ICS-10M-834

# SPATIAL RECOGNITION FOR HUMAN ROBOT INTERACTION

by

Lu Cao

Supervisor: Professor Yoshinori Kuno

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

Date of Submission: February 5, 2010

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

## ABSTRACT

Service robots are expected to carry out user requirements by intuitive instructions, though they have limited perceptual capabilities for surroundings and objects. We engage to develop a robotic system, assisting people to accomplish simple tasks in daily life (e.g., fetching objects for handicapped and elderly people). In these tasks, they are inevitably involved in various kinds of objects. The objects are described by their intrinsic attributes: color, size, shape etc in most cases and robots can recognize them successfully. However, it is difficult to detect objects if intrinsic attribute descriptions fall or be ambiguous. In this research, we aim to resolve this problem using position explanations and spatial relationships among objects.

Our robot system has been assumed to recognize some object classes and specific objects in autonomous object recognition. How human users, being aware of the limited object detection capability of their robot partner, describe objects in images is of primary interest. Two surveys have been conducted in this regard: one is to examine how human illustrates the position of target object among objects, the other one is to analyze how human chooses reference objects in a scene where several unknown objects are obtained. According to the results derived from the surveys, we first select seven prepositions which obtain high ranks in sum total and define a model for their mathematical presentations, and then a strategy on choosing optimal reference object is proposed which is alike in human cognition. To design an effective system to implement into the robot, we mainly utilize reference and relative reference systems to the tasks. We also use group-based reference system. It can be considered as the relative reference system using a group of objects as a reference object.

Moreover, a view-point problem arises while the speaker (human) and listener (robot) stand at different position in a scene. The diversity of information they hold results in varied descriptions to target and reference objects. We propose to eliminate this distinction by building up a database to store spatial relationships for every object or group in a scene. Preliminary experiments in simple scenarios with only a few objects have shown promising results.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1
# INTRODUCTION

## 1.1 Background and Motivation

Service robots are designed to be able to perform preprogrammed physical tasks, act under the direct control of a human or autonomously under the control of a pre-programmed computer. In recent years, with aging and social development, they are expected to carry out user requirements through intuitive instructions, for instance, fetching objects for handicapped and elderly people, rather than needing to be programmed by experts.

In a typical service robotics scenario, a robot is instructed by a human user to act upon a specific object. To achieve this aim, both participants need to negotiate their internal conceptual representations linguistically to identify the referent [1]. Between humans, such communication is fairly simple and straightforward, because humans have the ability to specify reference objects by their class names such as “the cup is on the table”. Even in a situation where it is difficult to name all of the trivial objects, humans naturally communicate by employing pointing gestures. Robots, however, have limited perceptual capabilities that often preclude accurate recognition of broadly similar objects and, moreover, may not have access to the necessary world knowledge that would identify the object by class [2]. A corresponding to the HRI instruction, an accommodated perceptual way to the robot may be more like that: “the white little reflecting object is on the brown round table (Assuming that the robot has known about table).” By means of this elaborate description, it complicates to establish an effective joint between human and robot. Thus, the spatial configuration and the position of the object relative to the robot itself can be used for linguistic reference. This work therefore concentrates on such an option. Namely, to interpret the user command, the robot locates the object by spatial differentiation among objects.

## 1.2 Problem statement and why it should be solved

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. Besides, humans have a vast cognitive database, which help them interpret requests without any error. We can detect objects in any kind

of their orientation, shape, size and texture once identify 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 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.

Furthermore, for most of existing approaches to the representation for objects in a finite space are accumulated a large number of datum to demonstrate the sizes, shapes and locations, for example, Object A is 34cm wide, 20cm high at (45,60). It is difficult to visualize the spatial configuration described by the sentence, even if paper and pencil are at hand. Thus, while quantitative approaches are useful in domains in which exact data is available, they have several drawbacks, such as complexity, falsifying effects partial and uncertain information, missing adequacy and transformational "impedance" [3].

Consequently, this research intends to overcome these drawbacks, making the robotic system involving spatial reasoning less complex, and more accessible to human users.

### 1.3 Qualitative Spatial Knowledge as a Communicative Method

The psychologist Jackendoff [4] has noted that there are many ways to describe what an object is, but few ways to describe where an object is. He has examined this disparity from both a language perspective and a cognitive perspective, and suggests that there is substantial filtering of information when going from a cognitive representation of a spatial scene to its linguistic representation. This may account for many spatial relationships going unexpressed, including those for describing what an object is in terms of its complex contours and textures. But describing where an object is appears to use even less complex geometric terms. Human can specify an object by its shape, color and texture but hard to describe the location. For example, Levin [5] has demonstrated a case where a person could image what an object looked like but could not image the spatial relationship of objects.

To address this problem, a powerful strategy for achieving reference in human-human communication should be considered more closely. Qualitative spatial reference then serves as a necessary bridge between the metric knowledge required by the robot, and

more ‘vague’ concepts that build the basis for natural linguistic utterances, as suggested by Hernàndez [3]. Whereas many objects may have some particular color, size or texture—which therefore give rise to more potential confusion, or ‘distracters’, for a referential expression—the position of an object is generally uniquely defined; if identified sufficiently restrictively, only one object is in a given place at a time. This could make the use of explicit positional information a good strategy for achieving unique reference in the human-robot communicative situation as well [2]. In this research, we focus on positional information for reference resolution and propose an appropriate solution to locate objects by means of spatial relationship representation.

### 1.4 Objectives

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. They are stated below:

i. To examine how human users describe location of objects applying semantic terms in images. What phrases are used and which ones are most commonly used gain more interests.
ii. To inspect how human users choose reference objects to explain location of objects in complex scenes. Under this circumstance, what objects are most liable to be chosen as reference objects and what features they have are crucial to the system.
iii. To analyze how to represent spatial relationships of objects and formulate patterns on position expressions are the core parts of this research.

# CHAPTER 2
# LITERATURE REVIEW

## 2.1 Linguistics and Spatial Reference Representation

Spatial reasoning and its application have been focused by researchers both in linguistics and robotics for many years. Authors in [6] are interested in how people talk about space and what they can do 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 [6] that, spatial terms cannot be derived simply from an interface between language and a set of sensory/perceptual maps. For example, the expressions "The cake is in the box on the table." The meaning of "**in"** does not simply map to surroundings in the visual display. One must appeal to some abstract relationship, such as a capacity for containment that box shares, 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 [7].

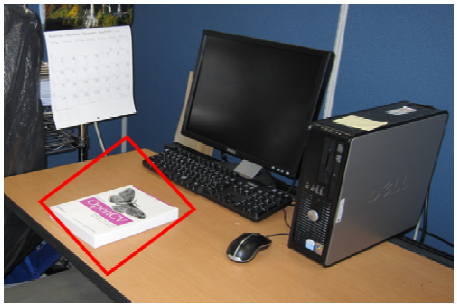
Although there have been considerable researches on the linguistics of spatial language for humans, only limited work done in using spatial language for interacting with robots. Some researchers have proposed a framework for such an interface. Moratz et al [8] investigated the spatial references used 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. 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 [9]. Thus, we investigated linguistics literature about spatial relationship usage.

## 2.2 Spatial Instructions

### 2.2.1 Using Intrinsic, Relative and Absolute

Humans use reference systems to describe object positions. The relation between human's spatial cognition and language expression has been well studied in the field of psychology, linguistics, and other related fields. Levinson [10] has proposed that humans use three kinds of reference systems: **intrinsic**, **relative**, and **absolute**. In **intrinsic** reference system, the relative position of one object (the referent) to another (the relatum) is described by referring to the relatum's intrinsic properties such as front or back. For example, the expression such as "the book in front of you" is good enough to describe the position of the book since the front of a human body is intrinsically determined. Note that intrinsic means external comparison is not needed [7]. On the other hand, in **relative** reference system, we use a position of a third entity as origin instead of referring to inbuilt features of the relatum. An example is "viewed from the cup, the pen is to the left of the box." In **absolute** system, neither a third entity nor intrinsic features are used for reference. Instead, we use some absolute direction specification terms, for example, such as north and south.

Fig. 2.1 shows an example scene. In this case, we may say, "The book (marked in red) is in front of the computer display," by using the intrinsic reference system, or "The book is to the left of the mouse (viewed from the speaker or listener)," by the relative reference system. The front direction of a computer display can be determined regardless of the viewpoint, whereas we need to specify the viewpoint to determine the left or right of a mouse. The viewpoint is usually omitted when it is either the speaker or listener.




ce object determines the
ıce system is intrinsic in
ference object, reference

### 2.2.2 Group-based Reference System

In addition to the above three reference systems, the group-based reference system has been proposed [1] [5]. When there are multiple same or similar objects in the scene, humans consider them as a group, describing the position of an object in the group by the spatial relation between the object and the total group. In Fig.2.2 (a), four books overlapped each other form a group, and the group can locate other objects around it using such reference system. Moreover, the notion of group can be extended, as the case may be. We could put objects together as a group even if they are different kinds. For example, we may indicate the book marked by the circle in Fig. 2.2(b) as the rightmost object by considering the coffee, the bottle and the book on the table as a group.

roup;
other

## 2.3 Previous Work Navigation

This subsection briefly summarizes our previous work.

### 2.3.1. Autonomous Object Recognition

Dating back to the work by Winograd [11], there has been a great deal of research that inquires into the ways robot systems understand the scene or task through interaction with the user [12][13][14]. These studies, however, have dealt with objects that can be described by simple word combinations such as 'blue box' or 'red ball'. In our application domain, objects are usually more complex, and are thus not so easily recognized by the robot. For example, we may want the robot to bring us a bag of potato chips, in which the package has various colors. However, we may not seem to use complex expressions to describe such complex objects. Thus, we performed observation experiments of human-human interactions and have found that humans usually use simple expressions to describe complex objects [15]. For example, humans usually use

only one major color to describe multicolor objects. The system should understand such human expressions to achieve user-friendly interaction. Based on this finding, we have proposed a vision system identifying multicolor objects even when the user mentions one color [15]. We have developed an integrated vision system by combining an autonomous object recognition method and an interactive one. The basic strategy is that the system first tries to recognize the object asked by its user. If the system cannot detect the object or make a mistake, the system turns into the interactive mode. Thus, we use the method developed by Dipankar and Mansur et al. [16] for autonomous object recognition. The method can deal with both specific object recognition (e.g., detection of 'Coke can') and category-level object recognition (e.g., detection of any can) by choosing an appropriate vision module depending on the recognition problem (either specific or category-level) and the target object. The method also learns which vision module is appropriate for particular objects in advance from various examples of the objects. Fig. 2.3 shows some examples of autonomous object recognition results.

Fig.2.3 Autonomous object recognition results

### 2.3.2. Object Detection by Color

Humans often describe multicolor objects by one color for the background or the largest area of the objects. Thus, we have developed a vision program to detect multicolor objects whose background or large area color is the color specified by the user [15]. The background color is determined as the color of the region whose convex hull is the largest among neighboring regions in the color segmentation result. Fig.2.4 explains this process. Fig.2.5 shows an example of color object detection result.

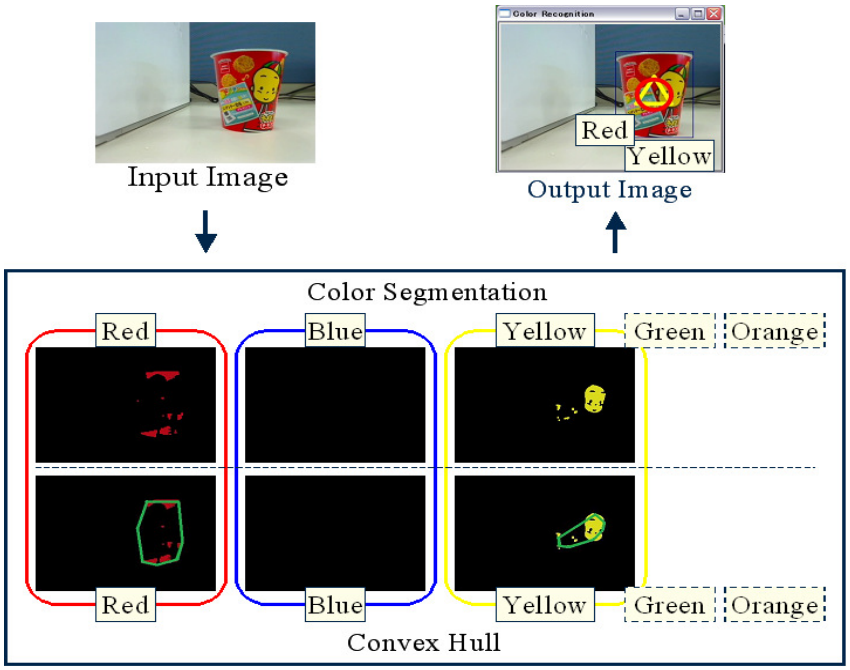



Fig.2.4 Object color decision. The color of this object is red (the color of the largest convex hull). The second candidate is yellow (the color of a large area)

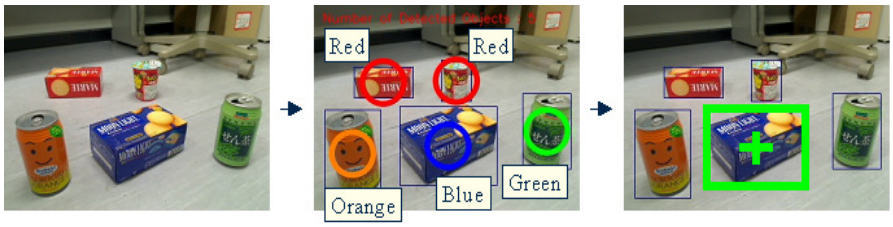



Fig.2.5 Color object detection example. If the user wants a blue object, the system can detect the object indicated by '+'.

# CHAPTER 3
# SURVEY WITH HUMAN PARTICIPANTS

## 3.1 Background

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, choose reference objects in 2D images is of our primary interest. In need of observing linguistic preference of humans, 2 surveys are designed and carried out.

## 3.2 Participants

90 participants in total, 50 of them are Japanese, 30 are Chinese, and 10 are English native speakers, in Saitama University.

## 3.3 Surveys

### Survey 1: Where is the target object?

#### ◆ Objective

It is impracticable to communicate with robot with neither too complex nor too simple semantic sentence, uncommonly used words also go beyond the knowledge base of robot. In neither case may cause interpretations. Thus, learning how human beings locate objects with concise and explicit words/phrases among similar objects and in what cases such words/phrases would be utilized are first two problems we need to take into consideration.

#### ◆ Role of Participants

45 pairs of participants stood in front of 2 scenarios in turn. One (Participant A plays a role of **Human User**) requested his/her partner (Participant B plays a role of **Robot**) to point out the objects under instructions. With a shared view of an image where there are several objects, the challenge of the User 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 solely speak in English. The only gesture for the robot user is to point to an object, 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. Fig.3.1 is

the representation of the experiment setting.

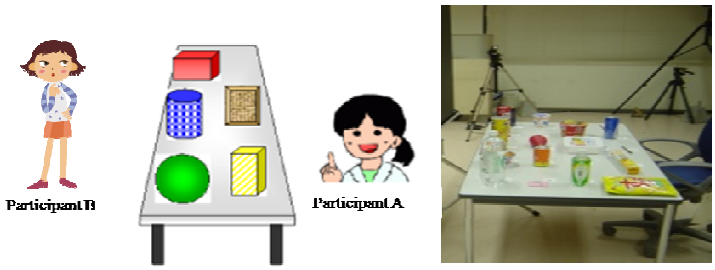



Fig.3.1 Experiment Setting

◆ **Scenarios and Design**

To obtain description of objects both in different situations and orientations of objects, 4 scenarios were chosen (shown in Fig. 3.2, from left to right). All of the objects are useful in our daily life. For each scene we decided the known and target objects. Known objects (labeled by "x" from now on in this paper) were reported to both users and "robots" so that only these objects are used for reference during conversation. Identity of target object (labeled by "o" from now on in this paper) was shown only to users in written form. The users were advised to choose proper words/phrases to depict the location.

Table 3.1 lists known and target objects defined by us for all Scenarios.

| Scenario | Target Object | Known Object |
|---|---|---|
| 1 | Scissors<br>Potato chips box | Floppy, Book, Remote Control, Dishwasher, Bag, Cup |
| 2 | Bottle, Box | 2 Bottles, Book, Box, Mouse, Tissue, Projector |

When describing the target, barely the words/phrases being bond up with orientations can be used, descriptions of detail are not recommended. We also suggested that the Users it would be best to apply straightforward words/phrases to picture the position of target objects. Fig.3.2 (ii) as an example to demonstrate the rules: the expression, such as: "The scissors is on the book/right to the floppy" was desired, but "The scissors is on the book, with white and red cover" was not advocated in this survey.

◆ **Procedure**

We noted down every conversation the pairs said. An excerpt of conversation for scenario 2(see Fig. 3.2(ii)) is shown below.

***(Target:*** Bottle***; Known Object:*** Book, Box, Mouse, Tissue, and Projector***)***

***User****: I want the bottle. Can you see it?*

***Robot****: Yes. I can see 3 bottles. Which one do you want?*

***User****: It is in front of the projector.*

***Robot****: Is this one? (Pointing to it)*

***User****: Yes.*

◆ **Results**

The conversations between 45 pairs of participants could be discussed from various viewpoints.

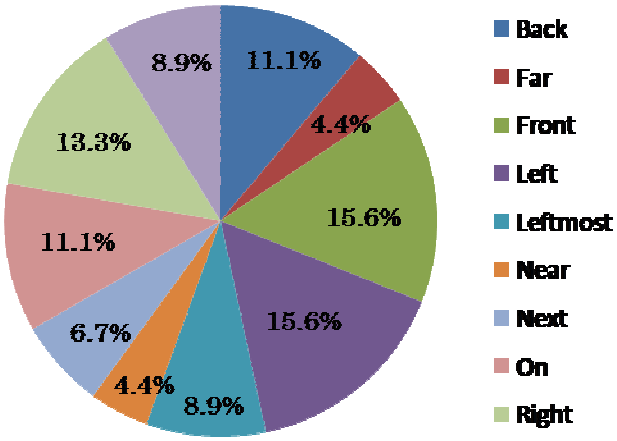
**(a) Vocabulary and Proportions**

All of the prepositions the participants mentioned are lined up in Table 3.2 by dictionary order. Subsequently, we counted up number of times they appeared and percentage usage of attributes. Table 3.3 shows the percentages of spatial relationship-based utterances.

| Prepositions | Example | Number of Times |
|---|---|---|
| Back | The bottle is at the back of the book | 5 |
| Far | The potato chips Box is far away from the | 2 |

| | | |
|---|---|---|
| | cup | |
| Front | Bottle is on the left side of the projector | 7 |
| Left | The box is on the left side of the tissue | 7 |
| Leftmost | The bottle is the leftmost one of 3 bottles | 4 |
| Near | The bottle is near the book | 2 |
| Next | The Scissors is next to the remote control | 3 |
| On | The Scissors is on the book | 5 |
| Right | The box is on the right side of the book | 6 |
| Under | The box is under the mouse | 4 |

ımbers




Apparently, the vocabularies which indicate direct orientations are used more frequently than others, such as **front, left.** As long as they are utilized, the listener can build up a map in mind and comprehend based on their natural representations. Correspondingly, the words **near, far,** and **next** are seldom applied due to their indirect implications. Whether the distance between two objects is near or far principally depends on the location where speaker stands in a scene and his viewpoint. If someone intends to explain with these words, he should ensure that his partner employ the same viewpoint with him, or else, at least the partner can obtain sufficient information from his

perspective. Particularly, **leftmost** implied identifying the intended object by its position relative to the other objects in the group using group as a whole as reference. The results revealed that more than half of the participants were liable to use direct orientations prepositions to instruct their partner as well. During the survey, by referring direct words, above 80% of the pairs would locate the target object successfully at their first attempt. Only 20% achieved the target by using the latter ones. The number of unsuccessful attempts varied greatly between users, and the reasons for changing instructional strategies also differed 70% of the speakers turned to replace the words with more distinct ones, whereas barely 30% repeated the previous instructions with the same words.

In this case, the ensuring instructions did not conform to our expectations. In scenario ii there were basically two kinds of reference systems that we expected to be employed to describe the bottle here: one is to simply refer to "**the bottle is in front of the projector**" by utilizing intrinsic system and another one is "**the bottle is on the left side of the projector**" using relative system. As the partner stood at the opposite side of the table, he could not be sure about that, it would be equally sensible to refer to the word as in "**the bottle is at the back of the projector**". Once the speaker add "From my position/your position" as a complimentary to avoid ambiguous, the partner experienced success at his second time.

Based on these results, we summarized several prepositions, they were: left-right, front-back, on-under, and leftmost. To adapt to our robotic system, we choose left-right, front-back and leftmost-between (middle)-rightmost as our domain. The last one is adequate to group reference system.

**(b) Binary Presentations**

The result also revealed the environment on how to utilize basic prepositions (front.etc). Summary of those prepositions that involve a reference object and a target object are defined as follow. We take **Front** as an example, which is illustrated in Fig.3.3.

**Front and Back,** respectively, requires that the projection on the y axis of the bounding box of the target object be above or below, respectively, the projection of the bounding box of the reference object.

**Left and Right** are defined analogously to **Front and Back**.

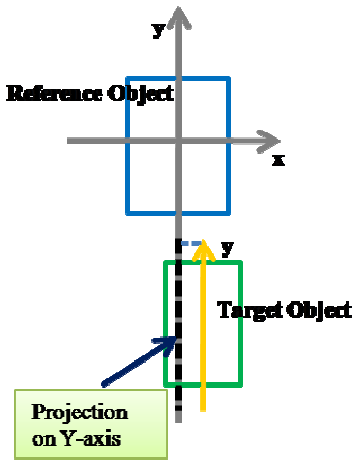



Fig.3.3 Binary Presentation of **Front**: y-axis projection of bonding box of target object is below the reference object

**Between (Middle)** is a ternary relationship. Ideally, the projection of the center of the target object is collinear with the centers of the two reference objects flanking it, and right at the midpoint of their connected line. In this case, "**middle**", a more elaborate preposition is used (refer to Fig3.4). In practice, between is defined when the projection of the center of the target object is on the connected line between two reference objects flanking it.

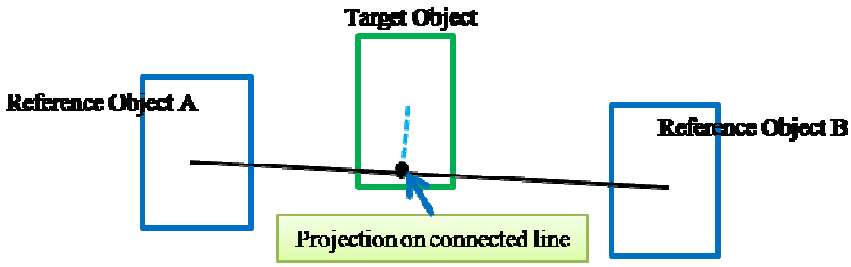



Fig.3.4 Binary Presentation of Middle

**(c) Use of Deictic Words**

We are also interested in use of deictic words such as this/that. In some cases, there are no noteworthy use is observed, but in most of cases, they indicate different distance from the speaker. "This" presents the nearer one, whereas "That" expresses the further one. Humans are accustomed to speaking out them with pointing gesture. In viewing of the fact that our robot is not able to recognize pointing gesture accurately, we prefer the user repeating full name of the object which he implied before. Furthermore, to remind the partner of the possible mistake, the robot made a query.

***User**: The book is right to the pen*

***Robot**: I can see two pens. Which one?*

***User**: The blue pen.* *(Instead of pointing and saying: "That one.")*

***Robot**: Is this one?* *(Pointing to it)*

***User**: Yes.*

**Survey 2: How does your reference object choose?**

◆ **Objective**

It is indicated that human favors to choose the object that he deems the most suitable one to express his intension. A human is to infer his partner even if they have different perspectives. However, one of the most distinctions between robot and human is robot be lack of the ability. Robot is not able to perceive the surrounding we feel, the color we identify, and the semantics our meaning implies unless human specially manipulate it. In this survey, we attempt to inspect how human chooses reference object to express his purpose obvious and straight in a complex scene.

◆ **Role of Participants**

20 participants were divided into 5 groups, 4 persons at a time in turn. We requested them to announce the reference object they chosen for by answering the question: "Where is xx?"

◆ **Scenarios and Design**

As Survey1, which we demonstrated above, 4 scenarios were arranged (shown in Fig.3.5, from left to right, up to down), only English was allowed as well. All of the objects are unique and identified in each scenario. We also determined target objects—in all of the scenarios, identical target was coffee can, only in Scenario 4, bottle(labeled in red) was the another target in that we wondered whether humans would switch reference object or not if there were multi-targets. The participants could select reference objects at will, and they were also allowed to use adjectives such as *black*, *round* to show us more depictions.

◆ **Procedure**

The question was concise and straightforward in first three scenarios. In Scenario 4, there were two queries: after locating the can, we inquired where the bottle was. An excerpt of conversation for scenario 2(Fig. 3.5(iii)) is shown below.

***Writer**: Please tell me where the can is.*

***Student**: It is next to a yellow cup.*

◆ **Results**

Table 3.4 shown the number of times which the objects to be as reference in each scenario. The first two numbers are noted by color of dark red.

(i) (ii)


(iii) (iv)

Fig.3.5 Scenario Scripts in Survey2

Obviously, result of Scenario 1 showed a fact that the objects which were closer to target is much more preferred than others. Scenario 2 indicated that salience objects which are large in size with a shape/color have much more tendency as reference than others. In result 3, there wasn't any quantitative difference among the number. The only one and most significant factor that we could not ignore was puppet not seemed to be effortless to depict for some participants like bottle, cup and tissue. Consequently, objects which are readily describe in semantics gain more advantages than others. The can which took the first place in result 4 was 4 times as many as the ball and blue box. It

applied that linguistic context is much more convince than select a new object as reference.

**Scenario 1:**

| Reference Object | Times of Used |
|---|---|
| Ball | 4 |
| Blue Box | **8** |
| Book | 2 |
| Bottle | noun |
| Cup | **7** |

**Scenario 2:**

| Reference Object | Times of Used |
|---|---|
| Ball | **10** |
| Book | noun |
| Bottle | noun |
| Cup | noun |
| Pencil | 4 |
| Tissue | **6** |

**Scenario 3:**

| Reference Object | Times of Used |
|---|---|
| Bottle | 3 |
| Cup | **8** |
| Puppet | 4 |
| Tissue | **5** |

**Scenario 4: Note that the result of locating coffee can was omitted as needed.**

*Target: Bottle*

| Reference Object | Times of Used |
|---|---|
| Ball | **3** |
| Book | noun |
| Blue Box | **3** |
| Can(As target in Q1) | **12** |
| Cup | noun |

| | |
|---|---|
| Pencil | 2 |
| Tissue | noun |

Table 3.4: Results of Survey 2: References and Occurrence Numbers

According to [7], entities can be salient by being very vivid, pervasive, and unique, or by being spoken about most recently. The higher the salience of an entity is the greater is its likelihood of selection during content planning.

**Vividness and Imageability**: [17]

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.

Hence, we set the criterion listed below:

1. Distance has the most priority: distance between target object and reference, i.e., objects closer to target are preferred.
2. Shape, size, color etc.
3. Objects which are easily expressed.
3. Linguistic context, i.e., objects which have been previously mentioned and which are in focus, are linguistically more salient.

We comply with the strategy in our experiments.

# CHAPTER 4
# PROPOSED METHODOLOGY

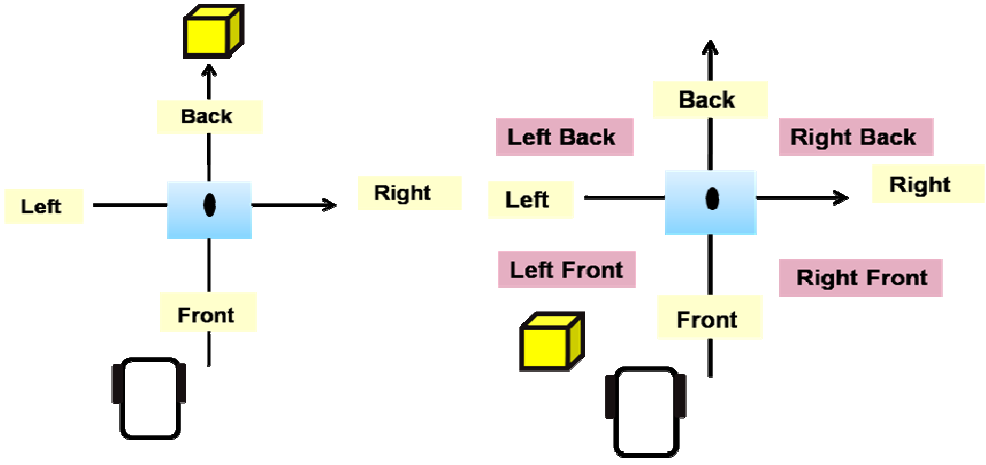
We present a qualitative method to describe spatial relationship among multiple objects. The absolute reference system may not be necessary in our service robot domain. Thus, we consider the relative, intrinsic, and group-based reference systems only.

## 4.1 The interpretation of bounding relations

An essential aspect of the robot's ability to execute instructions is its interpretation of the spatial relations specified between objects functioning as reference or the target objects. Different kinds of reference systems required for interpreting linguistic references according to the three options outlined in section 2 and for handling the corresponding instructions. Our survey results already allow us to exclude several theoretically possible alternatives that were not, in fact, selected as strategies by our experimental participants: for example, intrinsic and relative reference systems employing either the speaker or a salient object as origin.

The bounding expressions are then further resolved as follows. We have implemented the relative system, which may be most often used to represent spatial relationships. The relative system has three entities: **referent** (target object), **relatum** (reference object), and **origin** (viewpoint). The origin is often omitted and the default origin is usually the robot, sometimes the speaker. In our implementation, we assume that the origin is the robot. If the user (speaker) and the robot are looking in almost the same direction, the speaker origin coincides with the listener origin. If we take the robot's point of view as origin, all objects are represented in an arrangement resembling a plan view. Thus, the reference axis is a combination of two directed lines through the center of the object as a relatum, which is demonstrated in Fig. 4.1(i). The center of the bonding area can be used as point-like representation (marked in "●"). Note that the vertical divides the reference plan as left and right parts while the horizontal manages the front and back parts.

For more finely partition, the reference axis is rotated for 45 degrees, respectively, new orientation relations are found, which are called left-front, left-back, right-front and right-back. For combined expressions like "left-front" vs. precise expressions like "strict front", we use the partition presented in Fig. 4.1(ii).




(i) (ii)

Fig.4.1 Relatum and reference direction

Thus, to define the partitions formally, we take the angle Ø between the reference direction and the directed straight line from the relatum to the referent is defined (see Fig.4.2). The relations between spatial prepositions and Ø can be defined as:

referent *front* relatum : $-\pi/4 \leq \theta \leq \pi/4$

referent *back* relatum: $3/4\,\pi \leq 0 \leq 5/4\,\pi$

referent *left* relatum: $\pi/4 \leq \theta \leq 3/4\,\pi$

referent *right* relatum: $-\pi/4 > \theta > -3/4\,\pi$

referent left *front* relatum: $0 < \theta < \pi/2$

referent left *back* relatum: $\pi/2 < \theta < \pi$

referent *right front* relatum: $0 > \theta > -\pi/2$

referent *right back* relatum: $-\pi/2 > \theta > -\pi$

referent *strictly front* relatum: $\theta = 0$

referent *strictly left* relatum: $\theta = \pi/2$

referent *strictly behind* relatum: $\theta = \pi$

referent *strictly right* relatum: $\theta = -\pi/2$

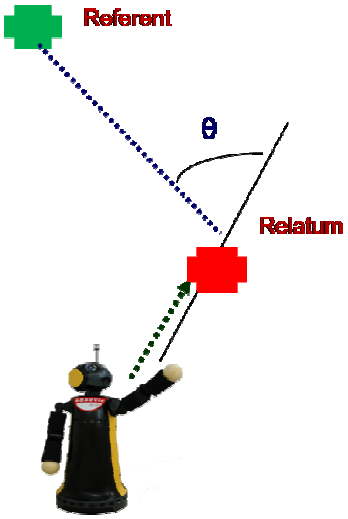



Fig.4.2 Relative reference model

## 4.2 Factors that May Influence Strategy Selection

According to our empirical results, speakers order their strategies in the way they do because of their hypotheses about baseness and difficulty. In particular, those speakers who did not try out the goal naming strategy at all may have assumed that this kind of complex instruction is too difficult for the robot. In the following, we look for further evidence that supports our hypothesis that in this particular situation, baseness and difficulty is relevant for the speakers. There are several observations that point in the same direction:

### 4.2.1 Viewpoint

Unlike in communication among humans, the speakers in our experiment 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.

In our experiment study, we consider relative reference system needs the viewpoint. Depending on the viewpoint, the orientation may be different. In Fig. 4.3, if the user and the robot are at the different locations, the can is described in two ways: for user, the can is left to the blue cookie box while for robot; the can is in front of the blue cookie box. This typical scenario indicates us that the diversity of information they hold results in varied descriptions to target and reference objects. We propose to eliminate this distinction by building up a database to store spatial relationships for every object or

group in a scene. We show the results in Chapter 6.

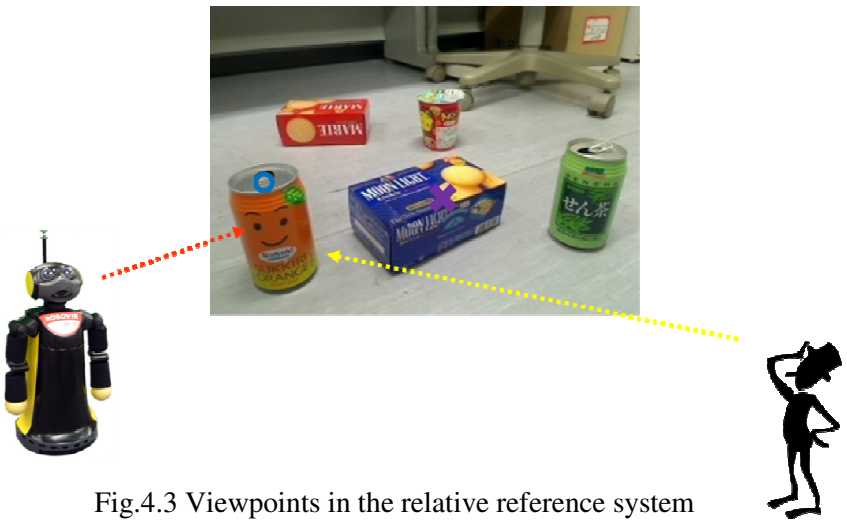
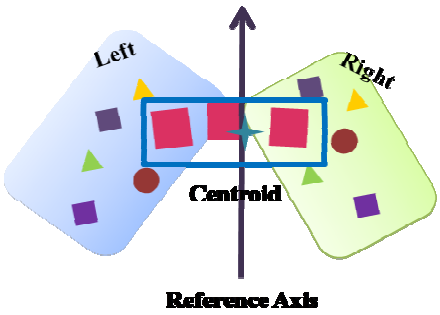


Fig.4.3 Viewpoints in the relative reference system

### 4.2.2 Group-Based Reference

As pointed out in previous surveys, many participants made use of the concept of a *group* in order to specify the position of one of its members. However, the question needs to be asked why many users did *not* use this concept, as it turned out to offer (in this scenario) an unambiguous referential strategy involving a linguistically simple kind of instruction. One reason for many users' failure to take advantage of this might be that the users did not expect the robot to be able to grasp the concept of a group, as this involves comparison, identification of similarity, and categorization.

In our strategy, the group-based reference system is considered to be the relative reference system using the group as a relatum, the centroid of the group serves as virtual relatum. Fig.4.4 shows the reference direction is given by the directed straight line through the center of the group. The objects labeled by blue box are considered as a group, same color implies they own the same attributes. The object closest to the group centroid can be referred to as the "middle object", then the left and the right one can be distinguished as well.




### 4.2.3 Linguistic Constructions

Speakers wondered both during the experiments and in the questionnaires about the linguistic capabilities of the robot, asking whether it understood particular words or syntactic constructions, such as relative clauses. Thus, they attended to the fact that the robot might have limited linguistic capabilities. Furthermore, most speakers employed ellipsis imperatives, a linguistic strategy rarely used in task-oriented human-to-human dialogues, as it completely lacks the kinds (sometimes rather complex) of elaborations which are loose to the letter of construction, such us: “right to the object that I indicated last time”. Under the restriction that object names cannot be mentioned, it was hard for robot to understand.

In our interactive reference resolution system, users provide information about the spatial information of target object to the robot only at the beginning of interaction. Providing informative context is essential for system detection. Based on the survey results we build a fundamental parser to analyze user command which have some restrictions listed as follow:

- **Vocabulary and Syntactic constructions**

As the robot has limited linguistic ability, simple and common use vocabularies would be better to analyze. In current system, theses vocabularies to describe position are legal:

Front – in front of
Back –at the back of
Left – on the left side of
Right – on the right side of
Left front/back, Right front/back

Leftmost – Between (Middle) –Rightmost

There are a lot of ways to describe location. In our system, we restrict user input command in the format like: **Get/Bring me A. A is xxxx B**. A is a delegates of target object while B indicates reference object, **xxxx** represents preposition. Note that complex syntactic structure such as clause is not supported at this stage.

◆ **Ellipsis**

Since robot cannot tell meaning of a word which is omitted from its context, our strategy imposes restriction on ellipsis. Experimentally, user is suggested that repeat the context which has been mentioned above in case of failure.

◆ **Punctuation and Notation**

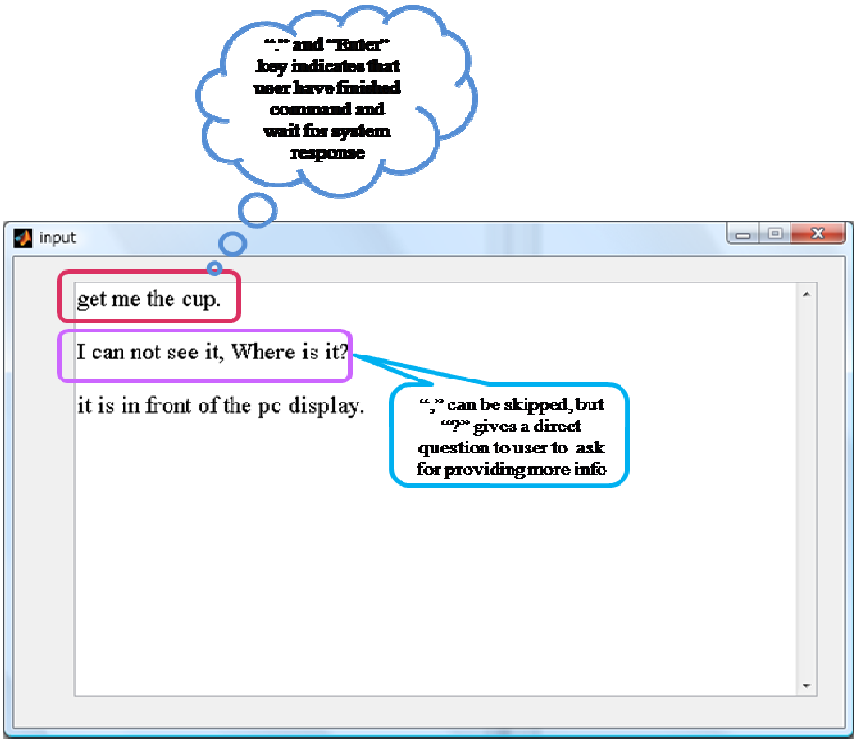
Notations are the symbols that they stand for or suggest something else. In our system, a few notations are applied to not only deliver different intends but also indicate user and robot.

Collectively, we provide that a **full stop** is placed at the end of sentence as defined in typography; a **comma** is principally for separating things; a **question mark** is used at the end of an interrogative sentence, Once question mark appears, the parser analyzes the sentence, either waiting user input or system response. A **space** is a blank area devoid of content, serving to separate words, and a **Carriage Return *is*** a control character that commands user to submit the input to the parser or switch to a new line. Once the carriage Return is entered, the parser detects the last full stop/question mark where it appears, and analyzes the sentence. The following table summarizes the functions:

An example is shown in Fig.4.5, the user input is marked in red and system response is in purple, respectively.

Note that only English semiangle charaters are valid, Japanese and Chinese symbols are not supported yet, but case insensitive.

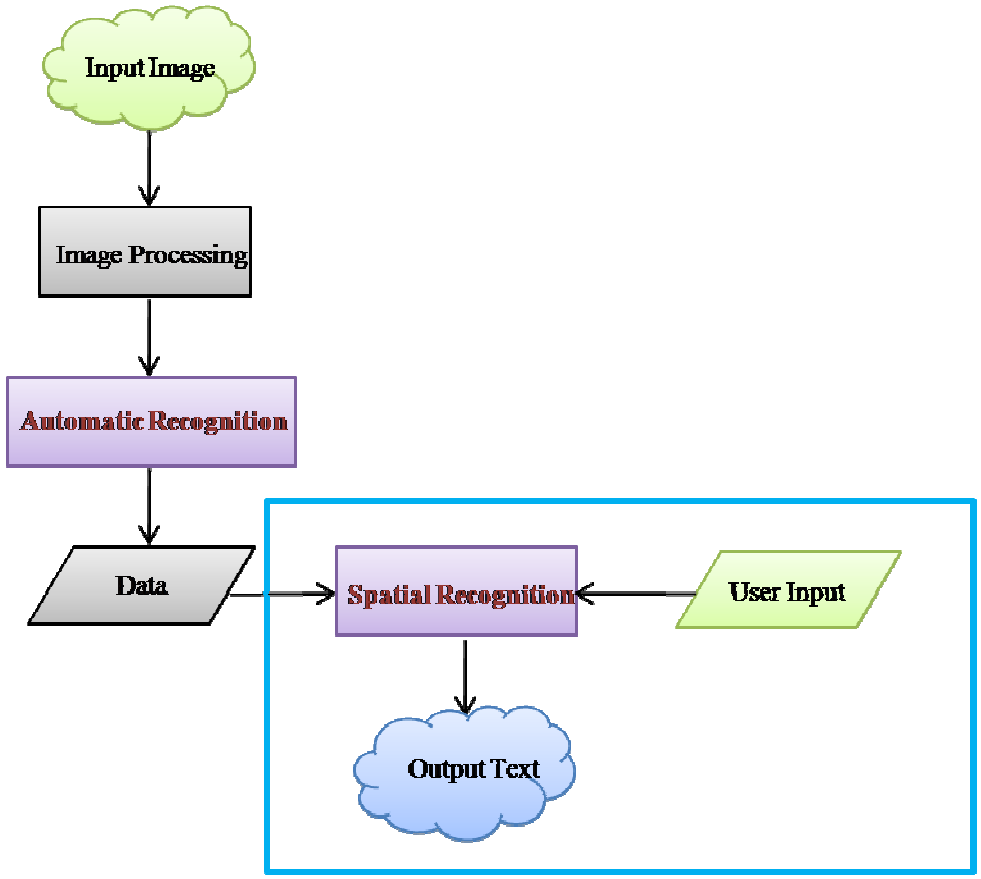
| | |
|---|---|
| , | Separating words and sentences |
| . | Placed at the end of a sentence, exclusive use only waits for the next step |
| ? | Used at the end of an interrogative sentence, exclusive use only waits for the next step |
| **Space** | Separate words |
| **Enter** | Switching to a new line |
| | Submit input command to the system |
| **“.”+ “Enter”** | Submit input command and wait for system response |
| **“?” + “Enter”** | Submit question and wait for system response |




ig.4.5 Command Input Parser

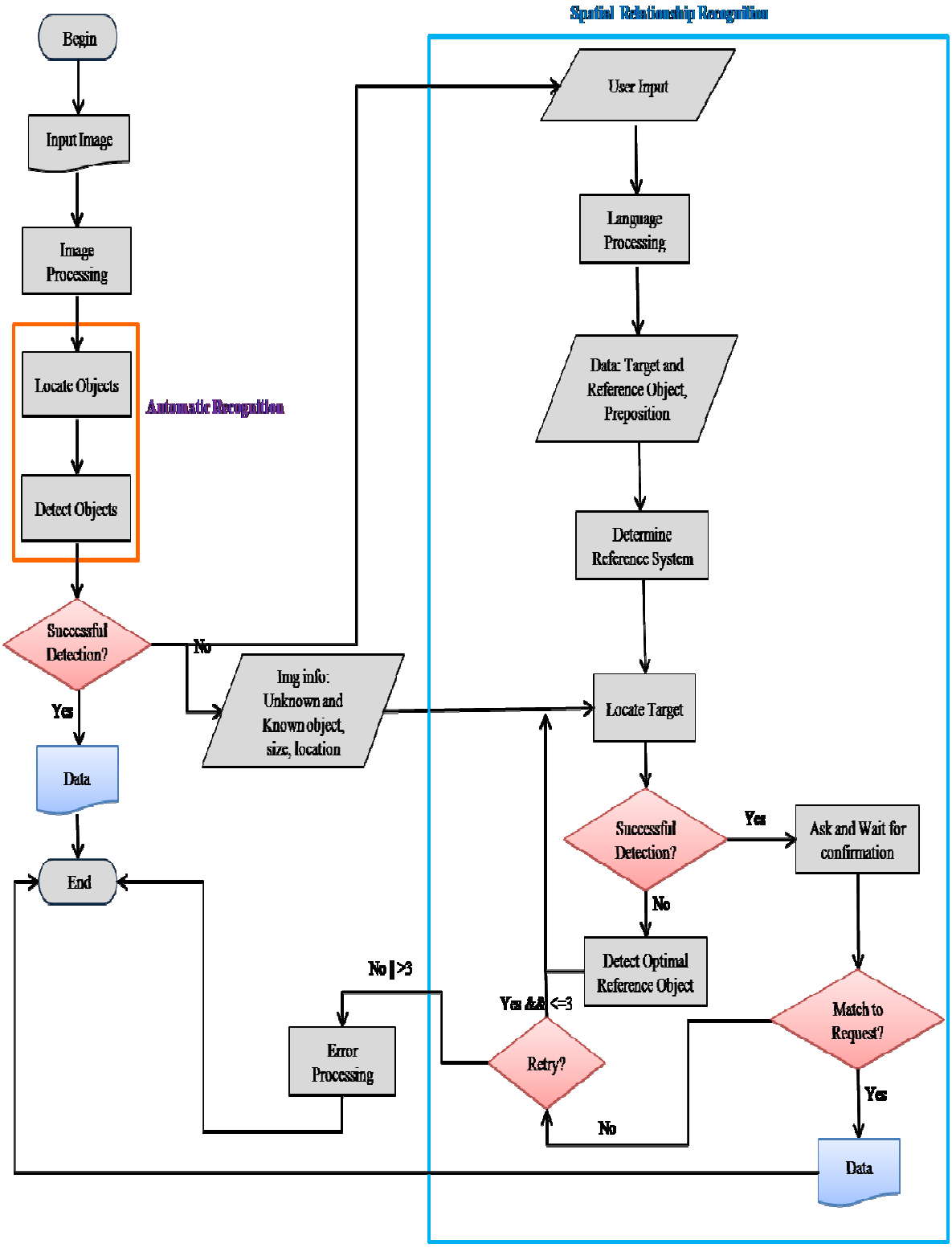
# CHAPTER 5
# SYSTEM ARCHITECTURE AND EXPERIMENTS

In this chapter, we present the approach for multiple objects spatial recognition. Fig.5.1 illustrates the system architecture. Generic module is denoted by light squares, input and output data by clouds, domain-specific modules by purple squares, and domain specific data by quadrangle. The input data and generic modules form the foundation of the system, and the majority of the theoretical contributions of this work are incorporated into these modules. System generality is attained through the use of the interchangeable domain-specific modules, which contain sufficient information about the particular domain necessary to produce meaningful interpretations of the spatial relationship of the object pairs. Our work mainly focuses on the spatial recognition processing which is labeled by blue rectangular box. In section 5.1, we propose flow-processing diagram, shown in Fig.5.2.




ıd output data

## 5.1 Flow-Processing Diagram




odel and Spatial

The input image is first processed to localize and detect objects by the automatic recognition model. The method is based on finding one or more probable locations of an object within an image using a generative model, and then evaluating these locations using a discriminative classifier. It combines the advantages of discriminative methods with those of probabilistic generative models [18]. At this stage, the objects are distinguished by their distinct features (bag of key-points) which have been recorded into our dataset. The ones, whose features do not belong to the dataset, the model noted them as "**unknown object**". It may happen that the target object's model is in dataset 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. Then, the system switches to interactive processing—recognizing **unknown objects** under user's instructions. In the meantime, a number of data are generated, which associates with the input image, known, and unknown objects (name, size, pixel, coordinate.etc).

The system waits for user's input, after that, analyzes the structure of the sentence and separates the phrase which refers to target object. 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. 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 a few 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 ten. 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. We prefer that user inquire about the object he attempts to be a reference whether the robot can "see" or not

(i.e. “Can you see cup?”). If the response is “Yes”, then the object is chosen as a reference, or else, user will not alternates another choice until the answer is “Yes”.

Generally, robot complies with the hints which are provided by user’s input command to locate target object. However, sometimes, user may give an inaccurate description, or sometimes robots may not comprehend the semantic context perfectly, for example, an object is described as “left to a cup” position. From other indicators it was clear to the robot that “in front of” should be the description. No matter what in either case, robot cannot detect the target object normally. Instead of re-asking user or treating as an error, robot still continue to locate the target object by detecting an optimal reference object processing based on the strategy, which we have listed in the last chapter. Then, robot locates target object again in relation to this reference object and gives feedback to user.

In addition, if the system doesn’t detect the proper target object, there are still 3 attempts for user to retry.

### 5.2 Object Recognition

As we expressed in Section 2, there are dozens of household objects already recognized at autonomous recognition phase such as foods and drinks. They are listed in Fig.5.3. Others but in the same scenario are bounded as unknown objects. Recognition of objects in an image is shown in Fig.5.4.

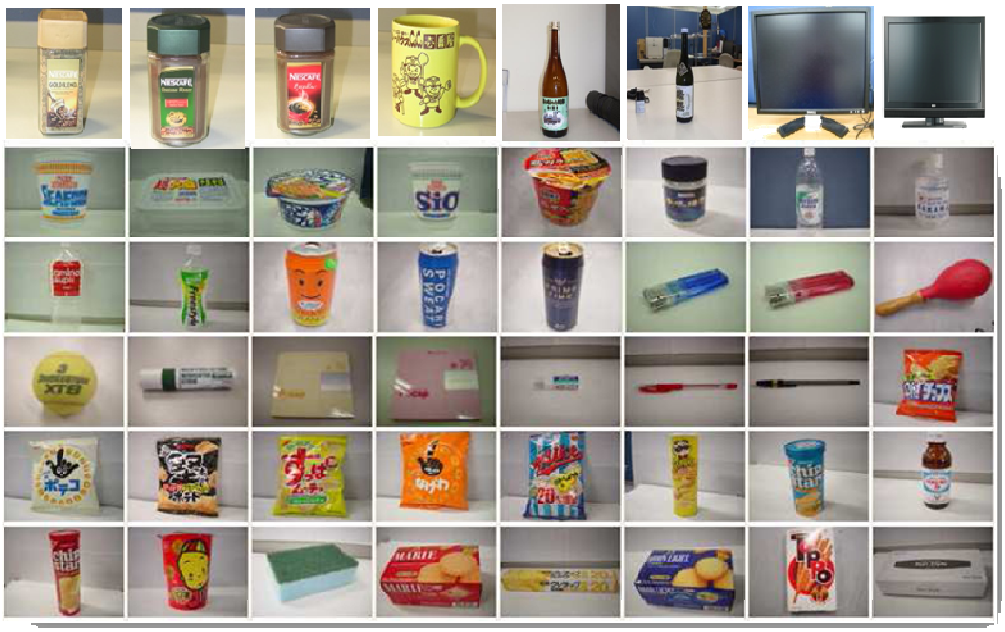
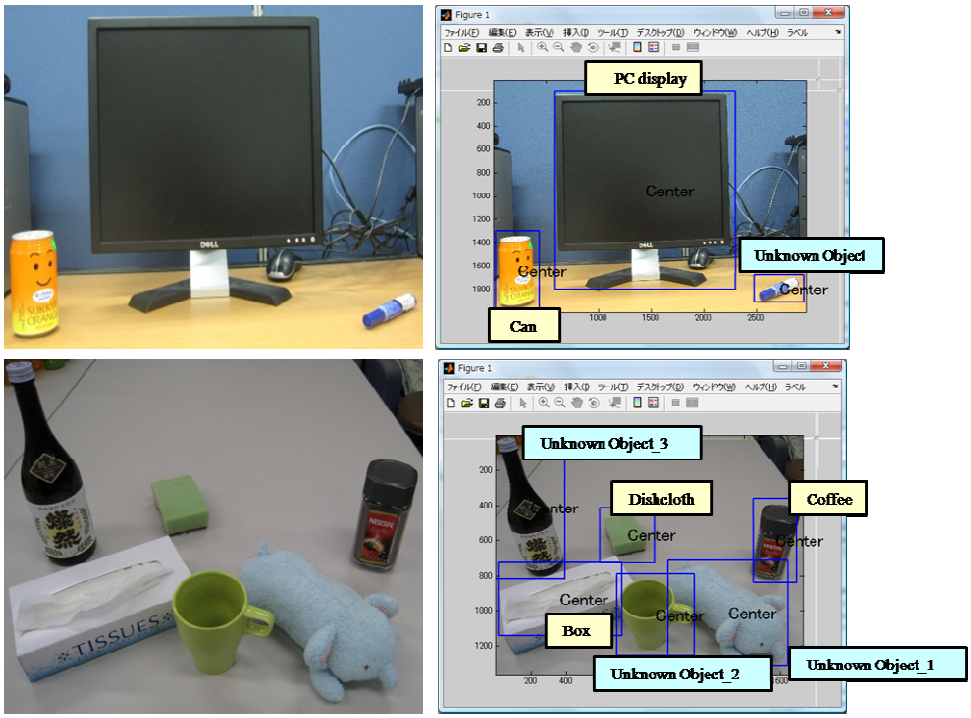


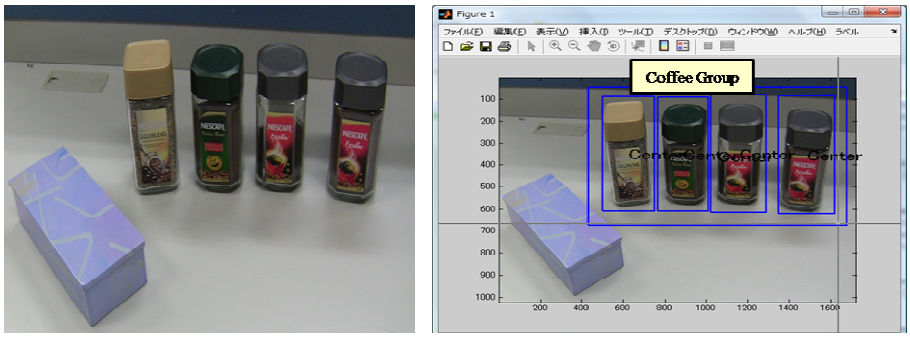
Fig.5.3 Object Examples

(i) (ii)




ject recognition.
objects, which is
nown objects are

Moreover, to gain more precision region, in group case, we generate average bounding box for similar objects. An example of group objects detection is shown in Fig.5.5. Four coffee jars have been detected and bounded, system viewed them as a group not only they have similar attributes which are recorded in the dataset, but also they are next to each other. Therefore, a larger bounding box which covered all of them is drawn by computing average widths and heights.




beling coffees,

### 5.3 Tentative Programme for Viewpoint

The problem in using the terms right/left may arise in relative system in that egocentric spaces depend on the direction the individual is facing [19]. When user and robot are facing to each other, the problem becomes much easier. That is in intrinsic system the regions to the left and right is interchanged by an $180^{o}$ rotation. However, in relative system, not only does interchange left and right, but also front and back. It indicates that for speaker and listener (robot), space to the right of the speaker lies to the left of the listener; verse visa front and back.

Our work concentrates more on such case: speaker and robot stands at the same side, but there are some distances between them. Under this circumstance, front/back is salient both in intrinsic and relative system. But depending on distinct reference object they select, different spatial relations they may obtain which profoundly result in detection. Thus, we attempt to seek an efficient resolution that robot is able to estimate where user is and collects sufficient information on user's viewpoint.

At current stage, we divide a plane into several particular points in a scenario, and build up a database which included four tables to record relationships of locations among objects. Robot is employed to search for the database according to the keywords which is provided by user. Note that user is forced to "tell" robot where he is, for example: "I am standing at Point A".

The database Design is demonstrated in Fig.5.6.

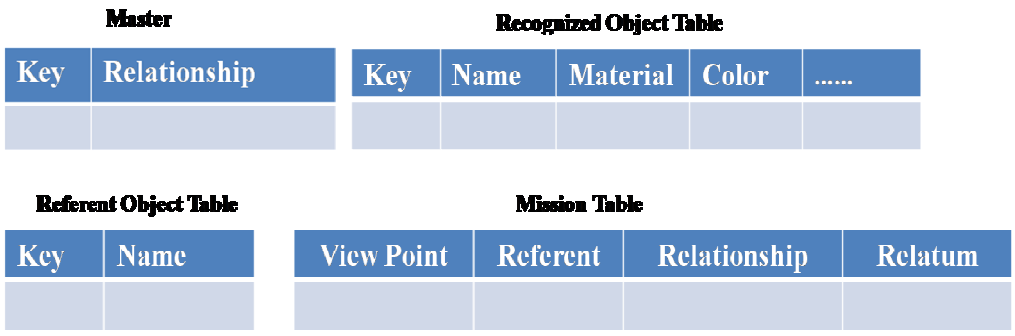



Fig.5.6 Database Design

In master table, prepositions are recorded in **Master Table**, such as front, left, left

front.etc. Known objects are stored in **Recognized Object Table**, with the name, material color and other attributes, while unknown object is in **Referent Object Table**, the item of Name is essential in case it is to be a keyword when interacting with database. Key in each table is an item which is abbreviation for prepositions and object names, for example, front can be defined as rs1 in **Master Table**. **Mission Table** is primarily used to store all the information at every point, for example, if at Point C, a spoon is left to a box, then the data is saved in this way: C, rf1, rs1, ro1.

### 5.4 User Input

We restrict user input based on regular expression. The parser detects user command whether it matches to a series of defined constrains. Then a sequence of keywords is output. The system continues to process or response to user with these. We showed it in Fig.5.7.

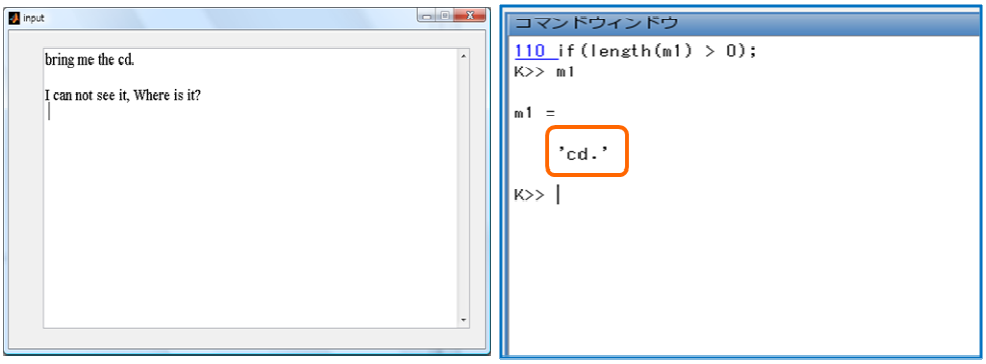



Fig.5.7 User Input. The matched keywords are output to the command window. System searches for the database to check if it is a recognized object and feeds back to user

### 5.5 Pay Attention to More Details

We regard the scenario, which involved in pc display, television and face as typical intrinsic case. For single object, it is perfect enough to describe the target object. However, for multiple objects, determining appropriate reference systems only by these objects is not rigorous. User who doesn't comprehend the knowledge of spatial representation may choose reference object at will, or take the handiest one to express. Considering the 2 cases below (Fig.5.8), in (i), assuming the target is blue mark pen, either can or pc display can be chosen as reference. If user describes in this way: "Mark pen is in front of the pc display", then it is a classical sample utilizing intrinsic system without question. Alternatively, if can is selected, applying intrinsic reference system notwithstanding may bring about ambiguous even failure. Likewise, in (ii), as target object is the staff toy, most of human users are flexible to take the box as reference

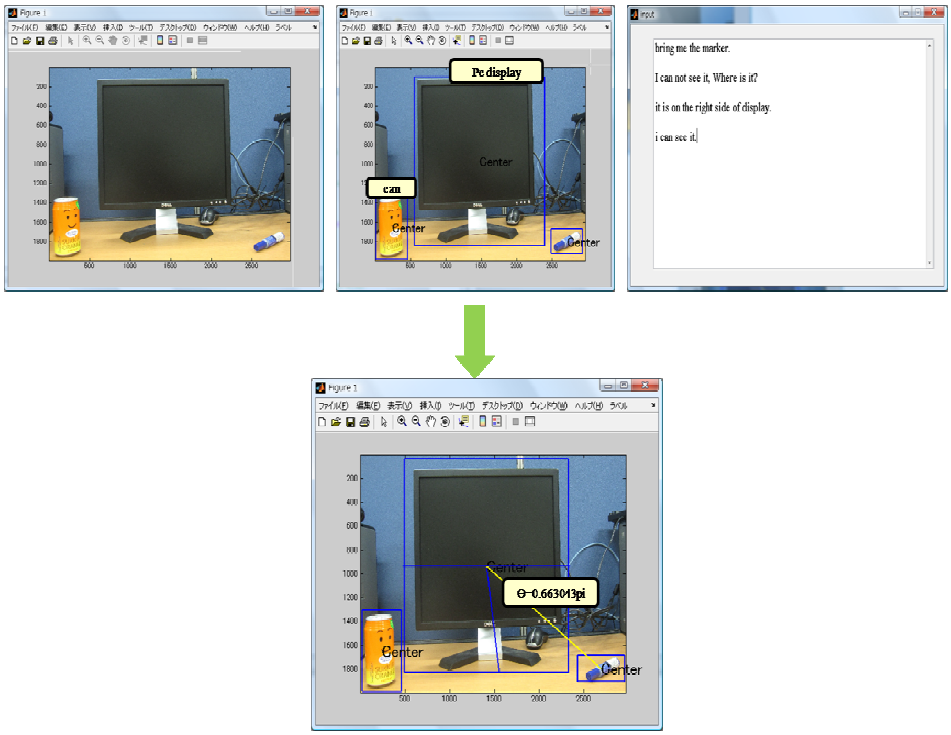
object due to its overwhelming distance advantage. In both of cases, inheriting relative reference system is the best way to fairly picture the target.

Hence, provided that pc **display/TV/face** + user input: "**in front of / at the back of**", intrinsic system will be applied. We don't take the combined prepositions such as **left front** into consideration. In our small area (table, desk.etc) experiments, "**in front of /at the back of**" is sufficient to achieve the anticipation. The tentative idea for large area (building, campus, .etc) is transforming reference system. Since Front-Back is salient (y-axis), applying intrinsic system first, then locating target object by relative system. It will concern viewpoint problem. We will discuss it in the future, too.

### 5.6 Experimental Study

Fig.5.9-Fig.5.11 shows the experimental results. In these three cases, we assume that robot has the same view point with user.

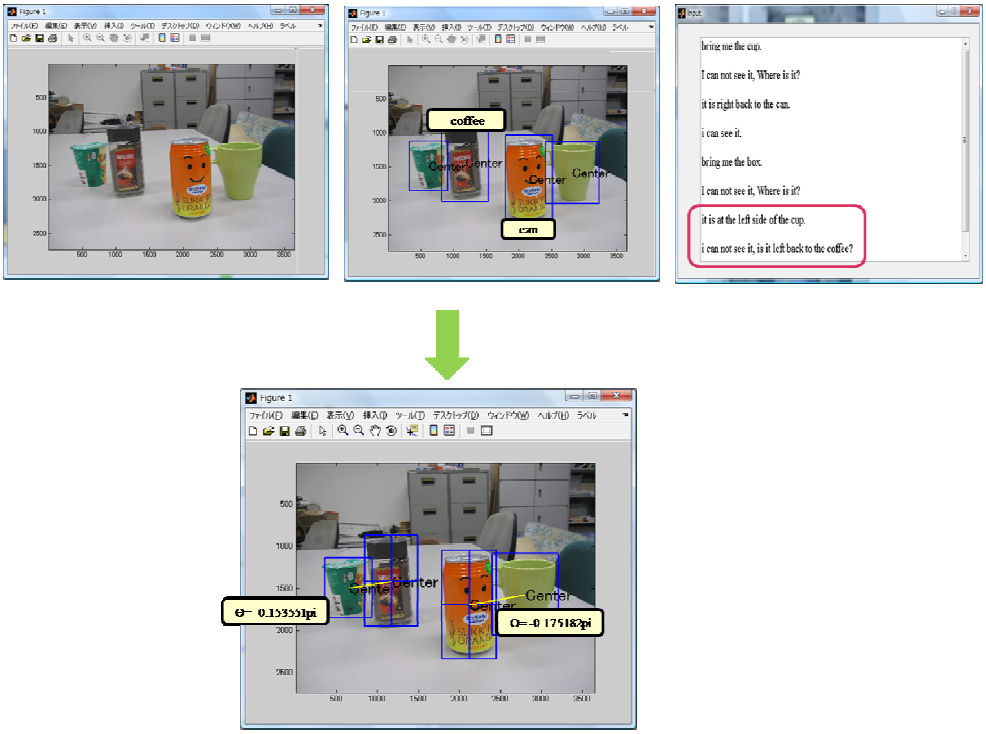
In Fig 5.9, from left to right on the top is: input image; system labeled known and unknown objects respectively; and user-system interaction. Successful detected result is shown at the bottom. Systems also labeled the centers of bounding boxes for convince.




or short

Experiment II in Fig.5.10, system was required to locate the green cup first, then, was the cylinder-shaped box. As the survey2 in Chapter 3, human likely to choose the object which had been used before as a reference. Thus, user provided the information in this way: "**it is at the left side of the cup**." However, the position was not precise to a successful detection, instead of notifying user the fault, it continued to choose an optimal reference object at the left side of the cup based on the reference strategy which is proposed in Section 3 and made a subsequent query (marked in pink box in Fig.5.10). In this case, then the coffee was located as an optimal reference.

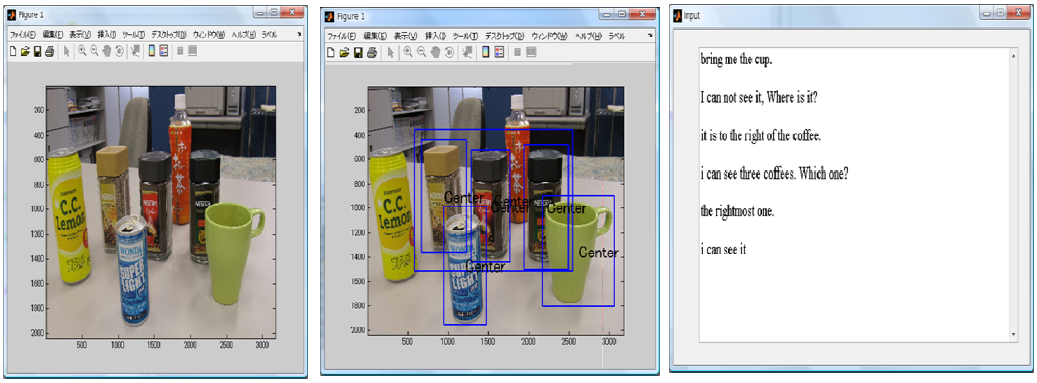
In our current strategy, system responses user as soon as the first unknown object was found. If there are several unknown objects compass the reference object, it won't report to user in turn until user inputs "Yes".

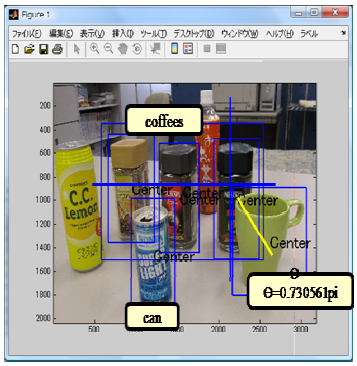


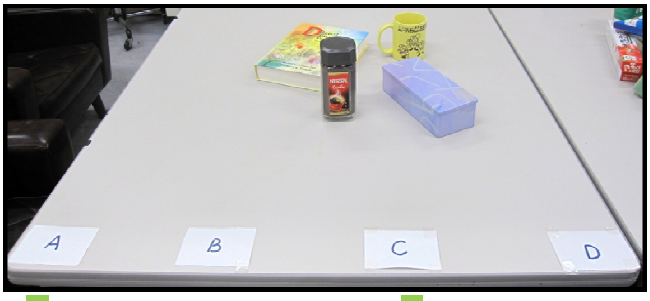

Experiment iii in Fig.5.11 is involved in group objects. As there were 3 coffees, the system marked them as a group, when it caught the keyword "**rightmost**", it commenced locating the cup with the previous keyword "right". We define the leftmost and rightmost by comparing the center coordinate on x axis. More elaborate definition will be left in the future.

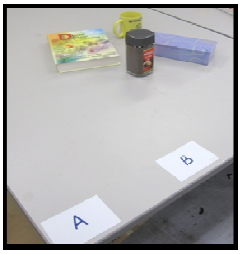
A preliminary instance in simple scenario with only a few objects to experiment viewpoint issue is shown in Fig.5.12 (a). Book and box were to be as unknown objects. Data was registered in DB forehead, listed in Fig.5.13. We assumed that user stood at point A and robot at Point C, the input image was at robot's angle of view.

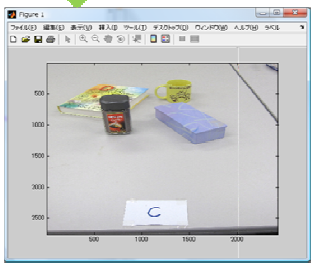
In robot's eye, the cup and the coffee had been recognized. He would like to wait for the user command and picked up the keywords: A, coffee and left searched into DB and located the target. The result was shown in Fig.5.12 (b). It was noticed that user should report to robot which points he stood nearby, for instance: "**I am standing at Point A.**" and robot was also informed where he was at manually.



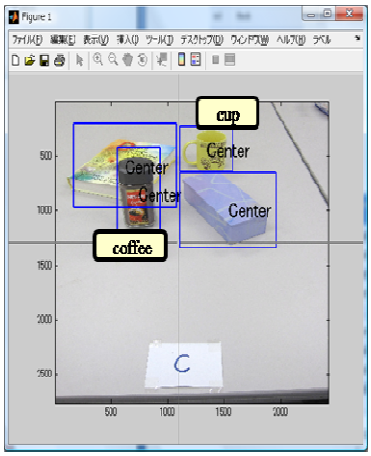



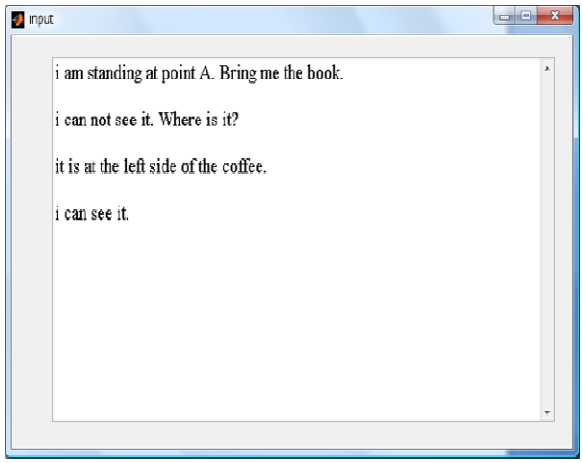




m Speaker's

the scenario




**Master**

| Key | Relationship |
|---|---|
| rs1 | Front |
| rs2 | Back |
| rs3 | Left |
| rs4 | Right |
| rs5 | Left Front |
| rs6 | Left Back |
| rs7 | Right Front |
| rs8 | Right Back |

**Recognized Object Table**

| Key | Name | Material | Color | ..... |
|---|---|---|---|---|
| ro1 | Cup | Iron | Yellow | ..... |
| ro2 | Coffee | Glass | Brown | ..... |

**Referent Object Table**

| Key | Name |
|---|---|
| rf1 | book |
| rf2 | box |

**Mission Table**

| View Point | Referent | Relationship | Relatum |
|---|---|---|---|
| A | rf1 | rs3 | ro1 |
| A | rf1 | rs3 | ro2 |
| A | rf2 | rs1 | ro1 |
| A | rf2 | rs4 | ro2 |
| ..... | ..... | ..... | ..... |
| C | rf1 | rs2 | ro2 |
| C | rf1 | rs3 | ro1 |
| C | rf2 | rs4 | ro2 |
| C | rf2 | rs1 | ro1 |
| ..... | ..... | ..... | ..... |

Fig.5.13 Data for Experiment IV

It is only applicable to limited simple sceneries by the method. And one of the most notable defects is it takes quantitative time to inquire DB. Besides, the system is supposed to be elastic to estimate the location where user and robot take, not only limited fixed points. We attend to improve our strategy to a model which is based on fuzzy rules. The issue will be left for future work.

# CHAPTER 6
# CONCLUSION AND PROSPECTION

## 6.1 Conclusions

As a final point, we have proposed a spatial relation model that is applied for the robot to interpret user's spatial relation descriptions. The model represents the target object position by its spatial relation with the relatum (reference object) in the reference system. The relatum can be a group of objects. We have considered two reference systems: the relative reference system and the intrinsic reference system. The robot asks the user the spatial relationship of the target object and some known objects automatically recognized. From the user's utterances, the robot determines the relatum and selects the appropriate reference system to detect the target object [20].

## 6.2 Limitations of this Research

Experimental results show promising results for our interactive approach. Nevertheless, we need to discuss two issues. In the current implementation, the robot system takes the initiative and asks the user questions about the target object. This may not seem to be a human-centered design. However, since the purpose of the system is limited to object recognition, this robot-initiative interaction can be acceptable way for users. The important point in our system is that the system can understand natural expressions in the user's answers. Thus, the user can make smooth interaction with the system even though in the robot-initiative way. The second issue is our current assumption that the robot vision can separate at least a part of target object from the background. In other words, the robot cannot detect an object through interaction if any part of the object is not separated from the background or all parts of the object are merged into other objects. This is related to an old but fundamental problem in object recognition: the relation between segmentation and recognition. In old conventional object recognition methods, segmentation comes first. Then, each segmented region is recognized. Segmentation errors are fatal in such methods. Recent object recognition methods [21] directly try to recognize objects. Our autonomous object recognition method [15] also adopts this approach. However, some segmentation is necessary to make interaction with the user. We think that proper interaction with the user enables the system to recover from segmentation errors. This is left for future work and we will also

consider other attributes such as object pose for intrinsic system.

### 6.3 Prospection

Humans employ a variety of para-linguistic social cues (facial displays, gestures, etc.) to regulate the flow of dialogue [22]. Merging these clues with our interactive method of reference resolution may produce better result. The determination of properties of image regions and spatial relationships among regions is critical for higher level vision processes involved in tasks [23] Now that humans may have an intuitive understanding of words, such concepts defy precise definitions, and it is our belief that they are best modeled by fuzzy set will yield realistic result.

Moreover, it is prospective to include “Situation awareness” into a service robot system. Situation awareness has been formally defined as in [24]. Limiting the object search database is possible if a robot can interpret a situation it exists in. Thus, mapping a situation to a pre-calibrated object database significantly reduces search domain.

## REFERENCES


*[1] T. Tenbrink, M. Reinhard, "Group-based Spatial Reference in Linguistic Human-Robot Interaction," Spatial Cognition and Computation, Volume 6, Issue1, pp.63-64, 2006.*

*[2] Reinhard Moratz, Thora Tenbrink, John Bateman, Kerstin Fischer, "Spatial Knowledge Representation for Human-Robot Interaction," Spatial Cognition III, Volume 2685, Springer Berlin / Heidelberg, 2003*

*[3] Daniel Hernandez, "Qualitative Representation of Spatial Knowledge," Springer-Verlag, pp.2-4, 1994.*

*[4] R. Jackendff, "Languages of the Mind," The MIT Press, 1992.*

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

*[6] Paul Bloom et al.,"Space and Language, Language and Space," the MIT Press, 1999.*

*[7] Pattabhiraman, Thiyagarajasarma,"Aspects of salience in natural language generation," Vancouver, B.C, Simon Fraser University, Ph.D. Thesis, 1992.*

*[8] R.Moratz, K.Fischer and T.Tenbrink,"Cognitive Modeling of Spatial Reference for Human-Robot Interaction,"In Intl, Journal on Artificial Intelligence Tools, vol, 10, no.4, pp.589-611, 2001.*

*[9] B.Tversky, P.Lee and S.Mainwaring, "Why Do Speakers Mix Perspective?," Spatial Cognition and Computation, Volume.1, pp.399-412, 1999.*

*[10] S.C. Levinson, "Frames of reference and Molyneux's question: Crosslinguistic evidence, "in Language and Space, MIT Press, 1999, pp. 109–170, 1999.*

*[11] T. Winograd, Understanding Natural Language, Academic Press, New York, USA, 1972.*

*[12] T. Kawaji, K. Okada, M. Inaba, and H. Inoue, "Human robot interaction through integrating visual auditory information with relaxation method," Proc. IEEE Int. Conf. Multisensor Fusion on Integration for Intelligent Systems, pp.323–328, 2003.*

*[13] P. McGuire, J. Fritsch, J.J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human machine communication for instruction robot grasping tasks," in Proc. IROS2002, pp.1082–1089, 2002.*

*[14] M. Takizawa, Y. Makihara, N. Shimada, J. Miura, and Y. Shirai, "A service robot with interactive vision- objects recognition using dialog with user," Proc. First Int. Workshop Language Understanding and Agents for Real World Interaction, 2003.*

*[15] A. Mansur, K. Sakata, Y. Kobayashi, and Y. Kuno, "Human robot interaction through simple expressions for object recognition," in Proc. 17th IEEE RO-MAN, pp. 647–652, 2008.*

*[16] A. Mansur and Y. Kuno, "Specific and class object recognition for service robots through autonomous and interactive methods," IEICE Trans. Information and Systems, vol.E91-D, no.6, pp.1793–1803, 2008.*

*[17] Hagen, C.H., "Journal of Psychology," 1949.*

*[18] Dipankar.DAS, Yoshinori Kobayashi, Yoshinori Kuno, "A Hybrid Model for Multiple Object Category Detection and Localization," MVA2009 IAP Conference on Machine Vision Applications, pp.431.*

*[19] Paul Bloom et al., "Spatial Perspective in Descriptions," Language and Space, 1999.*

*[20] L.Cao, Y.Kobayashi, and Y.Kuno, "Spatial relation model for object recognition in human-robot interaction," Proc International conference on Intelligent Computing, lncs 5754, Springer, pp.574-584, 2009.*

*[21] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, (Eds.), "Toward Category-LevelObject Recognition," Lecture Notes in Computer Science, LNCS4170, Springer, 2006.*

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

*[23] Raghu Krishnapuram, James M.Keller, and Yibing Ma, "Quantitative Analysis of Properties and Spatial Relations of Fuzzy Image Regions," IEEE Transactions on Fuzzy Systems, Vol.1.No.3, pp.222. 1993.*

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

**ACKNOWLEDGMENT**

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 Japanese and International 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 throughout the period of Master's degree. All friends deserve thanks and gratefulness as well.