The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

# Spatial Resolution for Robot to Detect Objects

Lu Cao, Yoshinori Kobayashi, and Yoshinori Kuno, *Member, IEEE*

*Abstract* **In this paper, we report on our development of a robotic system that assists people in accomplishing simple tasks in daily life (e.g., retrieving objects for handicapped and elderly people). These tasks, inevitably involve detecting various kinds of objects. In particular, here, we present an interactive method to detect objects using spatial information. Our experimental results confirm the usefulness and efficiency of our system. We also show how the approach can be improved and highlight necessary directions for future research.**

## I. INTRODUCTION

When describing an object in interaction, people often use spatial relationships, e.g., “the spoon is in front of the bowl”, or give directives, e.g., “bring me the book on the left”. Linguistic studies show the importance of spatial representation in language [1]. Our analysis also shows that people tend to use spatial concepts in performing navigation tasks, particularly, when they encounter objects that are similar or unfamiliar in category [2]. Spatial expressions partition the space into loose regions such as near, back and left. The location of the objects as well as the partitioning of the space is specified with a large degree of ambiguity [3]. Herskovits in [4] lists many of the factors that may be important for the interpretation of spatial expressions. Since robots have a limited cognitive capability for identifying objects solely by their intrinsic properties, such as shape, size, and color, spatial relationship expressions are a necessary approach that can establish a correspondence between human users and robots.

In previous work, Moratz et al. [5] developed a model of human robot communication that utilized spatial relationships that represented landmarks or referenced the target objects, analyzed spatial expressions, and provided navigation of orientations in a plan view. Skubic et al [6, 7] presented a multi modal robot interface that utilized spatial relationships, providing linguistic communication from the robot to the user. These studies, however, predominantly focused on how to map the overall environments and facilitate directives with objects using spatial terms. For instance, they controlled the robot with motion instructions such as, “Drive up to the right cube,” or “Go to the nearest object in front of you.” Moreover, the vision system was not sufficient enough to distinguish complex object categories with multiple objects within a scene.

While our approach is inspired by their methods, it differs from theirs in using spatial relationship strategies. In our application domain, objects are usually more complex, and not so easily recognized by the robot. For example, we may want the robot to bring us a toy that has various colors among several toys. Further, it does not use complex expressions to describe such objects. In addition to gesture recognition, the robot needs to have the capability to detect and locate the object mentioned in speech. In this paper, we present an experimental approach using spatial linguistic terms to identify various objects. The robotic system should distinguish and therefore select an appropriate interaction strategy, depending on the situation. Since we have developed an object recognition system combining autonomous object recognition [8, 9], when autonomous object recognition cannot detect the object or make mistakes, the system turns to the spatial recognition mode. The strategy is that a spatial description which describes the target object using spatial terms to assist the robot to accomplish tasks. The user is allowed to type instructions into a computer instead of using verbal commands. Note that our proposed system does not recognize the user’s pointing gesture because if the user is a bit far away from the object, the target object might not be identified from the user’s pointing gesture alone even though the robot uses the pointing gesture to ask the user for confirmation. Moreover, to achieve an intuitive interaction similar to communication with humans, users can choose reference objects freely according to the objects that have been recognized by the robot.

## II. SPATIAL KNOWLEDGE REPRESENTATION

This section briefly introduces our work. Now that the robotic system has been assumed to recognize some object classes and specific objects in autonomous object recognition, how human users, who are being aware of the limited object detection capability of their robot partner, describe objects in images is of primary interest.

### *A. Intrinsic, Relative and Absolute*

When referring to the position of an object in relation to another, humans alternate different resources which lead to the identification of three different reference systems.

Levinson [10] has proposed that humans use three kinds of reference systems: intrinsic, relative, and absolute. In the intrinsic reference system, the relative position of one object (the referent) to another (the relatum) is described by referring to the relatum's intrinsic properties such as front or back. For example, the expression “the book is in front of you” is good enough to describe the position of the book since the front of a human body is intrinsically determined. However, in the relative reference system, we use a position of a third entity as the origin instead of referring to the inbuilt features of the relatum. An example is “viewed from the cup, the pen is on the left side of the box.” In the absolute system, neither a third entity nor its intrinsic features are used for reference. Instead, we use some

L.Cao, Y. Kobayashi and Y. Kuno are with the Department of Information and Computer Sciences, Saitama University, Saitama city, Saitama 338-8570, Japan (e-mail: { caolu, kuno, yosinori}@cv.ics.saitama-u.ac.jp).

978-1-4244-6676-4/10/$25.00 ©2010 IEEE

absolute direction specification terms, such as north and south.

We mainly consider the relative and intrinsic reference systems in our domain. The viewpoint is usually omitted when it is either the speaker or listener.

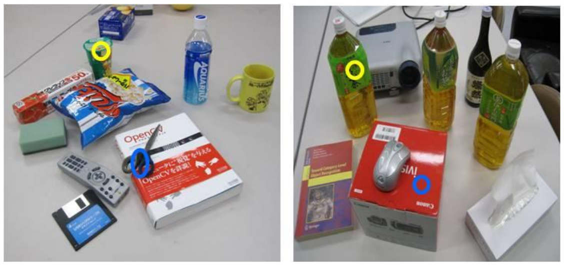
In addition to the above three reference systems, the group based reference system has been proposed by [11, 12]. This system is considered to be the relative reference system using the group as a reference object. When there are the same or similar multiple objects in the scene, humans consider them as a group, describing the position of an object in the group using the spatial relation between the object and the total group, such as “the second to left of those objects”. This notion of a group is extended in our work as well. There are cases suggesting we consider objects as a group even when they are different kinds if they are accumulated together.

### *B. Linguistic Spatial Terms Collection*

In order to adapt our robotic system, we conducted the following experiment to collect spatial terms.

**Participants** 90 participants in total: 50 Japanese, 30 Chinese, and 10 English native speakers, in Saitama University.

**Procedure** Participants were divided into 45 pairs (A and B) and requested to stand in front of 4 scenarios in turn. We asked participant A to describe an object, and asked participant B to choose the correct object. Participant A continued with the description until participant B selected the correct object. The only gesture allowed for the Participant B was to point to the object. Figure 1 is a representation of 2 scenario scripts out of the 4. For each scene we decided the target objects and Participant A could choose the reference object freely. The identity of target object (labeled by “o”) was shown only to Participant A in written form.



(i) (ii)

Figure 1. Two scenario scripts in experiment.

**Results** To obtain the terms many and various, throughout the experiment, we did not require the participants to employ the same instructions. We collected a total of 360 different linguistic instructions in total in our experiment. These are characterized into the following two types.

**Directional Instructions (272 occurrences)** [11] indicated that the position of the goal object is indentified as bounded linear oriented structures. The vocabularies used in relation to jointed verb and object varied, as did the positional adjective yielding front, back, on, under, left, right, leftmost, middle, between, and rightmost. In our experiment, the participants were not solely asked to specify the location of one object in relation to a different one, but rather, to specify the identity of one of several similar objects (Figure 2 (ii)). 174 utterances referred directly to the target object, such as “The box is under the mouse.” Another 98 utterances used the group as a relatum, such as “The leftmost bottle of the 3 bottles.” The listener’s intrinsic properties were used for instruction in 120 of the 360 instructions, using linguistic expressions such as “The book is in front of you.” Although the orientation of the robot is not stated explicitly in these commands, the speakers could not use an expression like “on your right side” without assuming the front of the robot [5]**.**

**Distance Instructions (88 occurrences)** Some users indicated the distance of the referent such as “The box is next to the bag.” 60 out of 88 occurrences applied the expression **next to**. The words near and far (28 occurrences) were least used in that the distance between two objects near or far principally depends on the location that the speaker stands in a scene and his or her viewpoint. If someone intends to explain this expression, he or she should ensure that the addressee employs the same viewpoint, or else, at least the partner can obtain sufficient information from his or her perspective.

In 36 of the 45 pairs, the participants started by referring directly to the target object, using instructions such as “Bring me the bottle on the left side.” 24 pairs located the referent successfully in their first attempts. When instructions of this kind were not successful, half of the participants turned to explaining the action in more details, using instructions such as “in front of the projector”. The other half of the participants described the tasks by using explicit instructions, such as “The bottle is in front of the projector,” which may be much simpler for Participant B to comprehend and carry out. 9 pairs chose distance instructions as their strategy and only 4 of these pairs located the referent on the first attempts. The 5 pairs that failed to locate the reference on their first attempt turned to employing the directional instructions and this led to successes.

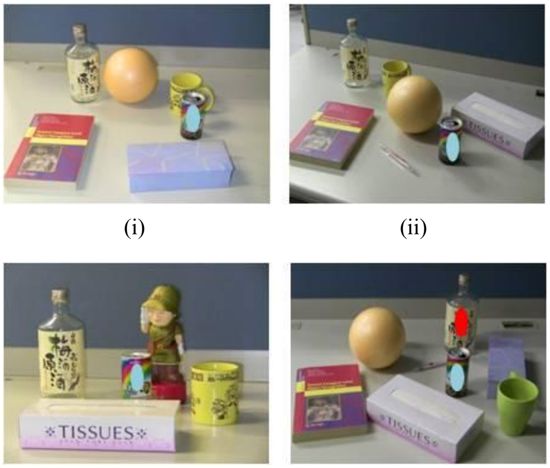
Altogether, the results showed that directional instructions are more explicit than distance instructions. This is an important result for designing the robotic system and it suggests that robots need to be able to understand simple directional instructions. In this regard, we employ 7 spatial terms: left right, front back and leftmost between (middle) rightmost as our domain. The last of these is adequate for group reference system. Although humans typically use their own point of view in spatial reference, they often adopt their interlocutor’s perspective if the actions by the listener or different cognitive abilities on the part of the listener are involved [14]. Consequently, in our robotic system, we adopt the reference system in this way: the speakers can refer to a salient object, if available, as relatum in a relative reference system. There are also some objects that can be employed using the intrinsic reference system, such as PC display, TV, refrigerator, etc. They may refer to the group as relatum in a relative reference system. In this case, they specify the object’s position relative to the rest of the group from the robot’s point of view [5].

### *C. Relatum Strategy*

It is observed that humans seem to favor choosing the object that deems the most suitable one to express intensions. Humans have the ability to infer their partner’s intentions even if they have different perspectives. However, robots lack such ability. Robots are not able to perceive the surrounding, the color, and the linguistic meaning unless humans manipulate it. In our Human Robot interactive design, the relatums should be easily recognized and unambiguous for both human and robot. We have been investigating the  preferable conditions for Relatums in [15]. In this experiment, we attempt to inspect how humans choose the optimal one to express their purposes straightforward in a complex scene.

**Participants** 20 participants were divided into 5 groups, 4 persons at a time in turn.

**Procedure** We requested the participants to announce the reference object they chose by answering the question: "Where is xx?" in 4 scenarios (Figure 2). All of the objects are unique and identified in each scenario. We also determined the target objects  in all scenarios, the target was a coffee can, though in Scenario 4, a bottle (labeled in red) was the other target (here, we wondered whether humans would switch the reference object when there were multi targets). The participants could select reference object on their own, and they were also allowed to use descriptors such as **black**, **round** to show us more depictions.



(i) (ii)

(iii) (iv)

Figure 2. Scenario scripts.

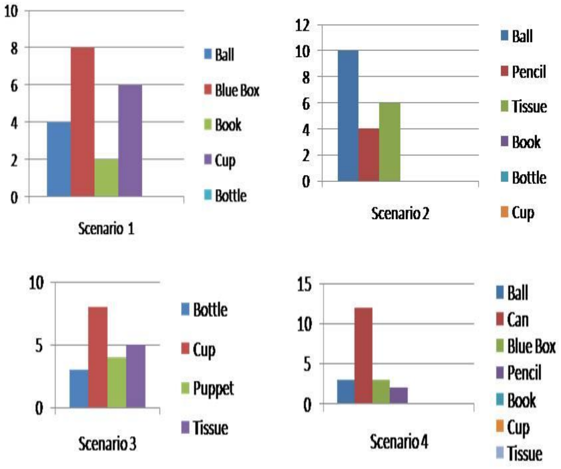


Figure 3. The number of objects used as Relatum in each scenario and their percentages.

**Results** Figure 3 shows the number of objects in each scenario and their percentages.

The result of Scenario 1 showed that the objects that were closer to the target were much more preferred over others. Scenario 2 indicated that salient objects, which are large in size with a shape/color, have greater tendency to become the relatums over others. In result 3, there was not any quantitative difference among the numbers. The only one and most significant factor was that the puppet did not seem to be effortless to depict for some participants like the bottle, cup and tissue. Consequently, objects that are readily describe in semantics gain more advantages than others. The can, which took first place in result 4, was 4 times more than the ball and blue box. This suggests that linguistic context is much more convincing than selecting a new object as reference.

According to [16], entities can be salient by being very vivid, pervasive, and unique, or by being spoken about most recently. The data indicated in Figure 4 is at the same level, since the robot has the capability of detecting simple objects and color, the criterion is set listed below:

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

When several similar relatums surround the referent, we employ the strategy to find the optimal relatum. For example, the user would briefly describe a referent by saying: "It is on the left side of the cup." Then, the robot will find the closest one.

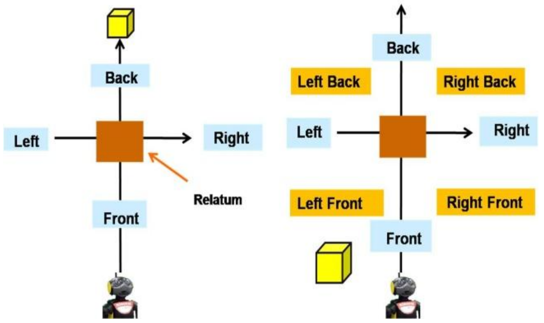
### *D. The Interpretation of Bounding Relations*

An essential aspect of the robot's ability to execute instructions is its interpretation of the spatial relations specified between objects as relatum and the referents [5]. Different kinds of reference systems required for interpreting linguistic references according to the three options were outlined in Section 2 and for handling the corresponding instructions.

The bounding expressions are then further resolved as follows. We have implemented the relative system, which may be most often used to represent spatial relationships. The origin is often omitted and the default origin is usually the listener, sometimes the speaker. In our implementation, we assume that the origin is the robot (listener). If the user (speaker) and the robot are looking in almost the same direction, the speaker origin coincides with the listener origin. If we take the robot's point of view as origin, all objects are represented in an arrangement resembling a plan view. Thus, the reference axis is a combination of two directed lines through the center of the object as a relatum, as illustrated in Figure 4(i). The center of the bonding area can be used as a point like representation. The vertical divides the reference plan as left and right parts while the horizontal manages the front and back parts.

For a finer partition, the reference axis is rotated for 45 degrees, respectively, and new orientation relations are found, which are called left front, left back, right front and right back. For combined expressions such as "left front" vs. precise expressions like “strict front”, we use the partition presented in Figure 4(ii).

In [11], the group based reference system is considered to be the relative reference system using the group as a relatum, and the centroid of the group serves as a virtual relatum. Figure 5 shows the reference direction given by a directed straight line through the center of the group. The objects labeled with a blue box are considered as a group, same color implies they own the same attributes. The object closest to the group centroid can be referred to as the “middle object”, then the left and the right ones can be distinguished as well.



(i) (ii)

Figure 4. (i): Relative reference model; (ii) after 45 degree rotation of the axes, a combined expression model.

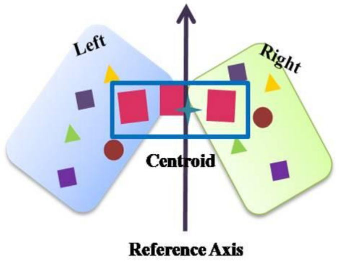


Figure 5. Group based reference.

## III. Using Natural Language Control System

The spatial reasoning and the Natural Language Processing collaborate to provide the capability of human robot interaction. For example, a user may issue a command directly to the robot: “Bring me the bowl.” The robot may respond: “I cannot see it. Where is it?” Then, the description of the bowl becomes necessary. It is appropriate that the user needs to describe the position of the bowl rather than its color and appearance, and then the user continues: “It is on the left side of the spoon.” And the dialog continues. In this way, providing an informative context is essential for system detection. Based on the survey results, we build a fundamental parser to analyze user commands, which have some restrictions listed as follow:

### *A. Vocabulary and Syntactic Constructions*

As the robot has a limited linguistic ability, simple and common vocabulary would be better to analyze. In the current system, theses vocabularies to describe position are legal: Front in front of, Back at the back of, Left on the left side of, Right on the right side of , Left front/back, Right front/back, Leftmost Between (Middle) Rightmost.

In our system, we restrict the user input command in the format such as: **Get/Bring me A. A is xxxx B**. A indicates the referent while B indicates the relatum, and **xxxx** represents preposition.

### *B. Ellipsis*

There are a lot of ways to describe location. In particular, humans often employ ellipsis imperatives in dialogue, which lacks elaborations and is loose in grammatical construction, such as: “right to the object that I indicated last time.” Under the restriction that object names could not be mentioned, it was hard for the robot to understand. Our strategy imposes a restriction on ellipsis. Experimentally, the user is suggested to repeat the context that has been mentioned earlier in case of making mistake.

We are also interested in the use of deictic words such as **this/that**. In some cases, no noteworthy uses are observed, but in most cases, participants indicate different distance from the speaker. “This” presents the nearer one, whereas “That” expresses the further one. Humans are accustomed to use them with pointing gestures. In light of the fact that our robot is not able to recognize pointing gesture, we prefer that the user repeats the full name of the object mentioned before.

## IV. Experimental Study

We performed a set of experiments to investigate our approach. Given a completely unlabeled image of multiple objects, our goal is to locate the object in instruction with in an image. In the first experiment shown in Figure 6, the robot is asked to locate the marker in the scene. The dialog is as follows.

User: Bring me the cup.
Robot: I cannot see it. Where is it?
User: It is in front of the PC display.
Robot: I can see it.

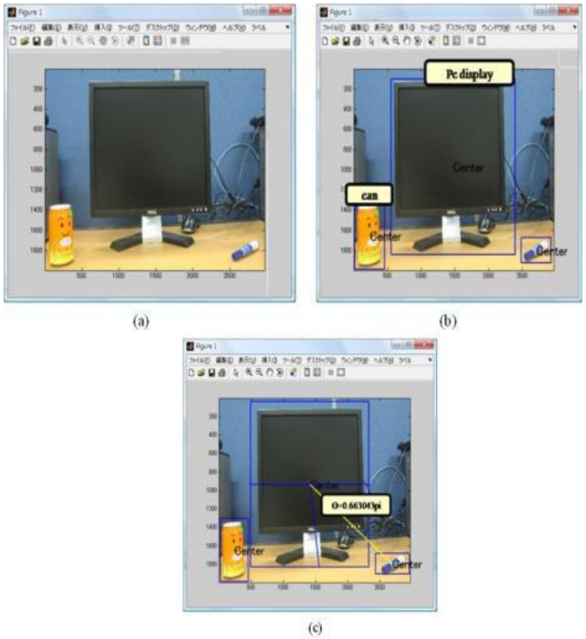
In the second experiment shown in Figure 7, the robot is required to locate the green cup first, then, the cylinder shaped box. The dialog is shown below.

User: Bring me the cup.
Robot: I cannot see it. Where is it?
User: It is to the right in back of the can.
Robot: I can see it.
User: Bring me the box.
Robot: I cannot see it. Where is it?
User: It is on the left side of the cup.
Robot: I cannot see it. Is it left in back of the coffee?

Experiment 3 in Figure 8 is involved in group objects. As there were 3 coffee bottles, the system marked them as a group, when it caught the keyword “**rightmost**”, it commenced by locating the cup with the previous keyword “right”. The dialog is:

User: Bring me the cup.
Robot: I cannot see it. Where is it?
User: It is in front of the bottle
Robot: I can see 3 bottles. Which bottle?
User: The rightmost bottle.

Robot: I can see it.

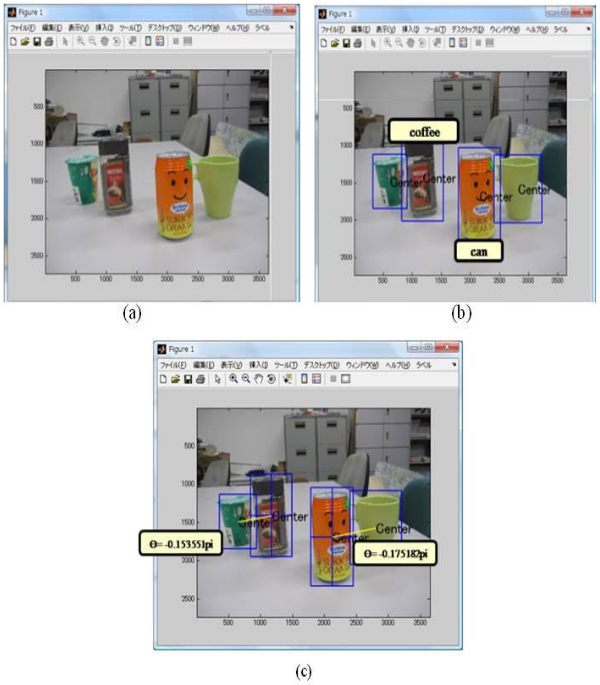


(a)

(b)

(c)

Figure 6. Experimental result 1.
(a) Input image; (b) Autonomous detection, the pc display and the can are detected, (c) Final result.

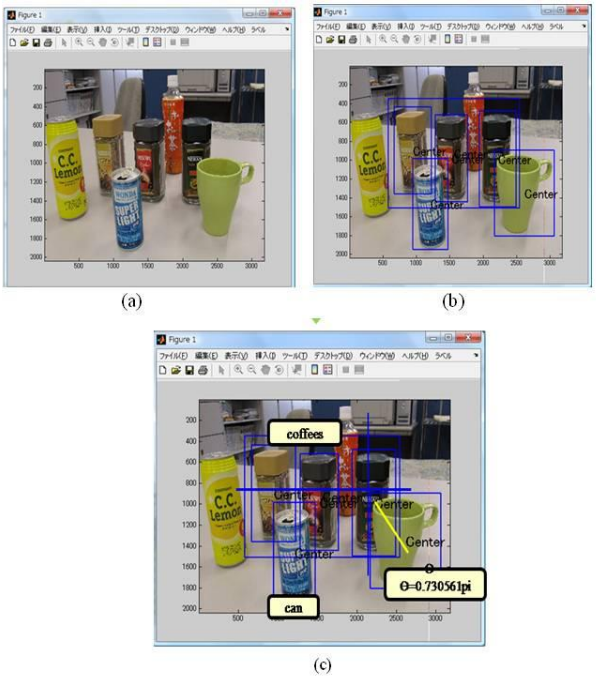


(a)

(b)

(c)

Figure 7. Experimental result 2.
(a) Input image; (b) Autonomous detection, the coffee and the can are detected, (c) Final result.



(a)

(b)

(c)

Figure 8. Experimental result 3.
(a) Input image; (b) Autonomous detection, 3 coffee bottles are recognized as a group, (c) Final result.

## V. Discussion

There are 2 specific areas that may influence the results.

### *A. Viewpoint*

We consider that the relative reference system needs a viewpoint. Depending on the viewpoint, the orientation may be different. The experimental results indicate to us that the diversity of information they hold results in varied descriptions to target and reference objects. In the experiments, the user consistently used the robot's viewpoint except the second one.

In the experiment, the user did not take the same viewpoint as the robot due to the optimal reference selection. Thus, the instruction was not carried out. Then, the robot used the user's perspective and made a subsequent query accordingly. To avoid ambiguity, the users may define the viewpoint when it is explicitly mentioned, for instance: "From my point of view" or "It is in front of you from your view point."

### *B. Group based Reference*

As pointed out earlier, many participants made use of the concept of a group in order to specify the position of one of its members. However, the question needs to be considered as to why many users did not use this concept. On the one hand, failures might be due to some participants not expecting the robot to be able to grasp the concept of a group and others, as this involves comparison, identification of similarity, and categorization [11]. On the other hand, others might over expect the robot to perform. The fact is that robots are not able to recognize a group of objects like humans' can. In our results, successful detection usually takes advantage of feature detection.

However, for different objects, barely by features is not accurate in that the viewpoint, distance in Euclidean Space. Generally, in human cognition, determining whether there is a group of objects or not depends on the distances between them. This involves another two linguistic concepts "near" and "far". How near is near and how far is far? It is difficult to define ranges in a numerical way. This is our next issue which will be discussed in the future. Further experiments will be necessary in order to more closely analyze humans' choices of group based reference.

## VI. Conclusion

This research analyzed how humans describe objects using spatial terms in tabletop environments. The purpose of the analysis is to offer a simple and nature way for humans to conduct robots to retrieve objects even if some of the objects may not be recognized by the robots. The core of the strategy is the reference system. This research employed two reference systems in this paper, they are: intrinsic and relative. Group based reference system can be viewed as a special case of relative reference. The distinction between intrinsic and relative system is not very pronounced but crucial. Therefore, this researched distinguished the two of reference systems by the combination of objects and spatial terms.

We illustrated the strategy of spatial reference that human users employ in interaction with a robot and it proposed an integrated interaction mode. In the interaction mode, the system took the initiative and asked the user information about the target object. Experiments were carried out in which human users instruct the robot employing the strategy. The experimental results showed that designing a human robot interaction system cannot solely rely on the way of human human communication. A current research direction was briefly sketched. We attempt to build up an explicit model in the future. This will need further evaluation of the approach and investigation of human robot interaction.

## Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology under the Grant in Aid for Scientific Research (KAKENHI 19300055).

## References

[1] P. Bloom, M.A. Peterson, L. Nadel, and M.F. Garrett (Eds.), *Language and Space,* MIT Press, 1999.

[2] L. Cao, Y. Kobayashi, and Y. Kuno, "Spatial Relation Model for Object Recognition in Human Robot Interaction," *Proc International conference on Intelligent Computing*, LNCS 5754, Springer, pp.574 584, 2009.

[3] S. Dobnik, S. Pulman, P. Newman and A. Harrison, "Teaching a Robot Spatial Expressions," *Second ACL SIGSEM*, Colchester, UK, 2005.

[4] A. Herskovits, *Language and Spatial Cognition: An Inderdisciplinary Study of the Prepositions in English*, *Studies in Natural Language Processing*. Cambridge: Cambridge University Press, 1986.

[5] R. Moratz, T. Tenbrink, J. Bateman and K. Fischer, "Spatial Knowledge Representation for Human Robot Interaction," in *Spatial Cognition III*, Volume 2685, Springer Berlin /Heidelberg, 2003.

[6] M. Skubic, G. Chronis, P. Matsakis and J. Keller, "Generating Linguistic Spatial Descriptions from Sonar Readings Using the Histogram of Forces", *in Proc.of the IEEE Intl.Conf on Robotics and Automation*, pp.465 490, 2001.

[7] M. Skubic, D. Perzanowski, A. Schultz and W. Adams, "Using Spatial Language in a Human Robot Dialog", *in Proc.of the IEEE Intl.Conf on Robotics and Automation*, Washington.D.C., 2002.

[8] A. Mansur and Y. Kuno, "Specific and Class Object Recognition for Service Robots through Autonomous and Interactive Methods," *IEICE Trans. Information and Systems*, vol.E91 D, no.6, pp.1793 1803, 2008.

[9] D. Das, Y. Kobayashi, and Y. Kuno, "A Hybrid Model for Multiple Object Category Detection and Localization", MVA 2009 IAPR Conference on Machine Vision Applications, 2009.

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

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

[12] D. Levine, J. Warach, and M. Farah, Two Visual Systems in Mental Imagery: Dissociation of 'what' and 'where' in Imagery Disorders due to Bilateral Posterior Cerebral Lesions, *Neurology*, Volume 35, pp.1010 1018, 1985.

[13] C. Eschenbach, L. Tschander, C. Habel, and L. Kulik, "Lexical Specification of Paths," *Spatial Cognition II, Lecture Notes in Artificial Intelligence*. Springer Verlag, 2000.

[14] T. Herrmann and J. Grabowski, "Sprechen:Psychologie der Sprachproduktion, " *Spektrum Verlag*, Heidelberg,1994.

[15] R. Kurnia, Md.A. Hossain, A. Nakamura, and Y. Kuno, "Using Reference Objects to Specify Position in Interactive Object Recognition," *International Conference on Instrumentation, Communication and Information Technology* Proc., pp710, 2005.

[16] T. Pattabhiraman, "Aspects of Salience in Natural Language Generation," *Vancouver, B.C, Simon Fraser University,* Ph.D. Thesis, 1992.