

# Comprehending Spatial Relations for Interactive Object Recognition

(対話物体認識のための空間位置関係理解)

A dissertation

Submitted to the Department of Computer Science and Engineering

And the Committee of Graduate Studies

Of Saitama University

For the Degree of Doctor of Philosophy

2013 年 9 月

Lu Cao

Supervisor: Professor Yoshinori Kuno

# Abstract

Human can effortlessly express their spatial experience and talk about *where* objects are located in relation some underlying objects. Since it is impossible for us to learn all the objects, such information has been critical to explore the visual world. Intuitively, if we know where the objects are, recognizing them will become easier. In computer vision, in order to mimic human's ability, an important and open problem is to endow robotic systems the ability to comprehend spatial relations as human does. This is somewhat like a school child does when learning to write a descriptive sentence, such as **the CD is to the left of the book**.

The primary goal of this work is to design and demonstrate spatial recognition methods to bridge the gap between visual information and human cognition. Towards this goal, we treat spatial relations as a kind feature as well as other visual features, such as *color*, *size*, etc and have developed computational templates to represent spatial relations. We propose a novel model to encode linguistic spatial expressions.

We first investigated how humans manipulate space by the action of natural language and classified basic class of relation. We then extracted the observations from cognitive systems to computer vision applications. We propose templates for recognizing spatial relation, translating linguistic expression into visual information, representing spatial terms in an angular fashion. The templates have been tested over 720 scenarios where 1-3 *unknown* objects within.

Comprehending spatial relations are beyond simply distinguishing them. It is noticeable that spatial relation needs a pair of objects. In determination of different class of relation, the underlying objects which are named as reference objects play a decisive role. Concretely, objects like *humans*, *animals* and *computer displays* are somewhat different with objects like *balls*, *boxes*, *cups* in that they have intrinsic *front* side. The former's *front* is independent from interlocutors' viewpoint whereas the latter's is not. It turns out the *front* orientation adjacent to the frontal-side of those objects are transformed accordingly if they are rotated from the frontal view. We then focus on introducing an estimation model for those objects, from estimating poses

transformations, to adjusting *intrinsic-front* orientation. The first step studies one prominent type of pose variation given viewpoint transformation in supervised fashion. Naïve Bayesian classifier is followed for prediction. The estimator performs highly competitively with the state of the arts on the ETH-80 database, and an everyday-object database that we collected on our own.

The models profit from an interactive interface, which is developed to understand some simple English words and grammatical structures. The ability makes our models are closer to the way of human-human interaction. Finally, we conduct experiments integrally within the system, which consists of an object detector, a spatial recognition model, a pose estimator and a user interface. The goal is recognizing *unknown* objects via comprehending spatial relations by interactive means. The simple yet effective models outperform in recognition tasks in the author's database.

To my family

# Acknowledgment

This thesis would not be accomplished without the support, guidance and encouragement from so many people around me.

In the five and half years I have spent at CV lab, it is a great fortune in my life. I am very grateful to have worked with quite a few wonderful people during this period. First with full respect, I would like to especially thank my advisor, Professor Yoshinori Kuno, for being a passionate teacher, a patient trainer, and an inspirational thinker. I highly appreciate his dedication to train me to be a researcher in every aspect, for improving presentation and writing skills. In addition, I met some unexpected difficulties when I first came to Japan; Prof. Kuno gave me countless help and invaluable advices in life.

I would like to thank Prof. Yoshinori Kobayashi, for his help and support during my research. Many thanks go to Dr. Dipankar Das for the insightful research discussions and advices. My research has also been greatly benefited from the collaboration with him.

I am also thankful to Dr. Zhongkui Wang and Dr. Kaiyuen Cheong, from whom I learned incredible advices in research, career and personal life.

I would like to sincerely thank the people in CV lab, past and present, for their kindness and friendship. Thank you for the parties and delicious Japanese food, which make my research journey exciting and enjoyable.

I would like to thank my parents for their unconditional love and support. My appreciation to them is beyond what I can express in words. Last but not least, especially want to thank my husband, Xi. His company, understanding and love make me a better person.

# Contents

Abstract .....	2
Acknowledgment .....	5
List of Figures .....	10
List of Tables.....	14
Chapter 1 .....	15
Introduction .....	15
1.1 Spatial Relations in Visual Recognition.....	15
1.2 Related Work .....	19
1.2.1 Spatial Comprehension in psychology, linguistics, and philosophy .....	20
1.2.2 Learning Spatial Relations for Robotic Systems.....	23
Chapter 2 .....	29
Towards Spatial Comprehension .....	29
2.1 Understanding Spatial Knowledge.....	29
2.1 .1 Terminology .....	29
2.1. 2 Classification of Frames of Reference .....	30
2.1.3 Spatial Templates and Their Acceptance Regions .....	35

2.2 Computational Model for Human Spatial Linguistic Expressions .....	36
2.2.1 The 2d Projective Model of Intrinsic and Relative Frames of Reference .....	37
2.2.2 Modifications: The 3-D Computational Model.....	43
2.2.3 The Model of Group-based Frame of Reference.....	49
2.3 Conclusion.....	57
Chapter 3 .....	58
Pose Estimation .....	58
3.1 Instruction .....	58
3.2 Related Work .....	59
3.3 The Model .....	61
3.3.1 Building Key-Pose Structure.....	61
3.3.2 Image Feature.....	62
3.3.4 Adjusting Front Orientation .....	65
3.4 Experiment .....	66
3.4.1 Pose Estimation Result.....	66
3.4.2 Adjusting front orientation .....	69
3.4.3 Spatial Recognition Experiment.....	70

3.5 Conclusion.....	72
Chapter 4.....	73
Constructing the Database.....	73
4.1 Instruction .....	73
4.2 Relative work .....	74
4.3 Collecting Candidate Objects.....	74
4.3.1 Collecting Candidate Objects for Visual Recognition .....	75
4.3.2 Designing Scenarios for Spatial Recognition.....	77
4.4 Conclusion.....	80
Chapter 5.....	81
Interactive Object Recognition.....	81
5.1 Integral System Overview.....	81
5.2 The role of Natural Language .....	81
5.3 Experiment 1: close linguistic form .....	83
5.4 Experiment 2: Comparison with the original model .....	89
5.5 Failure Case Study .....	91
Chapter 6.....	93



Conclusion.....	93
Related Publications .....	95
Bibliography.....	97

# List of Figures

Fig.1. 1: 2 scenarios from home object dataset. Spatial relations can simply distinguish the target object from the others. ....	16
Fig.1. 2: The integral system.....	18
Fig.1. 3: Spatial representations take as input information from vision, audition, and the haptic system, and provide information to the motor system and language and vice versa. ....	20
Fig.1. 4: Description in Intrinsic, Relative and Absolute frames of reference .....	23
Fig.1. 5: The robot system, two boxes and a barrel.....	26
Fig.1. 6 Configurations Experiment 1 .....	27
Fig.1. 7 Configurations Experiment 2 .....	28
Fig.2. 1: “Front” Examples in Intrinsic use.....	32
Fig.2. 2: “Left” Examples in Relative use.....	33
Fig.2. 3: 2 situations in the use of group-based frame of reference: (a) internal use; (b) external use .....	34
Fig.2. 4: (a) Good, acceptable, and bad regions for ‘front’ orientation in Logan’s template; (b) the acceptable and bad regions for ‘front’ orientation in our template. Here we merge the good and acceptable regions to form a large one. ....	36

Fig.2. 5: Templates in intrinsic and relative frames of reference, and acceptance regions .....	38
Fig.2. 6: Experiment results in single referent scenarios.....	39
Fig.2. 7: Experiment results in two referents scenarios.....	40
Fig.2. 8: Failure case study: interaction between angle and distance.....	41
Fig.2. 9: Vague vs. non-vague situations: the vagueness is occurred when a referent falls into two acceptance regions.....	42
Fig.2. 10: Distribution of the left, front and left front utterances in two scenarios in the experiment.....	42
Fig.2. 11: Modified <i>intrinsic, relative</i> relation templates, and enlarged acceptance regions for each orientation .....	44
Fig.2. 12: Compound expressional regions .....	45
Fig.2. 13: Canonical spatial relation experiment result. Row 1: the scenarios of 1 referent vs. 1 relatum. Row 2: the scenarios of 1 referent vs. 1 or 2 relatum(s). Row 3: the scenarios of 2 or 3 referents vs. 2 relatums. The referent and chose relatum is linked by → .....	47
Fig.2. 14: Compound spatial relation experiment result. The referent and chose relatum is linked by → .....	49
Fig.2. 15: Text Input box.....	53
Fig.2. 16: Grouping results by k-means algorithm.....	54
Fig.2. 17: Recognition result.....	55

Fig.2. 18: Deficiency of the resembling k-means algorithm: different results obtained .....	56
Fig.3. 1: Azimuth and zenith angle $\alpha$ , $\beta$ representation.....	61
Fig.3. 2: A car instance example. The key pose structure is built from 7 poses from the same viewpoint.....	62
Fig.3. 3: A PHOG representation of a car instance at pose $0^\circ$ . The ROI region is labeled with blue bounding box. With using a three-lever PHOG feature, a shape histogram of a ROI is a concatenation of histogram described at each level. ....	64
Fig.3. 4: axis rotation.....	66
Fig.3. 5: Example results of pose estimation. The last column of each category is the false estimation of category.....	67
Fig.3. 6: precision recall curves of 8 categories .....	68
Fig.3. 7: “Front” orientation adjustment .....	70
Fig.3. 8: Experiment results for recognizing spatial relation when adjusting “front” orientation .....	71
Fig.4. 1: A semantic structure of the container hierarchy .....	76
Fig.4. 2: Semantic structure in current version of the database.....	78
Fig.4. 3: A snapshot of the container abstract class containing two branches: the top row is from the can category; the bottom row is from the dispenser category. ....	79

Fig.4. 4: Designing scenarios with adequate relatum. The larger and more stable is usually treated as a relatum. If someone exchanges *referent* and *relatum*, saying the jar is to the left of the candle, it produces an odd-sounding result..... 79

Fig.5. 1: transcript without scene description given..... 85

Fig.5. 2: transcript with scene description given..... 86

Fig.5. 3: Instructional strategy in internal case in the use of group-based frame of reference.. 87

Fig.5. 4: Instructional strategy in internal case in the use of group-based frame of reference.. 88

Fig.5. 5: 7 example scenarios for comparative experiments ..... 90

Fig.5. 6: Failure case study: occlusion. (a) referent obscures the relatum; (b) the referent is hidden at the back side of the relatum..... 92

## List of Tables

Table 1: Prepositions in English.....	29
Table 2: Comparison with the two models on the author's database .....	69
Table 3: Grammar, symbols and linguistic form used in the interaction .....	82
Table 4: the results of comparison the approach with previous model .....	91

# Chapter 1

## Introduction

*Talking about space and understanding corresponding spatial relations are the fundamentals for humans. These abilities are so unique that can set humans apart from other species. In visual recognition, humans often use spatial relations to depict their visual surrounding environments and localize objects within. Intuitively, if we know where the objects are, recognizing them should be easier. Particularly, this strategy is extremely useful in the scenes in which the objects that we intend to reach for are not familiar with. In order to develop human-like robotic vision systems, it is necessary to endowing the machines with the same ability. This chapter will first describe the thesis motivations, objectives and challenges. An outline of the thesis is then given.*

### 1.1 Spatial Relations in Visual Recognition

Object recognition is an important yet challenging research direction. Over the years, much of the progress has been paid to develop algorithms for modeling and learning objects. Nearly all of these approaches are based on some kind of visual image features, e.g., *color, shape*. With the help of more and more sophisticated machine learning models, these features have achieved good success in high-level recognition tasks. But as the task level becomes higher and higher, the limitation of those features becomes more obvious. For example, they are highly influenced by illumination, scale, and object diversities. A slight variance in color can make an object thoroughly different. Different object instances can cause strong variations even they belong to the same category. Considering the numerous types of objects exist in our visual world and how fast the number grows every day. How the traditional object recognition algorithms going to be scalable to cope with these defects?

Consider the following scenarios in Fig. 1.1. In (a), a scenario contains several objects. A robotic system has already learnt a juicy box, and a cup noodle. The target object is labeled in yellow bounding box but we don't know what it is. How to instruct the robotic system successfully locate the target object? In scenario (b), 4 jars have been recognized by the robotic system. When the robotic system is supposed to locate one of them, what is the best way to distinguish it from the others? Obviously, our visual experience and intuition suggest a straightforward way that is, addressing *where* the object is. Intuitively, if we know *where* the objects are, recognizing them should be easier. One can say **the object is in front of the juice** in scene (a), and **the leftmost jar** in scene (b). In order to develop human-like robotic vision systems, it is necessary to endowing the machines with the same ability.



Fig.1. 1: 2 scenarios from home object dataset. Spatial relations can simply distinguish the target object from the others.

However, identifying spatial relations remains difficult for computers. There are two main reasons. First, duplicating the process from human perception is complicate. Somehow humans can effectively deal with this complexity—they can map natural language description onto nonlinguistic spatial representation. A listener



who perceives a description, must (1) find the reference and target objects in the description; (2) impose a proper frame of reference on the reference object; (3) identify the spatial relation that the speakers proved; (4) choose the target object, which best represents the relation; and (5) produce an answer. In terms of recognition tasks by verbal means, speakers must explicitly encode spatial relations into linguistic expressions, namely, people must have a non-verbal spatial representation of a perceived configuration so that they can map this onto a verbal one. The second difficulty is how to translate linguistic expression to visual information to non-intelligent computers. Since computers are good at processing numbers, how to present spatial terms in a numerical coordinate representation? For example, how *left* is *left*?

Validating spatial relation associates with a pair of objects, which is a significant defect is so that it cannot be independently used yet. Even though, we still believe identifying spatial relations is important and advantageous for 3 reasons: (1) spatial relations are important in many regions of cognitive science, including linguistics, philosophy, anthropology and psychology. In order to mimic humans' ability, it is propitious time to introduce the conception into computer vision field; (2) Spatial relation is stable. It implies a two-fold meaning. One is it is less influenced by illumination and scale changes than other kinds of visual features. On the other hand, it is independent of diverse object categories; and (3) Spatial relation is a unique form—relatively, only one relation exists in pair of objects.

The focus of this work, aims to address the very challenge of the spatial understanding, a subordinate territory in previous vision research. It poses the question of the how to translate natural language expressions to visual information, which is also the essence of the task. The ultimate goal is recognizing generic objects via spatial relations within an image in an interactive fashion. Despite the concerted effort in the last decade, the objective yet, remains challenging and unsolved well. Although numerical work for robot navigation, such as, instructing robotic systems moving towards to the certain objects, no satisfactory methods exist that work for localization tasks. In this work, we provide a novel framework. The sketch of our idea is following. An untrained image is first served by an object detector, where pre-learned object is detected. If target object is still undetected, human users manually annotate the object and instruct the system to recognize it via spatial

relations. User will benefit from a user interface, which requires typing simple instruction. The interface can systematically analyze semantic components, separating the target object as well as that of the reference object, and the relation between both. This is an interactive process that user continually gather information whether the system understands their intention. The integral framework illustrated in Fig. 1.2 includes a user interface, an object detector and object spatial recognition model. After processed by an object detector, a **dispenser** is highlighted. But what we want is the **toothpaste**, which still unknown to the robotic system. Then the system then switches to an interactive mode. User first inquires the system whether can see the target object, the **toothpaste**. If the answer is negative, then user prompts the positional information such as “**The toothpaste is to the right of the bottle**”. As long as the system detects a target object that best represents the relation provided by the user, it reports and waits for confirmation. If the answer is positive, the process is ended successfully with given a command, e.g. “**bring the toothpaste for me**”. Otherwise, it continues till the number of processing exceeds the total number of annotated objects in the image.

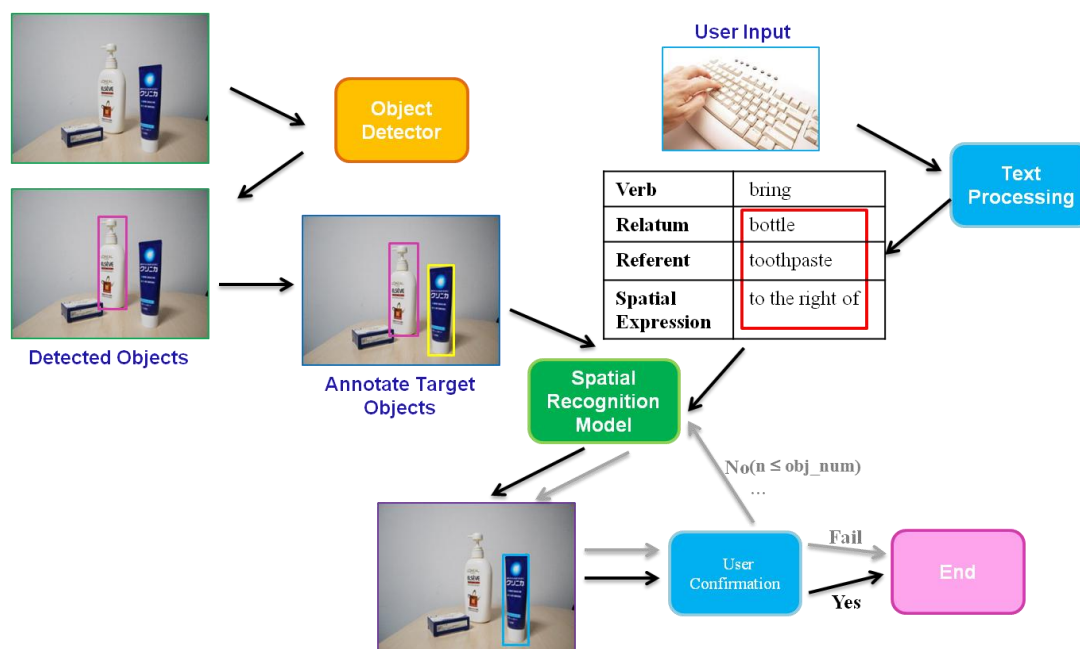


Fig.1. 2: The integral system

To validate our method, we have built a novel database specifically tailored to the task of spatial relation recognition. The dataset is twofold: **132** objects, in total of **5,124** images from **30** categories for the purpose of object recognition and in the total of **720** scene images for object localization task. The objects we grouped obeying the lexical meaning in English, which describe the meaningful concept in cognition and the perceptual apparatus,

We will show in our experiments that our model is capable of identifying fundamental spatial relations, e.g. *front*. To summarize, we highlight here the main contributions of our work.

- From a plethora of work in cognitive science, we make sense of how human manipulate space and conclude some perceptual limitations on talking about space.
- We then bring the conceptions into computer vision field and propose a model for identifying and comprehending some fundamental English spatial expressions
- Our work is designed as a visual-interactive fashion that reference objects can be visually detected which target objects are recognized through natural language. This scheme leverages the gap between language and vision.
- To our knowledge, it is the first approach to achieve visual recognition objectives via comprehending spatial relation.
- Since there is no publicly database available, we provide a new database specially tailored for our experiment design.

## **1.2 Related Work**

Spatial information can be derived from vision, audition, and haptic. That is, this representation is not exclusively visual or haptic or aural, but neatly incorporating with spatial. Researchers have been firmly believed that spatial representations can be translatable into a form of representation specific to the motor system that instinctively guides human behavior and vice versa. For example, we can touch what we see, find underlying targets according to what we hear see, and avoid obstacles as we navigate through space, see Fig.

1.3. In this section, we first review the related work in interdisciplinary domain, combining psychology, linguistics, and philosophy. The psychologist's contribution is a concern for how spatial relations are apprehended, a concern for the interaction of representations and processes underlying an individual's comprehension of spatial relations. These theories so are important to use that the theories and models they developed is the benchmark of our work. Then we focus on the robotic application in the AI field.

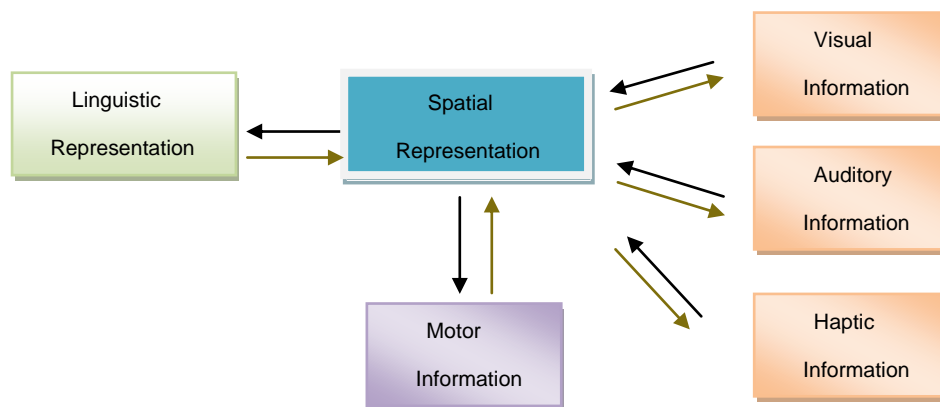


Fig.1. 3: Spatial representations take as input information from vision, audition, and the haptic system, and provide information to the motor system and language and vice versa.

## 1.2.1 Spatial Comprehension in psychology, linguistics, and philosophy

### 1.2.1.1 Spatial Representation

In the last decades, researchers are intrigued by the language that compatible with constraints on nonlinguistic spatial cognition. Specifically, there must be some level of visual representation that can be accessed by our linguistic system.

A canonical example of spatial language is the encoding of spatial relations between pairs of objects or object parts. Talmy[96] has demonstrated that spatial prepositions such as *on*, *above*, etc intuitively code only a schematic relationship, while disregarding many spatial properties, e.g. shape, color, size, orthogonal position. [7, 15, 80] suggested that objects are represented as structural descriptions, which are composed of spatial relations among parts. Some of these relations are defined according to their corresponding spatial prepositions; for example, *front*, *below*. This assumption gradually had been made more explicit. [41] has implemented a neural network for shape recognition that has nodes allocated specifically to the relations such as *left of*, *right of*. In such instances, there is no attempt to define the actual spatial configurations these relations encompass; rather such models of visual representation rely on our intuitive and linguistic conceptualization. Therefore, while spatial relations are a basic and essential element of several theories of object representation, they have been characterized mainly in terms of their linguistic counterparts and without direct evidence about their organization. Although a direct link between closed-class spatial forms, spatial prepositions in English, and visual representations may indeed exist, this connection is tenuous in that it has not been empirically validated, nor do there exist well-specified models of the underlying structure of particular spatial relations.

Meanwhile, connections between spatial prepositions and visual information have been studied predominantly through spatial language, but not both. For example, there is a strong tradition of implicitly addressing the nature of spatial representations through linguistic descriptions of spatial layouts and the study of cognitive maps [59, 60]. More relevant to our study, [38] explored the representation of the horizontal space surrounding oneself in terms of the spatial descriptors *front*, *back*, *left*, and *right*. When they required subjects pointing to the outer boundaries of different category regions in the experiments, they found that recall accuracy for object position relative to the subject varied between regions: *front* yielding the highest accuracy for object position and *back* yielding the poorest performance.

### 1.2.2.2 Classification of Frame of Reference Classification

The notion of frame of reference is crucial to the study of spatial cognition. To describe *where* a target object is with respect to a reference object, we need some way of specifying underlying coordinates systems on the space. [47] compared two systems of frame of reference: *deictic* and *intrinsic*, and specified how these frames of reference differed from the formal and the perceptual point of view, for example, analyzing the use of *left* versus *right*, *in front of* versus *behind*. [10, 83] ranged over the philosophical and psychological literature, and concluded that frames of reference come down to the selection of reference objects. Kant [42] argued elegantly that the human body frame is the source of our basic intuitions about the nature of space, as exemplified in a description such as **the glasses are to the right of the telephone**. [6, 18, 36, 37, 48, 49, 50, 61, 69, 72, 73, 84, 85, 102, 106] elaborated 3 frames of reference: *intrinsic*, *relative* and *absolute*, which can be thought of as different strategies for specifying the spatial relationship between the target and the reference objects. Fig. 1.4 illustrates the 3 frames of reference. In fact, the frequency and range of application of these frames of reference differ across languages. English speakers mainly use 2 different frames of reference to describe spatial relationships in table-top space: intrinsic frame of reference or relative frame of reference. In the use of *intrinsic* frame of reference, they say **the fork is beside the spoon** [58]. Or In the use of *relative* frame of reference, they say **the fork is to the left of the spoon**. They do not say **the fork is to the north of the spoon**. In Sect. 2.xx, we introduce the 2 frames or reference in more details.



Intrinsic: The fork is at the nose of the spoon  
Relative: The fork is to the left of the spoon  
Absolute: The for is to the north of the spoon

Fig.1. 4: Description in Intrinsic, Relative and Absolute frames of reference

## 1.2.2 Learning Spatial Relations for Robotic Systems

### 1.2.2.1 Qualitative Spatial Reference Learning

To specify positional information in human-robot interaction, qualitative spatial reference, a novel and powerful strategy, serves as a necessary bridge between the metric knowledge required by robot, and more abstract concept for natural language utterances. Increasingly sophisticated approaches to the computation of spatial relations have been developed in the last couple of decades [1, 2, 24, 68, 70, 76, 81]. In [30] the CSR-3D system, a model for the computational of topological and projective relations and 3D space, is presented. The model is enhanced to include composite spatial relations, such as *to the left of and behind* or *to the left of and near* which are very common used in German [31] and proved for cognitive plausibility [32, 33].

### 1.2.2.2 The Robotic Systems

In computer vision, little work has been done. As part of the VITRA project in Saarbrücken, Gapp [30, 31] developed a computational model for basic meaning of spatial relations. His model is used to generate linguistic spatial expressions based on simulated city scenes. For graded applicability regions he uses spline functions. [90, 91] developed a model which are often inspired by force-field models from physics. In their design, situations involving more complex object configurations (like walls, bent objects close to each other). Fuhr and his colleagues as part of the SFB 360 project “Situating Artificial Communicators” [29] developed the system, integrating speech and visual input. They used 3D acceptance *volumes* in their model. By introducing an intermediate representation layer independent of actual reference frames their approach can handle image sequences efficiently. Closer to our application is the system developed by Moratz [62-67, 98]. They have developed a computational framework to localize interested objects using geometry projection, focusing on configurations of simple, convex objects. They instruct the Pioneer robot to move towards particular objects pointed at by the experimenter. The test scene is designed simple in which two or three objects are placed on the floor together with the robot as shown in Fig. 1.5. Fig. 1.6 and Fig. 1.7 list in total of 13 controlled configurations in 2 experiments. 2 examples of a typical interaction situation are:



Dialog transcript in Experiment 1

USER: go back

ROBOT: *I don't understand*

USER: straight ahead

USER: start moving

ROBOT: *I don't understand*

USER: go

USER: roll

ROBOT: *I don't understand*

**(Prolonged failure - re-start: new configuration)**

USER to the left

ROBOT: *I don't understand*

USER: turn around to the left

ROBOT: *I don't understand*

USER: move to the left box

ROBOT: *successful robot movement to the goal*

Dialog transcript in Experiment 2

ROBOT: *I can see a barrel and two boxes. Where shall I go?*

USER: to the box at the outer left

ROBOT: *I don't understand the word "außen"*

USER: left

ROBOT: *please reformulate*

USER to the left box

ROBOT: *I start going*

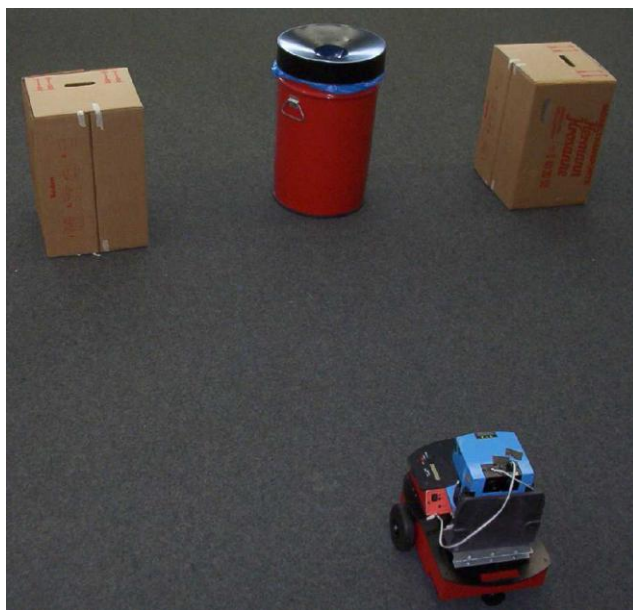


Fig.1. 5: The robot system, two boxes and a barrel



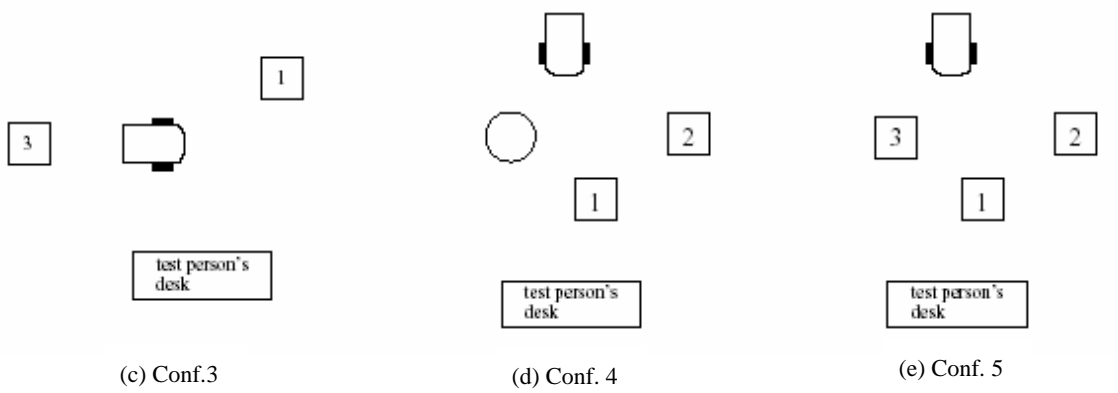
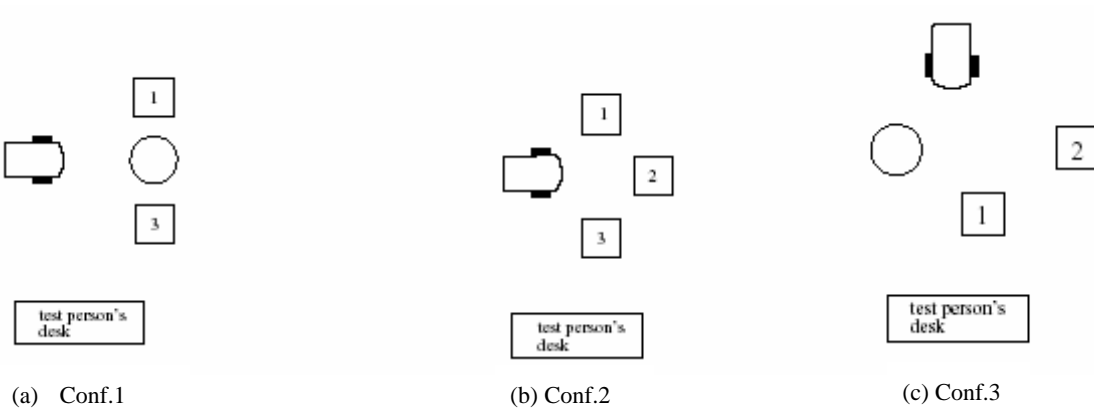


Fig.1. 6 Configurations Experiment 1



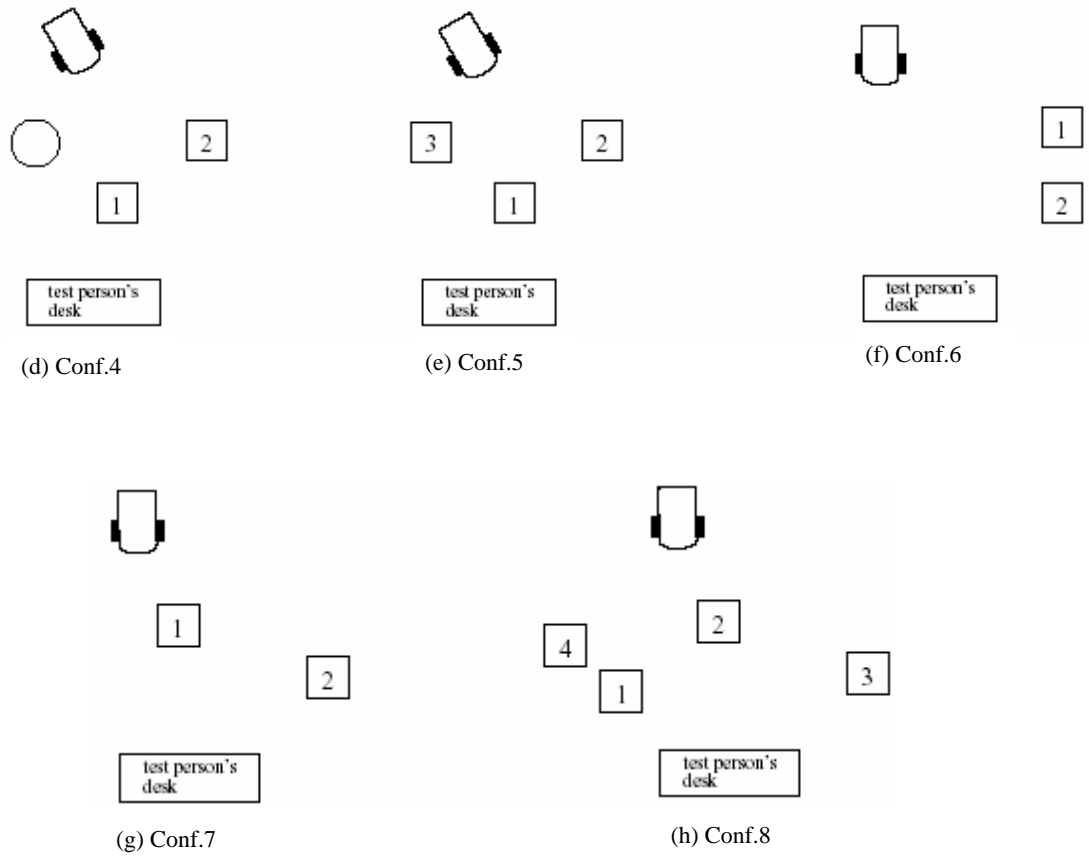


Fig.1. 7 Configurations Experiment 2

Motivated in part by their work, ours is different in two points: 1) we focus on the projection step, extending  $2d$  geometry projections to  $3d$ ; and 2) Unlike constructing a robotic system moving forward to *referents*, we prefer to locate the *referents* across a sliding window via spatial relations, instead.

# Chapter 2

## Towards Spatial Comprehension

*The only feature that distinguishes the target objects from others in scenarios is their position. This elicits the fundamental conception in spatial comprehension—the frame of reference. In this chapter, we classify 3 frames of reference—intrinsic, relative and group and elaborate how objects present in their usages. Then we propose spatial templates to model spatial linguistic expressions, from singular form to composites, to superlatives, from 2 dimensions to 3 dimensions. For each template, we conduct experiments over the author’s database.*

### 2.1 Understanding Spatial Knowledge

#### 2.1 .1 Terminology

---

Projective Prepositions	Topological Prepositions
front	above/top/upon/over
back	below/under/beneath/underneath
left	near/nearby/beside/around
right	far
leftmost	along/alongside
rightmost	here/ there
between	in/into
	on/onto
	at
	inside/outside

---

Table 1: Prepositions in English

In English expressions, the *referent* and *relatum* are encoded as noun phrases; the relation is encoded as a *preposition*, which is considered as a key part in a description. *Prepositions* are mainly divided into two classes: *topological*, and *projective*. Table.1 presents a list of the prepositions in English. In our work, we are interested in the projective ones.

In order to establish a validated spatial relation, it requires three entities: a located object, which is so-called *referent*, at least one reference object, which is so-called *relatum* and a prevailing reference coordinate, which is so-called *frame of reference*. In fact, *frame of reference* is the fundamental in spatial knowledge. It determines the direction in which the *referent* is located in relation to the *relatum*. In different kinds of scenarios the usage of frame of reference is impressively distinct. Next, we elaborate 2 basic classes of frames of references and distinguish the representation in spatial relations.

### 2.1. 2 Classification of Frames of Reference

Frame of reference is a 3-d coordinate system that defines an origin, orientation, and direction. English speakers most likely apply 3 principle axes: *top-down*, *left-right*, and *front-back*, which can be viewed as extending from the center of the reference object and providing 6 canonical directions. Specifically, direction determined by the *top-down* axis is given by gravitation, defining *over*, *above*, *under*, *below*, and *beneath*. Orthogonal against gravitation is the horizontal plane, which covers the other two axes and helps to define *front*, *back*, *left*, *right*, *besides*, *alongside*, and *next to*, etc. At current stage, we ignore the *top-down* axis so that the system is reduced to 2 dimensions. And also, our work barely focuses on projective relations, which, limits our interest on the various presentation of *front*, *back*, *left*, and *right*.

Projective prepositions can be used in different ways. One can easily identify the *front* side and find the *referent* as in the description like **the ball is in front of the car**. However, if we say **the ball is in front of the box**, it can mean the ball in relation to the box can be located either from the speaker's or the listener's viewpoint. This elicits the fundamental conception in spatial knowledge—the frame of reference. We should start from the classification by Levinson [50], who explicitly specified humans 'language and the proposal, is

the most widely approved systematic framework yielding spatially entities. In the first case, the relation is activated by *intrinsic* frame of reference. In the second case, the frame of reference we indicated is *relative*.

#### 2.1.2.1 Intrinsic frame of reference

One way is to refer the *intrinsic* side of a *relatum* e.g., the *front* side of car. In most of the cases, the *front*, *back*, *left* and *right regions* around a *relatum* are the regions adjacent to the *intrinsic front*, *back*, *left* and *right* side of the *relatum*, respectively. For example, a person's *front* is adjacent to their body that is on the opposite side to their back. When imposing *front-back* axis on a pair of objects, the *front* is the surface facing the speaker (or listener), and the *back* is the opposite surface. The relation in the *intrinsic* frame of reference only requires 2 arguments --a *relatum* and a *referent*. A description such as **the ball is in front of the car** is explicit enough for interlocutors to understand where the bike is in that the *front* side of car can uniquely distinguish from other side, see Fig. 2.1.

*Intrinsic* frame of reference is kind of an object-centered coordinate system and often treats as a complex form because it must be extracted from *relatums* by interlocutors in an appropriate way. Therefore, for the purpose of distinction, it is necessary to recognize some specific categories. Seeing an ambiguous figure as a *duck* or a *rabbit* leads the viewer to assign front to different regions of the object [75]. If you think the cars in Fig. 2.1 as a bottle, it may not felicitously say the ball is in front of the bottle. Concretely, objects like *people*, *houses*, *cars* and *televisions*, etc can serve as *relatums* because they have *intrinsic front* and *back* sides. However, objects like *boxes*, *balls* and *cans*, etc cannot because they don't have such characteristic.



Fig.2. 1: “Front” Examples in Intrinsic use.

#### 2.1.2.2 Relative frame of reference

*Relative* frame of reference can be employed in two different ways. If the object does not have an *intrinsic* side or its intrinsic orientation is not used for establishing the frame of reference, it can serve as a *relatum* in building *relative* relations. In this case, objects are often geometric symmetry in at least one dimension. And a *front* can still be contextually induced or projected on it [40, 60, 100]. Miller and Johnson-Laird [100] named it as an **accidental front**. It can be deduced through stating the **viewpoint**, which is an important contextual factor that explains the orientation imposed on the *relatum*, from where is viewed. Otherwise, it may give rise to ambiguity as in the example referred above. In particular, the speaker’s location can overwhelmingly serve as viewpoint, such as **from my viewpoint, the ball is in front of the box**, see Fig. 2.2. The listener’s position can also serve as viewpoint as well, but is used less frequently [83].

*Relative* frame of reference can apply on any objects. Objects like *balls*, *bottles* and *cups* belong to this domain because they don’t have *intrinsic* sides. We also can impose *relative* frame of reference on the object of which the *intrinsic* side is not referred, e.g. the *left-right* side of a computer display.



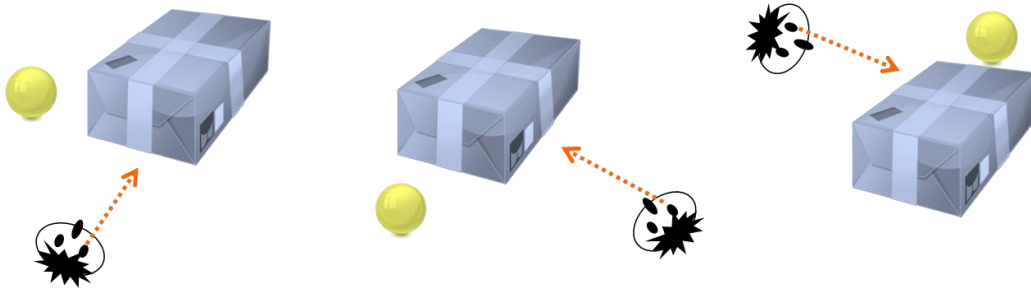


Fig.2. 2: “Left” Examples in Relative use

### 2.1.2.3 Beyond Relative Relation– Group-Based Frame of Reference

In the simplest case, a single object serves as the *relatum*, and can be detected in a straightforward manner by the speaker and the robotic system. However, spatial location descriptions can also serve to distinguish objects within a group of objects--a situation that occurs frequently in real-world environments, e.g. [98, 101, 107], but has been largely neglected in the literature. In addressing this issue, we focus on two particular questions: 1) since humans consider a group to constitute an integrated object, how can we enable a system to treat a group of objects likewise; and 2) what reference frames should be taken in these cases? Note that here the conception of “group” refers to either several identical or similar objects. In this section, beyond a single *relatum* scheme, we seek to distinguish unknown objects via simple and straightforward input instructions, such as **the pocket calculator is in front of the computer displays**.

When there are multiple identical or similar objects accumulated together in a scene, humans consider them to constitute a group. In this case, the group of objects can be viewed as a whole *relatum*. Intuitively, a conventional way for humans to specify which kind of spatial frame of reference should be imposed on a *relatum* is determine whether it has *intrinsic* part, especially, *intrinsic front*. However, this doesn’t make sense in *group-based* frame of reference. For example, a descriptive sentence such as **the pocket calculator is in front of the computer displays** is ambiguous because it can be interpreted in 2 ways. One is the calculator is

located *in front of* a group of monitors from the viewpoint of the speaker, with respect to the orientation of the monitors themselves, the other one is with respect to the *intrinsic* orientation of the displays. The first sense should employ the *relative* frame of reference with no doubt. It should be seemed to consider the *intrinsic* frame of reference but doesn't make sense in practice. Because every computer monitor within the group can have the *front* orientation so that it is impossible to find a mediate *front* to make them unified.

Thus, we resort for the first sense even if objects have intrinsic side. In the use of group-based frame of reference, Herskovits [39] claims that a situation depends on whether it is seen from the **inside** or the **outside**. In our work, we discuss these two situations, which are named as **internal use** and **external use**, respectively. **Internal use** refers to the interior spatial relations of the group in which the *referent* we intend to locate in relation to the *relatum* are both in the same group, see Fig. 2.3(a). In this case, human noticeably employ unary **superlative**. For example, they may use the terms such as *leftmost*, *middle* and *rightmost*. **External use** refers the exterior spatial relations in which the *referent* is independent from the group, see Fig. 2.3(b). In this case, we can treat as a derivation of *relative* frame of reference.

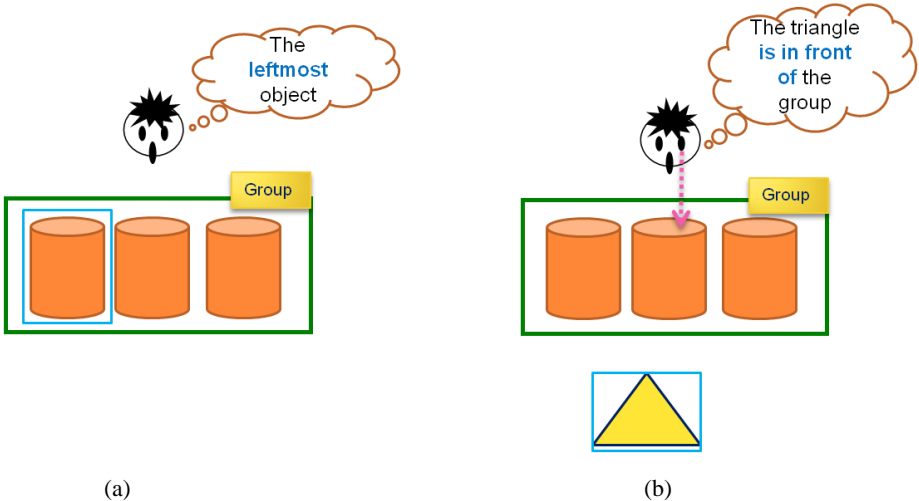


Fig.2. 3: 2 situations in the use of group-based frame of reference: (a) internal use; (b) external use

### 2.1.3 Spatial Templates and Their Acceptance Regions

A spatial template is a representation that is centered on the *relatum* and aligned with the frame of reference imposed on or extracted from the *relatum*. It is a 2- or 3-d field that best represent spatial between pairs or groups of objects appearing in space.

When determining whether a spatial relation can be applied on pair of objects what humans do in cognition is estimating the fitness. There are three main regions of acceptability. Roughly speaking, the position occupied by the *referent* is compared with the template to determine whether it falls into a *good*, *acceptable* or *bad* region: one reflecting good examples, one reflecting examples that are less than good but nevertheless acceptable, and one reflecting unacceptable examples [53]. If the *referent* falls into a good or an acceptable region when the spatial template is centered on the *relatum*, then the relation can apply to pair. It is should be noticed that the good and acceptable regions are not mutual or independent from each other, namely, there is no distinct with a sharp border between them. Instead, they gradually blend into each other. According to Logan [53], take the relation *front* as an example, any object that is aligned with the forward projection of the *front-back* axis (main axis) of the *relatum* is a good example. Any object parallel to a horizontal plane aligned with the *front* side of the *relatum* is an acceptable example, although not a good one. And any object opposite to a horizontal plane aligned with the *back* side of the *relatum* is a bad, unacceptable example; see Fig. 2.4(a). In our work, for the sake of simplicity, we merge the two regions into a larger one; see Fig. 2.4(b).

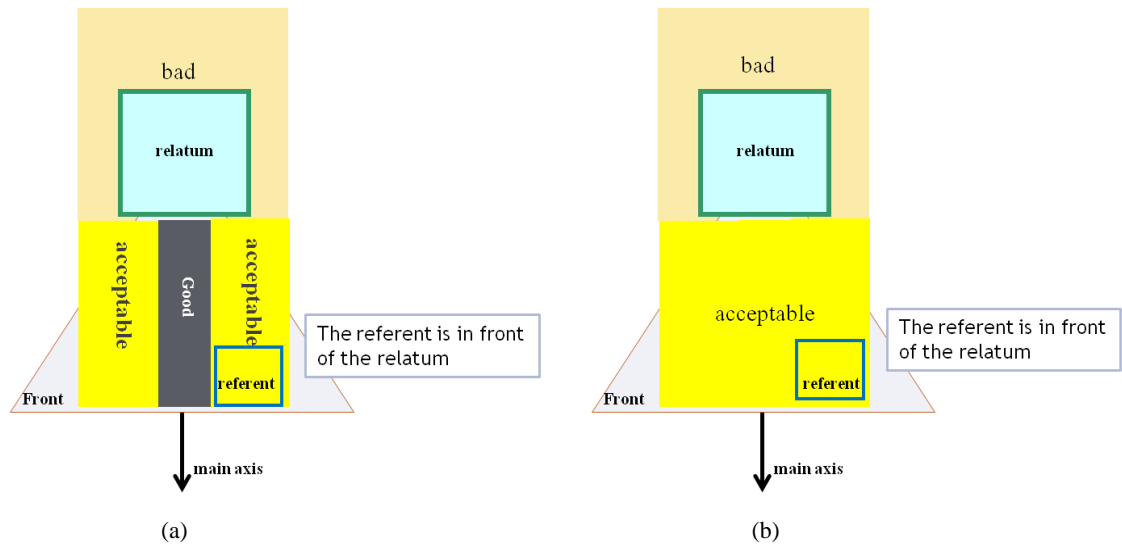


Fig.2. 4: (a) Good, acceptable, and bad regions for ‘front’ orientation in Logan’s template; (b) the acceptable and bad regions for ‘front’ orientation in our template. Here we merge the good and acceptable regions to form a large one.

## 2.2 Computational Model for Human Spatial Linguistic Expressions

The central functionality of our model is geometrically defined, but flexible mapping between projective linguistic expressions and spatial representation, allowing for 3 classes of frames of reference. The templates are based upon angular deviation. The basic idea is a listener projects a  $2d$  coordinate system onto a scenario. In *intrinsic* relations, it is aligned with the *intrinsic* frame of reference derived from the *relatum*. In *relative* relations, the spatial template is generated from the interlocutors’ viewpoint and projected onto the *relatum*.

## 2.2.1 The 2d Projective Model of Intrinsic and Relative Frames of Reference

### 2.2.1.1 Approach

We first proposed 2 templates on a 2d plane. The spatial templates of projective relation can be characterized as follows. For *intrinsic* frame of reference, the *intrinsic front* direction of *relatum* always serves as the main axis. While in *relative* frame of reference, classifying perceivers' viewpoint is crucial. The main axis is the connected line from the perceiver's center to the *relatum*. All objects are represented in a planar view (bounding box). Two diagonal axes through the centered of the *relatum* partition the plane into *front*, *back*, *left* and *right* parts. Each partition has a uniform triangular neighboring structure. Fig. 2.5 illustrates these two templates and the acceptance regions for each orientation.

The linguistic spatial expressions are represented  $\theta_{intr}$  and  $\theta_{rela}$ , respectively. It is the argument between the reference direction and the directed line from the *relatum* to the *referent*.  $\theta_{intr}, \theta_{rela}$  can be defined as:

$$\text{referent } \textit{front}_{\text{intri/rela}} \textit{ relatum} : 0 \leq \theta_{\text{intri/rela}} \leq 45^\circ \text{ or } 315^\circ \leq \theta_{\text{intri/rela}} \leq 360^\circ$$

$$\text{referent } \textit{back}_{\text{intri/rela}} \textit{ relatum} : 135^\circ \leq \theta_{\text{intri/rela}} \leq 225^\circ$$

$$\text{referent } \textit{left}_{\text{intri/rela}} \textit{ relatum} : 45^\circ < \theta_{\text{intri/rela}} < 135^\circ$$

$$\text{referent } \textit{right}_{\text{intri/rela}} \textit{ relatum} : 225^\circ < \theta_{\text{intri/rela}} < 315^\circ$$

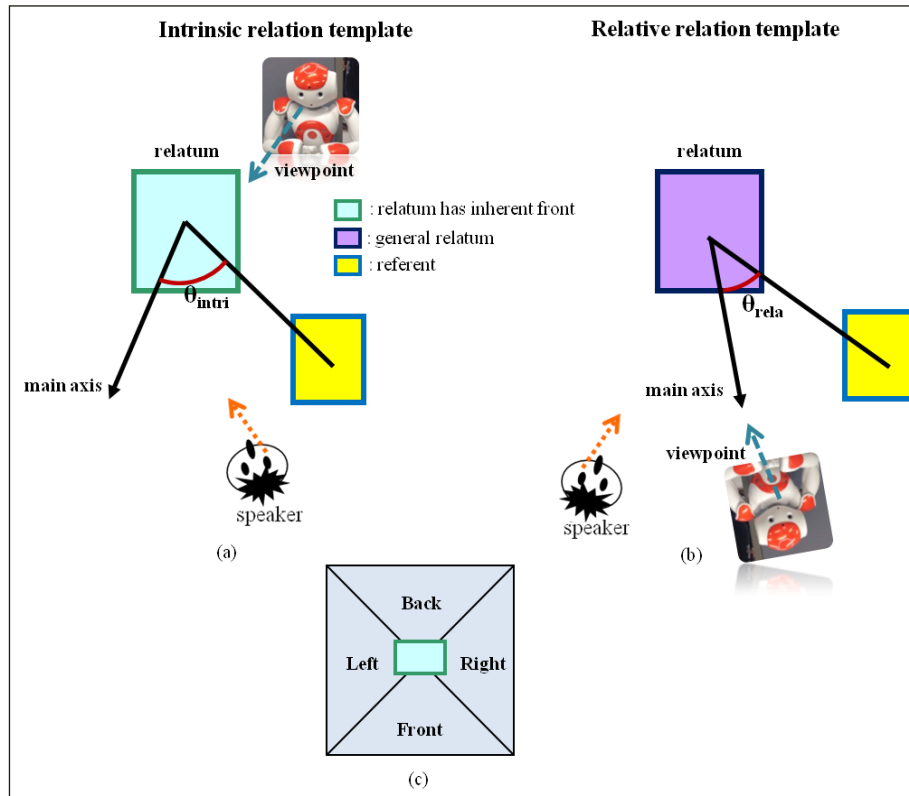


Fig.2. 5: Templates in intrinsic and relative frames of reference, and acceptance

### 2.2.1.2 Experiments

We made a series of preliminary experiment to evaluate the proposed approach on the author's database. The database is collected from 15 categories, in total of 450 images containing singular object for training and 60 scenarios for spatial recognition testing. The scene configuration is simple. Of the 100 images, 30 consist of 2 objects in which only 1 *referent* in relation to 1 *relatum*. There are another 30 images containing 3 objects in which 2 *referents* are surrounded by 1 *relatum*. The *relatums* are at their frontal view pose.

Experiments are done in three steps: (1) detection, where we use the automatic object detection model provided by [19] to split the *detected (known)* and *un-detected (unknown)* objects; (2) indexing *unknown*

objects, if multiple *unknown* objects are present; and (3) recognizing, where given the specified *relatum* and *referent*, we use our proposed method to detect the spatial relationship between them.

In practice, we invite 10 university students to write down spatial descriptions for every *unknown* object in relation to the *detected* objects. In experiments, we randomly select 6 scenarios per round; the experiments are conducted for 10 times. For the sake of simplicity, we have trained and tested the object detector in advance over the whole database, and ensure that all the relevant information, including the bounding boxes, the locations, widths and heights for the *detected* objects have been stored in cache. Thus, our experiment started from step (2) by manually labeling and indexing *un-detected* objects.

We first examined the simplest cases where only one *detected* and *unknown* objects exist. Among 30 images, our model can successfully achieve the same results with human participants write in advance in 18 images, which as shown in Fig. 2.6. Fig. 2.7 shows the observation of the result from the rest of 30 images. The precision is 80%, namely, 12 out of 60 pairs of spatial relationship is identified by mistake.

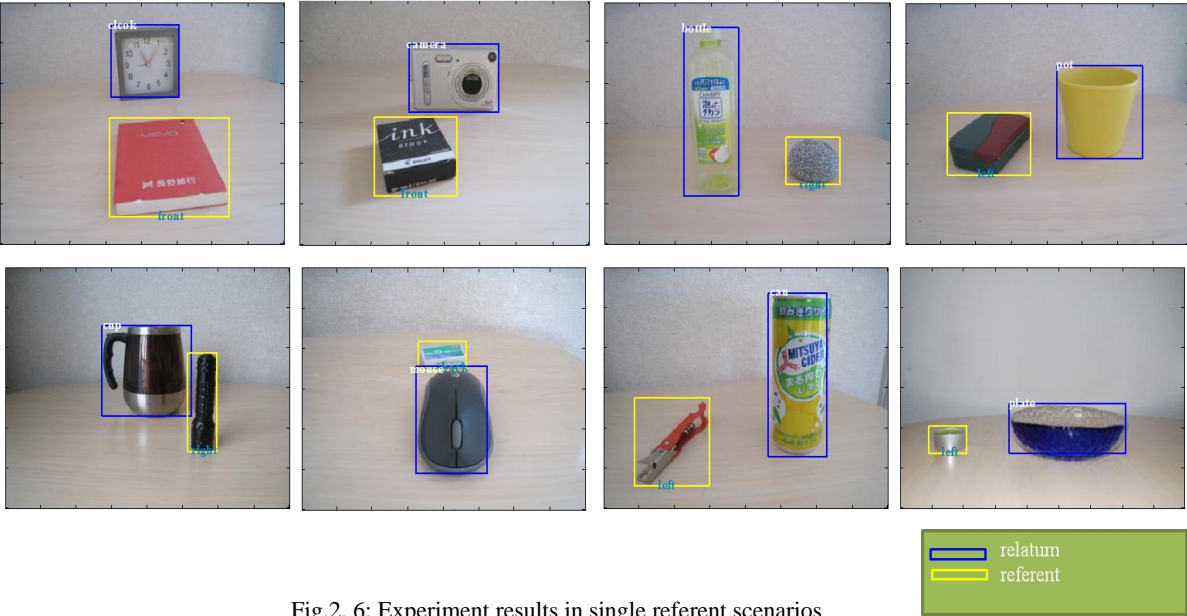


Fig.2. 6: Experiment results in single referent scenarios

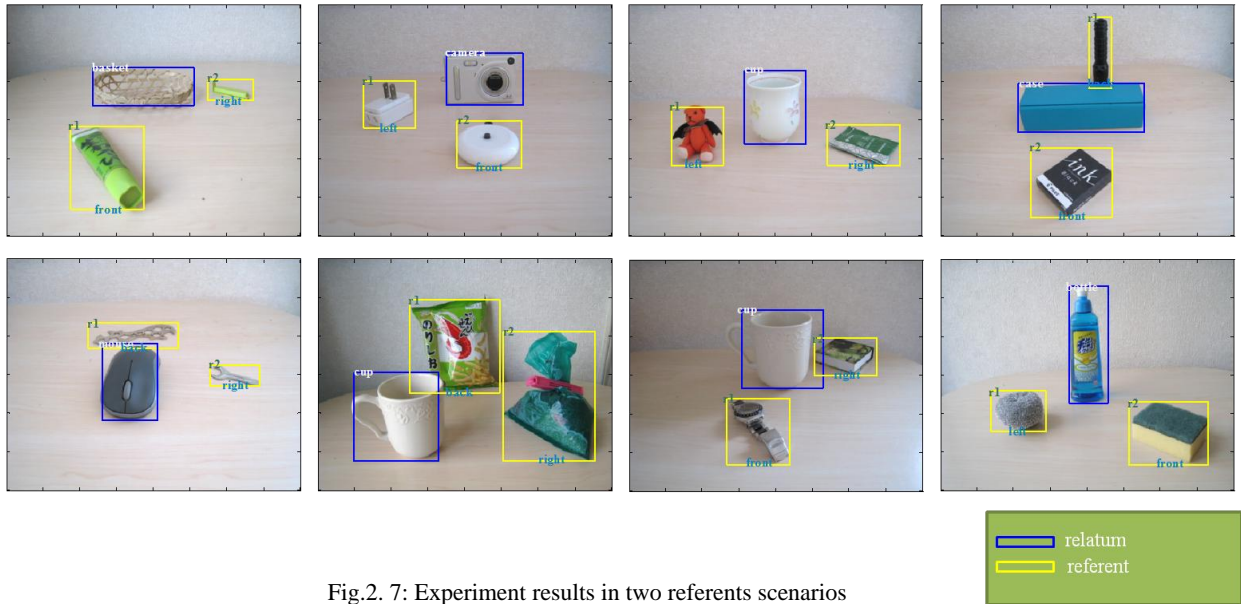


Fig.2. 7: Experiment results in two referents scenarios

### 2.2.1.3 Discussion

The result is competitive though, still far away from our expectation because of the simplicity of scene configuration.

In the following, we analyze two main issues that influence the success: distance and acceptance regions.

**Distance:** There is a significant interaction between angle and distance in our experiments. An evident from Fig. 2.8, the flashlight is moved slightly further away from the mug from scene (a) to (b). While being depicted from the same viewpoint, our model reckoned the configuration was the same in both of the scenes and evaluated the flashlight *to the right of* the mug whereas human users thought the flashlight became being *in front of* mug in scene (b). Actually, it is indeed considered as slightly different spatial configuration in perception. The deviation occurred in the process because of reducing the reference plane from  $3d$  to  $2d$ , which may give rise to a loss of finer details that are important for differentiating spatial relation. Therefore, we need to return to the  $3d$  world where we exist to cast eyes on the problem.



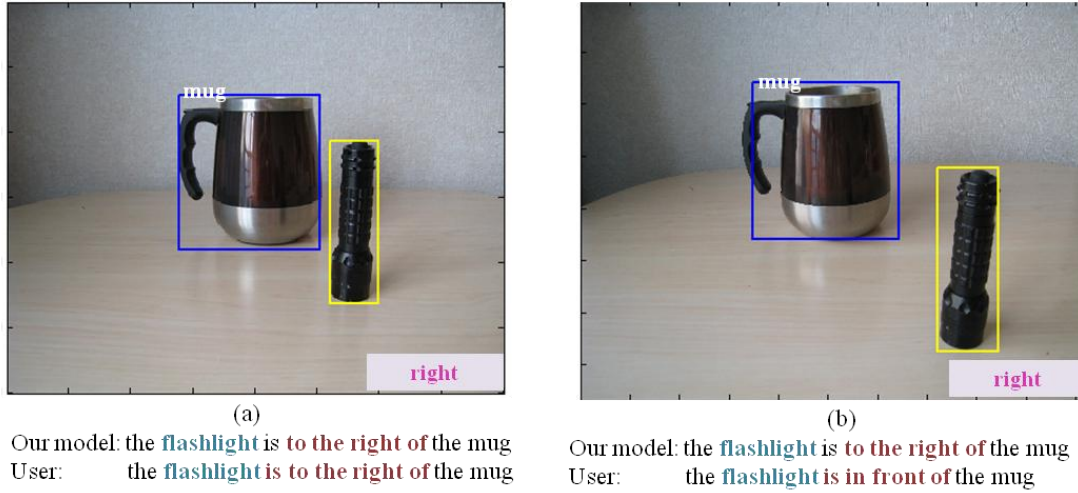


Fig.2. 8: Failure case study: interaction between angle and distance

**The Acceptance Regions:** A particular situation occurs in the case where a *referent* falls into two acceptance regions, and therefore gives rise to the system receptivity dilemma. Take the Fig. 2.9 as an example. Assuming a *referent* is being translated along the main axis relative to a *relatum*. Our model is able to distinguish (a) from (c) in accordance with humans' cognitive system in that most of the whole *referent* explicitly fell into one acceptance region with no doubt. However, a problematic result occurs in scene configuration (b) in which a *referent* is spanned two acceptance regions by coincidence. In Fig. 2.10, we collect in total of validate 36 utterances of which the model recognizes *left* in 7 utterances. There is another 7 utterances use the term *front*, which obtains the same portion, approximately 20%, with the *right*. The result reveals an interesting phenomenon that even human users can hardly to distinguish it was *front* or *left* at a glance. Perhaps, arguing either *front* or *left* being acceptable is unnecessary, because both of them are reasonable per se. Therefore, to reach a consensus with human's cognition, a conceivable way is improving the range of acceptance regions, making both fit.

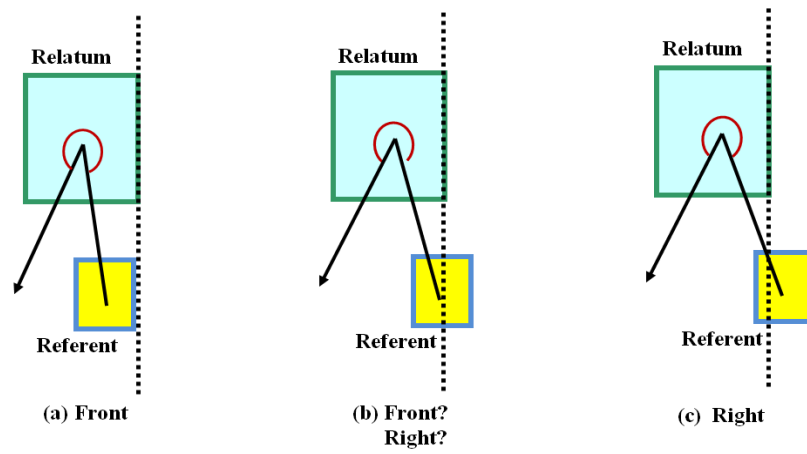


Fig.2. 9: Vague vs. non-vague situations: the vagueness is occurred when a referent falls into two acceptance regions.

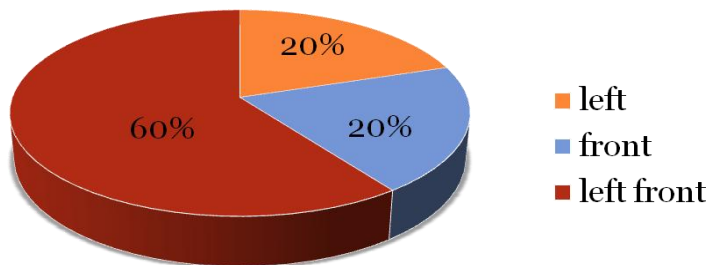
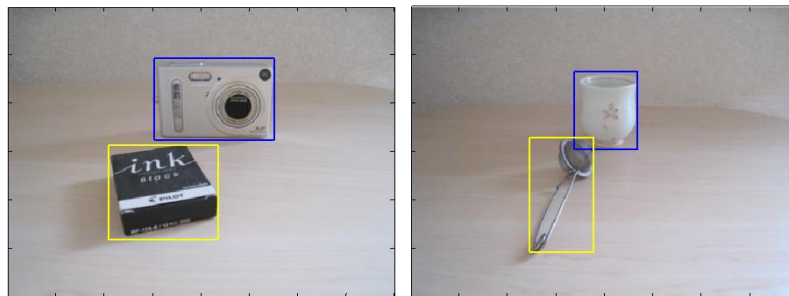


Fig.2. 10: Distribution of the left, front and left front utterances in two scenarios in the experiment

## 2.2.2 Modifications: The 3-D Computational Model

### 2.2.2.1 Approach

We defined a viewing sphere plane to represent *3d* information. The relationship between a particular *relatum* and *referent* in *intrinsic* and *relative* relations are represented in Fig. 2.11(a) and (b), respectively. Apparently, the center of the circular reference plane is the middle point on the base line of the bounding box. The most functional improvement of our modified model is the reference plane is not fixed, but flexible. We take advantage of the centre of the *referent* and find the perpendicular line, which is set as half as the Euclidean distance from the *referent* to *relatum*. Then the reference plane is determined with a radius from the foot point to the centre of the circle. This ensures the *referent* always in an acceptance region. The enlarged acceptance regions are shown in Fig. 2.11(c). The angle  $\theta_{intr}$  and  $\theta_{rela}$  thereby become the angle between the main axis and the straight line from the projection of *referent* onto the sphere plane. It should be noticed that it may acquire an *accidental front* by virtual viewpoint in *relative* frame of reference.

$\theta_{intr}$  and  $\theta_{rela}$  can be formulated as:

$$\text{referent } \textit{front}_{intri/rela} \textit{ relatum} : 0 \leq \theta_{intri/rela} \leq 60^\circ \textit{ or } 300^\circ \leq \theta_{intri/rela} \leq 360^\circ$$

$$\text{referent } \textit{back}_{intri/rela} \textit{ relatum} : 120^\circ \leq \theta_{intri/rela} \leq 240^\circ$$

$$\text{referent } \textit{left}_{intri/rela} \textit{ relatum} : 30^\circ \leq \theta_{intri/rela} \leq 150^\circ$$

$$\text{referent } \textit{right}_{intri/rela} \textit{ relatum} : 210^\circ \leq \theta_{intri/rela} \leq 330^\circ$$

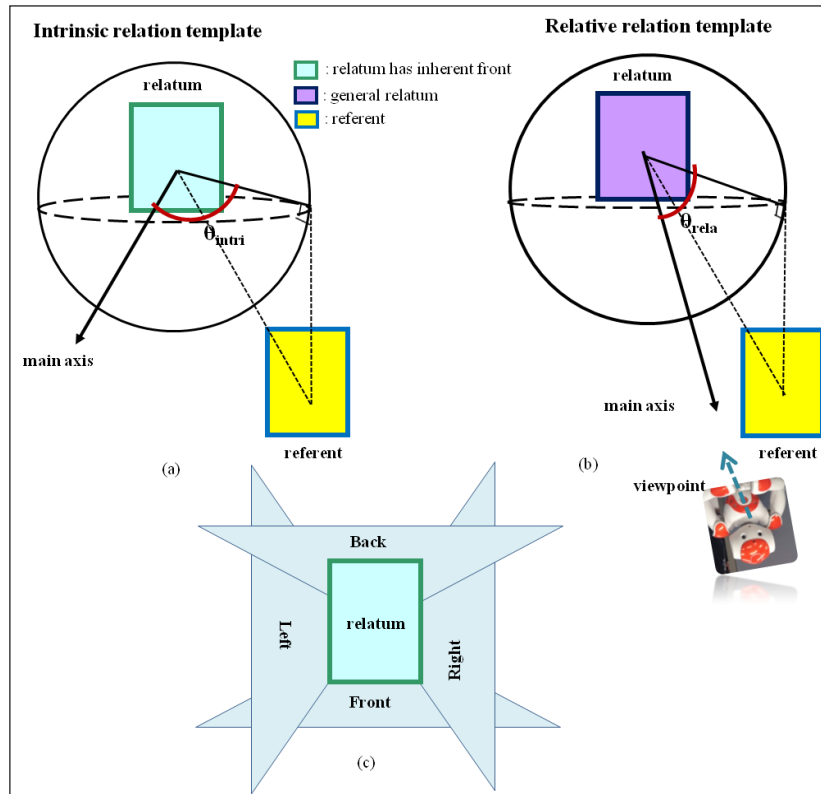


Fig.2. 11: Modified *intrinsic*, *relative* relation templates, and enlarged acceptance regions for each orientation

The enlarged acceptance areas can compromise with the paradox mentioned above. Alternatively, our visual experiences and intuition suggest a preference for both reasonable options, that is, asserting these two spatial relations together. Evidence is supported by the data we obtained from the experiment described above that rest of 22 collected utterances, approximately 60% combine two canonical terms together. In practice, such phenomenon is quite common in English. Normally, no more than two spatial relations are combined, e.g., *front* and *left* or vice versa; or more regularly, using compound expressions, for example, users can depict Fig. 2.9(b) as the (*referent*) is to the *right-front* of the (*relatum*).

Specifically, Fig. 2.12 shows four compound regions corresponding to *left-front*, *left-back*, *right-front* and *right-back*. We also note that not only are the composites overlapping regions generated from canonical expressions, but enlarged to the presentation of that expression in the best way.

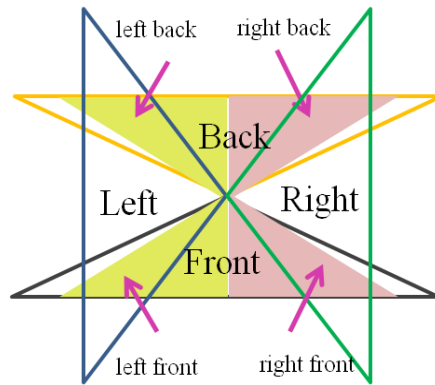


Fig.2. 12: Compound expressional regions

They can be formulated as:

$$\begin{aligned}
 \text{referent } \textit{left - front}_{\text{intri/rela}} \textit{relatum}: 0 \leq \theta_{\text{intra/rela}} \leq 45^\circ \\
 \text{referent } \textit{left - back}_{\text{intri/rela}} \textit{relatum}: 135^\circ \leq \theta_{\text{intra/rela}} \leq 180^\circ \\
 \text{referent } \textit{right - front}_{\text{intri/rela}} \textit{relatum}: 180^\circ < \theta_{\text{intra/rela}} \leq 225^\circ \\
 \text{referent } \textit{right - back}_{\text{intri/rela}} \textit{relatum}: 315^\circ \leq \theta_{\text{intra/rela}} \leq 360^\circ
 \end{aligned}$$

#### 2.2.2.2 Experiments

In this set of experiments, to evaluate our proposed model's precision, we designed it as is able to automatically select the *relatum*. If there are several candidates available, our model is designed to choose the optimal one.

The experiments are done in four steps: (1) detection, where we use the automatic object detection model to split the *detected (known)* and *un-detected (unknown)* objects; (2) indexing *unknown* objects, if multiple *unknown* objects are present; (3) choosing adequate *relatum* for each; and (3) recognizing, where given specified *relatum* and *referent*, we use our proposed method to evaluate spatial relationship between them.

To determining the adequate *relatum*, we followed the very simple criterion, that is, the distance from the *referent* to the *relatum*, which is considered a crucial factor in state-of-the-arts [103]. We chose the object, which owns the shortest distance to the *referent* as *relatum*. We applied two kinds: Euclidean distance and Manhattan distance. If a pair of distance belonging to the same kind has the same value by coincidence, we opt for both and evaluate the spatial relationship in turn.

We first test our model over canonical spatial expressions. The composites are not allowed. We invite 20 university students to write down spatial descriptions for every *unknown* object in relation to all *detected* object that can serve as a possible *relatum* in total of 100 scene images. We then count the highest votes for each pair of objects as a benchmark. In experiments, we randomly selected 10 images per round; the experiment is conducted for 10 times.

We evaluated the performance on 3 kinds of scene configurations, from easy to difficult, where 2-5 objects were placed on the table. We first examined the simplest cases where only one *known* and *unknown* objects exist. This indicates no need for evaluating the optimal *relatum*. Of all 30 images, our model successfully achieved the same results with what human participants wrote in advance in 28 images, as shown in row 1 of Fig. 2.13. Row 2 of Fig. 2.13 shows the observation of the result from another 30 images in which one *unknown* object is surrounded by two *known* objects. As we expected, because our modified model ensures the observation of finding the shortest projection from *referent* to the reference plane, the localization result is not unobtrusively influenced by the distance from *relatum* to *referent*, especially, in more than half of the cases the Manhattan distance we measured appears to be larger than the Euclidean distance. The precision is 97%, namely, only 2 out of 60 pairs of spatial relationship is identified by mistake. Finally, we tested the complicate case over 40 images with 2 or 3 objects being surrounded by 2 *known* objects. The result is shown in row 3 of

Fig. 2.13. Our model still performs stably. Among 220 spontaneous descriptions generated from our model, 200 pairs of spatial relations, yielding approximately 91%, conforms to human experimenters' answer.



Fig.2. 13: Canonical spatial relation experiment result. Row 1: the scenarios of 1 referent vs. 1 relatum. Row 2: the scenarios of 1 referent vs. 1 or 2 relatum(s). Row 3: the scenarios of 2 or 3 referents vs. 2 relatums. The referent and chose relatum is linked by →

Next, we test the performance on the usage of compound spatial expressions. Along with the evaluation method of object detection, the precision of spatial recognition can be defined as:

$$\text{Precision} = \frac{\# \text{ of true positive spatial relationships}}{\# \text{ of recorded spatial relationships}} \quad (2.1)$$

Again, we required the students who have took part in the previous experiment to write down spatial descriptions. We used 60 images including 45 positives and 15 negatives. The scenarios are designed as simple. The positive images are arranged by 3 kinds of configurations: (1) 1 *referent* and 1 *relatum* only; (2) 1 *referent* and 2 *relatums*; and (3) 2 *referents* and 1 *relatum*. In negative ones, we choose the simplest scene configuration with only 1 *referent* in relation to 1 *relatum*. Because canonical spatial expression is somewhat conflicted with the compound expressions, it is not allowed to utilize. We expect the model to alarm when confronting the negatives. The result is impressive, in total of 70 spontaneous descriptions generated from 45 ones, 68 obtain the success with the precision is 93%. Since there is no description generated from the negative examples, we barely count the alarm times when *referents* are beyond the scope of compound regions. By comparing with the total of 15 given answers, the true negative rate is  $\frac{13}{15} \times 100\% \approx 87\%$ . Some representative results are shown in Fig. 2.14.

The interactive mode result is elaborated in Chapter 5.





Fig.2. 14: Compound spatial relation experiment result. The referent and chose relatum is linked by →

## 2.2.3 The Model of Group-based Frame of Reference

### 2.2.3.1 Approach

The situation described earlier proves useful for scenes wherein only 1 object serves as a *relatum*. This study seeks to take this notion further, because we argue that it can also be applied in the case of groups of objects. When there are multiple identical or similar objects accumulated together in a scene, humans consider them to constitute a group. In this case, the group-relatum can be viewed as a whole bounding area, which functions as a rectangular sliding window across the group.

It has been observed that humans first determine whether the *relatum* has *intrinsic* parts in practice. Such cognitive behavior may give rise to conflict in the cases of those objects, e.g., computer display, accumulated together. For example, a description such as “**The pocket calculator is in front of the computer displays**” is ambiguous. It can be interpreted in several ways. The calculator can be located in front of a group of computer displays from the viewpoint of the speaker, with respect to the orientation of the group, or with respect to the intrinsic orientation of the computer displays. Herskovits [39] claims that a situation depends on whether it is seen from the inside or the outside of the group. In our approach, we discuss these two situations, which are called **internal use** and **external use**, respectively. **Internal use** refers to the interior spatial relations of the group in which the *referent* we intend to locate in relation to the *relatum* are both in the same group. In this case, human noticeably employ unary **superlative**. For example, they may use the terms such as *leftmost*, *middle* and *rightmost*. **External use** refers the exterior spatial relations in which the *referent* is independent from the group. In this case, we suggest that it could be viewed as a type of **relative** frame of reference.

### **Internal Use**

Speakers describe the position of an object in a group using the spatial relations between the object and the group as a whole, such as: “**the second from the left of those objects.**” Additionally, humans also noticeably employ unary **superlatives** to describe group of objects. For example, they may use the terms *leftmost*, *middle* and *rightmost*, which are binary-level spatial relations based on projections of bounding boxes on the x and y axes. Therefore, we define them based on the Manhattan distance between bounding boxes.

$$D_{manhattan} = |C_x - O_x| + |C_y - O_y| \quad (2.2)$$

where  $C(x, y)$  is the centroid of the bounding box of the total group, and  $O(x, y)$  is the centroid of the bounding box of each object in the group. Then, the superlatives can be defined as:

$$\begin{aligned}
 \textit{Leftmost} : O_x \ll C_x \textit{ and } \arg \max \{ D_{\textit{manhattan}} \} \\
 \textit{Middle} : O_x \approx C_x \textit{ and } \arg \min \{ D_{\textit{manhattan}} \} \\
 \textit{Rightmost} : O_x \gg C_x \textit{ and } \arg \max \{ D_{\textit{manhattan}} \}
 \end{aligned}
 \tag{2.3}$$

### External Use

Speakers can specify the position of one object which does not belong to the group using the whole group as a *relatum*. This is defined likewise in the way of a conventional relative reference system presentation.

#### 2.2.3.2 Grouping Objects

A core part of the model is grouping objects. This done by an agglomerative clustering algorithm resembles the K-means algorithm.

The k-means algorithm [57] is used to find features that are typical and representative for a give object category. It is one of the simplest and most popular clustering methods. It pursues a greedy hill-climbing strategy in order to find a partition of the data points that optimizes a squared-error criterion. The algorithm is initialized by randomly choosing  $k$  seed points for the clusters. In each iteration the data point is assigned to the closest cluster center. When all data points have been assigned, the cluster centers are recomputed as the means of all associated data points. In practice, this process converges to a local optimum within a few iterations.

In visual object recognition approaches [46, 87, 88, 89, 104, 105], many algorithms employ k-means clustering because of its computational simplicity, which allows to apply it to very large datasets. Similarly, in our work we inherit the idea to group concrete objects.

However, the most deficiency of the algorithm is it requires the user to specify the number of groups in advance, which is not what our human beings do. Thus, developing an efficient automatic grouping algorithm is our long-term goal.

---

### **The Grouping Objects Algorithm**

---

#### **Input:**

n-the number of groups you want

m- set of centroid vectors

#### **Local Variables:**

c- centroid coordinate matrix

g - current iteration group matrix

i -scalar iterator

---

#### **function y = groupObj(m n)**

```
% initializes value of centroid to start clustering
```

```
for i = 1:k
```

```
    c(i,1) = m(i,1);
```

```
    c(i,2) = m(i,2);
```

```
end
```

```
temp = zeros(maxRow, 1); % initialize as zero vector
```

```
While 1,
```

```
    d = computeDist(c,m); % calculate objects centroid distance
```

```
    [z,g] = min(d, [],1) % find group matrix g
```

```
    If g == temp,
```

```
        break; % stop the iteration
```

```
    else
```

```

    temp = g; % copy the group index
end
for i = 1: k
    c(i,1) = mean(m(find(g==i), 1);
    c(i,2) = mean (m(find(g==i), 2);
end
end
y =[m, g]
end

```

---

### 2.2.3.3 Experiment

The experiment is an interactive manner. It should start from grouping objects. We use 100 images on the author's dataset, containing several objects each. After objects are separated by *known* and *unknown* labels, the model enquires the user number of group. We show the textbox in Fig. 2.15.

We imply 2 kinds of scene configuration: non-unknown objects and 1 unknown object. Experiments are done in three steps: (1) object detection; (2) labeling *unknown* object if there is; (3) input number of group; (4) grouping objects; and (5) identifying spatial relations. Fig. 2.16 shows the grouping results. In Fig 2.17, we show some results. The compound expression result is represented in the last row.

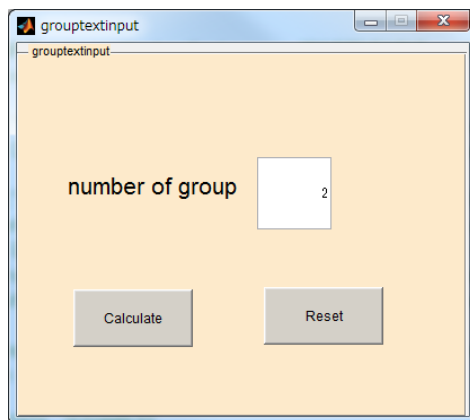


Fig.2. 15: Text Input box



Fig.2. 16: Grouping results by k-means algorithm

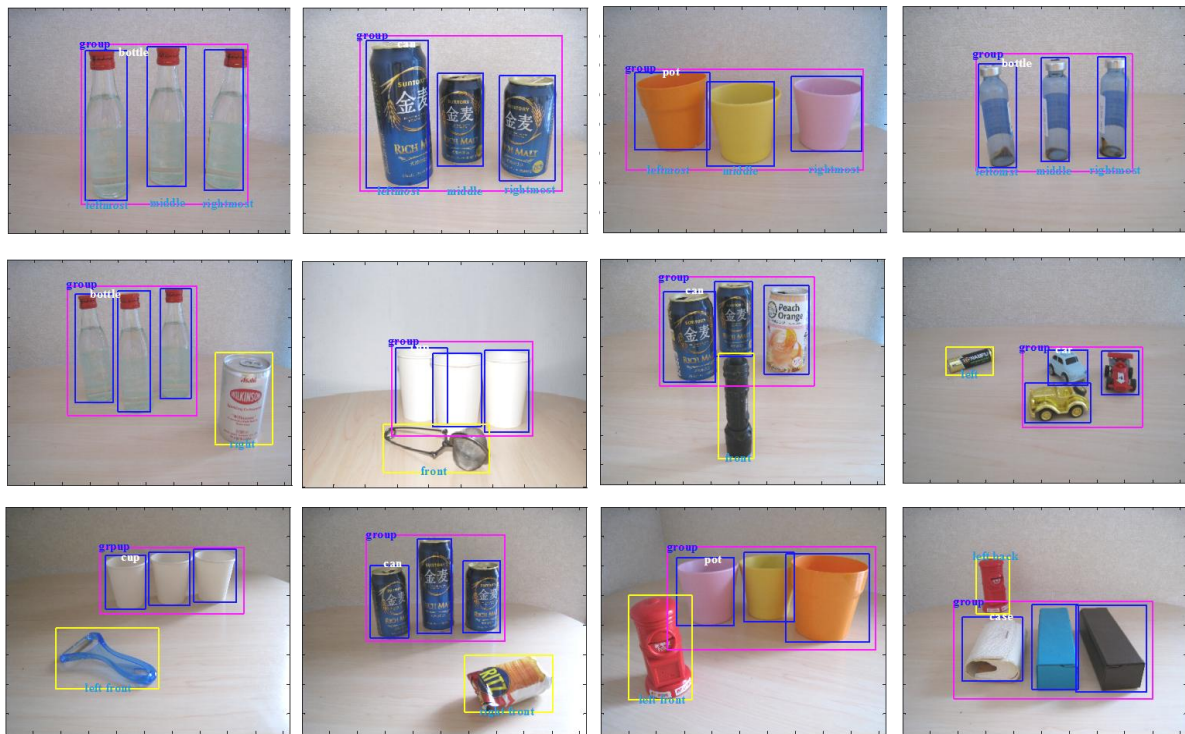


Fig.2. 17: Recognition result



However, such k-means algorithm has 2 known deficiencies. Firstly, it requires the user to specify the number of groups in advance, which is not what our human beings do for scene understanding. Secondly, the k-means procedure is only guaranteed to find a local optimum, so in some of cases, the results we obtain may be quite different from run to run. Fig. 2.18 shows a case in which the 4 objects are almost equidistance. We run the algorithm 3 times, it turns out that the result is different from time to time. Thus, developing an efficient automatic grouping algorithm is our long-term goal.

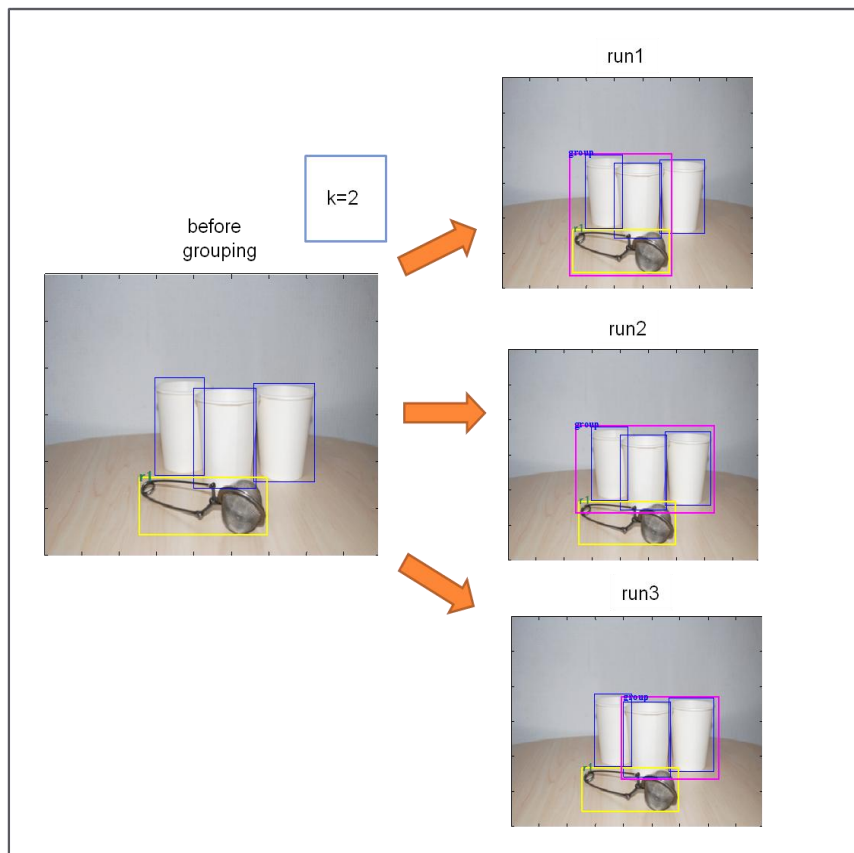


Fig.2. 18: Deficiency of the resembling k-means algorithm: different results obtained



## **2.3 Conclusion**

We proposed a novel approach for recognizing objects by distinguishing spatial relations. The approach contains 3 templates, which can adopt in intrinsic, relative and group-based frames of reference, respectively. Especially, the improved 3-D model is flexible so that can cover more complex situations and compound linguistic expressions. We evaluate the approach in 3 experiments, which has shown the ability to accurately identify spatial relations in various scene configurations. All the experiments hitherto are conducted without human interaction. In Chapter 7, we would like to show the interactive experiment result.

## Chapter 3

### Pose Estimation

*In natural scenes, objects are often arbitrarily placed. A typical case occurs in the use of intrinsic frame of reference. Since the intrinsic orientation is extracted from the relatum, irrespective of the user's viewpoint, when the relatum is rotated, its axis must be adjusted correspondingly. In ground plane level, finding intrinsic front orientation seems the most striking case. Because once the front orientation is determined, the back, left, and right orientation can be deduced. Thus, the core part of the problem to be addressed is in relation to the pose estimation. In this chapter we propose a two step approach that concerns estimating the correct object pose.*

#### 3.1 Instruction

In natural scenes, objects are often arbitrarily placed. A typical case occurs in the use of *intrinsic* frame of reference. Since the *intrinsic* orientation is extracted from the *relatum*, irrespective of the user's viewpoint, when the *relatum* is rotated, its axis must be adjusted correspondingly. The processes involve translate the origin of the frame of reference, rotate its axes to the relevant orientation, and choose a direction. Not all of these adjustments are required for every relation. *Near* requires setting the origin and the scale, whereas *above* requires setting origin, orientation, and direction [53]. In ground plane level, finding *intrinsic front* orientation seems the most striking case. Because once the *front* orientation is determined, the *back*, *left*, and *right* orientation can be deduced. Thus, the core part of the problem to be addressed is in relation to the pose estimation. In order to complete understanding spatial relation and objects representation in a visual scene, in this chapter we propose a two step approach that concerns estimating the correct object pose.

To reliably estimate object pose, we need to model their effect on the image features. We utilize significant image characteristics shared by object instances of the same category such as edge orientations and feature locations that can be used to constrain the pose and scale at which the object is imaged. Note that these estimators regardless of background. Our method shows that given appropriate training data and powerful feature, even a simple naïve Bayes classifier is sufficient.

Once we obtain such classifier for object pose, we can predict the same object instance pose with the similar one. The positive training sets for these view-tuned classifiers contain much less feature variation compared to that of a pose invariant approach. Therefore the feature statistics are much easier to model. In fact, this is similar in spirit to the one used in keypoint descriptors such as SIFT [54] to achieve scale and rotation covariant keypoint recognition. SIFT uses maxima of Laplacian and the gradient orientation histogram to estimate the scale and rotation of the keypoints to be matched, both of which can be directly computed from image features. Our pose estimators replace this direct computation with a probabilistic one. Furthermore, as in the case of keypoint descriptors, we do not require these estimates to be perfect. Approximate values are sufficient because we rely on histogram based feature representations that are largely invariant to small changes in bounding box and view angle.

Overall, ours is a layered approach. Here by ‘pose,’ we defined it by the viewing angle. We first build a view sphere which contains 7 pose transformations. Then we train the estimator only on the bounding box dimension instead of on the whole image. We quantify the performance on the author’s database. More details about the author’s database introduce in Chapter 4. Finally, we compare two state-of-the-arts to prove our result is competitive.

### **3.2 Related Work**

In the early days, vision researchers paid close attention to the 2D-to-3D correspondence, but many approaches were line-based and had many difficulties dealing with real-life images. The aspect graph of [48] presents a theory for modeling 3D objects with a set of inter-connected 2D views. This theory has a sound

psychological foundation (e.g. [12]) and has been very influential and underlies most approaches to 3D object recognition.

Estimating the 3D pose of objects is a classical problem, and many solutions have been developed using either local features (e.g. [20]) or shape outlines (e.g. [45]), usually assuming perfect knowledge of the object. With the maturation of local feature detection (as in SIFT and its variants), latest progresses on pose estimation have mostly been local-feature based (e.g. [14, 22]) and performed fairly well on instances of objects, preferably with texture. There has been an increasing interest lately in 3D object pose classification, which aims at predicting a discrete set of viewpoints. A variety of approaches have been explored (e.g. silhouette matching [17] or implicit shape models [3] or virtual-training [13]). At the same time, many works on category-level classification also address the issue of multiple views (e.g. [77, 99]). The series of work from Savarese and Fei-Fei [79, 94, 95] directly address the problem of 3D viewpoint classification at the category and are the most relevant for us. They have developed a number of frameworks for 3D viewpoints, most adopting the strategy of grouping local features into parts and learning about their relations. Similar approaches have been adopted in a number of other works (e.g. [52, 56]) that show promising results. The 3DObject dataset of Savarese et al [79] is a standard benchmark for viewpoint classification and has a systematic collection of object views. A number of categories from the PASCAL challenge [25], such as cars, are also annotated with viewpoints. We quantitatively evaluate our approach on these datasets.

The most recent progress sees the use of part-based templates [27]. These techniques have been shown to perform very well on real life cluttered images. The work of Gu[35] work is based on the mixture-of-HOG approach but focuses on viewpoints instead of categories. He explicitly handles viewpoints and train HOG models with a large number of viewpoints/components. His work also develops approaches for semi-supervised and unsupervised learning of viewpoints, and extends the discrete viewpoint model to the continuous case. Bao[4] introduce a new problem object co-detection. Given a set of images with objects observed from 2 or multiple images, the model is able to detect the objects, establish the identity of individual object instance as well as estimates the viewpoint transformation of corresponding object instances. He

measures appearance consistency between objects by comparing part appearance and geometry across images.

### 3.3 The Model

We use Naïve Bayesian classifier to learn the histogram distribution from a series of well-defined pose. For a novel pose, our objective is to match it with the closest samples. The result is represented by an azimuth angle.

#### 3.3.1 Building Key-Pose Structure

In our model, an object pose is denoted the azimuth and zenith angles:  $\alpha$ , see in Fig. 3.1. Practically, we take photos with a hand-camera by walking around an object to record pose changes. The canonical view of an object is defined as the most frontal view pose. Given a groups of images, we annotate them manually, first. We sample uniformly every 30-degree as a keypoint. For plane 180, we obtain 7 key poses in total. This is shown in Fig. 3.2.

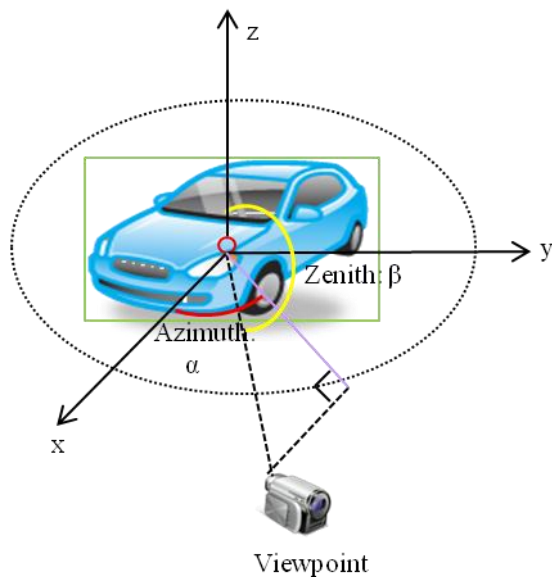


Fig.3. 1: Azimuth and zenith angle  $\alpha$ ,  $\beta$  representation

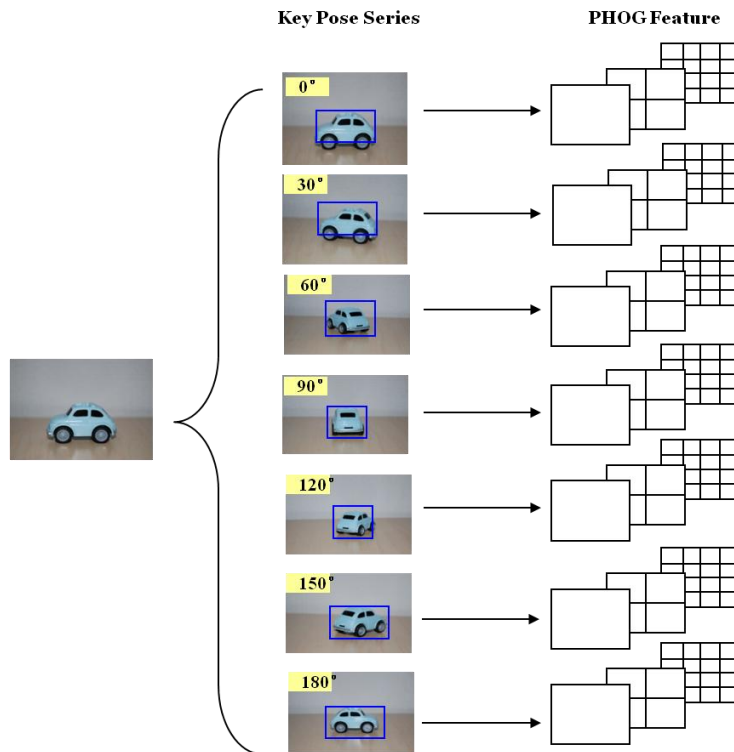


Fig.3. 2: A car instance example. The key pose structure is built from 7 poses from the same viewpoint.

### 3.3.2 Image Feature

Given a bounding box that defines a region of interested (ROI), we describe it in terms of histogram-based features, which have become the norm in object detection due to their ability to handle large intra-class variation and to provide robustness against errors in bounding-box size and location. In practice, we first compute at every pixel a SIFT descriptor [54]. We then assign to each pixel a cluster number to create label maps. The clusters centers are estimated using k-means. Finally, we create a spatial pyramid of gradient histograms (PHOG) [9]. Instead of from whole image area, we extract feature from a bounding area, which is

the smallest rectangular region enclosing the object of interest within the image such as the one in Fig. 3.3. The local shape is represented by orientations of an edge histogram within an object's sub-region quantized into  $k$ -bin and each edge's contribution is weighted by its magnitude. Therefore, each bin in the histogram represents the number of edges that have orientations within a given angular range. The spatial layout is given by tiling the object into regions at multiple resolutions (Fig. 3.3). As a result, at each level  $0 \dots l$ , the final shape descriptors consist of a histogram of orientation gradients over each object sub-region. In forming the pyramid the grid at level  $l$  has  $2^l$  cells along each dimension. Consequently, level 0 is represented by a  $k$ -vector corresponding to  $k$  bins of the histogram, level 1 by a  $2^1 \times 2^1 \times k$ -vector etc, and the PHOG descriptor of the entire image is a vector with dimensionality  $k \sum_{l \in L} 4^l$  and is normalized to sum to unity so that some objects (edge rich) are not weighted more strongly than others. For example, in our case, for levels up to  $l=2$  and  $k=40$  bins, it will be an 840-vector.

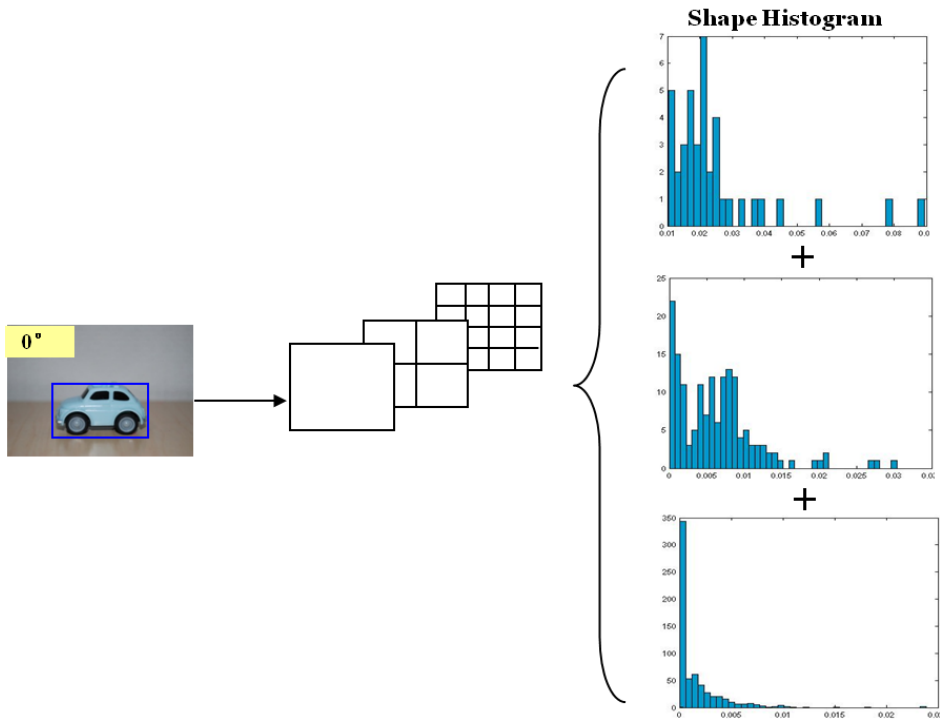


Fig.3. 3: A PHOG representation of a car instance at pose  $0^\circ$ . The ROI region is labeled with blue bounding box. With using a three-level PHOG feature, a shape histogram of a ROI is a concatenation of histogram described at each level.

### 3.3.3 Pose Estimation

Since the key pose structure is quantized into 16 poses. The  $i^{th}$  pose is denoted by  $p_i \in \{\alpha\}$  and  $p_0$  represents a frontal view object. We compute the azimuth angle  $\alpha$  for each image with respect to the canonical pose (frontal view).

We then use a Naïve Bayes classifier to learn the mapping from the PHOG feature to the probability of each key pose. Naïve Bayesian Classifier is a simple classifier that can be considered as maximum a *posterior* for a generative model. We then apply the rule to learn the distributions of PHOG feature from each key pose.



$$P(p_i | \text{ph}) \quad (3.1)$$

where  $\text{ph}$  represents the PHOG computed in the given bounding box. It is obtained by concatenating the histograms from all regions inside the bounding box,

$$\text{ph} = [\text{ph}^0, \text{ph}^1, \text{ph}^2, \text{ph}^3, \text{ph}^4, \dots, \text{ph}^m] \quad (3.2)$$

where  $\text{ph}^0$  is the histogram covering the whole bounding box,  $\text{ph}^1$  to  $\text{ph}^4$  are the four histograms that are computed on the second level, and so on. The mapping between the pyramid histograms and the object pose is approximately calculated by

$$P(p_i | \text{ph}) = \prod_0^l P(p_i | \text{ph}) \quad (3.3)$$

At run-time, given a bounding box in the image we compute the pyramid histograms and then use the learned mapping to estimate a distribution on the key pose. For the sake of simplicity, we take the object pose to be the one that maximizes the probability for the corresponding pose.

### 3.3.4 Adjusting Front Orientation

Since the pose variation from its frontal view pose can be estimated by

$$\alpha = \underset{p_i \in \alpha}{\operatorname{argmax}} \{P(\text{ph} | p_i)\} \quad (3.4)$$

the transformed *front* axis is the result of the rotation from the canonical position. Then the problem can be treat as the axis rotation problem, see Fig. 3.4. Assume a point in original coordinate is denoted as  $m(x,y)$ , then the transformed  $x$  and  $y$  coordinates can be represented by

$$\begin{cases} x_{\text{tran}} = m_x \cos \alpha + m_y \sin \alpha, \\ y_{\text{tran}} = m_y \cos \alpha - m_x \sin \alpha. \end{cases} \quad (3.5)$$

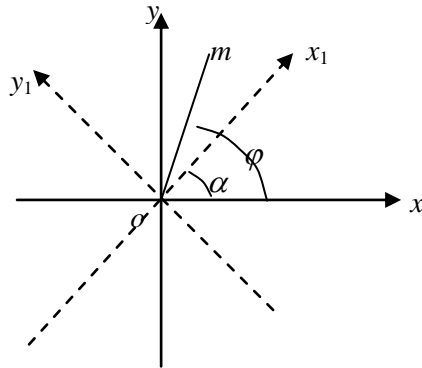


Fig.3. 4: axis rotation

## 3.4 Experiment

### 3.4.1 Pose Estimation Result

We use PHOG descriptor within the range 0 to 360-degree into 40 histogram bins stored in cache and Naïve Bayesian classifier for each sampled window. Fig. 3.5 shows some representative results. We also evaluate the performance with depicting the *precision/recall* curves drawn for the author's dataset in Fig. 3.6.

In our experiment, the PHOG feature can capture shapes of objects effectively and thereby is noteworthy for  $l = 2$  (0, 1, 2): with training examples, our method achieves an average performance of 93%, with the best results being over 98%. For cars, cellphones, and cameras, the accuracy is close to 91%, which make sense as these objects are more boxy. For shoes, race car and spoon, performance is average. We see the least successful cases gain for nozzle bottle and toothbrush, both of which are poorly captured with a rectangular window in that the front piece and the body of can be totally viewed as two parts. Besides, we found that the main pose confusion pairs are those off by 180-degree, for example, *front vs. back*.

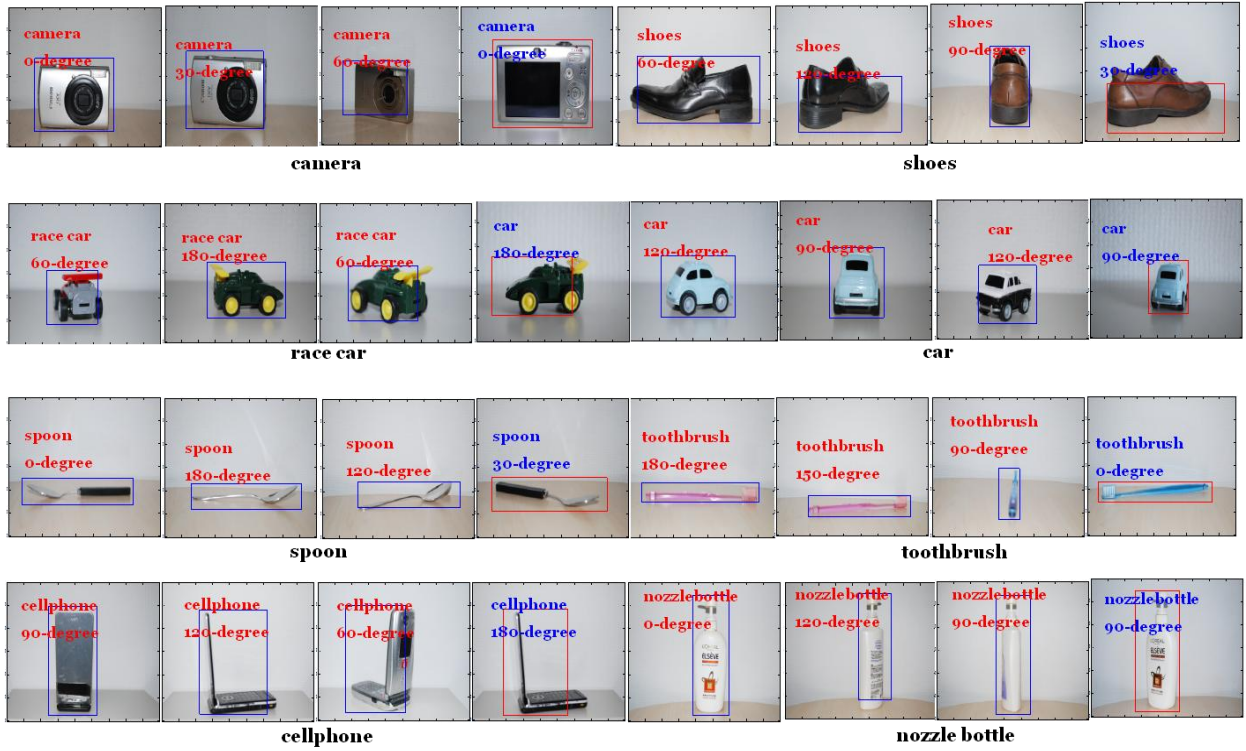


Fig.3. 5: Example results of pose estimation. The last column of each category is the false estimation of category.

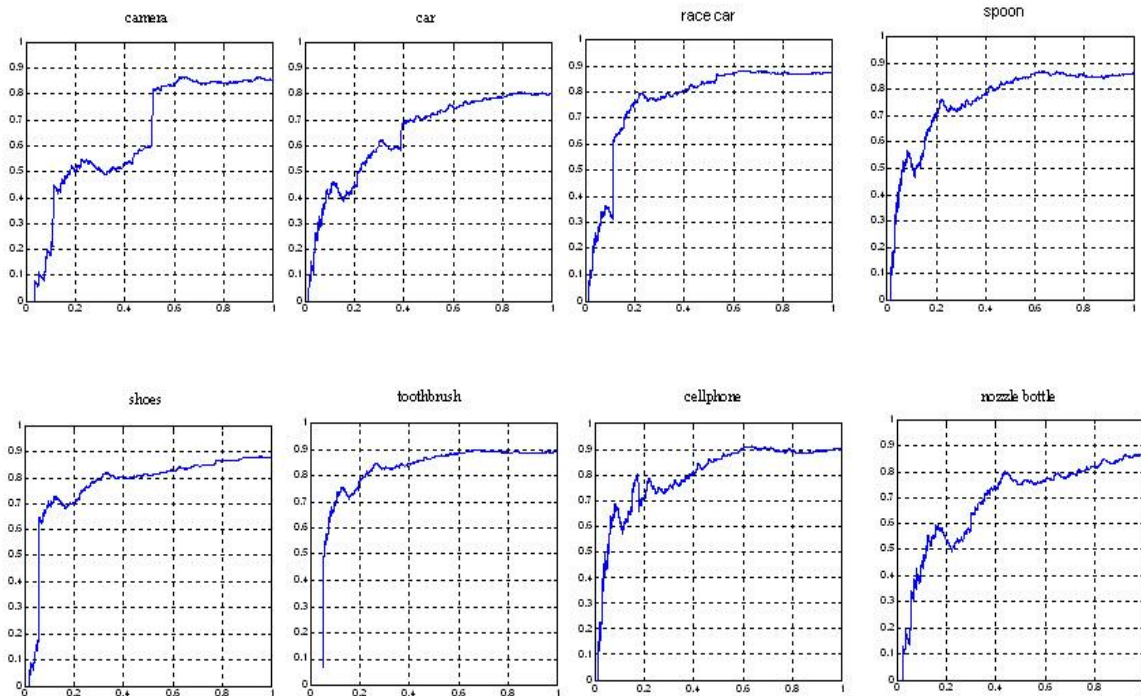


Fig.3. 6: precision recall curves of 8 categories

We compare the performance with M. Ozuysal[71] and Gu’s[35] model. The mean-average accuracy of ours in Table 2 is 80.1%, obtaining the second place. We observe that ours outperforms Ozuysal’s except the performance in car category. It is no coincidence that our results outperform his model. We use PHOG feature that can capture spatial distribution of edges. Moreover, in his model, rotation is limited to in-plane rotation on the ground. Gu’s model apply a discriminative template for pose classification based on a part-based templates [27]. The discriminative learning can address the classification problem directly and is very powerful in exploring noisy image data.

		Cameras	Computer Monitors	Shoes	Cars	Cellphones	Staplers	Nozzle Bottles	Mouse	Mean
Pose estimation accuracy(%)	M.Ozuysal's	93.1	80.8	85.4	95.3	60.8	63.3	71.2	65.4	76.9
	Chunhui Gu's	95.1	91.2	76.7	93.0	70.0	66.8	77.8	71.2	80.2
	Ours	96.4	84.5	87.2	94.1	67.3	65.4	72.0	73.6	80.1

Table 2: Comparison with the two models on the author's database

### 3.4.2 Adjusting front orientation

Next, we test the performance of adjustment. Fig. 3.7 shows the result. The last row is the false adjustment results.

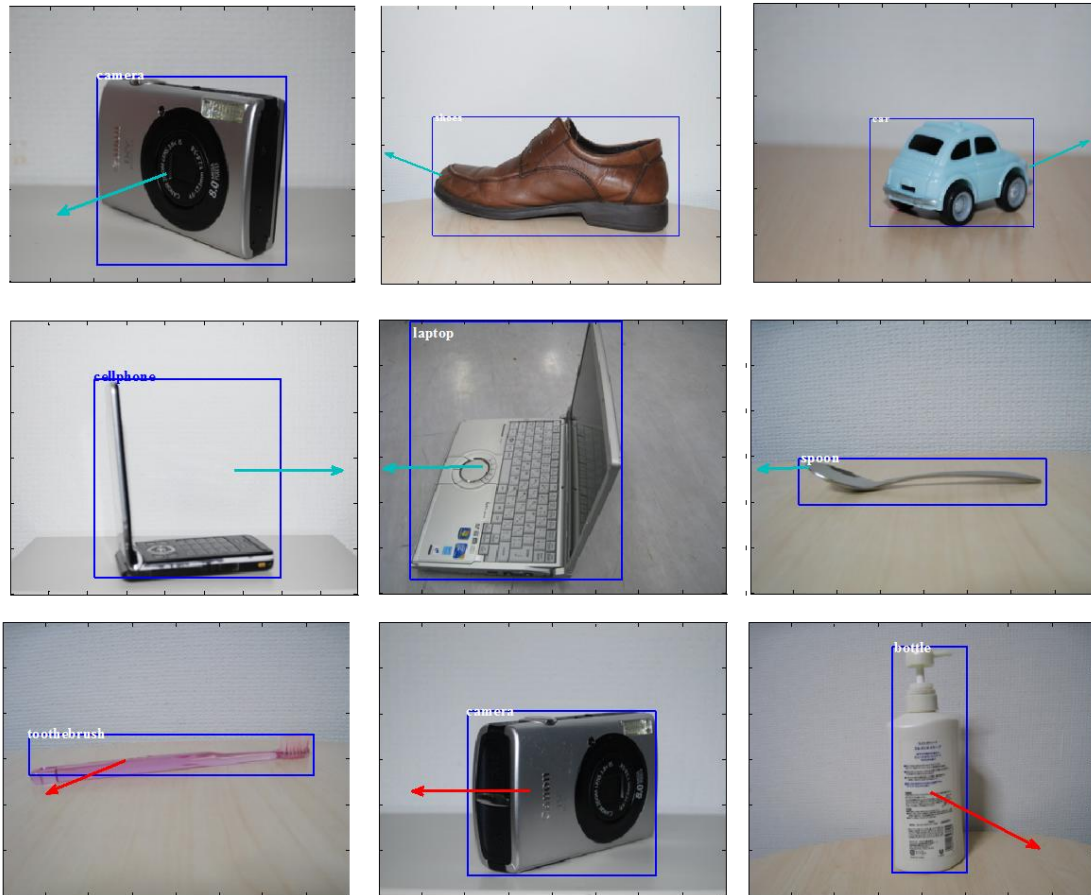


Fig.3. 7: “Front” orientation adjustment

### 3.4.3 Spatial Recognition Experiment

We use 100 images in the experiment. With using the same setting, the experiment is done in four steps: (1) detection, where we split the *detected (known)* and *un-detected (unknown)* objects; (2) indexing *unknown* objects, if multiple *unknown* objects are present; (3) estimating pose and adjusting the front axis if the *relatum* has *intrinsic* front; and (4) recognizing. Fig. 3.8 shows the result.



Fig.3. 8: Experiment results for recognizing spatial relation when adjusting “front” orientation

Overall, the approach was efficient and accurate in practice for all the images tested. On average, the accuracy is about 91% for singular object detection, and 81% for multiple objects. The response time was 0.2 seconds per round. Comparing to the cases with wide area of references (e.g. frontal-view of camera), the

accuracy was slightly decreased in the cases with narrow area (e.g. front piece of toothbrush). And it is important to notice that the most significantly “hardest” case is using compound expressions. In such cases, the accuracy was greatly decreased. Actually, in practice, it seldom utilizes compound expressions perhaps humans feel difficult to re-config the cognition map.

### **3.5 Conclusion**

In this chapter, we have introduced a pose estimator for adjusting front axis orientation extracted from the objects. We have acquired a database that contains continuous pose annotation and used it to train the estimator. By comparing their output to ground truth values obtained similarly, we showed that the estimator is accurate and stable. However, whether pose estimation can be performed reliably for an object category depends on the variation of image features as a function of object pose. The features, which are very significantly and in a statistically meaningful way, will be easier to design estimators and capture accurate posing information. Thus, a binary feature selection method will be discussed in the future.

Our approach only can estimate a limited number of pose. Therefore automatic ways to determine the extension of pose variation is a requirement for further developments. The latent variables used in deformable part-based model can represent a change in object pose and each individual part model can represent appearance from a separate viewpoint. As a result the model has a potential to be used in covariant pose detection and estimation of the viewpoint.



# Chapter 4

## Constructing the Database

*Database collection has been a critical part of computer vision research. A number of well collected databases have played an important role in visual recognition, such as Caltech 101/256[51, 34], PASCAL[25] However, none of them is suitable for spatial recognition tasks. In this chapter, we introduce a new database specially tailored for our experiment design. It is a two-fold dataset. The first part is used for specific object recognition tasks, which contains 122 objects from 24 categories. Each object is represented by 7 pose and slightly viewpoint and scale changes spaced evenly over the upper viewing hemisphere. In the second part of the database, there are in total of 400 scenarios where the learnt object appears together with several unknown objects. Using this database, we first learn and train the object model for recognizing specific object, then test it on the scenarios, separate the learnt and unknown. Our ultimate goal is recognize unknown objects collaborating with the user interface.*

### 4.1 Instruction

Our ultimate goal is recognizing *unknown* objects via spatial relation collaborating with user interface. But before that, the system is expected to learn and model some objects from training images, recognize those objects in multiple-object scenarios to separate *unknown* and *known* objects, and estimate pose transformation so that deducing and adjusting the *front* orientation once the *intrinsic* frame of reference is determined. Thus, unlike other *off-the-shelf* databases, orienting to single mission, our database should be eligible in 3 tasks simultaneously: (1) it should contain singular object samples for object recognition tasks; (2) the objects collected should cover multiple pose, and multiple views; and (3) it should provide scenarios in which consist of *undetected* and *detected* objects--at least, 1 recognized object and 1 undetected object.

## 4.2 Relative work

Database collection has been a critical part of computer vision research. A number of well collected databases have served as an important role in visual recognition. In this section, we introduce some databases directly related to our work.

As a number of well labeled small database, Caltech 101/256[51, 34], MSRC [86] have played an important role in visual recognition. Caltech 101 was essentially the first widely adopted dataset for object categorization. It has 101 categories and 9146 images in total. The PASCAL VOC datasets are a series of datasets used for the PASCAL Visual Object class Challenges. It contains 20 classes. The total number of images gradually is increased year by year.

Our dataset structure is similar to the ImageNet, but as a small one. ImageNet [21] is a large scale ontology, labeled image database. It uses a hierarchical structure and covers a wide range of visual concepts including animal, plant, artifact, geological formation, activities and materials. Currently, it consists of 21,841 concepts and there are a total of 14,197,122 images.

The most related to our endeavor, the ETH-80[46] database contains hemisphere views of 80 objects from 8 *basic-level* categories, but only singular object within, no multiple object scenarios. Moreover, by extending the basic level concept, we classify objects at 4 semantic levels, from the highest abstraction to the lowest exemplar.

Since there is no publicly database available, we aim to offer a new database both applicable to visual recognition and spatial recognition tasks simultaneously.

## 4.3 Collecting Candidate Objects

In this section, we present a new database with  $400 \times 300$  pixel resolution color images. All images are taken with a Canon IXY 910IS digital camera. The database is twofold: single-object images are for visual recognition. And multiple-object scenarios are for spatial recognition. We first describe how construct the database and link the concept into a semantic hierarchy for visual recognition tasks. This is the training and

evaluation benchmark. Then, we describe how we organize objects to set scenarios for spatial recognition tasks.

### 4.3.1 Collecting Candidate Objects for Visual Recognition

#### 4.3.1.1 Method

At the first step, we aim to collect objects for object recognition and pose estimation tasks. It is important to note that the objects we collected are not restricted to the categories because it does not exist per se in the world [46]. In fact, the conception of ‘category’ is a learned representation [78] and extremely relying on experience. Rather, the *basic-level* categories [46, 78], which, have shown that it is widely used in human categorization at which most knowledge is organized [78]. Taking an example from Brown’s work, a dog can not only be thought of as a *dog*, but also as a *boxer*, a *quadruped*, or in general a *animate being* [11]. Yet, *dog* is the term that comes to mind most easily, which is by no means accidental.

In addition, we explicitly do not intend to model functional categories, e.g. “*things that you can sit on*”, Even though those categories are important, they exist only on a higher level of abstraction and require a high degree of world knowledge and experience living in the real world [46]. Rather, we are interested in functional properties, e.g. “*things have intrinsic side*”. According to our empirical evidence, objects have *intrinsic front* and *back* side are more common in our daily life. Because of the symmetry in the *left-right* dimension, *intrinsic left* and *right* sides of objects are rare [100]. Thus, the most interesting case seems to determine *intrinsic front*. Once the *intrinsic front* is identified, the *back*, the *left* and the *right* side can be deduced accordingly. Previous works [5, 16, 23, 40, 55, 60, 92, 93, 96] have listed some principles to shed light on determination of *intrinsic front*. Those objects can be:

- The *intrinsic front* is the side lying in the direction of motion.
- The *intrinsic front* is the side containing the perceptual apparatus.
- The *intrinsic front* is the side which is most frequently used.

We construct the database via ontology concepts, which is a semantic structure supported by the asset of WordNet [26], a lexical database of English. Each meaningful concept in WordNet is explained by multiple words or phrases, is called “synset”. They are not only described in linguistics, but also in cognitive science. The synsets of images in the database are interlinked by two types of relations. Similarly to WordNet, the “is-a” relation is the most comprehensive and useful. Moreover, the functional property is associated with the “has-a” relation, which can semantically describe the meaning such as **a computer display has intrinsic-front side**.

The database organizes the different classes of objects into a hierarchy structure. It contains 3 levels. Apparently, *basic-level* category is the easiest for humans. Although it may be impossible to learn all the unique *basic level* for every object, there are objects that have become so much part of our daily life that their *basic level* is well-defined almost all over the world [46], e.g. *apples, cars*, etc. According to Rosch et al. [78] and Lakoff [44], this *basic level* is also

- The highest level at which category members have similar perceived shape.
- The highest level at which a single mental image can reflect the entire category.
- The highest level at which a person uses similar motor actions for interacting with category members.
- The level at which human subjects is usually fastest at identifying category members.
- The first level named and understood by children.

Subordinate category used in object identification can be found next to it. But we consider the superordinate categories which require a higher degree of abstraction and world knowledge as our starting point. For example, in Fig. 4.1, container is the highest level. The *basic-level* category can is adhering to it. The next lower level is a finer division in which we chose to include soda, cola, coffee and beer.



Fig.4. 1: A semantic structure of the container hierarchy

#### 4.3.1.2 Collecting the Candidate Objects

In our work, we intend to explore categorization for both nature and human-made objects. Inspired by ETH-80 and ImageNet, we include objects from six superordinate categories: “cloth & shoes”; “container”; “fruit & vegetable”; “instrument”; and “vehicle”. Fig. 4.2 lists the semantic structure in the current version of the database.

In principle, there are two ways of building a database. A category can either be set up by a representative distribution of member objects reflecting their probabilities of occurrence in practice, or by a few prototypes that approximately span the category [82]. We resort to the second option. Each object is represented by 42 images from pose spaced equally over the upper viewing hemisphere and slightly viewpoint and scale changes. Currently, the database contains 132 objects collected from 30 categories and a total of 5,124 images. Fig. 4.3 shows a snapshot of 2 branches of the container superordinate category.

#### 4.3.2 Designing Scenarios for Spatial Recognition

*Relatum* is the basis in establishment of spatial relation. In order to design quality and reasonable scenarios, we need to explore the question of what objects serving as *relatums*. In cognitive usage, a latent property governs the choice of *relatum* is the visual salience. Visual salience of a *relatum* depends on the interaction of basic features like size, shape and color, correlated to the corresponding attributes of the surrounding objects. As noted by [60, 96], if the objects are unequal in size or mobility, the larger and more stable is invariably encoded as the *relatum*. Consider the example in Fig. 4.3. This appears to be a predominant way of expressing spatial relations. If someone exchanges *referent* and *relatum*, saying **the jar is to the left of the candle**, it produces an odd-sounding result

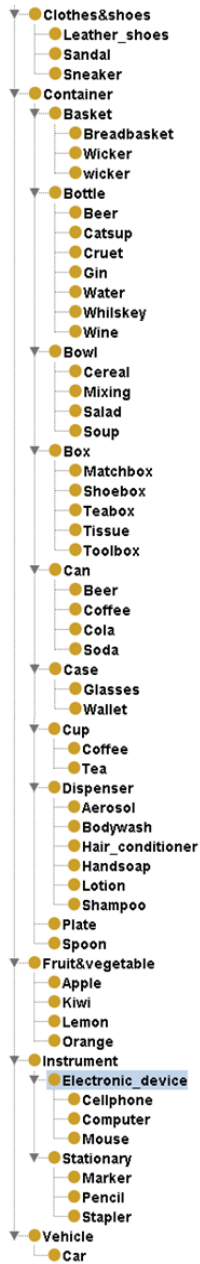


Fig.4. 2: Semantic structure in current version of the database

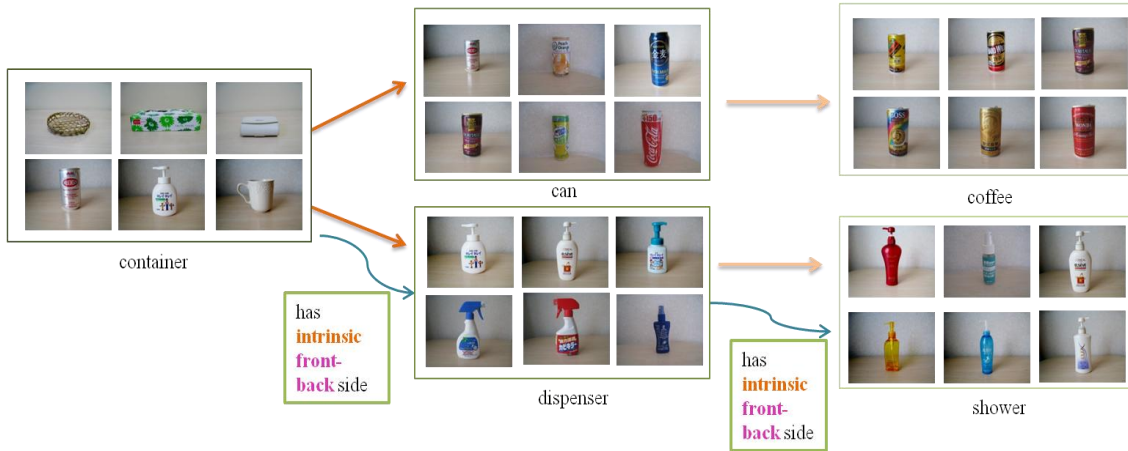


Fig.4. 3: A snapshot of the container abstract class containing two branches: the top row is from the can category; the bottom row is from the dispenser category.



Fig.4. 4: Designing scenarios with adequate relatum. The larger and more stable is usually treated as a relatum. If someone exchanges *referent* and *relatum*, saying the jar is to the left of the candle, it produces an odd-sounding

Currently, we have collected 720 scenarios in total with 1-3 unknown objects being placed around and 1-2 detected objects nearby. All the detected objects are basically at their frontal view. Example scenarios are shown in Fig. 2.6, 2.7, 2.13, 2.14, 2.16, and 2.17.

#### **4.4 Conclusion**

In this chapter, we introduce a new database specially tailored for the experiments. The database is built upon ontology structure, which can be easily expanded new domain or instance for the long run. We elaborate the criteria for us to collect the candidate objects. Then we describe how we design scenarios for spatial recognition tasks. We have conducted experiments on the database. In the next chapter, we introduce the experiments in interactive mode on the database.



# Chapter 5

## Interactive Object Recognition

*Having explained the structure and operation of the integral system, and validated on a variety of categories for each model, now we evaluate the performance on the proposed dataset experiments. We also perform experiments to make a comparison between of the original and improved models.*

### 5.1 Integral System Overview

We conduct experiments on the integral system. It contains a user interface, an object detector and a pose estimator and an object spatial recognition model. We use the detector provided by Dipnkar[19]. An untrained image is first served by an object detector, where pre-learned object is detected. If the target object is still undetected, human users manually annotate the object and instruct the system to recognize it via spatial relations. User will benefit from a user interface, which requires typing simple instruction. The interface can systematically analyze semantic components, separating the *referent* as well as that of the *relatum*, and the spatial relation between them. Corresponding to the object information recorded in the database, this step can make the system clear which frame of reference template should be taken. Once the *intrinsic* frame of reference is determined, we match the pose transformation and adjusting the *front* orientation.

### 5.2 The role of Natural Language

We use a user interface to control the system. The interface is developed to understand some simple English words and grammatical structures, such as **bring**, **how many**, **what** and **which**. To adopt this system to group use, the input toolbox was further developed to understand *noun-s(es)* rules. We restricted user input

commands to the following set format, e.g. *Bring me (referent), or It is (spatial expression) the (relatum)*

Table 3 lists the semantic form and grammatical structure.

Grammar used in the Transcript	
noun +s (es)	Plural noun
Can <div style="border: 1px solid black; padding: 5px; display: inline-block;">                     Yes, I can                      No, I can't                 </div>	Modal verb - positive answer - negative answer
noun[target object]+ <i>be</i> + preposition +noun(-s/es)[reference object]	Declarative sentence, indicating position of the target object
a/an/the/that + noun	Article, which combines with a noun to indicate the type of reference being made by the noun

Sentential Form used in the Transcript	
Can you see	Suitable for single reference object, asking whether the robot can “see” the target object at the beginning
How many	Suitable for multiple reference objects, asking numbers of detected objects when two negative answers response.
What are they	Suitable for multiple reference objects, inquiring about name of the detected objects by users
Bring/Get/ Give + noun	Verbs, which indicate motion starts.

Symbols used in the Transcript	
,	Slight pause
.	Full stop at the end of a declarative sentence
?	Ending for a sentence which indicates a question

Table 3: Grammar, symbols and linguistic form used in the interaction

### 5.3 Experiment 1: close linguistic form

In this set of experiments, the model collaborated with an interaction interface over 600 images. The experiments are done in 3 main steps: (1) detection, where using the same setting as in the Experiment 1; (2) annotating, where labeling *unknown* objects by users; (3) interaction, where users are required to stated intentions via input instructions if target objects are not detected yet. We note that users can interact with the system for several times till all the objects are detected if more than one object is in the image. And it is also allowed to use the objects which have been detected previously serving as *relatums* in the new round.

30 university students who are trained to familiar with our system participant the experiment. They are required to sit in front of a computer, and received 20 images at each round. Close linguistic form means we restrict the syntactic form format, but users can develop their own strategies thorough the procedure, for example, how to refer objects. During the instructions, the users are not directed as to which expression should or should not be used. The interface is able to process prepositional phrase including a projective term, such as “*to the left of*” or — simpler—“*left side*”. For utterances of compound expressions, we consider two possibilities. One is, combining two canonical expressions, e.g. “the (*referent*) is in front of and to the left of the (*relatum*)”; another one is using the composites, e.g. “the (*referent*) is to the left-front (side) of the (*relatum*)”. When the instruction could not be interpreted by the system, the users received the response like **I don’t understand**. They then try further to make their instructions understood. If users do not succeed till the number of processing exceeds the number of objects in the image, the system skipped and moved to the next image. Such a new start often encouraged the participants to reconsider their strategies [6]. The most substantial advantage of the experimental design is it guaranteed that the users chose the frame of reference which is most suitable for the query without any external disturbance. In addition, if more than one *referent* is detected at one acceptance region and therefore can share the same spatial relations in relation with a *relatum*, the strategy is enquiring whether it is the target object one by one.

The way of determination of group use is to detect whether a sentence contains the word **group**, or plural noun. If there is, the system enquires the number of group.

The experiments are conducted for 3 times and we collected a total of 3,352 instructions, corresponding to an average of 167.6 per person. Among the instructions, 436 utterances (13% of the total) cannot be interpreted because of thoroughly syntactic form or spelling errors. Of the 2,916 validated instructions, 2,450 (approximately 84%) lead to success, 466 instructions (16%) are not be interpreted successfully because the users either confused right and left regions, or used their own viewpoint.

Of every 20 images test by each participant, 5 are randomly selected to report *how many* and *what* objects have already “seen” at the beginning. Results therefore yield with respect to whether provide a scene description. Among the 783 validated instructions generated after a scene description is provided, the precision is as high as 88%. Meanwhile, 1533 validated instructions are produced without a scene description with the precision is slightly decreased to 80%. The reason for failure in most unsuccessful instructions is users named the objects that the system could not understand and perceive.

Regarding instructional strategy, when the scene description is provided, most of the users would like to directly refer the *referent*. For example:

**System:** I can see a book and a stapler.

**User:** The CD is *to the right of* the book.

If the scene description is not available, instead of directly referring the *referent*, users opt for other objects for help. This reflects user’s intention for finding out what is understandable for the system, for example, users sound out the system by asking “**Can you see the book?**” Fig. 5.1 and Fig. 5.2 show the complete transcripts with the scene description provided and unprovided. In the use of group-based frame of reference, we use 2 strategies. Fig. 5.3 and Fig. 5.4 show two strategies adopting **internal** and **external** manners, respectively. In the former case, we use the incremental strategy to guide the system, which make users more easily to be clear about how much the system understands the scene context. Whereas we apply a more

straight way to ask the system whether can find the group. If the system fails to find group of objects more than 3 times, it reports to the user that can't see the objects.

Non-scene description strategy

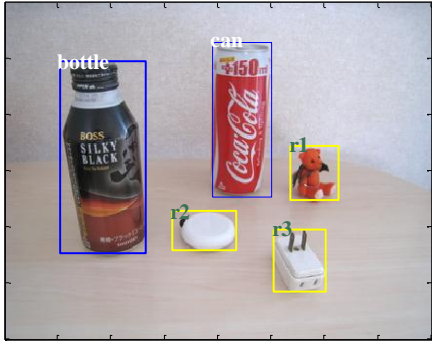
Scene1	Transcript1
	<p><b>User:</b> Can you see the <b>tape</b>?</p> <p><b>System:</b> No. I do not know what it is.</p> <p><b>Where</b> is it?</p> <p><b>User:</b> Can you see the <b>bottle</b>?</p> <p><b>System:</b> Yes, I can.</p> <p><b>User:</b> The <b>tape</b> is <b>at the right</b> side of the <b>bottle</b>.</p> <p><b>System:</b> Is this one?</p> <p><b>User:</b> Yes.</p> <p>-----</p> <p><b>User:</b> Bring me the <b>plug</b>?</p> <p><b>System:</b> No. I do not know what it is.</p> <p><b>Where</b> is it?</p> <p><b>User:</b> It is <b>to the right - front</b> the <b>tape</b>.</p> <p><b>System:</b> Is this one?</p>

Fig.5. 1: transcript without scene description given

Scene description strategy

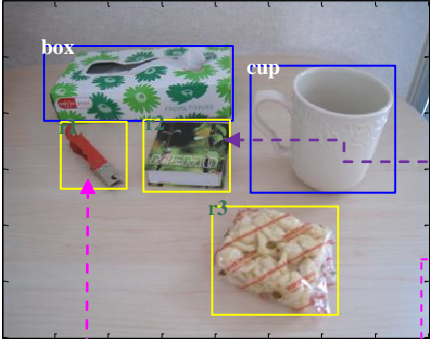
Scene1	Transcript2
	<p><b>System:</b> I can see 2 objects. Cup and box.</p> <p><b>User:</b> A can opener is in front of the box. Can you see it?</p> <p><b>System:</b> I see two. Is this one?</p> <p><b>User:</b> No.</p> <p><b>System:</b> Is this one?</p> <p><b>User:</b> Yes.</p> <hr/> <p><b>User:</b> Bring me the cake. It is <i>at the left -front side</i> of the cup.</p> <p><b>System:</b> Is this one?</p> <p><b>User:</b> Yes.</p>

Fig.5. 2: transcript with scene description given

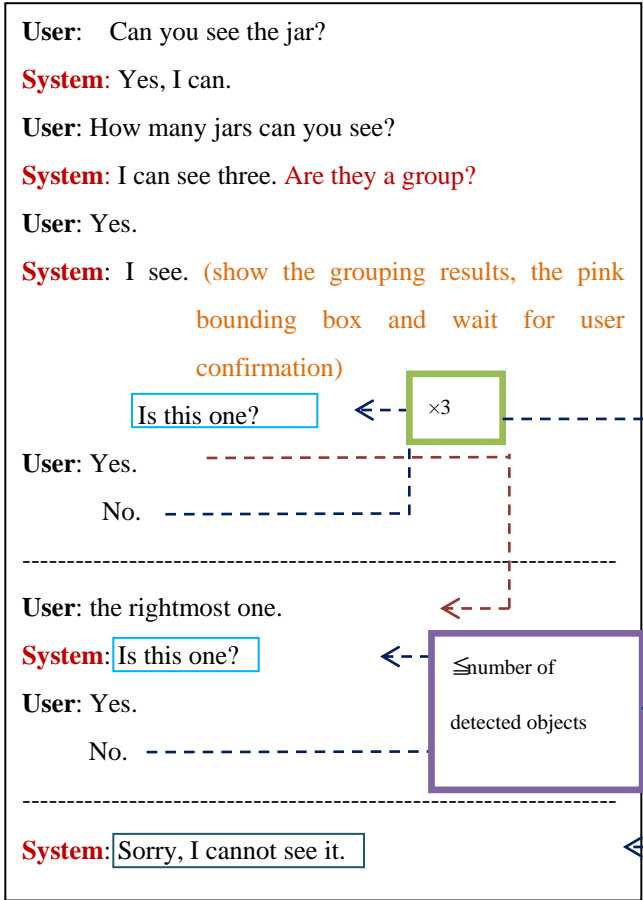
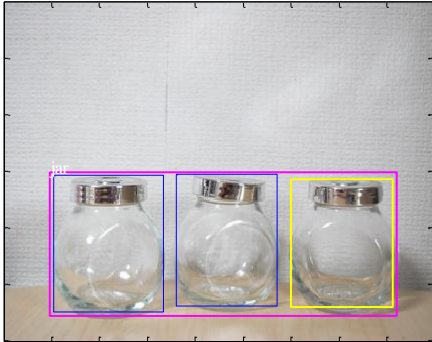


Fig.5. 3: Instructional strategy in internal case in the use of group-based frame of reference

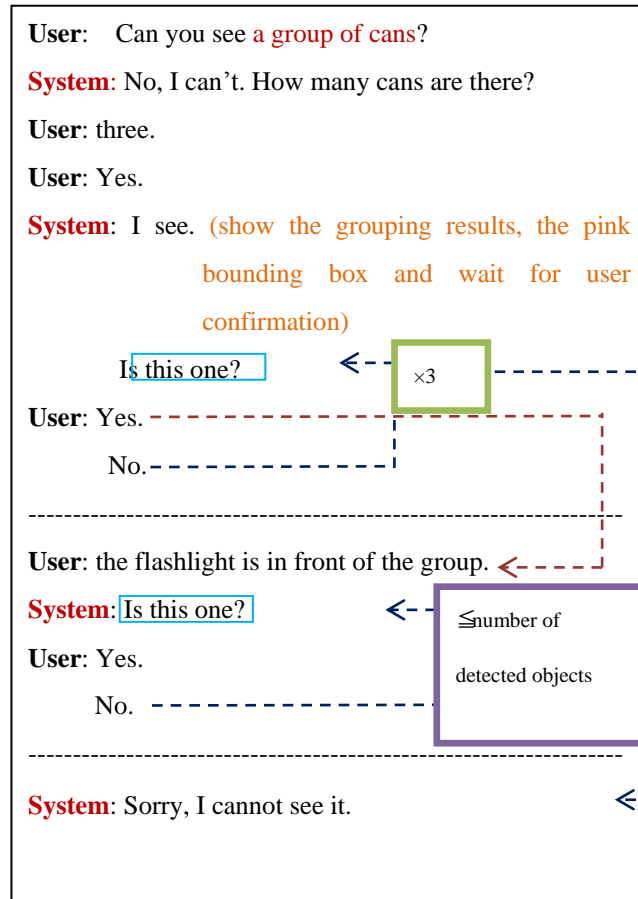


Fig.5. 4: Instructional strategy in internal case in the use of group-based frame of reference

We also calculate the search time. The system is implemented in Matlab and C/C++. The search time, which excludes the interaction time with human users, is measured from the instruction the user provides to the first sliding window shown. In average, operating on  $400 \times 300$  pixel images, an object is detected in 1.7 seconds on 2.80GHz Intel® Core(TM) i7 CPU. The kinds of objects do not influence the result, but their sizes do.



Our results show two interesting facets. One is users prefer to use incremental, e.g., *the (referent) is to the left of the (relatum)*, rather than goal-based instructions, e.g. *the left (referent)*, which is believed as easier for the system. However, the restricted linguistic setting that users employ somewhat limits the range of instructions, a major long-term goal is to employ a broader range of instructions, namely, unrestricted linguistic forms. The other one is for speakers, composites is more frequently used than canonical expression, although for the listener the canonical expression is more useful than the composite expression. The two observations are extremely valuable for us to investigate how spatial configurations influence instructional strategy and achieve natural, unconstrained communication in the future.

#### **5.4 Experiment 2: Comparison with the original model**

Finally, we compare the modified approach with previous version with implementing the compound template into the original one. 30 images from which have been used in previous experiments are selected. The test images consist of 2-3 *referents* and 1-2 *relatums*. 6 university students who have taken part in the second experiment are invited. They are required to run the two models on the same computer in turn. We first run the old one and record the validated and well-understood instructions, then the new model with the same instructions. The system reports to user *what* objects could be seen as long as the scene images are shown at the beginning. Fig. 5.5 shows 7 exemplars. The performance of our method, denoted as Ours-new and Ours-old, respectively in Table 4. As a result, our improved model gains superior performance than the previous one. All objects were localized successfully with users comments by the new model whereas only 1 test scene were achieved success by the old one.

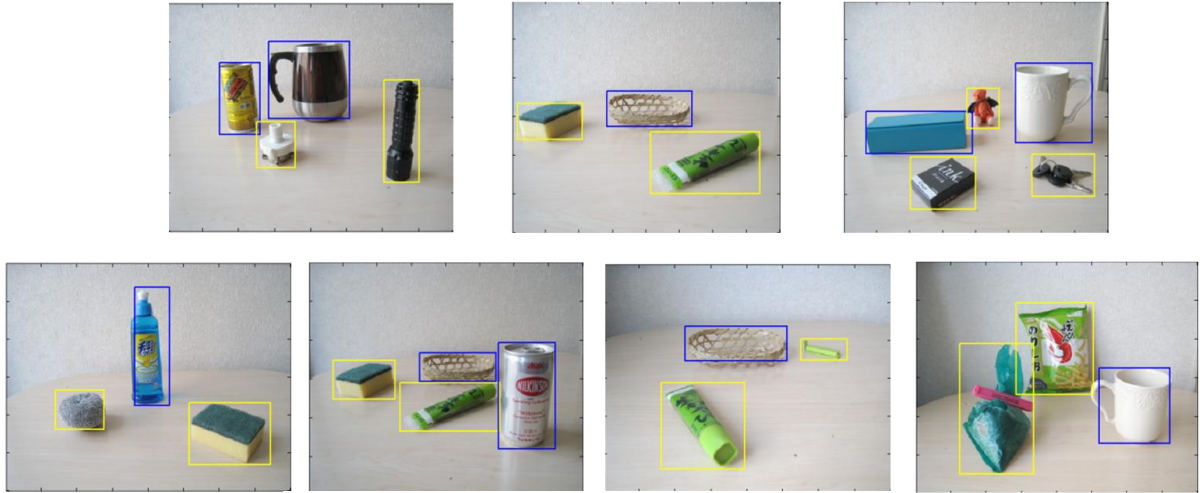


Fig.5. 5: 7 example scenarios for comparative experiments

Scenarios	Referents	Relatums	Used spatial expression	Recognition method	
				Ours-new	Ours-old
Scene 1	component	Coffee	right front	√	×
	flashlight	Mug	right	√	×
Scene2	wasabi	basket	right	√	×
	kitchen pad	basket	left	√	√
Scene3	teddy bear	case	right back	√	×
	ink	case	front	√	√
	keys	cup	front	√	√

Scene4	metallic scrub	bottle	left	√	×
	kitchen pad	bottle	right front	√	×
Scene5	wasabi	soda	back	√	√
	kitchen pad	basket	left	√	×
Scene6	wasabi	basket	front	√	√
	bag clip	basket	right	√	√
Scene7	green bag	cup	left	√	√
	snack	cup	left back	√	×

Table 4: the results of comparison the approach with previous model

## 5.5 Failure Case Study

We analyze the failure cases and found that the majority of error occurred in the *front* and the *back* cases. We ascribe the problem to the occlusion. The stricter *front/back* the *referent* situates, the worse the performance is, especially, when *referent* and *relatum* are almost collinear. In Fig. 5.6(a), the near one—*referent* often obscures the far object—the *relatum*, it can give rise to the insufficient coverage from the feature detector in the form of skew bounding box. In Fig. 5.6(b), on the contrary, the *referent* is hidden at the back side of the *relatum*, which results in an inaccurate annotation. Thus, we intend to employ range data method which is derived from images where the data is range or distance rather than intensity.

Moreover, the one-by-one-requiring instructional strategy doesn't practicable where 5 or even more objects are accumulated together. According to our experience, perhaps 3 or 4 objects are the maximum. A

conceivable way of remedying the problem is to clarify more details of the relevant *referent*, for example, color, which is believed as the most directly perceived information to identify objects.



(a)



(b)

Fig.5. 6: Failure case study: occlusion. (a) referent obscures the relatum; (b) the referent is hidden at the back side of the relatum

## Chapter 6

### Conclusion

In this work, we proposed an approach of integrating with a linguistic interface for object localization tasks. The goal is to offer a simple and natural way for human users to instruct a robotic system to locate intended objects. The core is a geometrical mapping function between spatial expressions and characteristic points on a reference plane. On the basis of previous work, we examined the existed problems and resolve the model on a 3-D plane. The modified model, on one hand, is flexible enough to ensure *referent* is always in the acceptance region in association with *relatum* irrespective of distance between them. On the other hand, the refined acceptance regions are designed to be capable of interpreting projective expressions, not only singular forms but also the composites. To validate our approach, we built a novel database specifically tailored to the task. We tested our approach on our database. The database is organized by its ontology concepts. We've shown convincing results that the ability to accurately localize objects obeying users' intension.

Since in nature scenes, objects are usually arbitrarily placed, a measure that would allow adjusting main axis direction when relevant objects are at an arbitrary pose or viewpoint would be an interesting direction, Perhaps finding *front* is the most interesting and crucial case as it weighs the highest priority in all 4 canonical directions. In this thesis, we have applied a classifier for estimation pose transformation. It can learn and capture the characteristics of different poses and can be directly used for pose classification. The main contribution is to show the applicability of finding virtue "front" orientation by the classifier. Our results are competitive on our database.

Although our work is still in a preliminary stage, we believe that our results are very important in visual recognition tasks. Our work discusses the problem from a novel perspective.

Possible directions for future work could include two domains. More candidate images with large scale and viewpoint changes should be collected for further experiment use. Furthermore, to address the occlusion problem, we opt for range data method for solution. Third, by integrating with other kinds of visual cues seeks for more to distinguish objects while a pair of spatial relation is available for several objects. Finally, we intend to carry out more experiments including more complicated configurations and diverse objects.

## Related Publications

- Journal (Under Review)

1. Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, " From 2-D to 3-D: An Integration of Spatial Recognition Model for Interactive Object Recognition" IPSJ of Information Processing.

- Lecture Notes

1. Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, "A Spatial-Based Approach for Groups of Objects", 8th International Symposium on Visual Computing (ISVC), Part II. LNCS, vol.7432, pp.597-608. Springer, Heidelberg (2012).

2. Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, "Spatial-Based Feature for Locating Objects" , 8th International Conference on Intelligent Computing(ICIC), LNAI, vol.7390, pp.128-137. Springer, Heidelberg (2012).

3. Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, "Spatial Relation Model for Object Recognition in Human-Robot Interaction" , 5th International Conference on Intelligent Computing(ICIC), LNCS, vol.5754, pp.574-584. Springer, Heidelberg (2009).

● International Conference

1. Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, "Object spatial recognition for service robots: Where is the front? ", 8th IEEE International Conference on Mechatronics and Automation(ICMA), pp. 875-880. Beijing, China (2011).

2. Md. Abdul Mannan, Hisato Fukuda, Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, "3D Free-Form Object Material Identification by Surface Reflection Analysis with a Time-of-Flight Range Sensor", 12th IAPR Conference on Machine Vision Application(MVA), pp. 227-230. Nara, Japan (2011).

3. Lu Cao, Yoshinori Kobayashi and Yoshinori Kuno, "Spatial Resolution for Robot to Detect Objects", 23th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.4548-4553. Taipei, Taiwan (2010).



## Bibliography

1. A. Abella, and J.R. Kender. Qualitatively describing objects using spatial preposition. Proc. of AAAI-93, PP. 536-540, Washington, DC, 1993.
2. E. André, G. Herzog, and T. Rist. On the simultaneous interpretation of real world image sequences and their natural language description: the system soccer. Proc. of the 8<sup>th</sup> ECAI, pp. 449-454, Munich, 1988.
3. M. Arie-Nachimison, and R. Basri. Constructing implicit 3D shape models for pose estimation. International conference on computer vision. 2009.
4. Y. Bao, Y. Xiang, and S. SAVERESE. Object co-detection. 12<sup>th</sup> European Conference of Computer Vision. 2012.
5. D. C. Bennett. Spatial and Temporal uses of English prepositions. London: Longman. 1975.
6. B.Bickel. Spatial operations in deixis, cognition and culture: where to orient oneself in Belhare. Language and Conceptualization, J.Nuyts, and E. Pederson(Eds.), pp. 46-83, Cambridge University Press. 1997.
7. I. Biederman. Recognition-by-components: a theory of human image understanding. Psychological Review, vol. 94, pp. 115-147, 1978.
8. S.Blisard, and M. Skubic, Modeling spatial reference language for Human-Robot Interaction, Proc. of the IEEE International Workshop on Robot and Human Interactive Commnication, Nashville, TN, 2005.
9. A. Bosch, A. Zisserman, and X. Munoz. Representing shape with spatial pyramid kernel. ACM International Conference on Image and Video Retrieval, pp. 401-408. ACM Press, Amsterdam. 2007.
10. B.Brewer, and J.Pears. Frames of reference. N. Eilan, R. McCarthy, and B.Brewer(Eds.), Spatial representation: Problems in philosophy and psychology, pp. 25-30. Oxford: Blackwell. 1993.
11. R. Brown. How shall a thing be called? Psychological Review, vol. 65, pp. 14-21. 1958.
12. H. Bulthoff, and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc. of the National Academy of Science, vol. 89(1), pp. 60-64. 1992.
13. H. Chiu, L. Kaelbling, and T. Lozano-Perez. Virtual-training for multi-view object class recognition.

Computer Vision and Pattern Recognition. 2007.

14. A. Collet et al. Object recognition and full pose registration from a single image for robotic manipulation. International conference on robotics and automation. 2009.
15. M. C. Corballis. Recognition of disoriented shapes. *Psychological Review*. vol. 95, pp. 115-123. 1988.
16. M.J. Cresswell. Prepositions and points of view. *Linguistics and Philosophy*, vol.2 pp. 1-41. 1978.
17. C. Cyr, and B. Kimia. A similarity-based aspect-graph approach to 3D object recognition. *International Journal of Computer Vision*, vol. 57(1), pp. 5-22. 2004.
18. E.Danziger. Pars and their counterparts: spatial and social relationships in Mopan Maya. *J.R.Anthropol. Inst*, vol. 2, pp. 67-82. 1996.
19. D. Das, Y. Kobayashi, and Y. Kuno. Multiple object category detection and localization using generative and discriminative models. *IEICE Transaction on Information and System*, vol. E92 (20), pp. 2112-2121. 2009.
20. D. DeMenthon, and L. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, vol. 15, pp. 123-141. 1995.
21. J. Deng, W. Dong, R. Socher, L.Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. *Computer Vision and Pattern Recognition*, pp. 248-255. 2009.
22. R. Detry, N. Pugeault, and J. Piater. A probabilistic framework for 3D visual object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.31 (10), pp. 1790-1803. 2009.
23. R. Dirven. Spatial relations in English. In *Kasusgrammatic und Fremdsprachendidaktik*. G. Radden, and R. Dirven (Eds.), pp. 103-132. Wissenschaftlicher Verlag. Trier, West Germany. 1981.
24. M.J.Egenhofer, and J.R. Herring. A mathematical framework for the definition of topological relationships. K. Brassel and H. Kishimoto (Eds.), *Proc. of the 4<sup>th</sup> International Symposium on Spatial Data Handling*, pp. 803-813. Zurich, 1990.
25. M. Everingham et al. The PASCAL visual object classes challenge 2006(VOC 2006) Results. <http://www.Pascal-network.org/challenges/VOC/voc2006/results.pdf>.
26. C. Fellbaum. *WordNet: An electronic lexical database*. Bradford Books. 1998.
27. P. Felzenszwalb, R. Girshick, D. McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.32(9), pp. 1627-1645, 2010.

28. F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, vol. 5, pp. 1531-1555. 2004.
29. T. Fuhr, G. Socher, C. Scheering, and G.Sagerer. A three-dimensional spatial model for the interpretation of image data. P.Olivier & K.P.Gapp(Eds.), *Representation and Processing of Spatial Expressions*, pp. 103-118, Mahwah, MJ: Lawrence Erlbaum Associates, Inc.
30. K. P. Gapp. Basic meanings of spatial relations: Computation and evaluation in 3d space. *Proc. of AAAI-94*, PP. 1393-1398, Seattle, WA, 1994.
31. K.P.Gapp. A computational model of the basic meanings of graded composite spatial relations in 3d space. M. Molenaar and S. DeHoop (Eds.), *Proc. of the AGDM'94 Workshop*, pp. 66-79, Delft, the Netherlands, 1994.
32. K.P.Gapp. Angle, distance, shape, and their relationship to projective relations. *Proc. of the 17<sup>th</sup> Annual Conference of the Cognitive Science Society*, Pittsburgh, PA, 1995.
33. K. P. Gapp. An empirically validated model for computing spatial relations. *Proc. of KI-95*, Berlin, Heidelberg, Springer. 1995.
34. G. Griffin, A. Holub, and P. Perona. Caotech-256 object category dataset. Technical Report 7694, Caltech. 2007.
35. C. Gu, X. Ren. Discriminative mixture-of-templates for viewpoint classification. 11<sup>th</sup> European Conference on Computer Vision. 2010.
36. J. B. Haviland. Anchoring, iconicity, and orientation in Guugu Yimithirr pointing gestures. *J. Linguist. Anthropol*, vol. 3, pp.3-45. 1993.
37. J. B. Haviland. Guugu Yimithirr cardinal directions. *Ethos* vol. 26, pp. 25-47. 1998.
38. L. Henkel, and N. Franklin. Dividing and remembering surrounding space. Presented at the 33<sup>rd</sup> ANNUAL Meeting of the Psychonomic Society, St. Louis, MO. 1992.
39. A. Herskovits. *Language and spatial cognition*. Cambridge, UK, Cambridge University Press. 1986.
40. C. Hill. Up/down, front/back, left/right. A constrastive study of Hausa and English. *Her and There. Cross-Linguistic Studies on Deixis and Demonstration*, J. Weissenborn and W. Klein (Eds.), John Benjamins, Philadelphia. 1982.
41. J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychological Review*, vol. 99, pp. 480-517. 1992.

42. I. Kant. On the first ground of the distinction of regions in space. The philosophy of right and left. J. Van Cleve, and R.E.Frederick(Eds.), PP. 27-34, Kluwer. 1991.
43. J. Koenderink and A. van Doorn. The Internal Representation of solid shape with respect to vision. *Biological Cybernetics*, vol. 32, pp. 211-216. 1979.
44. G. Lakoff. *Women, fire, and dangerous things- what categories reveal about the mind*. Univ. of Chicago Press, Chicago. 1987.
45. S. Lavallee, and R. Szeliski. Recovering the position and orientation of free-form objects from image contours using 3d distance maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.17 (4), pp. 378-390. 1995.
46. B. Leibe, and B. Schiele. Analyzing appearance and contour based methods for object categorization, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409-415. 2003.
47. W. Levelt. Some perceptual limitations on talking about space. In A., J., Van Doorn, W., A., Van de Grind, and J., J., Koenderink(Eds.), *Limits in perception*, pp. 323-358. Utrecht: VNU Science Press. 1984.
48. S. C. Levinson. Frames of references and Molyneux's question: cross-linguistic evidence. *Language and Space*. P.Bloom. et al.(Eds.), pp.109-169. MIT Press. 1996.
49. S. C. Levinson. From outer to inner space: linguistic categories and non-linguistic thinking. *Language and Conceptualization*. J.Nuyts, and E. Pederson (Eds.), pp. 14-45, Cambridge University Press. 1997.
50. S. C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*, Cambridge University press. 2003.
51. L. Fei-Fei, R.Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 594-611. 2006.
52. J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. *Computer Vision and Pattern Recognition*. 2008.
53. G. Logan, and D. Sadler. A computational analysis of the apprehension of spatial relations. *Language and Space*, pp. 493-529. MIT Press. 1996.
54. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60(2), pp. 91-110. 2004.
55. J. Lyons. *Semantics*, vol.2. Cambridge University Press. Cambridge, England. 1977.

56. A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3D object recognition. *Computer Vision and Pattern Recognition*. 2004.
57. J. MacQueen. Some methods for classification and analysis of multivariate observations, *Proc. of the 5<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297. 1967.
58. A. Majid, M. Bowerman, S.Kita, D.B.M. Haun, and S.C.Levinson. Can language restructure cognition? The case for space. *Trends in Cognitive Science*, vol. 8, pp. 108-114. 2004.
59. K. Mani and P.N.Jonson-Laird. The mental representation of spatial descriptions. *Memory & Cognition*, vol. 19, pp.181-187. 1982
60. G. A. Miller, and P. N. Johnson-Laird. *Language and Perception*. Cambridge. MA: Harvard University Press. 1976.
61. R.C.Mishra, P. R. Dasen, and S. Niraula. Ecology, language and performance on spatial cognitive tasks. *International Journal of Psychology*. 2003.
62. R. Moratz, H. Eikmeyer, B. Hildebrandt, A. Knoll, F. Kummert, G. Rickheit, and G. Sagerer. Selective visual perception driven by cues from speech processing. *Proc. of 7<sup>th</sup> Portuguese Conference on AI EPIA95, Workshop on Applications of AI to Robotics and Vision Systems*, pp. 63-72, Funchal, Madeira Island, Portugal. Trans Tech Publications Ltd.
63. R. Moratz, K. Fishcer, T. Tenbrink. Cognitive Modeling of Spatial Reference for Human-Robot Interaction. *International Journal on Artificial Intelligence Tools*, vol. 10, pp. 589-611.
64. R. Moratz, and T. Tenbrink. Instruction modes for joint spatial reference between naïve users and a mobile robot. *Proc. of RISSP IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, Special Session on New Methods in Human Robot Interaction*, Changsha, China. 2003.
65. R. Moratz, T. Tenbrink, J. Bateman, and K. Fischer. Spatial knowledge representation for human-robot interaction. C. Freksa, W. Brauer, C. Habel, and K.F. Wender(Eds.), *Spatial Cognition III*, PP.263-286. Springer, Berlin. 2003.
66. R. Moratz, M. Wünnstel, and R. Ross. Qualitative spatial arrangements and natural object categories as a link between 3d perception and speech. *Proc. of the 8<sup>th</sup> Pacific Rim International Conference on Artificial Intelligence*. 2004.
67. R. Moratz, and T. Tenbrink. Spatial Reference in Linguistic human-robot interaction: iterative,

- empirically supported development of a model of projective relations, *Spatial Cognition and Computation*, vol.6(a), pp. 63-107, Lawrence Erlbaum Associates, Inc. 2006.
68. A. Mukerjee, and G. Joe. A qualitative model for space. *Proc. of AAAI-90*, PP. 721-727, Boston, MA, 1990.
  69. S. Neumann, and T. Widlok. Rethinking some universals of spatial language using controlled comparison. *The Construal of Space in Language and Thought*. R.Dirven, M. Pütz (Eds.), pp. 345-372. De Gruyter. 1996.
  70. P. Olivier, T. Maeda, and J. Tsuji. Automatic depiction of spatial descriptions. *Proc. of AAAI-94*, PP. 1405-1410, Seattle, WA, 1994.
  71. M. Ozuysal, V. Leptetit, and P. Fua. Pose estimation for category specific multiview object localization. *Computer Vision and Pattern Recognition*. 2009.
  72. E. Pederson. Language as context, language as means: spatial cognition and habitual language use. *Cognitive Linguistics*, vol. 6, pp. 33-62. 1995.
  73. E. Pederson, E. Danziger, D. S.C. Levinson, S. Kita, and G. Senft. Semantic typology and spatial conceptualization. *Language*, vol. 74, pp. 557-589. 1998.
  74. E. Pederson. How many reference frames? *Spatial CognitionIII: Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning*, C. Freska et al., (Eds.), pp. 287-304, Springer Verlag. 2003.
  75. M.A.Peterson, F. J. Kihlstrom, P.M.Rose, and M.L.Glisky. Mental images can be ambiguous: Reconstruals and reference frame reversals. *Memory and Cognition*, vol. 20, pp. 107-123. 1992.
  76. R.Rajagopalan. A model for integrated qualitative spatial and dynamic reasoning about physical systems. *Proc. of AAAI-94*, PP. 1411-1417, Seattle, WA, 1994.
  77. F. Rothganger et al. 3D object modeling and recognition using local affine invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, vol.66 (3), pp. 231-259. 2006.
  78. E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, vol. 8, pp. 382-439. 1976.
  79. S. Savarese and L. Fei-Fei. 3D Generic object categorization, localization and pose estimation. *International Conference on Computer Vision*. 2007.

80. D. L. Schacter, L. A. Cooper, and S. M. Delaney. Implicit memory for unfamiliar objects depends on access to structural descriptions. *Journals of Experimental Psychology: General* vol. 119, pp.5-24. 1990.
81. J.R.J.Schirra, and E. Stopp. Antlima- a listener model with mental images. *Proc. of the 13<sup>th</sup> IJCAI*, PP.175-180, Chambéry, France, 1993.
82. S. Sclaroff. Deformable prototypes for encoding shape categories in image databases. *Pattern Recognition*, vol. 30(4), pp. 627-641. 1997.
83. G. Schmidt, Various views on spatial preposition, *AI Magazine*, vol. 9(2), pp. 95-105. 1988.
84. G. Senft. *Referring to space: studies in Austronesian and Papuan language*, Clarendon Press. 1997.
85. G. Senft. *Frames of reference in Kilivila*. *Studies in Language*, vol. 25, pp. 521-555. 2001.
86. J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. *9<sup>th</sup> European Conference on Computer Vision*. 2006
87. J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *9<sup>th</sup> International Conference on Computer Vision*. 2003.
88. J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. *IEEE Conference on Computer Vision and Pattern Recognition*. 2004.
89. J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping of video shots. *8<sup>th</sup> European Conference on Computer Vision*. 2004.
90. M. Skubic, P. Matsakis, G. Chronis, and J.Keller. Generating multilevel linguistic spatial descriptions from range sensor readings using the histogram of forces. *Autonomous Robots* vol. 14, pp. 51-69. 2003.
91. M. Skubic, D. Perzanowski, S.Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock. Spatial language for human-robot dialogs. *IEEE transactions on Systems, Man and Cybernetics, Part C*, pp. 154-167.2004.
92. N. K. Sondheimer. English as a command language for machines and the semantics of “Left” and “Right”. *Milwaukee Symposium on Automatic Control (and Autonomous Computing)*, Milwaukee, Wisc. 1974.
93. N. K. Sondheimer. Spatial reference and natural-language machine control. *International Journal of*

- Man-Machine Studies, vol. 8, pp. 329-336. 1976.
94. H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a Dense multi-view representation for detection, viewpoint classification and synthesis of object categories. International Conference on Computer Vision. 2009.
  95. M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3D object classes. Computer Vision and Pattern Recognition. 2009.
  96. L. Talmy. Figure and ground in complex sentences. In J.Greeberg, C.Ferguson, & E. Moravcsik(Eds.), Universals of human language, vol.4, pp. 625-649. Palo Alto, CA: Stanford University Press, 1978.
  97. L. Talmy. The relation of grammar to cognition: a synopsis. In D.Waltz(Ed.), TINLAP-2, pp.14-24. New York: Association for Computing Machinery, 1978.
  98. T. Tenbrink, and R. Moratz. Group-based spatial reference in linguistic human-robot interaction. Proc. of The European Cognitive Science Conference, pp. 325-330, Osnabruck, Germany. 2003.
  99. A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. van Gool. Towards multi-view object class detection. Computer Vision and Pattern Recognition. 2006.
  - 100.C. Vandeloise. Description of Space in French. Ph.D. diss., Dept. of Linguistics, Unive. Of California at San Diego. 1984.
  - 101.C. Vorweg, G. Socher, T. Fuhr, G. Sagerer, and G. Rickheit. Projective relations for 3d space: computational model, application, and psychological evaluation. AAAI'97, pp. 159-164.
  - 102.J. Wassmann, and P. R. Dasen. Balinese Spatial orientation: some empirical evidence of moderate linguistic relativity. J.R. Anthropol. Inst, vol.4, pp. 689-711. 1998.
  - 103.P. Wazinsk. Generating spatial descriptions for cross modal references. 3<sup>rd</sup> Conference on Applied Natural Language Processing, pp. 56-63. 1992.
  - 104.M. Webber, M. Welling, and P. Perona. Towards automatic object categories. IEEE Conference on Computer Vision and Pattern Recognition. 2000.
  - 105.M. Webber, M. Welling, and P. Perona. Unsupervised learning of object models for recognition , 16<sup>th</sup> European Conference on Computer Vision. 2000.
  - 106.T. Widlok. Orientation in the wild: the shared cognition of the HAI//OM Bushpepole. J.R. Anthropol. Inst, vol. 3, pp.317-332. 1997.
  - 107.H. D. Zimmer, H. R. Speiser, J. Baus, A. Blocher, and E. Stopp. The use of locative expressions in



dependence of the spatial relation between target and reference object in two-dimensional layouts. C. Freksa, C. Habel, K. F. Wender (Eds.), *Spatial Cognition, Lecture Notes in Artificial Intelligence*, pp. 223-240. Springer-Verlag, Berlin. 1998.