IIEEJ Paper

Understanding Spatial Knowledge: An Ontology-Based Representation for Object Identification

Lu CAO†, Antony LAM†, Yoshinori KOBAYASHI†,††, Yoshinori KUNO† (*Member*), Daisuke KACHI†††

†Department of Information and Computer Science, Saitama University,
††Japan Science and Technology Agency (JST), PRESTO,
†††Graduate School of Humanities and Social Sciences, Saitama University

<Summary> Spatial descriptions are one of the most effective methods to enable interlocutors to identify which object is being discussed in discourse. In this paper, we propose a framework that can identify an object whose positional relation with another object is indicated verbally by a human. To this end, we construct a spatial knowledge ontology. The ontology is enriched by Description Logic (DL) of concepts, which allows discovering hidden knowledge. We also propose a Spatial Object Dataset that is specifically tailored for our experiments with ontological structures. The dataset currently contains 130 objects and in total of 720 images for object recognition and 360 scenes for spatial recognition. Preliminary experimental results confirmed that the system was able to correctly recognizes human descriptions and identify unknown objects and that understanding human spatial descriptions is efficient for human-machine interaction.

Keywords: spatial knowledge, ontology engineering, human-machine interaction

1. Introduction

Spatial knowledge is ubiquitous in human communications. The comprehension and conveying shared information allow us learning new objects in a novel scene. However, this remarkable ability has still proven to be elusive task for visual object identification models. Unlike humans that can contiguously update the knowledge, it is impossible for robots to know all the objects that exist in the world. Robots may encounter new objects or recognize objects incorrectly even if they are able to obtain new knowledge on-the-fly. In order to perform tasks smoothly, e.g. pick up or deliver us a specific object, the most natural way is to simply describe unknown objects by spatial relations in relation to other known objects, rather than using some other finer concepts, such as fine-grained categories, and brand names. As suggested by [1], we advocate that spatial relations can provide strong cues in identification tasks in that: (1) a spatial relation is less influenced by illumination and scale changes; (2) a spatial relation is more stable than visual features, such as color and shape, etc. For example, a slight variance in color can make an object thoroughly different and may result in frustrating failure; and (3) a spatial relation is independent of object diversity. If an object is replaced by any other object, the spatial relation between the pair of objects will not be changed provided the objects are situated at the same position.

Spatial knowledge has been long an active research field in linguistics and cognitive science[2)-6)]. The frame of reference(we use FoR for short) concept plays an important role, which serves as a coordinate system that allows us to make references to identify target objects as well as to comprehend references made by others. In English culture, people often employ three categories of FoRs: absolute, intrinsic and relative[5)]. The absolute FoR generally refers to the earth's cardinal directions such as North and South, and thus is often used to describe large-scale, and geographical landmarks. In our work, we are interested in table-top space where intrinsic and relative FoRs are more commonly employed. We will introduce these two types of FoRs in Section 3.

How can we represent spatial knowledge? The difficulty is that spatial relations are regarded as somewhat of a *weak sense*. A spatial relation between entities is not something that the entity really 'has' like color or weight. For example, the rose is red can be interpreted as the rose in the object domain being associated with a particular RGB instance--red in the color domain. As a consequence, the color red is independent of time and place. In other words, the color red is always the rose's color. Moreover, spatial change occurs when objects possess different spatial attributes at different times and places. Assume there are two people--A and B

standing side by side. To A, B might be standing at his left side. However, to B, A becomes to be standing at his right side. Thus, to cope with the weakness and uncertainty, state-of-the-arts[7)~11)] in geography, computer vision and robotics present spatial knowledge in an ontology fashion. Although these works have already made significant breakthroughs, they still have limitations in that: 1) they do not provide an explicit mathematical formalism for spatial concepts; and 2) although the proposed ontologies are qualitative and can be used to infer primary spatial relations, they still far away to be applied in image interpretation and object identification.

In this paper, we propose a knowledge-based approach. The framework is able to identify unknown objects incorporating with spatial knowledge. The spatial ontology serves as an intermediate layer and thus enables to discover the hidden knowledge, such as spatial arrangements. To summarize, we highlight here the main contributions of this work.

First, we propose an ontology-based approach that allows to identify new objects. Unlike previous works focus on large-scale spaces, such as offices and corridors, our approach is suitable to table-top space.

Second, we conceptualize spatial knowledge in an ontological fashion. Our proposed ontology is different from previous works[12)~13)] which define spatial relations either object by object, e.g. wings are touched and at the left/right side of a plane, or image by image, e.g. sky is on the top of the image. Instead, our ontology represents spatial knowledge in a more generic way and thus can employ appropriate FoR with respect to different reference objects.

Third, we propose a spatial model for four *directional* spatial relations– front, back, left, and right, which is simply built upon angular deviation via a 2-D projections.

Since spatial relations cannot be used alone, each object within the image should be separately segmented. To do so, we implement off-the-shelf methods to segment[14)] and recognize pre-learned objects[15)]. In this research, we only focus on spatial ontology construction, spatial relation identification, and situated-language processing.

The rest of the paper is organized as follows. In Section 2, we review some relevant work. We introduce the fundamentals of spatial knowledge in Section 3. The details of ontology-based approach is elaborated in Section 4. In Section 5, we introduce the dataset tailored for our experiment. And in Section 6, we conduct human-machine experiments to show the effectiveness of our approach.

2. Related Work

2.1 Qualitative spatial knowledge representation

Spatial knowledge is an interdisciplinary topic combining linguistics and cognition studies. Due to the nature of the interaction between the agent and the environment, there are different types of spatial knowledge. Our work will focus on the *table-top* space, which is defined as a spatial environment that can be immediately and fully observed.

Our work is inspired by several seminal works. The core concept is the FoR which serves as a coordinate system that allows us to make references to identify target objects as well as to comprehend references made by others. Levinson[5)] clarifies English speakers use two distinct classes of FoRs existing for representing the spatial relations between manipulable and small-scale objects in the world: intrinsic and relative. Levelt[4)] analyzes the ambiguities might be arisen in an intrinsic FoR in natural language. G. Schmidt[6)] draws on earlier works and summarizes how intrinsic and relative FoRs are determined by reviewing objects' property.

Another concept that plays an important role relates to the spatial relations. In general, spatial relations can be grouped into three categories: topological, including relations like overlap, contain, and intersect; directional, including relations like front, back, left, and right; distance, including relations like near, and far. Here, we focus on directional relations that has been proven to gain the highest consistency of all between small-scale and manipulable objects in table-top space[16)]. The most commonly used relations are related to three axes of references: front, back, left, right, above, and below. In the 2-D field, we do not consider the top-bottom dimension. Currently, our work contains four main directional relations – front, back, left, and right corresponding to the projective prepositions: in front of, behind, to the left of, and to the right of in natural language.

2.2 Modeling spatial relations

A considerable body of research has been focused on modeling spatial relations. Wang et al.[17)] considers 2-D projections of 3-D spatial scenes and derives a three-level orientation relations from basic (front, back, left, and right) to compound relations (left-front, left-back, right-front and right-back). There are two major branches on generating spatial information from numerical data. Early works[18)~19)] define spatial relations by using fuzzy logic. A. Abella et al.[18)] propose a framework to describe qualitatively 2-D

objects. They define spatial prepositions using inequalities. Fuhr et al.[19] model six spatial relations based on acceptance volumes of a 3-D objects. The main contribution is that they not only take into account the FoR issue but also define a FoR simply by three distinct axes: the front-back, left-right, and bottom-top. In this case, each axis is represented by a pair of reference vectors that are inverse to each other. For example, the front-back axis is given by the vectors **fb** and **bf** directing from front to back and back to front, respectively. However, our proposed model is different from theirs. In this work, we model spatial relations by identifying angular deviations.

Gapp[20] clarifies the interdependencies between angle, distance, and shape with respect to the acceptability of directional relations. For each relation, he presents subjects stimuli with different shapes and requires them scoring the applicability with respect to the combination of 4 different angles (0°, 22.5°,45°, and 67.5°) and distances(130,240,350 and 460 pixels). Results show that the angular deviation gains the predominant effect with $F(3,608) = 521.82$, $p < 0.001$ and $F(3,608) = 487.15$, $p < 0.001$ in the horizontal and vertical experiments, respectively. He[21] accordingly utilizes the observation above on localizing landmarks in large-scale environments where the angle and distance between a reference object and a target object are mapped to a spline function so that the value is between 0 and 1. Another method close in spirit to ours is by Moratz *et al.* [22]. They develop a robotic system to localize and deliver objects placing on the ground (e.g. trash cans, briefcases etc.). The main contributions of their work are two-fold: 1) they simply define the directional relations based on angular deviations; and 2) the system is capable of distinguishing intrinsic FoR from the relative one. However, they only focus on the deictic case which is counted as a sub-category of the intrinsic FoR, such as the bucket is in front of me. More general cases in the intrinsic FoR are beyond the scope.

2.3 Ontology-based spatial knowledge representation

Ontology engineering is widely used to resemble knowledge in a specific domain. An ontology is defined as a set of explicit formal specifications of the terms in the domain and relations among them[23]. It is able to bridge the semantic gap between real-world domains to knowledge. Ontology OWL[24], which goes beyond others as a development language, enable to construct complex knowledge and allow data to be shared and reused across applications. Our ontology relies on OWL-DL, which is based on description logics (DL) [25].

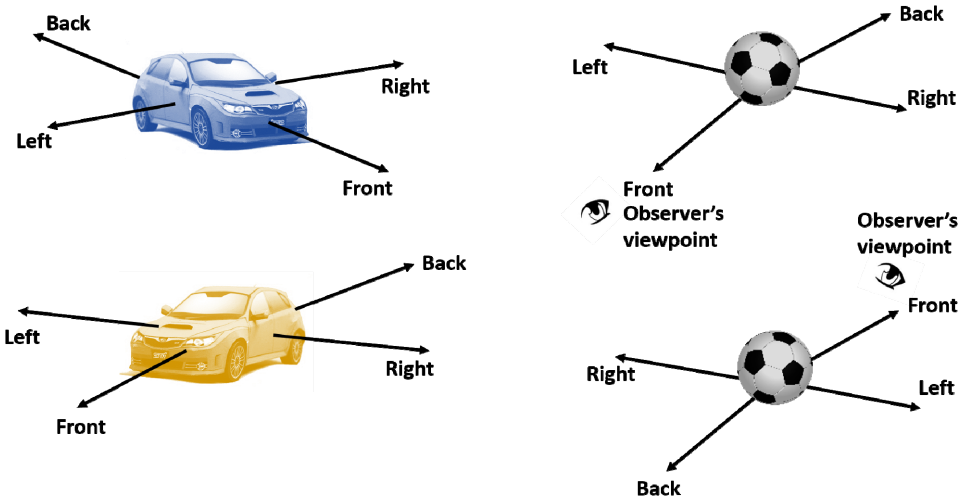
Two recent approaches are closer to our proposed ontology. Mailot et al.[12] present an ontology approach to categorize biological organisms by encoding high-level (color, texture, etc.) features and the spatial relations between the object and their subparts. Hudelot et al. [13] formalizes 6 directional relations: above, below, front, back, right, and left. They consider a spatial relation between 2 entities not as a concept, but as a property. And the corresponding FoR is treated as a concept and dedicated to the representation of uncertain and subjective spatial knowledge by integrating with a fuzzy temporal model. However, both of the works only perform on the objects or images with fixed patterns, where the spatial relation is defined either object by object, for example, wings are touched and at the left/right side of a plane, or image by image, for example, sky is on the top of the image.

3. Qualitative Spatial Knowledge Representation

Before elaborating our approach, it is useful to start with introducing some preliminary knowledge of spatial representation.

As we mentioned earlier, this work concentrates on the directional relations. Directional relations (a.k.a. orientation relations) specify where objects are located relative to one another. There are three elements essential: a target object (TO), a reference object (RO), and a certain frame of reference (FoR) which is a coordinate system that underlies the use of the relation between objects. According to [5), 26)], English speakers employ two distinct classes of FoRs between manipulable and small-scale objects in the world: intrinsic *and* relative.

Intrinsic FoR is a binary relation with respect to two elements: a TO and a RO. Intrinsic FoR requires that ROs should have intrinsic directions that act as a baseline analogous to the due north direction on the earth's surface, that is, the intrinsic front, back, left, and right. These directions are extracted from the corresponding inherent sides of ROs. Due to the asymmetry in the front-back dimension, the intrinsic front and back directions have a privileged status, while the intrinsic left-right directions of objects are rare. In [27], A. Galton points out that dolls and all of animal species including human beings that have perceptual apparatus have intrinsic front sides. Miller and Johnson-Laird[3] specify that objects such as cars, bullets, and arrows which possess characteristic of direction of motion have intrinsic sides, and objects such as cameras,



(a) Intrinsic FoR spatial arrangement (b) Relative FoR spatial arrangement

Fig. 2 Spatial arrangement in intrinsic and relative FoRs

chairs and screens which have functional parts also have intrinsic sides[3]. As in **Figure 1**, one might say the soda-can is in front of the robot, which means the soda-can is located at the front side of the robot's body where the front region of the body is inherent. We note that once the intrinsic front is determined, the back, left, and right directions can be deduced accordingly. As shown in **Figure 2**(a), the direction which is opposite to the front counts as the *back* direction. The left-front-right-back then follows an anticlockwise path around the RO.

Relative FoR is a ternary relation with respect to three elements: a TO, a RO, and a viewpoint (we call VP for short). The directions of ROs are extracted from interlocutors' viewpoint (either speaker or listener) from which the ROs are seen. In general, any objects existing in the world can be applied in relative FoR. In Figure 1, the man might say the soda-can is behind the tissue-box, which means the soda-can is located at the back side of the tissue-box when viewing from the man's viewpoint. Since there are no *intrinsic front*, *back*, *left* or *right* directions generated by the tissue-box in the horizontal and vertical dimensions, the *back* side is determined by the man's viewpoint. In other words, it is the position of the man that determines how the space around the soda-can is arranged. In principle, the space is arranged as the same as the intrinsic FoR. We emphasize that the front is facing directly towards the observer, which is parallel to the observer's viewpoint as illustrated in Figure 2(b).

4. The ontology - based approach

4.1 Overview

Our goal is to identify unknown objects in novel scenes by comprehending the spatial knowledge within. We use **Figure 3** and Algorithm 1(see **Table 1**) to illustrate the overall approach. Given a novel scene image, we first segment the image with semantic regions. The segmented objects are recognized into one of the categories, if the exemplars are pre-learned. For unlearned objects, they are labeled as unknown objects. This can be done either manually or automatically. With the recognition result, we begin the interactive process of identifying an unknown object. Our goal is to generate a tuple to represent the knowledge via ontological retrieving and inferring. With a referral sentence input by the user, e.g. the pen is in front of the can, we use the Stanford Part of Speech (POS) tagger[28] to tag every word. We note that the spatial prepositions are

Table 1 overall framework

Algorithm 1 Overall framework
Input: Image I
Segment image with semantic regions
Recognize objects within using trained models
repeat(if there are any unknown object)
Tag referral sentence input by the user
Retrieve in the ontology
Infer the hidden knowledge by rules
Identify TO using the spatial relation model
until user satisfied or all objects examined

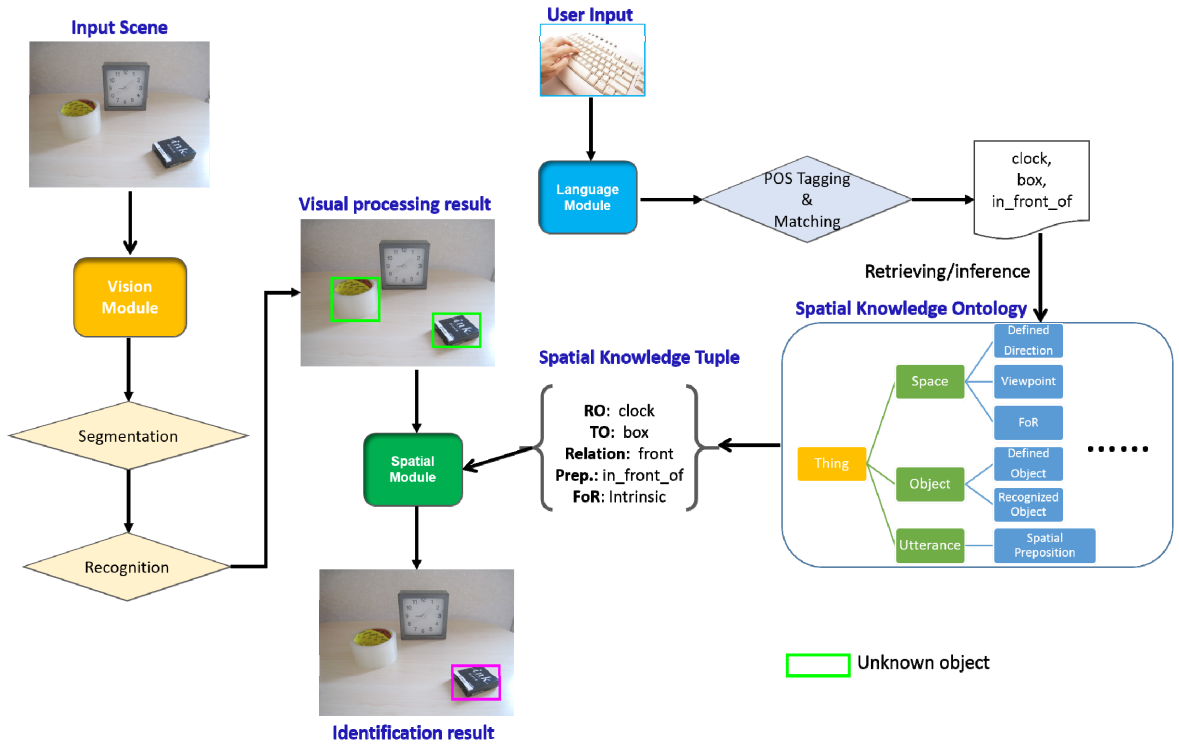


Fig.3 Illustration of the framework of the system

annotated as adjectives. The tagged words/phrases are retrieved in the ontology if they are defined spatial prepositions and objects. For a noun that cannot be retrieved in the ontology, we use WordNet[29] to determine if it is a defined *synset* and then infer it if is a RO or TO. With the tuple such as <TO spatial_relation RO>, we infer the underlying FoR by rules. Finally, we use the resulting tuple <TO [spatial_relation FoR] RO> to identify the TO with the computational spatial model. If there are several unknown objects, we repeat this process till user satisfies or we have identified all objects in the image.

4.2 The spatial knowledge ontology

In this section, we describe the spatial ontology in detail. Specifically, Sect. 4.2.1 – 4.2.3 describe the three concepts we defined. In order to infer the hidden spatial knowledge, Sect. 4.2.4 introduces the rules. By retrieving the ontology, our goal is to generate a triple repository to store the spatial information, such as <TO spatial_relation RO>. Algorithm 2 illustrates the procedure. The ontology is constructed based on DL [13, 25]. We briefly introduces the fundamental syntax in Appendix.

The spatial ontology is represented by three classes: Space, Object and Utterance. We note that the terminologies are different between spatial representation and natural language. In space domain, we cannot say an object has a front/back /left/right relation. Instead, it is feasible to say an object has a front/back/left/right direction. On the other hand, in language domain, we use spatial prepositions such as in front of and to the back of map the direction concept in the space domain.

4.2.1 Knowledge class – Space

The Space class consists of three concepts: FoR, Direction, and VP.

FoR – the frame of reference. Intrinsic and relative FoRs are instantiated in this domain.

Direction – the general concept. It subsumes defined topological, directional and distance directions.

VP — the viewpoint, which is an indispensable entity in relative FoR. Currently, we include two sub-concepts: the speaker's viewpoint and the listener's viewpoint with four instances: upper view, frontal view, left-profile view, and right-profile view.

As described in Sect. 3, a TO, RO, and at least one defined directions constitute the general concept FoR, which is defined by using the in Eq. (1).

$$\begin{aligned} FoR \equiv & \exists\ hasRO.Recognized\_Object \\ & \sqcap = 1\ hasRO \\ & \sqcap \exists\ hasTO.Defined\_Object \end{aligned} \tag{1}$$

$$\sqcap = 1\ hasTO$$
$$\sqcap \exists\ has\_Direction.Defined\_Direction$$

where the Defined_Direction belongs to the general concept Direction.

$$Defined\_Direction \sqsubseteq Direction \quad (2)$$

Eq. (1) contains two properties: hasRO and hasTO. The range of these two properties - Recognized_Object and Defined_Object can be found in the Object domain. For example, in the sentence the dry battery is in front of the camera has a property of hasRO with the value camera and a property of hasTO with the value dry battery.

Specifically, as sub-set of the FoR the intrinsic and relative FoRs can be described by DL as:

$$Intrinsic\_FoR \sqsubseteq FoR \quad (3)$$

$$Relative\_FoR \sqsubseteq FoR$$
$$\sqcap \exists\ hasVP.\ VP \quad (4)$$

We stress that VP is a necessity of relative FoR which can be either specified by the listener or the speaker. However, in most cases, the speaker's viewpoint is set as the default VP if no other VP is specified. Thus the VP can be defined as:

$$VP \equiv \geq 1\ hasVP$$
$$\sqcap \exists\ viewFrom.\ \{Speaker\}$$
$$\sqcap \ni\ hasVPInstance.\ \{\text{vp\_}value\} \quad (5)$$

where {vp_value} is the set of individuals.

We also specify some properties of the directions, such as symmetric, transitive, and functional. For example, the front-back and left-right are complement to each other. In this case, the Left direction can be defined as:

$$Left\_Direction \sqsubseteq Direction$$
$$\sqcap\ Directional\_Direction \quad (6)$$
$$\sqcap \exists\ inverse.Right\_Direction$$

4.2.2 Knowledge class – Object

Two concepts are defined in this domain: Recognized_Object and Defined_Object.

Defined_Object – a subset of the Object concept. Note that we do not elaborate any abstract concept, only objects that have not be recognized yet are assembled here.

Recognized_Object – a subordinate concept adhering the Defined_Object, which represents a set of ROs. They are basic-level categories of objects collected by ourselves, such as balls, screens, bottles, etc. (Sect. 5.1), and are recognized by the vision module.

Table 2 ontology retrieving and inferring

Algorithm 2 Ontology retrieving and inferring
Require: WordNet Dictionary *W*
Rule Set *R*
Input: Referral sentence *U*
Output: Tuple *T*
Initialize Tuple: *T*= < >
C ← SENTENCETAGGING(*U*)
for each ISNOTEMPTY(*C{i}*)
if ISNOUN(*C{i}*)
obj_candidate ← ONTOLOGYRETRIEVAL.DEFINED(*C{i}*)
if ISNOTEMPTY(*obj_candidate*)
ro_candidate ← ONTOLOGYRETRIEVAL.RECOGNIZED(*C{i}*)
if ISNOTEMPTY(*ro _candidate*)
RO ← *ro _candidate*
else
sim_score ← ONTOLOGYINFERRANCE (*W, R, C{i}*)
if *sim_score* < *threshold*
TO ← *C{i}*
else *RO* ← *C{i}*
else OUTPUT("Cannot resolve the sentence!")
else if ISADJECTIVE (*C{i}*)
sp_candidate ← ONTOLOGYRETRIEVAL (*C{i}*)
if ISNOTEMPTY(*sp_candidate*)
SP ← *sp_candidate*
else OUTPUT("Cannot resolve the sentence!")
end for
FoR ← ONTOLOGYINFERRANCE (*R, RO, SP*)
T ← GENERATETRIPLE (*R, RO, TO, SP, FoR*)

Given a name, to obtain semantic *synsets* (e.g. whether the query is an existing object identity), we import the WordNet database[29]. We make use of the hypernym (super-term) and meronym (contains) relations. We pick up some distinguished constituent parts from the meronym and determine whether a RO is able to generate an intrinsic direction by finding the individual's distinguished (inherent) part. For example, in the sentence of the battery is in front of the camera, the RO--camera can be described as a kind of equipment, and has a distinguished part of lens. This can be written as:

$$Camera \equiv \exists\ isRo.Recognized\_Object$$
$$\sqcap \exists\ hasSuperClass.Equipment \quad (7)$$
$$\ni\ hasDistinguishedPart.\ \{lens\}$$

In future work, we would like to extend the ontology to infer object affordance and functionality in an ontological hierarchy, where given a name, the ontology not only can recognize the object, but infer its hypernyms/hyponyms (super-term/sub-term), e.g., hypernyms of computer are machine, device, etc., and how to interact with it, e.g. type on.

4.2.3 Knowledge Class – Utterance

The Utterance class relates to the language module (see Sect. 4.4). It is used to parse the spatial prepositions of what humans use and capable of mapping the Direction concept onto the Space class. Note that this is not a strict *one-vs-one* mapping due to linguistic diversity. For example, both of the spatial prepositions such as in front of and at the front side correspond to front concept in the Direction domain. Thus, given a referral sentence, it can be described as:

$$\begin{aligned} &Spatial\_Description \\ &\quad \equiv \exists\ hasTO.Defined\_Object \\ &\quad \sqcap = 1\ TO \\ &\quad \sqcap \exists\ hasRO.Recognized\_Object \\ &\quad \sqcap = 1\ RO \\ &\quad \sqcap \exists\ hasSpatial\_Prepostions.Spatial\_Preposition \end{aligned} \tag{8}$$

where

$$\begin{aligned} &Spatial\_Preposition \\ &\quad \equiv \exists\ inRelationwith.Defined\_Direction \\ &\quad \sqcap = 1\ Defined\_Direction \end{aligned} \tag{9}$$

4.2.4 Reasoning about hidden knowledge

Logical inference is able to identify the hidden knowledge using well-defined rules. Currently, we define two types of rules: unidirectional R^U and bidirectional R^B. The R^U rules ensure the reasoning between properties, and instances within the same ontological classes, while R^B is able to transfer knowledge between different ontological classes.

Rule 1: matching of RO's name

People may use different terms to express the same sense. For example, both the monitor and display can express the sense of computer monitor. It is thus necessary to evaluate whether a query *X*, often a name of RO, is a synonym of the recognized object *Y*. To do so, we first apply the *unidirectional* rule:

IF X is RO

AND Y is recognized

AND similarity between X and Y is greater than Threshold T

THEN X is synonym of Y

We then use WordNet Similarity for Java (ws4j)[30]—a java reimplementation of WordNet-Similarity[31]. The WS4J provides several published semantic relatedness algorithms. We use the WUP[32] to estimate how semantically close between a querying and an existing *synset* defined in WordNet. We set the threshold T as 0.9. If T is less than 0.9, we treat the querying as a false negative, namely, maybe the observer names the object incorrectly. Assume one names a mug as coffee mug, the relatedness is Rela(mug, coffee_mug) = 0.957, or the relatedness between camera and digital camera is Rela (camera, digital_camera) = 0.952. However, the relatedness between cup and can is only Rela (cup, can) = 0.89.

Rule 2: matching TO's name

This rule is *unidirectional* as well as rule 1. If a query *X* cannot be retrieved in the Recognized_Object concept where the T is less than 0.9, but belongs to the Defined_Object, we assert that *X* is a TO. The rule is defined as:

IF X is Defined

AND X is *NOT recognized*

THEN X is a TO

Rule 3: Reasoning about the FoR

The bidirectional *rule* is applied across the different ontological classes. We consider the intrinsic FoR can be inferred from the following rule:

IF X is recognized object

AND X has distinguished part P

THEN X has Intrinsic Front

AND X generates Intrinsic FoR

Otherwise, we assert the FoR is relative. Once the *relative* FoR is determined, we find viewpoint from the utterance, and update the ontology.

4.2.5 Ontology implementation and the results

We use OWL Protégé 5.0[33] to construct the ontology. We implement the Apache Jena framework[34] as the underlying library to load and infer the OWL model. For instance, the referral sentence the pen is in front of the camera can be viewed as *concept* and denoted by DL as:

$$\begin{aligned} &C_0 \equiv \exists\ hasFOR.FOR \\ &\quad \sqcap \exists\ has\_Spatial\_Object.Defined\_Object \\ &\quad \sqcap \exists\ has\_Spaial\_Descritpion.Spatial\_Description \end{aligned} \tag{10}$$

The object individuals pen and camera can be retrieved from the Object class. And the spatial preposition in front of is represented in the Utterance class. After retrieving the Regonized_Object, we assert that the object individual

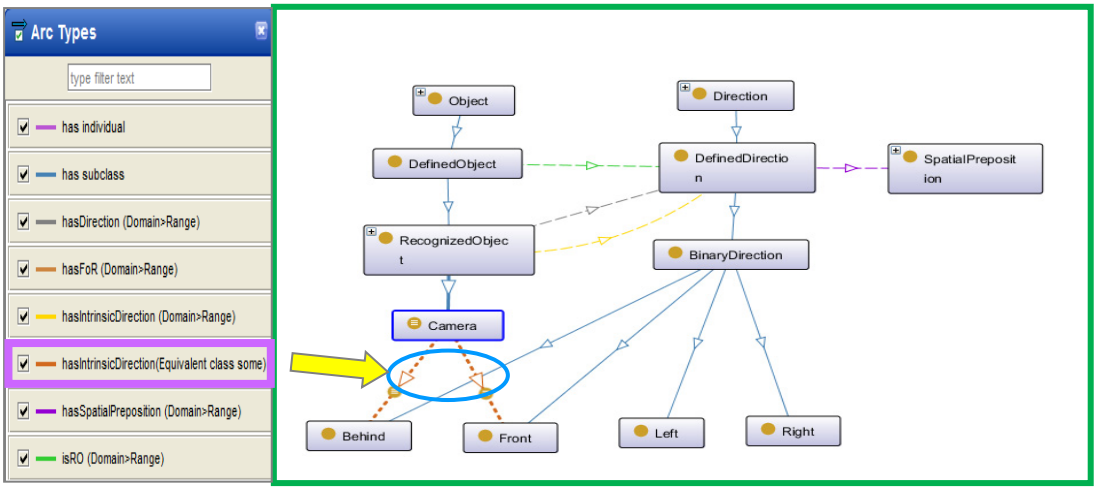


Fig.4 A FoR reasoning result. When camera serves as the RO in a referral utterance, by applying rule 3, the intrinsic FoR with intrinsic front and back directions can be inferred

camera serves as the RO, while pen is the TO (rule 2). Then we update the FOR concept in the Space class. As a result, Eq. (1) can be written as:

$$FoR \equiv \exists\ hasRO.\ Camera \sqcap \exists\ hasTO.\ Pen \sqcap \exists\ has\_Direction.\ Front \quad (11)$$

where the concept camera is defined in Eq. (7). Finally, by using rule 3, we can infer that camera has an intrinsic FoR. The result is shown in **Figure 4**.

4.3 Modeling spatial relations

Based on Hernandez[17] and Moratz[22]'s models, we introduce a geometric method to model the directional relations. The model is built upon 2-D view. With the scene being viewed from above, all the objects are represented in a planar view. According to Figure 2, the reference axis is along a RO's front direction and the reference plane, where a RO is centered, is thus partitioned into front, behind, left, and right regions. **Figure 5** illustrates

the configuration. In order to identify the partitions geometrically, we refer to the angle θ between the reference axis and the connected line from the TO to the RO in Eq. (12):

$$\begin{aligned} &TO\ front\ RO: && 0 \le \theta < \pi/4 \\ &\quad \text{or} && 7\pi/4 \le \theta \le 2\pi \\ &TO\ left\ RO: && \pi/4 \le \theta < 3\pi/4 \\ &TO\ behind\ RO: && 3\pi/4 \le \theta < 5\pi/4 \\ &TO\ right\ RO: && 5\pi/4 \le \theta < 7\pi/4 \end{aligned} \quad (12)$$

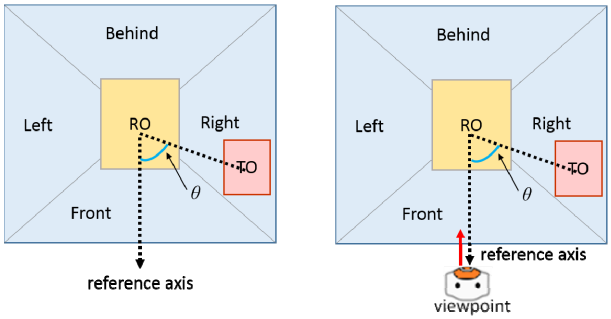


Fig.5 The computational model of intrinsic and

Table 3 Examples of inputs and the corresponding responses

Input	Response
How/how many objects can you see (?)	"[number of detected objects]"
What/what are they (?) What/what is it (?)	"[name of the detected objects]"
The toolbox is in front of the cup (.)	"Is this one?"/ "I can't see it"
Can/can you see the can (?) Can /can you see 3 cups (?)	"Yes, I can."/ "No, I can't."

4.4 Situated dialogue processing

The Language module treats an input as a natural language utterance, which contain references to objects by names and descriptions of spatial locations in relation to

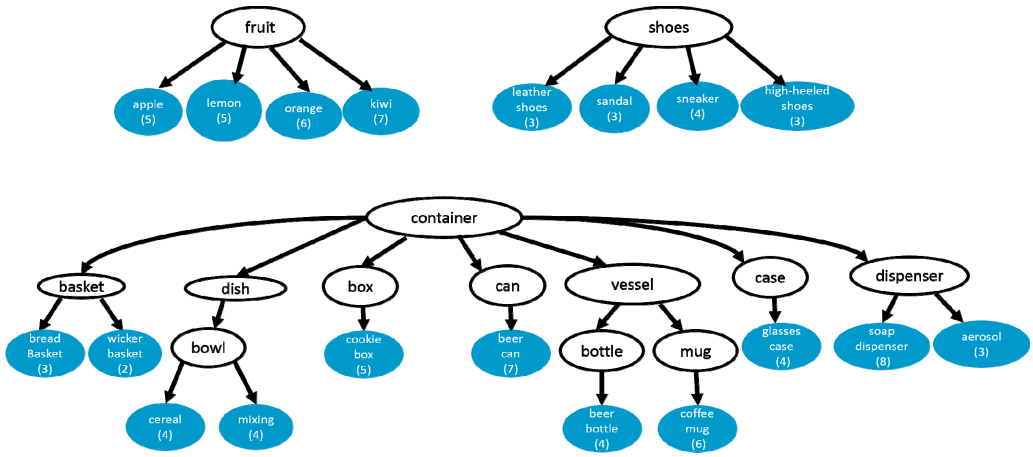


Fig.6 The fruit, shoes, container subtrees of the Spatial Object Dataset. The number of instances in each leaf category (shaded in blue) is given in parentheses

other objects. To extract names and spatial terms from an utterance, we rely on the Stanford Part of Speech tagger[28]. When executing inputs, we restrict a limited of syntactic formats to reduce matching complexity. For example, the query Can you see the soda-can (?) is matched by the keyword **Can** with case insensitivity. **Table 3** represents the syntactic formats. If the inputs cannot be interpreted, users will receive I don't understand as a response. Then users will reconsider and take further attempts.

5. The Spatial Object Dataset

Since there is no publicly dataset available, in order to tailor our tasks, we present a Spatial Object Dataset with samples of 130 objects and 300 scenes. The dataset is presented with 400 × 300 pixel resolution color images. All images are taken with a Canon IXY910IS digital camera.

5.1 Training dataset - Objects

Currently, the Spatial Object Dataset contains 130 objects and in total of 720 images. The chosen objects are commonly found in home and office environments, including office workspaces, living rooms, and kitchen areas. Objects are organized into a WordNet[29] hierarchy with hypernym/hyponym relations. The dataset contains 4 levels: basic, subordinate, superordinate and abstract. Apparently, the basic-level category (e.g., apples, cars, etc.) is the easiest for humans to organize knowledge. At the next lower levels, subordinate categories can provide fine-grained knowledge, such as soda can and coffee mug. The secondary-level subtrees in the hierarchy adhering to the basic level is superordinate categories, which are a higher degree of abstraction, such as vessel and dish. Categories such as device and container are the *abstract-level* subtrees in the hierarchy, which concentrates a high degree of world knowledge. We collected objects from 4 areas: fruit & vegetables; clothes & shoes; container and device. **Figure 6** shows the subtrees in the current version of the dataset. The leaf nodes are shaded in cyan, and the number of object instances in each category is given in parentheses.

We resort the way of building the dataset by a few prototypes. Each object is presented by 5 or 6 images from semi-upper viewpoint and scale slightly changed in canonical (frontal) pose. In **Fig. 7**, the first two rows show some example objects of the dataset.

5.2 Testing dataset-Scenes

We collected the scene images in home and office environments with the same manner as described in Sect. 5.1. In current version, we collected 360 scenes. Each scene has at 2-4 objects with at least one recognized object and one unknown object which indicate the RO and TO respectively that people can refer to. All of the objects are basically at their frontal view. A snapshot is also shown in Figure 7.

6. Experiment

6.1 Image segmentation and object recognition

In order to train and learn object models, we implement strongly-supervised deformable part-based model[15]. The model not only is able to category the objects, but also to

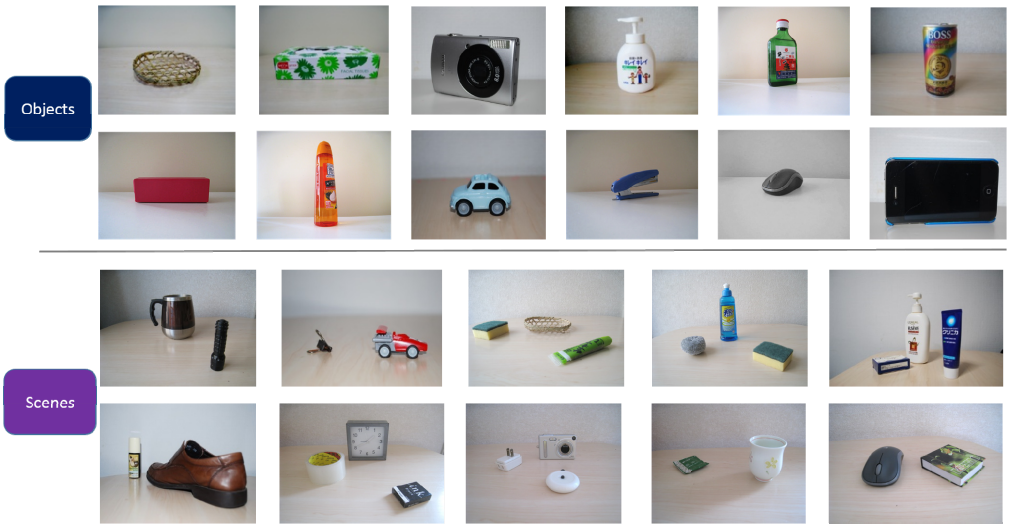


Fig.7 A snapshot of object and scene exemplars from the Spatial Object Dataset

recognize object parts such as camera lens, etc. For a scene image, we first use the multiscale combinatorial Grouping (MCG)[14] to semantically segment the image. The segmented objects are recognized into one of the categories, if the exemplars are pre-learned. For unlearned objects, they are labeled as unknown objects.

6.2 Ground truth

To evaluate the performance, we relied on human labors to verify each candidate scenes collected in the dataset. For a given RO and a TO, we required 15 students to select a best answer. The ground truth was defined by majority vote.

6.3 Experiment scenes and setting

200 scenes were chosen in this experiment in which recognized and unrecognized objects were well segmented and recognized. Each image contained one or two unseen objects positioned around a recognized object.

6.4 Procedure

Twenty university students, who were trained to familiar with our strategy were invited to take part in the experiment. They were required to sit in front of a computer, and received ten images at a time. Each image was presented for two minutes. If the user did not succeed within the prescribed time, the system skipped and moved to the next image. If a command cannot be interpreted, users received a response such as I don't understand so that they could reconsider the strategy in the further attempt.

6.5 Experiments

6.5.1 Experiment 1: one-shot experiment

This experiment was designed to evaluate the accuracy of the system. We only allowed the subjects to input a referral sentence which directly referring to the TO:

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

With all the testing images, we collected in total of 352 utterances, corresponding to an average of 17.6 per person. There were 24(6.8%) utterances that cannot be interpreted because of thoroughly syntactic form or spelling errors. In general, of the 328(93.2%) valid sentences, 5 cannot be executed correctly because participants confused right with left regions. In the rest of 323 utterances, there were 82 utterances corresponding to the *front* preposition in which 77 led to success with the accuracy being 94.0%. In the 5 unsuccessful trials, we noticed that the TO was placed at the proximate orthogonal region around the RO. This region is not considered as a good acceptance region of front by Logan and Sadler[35], but still acceptable so that it was difficult to distinguish it was a front or left. As a result, three subjects used the compound prepositions - left front. The other two people considered the TO was left to the RO. The same phenomenon occurred in the back trials. Of 80 utterances, 70 were achieved success with the accuracy being 87.5%, six people used the term right behind, and the other four people thought of the TO was being placed to the right of the RO. The accuracy was slightly increased in the left and right trials. Perhaps it was easier for human to

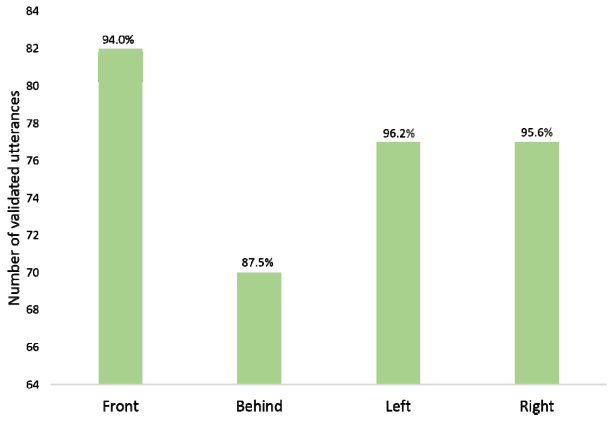


Fig.8 Results on four spatial relations. The accuracy is displayed on top of each bar

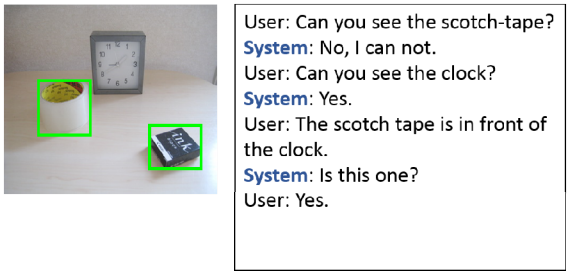


Fig. 9 Transcript of user-leading strategy

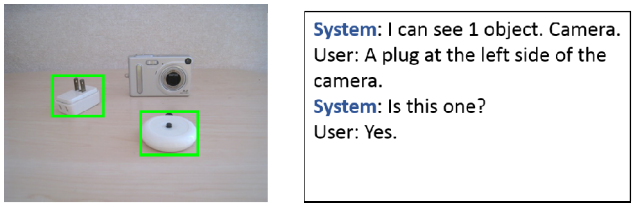


Fig.10 Transcript of system-leading strategy

identify directions in the horizontal dimension than the vertical one. There were 80 and 81 utterances corresponding to the left and right with the accuracy being as high as 96.2% (77 utterances) and 95.6% (77 utterances), respectively. **Figure 8** illustrates the accuracy.

6.5.2 Experiment 2: interactive experiment

In this experiment, subjects were allowed to interact with the system. Instead of performing on all the testing scenes, we picked up 100 images and suggested two types of interactive strategies to the users. We note that at current stage the input sentences are restricted with a fixed pattern.

Strategy 1: user-leading strategy

Users incrementally provide the scene information. The most advantage is that the strategy reveals users' intention and allows users to find out what the system is able to understand. For example, at the beginning, users often prompted the system by querying Can you see the scotch-tape? and waited for a feedback. **Figure 9** shows the transcript.

Strategy 2: system-leading strategy

Instead of user predominating interaction, the system first reports how many/what objects that can be seen at the beginning by saying, I can see a book and a stapler or I can see 1 object. Camera. Then users provide spatial information. We show the transcript in **Figure 10**.

We compared the accuracy by using two strategies. With the user-leading strategy, of 170 utterances, 156 led to successes with the accuracy approximately being 91.89%. By contrast, of 148 collected utterances by using system-leading strategy, there were 140 validated sentences led to success where the accuracy was increased as high as 95.83%. Despite of the syntax errors, the main reason was the objects named by users were not able to be understood by the system. For example, people may use object's names, such as Oreo rather than object categories - box or bag. Performance in the system-leading strategy outperformed the user-leading strategy with 91.89% vs. 95.83%. The main reason is that it allowed users to understand how capable of the system is. In some of the trials, we observed that the RO recognized by the system was not the one that seen by the users (false-positive cases). For example, an orange might be recognized as an apple due to its color variation. **Figure 11** shows the comparison result. We also evaluated the accuracy of four spatial prepositions on two strategies, which is shown in **Table 4**. As a result, the left and right outperformed the front and behind, which was in accordance with the result we observed in Sect. 6.5.1.

To summarize, if the system is capable enough to recognize RO candidates accurately, the user-leading strategy is a preference. Otherwise, users would like to take the system-leading strategy. Perhaps without any prior knowledge provided, the user-leading strategy increases the risk of failure. This observation is extremely valuable for us

Table 4 accuracy of 4 spatial prepositions on 2 strategies

Spatial Prep.	User-leading Strat.	System-leading Strat.	Avg.
Front	95.44%	95.38%	95.41%
Behind	92.24%	92.89%	92.57%
Left	95.67%	95.73%	95.70%
Right	95.43%	95.68%	95.56%

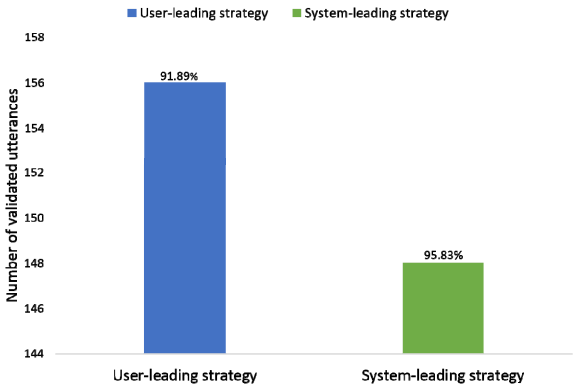


Fig.11 User-leading vs. system-leading strategy. The average accuracy is displayed on top of each bar

to design an interaction scheme and achieve natural, unconstrained communication in the future work.

7. Conclusion

In this paper, we presented a spatial recognition approach of integrating with ontology hierarchies. The geometric spatial model can map projective spatial prepositions in language onto characteristic points on a 2-D reference plane, and the ontology is able to infer underlying knowledge of space, for example, the FoR. We also proposed a Spatial Object Dataset that is specifically tailored for our experiments with ontological structures. Preliminary experimental results confirmed that the system was able to correctly recognizes human descriptions and identify unknown objects and that understanding human spatial descriptions is efficient for human-machine interaction.

Possible directions for future work may be as follows. First, we observed the failure cases mostly occurred in front and behind trials and attributed to the occlusion issue. The worst case occurred when the TO and RO are collinear. To address the problem, we need to obtain 3-D data such as RGB-D style. Second, we are interested in presenting a knowledge-based (KB) network to transfer knowledge. For instance, if object A is in front of B, and B is in front of C, we can infer that object A is also in front of C. Another example is if we know object A has an intrinsic direction, we can infer its hypernyms/ hyponyms also has intrinsic direction. This allows us to recognize objects by their attributes and parts, and learn the visual similarities. Third, in nature scenes, objects are usually arbitrarily placed, a measure that would allow adjusting the main axis direction. Perhaps finding the front direction is the most interesting and crucial case as it weighs the highest priority in all four directions. Meanwhile, more candidate images with large scale and viewpoint changes should be collected for further experiment use.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers 23300065, 26240038.

References

1) B. Landau, R. Jackendoff: "'What' and 'Where' in Spatial Language and Spatial Cognition", Behavioral and Brain Sciences, Vol.16, No.2, pp.217-238 (1993).
2) H. Clark, Space, Time, Semantics, and the Child, Academic Press, pp.28-64 (1973).
3) G. Miller, P. Johnson-Laird, Language and Perception, Cambridge University Press (1976).
4) W. Levelt, Some Perceptual Limitations on Talking about Space, VNU Science Press, pp.323-358 (1984).
5) S. Levinson, Frames of Reference and Molyneux's Question: Cross-linguistic Evidence, The MIT Press, pp.109-170 (1999).
6) G. Schmidt: "Various Views on Spatial Prepositions", AI Magazine, Vol.9, No.2, pp.95-105 (1988).
7) R. Casati, B. Smith, A. C. Varzi: Ontological Tools for Geographic Representation, IOS Press, pp.77-85 (1998).
8) E. Klien, M. Lutz: "The Role of Spatial Relations in Automating the Semantic Annotation of Geodata", Lecture Notes in Computer Science, Vol.3693, pp.133-148 (2005).
9) S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, M. Strintzis: "Knowledge-Assisted Semantic Video Object Detection", IEEE Trans. On Circuits Systems for Video Technology, Vol.15, No.10, pp.1210–1224 (2005).
10) P. Dominey, J. Boucher, T. Inui: "Building an Adaptive Spoken Language Interface for Perceptually Grounded Human–Robot Interaction", Proc. of IEEE/RAS International Conference on Humanoid Robots (IROS 2004), pp.168–183 (2004).
11) D. Han, B. You, Y. Kim, I. Suh: "A Generic Shape Matching with Anchoring of Knowledge Primitives of Object Ontology", Lecture Notes in Computer Science, Vol.3646, pp.473–480 (2005).
12) N. Mailot, M. Thonnat: "Ontology Based Complex Object Recognition", Image and Vision Computer Journal, Vol.26, No.1, pp.102-113 (2008).
13) C. Hudelot, J. Atif, I. Bloch: "Fuzzy Spatial Relation Ontology for Image Interpretation", Fuzzy Sets Systems, Vol.159, No.15, pp.1929-1951 (2008).
14) P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, J. Malik: "Multiscale Combinatorial Grouping", Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2014), pp.328-335 (2014).

A-1 DL syntax, example and interpretation

Constructor	Syntax	Example	Semantics
Atomic concept	C	Human	$C^{\mathcal{L}} \subseteq \Delta^{\mathcal{L}}$
Individual	a	Bob	$a^{\mathcal{L}} \in \Delta^{\mathcal{L}}$
Atomic role	r	has-sibling	$R^{\mathcal{L}} \subseteq \Delta^{\mathcal{L}} \times \Delta^{\mathcal{L}}$
Conjunction	$C \sqcap D$	Human ⊓ male	$C^{\mathcal{L}} \cap D^{\mathcal{L}}$
Disjunction	$C \sqcup D$	Human ⊔male	$C^{\mathcal{L}} \cup D^{\mathcal{L}}$
Negation	$\neg C$	¬ Human	$\Delta^{\mathcal{L}} \setminus C^{\mathcal{L}}$
Existential restriction	$\exists r.C$	∃has-sibling. Girl	$\{x \in \Delta^{\mathcal{L}} \mid \exists y \in \Delta^{\mathcal{L}}: (x,y) \in R^{\mathcal{L}} \wedge y \in C^{\mathcal{L}}\}$
Universal restriction	$\forall r.C$	∀has-sibling. Human	$\{x \in \Delta^{\mathcal{L}} \mid \forall y \in \Delta^{\mathcal{L}}: (x,y) \in R^{\mathcal{L}} \Rightarrow y \in C^{\mathcal{L}}\}$
Value restriction	$\ni r.\{a\}$	∋has-sibling. {Tom}	$\{x \in \Delta^{\mathcal{L}} \mid \exists y \in \Delta^{\mathcal{L}}: (x,y) \in R^{\mathcal{L}} \Rightarrow y = a^{\mathcal{L}}\}$
Number restriction	$(\geq nR)$	(≥ 2 has-sibling)	$\{x \in \Delta^{\mathcal{L}} \mid \lvert\{y \mid (x,y) \in R^{\mathcal{L}}\}\rvert \geq n\}$
	$(\leq nR)$	(≤ 2 has-sibling)	$\{x \in \Delta^{\mathcal{L}} \mid \lvert\{y \mid (x,y) \in R^{\mathcal{L}}\}\rvert \leq n\}$
Subsumption	$D \sqsubseteq C$	Man ⊑ Human	$D^{\mathcal{L}} \subseteq C^{\mathcal{L}}$
Concept definition	$C \equiv D$	Father ≡ Man⊓ ∃has-child. Human	$D^{\mathcal{L}} = C^{\mathcal{L}}$
Concept assertion	$a:D$	Bob : Man	$a^{\mathcal{L}} \in D^{\mathcal{L}}$
Role assertion	$(a,b):R$	(Bob, Mary) : has-sibling	$(a^{\mathcal{L}}, b^{\mathcal{L}}) \in R^{\mathcal{L}}$

15) H. Azizpour, I. Laptev: "Object Detection Using Strongly-Supervised Deformable Part Models", Proc. of European Conference on Computer Vision (ECCV 2012), pp.836-949 (2012).

16) X. Wang, P. Matsakis, L. Trick, B. Nonnecke, M. Veltman: "A Study on How Humans Describe Relative Positions of Image Objects", Lecture Notes in Geoinformation and Cartography, pp.1-18 (2008).

17) D. Hernandez: Qualitative Representation of Spatial Knowledge, Springer-Verlag Berlin Heidelberg (1994).

18) A. Abella, J. Kender: "From Images to Sentences via Spatial Relations," Proc. of the International Conference on Computer Vision Workshop on Integration of Image and Speech Understanding (ICCV Workshop 1999), pp.117-147 (1999).

19) T. Fuhr, G. Socher, C. Scheering, G. Sagere, A Three-dimensional Spatial Model for the Interpretation of Image Data, Lawrence Erlbaum Associates, Inc, pp.103–118 (1998).

20) K. Gapp: "Angle, Distance, Shape, and Their Relationships to Projective Relations", Proc. of Annual Conference of the Cognitive Science Society (CogSci 2015), pp.112-117 (1995).

21) K. Gapp: "From Vision to Language: A Cognitive Approach to the Computation of Spatial Relations in 3D Space", Proc. of the European Conference on Cognitive Science and Industry (ECCSI 1994), pp.339-358 (1994).

22) R. Moratz, 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, No.1, pp.63-107 (2006).

23) T. Gruber: "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal of Human-Computer Studies, Vol.43, No.5-6, pp.907-928 (1995).

24) OWL, http://www.w3.org/TR/owl2-primer/ (2012).

25) F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider: The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press (2003).

26) A. Majid, M. Bowerman, S. Kita, D. Haun, S. Levinson: "Can Language Restructure Cognition? The Case for Space", Trends in Cognitive Sciences, Vol.8, No.3, pp.108-114 (2004).

27) A. Galton, Qualitative Spatial Change, Oxford University Press (2000).

28) K. Toutanova, D. Klein, C. Manning, Y. Singer: "Feature-rich Part-of-speech Tagging with A Cyclic Dependency Network", Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003), pp.252-259 (2003).

29) C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books (1998).

30) WordNet similarity for Java (ws4j), https://code.google.com/archive/p/ws4j/ (2013).

31) T. Pedersen, S. Patwardhan, J. Michelizzi: "WordNet::Similarity - Measuring the Relatedness of Concepts", Proc. of National Conference on Artificial Intelligence (AAAI 2004), pp.1024-1025 (2004).

32) Z. Wu, M. Palmer: "Verb Semantics and Lexical Selection", Proc. of the Annual Meeting of the Associations for Computational Linguistics (ACL 1994), pp.133-138 (1994).

33) Protégé, http://protege.stanford.edu/ (2015).

34) Apache Jena, https://jena.apache.org/ (2011).

35) G. Logan, D. Sadler, A Computational Analysis of the Apprehension of Spatial Relations, The MIT Press, pp.493-530 (1999).

Appendix

This section introduces a formal representation of Description Logic (DL). In DL, a semantics is associated with *concepts*, *roles* and *individuals* in the form of $\mathcal{L} = (\Delta^{\mathcal{L}}, \cdot^{\mathcal{L}})$, where $\Delta^{\mathcal{L}}$ is a non-empty set and $\cdot^{\mathcal{L}}$ is an interpretation function that maps a concept C to a subset $C^{\mathcal{L}}$ of $\Delta^{\mathcal{L}}$ or a role r to a subset $R^{\mathcal{L}}$ of $\Delta^{\mathcal{L}} \times \Delta^{\mathcal{L}}$. A *Concept* C corresponds to a class in knowledge domain and is represented by a set of individuals. *Roles* are binary relations between objects. For instance, the concept BlackCat can be denoted as BlackCat $\equiv$ Cat $\sqcap \exists$hasColor.Black. **A-1** describes the main constructor and syntax in DL.

(Received May 3 ,2015)
(Revised Oct.16 ,2015)

Lu CAO

She received the B.S. degree in Computer Science at the Northern China University of Technology in 2005. From 2005 to 2007, she worked in NEC Corporation as a software engineer in China. In 2007, she joined the graduate school of Information and Computer Science at Saitama University, Japan, where she received the M.S. degree and the Ph.D. both in Computer Science in 2010 and 2013, respectively. After working at the National Institute of Advanced Industrial Science and Technology (AIST), She joined the Graduate School of Information and Computer Science as a postdoctoral researcher in 2015. Her research interests include object recognition, spatial recognition and reasoning, and ontological engineering.

Antony LAM

Antony Lam received the B.S. degree in Computer Science at the California State Polytechnic University, Pomona in 2004 and the Ph.D. in Computer Science at the University of California, Riverside in 2010. After working at the National Institute of Informatics, Japan, he joined the Graduate School of Science and Engineering at Saitama University as an assistant professor in 2014. His research interests are mainly in computer vision with emphasis on the areas of physics-based vision and pattern recognition.

Yoshinori KOBAYASHI

He completed the M.E. degree from the Department of Information Management science at Graduate School of Information Systems, the University of Electro-Communications in 2000, and joined the Design Systems Engineering Center of Mitsubishi Electric Corporation. He was in the doctoral program from 2004-2007 in Information and Communication Engineering at the Graduate School of Information Science and Technology, the University of Tokyo. He then joined the Department of Information and Computer Sciences, Saitama University, as an Assistant Professor. He is interested in computer vision for human sensing and its application for human computer interaction.

Yoshinori KUNO *(Member)*

He received the B.S., M.S. and Ph.D. degrees in 1977, 1979 and 1982, respectively, all in Electrical and Electronics Engineering from the University of Tokyo. After working with Toshiba Corporation and Osaka University, since 2000, he has been a professor in the Department of Information and Computer Sciences, Saitama University. His research interests include computer vision and human-robot interaction.

Daisuke KACHI

He received the B.A. degree in Liberal Arts and the M.A. in Philosophy both from the University of Tokyo. After accomplishing credits for the doctoral program in Philosophy at the University of Tokyo, he joined the College of Liberal of Arts at Saitama University in Philosophy as a lecturer, in 1993. In 1995 and 2005, he was promoted to an assistant professor and a professor in the Faculty of Liberal Arts, respectively. From 2009 to 2010, he also joined the Graduate School of Science and Engineering as a collaborating faculty member. Currently, he is a professor at the Graduate School of Humanities and Social Sciences. His research interests include analytic metaphysics, philosophical logic and formal ontology.