Grounding Spatial Relations in Natural Language by Fuzzy Representation for Human-Robot Interaction

Jiacheng Tan
School of Computing
University of Portsmouth
Portsmouth, UK
jiachen.tan@port.ac.uk

Zhaojie Ju and Honghai Liu
Department of Creative Technology
University of Portsmouth
Portsmouth, UK
zhaojie.ju, honghai.liu@port.ac.uk

***Abstract*—This paper addresses the issue of grounding spatial relations in natural language for human-robot interaction and robot control. The problem is approached by identifying two set of spatial relations, the image space-based and object-centered, and expressing them as fuzzy sets to capture the ambiguity inherent to the linguistic expressions for the relations. The sizes and shades of the scene objects have also been modeled as fuzzy sets for conditioning the spatial relations. To verify the validity of our approach and test its feasibility in a natural language-based interface, we have considered the typical scenarios of using the spatial relations in simple declarative and imperative sentences and designed simple grammars for parsing such sentences. Our experiment has shown that fuzzy spatial relation analysis provides a useful way for modeling the ambiguity or imprecision of the natural language in describing spatial relations and that it is possible to use the spatial relation models to support robot control and human-robot interaction in a natural language-based interface.**

Keywords—spatial relations; fuzzy set; human-robot interaction; artificial intelligence.

I. Introduction

Given the advances in computer science and artificial intelligence in the past few decades, it is still a huge challenge for robots or autonomous devices to survive unknown tasks in unknown environments. The problem is due to the inability of current robots in perceiving and understanding their environments and in identifying and organising their tasks. The situation is unlikely to improve greatly in the foreseeable future. For this reason, human operators are still indispensable in the control loops of critical robotic applications such as in space explorations, underwater servicing, landmine disposal, or household activities. In such missions, human operators, partially or completely, take over from the robots the responsibility of environment assessment and task planning, and the robots act passively on receiving the instructions from the operators. For this to happen seamlessly, effective human-robot interaction is a pivotal issue.

In the study of human-robot interaction, using natural language as an interface for human-robot interaction has gained a lot of momentum in recent years. Compared with programming languages or other specialised robot control interfaces, natural language enables people to plan and then communicate complex tasks in an intuitive and flexible way. To achieve this, the robots must be able to interpret the natural language sentences and act upon them. Such an interface would normally consist of at least two functional modules: a natural language parsing module that parses the commands in natural language into linguistic symbols and a grounding module that maps the linguistic symbols to the percepts that a robot perceives or senses in its environment or workspace, such as objects, places, relations or actions. These problems have attracted a lot of attentions in recent years and different approaches have investigated.

Within the context of robot control and human-robot interaction, the aim of natural language parsing has been to derive a formal representation from natural language utterances or commands. The representation could be obtained through observation [1], by parsing the sentences into spatial description clauses (SDCs) [2, 3], or through learning [4]. In comparison with the natural language parsing problem, the options for the solution of the natural language grounding problem are rather limited. The reason is because correct grounding depends very much on the correct perception and modeling of the work environments by the robots, which is still a very challenging problem.

In this paper, we focus on the issue of grounding spatial relations in natural language for human-robot interaction and commanding. The paper extends our work on fuzzy spatial relation representation [5] by introducing the group-related spatial relations and more powerful fuzzy qualifiers to better model the spatial relations used in natural language for object and spatial location description. We derive the formal representation of the simple natural language sentences that are normally used for describing the spatial relations in typical 2D workspaces or application scenarios. In using the relations to describe the objects in the workspace of a robot, we assume no prior knowledge about the workspace and the objects in it. The knowledge about the workspace, such as landmarks and object identities, can be obtained and accumulated through interactive instruction during task execution [6].

II. Related Work

Spatial reasoning is considered as the domain of spatial knowledge representation, in particular spatial relations between spatial entities, and of reasoning on these entities and relations [7]. Humans effortlessly use such knowledge in everyday life. Spatial reasoning can be used to solve sophisticated space-related problems and plays a very important role in many areas of science and engineering, for example, in image understanding and content-base retrieval [8] and robot navigation [9].

Spatial relation analysis and grounding is one of the important issues in the research of using natural language to control and interact with robots. Spatial relations are used to create a semantic map of the environment, which enables robots to acquire grounded representation of natural language. The representation is attractive because it allows unknown environment being gradually learnt and modeled without recourse to explicit 3D geometric data of the environment.

A command in natural language can be decomposed into a hierarchy of spatial description clauses. Each of the clauses consists of a subject, a verb for the action to take, a spatial relation that describes the relation between the subject and the reference object, and a landmark that is referenced by the relation [2]. The groundings of the linguistic constituents can be inferred from a grounding graph – a probabilistic graph that was trained on a corpus of natural language commands paired with groundings for each part of the command [3]. Guadarrama and et. al., [10] use a trained multi-class logistic regression model for the grounding of spatial relations. The model was trained with the spatial features computed between the bounding boxes of the landmarks and target objects.

Our approach differs from these approaches in that, instead of treating the grounding problem as purely a learning problem, we use membership functions to define the spatial relations. We do not rely on any prior knowledge or an explicit model of the environment to initialize the spatial relations among the objects. We represent the key attributes of an object, for example its identity and its spatial relations with other objects, in a knowledge base in an approach similar to [11, 12], although we have not fully investigated the knowledge acquisition and representation issue.

In our work, we assume no prior knowledge about the work environment. The knowledge about the environment such as the landmarks and their identities are learnt from the natural language control and operation commands. The spatial relations between objects are used as constraints to condition the learning problem. Our approach is similar to [6], where grounded language is learnt from interactive instructions. The knowledge about the workspace is transferred from a human instructor to a robot in the form of demonstrative sentences initiated by the human instructor or by the robot querying the instructor when it lacks the knowledge to comprehend an utterance or execute an action.

In modeling the spatial relations, we use the framework of fuzzy spatial reasoning. As an efficient tool for modeling the ambiguity in the linguistic definitions of regions and relations, the fuzzy set theory has been well studied [5, 13, 14, 15]. It is evident that the fuzzy representation of spatial regions and relations has provided an adequate framework for spatial knowledge representation and reasoning: it captures the imprecision inherent to the linguistic expressions of spatial regions and relations; it reduces the semantic gap between symbolic concepts and numerical information [16].

III. Grounding Spatial Relations

The aim of spatial relation analysis is to derive the spatial relations among a collection of objects and to establish an inference system in which the location of a single object can be uniquely defined and deduced, and subsequently the deduced location is used to ground the linguistic terms in natural language commands. In essence, by exploring the spatial relations we wish to establish a coordinate system equivalent to the Cartesian one but without recourse to the coordinates of real numbers (as far as specifying a location is concerned). The differences between the two are in that, one defines only a finite number of locations or regions by linguistic coordinates and a certain amount of imprecision is allowed, and the other defines an infinite number of locations by real numbers without any ambiguity.

As the system is regarded as a type of coordinate systems, we first need to consider the space it will address and the reference or “origin” with respect to which spatial relations can be defined. We consider using a robot-centered vision system to sense the workspace. At any one time, the workspace, or part of it, is represented by an image of the workspace. In the following discussion, we assume that the human operators share the same views of the workspace with the vision system of the robot, and for easy formulation we have ignored the effect of perspective foreshortening.

The workspace of the robot is considered as 2D scenes populated with objects. Two sets of spatial relations, image space-based relations and object space-based or object-centered relations, will be defined. It will become evident that the image space-based relations are necessary for ensuring any single location in the field of vision is accessible and that the object space-based relations, while not possessing the same attribute, can greatly improve the practicality of the reasoning system. As the spatial relations will be used for grounding natural language commands, the spatial relations must be so defined such that they are consistent with human’s conceptions of these relations and can tolerate the imprecision of the descriptions of these relations in natural language.

A. Image-Space Spatial Relations

Given an image of a workspace, in order to specify a spatial relation in the absence of any landmark it is necessary to partition the image and label the resulted partitions. When partitioning the workspace, we have paid attention to the following points. Firstly, the partitions must be conceptually consistent with human conceptions of the spatial regions and relations. This will make the language-grounding task easier.

Secondly, the partition scheme should allow partitioning be recursively applicable to the sub-regions. This is to ensure that every single location in the space is accessible. Finally, the partitioning must be geometrically complete, by which we mean that no holes or gaps are left by the partitioning. These considerations naturally lead to us to partition the entire image into nine *regions*: a *centre*, a *right*, a *left*, a *top*, a *bottom* and four *corners*, as shown in Fig. 1 (a). We name them purely for convenience of discussion.

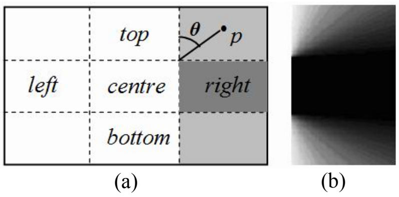
With respect to the region *centre*, we define four primitive spatial relations: *top*, *bottom*, *left*, and *right*. We use them to describe the spatial relation of a location with respect to the reference region, *centre*. Therefore, by relation *right* we actually mean a point is to the *right of* the centre region. We avoid using the term "*right of*", because we reserve it for naming its object-space counterpart. The relations have a second meaning: we use them to represent the set of all points on them a particular spatial relation holds. Therefore, there is a significant difference between the region *right* and the relation *right*. As we will see shortly, the relation *right* represents the whole region of the right hand side of the image. In the following discussions, when a relation is mentioned it normally refers to its second meaning.

Now, we consider the geometric meaning of these relations. In natural language, the meanings of these relations are rather vague. For example, a point anywhere within the shaded area of Fig. 1(a) could be considered more or less to have a *right* relation with the centre region. This relation is certainly valid for all the points within the area with darker shade – the region we have named as *right*. However, for the points within the two corners the degree of validity of the relation varies. If a point is very close to the top or bottom edges of the region *right*, the relation is almost certain. If a point is very close to the right edges of the *top* or *bottom* regions, the relation almost fails. Obviously, the distance between the point and the corresponding corner point of the *centre* region has no influence over the validity of the relation.

This variation in the degree of validity of the relation can be modeled as a simple linear function in θ, the angle between vertical line and the line joining the point P and the corner point of the *centre* region, as shown in Fig. 1(a). Being such modeled, the relation *right* refers to the set of points that maintain a *right* relation with the region *centre* – the entire shaded region in Fig. 1(a). The points within this region can be adequately represented by a fuzzy set, with each point being assigned a fuzzy membership to reflect the degree of validity of the relation at that point. Fig. 1(b) shows the fuzzy membership of each point of the region. The same argument applies to the relations *top*, *bottom* and *left*, and their membership can be similarly assigned [5].

We now consider the region *centre*. Being used as the reference to define the primitive spatial relations, the region *centre* itself is not a relation. However, the space described as *centre* in natural language tends to be very small and is accompanied by imprecision. Within this region, the degree of a point being considered to belong to *centre* varies according to its distance from the centroid of the region. This characteristic of the region also calls for a fuzzy representation. In fact, we can view *centre* as a region consisting of the set of points that hold a geometric relation to the centroid of the region.

Fig. 1. (a) An image is divided into nine regions: a *centre*, a *right*, a *left*, a *top*, a *bottom* and four corners; (b) The fuzzy membership of the relation *right*, where the total darkness represents 1.0 and white represents 0.0.



The fuzzy definition of this relation is straightforward. It consists of the set of points each of which has its distance to the centroid of the region *centre* as its membership:

$$\mu = 1 - \sqrt{2}\sqrt{(\frac{x - x_c}{w})^2 + (\frac{y - y_c}{h})^2}\ , \tag{1}$$

with $-w/2 \le x - x_c \le w/2$ and $-h/2 \le y - y_c \le h/2$, where (x_c, y_c) is the geometric centre, and w and h are the width and height of the region. This definition takes into account the aspect ratio of the region.

In addition to the above relations, another set of relations that are frequently used in natural language for describing spatial locations are *left-most, right-most, topmost* and *bottommost*. This set of relations are crisp and can be evaluated by the left most, right most, topmost and bottommost edges of the bounding boxes of the objects in the scene, which will be discussed in Section III. B.

The image-space relations discussed in this section normally serve as the starting point of spatial relation learning and reasoning when a suitable landmark or prior knowledge about the workspace is not available. The robot can use these relations to initiate queries about the environment. For example, the robot may ask: "*what is object at the centre*?" Such queries could be answered by a human instructor by saying "*It's a cup.*" The human instructor can also teach the robot by initiating a demonstrative statement: "*a cup is at the centre.*" In both cases, the knowledge about the environment, i.e., the object name and spatial relations, and may be more, is passed on to the robot.

B. Object-Space Spatial Relations

The object-centered primitive spatial relations such as *left of*, *right of*, *above*, *below* and a few more have been identified and widely used for describing 2D spatial relations [17]. In our work, because the image is a perspective projection of a 2D working space, we have defined the following primitive relations: *left of*, *right of*, *behind*, and *in front of*.

In defining these relations, one of the most important factors is to find an appropriate representation for objects of all possible shapes. A good representation should lead to unambiguous spatial relations and efficient evaluation of the fuzzy membership functions.

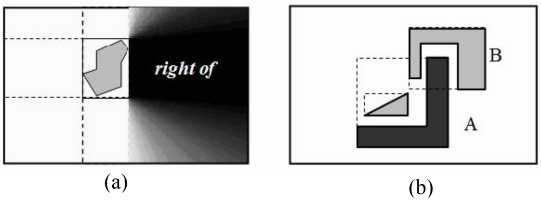
Bounding boxes and actual geometric shapes are two types of widely used representations in spatial relation analysis. Using the actual geometric shapes of the objects as a representation has been reported by some researchers, for example, in representing anatomic structures [16] and ground structures [8]. The main limitation of the representation is due to its computation demand in evaluating the spatial relations. The bounding box representation facilitates simple computation algorithms and therefore is more efficient. For this reason, it has mainly been used in applications where real-time performance is critical. The main concern about the bounding box representation is its representation accuracy.

For comparison purpose, we have investigated both representations. It has been shown that the actual shape representation gains no advantage over the bounding box representation when it comes to the evaluation of the primitive relations such as *left of*, *right of*, *behind*, and *in front of*. It results in the same spatial relations between two objects at varying distances as the bounding box representation does, but at much higher computation costs. In addition, the accuracy of image segmentation can seriously undermine the exact nature of the representation.

In our work, the bounding box representation is adopted. The centroids of the objects are used in evaluating the fuzzy membership functions of the primitive relations. Being such defined, the object-space primitive relations can be viewed as an extension of the image-space relations with the region *centre* being replaced by the bounding boxes of the objects. As an example, Fig. 2(a) shows the membership of *right of*.

From these primitive relations, we further define four derived object-centered spatial relations by using the logic conjunctions of the primitive relation and the distances to the reference object. They are *top left*, *top right*, *bottom left* and *bottom right*. Our observation and experience show that when dealing with a cluttered workspace these derived spatial relations are more frequently used than the primitive relations by the human operators in describing the objects with respect to a known landmark.

Fig. 2. (a) The bounding box representation and the membership of the relation *right of*; (b) a scenario where the bounding box representation fails and the relation *next to* applies.



(a) (b)

C. Distance-Related Spatial Relations

Our investigation has shown that the most useful distance-related relations are *next to* and *nearest to* when a single landmark is used, *between* when two landmarks are involved, and *within* when a group of more than two landmarks are referenced. The grounding of the linguistic elements associated with these relations necessarily involves evaluating the distances between the objects in question and the landmarks involved.

Distance computation can be efficiently done if two objects are sufficiently separated. As they are sufficiently separated, the influence of their actual shapes over the distance computation becomes less influential; therefore the distance between them can be better represented by their centroids. Of course, the computed distance between the centroids must be scaled afterwards to take into account the influence of the actual shapes of the objects or their bounding boxes. To save the computing costs, we choose to scale the distance by the bounding boxes of the objects. It can be shown, in most cases, this choice is harmless.

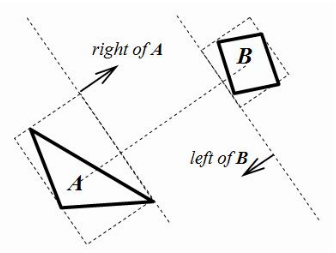
The relation *next to* describes the relations between the objects that are very close in space and the relations among them cannot be described by the object-space spatial relations *left of*, *right of*, *behind*, or *in front of*. The relation has a very limited scope; its uses are restricted to the close adjacency of a landmark. As illustrated in Fig. 2(b), with the bounding box representation, the triangular object would be considered as closer to *object A* than it is to *object B*. Obviously, the implementation of the relation *next to* calls for an actual shape-based distance computation.

Semantically, the relation *next to* is very similar to the relation *close to*, as reported by other authors [16, 17], but with *next to* having a much narrower scope and demanding for the actual object shapes being used in the evaluation of its membership functions. In contrast, the relation *close to* can be more efficiently implemented with the bounding box representation, as will be discussed in the next section.

The relation *nearest* differs from the relation *next to* in that the former has a wider scope. An object may have a *nearest* relation but not a *next to* relation with a landmark. The evaluation of the relations *nearest* and *farthest* requires the evaluation and ranking of the distances of objects to a reference landmark. Both *nearest* and *farthest* are crisp relations.

The spatial relations *between* and *within* in natural language have rather vague meanings. To implement them computationally, we have to bestow them more precise meanings. We define the relation *within* with respect to the centre of the convex of the centroids of the reference objects. The convex hull defines both the scope and the reference point of the *within* relation. Any object whose centroid falls within the convex hull will be accounted for in the evaluation of the *within* relation, and its distance to the centre of the convex hull is calculated as the measure of the “withinness” when multiple objects are involved.

Fig. 3. The relation *between* is the fuzzy conjunction of the object centered relation *right of* and *left of*.



The relation *between* is defined as the conjunction of the two object-space relations *left of* and *right of*. These two relations are evaluated with respect to the line that connects the centroids of the two landmarks, as shown in Fig. 3.

IV. NATURAL LANGUAGE QUALIFIERS

If we use the spatial relations to describe or retrieve an object in the workspace of a robot, the results can hardly be unique, unless the scene is sparsely populated. With a cluttered scene, it is more likely that a collection of objects will fit the same description. As each object will have a different membership with respect to different relations, one might attempt to use the numeric memberships as a type of coordinates to differentiate the objects within the collection. However, doing so would defeat our purpose of devising the spatial relations in the first place. To address this problem, we are going to introduce a few qualifiers.

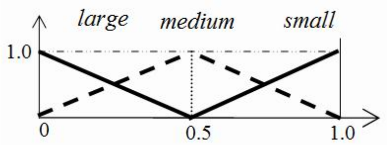
Humans use language to describe the features of an object to single it out when its spatial relations with other objects are ambiguous. Among the most frequently used features are the shapes, textures, colors and sizes of objects. Shapes are perhaps the most powerful and the most used feature in human object recognition. Unfortunately, a general solution for mapping 2D shapes to linguistic symbols that represent 3D objects is technically very difficult if possible at all. Therefore, in this section, we will discuss some linguistic qualifiers that can be readily detected. They are *size* and *shade qualifiers*.

A. Size Qualifiers

The frequently used qualifiers for size comparison are: *largest*, *large*, *medium*, *small* and *smallest*. We could have added two more qualifiers to the set, for example, *very large* and *very small*, but it has proved to be difficult for humans to perceive and differentiate the differences between *large* and *very large*. In this set of qualifiers, we want the qualifiers *largest* and *smallest* to be crisp ones because given a set of objects, the meanings of *largest* and *smallest* are normally unambiguous.

The areas of objects have been chosen as the measure for object sizes. The triangular functions are chosen as membership functions for the size qualifiers, as shown in Fig.4.

Fig. 4. The membership functions for the distance qualifiers.



B. Shade Qualifiers

Different from the size, the shade of an object refers to the fixed intensity values, for example, a shade of gray refers to an intensity value around 128. So the membership functions must be defined over the range of image intensity [0, 255] or its normalized equivalence [0.0, 1.0]. We noticed the fact that the visual perception of a shade is susceptible to the spatial configurations and the shades of surrounding objects [18]. However, without a quantitative analysis of the effect, it is difficult to account for it in this work. As a result, five qualifiers are defined over the range of the intensity: *black, dark*, *gray*, *light* and *white*. We also use the triangular functions to represent the qualifiers, as shown in Fig. 5.

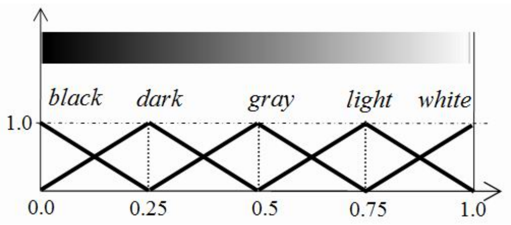
In this section, we have defined two sets of qualifiers, both crisp and fuzzy. Combined with the spatial relations, they form a set of tools for describing and querying a particular object in images.

V. WORKSPACE MODEL

For a robot to acquire and retain the knowledge about its workspace or environment, it is crucial to have a model of some kind for its world. The model plays the role of a memory and provides a resting place for the past and a starting point for predicting the future of its environment. A full model that accommodates the dynamic aspect of the workspace is beyond the scope of this paper. For simplicity reason, we consider a static scene to test our relation models. We assume that the relations between the objects are stable and the influence of the small variations in the viewpoint of the vision system is negligible. Under these conditions, we model the workspace as a knowledge base of which each entry corresponds to an object in the workspace. The robot will populate the knowledge base with the information elicited from the natural language commands about the scene objects.

The knowledge about an object can be very comprehensive. A full taxonomy, if possible, may contain information about its recognition, identification, manipulation, functional utility, interaction with and relations to other objects, and so on. As far as the work of this paper is concerned, the knowledge

Fig. 5. The membership functions of the shade qualifiers



about an object is minimal: its name and any synonyms and/or taxonomical information that allows words in natural language being mapped to the correct object. For example, both words *fruit* and *apple* should be grounded to an object named *apple*. The name and the taxonomical information can be hard-coded beforehand [12] or leant via interactive instruction at runtime [6].

Also included in the knowledge are the relations of an object with other objects. The use of such knowledge helps to avoid the unnecessary re-evaluation of the spatial relations. Ideally, certain ontological information should also be included in the knowledge base to facilitate automated decision making by the robot, but we did not explore this aspect of the workspace. In principle, the spatial relations between the objects become available once the image of the workspace is loaded. But we defer their evaluation to the command processing time. This is to avoid the problem of evaluating all the possible relations within a workspace at once, which is beneficial for reducing the overhead for processing the workspaces that consist of a large number of objects.

VI. NATURAL LANGUAGE COMMANDS

Our motivation for modeling the spatial relations has been to ground the linguistic symbols in robot control commands to the percepts of robots so that we can interact with the robots via natural language. In real world applications, to describe an operation task, very complex sentences may have to be used. Inferring the meanings of unconstrained natural language commands would require the deployment of a full-fledged natural language processing system, which is beyond the scope of this paper. However, to verify the validity and the feasibility of our approach, we have attempted to parse simple sentences that involve using the spatial relations for directing or instructing a robot. We have defined the simple declarative and imperative sentences, as shown by the grammar in Table I. In designing the grammar, we wish to be able to capture the natural language sentences that are likely to be used in our experimental scenario, but we assume in no way that the grammar will capture all the possible sentence patterns or in all possible situations.

The syntax for <*statement*> in Table I defines the declarative sentences used by the human operator to state a fact about the workspace. The fact is passed to the robot by issuing a command in the form of a demonstrative statement or by answering the robot-initiated queries. For example, "*an apple is at the centre*" and "*apple is at the right of the box*". The syntax also includes the rules for sentences that give taxonomical information such as "*an apple is a fruit*".

The imperative sentence, <*direction-cmd*>, defines the command that positions the robot or its end-effector. For example "*Move to the left of the box*". The other imperative sentence, <*action-cmd*>, models the sentences that instruct the robot to carry out the task operations specified by the non-terminal <*action-verb*>. The actual operations contained in <*action-verb*> depend on the nature of the operation tasks.

TABLE I GRAMMAR OF SIMPLE SENTENCES INVOLVING SPATIAL RELATIONS

```
<statement> ::= <qualified-noun><is ><noun>
            |<noun> "consists of" <noun>
            |<noun> "is at" <r_i>
            |<noun>"is at"<r_o><qualified-noun>
<direction-cmd> ::= <direction-verb> <location>
<action-cmd> ::= <action-verb><qualified-noun >
            |<action-verb><qualified-noun ><location>
            |< action -cmd> "and" < action-cmd>
<location> ::= <r_i>|<r_g>
            |<r_o><qualified-noun>
            |<r_o><qualified-noun>< r_s>
<qualified-noun> ::= <noun>|<q><noun>
<noun-group> ::= <noun-group>
            |<qualified-noun>"and" <qualified-noun>
            |<qualified-noun>","
<q> ::= " "|<q_s>|<q_c>|<q_s><q_c>
<r_o> ::= "left of"| "right of"| "behind"|"in front of"
            |"top left"| "top right"| "bottom left"
            |"bottom right" |"next to" |"nearest to"
<r_i> ::= "centre"| "left"| "right"| "top"| "bottom"
            | "left most" | "right most"
            | "top most" | "bottom most"
<r_i>::= "between"<noun-group>
            |"within"< noun-group>
<q_s> ::= "largest"|"large"|"medium"
            |"small"|"smallest"
<q_c> ::= "black"|"dark"|"gray"
            |"light"|"white"
<action-verb> ::= "pick up"|"put down"
<direction-verb> ::= "move to"
<noun> :: = "object" | "it" | <object name>
```

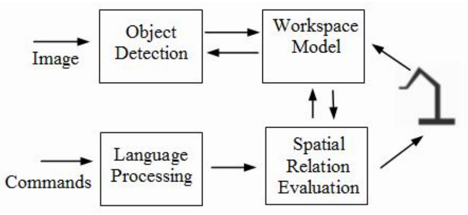
We consider two typical operations: "*pick up*" and "*put down*". The non-terminal <*noun*> contains a string "object", which is the placeholder for the un-named objects, and <*object name*>, the names of the objects whose identities are known. The names of objects in the workspace could be pre-programmed if they are known, or learnt from the demonstrative statements or the operation instructions at runtime. For example, in the sentence "*pick up the spoon next to the cup*", if the "*cup*" is known, then the object referred to by the name "*spoon*" is evident. At parsing time, any noun will be allowed in a sentence, and it will be checked against the set <*noun*>. The syntax allows an object be referred to by its name if it is known, e.g., "*put down the knife*", or by its location if its name is not known yet, e.g., "*pick up the object at the centre*". In the latter case, a generic name "*object*" is used as a placeholder. It also allows more than one actions being specified in a single sentence, e.g., "*pick up the knife and put it down on the left of the cup*".

VII. EXPERIMENT RESULTS AND OBSERVATIONS

To experiment with our approach, a system that implements all the relations, qualifiers and other necessary components has been developed. However, in the absence of a real robot manipulator, we can only simulate the response of the system. The architecture of system is as shown in Fig. 6.

The input image, as shown in Fig. 7, is processed by the object detection module. The processing segments the image

Fig. 6. The system diagram



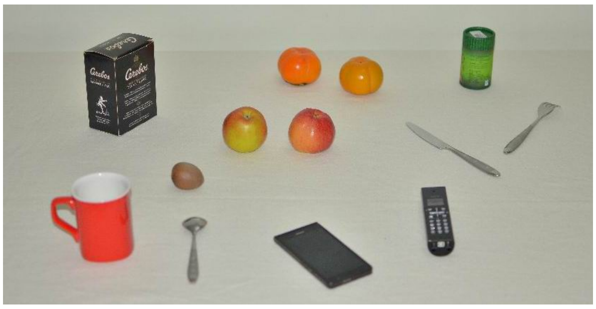
into regions that correspond to the objects in the workspace and computes their properties, which include the bounding boxes, centroids, convex hulls and pixel statistics of the regions. The regions and their properties constitute the complete knowledge of the robot about the workspace yet. At this stage, no spatial relation has been evaluated.

When a command is received from the operator, it is processed by the language-processing module, where the natural language sentences are parsed and assembled into concatenated function calls that evaluate the relevant relations and/or qualifiers within the spatial relation evaluation module.

When the system is initialized with an image, its knowledge about the workspace is no more than the segmented and indexed image regions that represent the scene objects. For convenience of discussion, we label the objects as shown in Fig. 8. Obviously, if the robot is working in a known or partially known environment, the basic information about certain objects such as their names and other properties can be pre-programmed into the system, or ideally, be obtained automatically by using vision or other sensing modalities.

In our experiment, we assume no prior knowledge whatsoever about the workspace. The system starts working by first being taught by the human operator (keyboard input) the landmarks which the robot can use as the references to derive spatial relations from. For example, we could start by issuing the sentence "*the largest object on the top left is a salt box.*" The sentence causes the evaluation of two image space relations, *left*, followed by *top*, and then followed by a qualifier *largest*. The evaluation of the two spatial relations retrieves all the objects that fall within the designated region. The application of the qualifier *largest* on the retrieved objects allows the box being selected.

Fig. 7. The image of workspace



After locating the object and checking the knowledge base against the name *salt box*, the name is assigned to *Object 2* and the object *salt box* can be used as a landmark thereafter. Of course, one may say "*Object 2 is a salt box*", but for the purpose of verifying the use of spatial relations, we have avoided referencing an object by its index. The usefulness of this simple learning mechanism might not be significant in this experimental scenario, but it would be useful for any robot that has been equipped with less than an ideal cognitive ability however has to work in an unknown environment.

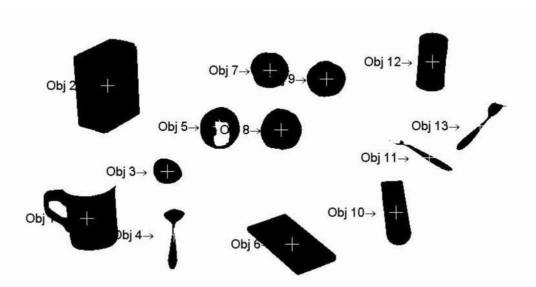
Another example for knowledge transfer from human operator to robot is given by a statement like "*the fruit behind the apple is a persimmon*". Suppose that object *apple* is known to be *Object 8*. Evaluating relation *behind* against object *apple* leads to *object 7*, therefore name *persimmon* is assign to it. In addition, the statement also allows the equivalence relation between the names *fruit* and *persimmon* being deduced and recorded, although the taxonomical relation between them is still to be clarified.

As the knowledge about and the number of landmarks of the workspace accumulating, the spatial relations among the objects become clearer and the ways of describing an object become more flexible and diversified. For example, in a command that involves the *pick up* operation, *Object 5* can be described as *"the apple on the left*", "*the apple on the right of the salt box*", "*the apple next to/closest to the nut*" and so on.

We have experimented the system with many such commands. It has been shown that all the objects in the test image can be correctly located or referenced by describing their spatial relations. This capability of mapping the linguistic symbols to the entities in the workspace provides a way of solving the natural language grounding problem and a means of compensation for the limited cognitive capability of a robot.

We noticed that given a workspace consisting of objects that lack distinctive characteristics in their sizes, shades or locations, it is not a straightforward process to select the "seed" object to start creating the referencing network of spatial relations. Some objects are hard to describe in natural language without referencing to any landmarks, for example, the *Object 3, 8* and *9* in Fig. 8. However, to our best knowledge, there is not a benchmark for us to use to assess the capability and efficiency of our system.

Fig. 8. The detected scene objects.



We also noticed that different user tends to use different spatial relations to describe the same object if the spatial relations of the object with the other objects are nontrivial. One of the implications of this observation is that given a workspace scenario, we may not be able to exhaust all the possible ways that an object could be described in natural language. As a result, given an arbitrary workspace configuration, it is not clear whether or not every object can be described by the sentences defined by the grammar in Table I.

VIII. Conclusions and Future Work

In this paper, we have studied the problems of modeling and grounding the spatial relations needed for commanding and interacting with robots via natural language. Using fuzzy sets, we have modeled the imprecise nature of the spatial relations and some language qualifiers in natural language. The fuzzy models of the spatial relations allow spatial relations being automatically extracted from the image of workspace. Performing fuzzy reasoning on the derived spatial relations enable us to ground the objects defined by complex spatial relations in natural language commands to the correct items in the workspace. We have assessed our approach by experimenting the implemented system with various natural language commands. It has been shown that grounding the spatial relations within a fuzzy framework provides a feasible and efficient way for facilitating a natural language-based interface for human-robot interaction. It has also demonstrated that correctly grounded spatial relations in natural language commands provide a mechanism for human operators to use natural language to teach robots to understand, and to guide them to operate in, unknown environments. This potential could be exploited as a way of compensating for the weaknesses of current robot systems in coping with unknown environments or tasks.

The work could be improved in several aspects. We have not considered or not fully considered the use of some important attributes of objects in conditioning the natural language commands. In describing objects, the shapes of objects provide a powerful descriptor. However, the lack of a pool of reliable detectors for shapes has discouraged us from using it, although it is practically possible for us to experiment with the qualifiers such as *round* or *elongated* and so on. Color is another important and frequently used attribute for defining objects. The qualifiers for colors such as *dark red*, *red*, *tint red* and so on would be very useful if they are well modeled. Another useful extension to this work would be to investigate the spatial relations in 3D space where the added extra dimension will drastically change the way we model the spatial relations. In this work, we have assumed that the human operators and the robots share the same static view of the workspace so that the same spatial relations hold valid for both the human operators and the robots over time. Therefore, a useful extension to the current work would be to model the spatial relations in a shared environment where human operators and robots cooperate and both perceive the same workspace but from different viewpoints.

References

[1] D. L. Chen, R. J. Mooney, "Learning to interpret natural language navigation instructions from observations", in *Proc. 25th AAAI Conf. on Artificial Intelligence*, pp.859-865, August 2011.

[2] T. Kollar, S. Tellex, D. Roy and N. Roy, "Towards understanding natural language directions", in *Proc. 5th ACM/IEEE Int. Conf. on Human-Robot Interaction*, pp259-266, 2010.

[3] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation", in *Proc. 25th AAAI Conf. on Artificial Intelligence*, August 2011.

[4] C. Matuszek, E. Herbst, L. Zettlemoyer and D. Fox, "Learning to parse natural language commands to a robot control system", *International Symposium on Experimental Robotics*, June 2012.

[5] J. Tan, Z. Ju, S. Hand, H. Liu, "Robot navigation and manipulation control based on fuzzy spatial relation analysis, *Int. Journal of Fuzzy Systems*, vol.13, no.4, pp.292-301, 2011.

[6] S. Mohan, A. Mininger, K. Kirk and J. Laird, "Learning grounded language through situated interactive instruction", *AAAI Fall Symposium on Robots Learning Interactively from Human Teachers*, 2012.

[7] I. Bloch, "Spatial reasoning under imprecision using fuzzy set theory, formal logics and mathematical morphology," *Int. Journal of Approximate Reasoning*, vol. 41, no. 2, pp. 77-95, 2006.

[8] J. M. Keller and X. Wang, "A fuzzy rule-based approach to scene description involving spatial relationships," *Computer Vision and Image Understanding*, vol. 80, no. 1, pp.21-41, 2000.

[9] M. T. Escrig, *Qualitative Spatial Reasoning: Theory and Practice, Application to Robot Navigation*, IOS Press, Amsterdam, 1998.

[10] S. Guadarrama, L. Riano, D. Golland, D. Gohring, Y. Jia, D. Klein, A. Abbeel and T. Darrell, "Grounding spatial relations for human-robot interaction", *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2013

[11] M. Tenorth, M. Beetz, "Knowledge processing for autonomous robot control", in *Proc. AAAI Spring Symposium on Designing Intelligent Robots*, 2012,

[12] D. Nyga, M. Beetz, "Everything robots always wanted to know about housework", in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.243-250, October, 2012.

[13] Y. Li and S. Li, "A fuzzy sets theoretic approach to approximate spatial reasoning", *IEEE Transaction on Fuzzy Systems*, vol. 12, no. 6, pp. 745-754, 2004.

[14] K. Liu and W. Shi, "Computing the fuzzy topological relations of spatial objects based on induced fuzzy topology," *Int. Journal of Geographical Information Science*, vol. 20, no. 8, pp. 857-883, 2006.

[15] U. Straccia, "Towards spatial reasoning in fuzzy description logics," In *Proc. of IEEE Int. Conf. on Fuzzy Systems*, pp. 512-517, 2009.

[16] C. Hudelot, J. Atif and I. Bloch, "Fuzzy spatial relation ontology for image interpretation," *Fuzzy Sets and Systems*, vol. 159, no. 15, pp. 1929-1951, 2008.

[17] J. Freeman, "The modeling of spatial relations," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 156-171, 1975.

[18] S. E. Palmer, *Vision Science: Photons to Phenomenology*, MIT Press, 1999.