ICS-08M-632

USING SPATIAL INFORMATION FOR INTERACTIVE REFERENCE RESOLUTION

by

Tajin Rukhsana Tarannum

Supervisor: Professor Yoshinori Kuno

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science.

Date of Submission: February 5, 2008

Department of Information and Computer Sciences

Graduate School of Science and Engineering

Saitama University
255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338-8570
Japan

ACKNOWLEDGEMENT

This research was carried out in the Graduate School of Science and Engineering of Saitama University, Japan under the supervision of Prof. Yoshinori Kuno who had been a vigilant and an enthusiastic supervisor from its embryonic stage. It would not have been possible for the author to accomplish this research without his supports. The author expresses her deep sense of gratitude and profound indebtedness to Prof. Yoshinori Kuno for his affectionate guidance, continuous support, encouragement, valuable suggestions and untiring efforts in this regard.

Sincere appreciation and gratefulness is expressed to acknowledge the valuable supports of all colleagues in the Computer Vision Laboratory. From time to time they have extended their helpful hand in providing resources and suggestions for this research.

The author is also grateful to the Bangladeshi students in Saitama University as they have participated in the survey, required for the thesis. Without their spontaneous involvement the survey would not be carried out smoothly.

Last but not the least, the author is highly indebted to her husband for his continuous encouragement and support through out the period of Master's degree. All friends of the author in Saitama University also deserve thanks and gratefulness for making her time enjoyable and memorable here in Japan.

ABSTRACT

Carrying out user commands entails target object detection for service robots. When the robot system suffers from a limited object detection capability, effective communication between the user and the robot facilitates the reference resolution. We aim to develop a service robot, assisting people inside home or small office environments where, most of the user requests are directly or indirectly linked to some objects in the scene. Objects can be described using features like color, shape, size etc. For simple objects on simple backgrounds, these attributes can be determined with satisfactory results. For complex scenes, position of an object and spatial relation with other objects in the scene, facilitate target object detection.

Our robot system is assumed to recognize some object classes and specific objects. How human users, being aware of the limited object detection capability of their robot partner, describe objects in images is of primary interest. A survey is conducted in this regard. Results show that color and spatial relation among objects are the mostly used attribute by the participants. Among several aspects of communication yielded from the experiment results, we have chosen “Feedback generation” to implement into our robot system. An algorithm has been constructed for this purpose. By analyzing the dialogs derived from the survey, vocabulary list for attributes and input style of users have been identified. Then a language parser has been designed to extract attributes from user input. Moreover, number of known objects has an impact on designing the query of the robot. This research also proposes a method to choose optimal reference among the known objects so that generated queries lead to target object detection.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background

Service robots assist human beings, typically by performing a job that is dirty, dull, distant, dangerous or repetitive, including household chores. They typically are autonomous and/or operated by a build in control system, with manual override options. Home automation is becoming a viable option for the elderly and disabled who would prefer to stay in the comfort of their home rather than move to a healthcare facility. . These systems make normal Activities of Daily Living (ADL) possible for the elderly and disabled who would otherwise not be able to live on their own. Service robots are also being developed to provide service in various places such as museums, shopping malls, care homes and restaurants. These are expected to carry out user requests. We assume that the service robot is assisting people inside home or small office environment where most of the user requests are directly or indirectly linked to some objects in the scene. For example, “Get me the coffee jar” or “Bring the middle book”. To interpret the user command, the robot has to locate the objects first. This work is associated with the target object localization.

1.2 What is Reference Resolution?

The term “Reference Resolution” has been used in a large body of work in linguistics. It is a process that accepts a term description and rewrites it to account for information available in the discourse context (Donna and James, 2002). Reference resolution algorithms are applicable for the processing of discourse and dialogue.

In this work we aim to develop an interactive method for a service robot that can be applied in identifying the desired object of a user. Since the interaction module entails comprehension of dialogs in the discourse context to identify the target, we refer to this “Target object identification” as “Reference Resolution”.

1.3 Significance of Interaction in Human-Robot Communication

Humanoid robots which are able to walk and behave human-like became very popular in the last few years. Now it is high time that they are able to use more natural communication means so that the human-robot interaction resembles more and more to human-human communication (Gieselmann, 2004). The role of unimodal (only through vision or speech) or multimodal (involving text, speech, vision, sonic sensor etc.) interaction between human and robot has been highlighted in research for many years. The results in (Fischer and Lohse, 2007) suggest that verbal robot output is a powerful means for guiding users into an understanding of robot’s capabilities. The authors also describe how homogenous users are in their beliefs about their interaction partner and how these beliefs determine the users’ linguistic behavior. In (Brenner, 2007) authors have stated that intelligent service robots are expected in the near future to understand NL (natural language) commands.

The authors of (Lopes and Teixeira, 2000) are involved in CARL, a project aimed at contributing to the development of task-level robot systems. This paper focuses on the human-robot interface. The main claim is that the only acceptable user interface for a

task-level robot is a spoken language interface.

In (Schultz and Trafton, 2006) authors have made a hypothesis that, a system using human-like representations and processes will enable better collaboration with people than a computational system that does not. Their findings suggest that similar representations and reasoning mechanisms make it easier for humans to work with the system. For close collaboration, systems should act "naturally" i.e. not do something or say something in a way that detracts from the interaction/collaboration with the human. Robot should accommodate humans; not other way around. Any interface which is to support collaboration between humans and robots must include a natural language component (Sofge et al., 2003).

1.4 Problem Statement

Humans have a vast cognitive database, which help them interpret requests without any error. For example, we can detect objects in any kind of their orientation, shape, size and texture once we learn what object it is. The limitation of a robot in this regard is its very small range of knowledge about the environment. Compared to humans, a robot has a significantly poor capability of storing the vast data of specific objects or object categories and manipulating those data to detect a target object. Moreover, no single recognition method is enough for all object classes. Even though an object model is in robot database, it may not be recognized in a scene. Instances of false detection are also very common.

To remove the load of object recognition for a robot, we utilize the spatial relations between objects, as well as other descriptors, in a scene where some objects are already recognized by the robot. Through generation of speech with the user, the robot attempts to detect the target object.

1.5 Objective

Our main goal is to incorporate positional relation among objects into the system of interactive reference resolution. To achieve this we proceed by fulfilling some objectives and secondary objectives. They are stated below:

1) To examine how humans describe objects in a scene
    a) Do humans find only spatial relation sufficient for objects description?
    b) What are the other attributes except positional information, used as descriptors?
    c) In what situation humans choose a specific descriptor?

2) To overview spatial understanding of human beings
    a) How do humans build model about the environment and how does this model affect generation of spatial language?
    b) Which perspectives and frame of references are used in spatial reasoning?
    c) How do people mentally represent spatial relationships?
    d) How do people naturally use language to communicate about spatial relationships?

3) To analyze the effect of communication partner on spatial reasoning of humans
    a) What are the factors that shape human understanding of partner's ability?

b) How do human-human and human-robot spatial communications differ?

4) To gain an understanding of linguistic choices of human in discourse
   a) How do human collaborate with an intelligent machine-like partner?
   b) How do humans ground their understanding about the environment?
   c) In which direction a discourse is continued to detect the target object in a scene?
   d) Which words are used as descriptors?

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Formation of spatial understanding and its application in human-human communication have been focused by cognitive psychologists for many years. Researchers in the field of linguistics also have worked on this issue. How human beings reason about an environment based on spatial information and how they produce spatial terms to fit a situation, are the key interest areas of the above-mentioned group of researchers. Scientists in robotics are also dealing with spatial reasoning. The more robots and intelligent systems are being involved in our daily life, the more spatial understanding is coming into attention. Human-robot interaction entails manipulating position of both the participants and the objects in environment. Following sections of this chapter survey work on linguistics and robotics that are related to spatial information.

2.2 Linguistics and Spatial Reasoning

2.2.1 How Our Language Corresponds to Environment

Authors in (Space and Language, 1999) are interested in how people talk about space and what they can and do choose to say about it. They state that one cannot learn a language unless one has an original language (language of thought) to structure the learning process.

Through detailed analysis it is shown in (Space and Language, 1999) that, spatial terms cannot be derived simply from an interface between language and a set of sensory/perceptual maps. For example, let us consider the expressions “The butterfly is in the jar/on the table/in the canyon.” Here the meaning of “in” does not simply map to surroundedness in the visual display. One must appeal to some abstract relationship, such as a capacity for containment that jars and canyon share, but tabletops do not. Moreover, although there is a relationship between category of nouns and the notion of object shape, it is mediated through a more abstract conceptual system of conceptual representations. When we name objects in day-to-day speech, we are most likely to choose a name which is neither too general nor too specific (Pattabhiraman, 1992).

2.2.2 Meaning of Locative Sentences

In (Meaning of Locative Sentences, 2005) authors have studied the meaning of locative sentences involving directional terms such as in front of, right of etc. First they contrasted two spatial communication tasks: pointing to objects in a layout and telling their direction. It has been observed that after imagining a body rotation, pointing was considerably slowed with respect to a physical body rotation, whereas performance in the verbal location task was similar under imaginary and physical rotation. The authors here propose that producing locative sentences, unlike pointing to objects, involves a second-order embodiment. That is, language spatial relations are represented and updated into a mental framework that is detached from body, but still preserves spatial relations analogically.

2.2.3 Frame of Reference

In localizing reference objects in space, humans have - broadly speaking - three kinds of reference systems at their disposal, which may be called *intrinsic, relative* and *absolute* (Levinson, 1996).

In **intrinsic** reference systems, objects are located by referring to the intrinsic properties of another entity, such as the speaker's front in *The ball is in front of me.*

Relative reference systems depend on the presence of a further entity (the so-called *relatum*), as in *The ball is in front of the table.*

Absolute reference systems depend on the earth's cardinal directions, such as *north* or *south.*

Additionally, speakers may variously employ either their own or their listener's *point of view* (also called *origin*) - or, which in some situations may also be useful, the perspective of a third entity (as in, *Viewed from the church's entrance, there is a bookshop on the right*).

In (Levinson, 1996) another kind of reference, deictic (reference based on the speaker) reference has been included in the same group as intrinsic. For intrinsic system ground and origin are the same object (spatial relation is binary, for deictic that is the speaker), while for relative system, the relation is ternary (figure, ground, origin of frame).

2.2.4 Influence of Functional Orientation on Frame of reference

In an experiment in (Taylor et al., 2000) it is shown that when objects were presented in a functional orientation, participants were more likely to use an intrinsic versus a deictic reference frame than when objects were presented in a non-functional orientation. This finding indicates that the nature of the objects being described, as well as relationships between them, are taken into account when people produce natural language descriptions of spatial arrangements. Thus, when designing systems which are intended to produce spatial descriptions in natural language, factors such as the functionality of the orientation between functionally related objects should influence the spatial description.

2.2.5 Problem Using "right" / "left"

Part of the problem in using the terms right/left may arise because egocentric spaces depend on the direction the individual is facing (Spatial Perspective in Descriptions, 1999). That is, the regions to the left and right are interchanged by a 180^{o} rotation. For a speaker and a listener who are facing each other, the space to the right of the speaker lies to the left of the listener.

The asymmetries of the front/back body axis are most salient because they separate the world that can be easily sensed and easily manipulated from the world that is difficult to sense or manipulate. The head/feet axis is next most salient, for its asymmetries, and the left/right axis is least salient.

2.2.6 In or On?

When should we use ‘in’ or when ‘on’ is better? This question is highlighted in (Geometry and Location Control, 2005). The container/contained, bearer/burden relationships underlie the representation of in, on. The meaning of ‘in’ or ‘on’ is related to the physical or functional relationship between the located object and the construction of location control. Location control is a relationship whereby the position of located object over time is determined by location of reference object.

Extrageometric (such as location control) and geometric (enclosure for in and contiguity with a surface for on) factors together determine the meaning of ‘on’ and ‘in’.

Children’s naming of novel objects could be influenced exclusively by functionally relevant properties if they had prior experience of interacting with the test objects. Without this direct experience global appearance is largely used. Young children are aware of and use both geometric and extra-geometric constraints when describing the relative positions of objects to container and supporting surfaces.

2.2.7 Figure, Ground and Prepositions that Fit Their Relations

In a locative sentence the entity which is referred to by positional description is called ‘**Figure**’. **Ground** is the entity related to which position of figure is expressed. Figure and ground are also mentioned as “**LO**” (Localizing Object) and “**RO**” (Reference Object) respectively. Some features of the figure and ground are associated with the direction or dimension tagged by the preposition (Meaning of Locative Sentences, 2005). For instance, part of bodies or machines in the experiment were more frequently associated to vertical directions, whereas animate entities were more frequently associated to horizontal terms.

Locative sentence construction with the format,

N1 (figure)-Verb-Locative Preposition-N2(ground)

stimulates a spatial layout. The meaning of locative construction has to fit into a coherent pattern. Thus, the sensory motor features of objects (animacy, solidity, part-whole, size etc) are retrieved from memory, they are placed into the slots for figure and ground, provided by the grammar and then simulation of layout marked by directional preposition is run.

Several semantic features of figure and ground are,

1. Animacy- animate/inanimate,
2. Part-whole: a noun is part of whole
3. Countable (many/few), uncountable (much/little)
4. Solidity
5. Mobility
6. Support relationship
7. Relative size between figure and ground

Sentences with horizontal prepositions (front, back) are more likely to involve figure and grounds that are solid, countable, animate and contain a projective view. Sentences with vertical prepositions (above,below) mostly involve partitive figures which are smaller than their grounds and there is a support relationship between them. Animacy of figure and ground are distinctive features in sentences with “front” “behind”. The pattern “figure smaller than ground” is a distinctive feature of vertical dimension. Figures and ground do not generally differ in size in horizontal direction.

2.2.8 Functional Features of Figure and Ground

Functional feature is a part of an object that is central to its use. Larger and less movable objects can act as more stable landmarks. People seem to prefer reference objects that depict a more functional interaction and that are more functionally related (Defining Functional Features, 2005). There is a preference for positioning objects in a manner that allows or affords their typical interaction while locating correct spatial relationship. In an experiment while subjects were asked to place the located object above or below the reference object, placement was significantly biased toward the functional part of reference object.

Both spatial and functional properties affect comprehension and production of spatial language (Form and Function, 2005). Schema is used to refer to a knowledge structure having certain attributes and values. People may have schemas not only for senses of words like prepositions but also for the things and situations they experience and talk about. Affordances are relations between prepositions and actions. A suspended upside down bowl can no longer function as container, and then it would not afford ‘in’.

Parts in their proper configuration determine the shapes of objects. They are features of function. Parts related high in goodness tend to be functionally significant as well as perceptually salient. Form and structure of many parts seem to suggest functions. Parts afford inferences (what the object can do) from structure to function.

2.2.9 Choice of Reference Object (RO)

According to (Gapp, 1995) size, color, shape, mobility, functional dependencies are some factor to choose reference object. Size should be considered in both vertical and horizontal dimension. Relative size of an object compared to the surrounding objects should be determined by comparing the measured differences in each dimension. It remains to be seen which dimension has the bigger influence. If in a certain context all potential candidates for a reference object have the same features of visual salience, except that one object is considerably smaller, then the smaller object might be more salient than the others. This paper (Gapp, 1995) also states that distance of localizing object (LO or figure) and RO (reference object or ground) is often the only criterion used to choose RO.

Experimental studies in (Mangold, 1986) showed that color dominates size and shape in object identification tasks. Visual salience of an object depends on the interaction of basic features like shape, size and color correlated to the corresponding attributes of the surrounding objects. Objects which are large in size with a salient shape/color are preferred as RO (Treisman, 1988). Shape is highly intrinsic, size is not as intrinsic as it is. Intrinsic means external comparison is not needed (Pattabhiraman, 1992).

More salient environmental features may have precedence over less salient ones (Taylor and Taversky, 1992). Salience of objects in pictures are determined by centrality, size, degree of unexpectedness (Conklin and McDonald, 1982).

Another study (Taylor et al., 2000) shows that relative stability of location may be the primary determinant of which of the pair of objects is chosen as the reference object.

In another study (Stopp et al., 1994) the criteria that guide the selection of the reference objects have been discussed. The criteria are:

1. Distance between LO and REFO, i.e., objects closer to LO are preferred.
2. Salience of the REFO, depends on factors like shape, size, color etc.
3. Linguistic context, i.e., objects which have been previously mentioned and which are in focus, are linguistically more salient.

Entities can be salient by being very vivid, by being pervasive, by being unique, or by being spoken about most recently (Pattabhiraman, 1992). The higher the salience of an entity is the greater is its likelihood of selection during content planning.

Vividness and Imageability: (Haagen, 1949)
Vivid words evoke attitudes and feelings quite like those created by the actual experience. Imageability is the ability to evoke clear internal visual representation. For vividness contrast is necessary.

Uniqueness:
Being exceptional or rare in a group.

Pervasiveness**:** abundant, frequent or probable in a context.

2.3. Using Spatial Information in Human-Robot Interaction

2.3.1 Human and Spatial Relation

Humans do not naturally define the exact metric position of a goal object in terms of its distance or angle with respect to a different entity, such as the robot, as their perceptual abilities do not allow for such precision. However, from results in human-human interaction authors in (Moratz and Tenbrink 2003) have states that it is quite natural for humans to refer to objects by means of qualitative descriptions of their location, such as “left” or “right”. Here they have used such results to develop a computational model of qualitative spatial knowledge in order to enable their robot to resolve this kind of spatial reference. In their approach, qualitative spatial reference serves as a bridge between the metric knowledge required by the robot, and the more vague concepts that build the basis for natural linguistic utterances.

2.3.2 Need for Spatial Relation

To reason about space is a fundamental human ability pervading our everyday life (Claus et al., 1998). In a typical service robot scenario, there is an object and an action that the robot must perform with that object. In natural language interaction between humans, the goal object is usually specified by its class name or by a description of its

features, or both. However, in an open scenario in which the robot has no detailed *a priori* knowledge about all of the relevant objects, the current state of the art does not allow correct or guaranteedly unambiguous object categorization (Moratz and Tenbrink, 2002). Then, the spatial configuration and the position of the object relative to the robot itself can be used for linguistic reference. The main advantage of this solution is that, while objects may share many of their physical characteristics such as size, colour etc., only one object can occupy one position at a time.

2.3.3 Spatial Reasoning in Human Robot Interaction

Spatial reasoning is important not only for solving complex navigation tasks, but also because we as human operators often think in terms of the relative spatial positions of objects, and we use such relational linguistic terminology naturally in communicating with our human colleagues (Sofge et al., 2003). For example, a speaker might say, “Hand me the wrench on the table.” If the assistant cannot find the wrench, the speaker might say, “The wrench is to the left of the toolbox.” The assistant need not be given precise coordinates for the wrench but can look in the area specified using the spatial relational terms.

In a similar manner, this type of spatial language can be helpful for intuitive communication with a robot in many situations. Relative spatial terminology can be used to limit a search space by focusing attention in a specified region, as in “Look to the left of the toolbox and find the wrench.” It can be used to issue robot commands, such as “Pick up the wrench on the table.” A sequential combination of such directives can be used to describe and issue a high level task, such as, “Find the toolbox on the table behind you. The wrench is on the table to the left of the toolbox. Pick it up and bring it back to me.” Finally, spatial language can also be used by the robot to describe its environment, thereby providing a natural linguistic description of the environment, such as, “There is a wrench on the table to the left of the toolbox.”

In all of these cases the spatial language increases the dynamic autonomy of the system by giving the human operator a less restrictive vernacular for communicating with the robot. However, the examples above also assume some level of object recognition by the robot. Although there has been considerable research on the linguistics of spatial language for humans, there has been only limited work done in using spatial language for interacting with robots. Some researchers have proposed a framework for such an interface (Müller et al., 2000). (Moratz et al., 2001) investigated the spatial references used by human users to control a mobile robot. An interesting finding is that the test subjects consistently used the robot’s perspective when issuing directives, in spite of the 180-degree rotation. At first, this may seem inconsistent with human-to-human communication. However, in human-to-human experiments, (Tversky et al., 1999) observed a similar result and found that speakers took the listener’s perspective in tasks where the listener had a significantly higher cognitive load than the speaker.

To address the object recognition problem, we use the spatial relational language to assist in recognizing and labeling objects, through the use of a dialog. Once an object is labeled, the user can then issue additional commands using the spatial terms and referencing the named object. An example is shown below:

Human: “How many objects do you see?”

Robot: “I see 4 objects.”

Human: “Where are they located?”

Robot: “There are two objects in front of me, one object on my right, and one object behind me.”

Human: “The nearest object in front of you is a toolbox. Place the wrench to the left of the toolbox.”

Establishing a common frame is necessary so that it is clear what is meant by spatial references generated both by the human operator as well as by the robot. Thus, if the human commands the robot, “Turn left,” the robot must know whether the operator refers to the robot’s left or the operator’s left. In a human-robot dialog, if the robot places a second object “just to the left of the first object,” is this the robot’s or the human’s left?

Currently, commands using spatial references (e.g., go to the right of the table) assume an extrinsic reference frame of the object (table) and are based on the robot’s viewing perspective to be consistent with Grabowski’s “outside perspective”(Grabowski, 1999). That is, the spatial reference assumes the robot is facing the referent object.

There is some rationale for using the robot’s viewing perspective. In human-robot experiments, (Moratz et al., 2001) found that test subjects consistently used the robot’s perspective when issuing commands. We are currently investigating this through use of human-factors experiments where individuals who do not know the spatial reasoning capabilities and limitations of the robot provide instructions to the robot for performing various tasks where spatial referencing is required. The results of this study will be used to enhance the multimodal interface by establishing a common language for spatial referencing which incorporates those constructs and utterances most frequently used by untrained operators for commanding the robot.

2.3.4 Spatial Strategies in HRI (Considering Robot as a Communication Partner)

In dialogue, speakers react to their interaction partner’s contributions, and they attune their linguistic choices to what they believe to be suitable for their partner in the situation at hand (Tenbrink et al., 2002). Human-robot interaction also provides us with a number of additional data not usually available in human-to human communication. Users often produce self-talk in which they give accounts of their strategies, and in which they reveal their interpretations and explanations about what is going on. The order of instructions employed by the users in (Tenbrink et al., 2002) revealed the following hierarchy of instructional strategies:

<goal description
< direction description
< movement description
< description of actions instrumental to movement

The speakers in experiment (Tenbrink et al., 2002) *consistently* took the robot’s perspective, unless there was (or seemed to be) evidence that this could not be the right strategy. This linguistic behavior may indicate that the speakers regarded the robot as a

communication partner who is not capable of taking the speaker's perspective, i.e., who should receive as simple instructions as possible. These findings suggest that the interaction in experiments was influenced by the speakers' conceptualization of the robot as a communication partner with non-humanlike capabilities.

Throughout the experiments in (Moratz et al., 2003), the participants employed the robot's perspective, i.e., there were virtually no instructions in which the user expected the robot to use a reference system based on the speaker or a further object as origin (except for one case in which after a mistake the user explicitly stated that she assumed the robot to be using her point of view). Furthermore, whenever the users referred to the goal object, they overwhelmingly used basic level object names such as [cube], and there was also a very consistent usage of imperatives rather than other, more polite, verb forms.

However, the participants in the experiment (Moratz et al., 2003) nevertheless showed considerable variation with regard to the instructional strategies employed. Half of the participants started by referring directly to the goal object, using instructions such as [drive up to the right cube]. When instructions of this kind were not successful—because of orthographic, lexical, or syntactic problems—the participants turned to directional instructions; if successful, they re-used this goal-naming strategy in later instructions.

The other half of the participants started by describing the direction the robot had to take, for instance, [drive 1 meter straight ahead]. If they were unsuccessful with this type of instruction, some users turned to decomposing the action into even more detailed levels of granularity, using instructions such as [turn your right wheel].

2.3.5 Need for Reporting the Ability of Robot to Human

The paper (Thora, 2003) outlines some basic aspects of what characterizes human-robot interaction in contrast to other kinds of interaction, such as communication with children or foreigners. Here, the robot's looks – humanoid or not – does not play a major role. The question at hand is rather whether or to what degree humans expect a robot to behave linguistically like a human being. The key points author describe in this work are mentioned below:

1) In human-robot interaction inadequate conceptualizations of the robot's functionality and lack of knowledge concerning its linguistic and technical features may influence the success of the dialogue to a high degree. Thus, if there is an unconscious underlying assumption that robots need to understand as much as they can do, more than the usual discourse adaptation processes might be needed to rule out such an assumption.

2) In order to communicate, human users need information concerning some general features of robots, and, more specifically, the linguistic and functional abilities of the robot they are dealing with. If the ongoing discourse does not provide any specific and well-tuned clues concerning these facts, the human users might have to try out many different kinds of variation regarding differing levels of linguistic interaction before they find out how to communicate with their artificial interaction partner.

3) In a different scenario, however, the robot, instead of waiting passively for the user to give instructions, might initially ask a question such as: "Which of the three boxes shall I go to?". By way of this one short question, users can extract information about several useful issues at once: First, they can conclude that the robot already perceives a group of objects such that they may instruct it as to which one of them it should approach. Second, they do not have to worry about goal instructions being too complex for the robot to fulfill, because the robot already asked about the goal itself. Third, the kind of language to be used does not need to be wondered about, as the question is stated in English, and the object is identified by an intuitively suitable label, namely, "box".

4) However, not all communication problems are solved this easily. In more complex robot instruction scenarios, language understanding systems are needed which react to the users' linguistic input so that the users are, where necessary, informed about the specific features of the system they are dealing with. This enables them to address the robot in an adequate manner. Thus, a robot that is not equipped to understand comparisons might answer to an instruction like "Go to the larger box" by: "Sorry, I didn't understand. Shall I go to the leftmost box from my point of view?", thereby telling the user unobtrusively that it can understand qualitative directions but not comparisons. Regarding the ambiguity of spatial reference systems, a robot needs to be able to determine which point of view human users are likely to use, even if they do not explicitly state it. Thus, "go to the left" is ambiguous in itself; however, if human users are predisposed to use the robot's point of view, the ambiguity can be accounted for by using the robot's perspective per default.

2.3.6 Group Based Reference

In cases with a group of similar objects, the centroid of the group serves as virtual relatum (Moratz and Tenbrink 2003). Here the reference direction is given by the directed straight line from the robot center to the group centroid (Fig. 2.1). The object closest to the group centroid can be referred to as the “middle object”.

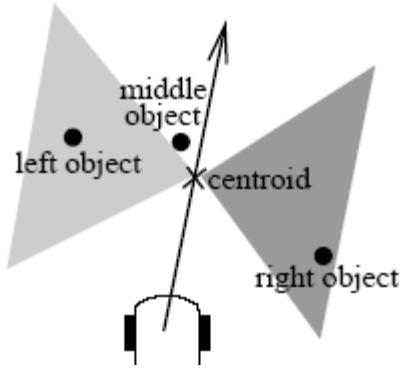


Figure 2.1: Group Based References

In (Thora and Moratz, 2003) to model reference systems that take the robot’s point of view as origin, all objects are represented in an arrangement resembling a plane view (a scene from above). The reference axis is a directed line through the center of the object used as relatum (see figure 2.2), which may be the robot itself, the group of objects, or other salient objects.

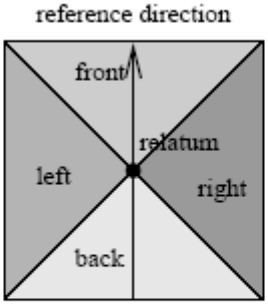


Figure 2.2: Relatum and Reference Direction

After analyzing the errors of users in goal description, authors have modified the region for left, right, front, back regions (Fig. 2.3).

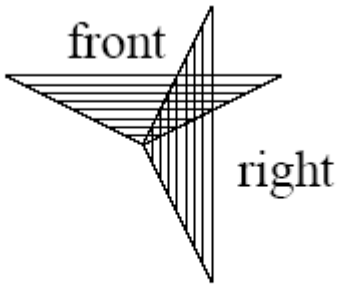


Figure 2.3: Enlarged Acceptance Areas (from 90^0 to 120^0)

2.3.7 Reference Frame and Direction in HRI

In experiment scenario (Moratz and Tenbrink, 2002) involving a group of objects and (in some cases) a further, salient object, three different kinds of linguistic spatial reference offered themselves for communication, and were used in their study when users referred directly to the goal object. First, speakers could employ an intrinsic reference system using the robot's position as both relatum and origin. In this case, they specified the object's position relative to the robot's front. Second, they could refer to a salient object, if available, as relatum in a relative reference system. Then, they specified the object's position relative to the salient object from the robot's point of view. Finally, they could refer to the group as relatum in a relative reference system. In this case, they specified the object's position relative to the rest of the group from the robot's point of view (Fig. 2.4).

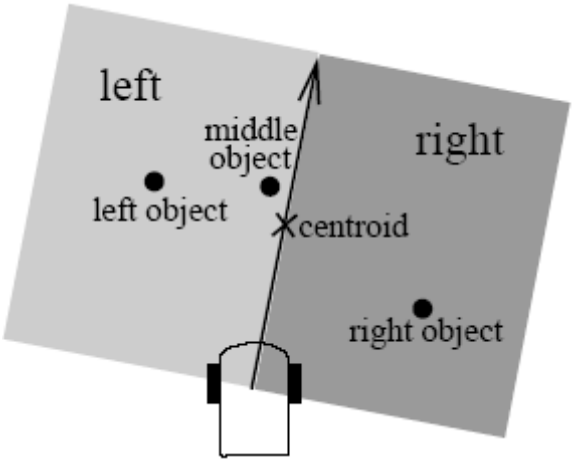


Figure 2.4: Left, Right in Group Based Reference

To define the partitions formally for all three options they refer to the angle Ø between the reference direction and the directed straight line from the relatum to the referent (see Fig. 2.5).

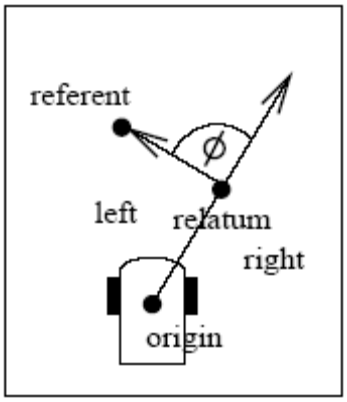


Figure 2.5: Relative Reference Model

Relation between spatial prepositions and Ø can be defined as:

$$
\begin{aligned}
\textit{referent} \text{ front } \textit{relatum} &:= -\pi/2 < \phi < \pi/2 \\
\textit{referent} \text{ left } \textit{relatum} &:= 0 < \phi < \pi \\
\textit{referent} \text{ back } \textit{relatum} &:= \pi/2 < \phi < 3/2\pi \\
\textit{referent} \text{ right } \textit{relatum} &:= -\pi < \phi < 0 \\
\textit{referent} \text{ left front } \textit{relatum} &:= 0 < \phi < \pi/2 \\
\textit{referent} \text{ left back } \textit{relatum} &:= \pi/2 < \phi < \pi \\
\textit{referent} \text{ right front } \textit{relatum} &:= -\pi/2 < \phi < 0 \\
\textit{referent} \text{ right back } \textit{relatum} &:= -\pi < \phi < -\pi/2
\end{aligned}
$$

2.4 Manipulation of Spatial Terms

2.4.1 Toward a Quantitative Measure of Prepositions

To transform the meaning of qualitative spatial words like left, right, front, near etc. into quantitative measurements, a number of researchers have investigated. Authors in (Abella and Kender, 1994) have addressed the issue how computer vision and natural language processing can be used to address the problem of object localization in a 2D image. The ultimate goal is a system capable of generating descriptions that relate the spatial arrangement of the objects through the use of spatial prepositions. This paper presents a semantic representation of spatial prepositions based on an image's visual properties for the purpose of describing the spatial relationship of objects in an image.

In the paper (Skubic et al., 2002), the work on robot spatial relationships is combined with a multimodal robot interface developed at the Naval Research Lab. Authors show how linguistic spatial descriptions and other spatial information can be extracted from an evidence grid map and how this information can be used in a natural, human-robot dialog. The robot spatial reasoning and the NRL Natural Language Processing system are combined to provide the capability of natural human-robot dialogs using spatial language. For example, a user may ask the robot, "How many objects do you see?" The robot responds, "I am sensing 5 objects." The user continues, "What objects do you see?" The robot responds, "There are objects behind me and on my left." and the dialog continues.

Authors in (Roy, 2002) has been developed a spoken language generation system that learns to describe objects in computer-generated visual scenes. The system is trained by a 'show-and-tell' procedure in which visual scenes are paired with natural language descriptions. Learning algorithms acquire probabilistic structures which encode the visual semantics of phrase structure, word classes, and individual words. Using these structures, a planning algorithm integrates syntactic, semantic, and contextual constraints to generate natural and unambiguous descriptions of objects in novel scenes.

Describing the semantics of natural language spatial expressions such as (1) 'moving forward slowly' and (2) 'you're in front of the desk' is not a straightforward task. In the paper (Dobnik and Pulman, 2005) authors describe a setting with a mobile robot where the meanings of such expressions are learned from a set of robot data and natural language descriptions made by a human commentator. They start with simple robot-centered spatial expressions like (1). These do not make reference to the environment external to the robot. Authors then extend the learning to prepositional expressions like (2) which denote relations between the objects in the environment.

In (Keller and Wang, 1995) authors examine three methods for defining spatial relations of image subsets on a group of standard synthetic images. In particular, they consider the centroid method, the compatibility method and the angle aggregation method in order to gain insight into the mechanisms to accurately determine spatial relationships of objects for computer vision applications.
A subset of spatial prepositions is chosen and an appropriate quatification is applied to each of them that capture their inherent qualitative properties in (Abella and Kender, 1993). The quantification use such object attributes as area, center of mass and

elongation properties. A technique for fuzzifying the definition of the spatial preposition is also explained.

2.4.2 Finding Region Around a Reference

To support robot commands such as "Go to the right of the object", we must first compute target destination points in unoccupied space, which are referenced by environment objects. In (Skubic et al., 2004) these spatial reference points are computed for the four primary directions, left, right, front, and rear of an object, from the robot's view. In this paper, authors consider a method of finding these destination points called the *Intersecting Ray Method*. This method uses the main direction from the constant forces histogram to calculate reasonable target points. The main direction is also used for the viewing perspective of the robot. That is, the spatial reference points are computed, as if the robot is facing the target object along the main direction.

Fig. 2.6 shows a diagram for the *Intersecting Ray Method*. As shown in the figure, a bounding box is constructed by considering the range of (x, y) coordinates that comprise the object contour. The bounding box is used as a convenient starting point for a search of key points along the object contour.

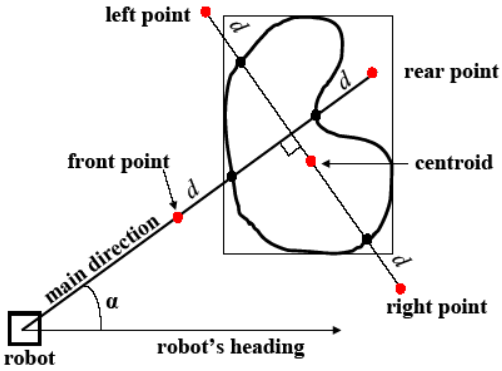


Figure 2.6: The Intersecting Ray Method

The front and rear points are computed to lie on the main direction vector, at a specified distance, *d*, from the object boundary. Consider first the front point. Coordinates are calculated along the main direction vector using the following equations:

$$x = r\cos(\alpha)$$
$$y = r\sin(\alpha)$$

where α is the main direction, (x,y) is a point along the main direction, and r is the distance of the vector from the robot to the (x,y) point. Coordinate points are computed incrementally, starting from the robot and checked for intersection with the object contour until the intersection point is identified.

Front: When the intersection point is found, the front point is computed by subtracting the distance, *d*, from v_F, the vector length of the front intersection point, and computing a new coordinate.

Rear: In computing the rear point, this method again searches for the intersection point of the contour along the main direction vector, this time starting from behind the object. The bounding box of the object is used to compute a starting point for the search. The algorithm first determines the longest possible line through the object by computing *l*, the diagonal of the bounding box. The starting vector length used in the search is then $v_F + l$. Once the rear contour intersection point is found, the rear point is computed by adding *d* to the vector length of the rear intersection point and computing a new coordinate.

Left and Right: The left and right points are computed to lie on a vector that is perpendicular to the main direction and intersects the centroid (*xC,yC)* of the object. Again, a search is made to identify the contour point that intersects this perpendicular vector. The starting point for the search of the right intersection point is shown below:

$$x = x_C + l\cos(\alpha - \frac{\pi}{2})$$
$$y = y_C + l\sin(\alpha - \frac{\pi}{2})$$

Once the intersection point is found, a new vector length is computed by adding the distance, *d*, and computing the new coordinate.

2.4.3 Meaning of Prepositions

The paper (Abella and Kender, 1999) proposes a computational understanding of spatial prepositions that integrates visual and linguistic ideas to generate natural language descriptions of images. In this section we briefly discuss the key points of this work.

2.4.3.1 Notations and Definitions

1) The set of spatial relations chosen, usually covered in language by prepositions (P) is: near, far (from), above, below, aligned (with), next (to), inside, left (of), right (of), between.

2) Also defined is a computational model for intrinsic object properties such as shape or color, usually covered in language by adjectives (A): small, medium and big.

3) Additionally included are superlatives (S) both relational and intrinsic: nearest, farthest, leftmost, rightmost, topmost, bottommost, biggest, smallest.

2.4.3.2 Bounding Box

The parameterization of an object as a blob leads to the concept of the blob's bounding box. The dimensions of the box, w and h, are defined in terms of the maximal and minimal moments of inertia. The maximal moment of inertia is given by:

$$I_{\max} = \int\int_A u^2 dudv,$$

where v and u are axes of maximal and minimal moments of inertia, 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. The half-width of the bounding box, w, is set to be $w = k\bar{u}$, where k is such that the bounding box of an object that is a rectangle should be the rectangle itself. Since the maximal moment of inertia for a rectangle oriented in the x and y direction with width a and height b is $I_{\max} = 1/12a^2A$, it is easy to show that k = $\sqrt{3}$. In the more general case, an object blob's w and h are then projected onto the x and y axis of the image in the usual way.

The choice of a bounding box that is aligned with the x and y coordinates of the image frame is based on human preference for alignment, and is confirmed in the two task domains. In the radiography domain, images are aligned with the spine, which is oriented vertically and serves as an axis of symmetry. In the landmark domain, maps are drawn with a “virtual north" preferred.

2.4.3.3 The Semantic Representation

The semantic representation of prepositions is based here on object area, centers of mass, and elongation properties calculated from their moments. Each preposition is then defined through a set of inequalities, resulting in sets $U_{\bar{p}}$ having nonzero measure in R^{12}.

2.4.3.4 Object Properties

Object properties such as **size** are useful for locating objects. The qualitative properties small, medium, and big are encoded in the following way, based on a sorting of observed object blob areas. An object a is small if its area A_a is not bigger than $s_{small}A_{min}$, where $s_{small} > 1$ and A_{min} is the area of the smallest object in the image.

Similarly, an object b is big if its area A_b is not smaller than $(1/s_{small})\ A_{max}$, where $s_{big} > 1$ and A_{max} is the area of the largest object in the image. Parameters s_{small} and s_{big} are calibrated beforehand for the particular domain and/or image by user studies. An object is medium if it is not small and not big. Since it appears approximately true from user studies, $s_{small} = s_{big} = s$. To preserve ordering from small to big, it must be that $sA_{min} <= (1/s)\ A_{max}$ and therefore $s^2 <= .\ A_{max}/\ A_{min.}$

2.4.3.5 Binary Prepositions

Those prepositions from P that involve a reference object r and a figure object f are defined as follows.

Near is defined when the object bounding boxes, suitably enlarged, have a non-empty intersection. Each bounding box's width is increased by a fraction of its own height, and vice versa. The value of this fraction, ρ, is a statistical parameter determined from human psychology studies, of approximately 0:6. The enlarged bounding boxes appear necessary to accommodate the observed human descriptions, particularly in the case for long narrow parallel objects.

Far is not the complement of near; an object pair may be neither. It is defined when the distance between the two enlarged bounding boxes in either the x extent or the y extent is larger than the maximum dimension of the two objects in that same x or y extent. Inside requires that the bounding box of one object be completely embedded within the bounding box of another.

Above and below, respectively, requires that the projection on the y axis of the bounding box of the figure object f be above or below, respectively, the projection of the bounding box of the reference object r. As with near and far, above and below are mutually exclusive prepositions, and an object pair may be in neither relation.

2.4.3.6 Ternary Prepositions

Between is a ternary relationship. Ideally, the center of the figure object is colinear with the centers of the two reference objects flanking it, and at the midpoint of their common line. In practice, the centers form a triangle and between is defined when the distance of the figure object center to the line formed by the other two object centers is small. “Small" is defined to mean that the projections t_x and t_y (refer to Figure 2.7) 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.

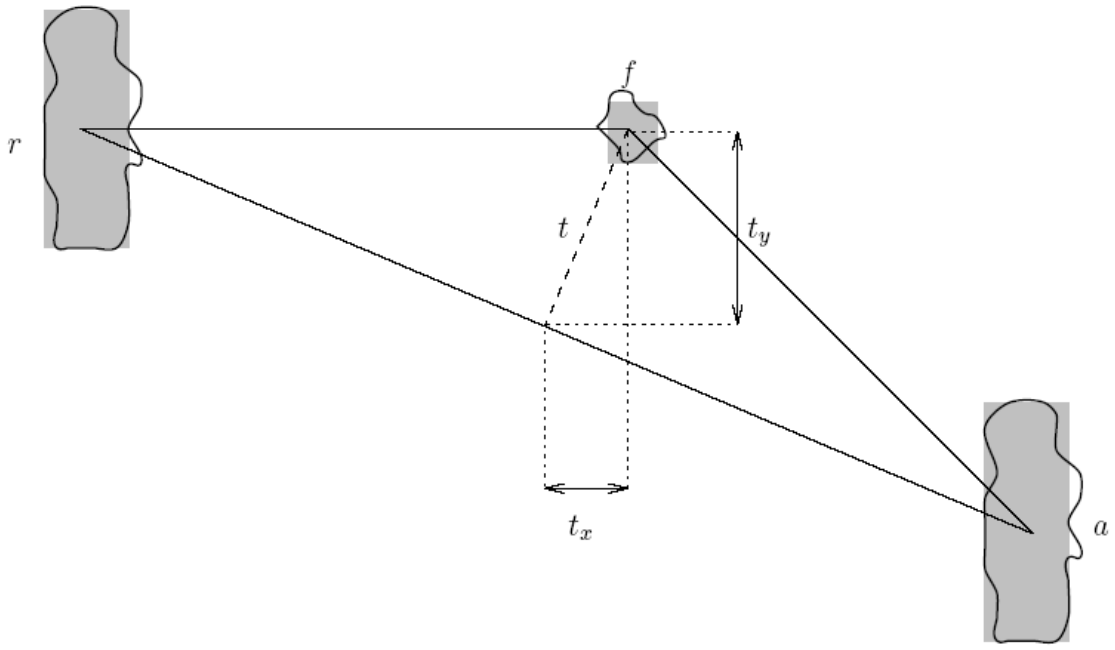


Figure 2.7: Definitions of *tx* and *ty*

2.4.3.7 Superlatives

Smallest and biggest, which are unary superlatives of object description, are defined in the natural way based on the areas of objects. Nearest, Farthest, Leftmost, Rightmost, Topmost, and Bottommost, which are unary superlatives of relationship description, are de-fined likewise in the natural way based on the **Manhattan distance between object bounding boxes.**

CHAPTER 3
SURVEY WITH HUMAN PARTICIPANTS

3.1 Background

Object searching has been addressed in several researches. (Torralba et al., 2003) presents a global image representation that provides relevant information for place recognition and categorization. Such contextual information has been used to simplify object recognition. Global image features have been shown to benefit object search mechanisms while providing an efficient shortcut for object detection in natural scenes (Torralba et al., 2007; Torralba et al., 2006). In (Torralba et al., 2007), authors have built up multiclass and multiview object detection mechanism by sharing the common features across classes. This mechanism has reduced the computational complexity in a large manner comparing to related works.

While the aforementioned researches use statistical approaches to detecting objects, we are interested in discourse-based object detection. Verbal input and natural language understanding are described as indispensable part of human-robot interface in (Fong et al., 2007). Communication of human and robot through dialogue is also used in (Fransen et al., 2006). The motivation behind our research is to generate an interactive object detection model in presence of some known (can be detected by robot) objects. Our primary interest is to observe how humans describe a target object to a robot which can recognize very few objects in the scene. Researches have been found which address various issues of referring behavior of humans. Mutual responsibility of participants in the making of a “definite reference” is addressed in (Clark and Wilkes-Gibbs, 1986), whereas, (Clark and Brennan, 1991) describes how “Grounding”, which is very basic to communication, gets shaped. But these works refer to general referring behavior in conversation. We could not find any previous work specifically in our interest area, i.e. object description using attributes.

3.2 Objective of the Survey

When the robot system suffers from a limited object detection capability, effective communication between the user and the robot facilitates the reference resolution. How human users, being aware of the poor object detection capability of their robot partner, describe objects in 2D images is of our primary interest. In need of observing linguistic preference of humans, this survey has been designed and carried.

3.3 Survey Basics

The participants are non-native speakers of English and graduate students of Saitama University. To receive user input, a Visual Basic program is developed. Screenshot of the program is shown in Fig. 3.1.

3.4 Role of the Participants

A pair of participants sits in front of the screen. One plays the role of a robot (we refer to him as “Robot” from now on in this paper) and his partner acts like a human (referred to as “Human”). With a shared view of an image where there are several objects, the challenge of the Human is to describe a target object to his Robot partner providing efficient and effective hints. Through a conversation, Robot will endeavor to detect the target object as soon as possible, using those hints. Both users only input text in English. No oral input is allowed. All input was saved in a text file. The only gesture input permitted for the robot user is to point to an object on the screen, which he thinks the target. Since the experiment was not videotaped, the pointing gesture of robot was recorded by taking note of the name of object pointed.

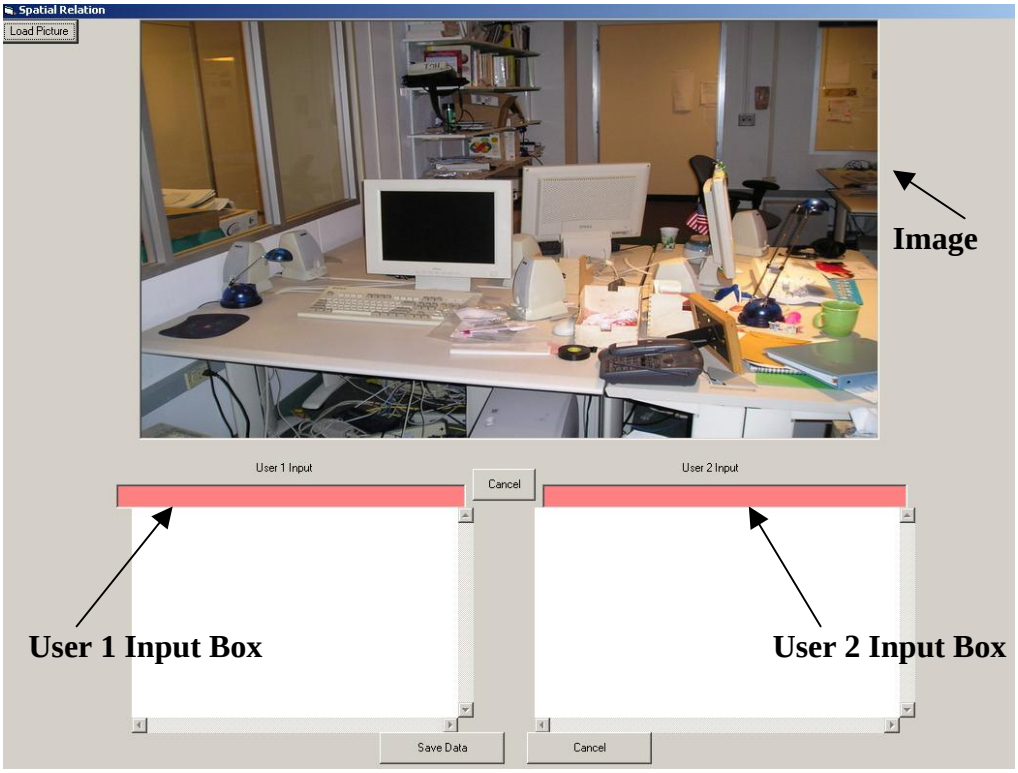


Figure 3.1: Screenshot of the Survey Software

3.5 Need for Reporting Known Objects to ‘Human’

Now, there arises an issue of human understanding of the knowledge level of his partner. As mentioned earlier, in our research a Robot system is assumed to have limited object recognition capability. Our intention is to grasp human preference of description when their partner has such kind of limitation. To help both users restrict their choice of words during a conversation, they are informed beforehand that “Robot” will pretend to have a very limited knowledge about the objects in the environment. So, in a 2D image he is supposed to be able to recognize only two types of objects. The cognitive and linguistic database of Robot is limited to name of these known objects and basic properties of objects i.e. color, shape, size and positional relationship among objects in an image.

3.6 Known and Target Objects for Images

For a total of 15 images, 45 pairs of participants engaged in conversation i.e. 3 pairs for each image. To obtain description of objects in different situations and different orientation of objects, 15 images were chosen. Number of participants was 20 and some of them were involved in trial more than one time with the same or different partner. All the images are shown in Fig 3.2. For each image we decided the known and target objects. Known objects were reported to both users so that only these objects are used for reference during conversation. Identity of target object was shown only to Human in written form. Table 3.1 lists known and target objects defined by us for all images. Images in Fig. 3.2 are numbered from 1 to 15 (left to right, increasing with successive rows). Unknown objects are marked by 'X'.

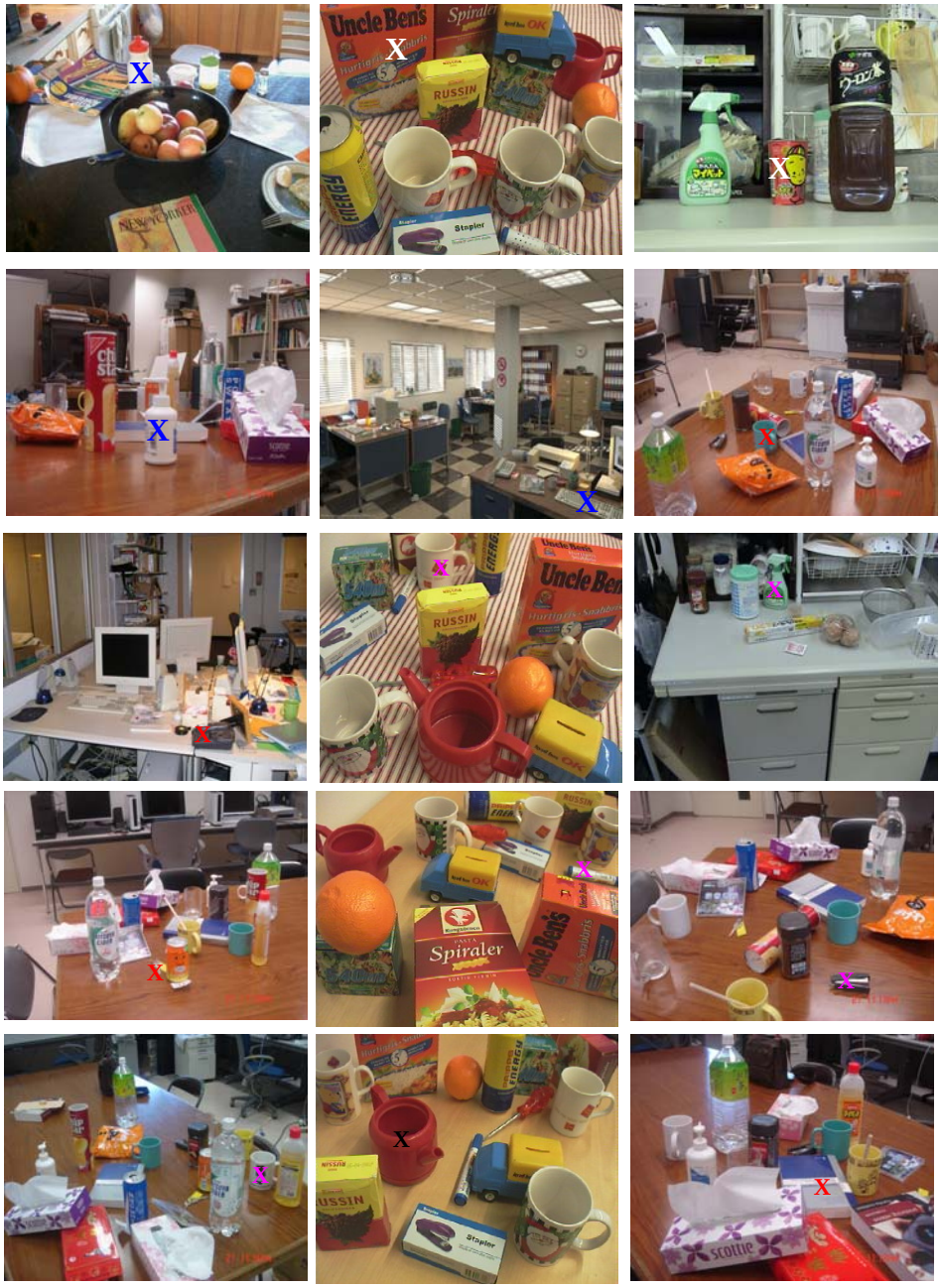


Figure 3.2 Images Used in the Survey

Table 3.1: Known and Target Objects in Images of Survey

Image	Known Objects	Target object
1	Magazine, orange	White jar with red top
2	Orange, can	Uncle Ben's box
3	Urong tea	Ramen
4	Red Chips, Blue can	White Hand Soap
5	Telephone, table lamp	Keyboard
6	Ocha bottle, Cider bottle	Light Blue cup
7	Table Lamps, Monitors	Telephone set
8	Uncle Benz, Teapot	Red and white cup
9	Potato, Umbrella	Light Green Dish cleaner
10	Coffee jar, Two bottles	Key
11	Teapot, cups	Marker pen
12	Orange chips, Blue can	Stapler
13	Books, cans	White cup with a cartoon
14	Three cups	Teapot
15	CD, Bottles	Blue book

3.7 Exemplary Descriptors

We made a list (Table 3.2) of exemplary descriptive words for object properties and positional relationship. Both participants were shown this list before starting the conversation. This was done not to bias their word choice, but to help them have a clearer view of range of possible descriptions.

Table 3.2: Example of Candidate Words for Description

Category	Examples
Color	Red, Yellow and Green, Light Blue etc.
Shape	Rectangular, Round, Cylindrical, Square etc.
Positional relation	Left, Front, Near, Middle, Front-right etc.
Size	Big, Small etc.
Superlatives, Comparatives	Rightmost, Nearer, Smallest etc.

An excerpt of conversation among users for image 4 (see Fig. 3.2) is shown below.

(Target: White hand soap; Known: Red chips, Blue can)

Robot: I can see red chips and blue can.

Human: I want to have a round object.

Robot: I can see many round objects, which one do you want?

Human: I want to have the white one, which is in front of red chips.

Robot: Ok I found it. Do you want this? (pointed to the white hand soap).

Human: Yes.

3.8 Rules for Encoding Dialogs

To analyze text input of users, we did not use any state-of-the-art language processing tool. Instead, our own simple strategy was followed. At first we encoded the dialogs with symbols used for four primary categories of object attributes; color (C), shape (Sh), posiotnal relation (P) and size (Si). We divided the positional relation terms into some sub-categories (Table 3.3).

Table 3.3 Different Types of Positional Relations

Type of relation	Example
Group based, Pg	(Leftmost cup)
Relative, Pr	(In front of the orange)
Environmental direction, Pd	(North, south etc.)
Image plane as reference, Ps	(Upper portion, in the middle etc.)

Initial reporting of known objects by the Robot is excluded from encoded data. No other word except those already mentioned, was included in encoding. To count the frequency of mentioning any specific category (Color, Shape, Size, Positional relation) in a conversation, we just calculated the frequency of symbol used for that category.

3.9 Results

The conversations between 45 pair of robot and human users can be discussed from various viewpoints.

3.9.1 Proportions and Vocabulary

A total of 198 words were mentioned by the participants; 137 by Human and 61 by Robot. This was obvious because Human contributed to the conversation describing the target object, notably more than the Robot.

Among all words, color was the mostly mentioned one. Positional relation is the follower (Table 3.4). Percentage usage of attributes are, color 38%, shape 20%, position 35% and size 7%. It can be summarized from data in Table 3.4 that, both Human and Robot, across all trials, were consistent in using color and positional relations almost equally.

Table 3.4: Individual Preference for Object Attributes Used (in percentage)

	Color	Shape	Position	Size
Human	39	19	34	8
Robot	37	22	37	4

Now we analyze the vocabulary for object properties used by participants. With no doubt, the list can not depict general linguistic preference of humans because all conversations are solely biased by the features and orientation of target, known objects and all other objects in the image. Experiment with another set of images with different target and known objects might change the list obtained from our experiment.
Along with standard colors such as red, green, yellow etc., some modifiers like light, deep, almost, -ish etc. were used. For multicolor objects users mentioned all prominent colors when needed. Words used for size and shape across all trials are shown below:

Property	Words used
Size	Big, small, medium, short, long, half, large, thin
Shape	Rectangular, round/circular, flat, square, cylindrical, triangular

Superlatives and comparative forms of size were also used. In response to a question of the robot, “Is it the highest one?”, the human user answered, “Half of the highest”. The word “half” was used only in this case among 45 trials. Users found it easier to use rectangular and round than other shapes. For an object having two parts of two different shapes and colors, one user mentioned the shapes separately. This also was the only instance of using the word ”triangular”.

For positional relations, there were a wide variety of words used (Table 3.5).

Table 3.5 Vocabulary for different Positional Relations

Type of reference	Vocabulary
Relative, Pr	Left, right, in front of, behind, near/close/nearby, far, middle/ between,
Group based, Pg	Middle/center, leftmost, rightmost
Directional, Pd	South-west
Image plane as a reference, Ps	Left upper corner, leftmost, center/middle

We found one instance of using the modifier “little” with “far”. If we compare the use of relative terms of three groups, left/right (29), near/far (19) and front/behind (11), the first group was the mostly used one. The proportions of using different types of reference are Pr 80%, Pg 12%, Pd 2% and Ps 6%.

3.9.2 Necessary Information Provided by Human at the Beginning

In (Hossain et al., 2006; Kurnia et al., 2006; Kurnia et al., 2006; Mansur and Kuno, 2007), robot leaded the conversation with user, generating efficient queries. In order to do this, huge primary manipulation was required for scene understanding. In our experiment, however, we have instructed the Human to provide as many information about the target object as they think useful at the start of the conversation. We have seen throughout the experiment that, Human users could describe attributes of target object avoiding unnecessary details. Using the information obtained from Human at the very beginning, Robot has been able to squeeze the horizon of candidate objects in an efficient manner. Thus, not only he was relieved from the primary manipulation, but also generation of subsequent queries was easier for him. Here we give an example.
In image 11 (Fig. 3.3), Robot is supposed to be able to recognize “teapot” and “cup” (denoted by O). The target is “marker pen” (denoted by X).

Figure 3.3: Known and Target Object in Image 11

The dialogs to detect the target in this image are depicted below.

Robot: I can see teapot and cups
Human: The object is in front of cup.
Robot: Which cup?
Human: Rightmost cup.
Robot: Nearer to that cup?
Human: Yes

The Robot then pointed to the intended target object. Here we see that Robot did not have to decide which way he should ask for information about the target. Rather, Human provided positional information, which he regarded as the best strategy to describe the target.

3.9.3 Nature of Dialogs

The participants in almost all the trials showed a tendency to use short, communicative form of English sentences, although not instructed to do so. The expressions are, most of the time, grammatically incorrect but effective to solve the purpose.

3.9.4 Descriptors Not in Knowledge Domain

Although the participants were given a brief outline of the knowledge domain of the robot, i.e. candidate properties and word families for describing objects, we found that in some trials they introduced descriptors from outside the domain. There are, however, only four instances of this throughout all of the 45 trials.

3.9.5 Use of Deictic Words

We were interested in use of deictic words such as this/that, but no noteworthy use was observed. Among 198 words elicited from all the trials, demonstrative pronoun "that" was used only once by the human user and once by the robot (Table 3.6).

Table 3.6 Instance of Using "that" in Survey

Example	Used by	"that" refers to
Left of **that** one.	Human	Object pointed to by Robot
Nearer to **that** cup?	Robot	"Rightmost cup", mentioned by Human in the previous dialog.

3.9.6 Error Correcting Strategy

It reveals from this experiment that, in a conversation mutual agreement between partners about any physical property of objects, plays an important role in identifying the target. Some examples can be given in this regard. In one trial, the Human described a target object as "purple". But, the Robot did not find any purple object in the scene and he reported it to the Human. Consequently, the Human inferred that what he thinks "purple" is not actually the same for his partner. He, then, relied on another property to describe the object. In another trial, the Robot agreed to all other properties except color for the target object. He considered it to be black, which was described as gray by his partner. To remind the partner of the possible mistake, the Robot made a query.

Robot: Is it gray or black?
Human: Deep gray, not black.

Although the Human still sticks to "gray" rejecting the possibility of "black", the word "deep" helps the Robot infer that the black object in his mind is being referred to as "deep gray".

In one trial an object was described as at "southwest" position. From other indicators it was clear to the Robot that "south-east" should be the description. But instead of making his partner aware of the mistake, he continued to look for the objects at southwest position and made subsequent queries. The trial was not, as can be easily understood, an efficient one in terms of length of conversation. After the trial the participant who played the role of Robot told that, he could detect the target object at the very first moment, but he was not sure if he was allowed to correct the mistake of

his partner. We see here that, correcting an error or at least asking the partner about alternative choices, as soon as the error is revealed, makes the reference resolution efficient and reduces further complexity.

3.9.7 Feedback from Partner

Throughout the 45 trials, there are various occurrences of feedback from both users. We summarize below the strategies followed to provide feedback.

Strategy 1: The robot finds several candidate objects based on the description and generates group-based query to elicit the target. For example, “There are two round yellow objects. Which one do you want?”, “Which object is near the target?”

Strategy 2: Robot reports that no object of a given description is found. For example, “No red object”.

Strategy 3: Expressing inability to describe a specific property of an object (example 1) or inability to resolve a given property (example 2).

> Example 1
> Robot: What’s the shape?
> Human: I can’t describe.
>
> Example 2
> Human: Do you find anything round?
> Robot: I don’t know.

Strategy 4: Asking about specific choices for a property in order to reduce the number of candidates. For example, “I found two yellow objects. Is the target square or cylindrical?”

Strategy 5: Asking for more information when the robot finds the given description insufficient. For example, “Tell more”, “Give more information” etc.

CHAPTER 4
PROPOSED METHODOLOGY

4.1 Block Diagram

We propose here a methodology for interactive object detection (Fig. 4.1).

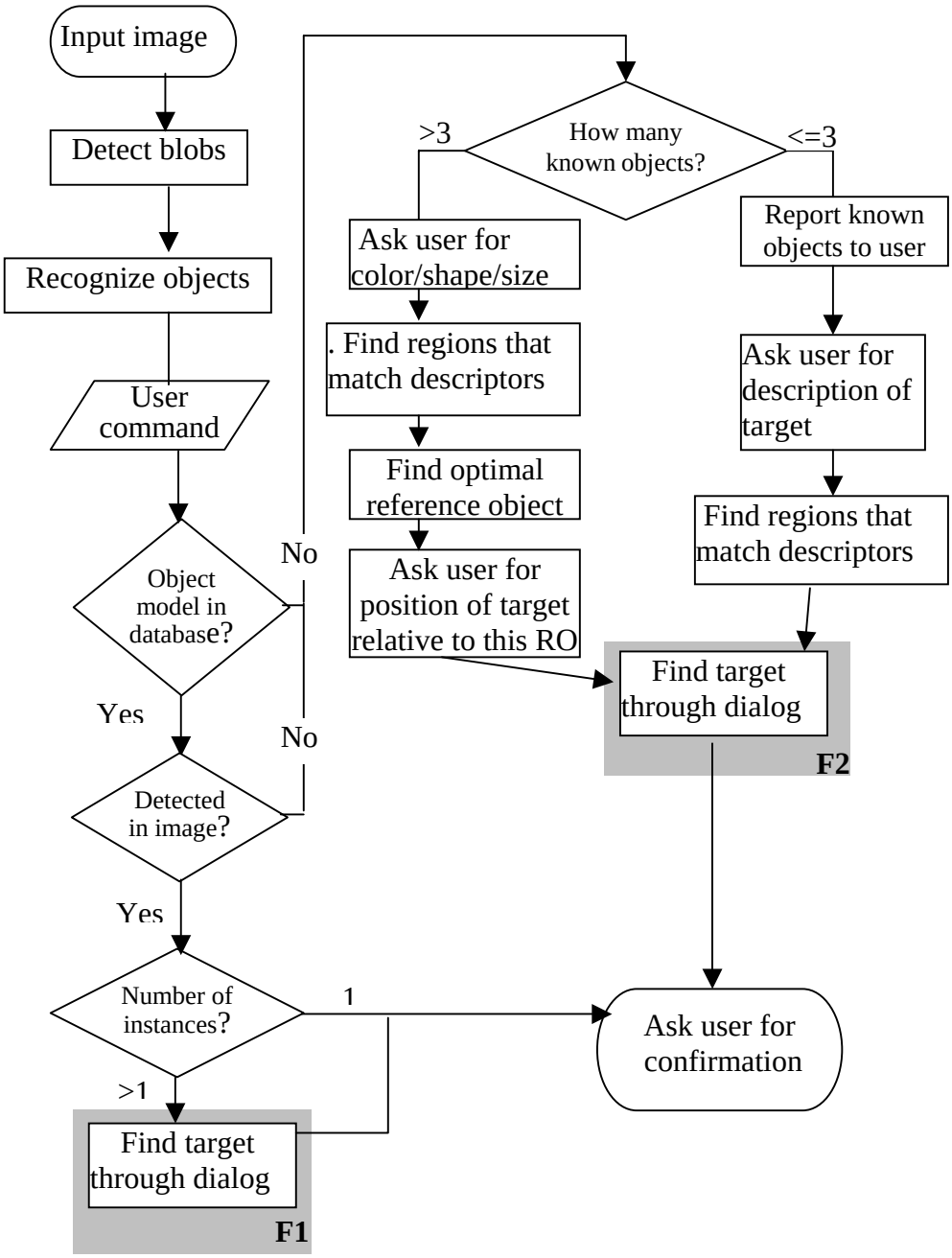


Figure 4.1 Overview of the System

4.2 Description of Various Modules

The input image is first processed to detect blobs (Lindeberg, 1988). Some blobs may be recognized as objects using methods described in (Mansur et al., 2007). After the user's command input, the system looks for the phrase which refers to target object. When the target is one of the objects for which there is an object model in robot's database and the object is correctly recognized in image, robot determines the number of instances found. If there is only one instance found, robot detects it as target and ask user for confirmation. If there is more than one instance of the object, robot tries to confirm which one the target is among the group of objects. This is resolved through asking user a question like, "Which one among the three?". Answer may be the biggest, middle one, the nearest, the leftmost etc. This interaction part is shaded region F1 in Fig. 4.1.

It may happen that the target object's model is in database but no instance of it is recognized in image (i.e. false negative). In another situation, robot may not be trained to detect the intended object of user. For both these cases, robot will proceed to conversation with user.

Throughout the previous chapters it has been mentioned that our proposed robot system is assumed to have limited object detection capability. Hence, user's description of target object is necessary. In the survey carried out for this research, the participants followed the principle that "Robot" can recognize two objects in the image. Prior to any description of target object, "Robot" reported the name of known objects to "Human". The reason behind this is that having known the name of recognized objects, "Human" will not refer to any unknown object in the description. It is very natural that user can describe position of the target in relation to an unknown object and hence robot is unable to locate the reference object. Reporting the known objects before stating any description is thus convenient for both robot and user.

Then, robot asks the user for a description of the target. Based on the description it then removes candidate objects through dialog generation and finally confirms the detected object as target, given user feedback.

While this reporting is natural and easy for few (we consider the number as less than three) known objects (i.e. "I can see a cup, a pen and a book), it does not seem to be efficient when number of known objects is larger than three. User may find it troublesome to remember names of all objects that the robot can recognize. In cases like this, there should be a way out so that user can describe the position of target making reference to a known object. In chapter 7, we have described the method of optimal reference object selection in details.

When robot can recognize more than three objects, it takes the control of the flow in conversation. As user is not informed at this stage which objects are known to robot, he cannot make a description involving spatial position of target. That is why, robot asks for other attributes of the target except spatial information and find regions that match

these attributes. In this way search region gets reduced. Now, robot decides optimal reference and asks for position of target in relation to this reference object. The method of target detection then proceeds through interaction.

4.3 Feedback Generation

The step “Find Target through Dialog”(shaded region in Fig.4.1, marked by F2) reflects our findings about “Feedback generation” from the survey. All consisting steps of it are described below in details:

F.1. Receive user input and decide number of attributes used in it.

F.2. If one attribute is used, find whether it is positional relation or not. Otherwise, go to step F.4.

F.2.1 Positional relation, P

If Pr is used, find reference object. Otherwise, decide the region where objects will be searched later. Then, find objects matched with given positional information.

F.2.2 Attribute other than P

Find objects of the given attribute.

F.3. Decide how many objects found which match the given attribute.

F.3.1 One object found

Show the user and asks for confirmation.

F.3.2 More than one object found

Consider the objects that match given attribute as a group and ask user to tell which one the target is, among these objects. Upon user input, find the target and asks user for confirmation.

F.3.3 No object found

Report to user as “Not found” and request for more information. Repeat from step F.1.

F.4. Decide whether the attributes include positional information or only are the combination of color, shape and size. If they include positional information go to step F.6.

F.5. Decide whether any object satisfying all given attributes is found.

F.5.1 Objects found matching all given attributes

If there is one such object, execute step F.3.1. Otherwise, execute step F.3.2.

F.5.2 No object found that match all attributes

F.5.2.1 Report the attribute, which is not matched, to the user.

F.5.2.2 Keep a record of all objects that satisfy some attributes.

F.5.2.3 Then ask the user to describe position of the target in relation to any of the known objects.

F.5.2.4 Find objects using both F.5.2.2 and F.5.2.3. Execute step F.3.

F.6. Execute step F.2.1. Use attributes other than positional relation to reduce number of candidates from the result of step F.2.1. Then execute step F.3.

CHAPTER 5
OVERALL ARCHITECTURE OF THE UNDERLYING SOFTWARE

5.1 Introduction

We have integrated the entire process of blob detection, object recognition, user input reception, parsing and finding objects based on the given descriptors into a single system. In this chapter overall design of the software system will be described. Functions of various modules of the system will be discussed in brief, being left for further discussion in following chapters.

5.2 Block Diagram

In Fig. 5.1 an overview of the software implementation of our system is shown. 'LO' means 'Localizing Object' i.e. the target and 'RO' means 'Reference Object'. If user provides any information about the position of target in relation to any known object, then RO will be used.

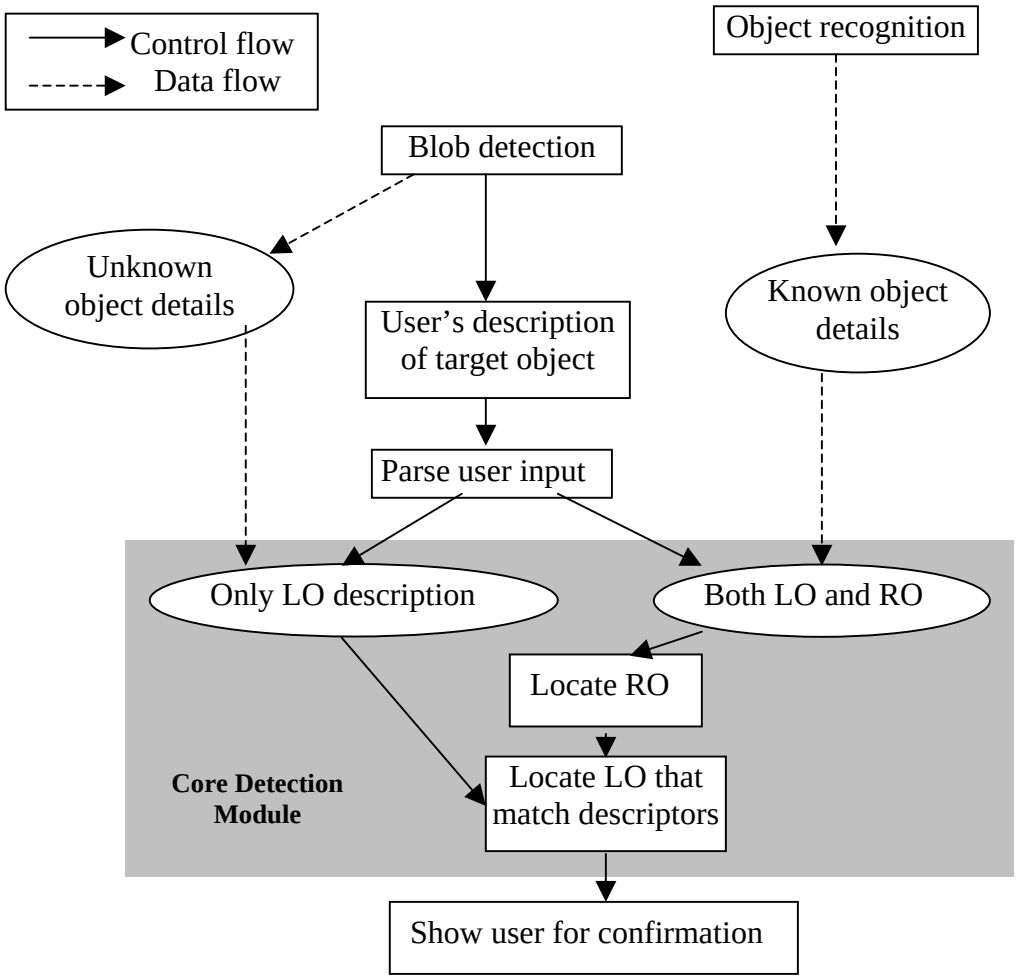


Figure 5.1: Architecture of the Software System

5.3 Object Recognition

Object recognition is an independent module in Fig. 5.1 that runs at background. For object recognition we use SIFT (Lowe, 1999). Our system is trained to recognise some household objects. Recognition of objects in image of Fig. 5.2a is shown in Fig 5.2b.

Figure 5.2: (a) Original Image (b) Objects Recognized

5.4 Blob Detection

By ‘blob’ we refer to the regions in image that are either brighter or darker than the surrounding. Detected blobs are candidate object regions in our case. We assume that there will be some objects in the image that are unrecognized by the robot. To keep track of these regions we need to find blobs. In interaction with human user, properties of these blobs are used. We use the blob detection JAVA library obtained from (“Blob”). This library is developed using method described in (Lindeberg, 1988). In Fig 5.3 result of detection is shown. Regions separated by green edge boundary are regarded as object regions in the image of Fig. 5.2a.

Figure 5.3: Blob Detection Separates Regions of Fig. 5.2a

In Fig. 5.3 probable regions occupied by objects are identified as blobs. There are some limitations of this detection method. Some of these regions are not objects as well as whole of the region occupied by an object is also not inside the boundary. Although blob detection in this research does not produce expected results in some cases, these limitations can be overcome through interaction with user. Image coordinates of both recognized and nonrecognized objets are kept in separate lists.

The Laplacian (LoG)

Blob detector in (Lindeberg, 1998) is based on the Laplacian of the Gaussian. Given an input image f(x,y), this image is convolved by a Gaussian kernel

$$g(x, y, t) = \frac{1}{2\pi t} e^{-(x^2+y^2)/(2t)}$$

at a certain scale t, to give a scale-space representation

$$L(x, y, t) = g(x, y, t) * f(x, y).$$

Then, the Laplacian operator $\nabla^2 L = L_{xx} + L_{yy}$ is computed, which usually results in strong positive responses for dark blobs of extent $\sqrt{t}$ and strong negative responses for bright blobs of similar size. A main problem when applying this operator at a single scale, however, is that the operator response is strongly dependent on the relationship between the size of the blob structures in the image domain and the size of the Gaussian kernel used for pre-smoothing. In order to automatically capture blobs of different (unknown) size in the image domain, a multi-scale approach is therefore necessary.

A straightforward way to obtain a multi-scale blob detector with automatic scale selection is to consider the scale-normalized Laplacian operator

$$\nabla^2_{norm} L(x, y; t) = t(L_{xx} + L_{yy})$$

and to detect scale-space maxima/minima, that are points that are simultaneously local maxima/minima of $\nabla^2_{norm} L$ with respect to both space and scale (Lindeberg, 1998). Thus, given a discrete two-dimensional input image f(x,y) a three-dimensional discrete scale-space volume L(x,y,t) is computed and a point is regarded as a bright (dark) blob if the value at this point is greater (smaller) than the value in all its 26 neighbours. Thus, simultaneous selection of interest points $(\hat{x}, \hat{y})$ and scales $\hat{t}$ is performed according to

$$(\hat{x}, \hat{y}; \hat{t}) = \operatorname{argmaxminlocal}_{(x,y;t)}(\nabla^2_{norm} L(x, y; t))$$

This notion of blob provides a concise and mathematically precise operational definition of the notion of "blob", which directly leads to an efficient and robust algorithm for blob detection. Some basic properties of blobs defined from scale-space maxima of the normalized Laplacian operator are that the responses are covariant with translations, rotations and rescalings in the image domain. Thus, if a scale-space maximum is assumed at a point $(x_0,y_0;t_0)$ then under a rescaling of the image by a scale factor s, there will be a scale-space maximum at $(sx_0,sy_0;s^2t_0)$ in the rescaled image (Lindeberg, 1998).

5.5 User Input

Now this is the time to accept user request. If the user directly requests for any known object, the system can easily identify the target using data of known object list. If the requested object name is unknown to it, user is requested to provide description of the target.

5.5.1 Parsing User Input

To break down the input into meaningful parts for our software system, a parser has been designed. Details of it will be found in chapter 0. Output of the parser is a parsing tree with descriptors of LO and RO are positioned at different nodes. From the outcome of survey carried out in our research, it has been evident that user's description takes one of the two basic forms:

a) Description of target include only color, size, shape and spatial information does not refer to any reference object,

   Example: `Rightmost square thing.`

b) Description of target include spatial information along with others. Here, spatial information makes reference to known objects.

   Example: `Blue round thing, in front of cup.`

In case of (b), reference may be made to one or more known objects. The parser is made taking both cases into account. So we can decide from the parse tree if the user input is of type (a) or (b).

5.6 Core Detection Module

In this section we describe the process of target detection considering the case of less than three known objects (refer to Fig 4.1). The other case is described in chapter 7 in details.

Shaded area in Fig. 5.1 is the core detection module of this system. We have specialized this module since it entails elaborate description. As can be seen in Fig 5.1, this module first checks if the case is (a) or (b) of section 5.5.1. If the case is a, system proceeds by matching descriptors to unknown object regions in the image. If the case is b, it first locates reference object and then matches spatial or other information of the target object. Finally the searching result is shown to user for confirmation.

Before we go further into describing the algorithm underlying core module, we will introduce and explain two necessary data structures for this module. One is READY QUEUE and the other is PROCESSED STACK.

5.6.1 READY QUEUE

Queue is a FIFO (first in first out) data structure. Element which enters into it first (called 'front element') is processed first and then removed. New values are added to the rear of the queue. We need to use a queue in the detection module in order to maintain the list of incoming object name and descriptions as the conversation continues. Node values of parse tree (described in details in chapter 6) are added into READY queue in order of their appearance in discourse.

5.6.1.1 Input Sequence in READY QUEUE

Consider the user input as "Blue, round object in front of red cup". Here both LO and RO description are present (case (b) of section 5.5.1). At first RO name ('cup') will be inserted into the queue. If RO comes with any descriptor (in this exaple, 'red'), it will follow RO name. At this stage, READY queue will look like following (Fig. 5.4):

Figure 5.4: READY Queue after Adding RO Information

Next, LO descriptor is added into READY. If there is positional information (as in example sentence) it will precede other descriptors. The reason behind this is that, spatial preposition will link LO to RO, that is already in the queue. READY now looks like Fig. 5.5.

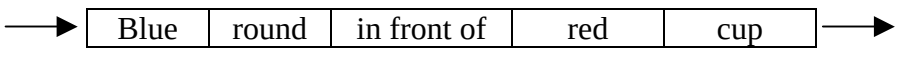


Figure 5.5: READY Queue of Fig. 5.4 after Adding LO Information

If there were no RO in user input only LO descriptors would have been added.

5.6.1.2 Processing Elements of READY

In order to match the descriptors with blobs and known objects in the image, front element of READYwill be being processed sequentially and FRONT (Fig 5.4) will move to next element. Output of this processing will be saved into PROCESSED STACK. So, at this stage we describe another data structure PROCESSED.

5.6.2 PROCESSED STACK

In contrast to a queue, a stack is a LIFO (last in first out) data structure. Element which enters into it at last (called 'top element') is processed first and then removed. New values are added to the top of the stack. Both insertion and deletion are done at the same end of a stack.

At the moment an element of READY is processed, it gives us one or more objects in the image. The more and more description user provides, the more gets the number of candidate objects reduced and finally it leads us to detection of target object. That is why, after matching each description of objects, we need to keep track of the objects resulted from this matching. Moreover, a task of matching may require some subtasks of matching. For example, to interpret the descriptor 'nearmost' system fist need to find objects near a certain RO. Then it search for the nearmost object among ones which are near the RO. Thus, domain of later matching will be output of the previous matching. This is the idea behind introducing a stack PROCESSED in our system.

5.6.2.1 Structure of PROCESSED STACK

PROCESSED is designed in a way that allows to save both description of object that is taken from READY queue and identities of the objects that match the description. Structure of PROCESSED is shown in Fig. 5.6.

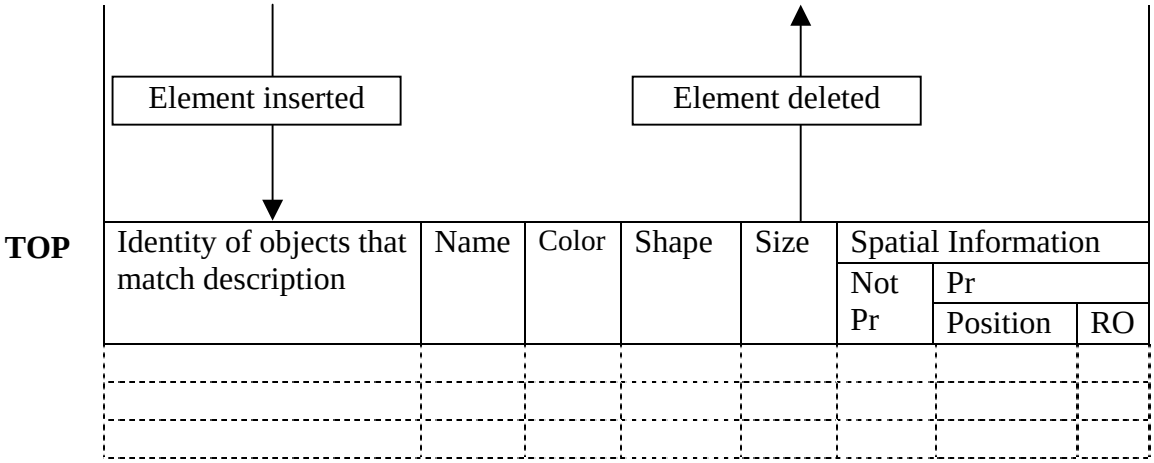


Figure 5.6: Structure of the Stack PROCESSED

Each element of the stack has five main fields. Reason for keeping object identities (first field in a stack element) has been described in section. An object region is confirmed to match a given description if it is the only member in this field. Otherwise, query will be being made by robot to reduce candidate regions to only one.

Other fields except the first one also demand to be included in the stack in order to refrain our system from reinventing the wheel. Let us further explain this. At times, in a discourse there may come a description which also appeared before, then we need not match objects with the given description again. By searching in PROCESSED we can easily detect if the same descriptor for an attribute (color, shape, position etc.) was matched before. Thus keeping record of previously executed descriptions saves time.

It can be seen from Fig. 5.6 that, 'Spatial information' field has two subfield, 'Not Pr' and 'Pr'. Pr means relative reference (please refer back to section 3.8 and 3.9 of chapter 3). If positional information makes reference to other object then we call it Pr (i. e. 'Left to cup'). Examples of non-Pr positional information may be like this, 'Frontmost object', 'Object in the middle' etc. In case of Pr, both positional term and RO will be recorded in PROCESSED.

5.6.3 Feedback Generation

By looking carefully at the corpus of discourse used in the survey, we have identified some ways 'Robot' followed to ask 'Human' further questions (refer to section 4.3). Whenever given description was felt to be inadequate or ambigous, 'Robot' produced feedback by generating a query. Three main types of generated feedback were observed:

i) When given description results to less than three object regions, 'Robot' asks "I found two (or three) such objects. Which one between them?"
ii) When given description results to more than three object regions, 'Robot' asks about an attribute that was not mentioned in previous description. In this way it attempts to reduce candidates for target. One such question is, "What color is the object?"
iii) When given description results to more than three object regions and no attribute was missing in previous description, 'Robot' throws open questions like this, "Tell me more."

The reason we discuss feedback generation here is that, from different fields of PROCESSED stack, the system can find a clue to decide which type of query should be generated. Number of resulted object regions from a description can easily be found when we look at the first field of the stack. Missing attributes in previous description are those for which the corresponding field in PROCESSED contains no value. So query can be generated about a missing attribite. When all attribute fields are filled in PROCESSED, it means nothing was missed previously and the system asks for more information.

5.6.4 How LO and RO Processed Simultaneously in READY and PROCESSED

Depending on the type of user input (case a or b in section 5.5.1) one point should be taken into account. If the type is 'b', i.e. target object description includes RO, then the system must find RO at first. Until it decides RO, no descriptor of LO from READY

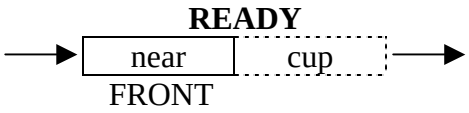
should be picked up and processed. Now comes the question how the system performs this manipulation. There are two things to be done for this:

i) As long as the first field of PROCESSED (Fig. 5.6) does not contain a single object, no more element of READY will be processed.

ii) When given description of RO results in ambiguity, robot asks the user for clarification. So, there comes more description about RO which also should be included in READY. But this will lead to conflict with descriptors in LO. Consider the following excerpt (in italics) from a trial discourse in our survey.

1 User: I want object near cup.

[READY and STACK at this stage are shown below.]

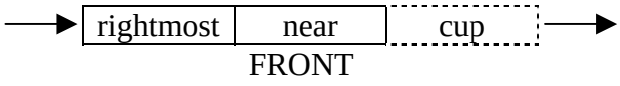


PROCESSED

Information of 3 cups	Name (cup)	Color	Shape	Size	Spatial Information		
					Not Pr	Pr	
						Position	RO

(Robot finds 3 cups, so needs to clarify which cup the RO is)
2 Robot: Which cup?

3 User: Rightmost cup.
Where this term 'rightmost' should be placed in READY? If placed at end READY will look like:



This will create a problem because next element to be processed is 'near'. We rather want FRONT to be moved to 'rightmost' instead of 'near'. To solve this problem READY should be treated now as a stack. So, instead of adding 'rightmost' to the rear of READY queue, it can be pushed on top of the READY stack. So READY now looks like the following.

Thus, ‘rightmost’ will be picked up now from READY queue and ‘cup’ that matches this property will be added into PROCESSED stack.

5.6.5 Data Flow Between READY and PROCESSED

As stated before READY is used to maintain list of all object names (in case of RO only) and descriptions mentioned in the discourse. PROCESSED keeps track of object regions that match a description. The inherent functionality of READY and PROCESSED will be discovered when we describe the data flow between them. We attempt to do this by showing an example of user input taken from our survey.

Target object in image of Fig. 5.7 is stapler (denoted by ‘X’) and known objects are ‘orange chips’ and ‘blue can’ (both denoted by ‘O’). Known objects are reported to the user at first. To make the discussion clear we have marked all blobs in the image with numbers. Here are 14 unknown object regions. We will show the state of READY and PROCESSED after each dialog. This will reveal how a robot system can proceed to successful reference resolution.

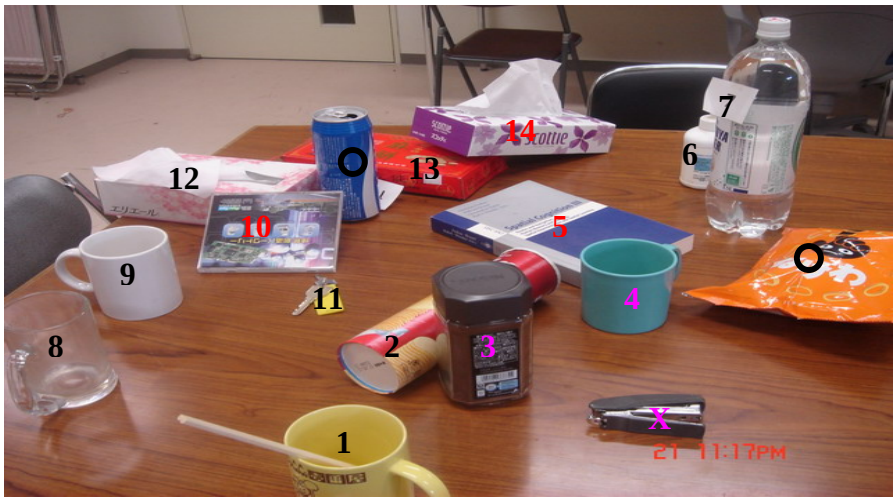


Figure 5.7: Blobs, Target and Known Objects in an Image of Survey

1 User: Near the orange chips.

READY **PROCESSED**

near	orange chips

Objects found (X, 4,5,7)	Name	Color	Shape	Size	Near	Orange chips

2 Robot: What is the color?
3 User: Black.

READY

black	near	orange chips

PROCESSED

Objects found (X,3)	Name	Color (Black)	Shape	Size	Near	Orange chips
Objects found (X,3,4,5,6,7)	Name	Color	Shape	Size	Near	Orange chips

4 Robot: Two black objects (X, 3). Which one?
5 User: Smaller one.

READY

smaller	black	near	orange chips

PROCESSED

Objects found (X)	Name	Color (Black)	Shape	Size (Smaller)	Near	Orange chips
Objects found (X,3)	Name	Color (Black)	Shape	Size	Near	Orange chips
Objects found (X,3,4,5,6,7)	Name	Color	Shape	Size	Near	Orange chips

It can be seen from the TOP of PROCESSED that, there is only object now in the first field, which is ‘X’. Robot thus reaches a decision about the target and asks user for confirmation.

CHAPTER 6
NATURAL LANGUAGE PROCESSING

6.1 Introduction

User input in the interaction contains information about the target object and reference object. In order to map the description of objects to regions in image, an interface of language processing is required. In this chapter we describe the parser, developed for user input understanding.

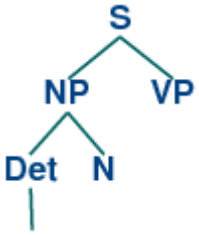
What is NLP?

From the Natural Language Processing Research Group at the University of Sheffield Department of Computer Science. "Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself". Natural Language Processing (NLP) is the use of computers to process written and spoken language for some practical, useful, purpose: to translate languages, to get information from the web on text data banks so as to answer questions, to carry on conversations with machines, so as to get advice about, say, pensions and so on.

6.2 The Architecture of Linguistic and NLP Systems (Bird et al.)

Generative grammar. contains a set of rules that recursively specify (or *generate*) the set of well-formed strings in a language. In the Chomskyan tradition, it is claimed that humans have distinct kinds of linguistic knowledge, organized into different modules: for example, knowledge of a language's sound structure (**phonology**), knowledge of word structure (**morphology**), knowledge of phrase structure (**syntax**), and knowledge of meaning (**semantics**). In a formal linguistic theory, each kind of linguistic knowledge is made explicit as different **module** of the theory, consisting of a collection of basic elements together with a way of combining them into complex structures. For example, a phonological module might provide a set of phonemes together with an operation for concatenating phonemes into phonological strings. Similarly, a syntactic module might provide labeled nodes as primitives together with a mechanism for assembling them into trees. A set of linguistic primitives, together with some operators for defining complex elements, is often called a **level of representation**.

A simple parsing algorithm for context-free grammars, for instance, looks first for a rule of the form $S \rightarrow X_1 ... X_n$, and builds a partial tree structure. It then steps through the grammar rules one-by-one, looking for a rule of the form $X_1 \rightarrow Y_1 ... Y_j$ that will expand the leftmost daughter introduced by the S rule, and further extends the partial tree. This process continues, for example by looking for a rule of the form $Y_1 \rightarrow Z_1 ... Z_k$ and expanding the partial tree appropriately, until the leftmost node label in the partial tree is a lexical category; the parser then checks to see if the first word of the input can belong to the category. To illustrate, let's suppose that the first grammar rule chosen by the parser is $S \rightarrow NP\ VP$ and the second rule chosen is $NP \rightarrow Det\ N$; then the partial tree will be:



If we assume that the input string we are trying to parse is *the cat slept*, we will succeed in identifying *the* as a word that can belong to the category DET. In this case, the parser goes on to the next node of the tree, N, and next input word, *cat*. However, if we had built the same partial tree with an input string *did the cat sleep*, the parse would fail at this point, since *did* is not of category DET. The parser would throw away the structure built so far and look for an alternative way of going from the S node down to a leftmost lexical category (e.g., using a rule S → V NP VP). The important point for now is not the details of this or other parsing algorithms; we discuss this topic much more fully in the chapter on parsing. Rather, we just want to illustrate the idea that an algorithm can be broken down into a fixed number of steps that produce a definite result at the end.

6.3 Designing a Parser for Analyzing User Input

In our interactive reference resolution system, users time to time provide information about the attributes of target object. Depending on the input the robot system looks for candidate objects in image and thus proceeds to reference resolution. Based on the survey results we have built a structure for potential user inputs and vocabulary choice for descriptors. These serve as the lexicon for our parser.

We have used a parser-development tool ProGrammar ("ProGrammar,") developed by NorKen Technologies to build the parser. ProGrammar is a visual environment for building parsers that are platform-independent, programming language-independent and reusable. Throughout the remaining sections of this chapter, we will use the notations used in ProGrammar .

6.3.1 ProGrammar Grammar Definition Language Notation

The following table summarizes the notation used in GDL.

Table 6.1: Notation Used in GDL of ProGrammar

Notation	Description
::=	Defines a production rule
;	A semicolon marks the end of a production rule
|	A vertical bar denotes *disjunction* (logical-OR)
" "	Double quotes delimit a *literal term*
' '	Single quotes delimit a *regular expression*
[ ]	Square brackets delimit an *optional term*
{ }	Curly brackets delimit a *repeater*
*	An asterisk denotes a *wildcard*
< >	Angle brackets delimit either a *length constraint*, or *symbol*

	attributes
^	A caret denotes *negation*
( )	Parentheses are used for grouping terms
(?)	A question mark, within parentheses, delimit a *parse constraint*

Grammars

A grammar consists of nonterminal symbols and other grammars. Collectively, the grammars and symbols in a project comprise a hierarchical *namespace*, in which each grammar and symbol has a unique qualified name. The namespace is used by the parse engine at runtime to resolve symbol references.

Notation

grammar *grammar-name* [**extends** $parent_1$, ..., $parent_n$]
{
declarations
};

or

grammar *grammar-name* < *attributes* > [**extends** $parent_1$**, ...,** $parent_n$]
{
declarations
};

Where:

grammar-name	Name of the grammar.
$parent_1$, ..., $parent_n$	List of the qualified names for parent grammars, separated by commas.
attributes	List of grammar attributes, separated by commas.
declarations	One or more declarations. Each declaration may be a production rule or another grammar.

Production Rules Notation

nonterminal **::=** *production* **;**

or

nonterminal < *attributes* > **::=** *production* **;**

Where:

nonterminal	Name of the nonterminal symbol.
attributes	Optional set of attributes for the symbol. See the description for *Symbol Attributes* for more information.
production	Set of instructions for how to parse the symbol; i.e., how to match it with data from the input.

6.3.2 What is a Parser?

A parser breaks data into smaller elements, according to a set of rules that describe its structure. Most data can be decomposed to some degree. For example, a phone number consists of an area code, prefix and suffix; and a mailing address consists of a street address, city, state, country and zip code.

Consider the following data:

```
(800) 555-1234
(123) 555-4321
(999) 888-7777
```

Because of the way these items are formatted, we recognize them as a list of phone numbers. The structure of these items may be described informally as:

```
"A phone number consists of a three-digit area code, enclosed by
parentheses, followed by a three-digit prefix, followed by a
dash, followed by a four-digit suffix."
```

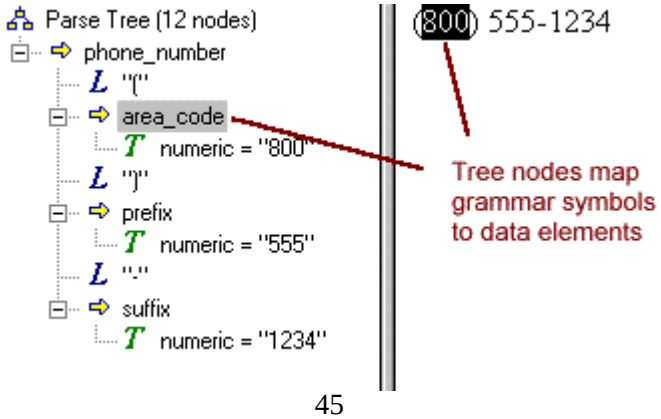
This description can be expressed more formally as a *grammar*. A grammar is a set of rules that describe the structure, or *syntax*, of a particular type of data. The following grammar describes the syntax of phone numbers:

```
phone_number ::= "(" area_code ")" prefix "-" suffix;
area_code    ::= numeric<3>;
prefix       ::= numeric<3>;
suffix       ::= numeric<4> ;
```

Each rule in the grammar, known as a *production rule*, describes the composition of a named symbol. The "**::=**" notation may be interpreted as "is composed of". Hence, the first production rule states that a **phone_number** is composed of a left parenthesis, followed by an **area_code**, followed by a right parenthesis, and so on. The next rule states that an **area_code** is composed of exactly three digits. Note how closely the grammar corresponds to the informal description of phone numbers.

Once the syntax of a data source has been described by grammar rules, a parser can use the grammar to parse the data source; that is, to break data elements such as phone numbers into smaller elements, such as area codes.

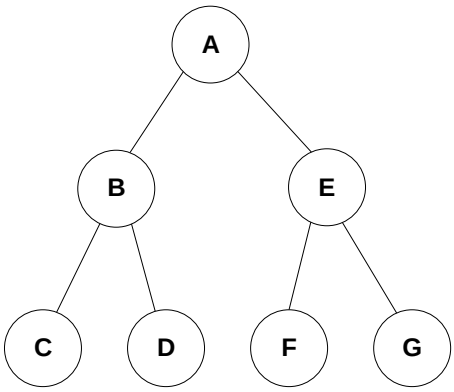
The output of the parser is a *parse tree*. The parse tree expresses the hierarchical structure of the input data. For example, the following parse tree is generated when phone number "(800) 555-1234" is parsed, using the grammar shown above:



Parsing is the process of matching grammar symbols to elements in the input data, according to the rules of the grammar. The resulting parse tree is a mapping of grammar symbols to data elements. Each node in the tree has a label, which is the name of a grammar symbol; and a value, which is an element from the input data.

6.3.3 Basic Tree Terminology

A *tree* is a data structure that consists of a set of nodes, which are connected to each other by parent/child relationships to form a hierarchy. Each node can have any number of children, but may have at most one parent. A generic tree structure is shown below:



Nodes are characterized by their positions within the tree, based on relationships to other nodes. The following table summarizes the terminology used to refer to tree nodes:

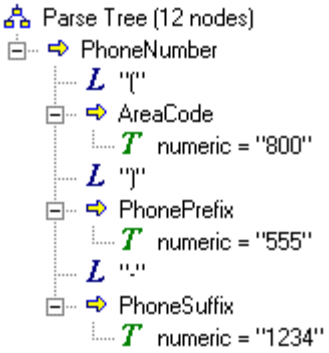
Ancestor	The *ancestors* of a node are its parent, plus it parent's parent, etc. In the above example, the ancestors of G are E and A.
Descendant	The *descendants* of a node are all of its children, plus all of their children, and so on. In the above example, the descendants of A are B, C, D, E, F, G
Sibling	Nodes having the same parent are *siblings*. In the above example, the following pairs of nodes are siblings: (C, D), (F, G), and (B, E).
Root	The *root* node is the only node in the tree that doesn't have a parent. In the above example, A is the root node.
Leaf node	A *leaf node* doesn't have any children. In the above example, nodes C, D, F, and G are leaf nodes.
Non-leaf node	A *non-leaf node* has one or more children. In the above example, nodes A, B, and E are non-leaf nodes.
Subtree	Each node in the tree defines a logical *subtree*, for which it is the root. In the above example, the subtree for B contains nodes B, C, and D; and the subtree for E contains nodes E, F, and G.

Child Index — The children of a node are numbered left to right from 0 to *n*-1, where *n* is the number of children. In the above example, the child index of B is zero (0), the child index of E is one (1), and the child index of D is one (1).

6.3.4 Parse Trees

A *parse tree* is a tree data structure that is built by the parse engine to represent the hierarchical structural of the input data. As the parser consumes data from the input file, it adds nodes to the parse tree. Each node associates a symbol in the grammar to a subset of the input file. Each node *label* contains the name of a grammar symbol, while its *value* is the range of characters from the sample file that corresponds to the value parsed by that symbol.

Consider the following parse tree:



In this tree, the child node of **AreaCode** (which is labeled "numeric") has a value of "800". The value of any non-leaf node is the concatenation of all the values in its subtree. For example, the node for **AreaCode** has a value of “800”, while the node for **PhoneNumber** has a value of "(800) 555-1234". The value of the root node is the entire input buffer.

6.4 Code for Parser Design

Using the tool ProGrammar we have designed a parser to break user input into various nodes. The values of different nodes of the parse tree are used as input for object detection module. The code developed is shown below. C like commented text (/*….*/) after each command explains the command.

```
grammar input<IGNORECASE,SPACE=" ,">
{
  file ::= [LO][{spatial_info, "and"} LO]
            [LO {spatial_info, "and"}];

  /*
  LO part contains attribute values other than relative positional information,
  spatial_info. User command can take any of the three forms;

  1. [LO]:Only attributes like color, shape, size and no positional information.
```

2. [{spatial_info, "and"} LO]: Both relative positional information and other attributes present in user input with positional information placed first.
3. [LO {spatial_info, "and"}]: Same as 2 but changed order of LO and spatial_info.
*/

```
LO::= lo_attr [object];

/*
 lo_attr are the descriptors, [object] are extra words. Example: Blue, round,
 rightmost thing. Here,words before 'thing' are included in lo_attr.
*/

lo_attr ::= [{pos_term_o}][{color}][{shape}][{size}];

/*
 pos_term_o means positional information of LO that does not make reference to
 any RO, such as 'Leftmost rectanular thing' or 'Black square thing at middle'.
 Some attributes can be omitted as well as some can be mentioned more than
 once.
*/

object <HIDDEN> ::= alpha;
color <IGNORECASE> ::= [modifier]color_name;

/*
 Color names can appear with optional modifier such as 'Light blue', 'Almost
 green'.
*/

modifier ::= "light" | "dark" | "almost";
color_name ::= "red" | "blue"|"black"|"white"|"green"|"yellow"
                          | "magenta" | "gray" | "orange";
shape <IGNORECASE>::= "round" | "circular" |"square"|
              "cylindrical"|"rectangular"|"triangular"|"flat";
size <IGNORECASE>::= "small"|"medium"|"big"|"large"|"half"
             |"smallest"|"biggest"|"largest"|"short"|"thin";
spatial_info ::= pos_term skip_to_RO {Reference, "and"};

/*
 pos_term is the relative term that connects LO to an RO. Reference is RO which
 can appear more than once as in the case of 'between'.
*/

Reference ::= [ro_attr] ro_name;

/*
 An RO itself can be described with some descriptors, ro_attr. For example, 'Left
 to the frontmost red cup'. Here 'frontmost red' is ro_attr.
*/

RO <HIDDEN> ::= "cup"|"pen";

/*
 RO refers to the name of reconized objects in a scene. Here 'cup' and 'pen' are
 just examples. Values for RO will be determined by object recognition module
```

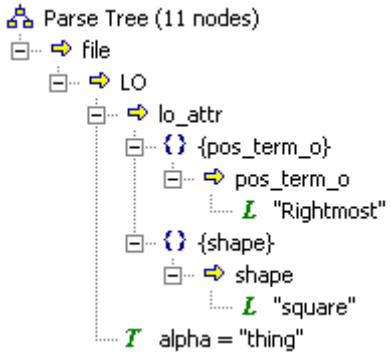
```
    of the system.
  */

  ro_attr ::= [{pos_term_o}][{color}][{shape}][{size}];
  ro_name ::= RO;
  skip_to_RO <TERMINAL , HIDDEN> ::=  *(Reference);
  pos_term <IGNORECASE> ::= "left" | "right"|"in front" |"near"|
                      "far"|"between"|"close"|"behind";
  pos_term_o <IGNORECASE> ::= "middle"|"leftmost"|"rightmost"
        |"nearest"|"closest"|"center"| "upper"||"frontmost"
        "lower"|"back"|"north"|"south"|"east"|"west"|"front";
} ;
```

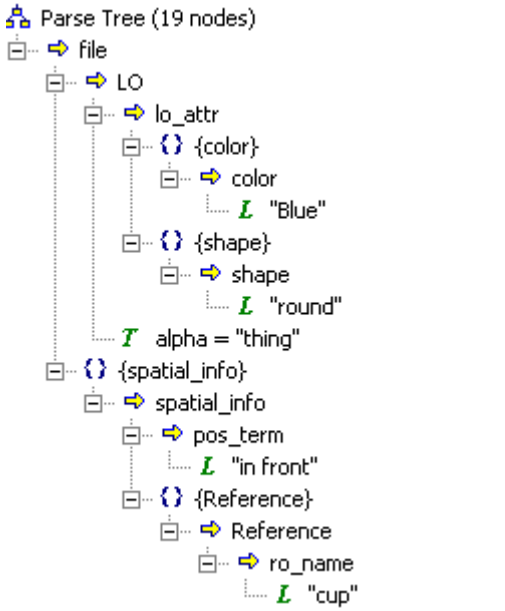
6.5 Parse Tree for Sample Input

In this section we show parse trees generated by our parser for some user inputs.

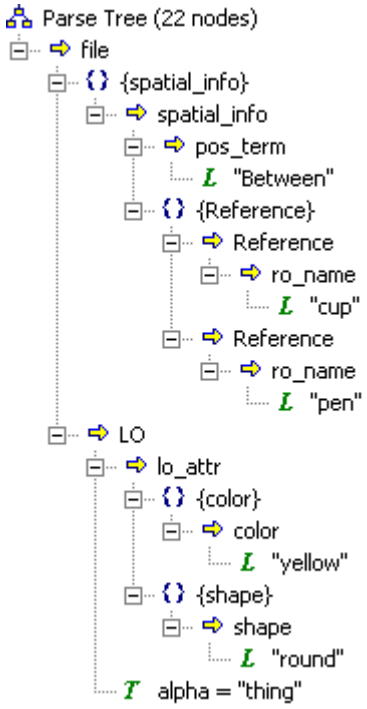
1) Rightmost square thing



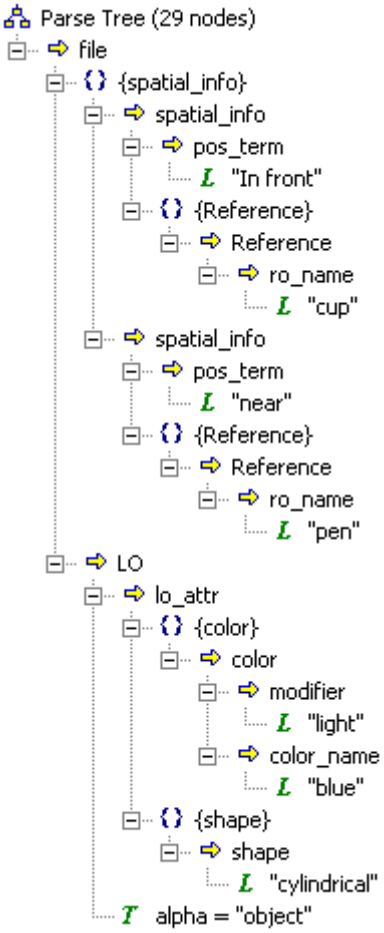
2) Blue round thing, in front of cup



3) Between cup and pen, yellow round thing



4) In front of cup and near pen, light blue cylindrical object



CHAPTER 7
SELECTION OF OPTIMAL REFERENCE

7.1 Procedure of Selection

In this chapter we describe how to find optimal reference object when known objects are more than three. We propose an algorithm to generate queries by robot so that user can describe position of the target in relation to the mentioned known object. As in the case of less than three known objects, user can tell the color, shape and size of target. The robot will not tell the name of known objects in this case. Instead, it will find the optimal reference object among the known ones and analyze the relative position of unknown objects. The idea is to omit large portion of unknown objects so that search domain reduces successively.

The following figure (Fig. 7.1) depicts the situation where there are more than three known objects (referred to as R) and some unknown objects (referred to as U).

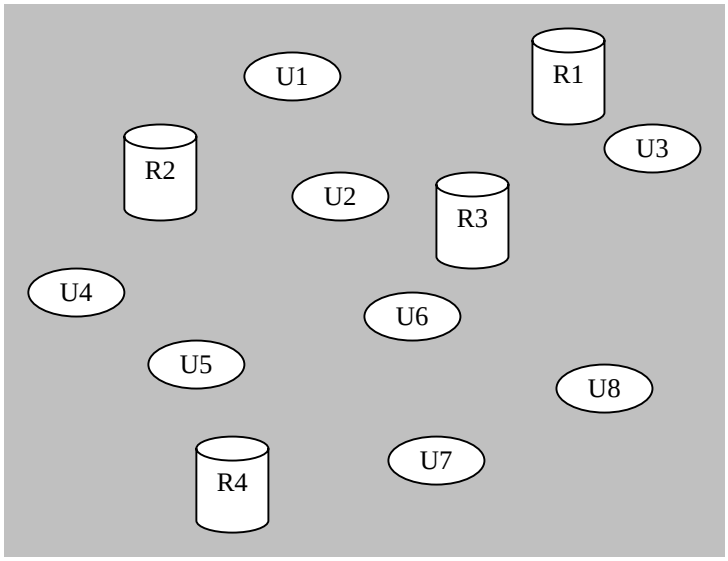


Figure 7.1 A Situation where Known Objects are more than Three

Robot first asks for color, shape and size of the target. Objects which match the given attributes can be considered at the next step. Suppose that in Fig. 7.1 U4 and U6 do not match the descriptors given by user, so they are eliminated in Fig. 7.2.

We consider an object as optimal reference if it is closest to the centroid of the group of unknown objects. In Fig. 7.2 the centroid is C1 and hence known object R3 is chosen as optimal reference. We show the left, right, front, behind regions by dotted lines centered

Let us assume that R3 is a cup and unknown object is U2. The robot then generates the query:

Robot: “Which side of the cup is the target?”
User: Left.

At this stage there may be one, two or more than two objects to the left of the cup. If there is only one object, robot shows it to user for confirmation. If there are two objects robot decides in which direction (horizontal or vertical?) these objects are more dispersed from each other. The next query will involve either left/right (in case of horizontally dispersed) or front/rear (in case of vertically dispersed).

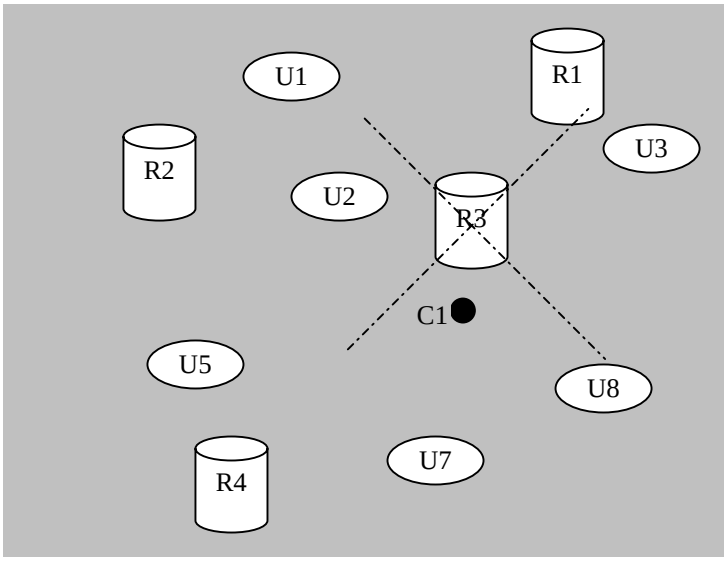


Figure 7.2: Some Objects Eliminated from Fig. 7.1. C1 is Centroid of Group of Blobs, R3 is Optimal Reference Here.

For more than two objects step to decide optimal reference object will be repeated. For the example in Fig 7.2, candidate objects for target have been reduced to U1, U2 and U5. Centroid for U1, U2 and U5 is C2 and R2 is closest to it (Fig. 7.3). While being asked “Which side of R2 is the target?”, the user will answer “right”. Now, there are two objects U1 and U2 to the right of R2 and next question of robot will be “Front one or rear one?”. Thus the robot confirms that front object U2 is the target.

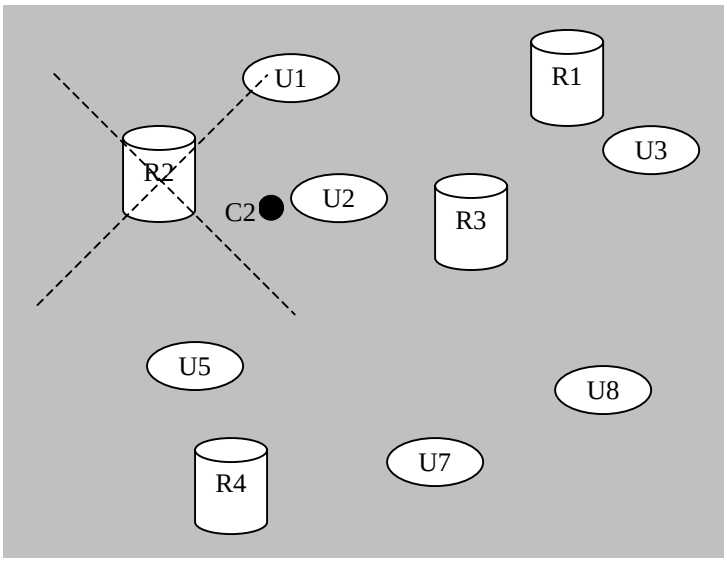


Figure 7.3: A new Centroid C2 is calculated. R2 is Optimal Reference Here.

7.2 Algorithm

The whole procedure of optimal reference selection and target object detection described above is included in an algorithm (Table 7.1).

Table 7.1: Algorithm for optimal reference selection

1. Ask about color, shape and size of the target.
2. Find objects which match the given descriptors and exclude others.
3. Locate centroid of the group of unknown objects.
4. Find known object closest to this centroid and consider it as reference.
5. Ask for position of target in relation to the reference.
6. Find objects in given region.
7. If one object in this region show this to user for confirmation and go to step 12.
8. If two objects, go to step 9. Otherwise go to step 11.
9. If $\|Cx1-Cx2\| > \|Cy1-Cy2\|$ ask, “Left or right one?” Else ask, “Front or Rear one?”. [(Cx,Cy) = Center of mass of objects]
10. Find object in step 9, show this to user for confirmation and go to step 12.
11. Repeat from step 3.
12. Exit.

CHAPTER 8
CONCLUSION

8.1 Summary

Household service robots are intended to assist elderly and disabled persons and hence, typical objects present in a house need to be located. While object recognition is the prior condition to carrying out user command, no robot system can recognize all objects in a scene with the existing methods. Moreover, from previous research in this field, it is evident that dialogical interaction helps human users feel the robot more like a real interaction partner. To assist service robots in detecting intended objects, an interactive reference resolution is developed in this study. Through conversation with a user, this system collects information about position of the target in image as well as color, shape and size. It then implements these descriptors to locate target in a scene. A survey is carried out in order to acquire an understanding of vocabulary for descriptors used by human participants. The outcome of the survey also suggests in which situation a certain response is emerged. Various aspects of the result lead to designing a framework for interaction with user. Qualitative positional descriptors are turned into quantitative measures to find regions around an object in the image. Selection of optimal reference object among several known objects is also covered in this study.

8.2 Limitations of this Research

This study is a step towards assisting a service robot to locate intended household objects. The deployment of a service robot has to cover many areas like understanding of 3D world, navigation, map building, path-planning, grasping objects, interested person detection and so on along with target object detection. This study merely covers object localization in images. Moreover, separation of object regions and detection of color and shape did not always produce expected results, especially for complex background and objects. Although these are the basic modules before starting any interaction, this study did not make much effort to improve these modules. Rather, we concentrated more on spatial reasoning. Moreover, using only 2D information for image does not provide information about actual orientation of objects in environment.

In survey with human participants, number of known objects was two for all trials. This choice does not remain open the opportunity for assessing user's choice of optimal reference. Furthermore, it is not possible to comprehend from the survey how a robot might lead the conversation from the very beginning to detect a target. Besides this, some participants in the survey playing the role of "Robot" were unable to decide what should be their word choice in a conversation. It happened because as a human being it was difficult to portrait themselves with a low cognitive ability. It was not well defined to them to what extent they should show "intelligence". For the same reason, "Human" participants also faced difficulty in deciding which utterance could be understood by their "Robot" partner. The survey would have been more realistic if we could employ a real robot whose "intelligence" is well explained to "Human" partner.

Although we propose an interactive methodology, we have not implemented it online during the period of research. We also do not suggest any evaluation criteria of proposed method. Performance of this method is also not measured. Moreover, no speech processing is done and no speech input is handled.

8.3 Recommendations for Future Study

Humans employ a variety of para-linguistic social cues (facial displays, gestures, etc.) to regulate the flow of dialogue (Fong et al., 2003). Merging these clues with our interactive method of reference resolution may produce better result. To detect blobs and to recognize objects, recent statistical methods of feature extraction can be utilized. A concrete “Dialog Manager” module is essential to implement an interaction system. The purpose of this module will be to control flow of conversation, interpret dialogs and trigger appropriate response.

Moreover, it is prospective to include “Situation awareness” into a service robot system. Situation awareness has been formally defined as "the *perception* of elements in the environment within a volume of time and space” (Endsley, 1988). Limiting the object search database is possible if a robot can interpret a situation it exists in. Thus, mapping a situation to a pre-callibrated object database significantly reduces search domain.

REFERENCES

Abella, Alicia and Kender, J.R. (1993). Qualitatively describing objects using spatial prepositions. IEEE Workshop on Qualitative Vision, pp. 33 - 38

Abella, Alicia and Kender, John R. (1994). Conveying spatial information using vision and natural language. AAAI Integration of vision and natural language processing workshop.

Abella, A. and Kender, J.R. (1999). From Images to Sentences via Spatial Relations. Integration of Speech and Image Understanding, pp. 117 - 146

The Bicycle pedal is in Front of the Table. Why some Objects do not Fit into some Spatial Relations (2008). Functional Features in Language and Space. Laura Carlson et al., Oxford University Press.

Bird, Steven, Klein, Ewan and Loper, Edward. "Introduction to Natural Language Processing." from http://nltk.org/book/.

"Blob" from http://www.v3ga.net/processing/BlobDetection/

Brenner, Michael (2007). Situation-Aware Interpretation, Planning and Execution of User Commands by Autonomous Robots. 16th IEEE International conference on Robot & Human Interactive Communication.

Clark, H. and Brennan, S. (1991). "Grounding in Communication." Perspectives on socially shared cognition , pp. 127-149.

Clark, H. H. and Wilkes-Gibbs, D. (1986). "Referring As a Collaborative Process." Cognition Vol. 22, pp. 1-39.

Claus, Berry, Eyferth, Klaus, Gips, Carsten, Hörnig, Robin, Schmid, Ute, Wiebrock, Sylvia and Wysotzki, Fritz (1998). "Reference Frames for Spatial Inference in Text Understanding." Spatial Cognition - An interdisciplinary approach to representing and processing spatial knowledge, pp. 241-266.

Conklin, EI. and McDonald, D.D. (1982). Salience: The Key to the Selection Problem in Natural Language Generation. The 20th ACL Conference, Toronto, pp. 129-138

Defining Functional Features for Spatial Language (2008). Functional Features in Language and Space. Laura Carlson et al., Oxford University Press.

Dobnik, Simon and Pulman, Stephen (2008). Teaching a robot spatial expressions. Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, UK, pp.

Donna, K. Byron and James, F. Allen (2002). What's a Reference Resolution Module to do? Redefining the Role of Reference in Language Understanding Systems. 4th Discourse Anaphora and Anaphor Resolution Colloquium, pp. 80-87

Endsley, M. R. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. The Human Factors and Ergonomics Society 42nd Annual Meeting, pp. 82-86.

Fischer, Kerstin and Lohse, Manja (2007). Shaping naive user's model of robot's situation awareness. 16th IEEE International conference on Robot & Human Interactive Communication. Korea.

Fong, T., Kunz, C., Hiatt, L. M. and Bugajska, M. (2007). Using Vision, Acoustics, and Natural Language for Disambiguation. ACM/IEEE International Conference on Human Robot Interaction, pp. 73-80

Form and Function (2008). Functional Features in Language and Space. Laura Carlson et al., Oxford University Press.

Fransen, B., Morariu, V., Martinson, E., Blisard, S., Marge, M., Thomas, S., Schultz, A. and Perzanowski, D. (2006). The Human-Robot Interaction Operating System. ACM conference on HRI, pp. 41-48

Gapp, K. P. (1998). Object Localization: Selection of Optimal Reference Objects. 2nd International Conference on Spatial Information Theory, Berlin.

Gieselmann, Petra (2004). Reference Resolution Mechanisms in Dialogue Management. CATALOG '04, The 8th Workshop on the Semantics and Pragmatics of Dialogue Barcelona, Spain.

Grabowski, J. (1999). "A Uniform Anthropomorphological Approach to the Human Conception of Dimensional Relations." Spatial Cognition and Computation Vol. 1, pp. 349-363.

Haagen, C.H. (1949). Journal of Psychology.

Hossain, M.A., Kurnia, R., Nakamura, A. and Kuno, Y. (2006). "Interactive Object Recognition through Hypothesis Generation and Confirmation." IEICE Transactions on Information and Systems Vol. E89-D(7), pp. 2197-2206.

Is it in or is it on? The Influence of Geometry and Location Control on Children's Descriptions if Containment and Support Events (2008). Functional Features in Language and Space. Laura Carlson et al., Oxford University Press.

Keller, J. M. and Wang, Xiaomei (1998). Comparison of spatial relation definitions in computer vision. 3rd International Symposium on Uncertainty Modelling and Analysis IEEE Computer Society pp. 679

Kurnia, R., Hossain, M.A., Nakamura, A. and Kuno, Y. (2006). "Generation of Efficient and User-friendly Queries for Helper Robots to Detect Target Objects." Advanced Robotics Vol. 20, pp. 499-817.

Kurnia, R., Hossain, M.A., Nakamura, A. and Kuno, Y. (2006). Use of Spatial Reference Systems in Interactive Object Recognition. The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), pp. 62

Levinson, S. C. (1996). Frames of Reference and Molyneux's Question: Crosslinguistic Evidence. Language and Space. Paul Bloom et al.
Lindeberg, T. (1988). "Feature Detection with Automatic Scale Selection." International Journal of Computer Vision Vol. 30(2), pp. 77-116.

Lopes, L. Seabra and Teixeira, A. (2000). Human-Robot Interaction through Spoken Language Dialogue. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems.

Lowe, David G. (1999). Object recognition from local scale-invariant features. International Conference on Computer Vision, pp. 1180-1187

Mangold, R. (1986). "Sensorische Faktoren beim Verstehen " erspezifizierter Objektbenennungen Vol. 188.

Mansur, A., K., Sakata and Kuno, Y. (2007). Recognition of Household Objects by Service Robots Through Interactive and Autonomous Methods. International Symposium on Visual Computing, pp. 140-181.

Mansur, A. and Kuno, Y. (2007). Integration of Multiple Methods for Robust Object Recognition. SICE Annual Conference, Kagawa, Japan.

Moratz, R., Fischer, K. and Tenbrink, T. (2001). "Cognitive Modeling of Spatial Reference for Human-Robot Interaction." International Journal on Artificial Intelligence Tools Vol. 10(4), pp. 889-611.

Moratz, R. and Tenbrink , T. (2003). Instruction modes for joint spatial reference between naive users and a mobile robot. IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, pp. 43 - 48.

Moratz, Reinhard and Tenbrink, Thora (2002). Natural Language Instructions for Joint Spatial Reference between Naive Users and a Mobile Robot. 11th IEEE International Workshop on Robot and Human Interactive Communication, pp. 229-234

Moratz, Reinhard, Tenbrink, Thora, Bateman, John and Fischer, Kerstin (2003). "Spatial Knowledge Representation for Human-Robot Interaction." Spatial Cognition III: Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Reasoning.

Müller, Rolf, Röfer, Thomas, Lankenau, Axel, Musto, Alexandra, Stein, Klaus and Eisenkolb, Andreas (2000). "Coarse Qualitative Descriptions in Robot Navigation." Spatial Cognition II. Lecture Notes in Artificial Intelligence Vol. 1849, pp. 268-276.

Pattabhiraman, Thiyagarajasarma (1992). Aspects of salience in natural language generation. Vancouver, B.C, Simon Fraser University. **Ph.D. Thesis**.

"ProGrammar." from http://www.programmar.com.

Roy, Deb K. (2002). "Learning Visually-Grounded Words and Syntax for a Scene Description Task." Computer speech and language Vol. 16, pp. 383-388.

Schultz, Alan C. and Trafton, J. Gregory (2006). Using Computational Cognitive Models for Better Human-Robot Collaboration. ICRA 2006 Workshop on Cognitive Robots and Systems. Florida, USA.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M. and Brock, D. (2004). "Spatial Language for Human-Robot Dialogs." IEEE Transactions on Systems, Man, and Cybernetics Vol. 34(2), pp. 184 - 167.

Skubic, Marjorie, Perzanowski, D., Schultz, A. and Adams, W. (2002). Using Spatial Language in a Human-Robot Dialog. IEEE International Conference on Robotics and Automation, pp. 4143 - 4148

Sofge, D., Perzanowski, D., Skubic, M., Cassimatis, N.L., Trafton, J.G., Brock, D., Bugajska, M., Adams, W. and Schultz, A. (2003). Achieving Collaborative Interaction with a Humanoid Robot. Second International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS).

Space and Language (1999). Language and Space. Paul Bloom et al., The MIT Press.

Spatial Perspective in Descriptions (1999). Language and Space. Paul Bloom et al.

Stopp, E., Gapp, K.P., Herzog, G., Laengle, T. and Lueth, T. C. (1994). Utilizing Spatial Relations for Natural Language Access to an Autonomous Mobile Robot. The 18th German Annual Conference on Artificial Intelligence, pp. 39-80.

Taylor, H. J. and Taversky, B. (1992). "Descriptions and Depictions of Environments." Meomory and cognition Vol. 20(8), pp. 483-496.

Taylor, Tamsen E., Gagné, Christina L. and Eagleson, Roy (2000). Cognitive Constraints in Spatial Reasoning: Reference Frame and Reference Object Selection. "Smart Graphics", AAAI 2000 Spring Symposium. Stanford, CA, USA

Tenbrink, Thora, Fischer, Kerstin and Moratz, Reinhard (2002). "Spatial Strategies in Human-Robot Communication", KI, Vol. 16(4), pp. 19-23.

Thora, Tenbrink (2003). "Communicative Aspects of Human-Robot Interaction." Languages in development.

Thora, Tenbrink and Moratz, R. (2003). Group-based Spatial Reference in Linguistic Human-Robot Interaction. The European Cognitive Science Conference, Germany.

Torralba, A., Murphy, K. P. and Freeman, W.T. (2007). "Sharing Visual Features for Multiclass and Multiview Object Detection." IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. 29(8), pp. 884-869.

Torralba, A., Murphy, K.P., Freeman, W.T. and Rubin, M.A.: (2003). Context-based Vision System for Place and Object Recognition. IEEE International Conference on Computer Vision, pp. 273-280

Torralba, A., Oliva, A., Castelhano, M. S. and Henderson, J. M. (2006). "Contextual Guidance of Eye Movements and Attention in Real-world Scenes." Psychological Review Vol. 113, pp. 766-786.

Treisman, A. (1988). "Features and objects." Quarterly Journal of Experimental Psychology Vol. 40A, pp. 201-237.

Tversky, B., Lee, P. and Mainwaring, S. (1999). "Why Do Speakers Mix Perspective?" Spatial Cognition and Computation Vol. 1, pp. 399-412.

APPENDIX

CONVERSATION FOR REFERENCE RESOLUTION IN SURVEY

Image 1 1 Robot: I know magazine and orange. 2 User: I want a white cylindrical object. 3 Robot: Many cylindrical objects. Which one? 4 User: It is close to the magazine. 5 Robot: Give more information. 6 User: Its upper portion is red. 7 Robot: Yes, I find it. 8 User: Ok.	Image 2 1 Robot: I know yellow can and orange. 2 User: Give me rectangular leftmost thing. 3 Robot: Is it almost white? (Pointed to the stapler box). 4 User: No, it's orange color. 5 Robot: The thing placed at left upper corner? 6 User: Yes, correct.
Image 3 1 Robot: I see tea bottle. 2 User: Object is cylindrical. 3 Robot: What color? 4 User: Purple. 5 Robot: I cannot see anything purple. Where is it? 6 User: Near most of the tea bottle. 7 Robot: Left or behind the tea bottle? 8 User: Left. 9 Robot: Do you want red thing? 10 User: Yes. 11 Robot: I found it.	Image 4 1 Robot: I can see red chips, blue can. 2 User: I want to have a round object. 3 Robot: I can see many round objects, which one do you want? 4 User: I want to have the white one, which is in front of red chips. 5 Robot: Ok I can understand. This one? (Pointed to white hand soap). 6 User: Right, thanks
Image 5 1 Robot: I can see telephone and table lamp. 2 User: Just right of telephone and front of table lamp. 3 Robot: What color? (Confused with the yellow memo pad). 4 User: Gray. 5 Robot: Big or small? Square? 6 User: No. Large, thin. 7 Robot: I found the thing.	Image 6 1 Robot: I can see Ocha bottle and Cider bottle. 2 User: Middle of these two bottles, blue color. 3 Robot: Do you want cylindrical blue thing? 4 User: Yes. 5 Robot: I found it.

Image 7 1 Robot: I see monitor, table lamp. 2 User: Medium size, gray, right position, square object. 3 Robot: You want this? (pointed to gray rightmost file). 4 User: No. Left to that one. 5 Robot: Is it gray or black? 6 User: Deep gray, not black. 7 Robot: Do you want this? 8 User: Yes.	Image 8 1 Robot: I see Uncle Benz and teapot. 2 User: I want white object that is at the left of Uncle benz. 3 Robot: Is it cylindrical or rectangular in shape? (Had both cup and stapler box in mind). 4 User: Cylindrical. 5 Robot: Ok, I can find the object. 6 User: Yes, correct.
Image 9 1 Robot: I know potato and umbrella. 2 User: I want object in center.. 3 Robot: What color? 4 User: Light green. 5 Robot: What is the shape? 6 User: I can't describe. 7 Robot: Which object is near the target? 8 User: Potato. 9 Robot: Do you want this? 10 User: Yes.	Image 10 1 Robot: I see coffee and two bottles. 2 User: I want small and yellow object. 3 Robot: Is the target between the two bottles? 4 User: Yes. 5 Robot: What is the target's shape? 6 User: One part is square and another triangular. 7 Robot: Do you want this? 8 User: Yes.
Image 11 1 Robot: I can see teapot and cups. 2 User: In front of cup. 3 Robot: Which cup? 4 User: Rightmost cup. 5 Robot: Nearer to that cup? 6 User: Yes. 7 Robot: I found the thing. 8 User: Ok correct.	Image 12 1 Robot: I can see orange chips and blue can. 2 User: Nearer to the orange chips. 3 Robot: What is the color? 4 User: Black. 5 Robot: Two black objects. Which one? (Pointed to coffee jar and stapler). 6 User: Smaller one. 7 Robot: Ok, found it.

<table>
<tr>
<td>Image 13

1 Robot: I know books and cans.
2 User: I want a white object.
3 Robot: White can? (misunderstood what is meant by can).
4 User: No white can here. It’s cylindrical.
5 Robot: Which side of blue book?
6 User: Right.
7 Robot: Is it this white object? (pointed to handsoap).
8 User: No. Its near the orange can.
9 Robot: Tell more.
10 User: White, cylindrical and right side of orange can.
11 Robot: This one? (pointed to correct target.)
12 User: Yes, it is.</td>
<td>Image 14

1 Robot: I see cups here.
2 User: Red, cylindrical, middle of the image.
3 Robot: Left side of nearest cup?
4 User: Yes.
5 U1: You want this object?
6 User: Yes.</td>
</tr>
<tr>
<td colspan="2">Image 15

1 Robot: I can see CD and bottles.
2 User: Object is close and left of CD, blue, rectangular.
3 Robot: Do you want this?
4 User: Yes.</td>
</tr>
</table>