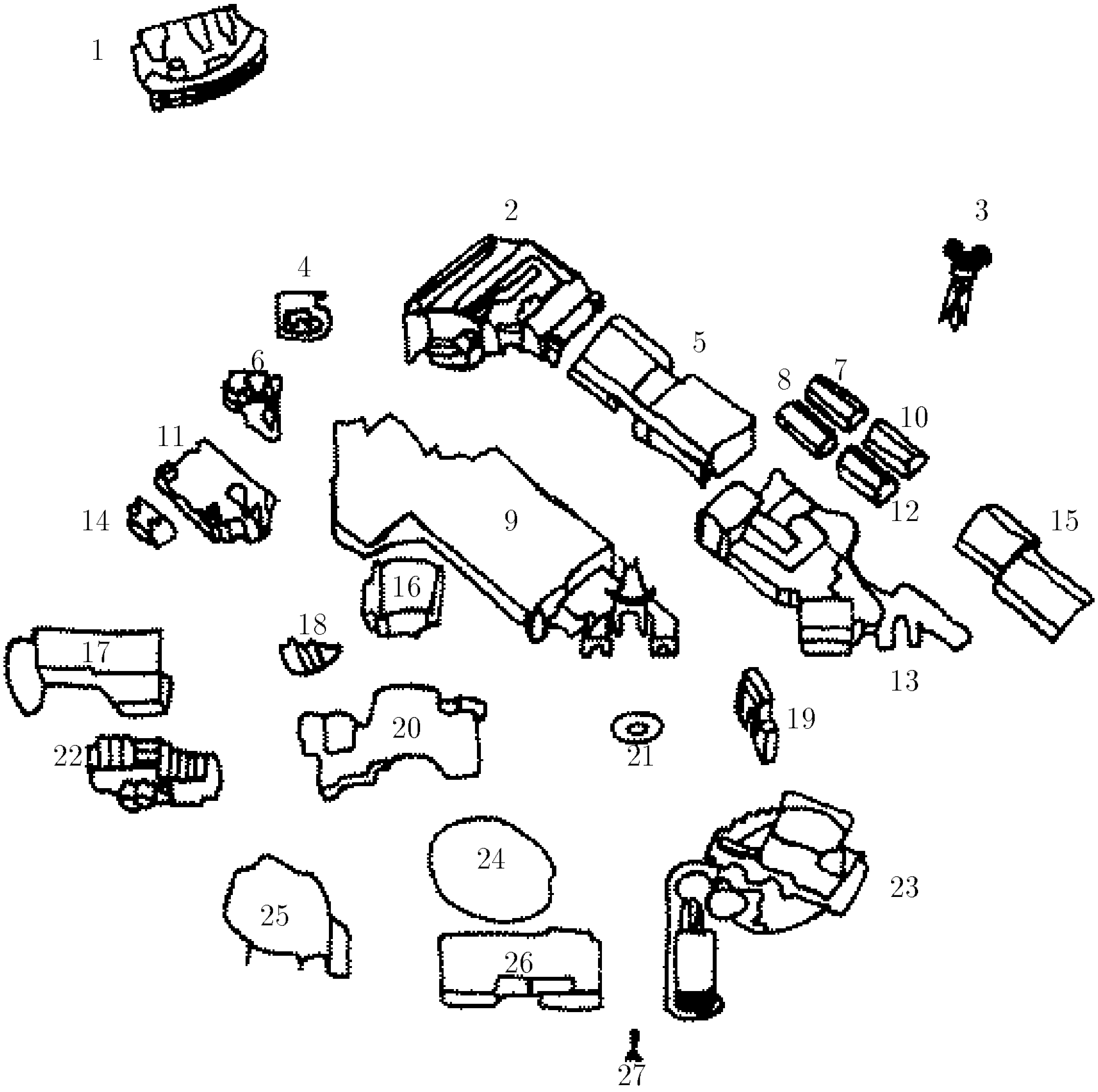
Figure 6.6: Landmarks from figure 6.1. A landmark is considered to be an object that appears the most in generating path descriptions. The graph illustrates the landmarks for $w = 0, w = 0.5, w = 1$.

the strongest landmark. It was selected as an intermediate object the most. It was also consistently a landmark for all three values of w. The Living Seas(a_{24}) was also a fairly strong landmark, although it was not as consistent as the International Gateway Pavilion. Both of these objects are describable using superlatives. The International Gateway Pavilion is the *rightmost* one, and the Living Seas(a_{24}) is the *bottommost* one.

6.2.2 Example 2

For the upcoming set of examples we have chosen Walt Disney World's MGM studios theme park map given in figure 6.7. As with the examples of subsection 6.2.1 we will illustrate the sentences generated for several examples along with the vagueness values associated with the chosen path. Figure 6.8 contains all the relevant data associated with each object pair in the examples.

Let us begin at Earffel Tower (a_3)). We would like to know where the Muppet Vision 3D (a_{14}) is. As before the system looks for the path with minimum vagueness. For $w = 0, w = 0.5$ and $w = 1$ it finds a path that goes through



1. Catastrophe Canyon
2. Inside the Magic
3. Earffel Tower
4. Teenage Mutant Ninja Turtles
5. Soundstage
6. New York Street
7. Production Center 1
8. Production Center 2
9. The Great Movie Ride
10. Production Center 3
11. Muppet Vision 3D
12. Production Center 4
13. Voyage of the Little Mermaid
14. Muppet Vision 3D
15. Backstage Studio Tour
16. The Movie Set Adventure
17. Star Tours 1
18. The Monster Sound Show
19. Arch
20. Superstar Television
21. fountain
22. Star Tours 2
23. Hollywood Boulevard Shop 1
24. Echo Lake
25. Indiana Jones Spectacular
26. Hollywood Boulevard Shop 2
27. statue

Figure 6.7: Walt Disney World's MGM studios theme park

r	f	superlative	V	probability	l
3	14		3.17	11%	far, left
3	15	rightmost	0	100%	
15	14	leftmost	.29	81%	aligned
2	24		1.94	26%	far, below, between
2	1	topmost	0	100%	aligned
1	17	leftmost	.03	97%	
17	26	bottommost	.04	97%	aligned
26	24	nearest	.57	67%	
2	26	bottommost	0.04	90%	aligned
17	22	leftmost	.28	82%	
26	4		1.95	25%	far, left, between
26	1	topmost	0	100%	aligned
1	2	topmost	0	100%	aligned
2	4	nearest	.36	77%	left
26	2	topmost	.14	90%	big
23	5	topmost	.64	64%	medium, above
23	13	rightmost	0.07	95%	big
13	5	topmost	0.17	88%	next
2	16	nearest	1.43	37%	medium, below
1	23	rightmost	0	100%	
23	3	topmost	.01	99%	right
3	15	rightmost	0	100%	
15	25	bottommost	0	100%	aligned
25	16	rightmost	.89	53%	far,above, between
2	15	rightmost	.02	98%	

Figure 6.8: The superlative, vagueness value, probability of getting to f and l for each of the object pairs discussed in the image of figure 6.7

one intermediate objects with a vagueness of $V(a_3, a_{15}, a_{14}) = .3$ as opposed to $V(a_3, a_{14}) = 3.2$. The sentences generated are:

```
First, find the Backstage Studio Tour which is the rightmost one.
The Muppet Vision 3D 2 is the leftmost one aligned with the
Backstage Studio Tour.
```

Suppose now that we are at Inside the Magic (a_2) and we want to locate the Echo Lake(a_{24}). For $w = 0$ the system finds it necessary to go through three intermediate objects. The vagueness of the path is $V(a_2, a_1, a_{17}, a_{26}, a_{24}) = 0.64$. The sentences generated are:

```
First, identify the Catastrophe Canyon which is the topmost one.
Then, locate the Star Tours 1 which is the leftmost one.
aligned with Catastrophe Canyon.
Then, find the Hollywood Boulevard Shop 2 which is bottommost one
aligned with the Star Tours 1.
The Echo Lake is the nearest one.
```

This example certainly exemplifies the need for some control over the number of intermediate objects that are produced. When we set $w = 0.5$ the system finds the same path. However, when we set $w = 1$ the system finds it necessary to go through only one intermediate object. The vagueness of the path is $V(a_2, a_{26}, a_{24}) = 0.7$. If we go directly from a_2 to a_{24} the vagueness is $V = 1.9$. The sentences generated are:

```
First, identify the Hollywood Boulevard Shop 2 which is the bottommost
one aligned with Inside the Magic.
The Echo Lake is nearest one.
```

As another example, suppose we are at Star Tours $1(a_{17})$ and we want to know where Star Tours $2(a_{22})$ is. The system finds no intermediate objects and generates the following sentence with a vagueness of $V = 0.28$

```
The Star Tours 2 is the leftmost one.
```

The interesting thing to note here is that, although the Star Tours 1 is the leftmost object on the map, Star Tours 2 is the leftmost object with respect to Star Tours 1.

Next suppose we are at the Hollywood Boulevard Shop $2(a_{26})$ and we want to know where the Teenage Mutant Ninja Turtles(a_4) are. With $w = 0$ the system finds it necessary to go through two intermediate objects with a vagueness

$V(a_{26}, a_1, a_2, a_4) = 0.37$. The sentences generated are:

```
First, identify the Catastrophe Canyon which is the topmost one.
Now, find the Inside the Magic which is the topmost one
aligned with the Catastrophe Canyon
The Teenage Mutant Ninja Turtles is the nearest one
to the left of the Inside the Magic.
```

When we set $w = 0.5$ or $w = 1$ the system generates the following sentences with a vagueness of $V(a_{26}, a_2, a_4) = 0.5$:

```
First, identify the Inside the Magic which is the big topmost one.
The Teenage Mutant Ninja Turtles is the nearest one
to the left of the Inside the Magic.
```

After shopping at the Hollywood Shops 1(a_{23}) we are interested in visiting the Soundstage(a_5) so the system generates the following sentences for $w = 0$:

```
First, find the Voyage of the Little Mermaid which is the big
rightmost one.
The Soundstage is the topmost one next to the Voyage
of the Little Mermaid.
```

The vagueness is $V(a_{23}, a_{13}, a_5) = 0.24$ as opposed to $V(a_{23}, a_5) = 0.64$. If we set $w = 0.5$ the system finds the same path. However, when we set $w = 1$ the system generates the following sentence:

```
The Soundstage is the medium-sized topmost one above
the Hollywood Boulevard Shop 1.
```

To emphasize again the tradeoffs that occur when searching for a path with minimum vagueness, observe the sentences generated if the reference object is Inside the Magic(a_2) and the figure object is The Movie Set Adventure(a_{16}) and

$w = 0$:

```
First, find the Catastrophe Canyon which is the topmost one.
Then, note the Hollywood Boulevard Shop 1 which is the rightmost
one aligned with the Catastrophe Canyon.
Now, note the Earffel Tower which is the topmost one
to the right of the Hollywood Boulevard Shop 1.
Afterwards, find the Backstage Studio Tour which is the rightmost one.

Then, find the Indiana Jones Spectacular which is the bottommost
one aligned with the Backstage Studio Tour.
The Movie Set Adventure is the rightmost one above and far from
the Indiana Jones Spectacular.
It also lies between the Indiana Jones Spectacular and another building.
```

The system found it necessary to describe five intermediate objects for this example with a vagueness of $V(a_2, a_1, a_{23}, a_3, a_{15}, a_{25}, a_{16}) = 0.9$ as opposed to $V(a_2, a_{16}) = 1.4$. If $w = 0.5$ the path description changes, it becomes:

```
First, identify the Backstage Studio Tour which is the rightmost
one.
The, identify the Indiana Jones Spectacular which is the bottommost
one aligned with the Backstage Studio Tour.
The Movie Set Adventure is the rightmost one above and far from
the Indiana Jones Spectacular.
It also lies between the Indiana Jones Spectacular and another building.
```

The vagueness in this case is $V(a_2, a_{15}, a_{25}, a_{16}) = 0.91$. Finally for $w = 1$ the system finds a direct path with vagueness $V(a_2, a_{16}) = 1.4$. The sentence generated is:

```
The Movie Set Adventure is the medium-sized nearest one
below the Inside the Magic.
```

Landmarks

Figure 6.9 illustrates how many times each object from the image in figure 6.7 was used as an intermediate object. Object 1 (Catastrophe Canyon) is the clear

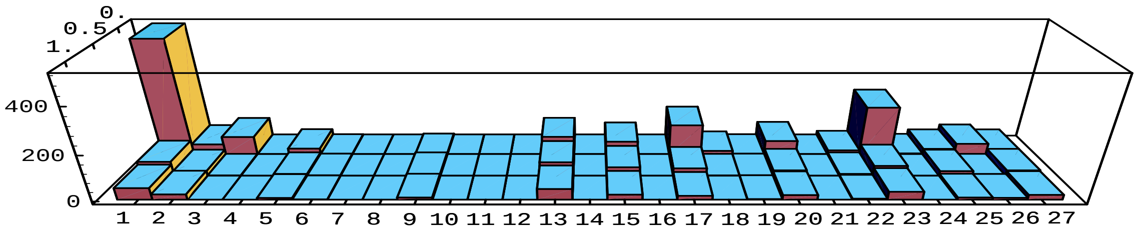


Figure 6.9: Landmarks from figure 6.7. The graph illustrates the landmarks for $w = 0, w = 0.5, w = 1$.

winner for $w = 0$. However, for $w = 0.5$ and $w = 1$, there is not a clear winner. As a matter of fact, even the count for object 1 decreased significantly. What examining a graph such as the one in figure 6.9 may reveal is whether the objects in the image are easy to describe. If there are several landmarks, in the image then they can serve as an anchor for describing other objects in the image. The anchor could be used in the descriptions since it would presumably be easy to locate. However, in an image where there is no obvious anchor object this task becomes more difficult. In this particular case the count for object 1 is so much smaller for $w = 0.5$ and $w = 1$ because the vagueness of the path that included object 1 as an intermediate object was not that much lower than for those paths that did not include object 1. What is interesting to note about the graph of figure 6.6 and figure 6.9 is that the objects with the highest count are all objects that are extremes; they are all describable with superlatives or object features in combination with superlatives. In figure 6.7 object 1 is the *topmost* one, object 23 is the the *big bottommost* one, object 17 is the *leftmost* one, object 13 is the *big rightmost* one, and object 15 is the *rightmost* one.

6.3 Implementation Issues

The sentences in subsection 6.2.1 and 6.2.2 were generated "on the fly" by the NL generator with the help of the language generation preprocessor. Before the NL generator can generate a sentence it must be given the semantic input that is generated by the language generation preprocessor. The language generation preprocessor uses the locative expressions generated by the locative expression generator to produce the semantic input. The language generation preprocessor converts a locative expression into a form that is acceptable for the NL generator. The NL generator uses FUF [Elhadad, 1993], a natural language generator program that uses the technique of unification grammars. It consists of two modules, the unifier and the linearizer. The unifier takes the semantic input of the text to be generated and a unification grammar and produces a syntactic description of the text. The linerizer interprets the syntactic description and generates the English sentence. The unification grammar is called SURGE(Systemic Unification Realization Grammar of English) and was developed by the natural language processing group at Columbia University, [McKeown *et al.*, 1990], [Robin, 1994]. It is capable of generating a large variety of sentences.

Before the natural language generator can generate the English sentence the language generation preprocessor must translate the output of the locative expression generator. The locative expression generator's output is a vector of 0's, 1's, and X's. The language generation preprocessor takes this vector and produces a preliminary semantic input. The vector contains a value of 0, 1 or X for each of the following prepositions and features, (*small, medium, big, near, far, above, below, aligned, next inside, left, right, between*). An example of a vector produced by the locative expression generator is (1 X X X X X 1 X X X X X X), where each entry corresponds to a feature or preposition aforementioned. The language generation preprocessor then takes this output and generates the preliminary semantic input as shown in figure 6.10.

```
((size small)
 (vertical-orientation below))
```

Figure 6.10: An example of the preliminary semantic input generated by the language generation preprocessor

```
((cat clause)
 (proc ((type locative) (mode attributive)))
 (partic ((carrier ((cat common) (lex "DNA statue")))
          (attribute ((cat common)
                      (lex "one")
                      (describer ((lex "small")))
                      (qualifier ((cat pp)
                                  (prep ((lex "below")))
                                  (np ((cat common)
                                       (lex "Horizons"))))))))))
```

Figure 6.11: Enriched semantic input of figure 6.10. It is used to ultimately generate the final sentence, which is “The DNA Statue is the small one below Horizons.”

The language generation preprocessor then gives this preliminary semantic input along with a preliminary grammar to the NL generator. The NL generator then calls upon FUF to enrich this preliminary semantic input. FUF then uses this enriched semantic input and SURGE in order to generate the final sentence. The enriched semantic input of figure 6.10 is shown in figure 6.11.

The reason for the preliminary semantic input and grammar is to allow the language generation preprocessor the ability to generate a wider range of sentences with a simpler semantic input. Instead of generating the semantic input of figure 6.11 it needs to only generate the semantic input of figure 6.10 and the preliminary grammar handles producing the semantic input required by SURGE. Being able to use this preliminary grammar also means that the language generation preprocessor can generate the same semantic input for different lexical items. For example, the semantic input of figure 6.12 contains the same form as the semantic input of figure 6.10 but for two different lexical items. Instead of *small*

```
((size big)
 (vertical-orientation above))
```

Figure 6.12: A preliminary semantic input similar to the semantic input of the figure 6.10 except for the lexical items, *above* and *big*

and *below*, the preliminary semantic input of figure 6.12 contains *big* and *above*. The appropriate reference object and figure object are chosen by the preliminary grammar. FUF allows external calls to functions, and the preliminary grammar does so to retrieve the reference object and figure object.

Some other examples of preliminary semantic input and the sentences they represent are shown in figure 6.13.

6.4 Verification

We showed in subsections 6.2.1 and 6.2.2 some of the sentences the system generated. In this section we take a sample of sentences generated by the system and test their ability to convey a proper description of a figure object. We gave the map of Epcot Center to sixteen people along with ten examples describing how to get from a start location to a goal location. The map and sentences are shown in figure 6.14, and a table of results is shown in table 6.15.

The intended figure object for the first three examples were found by all sixteen participants. The figure object in the fourth example was found by 11 of the participants. Two of the participants felt that object number 23 was the *small* farthest one, one participant thought the intended figure object was 18, and two participants chose not to respond.

In the fifth example, 15 participants agreed with the system's choice of figure object and one thought it was the object numbered 23. This person may have

The American Adventure Pavilion is the nearest one to the right of the Italian Pavilion.

```
((superlative nearest)
 (horizontal-orientation right))
```

First, find the Chinese Pavilion which is the leftmost one near the German Pavilion.

```
((sent-pos first)
 (superlative leftmost)
 (distance near))
```

Afterwards, find the Backstage Studio Tour which is the rightmost one.

```
((sent-pos intermediate)
 (superlative rightmost))
```

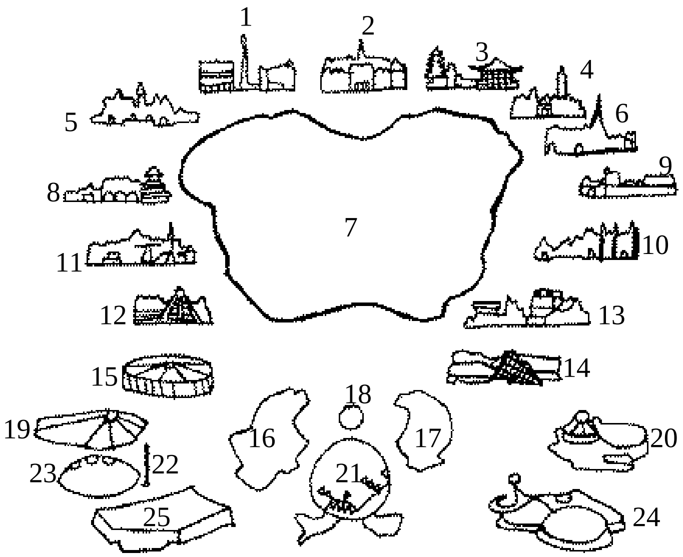
The Norwegian Pavilion is the biggest one far from, aligned with and to the left of the International Gateway Pavilion.

```
((superlative biggest)
 (distance far)
 (orientation aligned)
 (horizontal-orientation left))
```

Figure 6.13: Examples of preliminary semantic input generated by the language generation preprocessor

thought that object 22 was not a building which is why they chose 23.

In the sixth example, 10 participants were able to locate the intended figure object. Those participants that did not were unable to do so because they used the concept of *strictly above* when answering the first part of the description and the system was using the computational model of *above* that did not restrict *above*. The six participants either chose the object numbered 3, 4, or 6 as the one that is



The following set of directions is intended to get you from a start location to a goal location. Provide the number of the building you think is the goal that is being described for each of the following 10 directions.

1. Find the building which is the nearest one to the left of 2. __________
2. First locate the building which is the nearest one below 8. __________
   Then find the building which is the nearest one below the one you found in the previous sentence. __________
3. First find the building which is the rightmost one. __________
   Then locate the building which is the nearest one below the one you found in the previous sentence. __________
4. Find the building which is the small farthest one from 7. __________
5. Find the building which is the small one near 19. __________
6. First find the building that is the leftmost one above 10. __________
   Then find the building which is the nearest one below the one you found in the previous sentence. __________
7. Locate the building which is the rightmost one below 10. __________
8. First find the building which is the nearest one above 19. __________
   Then find the building which is nearest to the one you found in the previous sentence. __________
9. First locate the building which is the rightmost one below 12. __________
   Then find the building which is the nearest one above the one you found in the previous sentence. __________
10. First find the building which is the rightmost one far from 6. __________
    Then find the building which is the nearest one above the one you found in the previous sentence. __________

Figure 6.14: The map and sentences used to test the ability of people to find intended figure objects.

Question	1	2	3	4	5	6	7	8	9	10
System's response	1	11, 12	9, 10	22	22	8, 11	20	15, 12	20, 14	20 14
% of correct responses	100	100	100	69	94	62	94	6	50	50

Figure 6.15: This table indicates for each question in figure 6.14, the figure object the system was referring to and the percentage of people that were able to find it correctly.

the *leftmost one above* 10.

In the seventh example, 15 participants correctly located the figure object and the one that did not must have confused *above* with *below* because he/she choose the object numbered 9, which is clearly not *below* 10.

The eighth example was only answered correctly by 1 participant. All of the other 15 participants gave an answer of 19 as the intended figure object. The confusion with these sets of descriptions was due to the lack of understanding that we want to describe a goal object that is not the same as the figure object. A more detailed explanation of the task would probably have eliminated this confusion.

In the ninth example, 8 of the participants found the intended figure object. 6 out of the 8 that did not find the intended figure object used the concept of *strictly below* when searching for the figure object; the objects numbered 25, 19, or 16, were chosen as the object that is the *rightmost one below* 12. One participants felt that the object numbered 24 was the *rightmost one below* and the other participant chose the object numbered 9 as the *rightmost one below.* The apparent confusion with this participant is in the use of two spatial relation together in one statement. While it is true that object 9 is the *rightmost* one, it is not *below* 12, and hence can not be the *rightmost one below* 12.

For the last example, 12 of the participants found the intended figure object. Two left it blank and two felt that the object numbered 9 was the *rightmost one far from 6.* It would seem that these two participants, like the one in the previous example, find the concept of two spatial relation in one statement confusing, and they simply defaulted to using one of the statements. It may have been the that believed that 9 is *far* from 12, but this wouldn't seem to be the case, because the two objects are practically touching one another.

In summary, the generated descriptions were interpreted rather well (over 70% overall). The most interesting misunderstanding stemmed from the concept of *strictly above* or *strictly below*. It is impossible to come up with a computational model that can capture everybody's impressions about what is considered what. The best we could have done to handle this situation would have been to generate two descriptions, one that used the computational model of *strictly above* and one that did not, but this would not be very feasible. Instead of clarifying a description we may confuse the user by introducing too many descriptions.

6.5 Conclusion

In this chapter we examined the results of applying the techniques of chapters 3 and 4 on the task of landmark navigation. The goal of generating the "best" path description relies on the vagueness value associated with each description along the path. As we saw it may be the case that it is better to describe how to get to the figure object if we first describe how to get to an intermediate object. The reason for this is that the intermediate object has higher probability of being selected than the figure object. Of course, the ultimate deciding factor is what the vagueness is along the entire path. However, it may be the case that the vagueness associated with the description for the intermediate object and the figure object may be greater than the vagueness value for the description of the reference object and figure object. In this case we choose to describe the figure object directly since it has a lower vagueness value than the combined vagueness values along the path.

We showed in the two theme park maps that objects that tended to be chosen the most as intermediate objects were those objects that were extremes, i.e. those objects that were described by global superlatives, like *leftmost*. These yielded the highest vagueness values. These objects fit the description of a landmark since they can be easily located and are used as reference objects to describe other objects in the image. A landmark may also be an object that when used as a reference object produces locative expression with the least number or prepositions (this will be the topic of future study).

We examined the way in which the language generation preprocessor produces the semantic input required by the natural language generator. The simplicity of the preliminary input that it creates is important if we want to generate sentences that differ lexically but not structurally. Instead of writing routines to generate the entire semantic input that the natural language generator requires we wrote a preliminary grammar that FUF can then interpret to produce the final semantic input. We let FUF do all the work of generating semantic inputs like the one in figure 6.11.

Chapter 7

Description Generation of Abnormal Densities found in Radiographs

In this chapter we will describe how the methodologies from chapters 3, 4, and 5 are used to generate descriptions of abnormal densities found in radiographs. We will illustrate the various steps that the image preprocessing and image processing module must make in order to locate the stone and register the X-ray with a model of the X-ray. We will also demonstrate how the language generation preprocessor adheres to the appropriate medical terminology when generating descriptions. Several examples will be given to illustrate the results of the image processing and language generation preprocessor [Abella and Kender, 1994b].

7.1 Introduction

The goal of this task is to describe renal stones found in radiographs - the precise goal of a radiologist examining an X-ray. The issue in the exam is to first determine if a stone is present in the radiograph, and if so to identify its location for possible treatment. This task is an interesting one because it uses the principles laid out in this thesis; that is the qualitative spatial description of

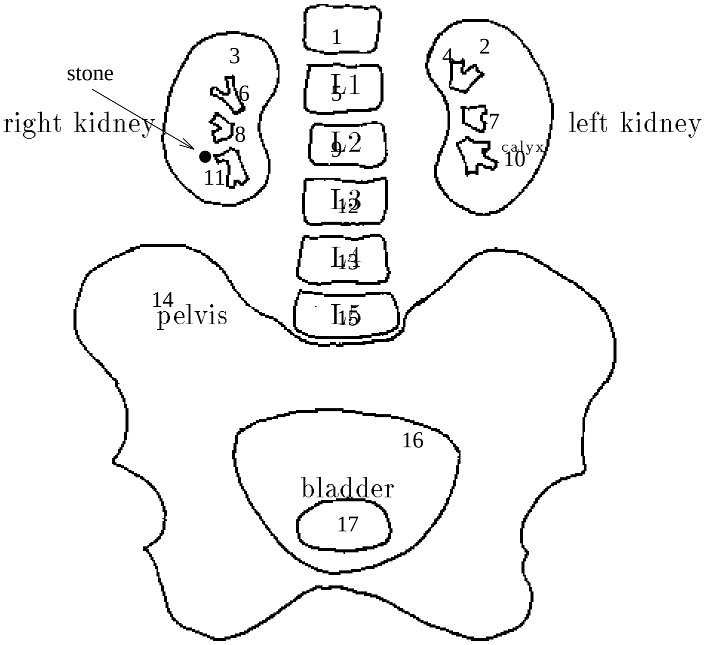


Figure 7.1: The model of the urinary system used to find the spatial relationship of stones found in radiographs

objects in images. Radiologists categorize stones according to their locations on the X-ray. The stones are described by observing the spatial relationship of the stones to other parts of the radiograph. Those parts include the spinal cord, the various lumbar bodies of the spinal cord, the bladder, the kidney, the calyx, and a few others that are not usually visible on the X-ray. For those descriptions that include references to landmarks that are not visible on an X-ray the radiologists use a mental map of where certain landmarks are located. Hence we too use a model of the X-ray image to determine some of the descriptions that rely on items not present in the X-ray. This model is illustrated in figure 7.1.

Each radiograph differs in quality; the better the quality the more visible the stones and other major landmarks. The radiographs may also be slightly translated, rotated or scaled when they were taken. The image preprocessing takes this into account and first registers the X-ray with the model of the X-ray. It searches for the best affine transformation that fits the model and then applies that transformation on the X-ray. A more detailed exposition of this process will be given in section 7.2.

Unlike the landmark navigation task this task requires that the language generation preprocessor decipher not only the spatial relationship of the stone with respect to a reference object, but also that it decipher the relationship of the stone to various reference objects in the X-ray. Most of these reference objects are shown in figure 7.1 but some are inferred from the spatial relationship. For example, if a stone is *between* the kidney and the bladder then it is very probable that the stone is in the ureter and depending upon which lumbar body it is level with it will be categorized as a pelvic ureteral stone, a mid-ureteral stone, or a distal ureteral stone. A more detailed explanation of how the stones are categorized will be given in section 7.3.

There has been a substantial amount of research effort in medical imaging. Most of this research is concerned with devising algorithms to effectively locate boundaries, measure temporal changes in images, and improve the quality of the images. Pattern recognition has also seen its application in medical imaging – locating skin lesions, tumors, organs etc. and is seen as a means of performing some interpretation of the image. What has not seen activity is integrating image processing with tools for generating qualitative descriptions of images, which is what will be the main focus of this chapter. The closest area is suggested by [Fox and Walker, 1989]. They believe that a useful role of computers in medicine is for imaging systems to be combined with methods for interpreting clinical data. They feel that expert systems can be used to integrate imaging techniques with tools for clinical decision making and problem solving. Fox favors programs that can accept pixel arrays and output explicitly symbolic descriptions.

More along the lines of medical imaging systems can be found in [Kobashi and Shapiro, 1992]. They describe a knowledge-based recognition system that utilizes knowledge of anatomy and CT (computerized tomography) imaging for organ identification of the abdomen. Their system does not use standard shape-matching techniques, and uses negative shape constraints instead, (e.g. what shape an organ does not take). It also uses knowledge-based segmentation/extraction guided by a feedback control system based on information about various constraints. Instead of performing segmentation and extraction sequentially their system performs them at the same time. This enables their system to utilize a knowledge-based feedback procedure. They define a model of the organs based on several properties, such as, gray tone levels among organs and location in terms of the coordinate system based on some stable landmarks.

In [Tagare *et al.*, 1993] a concept called an "arrangement" is defined that is used to retrieve MR images of the heart from a database. An arrangement is a qualitative spatial relation that describes the sequence in which neighbors of each part of an image are situated around it. They also define a metric for comparing arrangements. The goal of their system is to be able to use arrangements to retrieve images from a medical image database which have the same or similar tomographic section to an example image. They argue that a geometric model of an example image is not useful for retrieving images from a database because the images in the database will always differ slightly. They argue that their approach is better because it allows for a similarity measure in terms of arrangements, so that similar images may be retrieved.

A method for automatically detecting boundaries of brain tumors is given in [Lu *et al.*, 1992]. The paper describes a knowledge-guided boundary detection algorithm. The choice of a knowledge-guided algorithm is made because the authors believe that it will more accurately detect boundaries. A similar paper is [Selfridge and Prewitt, 1981] that describes two boundary-delineating algorithms for detecting kidneys in tomographic images. The first algorithm operates on the entire image in order to establish an initial segmentation and the second allows for the incorporation of anatomical knowledge.

Most medical image processing involves standard techniques like those found in [Wechsler and Sklansky, 1977]. In this paper the authors describe a system for finding the rib cage in chest radiographs. They do not use a knowledge-based approach to the segmentation and extraction of desired features. The steps invoked by the system include digitizing and filtering the input radiograph, applying a local edge detector, applying a global boundary detector, and joining the dorsal and ventral segments found to be part of the rib cage.

7.2 Image Processing

Before we can describe a density, that is a potential stone, we must be able to find it in the image. This requires various levels of image processing. In general we need to accomplish two separate yet interdependent goals, the first is to register the image with the model of the urinary system shown in figure 7.1 and the second is to find the stone once the image has been registered. This is necessary because each image is slightly distorted; each may either be slightly translated, scaled or rotated and may differ in brightness. The reasons for the distortions are many, but some of the primary reasons involve the size of the person, the duration of the exposure to the X-ray, the type of film used, the type of enhancing screen used (e.g. a phosphorent sheet improves efficiency of the conversion of X-ray to measurable photons), the energy of the X-ray photons (e.g. depending on the voltage, the relative absorption by different tissues differs), the filters used (e.g. thin sheets of metal are used to filter out some X-ray, resulting in a more homogeneous beam) and the size of the electron beam used.

Accurately describing the location of the stone will depend on how well we were able to register the image. If the registration process was not accurate then when we find the stone and superimpose it on the model of the image we will find that

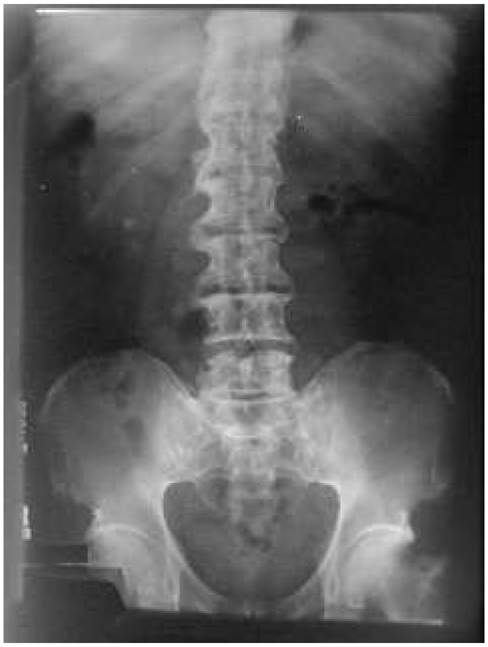


Figure 7.2: X-ray image

it may not be in the place it was on the original image. Of course, the algorithm for actually finding the stone also impacts on whether or not we describe a stone or some miscellaneous artifact.

Generating the binary image. The image processing module first converts an X-ray image into a black and white (binary) image in order to locate the spinal cord and pelvis. The spinal cord and pelvis are the two landmarks that we extract from the X-ray image in order to register it with the model. This process requires a threshold value that needs to be chosen beforehand. It is not possible to fix this threshold value for all the images as explained in the introduction. In order to find an appropriate threshold value the system uses the fact that the spinal cord and the pelvis occupy an approximately constant fraction of the X-ray. We experimentally determined this fraction to be 30%. The system first creates an image brightness histogram and then chooses the threshold value for which 30% of the image pixels have higher brightness. This is exemplified in figures 7.2-7.5.

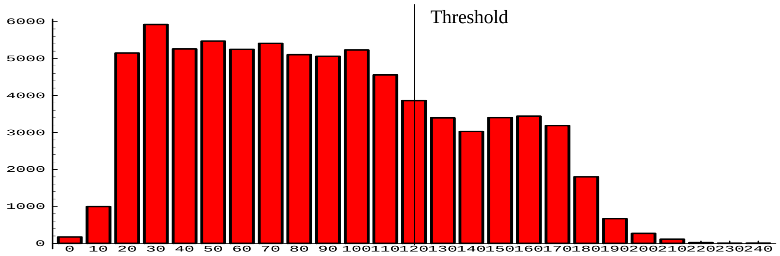


Figure 7.3: Image histogram

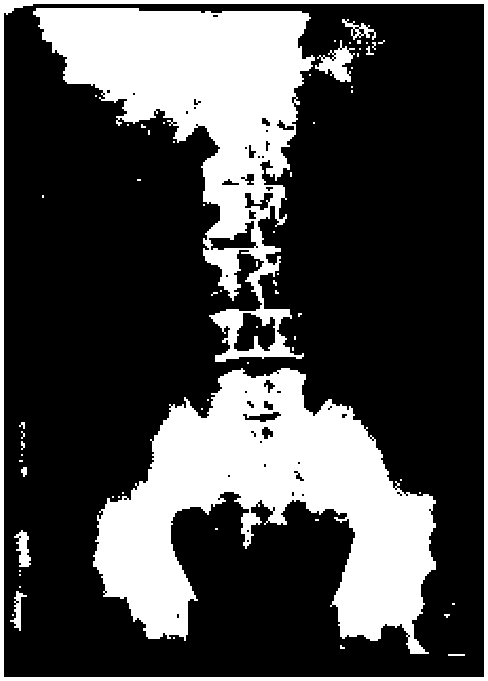


Figure 7.4: Thresholded image

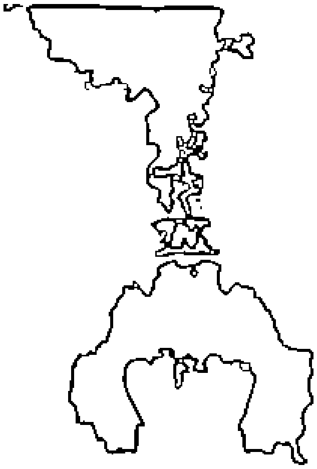


Figure 7.5: Edge image

Edge Detection. Once we've created the binary image we need to find the edges. The edges are small regions in an image that have a rapid change in image intensity. To find the edges we use a 5x5 edge operator, patterned after the Sobel 3x3, as a discrete approximation for the partial derivatives that measure the gradient.

Registering the image. In this step the image processing module needs to find a transformation that maps the edge image into the model. A set of points that outline distinctive features of the model, such as the pelvis and the spine, is taken from the model and used in this process. See figure 7.6.

The system looks for a six parameter affine coordinate transformation. The six parameters are the rotation angle θ, translation p_x and p_y, and the scaling parameters $\rho_{xx}, \rho_{xy}, \rho_{yy}$. The homogeneous rotation, translation, and scaling matrices are given by:

$$T_R = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad T_T = \begin{bmatrix} 1 & 0 & p_x \\ 0 & 1 & p_y \\ 0 & 0 & 1 \end{bmatrix} \quad T_S = \begin{bmatrix} \rho_{xx} & \rho_{xy} & 0 \\ \rho_{xy} & \rho_{yy} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

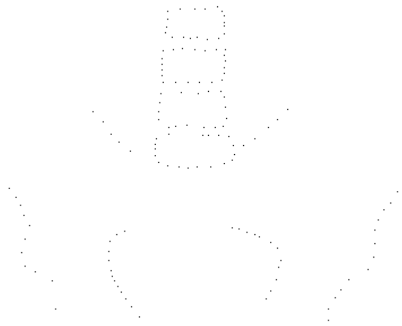


Figure 7.6: The representative points taken from the model

The affine transformation is given by:

$$T_A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} = T_R T_T T_S$$

or

$$T_A = \begin{bmatrix} \rho_{xx}\cos\theta - \rho_{xy}\sin\theta & \rho_{xy}\cos\theta - \rho_{yy}\sin\theta & p_x\cos\theta - p_y\sin\theta \\ \rho_{xy}\cos\theta + \rho_{xx}\sin\theta & \rho_{yy}\cos\theta + \rho_{xy}\sin\theta & p_y\cos\theta + p_x\sin\theta \\ 0 & 0 & 1 \end{bmatrix}$$

We are looking for T_A that maximizes the following criterion:

$$J(T_A) = \|\{(\mathbf{x}_i, \mathbf{m}_j)|\ \|T_A\mathbf{x}_i - \mathbf{m}_j\| < \delta\}\|$$

where $\mathbf{x}_i$ are edge points from the edge image and $\mathbf{m}_j$ are the representative points from the model. A match between a transformed point and a model point is found if their distance is less than δ. The criterion states that the transformation should bring the image as close as possible to the model image, meaning that it should have as many points matching with or being very close to points in the model. We assume that T_A is "small" because the X-rays themselves vary only moderately from one another. Therefore we assume tight intervals for T_A's parameters. The "best" transformation is found by using a best-first search. The search begins by splitting a 6D volume defined by intimal intervals for each 6 parameters $p_x \in [p_x^{min}, p_x^{max}], p_y \in [p_y^{min}, p_y^{max}], \ldots, \rho_{yy} \in [\rho_{yy}^{min}, \rho_{yy}^{max}]$ into $3^6 = 729$ regions. Each region is assigned a number that corresponds to the value of $J(T_A)$ where T_A is defined by the center of the region. We sort all the regions according to the criterion J and then split that region which yielded the maximum value for J. This process continues until the region with the maximal J has the translation component in the x and y direction below 1 pixel.

When we have found a transformation we apply it to the image. We are then ready to use this transformed image to locate the stones. The search for an optimal affine transformation can be done using other methods such as the gradient descent method. We used the method aforementioned because we had the existing software to implement it and it yielded positive results.

Locating the stones.

To locate the stones we use the technique sketched out in [Kimme *et al.*, February 1975] for circle finding. Since a majority of the stones are circular in nature we may use this technique. The principle behind the technique is the hough transform. Rather than searching all of the image we improve efficiency and reduce the number of erroneous circles found by looking for circles in a prescribed area. This area is illustrated in figure 7.7. After the search is complete the maximal entry in the accumulator array of the Hough transform defines the stone. If more than one entry is maximal than there is more than one stone found in the image.

Once the stone is found it is superimposed on the model and the spatial relations

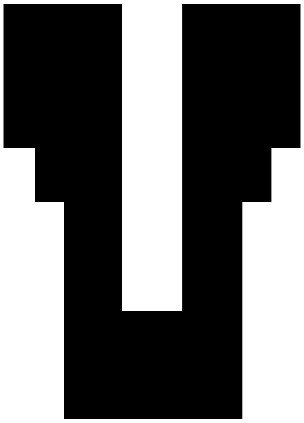


Figure 7.7: The area used to restrict where the circle finding algorithm looks to for potential circles

of the stone to all the reference objects in the model are computed. The next step is to choose which of the spatial relations are sufficient to describe the stone.

7.3 The Natural Language Generation Aspects of Density Descriptions

In this section we will discuss how the spatial relations of the previous section are translated into the proper medical terminology. Each preposition or combination of prepositions can potentially map into a particular description of a density. For example, the sentence that would be generated for a stone that is found to be to the left of L4 is

```
A stone is seen level to L4 on the left.  This probably represents a
mid-ureteral stone.
```

More than one preposition can be combined to produce a description, for example, if a stone is found to be *near* the middle calyx in figure 7.1 and *inside* the right kidney, then the system generates the sentence

```
A calyceal stone is seen in the right kidney.
```

Whenever the system encounters a set of spatial relations that it can not interpret it generates a default sentence. The default sentence conveys the spatial relations without the translation into the medical terminology. For example, a possible sentence that can be generated under this scenario is “A density is seen inside L3, near the right kidney, and to the right of the calyx.”

The locative expression generator supplies all the spatial relations it found to be necessary and sufficient to describe the stone. It is the job of the language generation preprocessor to compose the appropriate input to the natural language generator so that it may produce a meaningful sentence similar to the types of sentences that could be found in actual pathology reports. The language generation preprocessor is an embryo of a rule-based system for translating spatial relations into proper medical terminology.

The language generation preprocessor takes as input the result produced by the locative expression generator. An example of input to the language generation preprocessor is

```
((inside right-kidney) (near calyx) (above middle-calyx))
```

The language generation preprocessor then expands this input and creates the input needed by the natural language generator. For this particular example the fact that the stone is *above* the middle-calyx signals the language generation preprocessor that the stone is in the upper portion of the right kidney. The language generation preprocessor translates this into the proper medical term *upper pole*. The rule associated with translating *above* to *upper* and *below* to *lower* is shown in figure 7.8. The numbers in the rules correspond to the body parts numbered in figure 7.9. (Note that the rules that we will show are only a partial view of the

```
(def-alt upper-lower
  ((({above} given)
    (alt above-what
 ((({above} 8)
   (lex "upper"))
  (({above} 7)
   (lex "upper")))))
   (({above} none)
    (alt below-given
 ((({below} given)
   (alt below-what
((({below} 8)
  (lex "lower"))
 (({below} 7)
  (lex "lower")))))
  (({below} none)))))))
```

Figure 7.8: Rule for translating *above* to upper

rules used to produce the proper semantic input that SURGE and FUF need.)

The pair `(near calyx)` causes the language generation preprocessor to generate the semantic input that will produce the phrase *upper pole calyx* or *lower pole calyx* depending on whether the stone is *above* or *below* the middle calyx. The rule associated with this translation is shown in figure 7.10 and the partial semantic input that this rule creates is shown in figure 7.11.

The output produced by the language generation preprocessor for this example is shown in figure 7.12.

The final sentence produced by the natural language generator is

```
The right upper quadrant contains a density which probably represents
a stone in the upper pole calyx.
```

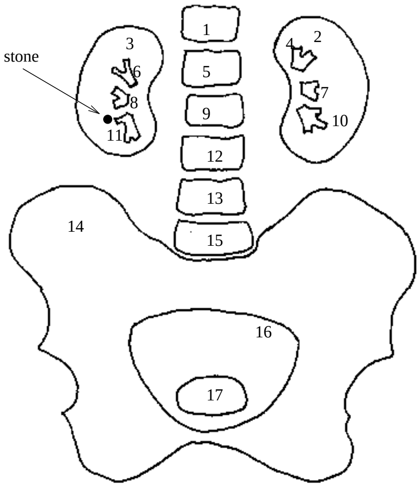


Figure 7.9: The model of the urinary system.

```
(alt calyceal
     ((({near} given)
       (alt near-what
    ((({near} 6)
      (:& calyceal-upper-lower-stone))
     (({near} 8))
     (({near} 7))
     (({near} 11)
      (:& calyceal-upper-lower-stone))
     (({near} 4)
      (:& calyceal-upper-lower-stone))
     (({near} 10)
      (:& calyceal-upper-lower-stone)))))
      (({near} none))))
```

Figure 7.10: Rule associated with translating *(near calyx)* to *upper pole calyx* or *lower pole calyx*

```
(def-conj calyceal-upper-lower-stone
  (circum
   ((location
     ((cat pp)
      (prep ((lex "in")))
      (np ((cat common)
   (lex "calyx")
   (describer
    ((cat list)
     (distinct
      ~((((:! upper-lower) (cat adj))
((lex "pole") (cat adj))))))))))))
```

Figure 7.11: The partial semantic input for generating the phrase *upper pole calyx* or *lower pole calyx*

7.4 Examples

7.4.1 Example 1: Distal Left Ureteral Stone

Figure 7.13 shows an original patient's X-ray. Figure 7.15 illustrate the edges found in figure 7.14. The image in figure 7.16 illustrates the transformation that was found for this X-ray. It is superimposed on the model to show the match. Figure 7.17 illustrates what the X-ray looks like after applying the transformation. Figure 7.18 shows the density that was found.

After applying the inference network minimization technique the spatial relations that resulted were `((inside inner-pelvis) (right inner-pelvis))`. This was then translated by the language generation preprocessor and the proper semantic input was sent to the language generator to produce the following sentence:

```
A density is seen in the distal left ureter.
```

```
(setq med7 '
      ((cat clause)
       (proc ((type locative)
              (mode equative)
              (lex "contain")))
       (partic ((identified
                 ((cat common)
                  (lex "quadrant")
                  (describer
                  ((cat list)
                   (distinct
                    ~(((cat adj) (lex "right"))
                      ((cat adj) (lex "upper")))))))
                  (identifier
                  ((cat common)
                   (lex "density")
                   (definite no)
                   (qualifier
                   ((cat clause)
                    (scope {^ partic processor})
                    (epistemic-modality "may")
                    (proc ((type mental)
                           (lex "represent")))
                    (partic ((phenomenon
                             ((cat common)
                              (definite no)
                              (lex "stone")))))
```

7.4.2 Example 2: Calyceal Stone

Figure 7.19 is an original patient's X-ray. Figure 7.20 is the image after being thresholded. Figure 7.21 illustrates the result of edge detection on figure 7.20. Figure 7.22 shows the result of applying the transformation on the edge image of figure 7.21. Figure 7.23 is the result of applying the transformation on the original X-ray of figure 7.19. Figure 7.24 depicts the location of the stone that was found. The spatial relations found were: `((inside right-kidney) (near calyx) (below middle-calyx))`. The sentence generated was:

```
                        (circum
                         ((location
                          ((cat pp)
                           (prep ((lex "in")))
                           (np ((cat common)
                                (lex "calyx")
                                (describer
                                ((cat list)
                                 (distinct
                                   ~(((cat adj) (lex "upper"))
                                     ((cat adj) (lex "pole"))))))))))))))))))
```

Figure 7.12: Output produced by the language generation preprocessor. It is the input to the natural language generator

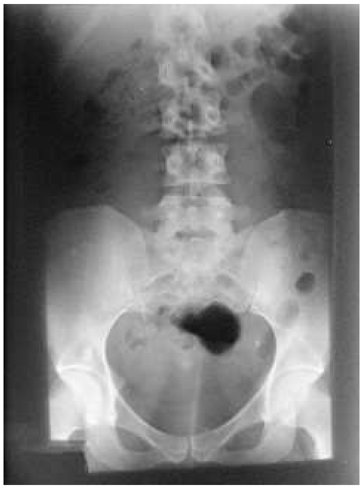


Figure 7.13: Example 1: Original X-ray

Figure 7.14 is the result of applying the binary algorithm on the original image.

```
   The right lower quadrant contains a density which may represent a stone

in the lower pole calyx.
```

7.4.3 Example 3: Mid-Ureteral Stone

Figure 7.25 is a third X-ray. Figure 7.26 is the image of figure 7.25 after it has been thresholded. Figure 7.27 illustrate the edges found in figure 7.26. Figure 7.28

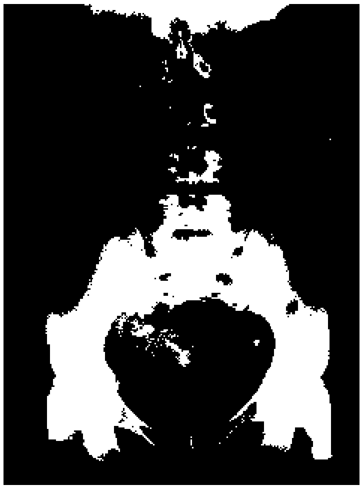


Figure 7.14: Example 1: Thresholded image

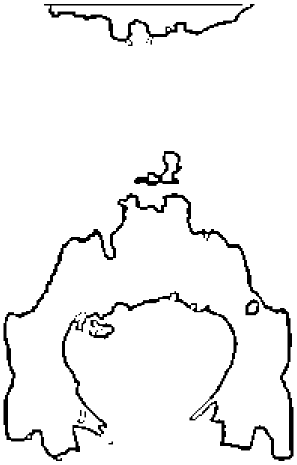


Figure 7.15: Example 1: Edge image

depicts the transformed edge image superimposed on the model to illustrate the match. Note that the match it made aligned the spine the most and not the pelvis region. The reason for this is the unusual shape of this patient's pelvis. In most cases the pelvis has a pronounced roundedness to it but the pelvis in this image is rather oblong therefore the model did not match very well with the pelvis but did a good job at matching the spine. For this example this was a good enough match to superimpose the stone correctly on the model and generate an appropriate description.

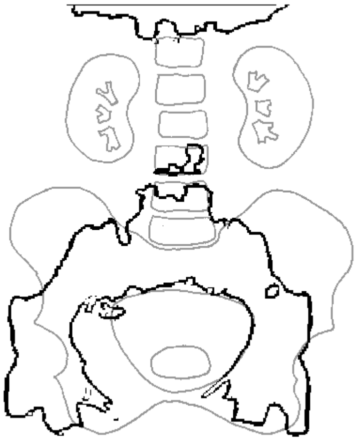


Figure 7.16: Example 1: Transformed edge image superimposed on the model

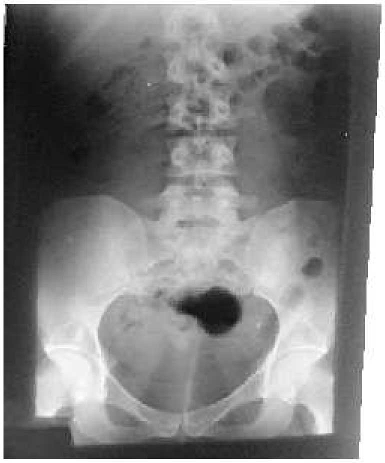


Figure 7.17: Example 1: Transformed X-ray

Figure 7.29 illustrates the transformation on the X-ray.

The spatial relation found for this example was `((left L4) (near right-kidney) (inside pelvis))`. The sentence that generated was

```
A density is seen at the level of L4 on the right which may represent
a stone in the right mid-ureter.
```

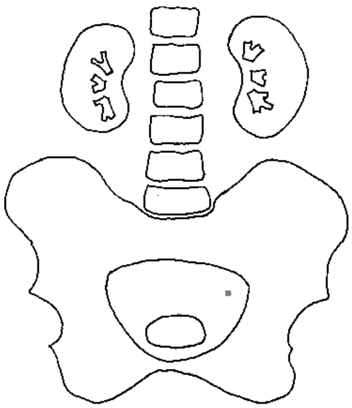


Figure 7.18: Example 1: Stone superimposed on the model

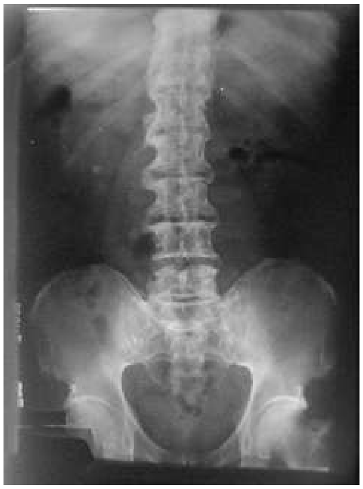


Figure 7.19: Example 2: Original X-ray

7.4.4 Example 4: Renal Stone

Figure 7.31 is the original X-ray. Figure 7.32 is the thresholded image of figure 7.31. Figure 7.33 contains the edges extracted from figure 7.32. Figure 7.34 is the transformed edge image superimposed on the model.

The spatial relation found for this example was `((inside right-kidney) (left calyx))`. The sentence that the natural language generator produced for this example was:

`The right kidney contains a density which may represent a right renal`

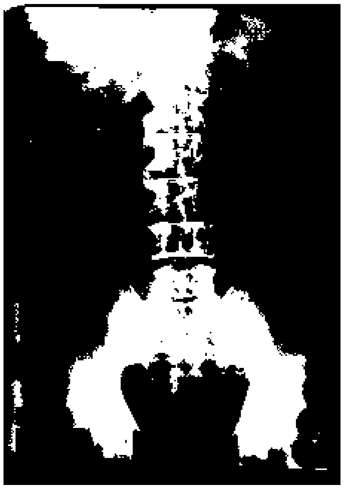


Figure 7.20: Example 2: Thresholded image

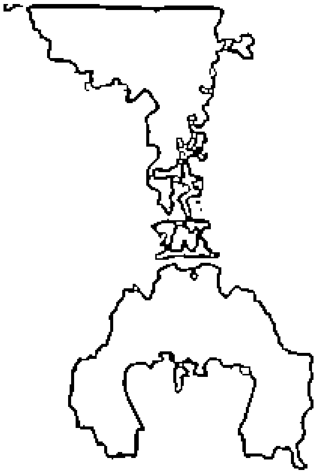


Figure 7.21: Example 2: Edge Image

`stone.`

7.5 Conclusion

This chapter covered the material necessary to gain an understanding of what is involved in the image processing and language generation preprocessing component of the system as it applies to the radiography domain. Because these images are noisier and objects in them are not separable, as they were in the landmark

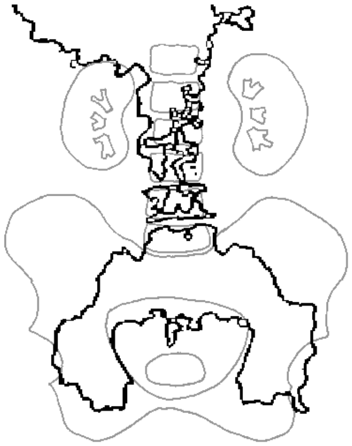


Figure 7.22: Example 2: Transformed edge image superimposed on the model

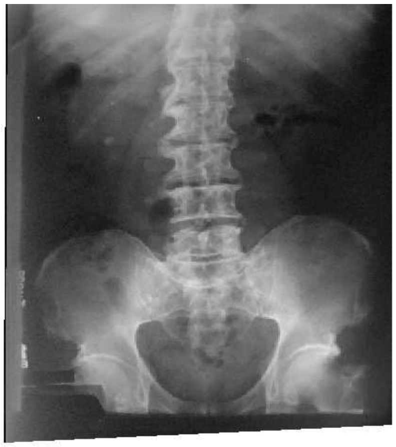


Figure 7.23: Example 2: Transformed X-ray

navigation domain, more work needs to be done by the image preprocessing module. It is this module that registers the image with a model of the image in order to transform the image. The registration process is essential in enabling a correct description to be generated. If the transformation that is found is not a good one then any densities that are found will not be superimposed in their appropriate place and the description that is generated may be incorrect and misleading. We were able to produce correct descriptions in four out of five cases. In the fifth case the stone was too small to be discriminated from random noise in the image. How-

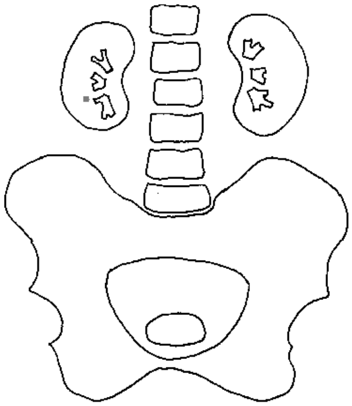


Figure 7.24: Example 2: Stone superimposed on the model

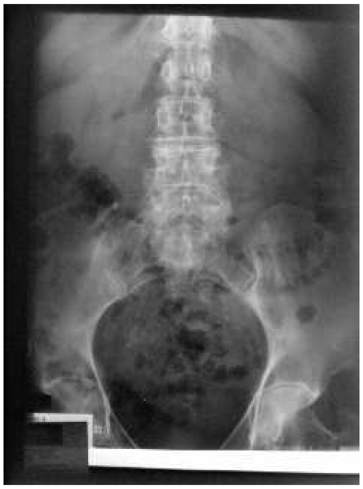


Figure 7.25: Example 3: Original X-ray

ever, this is not a weakness of the system since even the radiologist given no other information except the radiograph was not able to find the stone. Radiologists generally use clinical information to guide their examination. A future version of this system can do the same. Instead of just examining the radiograph it would also consult a knowledge base of clinical information. For example if the clinical information states that the person is suffering pain in his/her lower left side, then the system may narrow in on that portion of the radiograph and apply further imaging techniques in its search for some abnormality. The system does this now

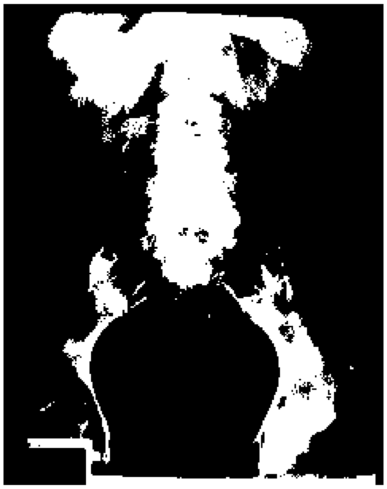


Figure 7.26: Example 3: Thresholded image

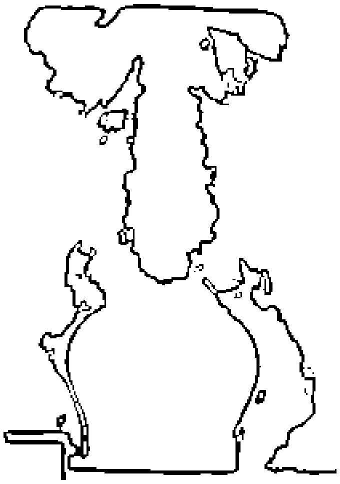


Figure 7.27: Example 3: Edge image

when it searches for a circle-like object. It searches the area outlined in figure 7.7 in order to narrow the search space and hence speed up the processing.

The language generation preprocessor consists of a preliminary grammar, like the one used for the landmark navigation domain, that takes a very simple input like `(left L4) (near L4)` and converts it into a form that can be understood by SURGE. The reason for using this preliminary grammar instead of generating the input that SURGE can understand is ease. FUF is designed to make this type of thing possible rather effortlessly. Without the use of FUF we would have

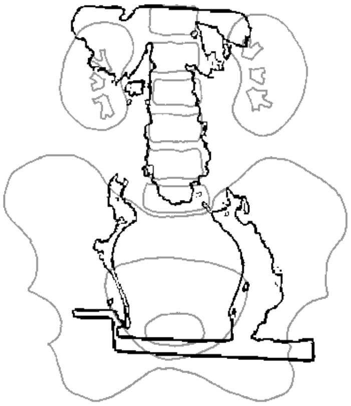


Figure 7.28: Example 3: Transformed edge image superimposed on the model

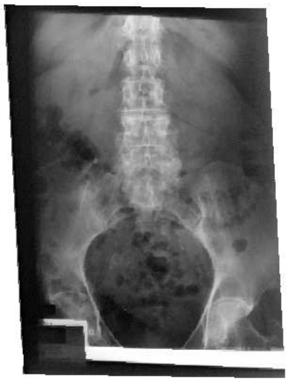


Figure 7.29: Example 3: Transformed X-ray

to resort to building a routine that would generate an input like the one shown in figure 7.12. This routine would quickly become very big, complicated, and probably not as efficient as FUF.

We included in the preliminary grammar the capability of generating the most commonly occurring descriptions. Further work to enhance the preliminary grammar could lead to generating less common descriptions. This would entail more consultation with the radiologists and additions to the existing preliminary grammar.

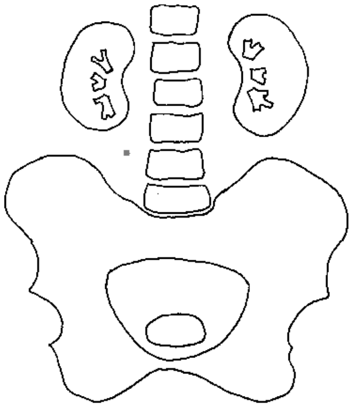


Figure 7.30: Example 3: Stone superimposed on the model

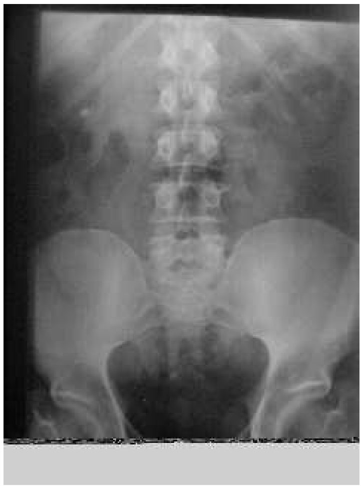


Figure 7.31: Example 4: Original X-ray

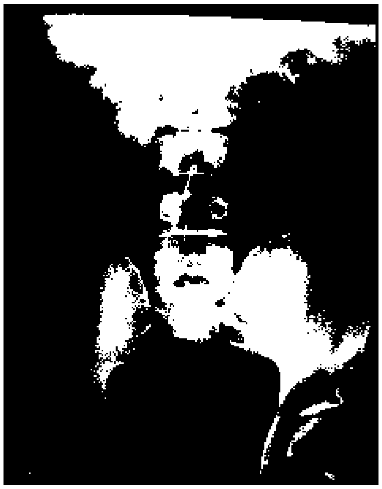


Figure 7.32: Example 4: Thresholded image

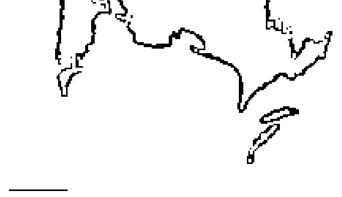


Figure 7.33: Example 4: Edge image

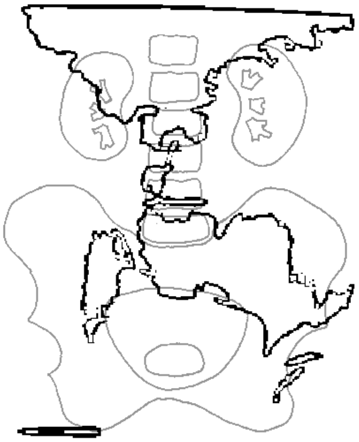


Figure 7.34: Example 4: Transformed X-ray superimposed on the model

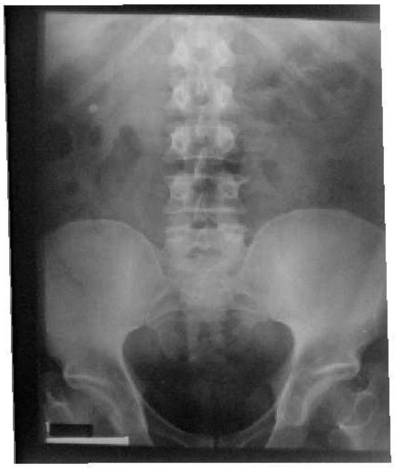


Figure 7.35: Example 4: Transformed X-ray

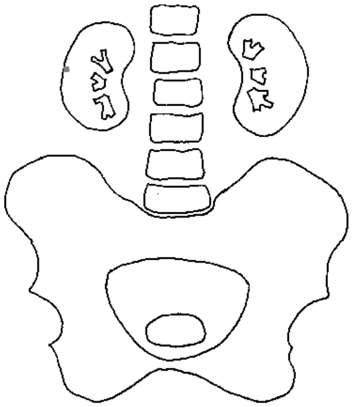


Figure 7.36: Example 4: Stone superimposed on the model

Chapter 8

Conclusion

This thesis developed a computer system that integrates image and natural language processing techniques for performing tasks that involve communicating visual information. The visual information that the system conveys is the location of objects in an image. It conveys this information through the use of prepositions, intrinsic object properties and superlatives. In order to be able to use these language constructs we developed a semantic representation for each of them based on the fundamental assumption that objects may be represented as "blobs". Each representation may be calibrated according to the specific domain in order to adjust to the variability of users in different domains. For example, what is *near* in one domain may not be *near* in another. This was the case in the example associated with figure 3.20, where the people reading the description are expected to be traveling by car and not by foot, therefore the interpretation of *near* is different.

With these semantic representations the system is able to compute all the possible ways that the location of an object can be described given the set of semantic representations available. All these spatial relationships are not necessary, in fact, too many of them lead to incomprehensible descriptions. We devised two methods for optimizing these relationships. One method is based on Boolean formula minimization and the other is based on inference network minimization. These technique are a novel approach to the problem of generating the smallest yet most discriminating set of spatial relations for object pairs in an image.

The Boolean formula minimization method was designed to be used in situations where the recipient of the description is a novice. A novice is one who requires an explicit set of descriptions regarding the environment. A novice does not know anything about any of the objects in the image except for his/her current position in the environment. This means the system needs to generate descriptions for as many referents (landmarks) in the image as needed. The inference network minimization method was designed to be used in situations where the recipient of the description is an expert. An expert is familiar with the environment and only requires that the intended figure object be described in terms of the least number of referents as possible. Also the middle ground exists, in which it would be necessary to combine the two methods. Perhaps a person is an expert in a portion of the environment but not all. Certainly this person does not require that what he/she knows be made explicit, but it is necessary to make explicit what he/she is not familiar with. This would require interaction between the system and the user. The system would need to know what the user is familiar with and what he/she wishes to know.

It is not always the case that the minimized set of spatial relations that the system finds is all that discriminating. It can be the case that this set may convey information about several object in the image. We asked the question, "How likely will a recipient choose the intended figure object?" To answer this question required that we model user behavior and error. Based on this model we defined a vagueness measure that measures the discriminating ability of the set of spatial relations: the higher the vagueness measure the higher the chances of misinterpretation on the part of the recipient.

Each method for generating the smallest set of spatial relations factors in context. We are able to quantitatively capture context via both minimization techniques as well as some domain specifics. This is especially true in the radiography domain where certain spatial relations have been promoted into medical terminology. For example, if a stone is *near* the fourth lumbar body and it is to the *left* of it then this is considered a "mid-ureteral stone lateral to the L4 process". Nowhere do the words *left* and *near* appear in the description. This phenomenon is captured by the language generation preprocessor, so that the system can convey this type of information.

Because of the nature of the minimization techniques, adding new language constructs requires minimal effort. The Boolean formula minimization technique would proceed in the same fashion irrespective of the number or type of language constructs. The implication network, on the other hand, requires that inferences be possible. If we can find implications that we can prove the way we proved them in section 5.3 then we may add them to the inference network of figure 5.3. Object properties like color or size could not be incorporated into such an implication network simply because an object's color can not imply any information regarding the color of another object.

The methods presented in this thesis can also be applied to other domains like weather forecasting where clouds and rain storms can be represented as "blobs" Their locations are with respect to the area that they are covering which may include landmarks, like cities, towns, reservoirs, mountains, lakes etc. For example, "A storm cloud is sitting *over* the Adirondack Mountains." We have the "what" addressed by size, and the "where" addressed by prepositions, but we may also apply our techniques to "when". We may talk about motion in terms of temporal prepositions, like *towards*. Instead of one static image we may use several images taken at different time instances and then find the optimal set of temporal prepositions that describe the changes seen in the image. This, for example, may be used in studying radiographs taken over time. Radiologists will sometimes examine several radiographs of a patient over time to track and record the movement of stones.

While we have only included size as an object's intrinsic property, we may elaborate this to include such properties as color, texture, reflectance (shiny, dull), and shape (circular, trapezoidal, especially for the radiography domain, where radiologists on occasion will describe the shape of a stone as trapezoidal as opposed to circular or elliptical).

There are numerous possible research directions. Some of the global topics are discussed in the following paragraphs.

The integration of Boolean formula minimization and inference network minimization This integration is necessary if we want to be able to handle the middle ground. The user in the middle ground is familiar with some parts of the environment and not others. In this scenario we would need, for example, a dialog manager that would acquire the knowledge that the user possess about the environment. This would force the system to generate descriptions on the fly. In the middle ground the generation of descriptions must be interactive since the computational complexity of anticipating user knowledge is infeasible. While this thesis did not address this issue it developed methods on each end that could provide the basis for their integration.

Development of computational models for different language constructs An augmentation of the computational models that we have would include temporal prepositions; prepositions such as *towards, along, away.* A potential application of temporal prepositions is for describing the movement of kidney stones. Radiologists often take various radiographs of a patient over time to track the movement of kidney stones. To be able to incorporate temporal prepositions into our vocabulary does not necessitate the development of any new computational models. It would necessitate a bit more work on the image processing module. What we may do is to take snapshots of the environment at different time instances and determine if any objects have moved, and if so in what manner. This requires that we perform object recognition since we need to know which object has moved so that we can talk about it. Once we have this we may use, for example, the preposition *towards* if we detect that an object in a previous image is now nearer to an object than it was before. Likewise, we may use the preposition *away* if we detect that an object in a previous image is now farther away from an object. We may choose to create computational models for language contructs other than prepositions, for example, verbs of motion, or additional intrinsic properties, like texture and color. Again it would seem that the brunt of the work lies in the image processing module.

Psychological confirmation of the computational models of the language constructs. In section 3.4 we performed a preliminary validation of the computational models of our language constructs. Validation is an important part of generating results that are as intuitive as possible and would require a thesis' worth of work to see more precisely how people view shapes and relations of various sizes. For example, it would be of interest to see when the six parameter representation of shape fails, possibly for near-linear objects like roads. Objects like roads or rivers may have a distinct representation, and interaction with other objects: a road next to a river is different from a building next to a river, etc. The work described in section 3.4 is a first step in determining if the computational models that we defined capture the meaning of the various language constructs.

Application of the system to different specific domains Meteorology is an area that can see direct application of the methods discussed in this thesis. Weather maps contain objects that are already blobs, these objects are clouds. Descriptions are generally given in terms of cloud coverage. For example, "There is heavy cloud coverage *over* most of the Metropolitan Area", "A storm is heading *towards* the East coast." The clouds in the weather maps are the objects of interest along with their spatial relationship to geographical landmarks, such as mountain ranges, cities, towns, etc. We can expand the vocabulary to include prepositions such as *over* and *towards*, but those prepositions that we have can also apply.

Verification of the error model The error model defined in chapter 4 was based on some preliminary findings that indicate that people are in general agreement when an object pair certainly satisfies a locative expression and when it certainly does not. The uncertainty lies somewhere in the middle when values for the locative expression are around 0.5. Further experiments, such as asking people to rate, on a scale from 1 to 10, their belief that certain object pairs satisfy given prepositions will give us more of an idea of, for example, where the largest interval of discrepancy in the error model should be. We can also use variations of the error model and see how different error models affect the descriptions that are generated. To be more precise will require that we study and apply the methodology used in psychology and apply statistics rather than an estimate. However, the system as demonstrated would not have to be substantially changed, since the error model is a plug-in function.

Verification of user preferences It would be interesting to determine how people would choose to combine prepositions, superlatives and intrinsic object properties to describe where objects are. For example, how likely are people going to describe the figure object of figure 4.11 as the *topmost and leftmost* one? Do people combine more than one superlative in a sentence to describe an object? Are descriptions of the form $s\bar{l} \wedge s\bar{l} \ldots \wedge s\bar{l}$ (where s is a superlative and $\bar{l}$ is a locative expression) natural, or do people combine superlatives and locative expressions differently? we can answer some of these questions by performing more tests on people or by using the statistical tools of natural language processing, which scans corpora of text looking for such occurrences. Statistical natural language processing is a new and growing, and some of the methodology may not be appropriate, particularly since its emphasis does not yet include prepositions.

The following paragraphs contain some of the more local areas of improvement.

Improvements to the image processing module The biggest bottle neck in the radiography domain is image registration and the stone localization algorithm. Additional algorithms for cleaning up the images and properly registering the image are necessary. Using the gradient descent search methods on matching the image to models would increase speedup enormously. If the image is not registered properly it is likely that the description that is generated will be incorrect since the spatial relationships will not be correct. Labeling an object in the image as a stone when in reality it is not a stone is an obvious concern. Besides making the image processing aspects of stone localization more sophisticated we may also incorporate clinical information. Radiologist are not simply given a radiograph and asked to find any anomalies, they are also given clinical information, such as the possible source of a patient's pain, which assists the radiologist in narrowing the search for possible densities. Being able to access this kind of information would help reduce the search space and thus improve efficiency and improve the chances of locating the correct stone.

Expansion of the preliminary grammar Currently the system can generate descriptions of stones that are found in typical scenarios. Additional rules that translate the spatial relations into proper medical terminology for atypical scenarios would enhance the usability of the system. In general, this work derived its results from only about a dozen images. "Typical" would have to be defined by talking to the experts. It is not clear what the limits to this process would be, and when there would come a point of diminishing returns.

Different kidney diseases The ability to describe different kidney diseases such as tumors or phleboliths is another enhancement. Pathology reports will contain more than just the location of a kidney stone if the radiologist also finds phleboliths.

Phleboliths are characteristically very small white calcific densities that have the property of occuring in groups. Occasionally a single phlebolith will occur and it may be difficult to distinguish it from a stone.

What we have realized in this system is the potential available for integrating vision and natural language. This is a relatively unexplored area that certainly merits more exploration especially now when we have reached the technological stage where multimedia is no longer a thing of the future. It is important to integrate different modes of communication (audio, visual, textual) if we are to build systems capable of interacting with a user in a humanly fashion. Certainly the time has come to integrate all those facets that we have been studying for years and create systems capable of doing what people have been doing for centuries.

Bibliography

[Abella and Kender, 1993] Alicia Abella and John R. Kender. Qualitatively describing objects using spatial prepositions. In *Eleventh National Conference on Artificial Intelligence*, 1993.

[Abella and Kender, 1994a] Alicia Abella and John R. Kender. Conveying spatial information using vision and natural language. In *AAAI Integration of vision and natural language processing workshop*, 1994.

[Abella and Kender, 1994b] Alicia Abella and John R. Kender. From pictures to words: Generating locative descriptions of objects in an image. In *Image Understanding Workshop*, 1994.

[Abella, January 1992] Alicia Abella. Extracting geometric shapes from a set of points. In *Image Understanding Workshop*, January 1992.

[Acredolo and Evans, 1980] L. P. Acredolo and D. Evans. Developmental changes in the effects of landmarks on infants' spatial behavior. *Developmental Psychology*, 16:312–318, 1980.

[Acredolo, 1978] Linda P. Acredolo. Development of spatial orientation in infancy. *Developmental Psychology*, 14:224–234, 1978.

[Aho *et al.*, 1974] Alfred V. Aho, John E. Hopcroft, and Jeffrey Ullman. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, 1974.

[Becker and Arms, 1969] A.L. Becker and D.G. Arms. Prepositions as predicates. *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, 1969.

[Bennett, 1968] D.C. Bennett. English prepositions: A stratificational approach. *Journal of Linguistics*, 4:153–172, 1968.

[Bennett, 1972] D.C. Bennett. Some observations concerning the locative-directional distinction. *Semiotica*, 1972.

[Bremner, 1978] J.G. Bremner. Spatial erros made by infants: Inadequete spatial cues or evidence of egocentrism? *British Journal of Psychology*, 69:77–84, 1978.

[Clark, 1977] Eve V. Clark. Strategies and the mapping problem in first language acquisition. In *Language Learning and thought.* Academic Press, 1977.

[Dorr and Voss, 1993] Bonnie J. Dorr and Clare R. Voss. Machine translation of spatial expressions: Defining the relation between an interlingua and a knowledge representation system. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 1993.

[Elhadad, 1993] Michael Elhadad. *FUF: The Universal Unifier*, 1993.

[Farah *et al.*, 1988] M. Farah, K. Hammond, D. Levine, and R. Calvanio. Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology*, 20:439–462, 1988.

[Fillmore, 1968] C. Fillmore. Universals in linguistic theory. In E. Bach and R Harms, editors, *The case for case.* Rinehart and Winston, 1968.

[Fox and Walker, 1989] John Fox and Nicholas Walker. Knowledge based interpretation of medical images. In *Medical Imaging*, 1989.

[Glasgow and Papadias, 1992] J. Glasgow and D. Papadias. Computational imagery. *Cognitive Science*, 16:355–394, 1992.

[Herskovits, 1986] A. Herskovits. *Language and Spatial Cognition: An interdisciplinary study of the prepositions in English.* Cambridge University Press, 1986.

[ichi Takami, 1992] Ken ichi Takami. *Preposition Stranding From Syntactic to Functional Analyses.* Walter de Gruyter & Co., 1992.

[Jackendoff, 1987a] R. Jackendoff. *Consciousness and the Computational Mind.* MIT Press, 1987.

[Jackendoff, 1987b] R. Jackendoff. *Consciousness and the Computational Mind.* The MIT Press, 1987.

[Jackendoff, 1992] R. Jackendoff. *Languages of the Mind.* The MIT Press, 1992.

[J.Piaget and Inhelder, 1956] J.Piaget and B. Inhelder. *The child's conception of space.* Routledge & Kegan Paul, 1956.

[Kandel and Lee, 1979] Abraham Kandel and Samuel C. Lee. *Fuzzy Switching and Automata: Theory and Applications.* Edward Arnold, 1979.

[Kautz, 1985] Henry A. Kautz. Formalizing spatial concepts and spatial language. Technical report, Stanford University, 1985.

[Keating and McKenzie, 1986] M. B. Keating and B. E. McKenzie. Constancy in a square and circular room with and without a landmark. *Child Development*, 57:115–124, 1986.

[Kender and Leff, 1989] J.R. Kender and Avraham Leff. Why direction-giving is hard: The complexity of linear navigation by landmarks. *IEEE Transactions on Systems, Man and Cybernetics*, 1989.

[Kender *et al.*, 1990] J. R. Kender, Il-Pyung Park, and David Yang. A formalization and implementation of topological visual navigation in two dimensions. In *SPIE International Symposia*, 1990.

[Kimme *et al.*, February 1975] Carolyn Kimme, Dana Ballard, and Jack Sklansky. Finding circles by an array of accumulators. In *Communications of the ACM*, volume 18, February 1975.

[Klir and Folger, 1988] G. J. Klir and T. A. Folger. *Fuzzy Sets, Uncertainty and Information.* Prentice Hall, 1988.

[Kobashi and Shapiro, 1992] Masaharu Kobashi and Linda G. Shapiro. Knowledge-based organ identification from ct images. In *Medical Imaging VI: Image Processing*, 1992.

[Kosslyn, 1980] S.M. Kosslyn. *Image and Mind.* Harvard University Press, 1980.

[Kuipers and Levitt, 1988] Benjamin J. Kuipers and Tod S. Levitt. Navigation and mapping in large-scale space. *AI Magazine*, 1988.

[Kuipers, 1978] Benjamin J. Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2:129–153, 1978.

[Landau and Stecker, 1990] Barbara Landau and Deanna S. Stecker. Objects and places: Geometric and syntactic representations in early lexical learning. *Cognitive Development*, 5:287–312, 1990.

[Lea, 1975] G. Lea. Chronometric analysis of the method of loci. *Journal of Experimental Psychology: Human Perception and Performance*, 2:95–104, 1975.

[Leech, 1969] G. Leech. *Towards a semantic description of English.* Longman Press, 1969.

[Levine *et al.*, 1985] D. Levine, J. Warach, and M. Farah. Two visual systems in mental imagery: Dissociation of 'what' and 'where' in imagery disorders due to bilateral posterior cerebral lesions. *Neurology*, 35:1010–1018, 1985.

[Lindvist, 1976] K. Lindvist. *Comprehensive study of conceptions of locality in which English prepositions occur.* Almqvist & Wiksell International, 1976.

[Lu *et al.*, 1992] Yi Lu, Lucia Zamorano, Federico Moure, and Steven Schlosser. Automatic detection of boundaries of brain tumors. In *Medical Imaging VI: Image Processing*, 1992.

[Lynch, 1960] K. Lynch. *The image of the city.* The Technology Press and Harvard Press, 1960.

[Macnamara, 1978] J. Macnamara. *How do we talk about what we see?* Mimeo, McGill University, 1978.

[Mandler, 1988] J. Mandler. The development of spatial cognition: On topological and euclidean representation. In *Spatial Cognition.* Hillsdale, NJ: Erlbaum, 1988.

[Marburger *et al.*, 1981] H. Marburger, B. Neumann, and H-J Novak. Natural language dialogue about moving objects in an automatically analyzed traffic scene. In *Proceedings 7th International Joint Conference Artificial Intelligence*, 1981.

[McKeown *et al.*, 1990] K. R. McKeown, M. Elhadad, Y. Fukumoto, J.G. Lim, C. Lombardi, J. Robin, and F.A. Smadja. Text generation in comet. In R. Dale, C.S. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation.* Academic Press, 1990.

[Miller and Johnson-Laird, 1976] G.A. Miller and P.N. Johnson-Laird. *Language and Perception.* Harvard University Press, 1976.

[Mukerjee and Joe, 1990] Amitabha Mukerjee and Gene Joe. A qualitative model of space. In *Proceedings Eighth National Conference on Artificial Intelligence*, pages 721–727, 1990.

[Nagel, 1988] H-H Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2), 1988.

[Neumann and Novak, 1983] Bernd Neumann and Hans-Joachim Novak. Event models for recognition and natural language description of events in real-world image sequences. In *Proceedings 8th International Joint Conference Artificial Intelligence*, 1983.

[Neumann, 1984] Bernd Neumann. Natural language access to image sequences: event recognition and verbalization. In *Proceedings 1st Conference Artificial Intelligence Applications*, 1984.

[Park, 1993] Il-Pyung Park. *Qualitative Environmental Navigation: Theory and Practice.* PhD thesis, Columbia University, 1993.

[Pinker, 1984] S. Pinker. Visual cognition: An introduction. *Cognition*, 18:1–63, 1984.

[Retz-Schmidt, Summer 1988] Gudula Retz-Schmidt. Various views on spatial prepositions. *AI Magazine*, Summer 1988.

[Rieser and Heiman, 1982] J. J. Rieser and M. L. Heiman. Spatail self-reference systems and shortest-route behavior in toddlers. *Child Development*, 53:524–533, 1982.

[Robin, 1994] Jacques Robin. *Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation and Evaluation.* PhD thesis, Columbia University, 1994.

[Rueckl *et al.*, 1989] J. Rueckl, K. Cave, and S. Kosslyn. Why are 'what' and 'where' processed by seperate cortical visual systems? a computational investigation. *Journal of Cognitive Neuroscience*, 1:171–186, 1989.

[Selfridge and Prewitt, 1981] Peter G. Selfridge and Judith M. S. Prewitt. Organ detection in abdominal computerized tomography scans: Application to the kidney. In *Computer Graphics and Image Processing*, volume 15, pages 265–278, 1981.

[Srihari, 1991a] Rohini K. Srihari. *Extracting Visual Information from Text: Using Captions to Label Faces in Newspaper Photographs.* PhD thesis, State University of New York at Buffalo, 1991.

[Srihari, 1991b] Rohini K. Srihari. Piction: A system that uses captions to label human faces in newspaper photographs. In *Procceddings of AAAI-91 (Anaheim)*, 1991.

[Suppes, 1992] Patrick Suppes. *Language for Humans and Robots.* Blackwell Publishers, 1992.

[Tagare *et al.*, 1993] Hemant D. Tagare, Frans Vos, Conrade C. Jaffe, and James S. Duncan. Arrangement: A spatial relation comparing part embeddings and its use in medical image comparisons. In *bla bla*, 1993.

[Talmy, 1983] L. Talmy. How language structures space. In *Spatial Orientation Theory, Research, and Application.* Plenum Press, 1983.

[Ungerleider and Mishkin, 1982] L.G. Ungerleider and M. Mishkin. Two cortical visual systems. In *Analysis of Visual Behavior*, pages 549–586. MIT Press, 1982.

[Urban, 1939] W.M. Urban. *Language and reality: the philosophy of language and the principles of symbolism.* Allen & Unwin, 1939.

[Wahlster *et al.*, 1983] Wolfgang Wahlster, Heinz Marburger, Anthony Jameson, and Stephan Busemann. Over-answering yes-no questions: Extended responses in a nl interface to a vision system. In *International Joint Conference Artificial Intelligence*, 1983.

[Wechsler and Sklansky, 1977] H. Wechsler and J. Sklansky. Finding the rib cage in chest radiographs. In *Pattern Recognition*, 1977.

[Winograd, 1972] Terry Winograd. *Understanding Natural Language.* Academic Press, 1972.

[Winograd, 1973] Terry Winograd. A procedural model of language understanding. In *Computer Models of Thought and Language.* W.H. Freeman and Company, 1973.

[Wishart and Bower, 1982] Jennifer G. Wishart and T. G. R. Bower. The development of spatial understanding in infancy. *Journal of Experimental Child Psychology*, 33:363–385, 1982.

[Zernik and Vivier, 1988] Uri Zernik and Barbara Vivier. How near is too far? talking about visual images. *Proceedings of the 10th Annual Conference of the Cognitive Science Society*, pages 202–208, 1988.