

Qualitative spatial logic descriptors from 3D indoor scenes to generate explanations in natural language

Zoe Falomir¹ · Thomas Kluth²

Received: 12 November 2016 / Accepted: 14 June 2017

Abstract The challenge of describing 3D real scenes is tackled in this paper using qualitative spatial descriptors. A key point to study is which qualitative descriptors to use and how these qualitative descriptors must be organized to produce a suitable cognitive explanation. In order to find answers, a survey test was carried out with human participants which openly described a scene containing some pieces of furniture. The data obtained in this survey is analyzed, and taking this into account, the *QSn3D* computational approach was developed which uses a XBox 360 Kinect to obtain 3D-data from a real indoor scene. Object features are computed on these 3D-data to identify objects in indoor scenes. The object orientation is computed and qualitative spatial relations between the objects are extracted. These qualitative spatial relations are the input to a grammar which applies saliency rules obtained from the survey study and generates cognitive natural-language descriptions of scenes. Moreover, these qualitative descriptors can be expressed as first-order-logical facts in Prolog for further reasoning. Finally, a validation

study is carried out to test whether the descriptions provided by *QSn3D* approach are human readable. The obtained results show that their acceptability is higher than 82%.

Keywords Qualitative spatial descriptors · RGB-depth data · Kinect · Spatial language · Grammar · Machine learning · SVMs · Spatial cognition · Survey · Validation · Human readable · Human-computer interaction

1 Introduction

Imagine the following scenario: It is 2056 and you have a robot at home that every morning after you leave arranges the furniture in your living room that you untidied the previous night. Before leaving, you provide the following instruction: *Please, tidy the living room*. To clarify, your robot asks back: *Should the new stool go in front of the armchair or down the table?* And you answer: *To the right of the armchair, it is fine*. Or imagine another scenario in which you move to a new house and a decorator tutor application in your tablet helps you to arrange new furniture in your rooms in a functional and fashionable way. Those situations would both involve spatial intelligence. In the first scenario, your robot at home would need to understand the scene (i.e. identifying the objects and their spatial locations in the living room), detect changes and then interact with the environment to get the previously defined arrangement. In the second scenario, the decorator tutor would engage in human-computer interaction. It would need to produce natural language descriptions to provide the user with instructions, then it would need scene understanding for interpreting if the user has nicely done the described task (or not), and finally it would also need to interpret the changes to provide the user with feedback about how to improve. These scenarios are still “imagined”, but there is much research effort in the literature focusing on solving

Handling editor: Antonio Bandera (University of Malaga);
Reviewers: Andrea Torsello (Ca Foscari University Venice), Ricardo Vzquez Martn (University of Malaga), Rebeca Marfil (University of Malaga).

This article is part of the Special Section on Cognitive Robotics guest-edited by Antonio Bandera, Jorge Dias, and Luis Manso.

This is an author generated manuscript with the same content as the final publication which is available at Springer via <https://dx.doi.org/10.1007/s10339-017-0824-7>

✉ Zoe Falomir
zfalomir@uni-bremen.de

¹ Bremen Spatial Cognition Centre (BSCC), University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany

² Language and Cognition Group, Cognitive Interaction Technology Excellence Cluster (CITEC), Bielefeld University, Inspiration 1, 33615 Bielefeld, Germany

those challenges. This paper deals with the topic of scene understanding in order to detect pieces of furniture in a 3D scene and describe its location using natural language descriptions based on qualitative spatial descriptors. For human-computer interaction qualitative representations are useful because they can deal with abstractions and uncertainty. Moreover, qualitative descriptors are based on reference systems which align with human perception and thus help establishing a more cognitive communication.

Recognizing objects in real scenes (i.e. any cup in different homes or offices) is a challenging task in the field of 3D computer vision and robotics (Olszewska, 2015a). However, research in the field of 3D object recognition has been fostered by the availability of low-cost, consumer depth cameras based on structured infrared light (also called RGB-Depth, RGB-D, cameras) incorporated in the Microsoft Kinect sensor, the Asus Xtion sensor and Google Tango tablet¹. A large number of research efforts has been carried out to improve RGB-D perception (i.e. overcome 3D data distortions due to noisy sensors, viewpoint changes and point density variations) for enabling robots to operate in unstructured real-world environments. One of the key challenges in this direction is to understand humans and their environments, since robots are envisioned to operate and perform various tasks at homes and working places in the near future. In the literature, interesting progress in this direction was made by dealing with 3D modelling of objects and environments: e.g., indoor modelling (Henry et al, 2010), dynamic scene modelling (Herbst et al, 2011a), autonomously learning new object models by meta-data sharing between robots (Krainin et al, 2011) or interactive modelling and 3D visualization (Du et al, 2011). Moreover, 3D robust recognition of everyday objects have been achieved by applying different machine learning techniques: e.g., depth kernel descriptors (Bo et al, 2011b), hierarchical kernel descriptors (Bo et al, 2011a), sparse distance learning (Lai et al, 2011a), scalable and hierarchical recognition (Lai et al, 2011b) or multi-scene analysis (Herbst et al, 2011b). Among other things, RGB-D cameras have been used to recognize human poses (Shotton et al, 2011), to build and maintain semantic maps of scenes using probabilistic graphical models for recognizing objects and rooms (Ruiz-Sarmiento, 2016; Ruiz-Sarmiento et al, 2015), etc.

In the first presented scenario, let us highlight that the spatial terms such as *in front of* and *to the right of* are qualitative and define a vague relation in space instead of a precise numerical location (e.g., Landau, 2016). In the literature, researchers showed the usefulness of converting qualitative models to natural language for: generating navigation instructions from sketched route maps (Skubic et al, 2004); com-

municating about vehicle traffic situations and traffic maneuvers (Steinhauer, 2005); describing the qualitative shape of geometric objects in sketches (Museros et al, 2014); and describing digital images using qualitative descriptors (Falomir, 2013). Other approaches also showed the adequacy of cognitive narratives in: (1) describing virtual scenes by selecting reference objects for generating location expressions (Barclay and Galton, 2013); (2) human-robot interaction (Moratz and Tenbrink, 2008, 2006) which takes into account how people choose perspective and *relatum* to describe object arrangements in space (Tenbrink et al, 2007, 2011); and (3) selecting *salient* features to describe objects depending on the context (Mast et al, 2016).

All these previous research works have inspired the current paper, where an approach for describing 3D scenes using qualitative descriptors (*QSn3D*) is presented. The *QSn3D* approach involves detecting real objects in indoor scenes and then describing their spatial arrangements logically and in natural language. For that, previous results of cognitive studies (Tenbrink et al, 2007, 2011; Zhang et al, 2014) as well as our own survey study are taken into account.

As far as we are concerned, there are very few works that obtain spatial logics and produce a narrative for a 3D scene. Olszewska (2015b) developed a system able to automatically process a 3D scene (grounded by its 2D views) and to provide high-level specifications of the scene (3D directional relations and 3D far/close spatial relations) using description logics that allowed reasoning about the 3D scene and its 2D views. Olszewska (2016) presented a computer-vision system to determine the object of reference in a conversation between multiple agents where the saliency of this reference object is computed from the visual interest points in each agent view. Huo and Skubic (2016) proposed a spatial language generation system to find short accurate human-like descriptions for robots to communicate the location of an object to a human user in an indoor environment, where the rooms and the pieces of furniture are described. They also dealt with oriented objects and obtained good results in a simulated environment using Gazebo3D platform. Although these works are related to the approach presented here, none of them manage raw data extracted from real 3D scenes using RGB-D cameras. Moreover, they do not apply the qualitative spatial descriptors (*QSn3D*) and the saliency rules proposed in this paper.

The rest of the paper is organised as follows. In Section 2, the challenge of providing spatial instructions in natural language is explained. Section 3 presents the survey test carried out in order to investigate the preferred narratives for people when describing an indoor scene that includes oriented and non-oriented objects. Section 4 presents the techniques used for detecting objects in 3D point clouds obtained by a Kinect RGB-D camera. It also explains the proposed qualitative spatial descriptors (*QSn3D*) and the provided logics.

¹ Trade and company names are included for benefit of the reader and imply no endorsement or preferential treatment of the product by the authors.

Section 5 explains the grammar and saliency rules used for describing a 3D scene in natural language using the proposed *QSn3D*. Section 6 shows the experiments carried out and the obtained narratives and logics. Section 7 presents a validation study done with human participants to test if the provided descriptions are human readable. Section 8 provides a general discussion. Finally, Section 9 presents conclusions and future work.

2 The challenge

Understanding and generating spatial instructions can be a challenge for robots because they perceive the world in terms of numerical values – through their sensors – instead of using concepts or spatial representations that they may use in human-robot communication. Moreover, different spatial terms might be used in the same communicative situation to describe a scene from different perspectives. Thus, robots need strategies for effective interpretation.

When humans describe scenes in natural language, they often use spatial expressions. Such expressions contain at least one spatial term, such as *on*, *near*, *beside*, *left*, *right*. An important group of spatial terms are the so-called *projective prepositions*. Projective prepositions describe spatial relationships among objects, such as *above*, *below* or *to the left of*, including a direction in which one object (the *located object*) is located with respect to another object (the *reference object*). Therefore, projective prepositions need a reference frame in which they are interpreted. According to Levinson (2003) three types of reference frames can be distinguished²:

- Intrinsic reference frames are established by inherent properties of the reference object (i.e. the front side of an object). For example: *the car is in front of the house*, where the *car* is the located object and the *house* is the reference object.
- Deictic reference systems are given by an observer's perspective on the reference object. Taking the previous example, the same situation can be described from the point of view of somebody standing next to the car: *the car is on my right*, where the *car* is the located object and the speaker is the reference object, who has their own front/ back and left/right sides.
- Extrinsic reference frames are imposed on the reference object by external factors, e.g., the Earth's gravitation, a geo-reference system, etc. For example: *the car is oriented to the West*, where the reference object is the Cardinal system.

Fig. 1 shows a situation in which different utterances can be used, such as: “*The rubbish bin is on my left*” (Utt-1) or “*The rubbish bin is in front of the office chair*.” (Utt-2). The Utt-



Fig. 1 An example of an indoor scene where different reference frames can be used for describing its spatial arrangement.

1 is produced considering a deictic reference frame centered on the viewpoint of an observer describing the scene. The Utt-2 is produced considering an intrinsic reference frame centered on the (oriented) reference object. Another possible utterance might be “*The rubbish bin is on the East*”, which would be used by speakers oriented to the geographical south of the Earth who are using an extrinsic reference frame. This kind of reference frame is not common in English, so it is not further considered in this paper.

In a communicative process, the capabilities assumed for the addressee depend if they are a human or a robot. Tenbrink et al (2002) showed that speakers have a conceptualisation of a robot as “*a communication partner who needs comparably simple instructions*” (p.22). This capacity of adaptation in humans in interactive situations facilitates our task. However, the more the robot can use human-similar utterances and adapt to the human user, the more natural the interaction gets.

3 A survey study

In order to study how people describe a scene, we carried out a survey. The scene in Fig. 2 was shown to the audience in the *JARCA Workshop on Qualitative Systems and Applications in Diagnosis, Robotics and Ambient Intelligence*³, who was asked to answer the following question on a piece of paper: *What do you see in the image?*

Thirty-eight participants provided their descriptions openly in a white paper: 34% women and 66% men between 25 and 60 years old, among them undergraduate students, master students, PhD students, doctors and professors mainly in the field of engineering and technology. The language of the answers were Spanish and English.

Speakers usually start descriptions by mentioning the object that captured their attention most, that is, by the most

² For a cross disciplinary taxonomy of reference frames see the work by Pederson (2003).

³ JARCA workshop: <http://madeirasic.us.es/jarca16/?lang=en>



Fig. 2 Image provided to the participants in the survey test in which they should answer the following question: *What do you see in the image?*

salient object for them. In this survey, a corpus of 38 spatial language descriptions was gathered. Usually participants generated different utterances by starting their descriptions from different objects. However, after analysing all the answers, four groups were differentiated according to the level of elaboration in the provided descriptions. Fig. 3 provides an iconic representation of these groups, which are described as follows:

- Group 1: Objects were enumerated or listed without spatial connectors to link them. Although no spatial prepositions were provided (see Fig. 2) some spatial order was found since participants mentioned objects following an order: (1) from the background-left to the foreground-right and vice versa; and (2) from the background-right to foreground-left and vice versa. An example of a description belonging to this group is the following: *There is a blue office chair, a wooden grey chair, a wooden stool and a green rubbish bin with papers inside.*
- Group 2: Objects were grouped/listed together depending on their properties or utility. As Fig. 2 shows, usually the office chair and the armchair were grouped together. Another group that we found was that made by 4-legged objects. An example of an utterance belonging to this group is the following: *There are 3 seats. A 4-legged stool, a 4-legged grey chair, a blue chair with wheels and a green rubbish bin with papers.*
- Group 3: Objects were grouped and described according to their spatial locations. As Fig. 2 shows, objects were grouped according to their category (i.e. chairs) or according to the distance to the observer. In the following example even the description of the ordering is given:

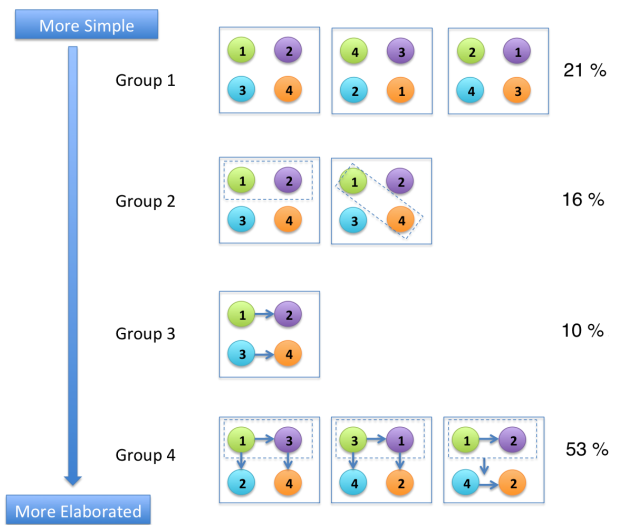


Fig. 3 Iconic representation of the utterances obtained from the survey.

From the back to the front, and from the left to the right there are: two chairs, a rubbish bin and a stool.

- Group 4: Objects were grouped and described according to their orientations and spatial locations. As Fig. 2 shows, the furniture which has a *front* and a *back*, such as chairs, are used as the reference object to describe the rest of the objects in the scene. An example is the following: *There are an armchair and an office chair in the background. There is a green rubbish bin in front of the armchair and a wooden stool in front of the office chair.* Another utterance which combines grouping by category and intrinsic orientations is the following: *In the background, there are two chairs and in front of the chairs there is a rubbish bin and a stool.*

Other curious results obtained from the survey were the following:

- some participants named the stool as a *table*. They probably were confused because they grouped it with the chair behind it.
- some participants described the furniture in detail by mentioning its manufacturer and its cost (i.e., *a IKEA-style-69-euro armchair*) or by mentioning their detailed composition (i.e. *a wooden beech stool, a blue office chair with black plastic parts*).
- some participants made some subjective observations regarding their experiences with those kind of furniture, such as: *the chair on the left is more comfortable than the one on the right.*
- some participants grouped the objects by their commonalities and then they explained their differences. For example, some participants named the office chair and the armchairs as “two chairs” and then they differentiated them by describing their colours or their parts, such as

the height of its back or by mentioning that one chair has armrests while the other not.

In summary, the most popular results in the survey were those reported in Group 4 (53%). Thus, this is the strategy selected by *QSn3D* to describe the scenes in natural language.

4 A qualitative 3D scene descriptor (*QSn3D*)

In order to obtain a qualitative descriptor of a real 3D scene, first the objects in the scene must be detected and categorized. For that, 3D point clouds are extracted using a RGB-D camera and machine learning techniques are used to identify patterns of features which characterize objects in the scene (Section 4.1). After detecting each object in the scene, their location (Section 4.2) and their orientation (Section 4.3) are computed. Then, logics are provided describing each scene (Section 4.4).

4.1 Recognizing objects in 3D scenes

To identify objects in a scene, their corresponding 3D point clouds are obtained and segmented into disjoint parts ($S = \{P_0, P_1, \dots, P_N\}$), each one corresponding to one object in the scene. For that, the following steps are followed:

1. the floor in the scene is extracted by applying a RANSAC-based segmentation (RANdom Sample And Consensus) (Fischler and Bolles, 1981).
2. to distinguish different objects, an Euclidean Cluster Extraction process is carried out. For each extracted cluster, two geometrical 3D-features are calculated:
 - (a) the Viewpoint Feature Histogram (VFH, Rusu et al, 2010) which is scale invariant but viewpoint variant. The main idea of VFH is to calculate three different angles between two points, using the normal vectors and the viewpoint direction.
 - (b) the Global Radius-based Surface Descriptor (GRSD, Marton et al, 2010). The goal of GRSD is to approximate 3D-objects by searching for best-fitting feature circles at each point.

As Fig. 4 shows, for each object, point clouds from different views are obtained, recorded and labeled with the name of the object. These point clouds contain different orientations and scales of the objects. The aforementioned 3D-features are calculated on each point cloud. With these labeled feature vectors, a Support Vector Machine (SVM)⁴ is trained using LIBSVM (Chang and Lin, 2011). This training procedure is done only once for each environment.

⁴ In general, one could apply different classification algorithms as well. In particular, zero-shot learning (e.g., Ji et al, 2017; Socher et al, 2013) might prove as a useful improvement to the current implementation, as these methods do not require a training phase. This allows to more easily add new objects to the system.

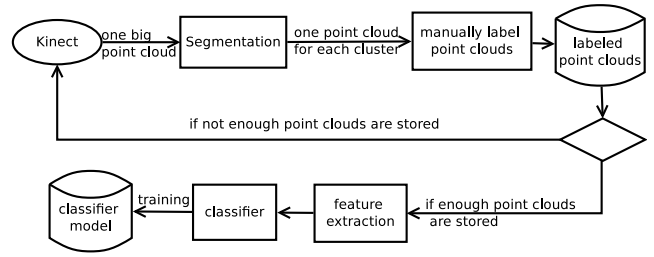


Fig. 4 The *training-mode* for the recognition of objects in *QSn3D*.

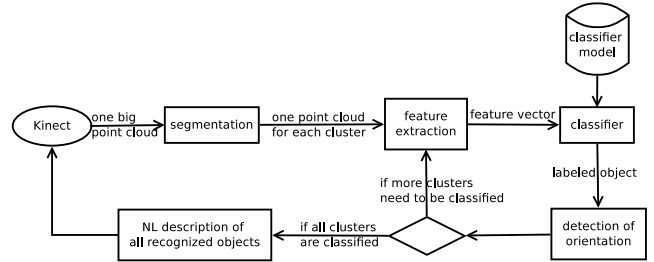


Fig. 5 The *live-mode* in the *QSn3D* approach.

As Fig. 5 shows, after the SVM-model is generated, the *QSn3D* approach can run in *live-mode*, that is, it can recognize objects and describe the complete scene in real time. The point cloud of the scene is segmented as explained above and a cluster for each object in the scene is obtained and matched. On each cluster the VFH and the GRSD features are computed and one 3D-feature vector for each object in the scene is obtained. Moreover, for each cluster, its centre point is also computed. Each feature vector is classified using the SVM-model previously trained with the corresponding object names. If the SVM is not able to classify a cluster, it is not further processed. As a result of this recognition process, each object in the scene is described as a 3D-centre point with a name.

4.2 Qualitative location in *QSn3D*

The location of an object can be obtained using a Location Reference System or LoRS = $\{\circ, LO_{LAB}, LO_{INT}\}$ where, degrees (\circ) indicate the unit of measurement of the angular location of the object; LO_{LAB} refers to the set of labels for the locations; and LO_{INT} refers to the values in degrees (\circ) related to each label. The LO_{LAB} and LO_{INT} used in *QSn3D* are the following:

$$LO_{LAB} = \{left, centre, right\}$$

$$LO_{INT} = \{(0, 80], (80, 100], (100, 180]\}$$

In general,

$$LO_{LAB_g} = \{L_1, L_2, \dots, L_{K\ell}\}$$

$$LO_{INT_g} = \{[0, \ell_1], (\ell_1, \ell_2], \dots, (\ell_{K\ell-1}, 180]\}$$

where $K\ell$ is the number of concepts used for defining locations in intervals of angles. The higher $K\ell$, the finer the granularity in the *LoRS*, while the lower $K\ell$, the coarser the granularity of this reference system.

The distance between an object and the observer also establishes a spatial relation in the scene. The closeness to an object can be described using a Distance Reference System or $DRS = \{m, D_{LAB}, D_{INT}\}$ where, m indicates the unit of measurement of the distance (meters); D_{LAB} refers to the set of labels for the distances; and D_{INT} refers to the interval values related to each label. The D_{LAB} and D_{INT} selected for *QSn3D* are the following:

$$D_{LAB} = \{foreground, background\}$$

$$D_{INT} = \{(0, d_n], (d_n, \infty)\}$$

where d_n is the distance threshold used in a scene. It can be parameterised depending on the environment.

In general,

$$D_{LAB_g} = \{D_1, D_2, \dots, D_{Kd}\}$$

$$D_{INT_g} = \{[0, d_1], (d_1, d_2], \dots, (d_{Kd-1}, 180]\}$$

where Kd is the number of concepts used for defining angles.

Qualitative distances were studied by [Hernández et al \(1995\)](#) defining reference systems with different granularity. Qualitative distances at a finer granularity level ($D_{LAB} = \{very\ close, close, quite\ near, near, medium, quite\ far, far, very\ far, too\ far, extremely\ far\}$) were previously used in robotics to integrate patterns from different kind of sensors (i.e. sonar and laser) to detect special obstacles (i.e. glasses or mirrors) and to categorize corner reference systems for orientation ([Falomir et al, 2011a](#)). The *QSn3D* uses a coarse reference system which distinguishes between *background* and *foreground*, since this distinction was that used by the participants in the survey study carried out.

Computationally, the space division depicted in Fig. 6 is used in *QSn3D* where the z -axis represents the depth-information and the x -axis is the horizontal information delivered by the RGB-D camera. An object is computationally located in the *background/foreground*, if its z -value is higher/lower than a scene-specific threshold, d_n , which is represented by the dashed line in Fig. 6. An object is computationally placed on the *right*, if its location is to the right of the right dotted line in Fig. 6. That is, if the object is located at an angular position included in the interval $(100, 180]$. An object is computationally placed on the *left*, if its location is to the left of the left dotted line in Fig. 6. That is, if the object is located at an angular position included in the interval $(0, 80]$. Otherwise, it is determined that the object is placed in the *centre* location. The dotted lines are defined by an angular location at the origin, that is, where the RGB-D camera is placed.

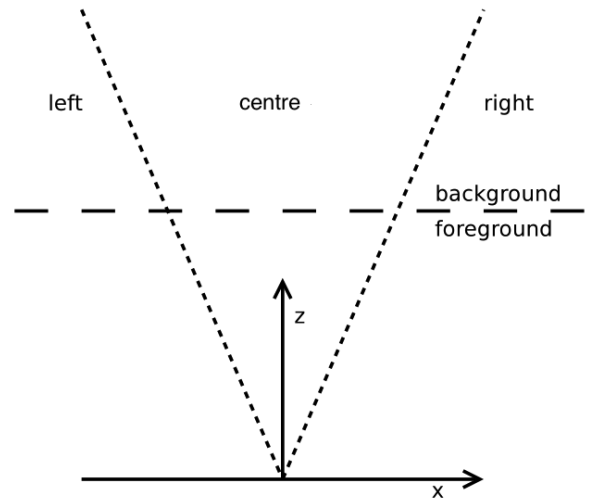


Fig. 6 Deictic reference systems for dividing the space observed from a RGB-D camera (*LoRS* and *DRS*).

4.3 Qualitative orientation in *QSn3D*

The *QSn3D* uses a deictic reference system (see Section 2) to define the location of the objects with respect to the point of view of the observer (i.e. a robot with a RGB-D camera), whereas an intrinsic reference system is used in oriented objects to describe the location of their neighboring or close objects with respect to them.

By combining the Location and Distance Reference Systems (*LoRS* and *DRS*), a reference system with finer granularity can be defined (see Fig. 7) and named as: $LoDRS = \{LoD_{LAB}, LoD_{INT}\}$ where LoD_{LAB} refers to the set of labels for regions in space; and LoD_{INT} refers to the interval values related to each label. The LoD_{LAB} and LoD_{INT} obtained for *QSn3D* are the following:

$$LoD_{LAB} = \{front - right, back - right, front - centre, back - centre, front - left, back - left\}$$

$$INT_{LoD} = \{(0, 80] \times (0, d_n], (0, 80] \times (d_n, \infty], (80, 100] \times (0, d_n], (80, 100] \times (d_n, \infty], (100, 180] \times (0, d_n], (100, 180] \times (d_n, \infty)\}$$

In general,

$$LoD_{LAB_g} = \{Lo_{LAB} \times D_{LAB}\}$$

$$LoD_{INT_g} = \{[0, \ell_1] \times (0, d_n], [0, \ell_1] \times (d_n, \infty], \dots, (\ell_{K\ell-1}, 180] \times (0, d_n], (\ell_{K\ell-1}, 180] \times (d_n, \infty)\}$$

where LoD_{LAB_g} is the result of combining the concepts previously defined in Lo_{LAB} and D_{LAB} ; the intervals in LoD_{INT_g} are the result of the Cartesian product of the intervals in Lo_{INT} and D_{INT} ; d_n is the threshold used in D_{INT} and $K\ell$ is the number of concepts used for defining angles in Lo_{INT} .

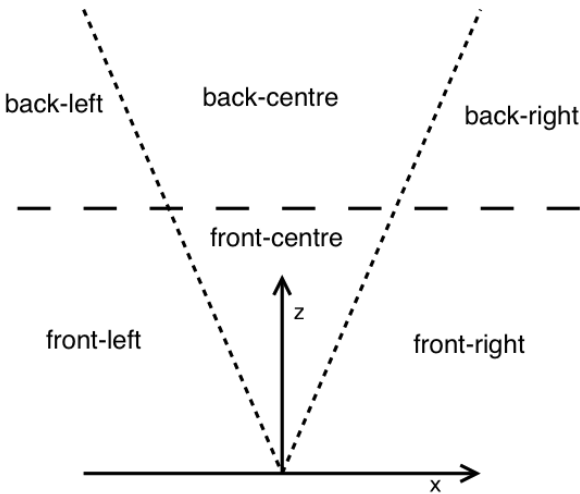


Fig. 7 Reference system for describing distance and orientation ($LoDRS$).

Computationally, in order to detect the orientation of an oriented piece of furniture (i.e. a chair) every point cloud cluster is searched for a horizontal plane (i.e. seat part) and a vertical plane (i.e. resting back part). If both planes exist, the cluster is recognized as *oriented* and its orientation is calculated as the normal of the vertical plane showing the front of the chair. If only a vertical plane is detected, the object is treated as an object *oriented towards the background*. If only a horizontal plane or no plane at all is detected, the object is considered as *not oriented*.

Both reference systems used in $QSn3D$, deictic and intrinsic to oriented objects, are computed by $LoDRS$. The deictic locations are obtained by locating the $LoDRS$ where the observer is placed (i.e. a robot with a RGB-D camera or a human user with a Tango tablet). And the intrinsic locations are obtained by locating the $LoDRS$ at the 3D centre of the object and matching its front to the object front. To determine the spatial relation between two objects, the coordinates of the 3D centres of both objects are compared and an output is generated with respect to the $LoDRS$ of the oriented object.

4.4 Logics for $QSn3D$

In order to describe a scene using spatial logics, a first-order knowledge base (KB) is built as a set of formulas in first order logic (Genesereth and Nilssson, 1987) using four types of symbols (variables, constants, predicates, and functions). Variable symbols range over the objects in the domain (i.e. *Location* range over *left*, *right*, *centre*). Constant symbols can represent objects in the domain of interest (i.e., *stool*, etc.). Predicates represent relations among objects in the domain (i.e., *close_object*) or attributes of objects (i.e., *is-oriented*). Function symbols can be formulated for inferring

Table 1 Logic facts extracted by $QSn3D$ for the objects in the scenes.

$QSn3D \subseteq \forall Object \in Scene \exists QD$
$QD \subseteq \exists is_categorized(ObjId, ObjName).$
$QD \subseteq \exists location_wrt_observer(ObjId, LoLAB).$
$QD \subseteq \exists distance_wrt_observer(ObjId, D.LAB).$
$QD \subseteq \exists close_object(ObjId, Obj2Id).$
$QD \subseteq \exists is_oriented(Answer, ObjId, LoD.LAB).$

new situations using the predicates defined. First-order KB are usually built using Horn clauses, which contain at most one positive literal. The Prolog programming language is based on Horn clause logic (Lloyd, 1987).

$QSn3D$ generates first order logic facts related to all the objects in the scene. Prolog syntax is used for expressing these logics as described in Table 1. Variable *Scene* represents a 3D point cloud scene, and variable *Object* represents any object detected in the point cloud.

The predicate *is_categorized* relates the object identifier with its name. The predicate *location_wrt_observer* describes the location of an *Object* with respect to an observer (i.e. robot with RGB-D camera or a person with a Tango tablet) using the reference system showed in Fig. 6 and the concepts defined by the variable $LoLAB_1$ (i.e. *left*, *centre*, *right*). The predicate *distance_wrt_observer* describes the closeness of an *Object* with respect to an observer which is defined by the variable D_{LAB_1} (i.e. *background*, *foreground*). The predicate *close_object* relates an object with the objects which are close to it. The predicate *is_oriented* says if an object is oriented or not, and provides its orientation using the reference system shown in Fig. 7 and the spatial concepts defined by the variable $LoDLAB$.

5 Generating Narratives for $QSn3D$

Considering the studies in the literature and the results of the survey test, the $QSn3D$ is designed to generate two types of natural language utterances which describe location by taking into account:

1. a deictic reference system located at the point of view of the observer (i.e., robot with RGB-D camera or user with tablet) from which the objects in the scene are described, and
2. an intrinsic reference system between objects in the scene that have clear orientations, as for example, chairs, sofas, armchairs, etc.

Section 5.1 presents how the $QSn3D$ computes the saliency of the objects in the scene. Section 5.2 describes the grammar used to produce the narratives for the scenes.

5.1 Saliency

As the survey showed (Section 3), there are multiple ways to describe the same scene regarding the order in which objects are mentioned. Usually, the first described object is considered the most salient for the speaker. Several features might be used to compute the saliency of an object. After the analysis of the survey results, *QSn3D* focused on two possible saliency measures: (1) the object size and (2) the distance to the observer. These are also the measures used in other works in the literature such as that by Lison (2010).

Therefore, first the most salient object is chosen, that is, either the biggest object in the scene or the closest object in the scene. Then, the rest of the objects are described, first considering those objects located at the same proximity to the observer as the most salient object (i.e. background), and then the other objects. If objects located at the same proximity belong to the same category, they can be referred together as a group (i.e. *two chairs, three tables, a group of cushions*, etc.). In Algorithm 1 this strategy is described in pseudocode.

```

firstObject = findMostSalientObject();
ObjectList = [firstObject, closeObject(1), closeObject(2),...,
closeObject(m)];
for Object i in ObjectList do
  if not(i.Described()) then
    if i.isInGroup() then
      describeGroup();
      describeObject_wrt_Observer(i);
    else
      describeObject_wrt_Observer(i);
    end
  end
  if i.isOriented() then
    for Object j in ObjectList do
      describeObject_wrt_CloseObject(j, i);
      setDescribed(j);
    end
  end
  setDescribed(i);
end

```

Algorithm 1: *QSn3D* Strategy for describing objects in scenes.

If a non-oriented object is selected as the most salient object, the deictic reference frame is selected (Fig. 6) to describe the location of the rest of the objects in the scene. In contrast, if an oriented object is selected as the most salient object, the *QSn3D* uses the corresponding intrinsic reference frame (Fig. 7 located at the object 3D centre and matching the object front) to describe the location of the rest of the close objects. If there are two oriented objects, two relations are created. However, the spatial relation corresponding to the biggest oriented object is used.

5.2 A grammar for *QSn3D*

In order to obtain a description in natural language of a real scene, the qualitative descriptors defined by the *QSn3D* are used and organized in a context-free grammar (G) built on the following parameters:

$$G = (V, \Sigma, P, \langle QSn3D \rangle) \text{ where,}$$

- V is an alphabet of symbols that are non-terminals;
- Σ is an alphabet of terminal symbols (qualitative labels or words), disjoint with V ;
- $P \subseteq V \times (V \cup \Sigma)^*$ is the set of production rules;
- $\langle QSn3D \rangle \in V$ is the initial symbol of the grammar;
- λ is the empty string.

The grammar $G(QSn3D)$ is defined as follows:

$$\begin{aligned}
 \langle QSn3D \rangle &\rightarrow \langle SaliencyBySize \rangle \langle SaliencyByProximity \rangle \\
 \langle SaliencyBySize \rangle &\rightarrow \text{In the } \langle D_{LAB_1} \rangle, \text{ there is } \langle ObjDesc \rangle. \\
 \langle ObjDesc \rangle &\rightarrow \langle OrientedObj \rangle \langle RestofObjOr \rangle \\
 \langle ObjDesc \rangle &\rightarrow \langle NonOrientedObj \rangle \langle RestofObjNonOr \rangle \\
 \langle OrientedObj \rangle &\rightarrow \langle ObjName \rangle \langle Lo_{LAB_1} \rangle \langle Orientation \rangle \\
 \langle NonOrientedObj \rangle &\rightarrow \langle ObjName \rangle \langle Location \rangle \\
 \langle ObjName \rangle &\rightarrow \langle Article \rangle \langle ObjName \rangle \\
 \langle Article \rangle &\rightarrow a \mid an \mid the \\
 \langle Name \rangle &\rightarrow \text{armchair} \mid \text{rubbish bin} \mid \text{stool} \mid \text{office chair} \mid \\
 &\text{white-chair} \mid \text{white-table} \mid \text{wooden chair} \mid \text{two chairs} \\
 \langle D_{LAB_1} \rangle &\rightarrow \text{foreground} \mid \text{background} \\
 \langle Lo_{LAB_1} \rangle &\rightarrow \text{in the centre} \mid \text{on the left} \mid \text{on the right} \\
 \langle Orientation \rangle &\rightarrow (\text{oriented to the } \langle LoD_{LAB_1} \rangle) \\
 \langle LoD_{LAB_1} \rangle &\rightarrow \text{front-right} \mid \text{back-right} \mid \text{front-centre} \mid \text{back-} \\
 &\text{centre} \mid \text{front-left} \mid \text{back-left} \\
 \langle RestofObjOr \rangle &\rightarrow \langle ObjName \rangle \text{ has } \langle CloseObj \rangle \langle IntrinsicLoc \rangle \\
 \langle RestofObjOr \rangle &\rightarrow \lambda \\
 \langle IntrinsicLoc \rangle &\rightarrow \text{on its right} \mid \text{on its left} \mid \text{on its front} \mid \text{at its} \\
 &\text{back} \\
 \langle RestofObjNonOr \rangle &\rightarrow \text{In the } \langle D_{LAB_1} \rangle, \text{ there is } \langle ObjDesc \rangle. \\
 \langle RestofObjNonOr \rangle &\rightarrow \lambda \\
 \langle CloseObj \rangle &\rightarrow \langle ObjName \rangle \langle MoreCloseObjs \rangle \\
 \langle MoreCloseObjs \rangle &\rightarrow \text{and } \langle ObjName \rangle \mid \langle MoreCloseObjs \rangle \\
 \langle MoreCloseObjs \rangle &\rightarrow \lambda \\
 \langle SaliencyByProximity \rangle &\rightarrow \text{In the } \langle D_{LAB_1} \rangle, \text{ there is } \langle ObjDesc \rangle.
 \end{aligned}$$

The requirements for activating the production rules of this grammar are:

- The scene is described starting by its most salient object, that is, its biggest object (firing the rule $\langle SaliencyBySize \rangle$) or the closest object to the observer (firing the rule named $\langle SaliencyByProximity \rangle$) and passing the corresponding object, which can be oriented or non-oriented.
- If the most salient object ...

... is oriented, then its orientation is described with respect to the point of view of the observer and the locations of the close objects (if any) are described using an intrinsic reference system situated on the oriented object (the rule $\langle \text{OrientedObj} \rangle$ is fired),

... is non-oriented, then the location of the rest of the close objects is described according to the point of view of the observer (deictic reference frame, rule $\langle \text{NonOrientedObj} \rangle$ is fired).

- The rest of the objects in the scene are described according to the proximity to the most salient object, and their descriptions also depend on whether they are oriented objects or not.

The language generated by the $G(QSn3D)$ grammar is defined as follows:

$$G : L(G) = \{x \in \Sigma^* \mid \langle QSn3D \rangle \xrightarrow{*} x\}$$

The $G(QSn3D)$ language describes scenes using two different options: (1) starting by the biggest object as the most salient object; and (2) starting by the closest object to the observer as the most salient object. Both narratives are obtained, so that the system can recognize any situation or produce any utterance. If the most salient object is an oriented object, an intrinsic reference system is used for describing its location, otherwise a deictic reference system is applied. Example utterances produced in this language are given in Tables 6, 7, 8, and 9.

6 Experimentation and results

The $QSn3D$ approach was tested in two indoor environments: a common home scenario and an office scenario. The sensor used for extracting the 3D point cloud from each scene was a Microsoft Xbox 360 Kinect.

The $QSn3D$ approach developed is written in C++ and build upon the Robot Operating System (ROS) framework⁵. In order to receive the 3D-data from the MS Kinect device we used the openNI-driver⁶, included in ROS. For processing the obtained point clouds, we used the Point Cloud Library (PCL) framework⁷ which is also integrated in ROS. The LIBSVM library (Chang and Lin, 2011) was used to train the SVM-model with the labeled 3D-feature vectors extracted from the clusters, each one corresponding to one piece of furniture. Fig. 8 shows an example of a scene in the home scenario where four pieces of furniture are detected: an armchair, an office chair, a rubbish bin and a stool. Fig. 8 (b) shows the point clouds obtained by the RGB-D sensor which are the input to the SVM-model. Fig. 8 (c) shows the results after applying our 3D object recognition process.

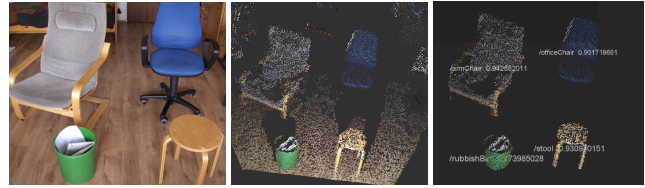


Fig. 8 Example: (a) Scenario in our testing; (b) Point clouds of the scene extracted by the RGB-D sensor; (c) Object recognition in the scene: output of the classification system.

The logics created by $QSn3D$ for the scene shown in Fig. 8 and also the narratives generated for the $G(QSn3D)$ grammar are presented in Table 2. Note that this scene has two oriented objects (the office chair and the armchair) and two non-oriented objects (a stool and a rubbish bin. Further results obtained by $QSn3D$ (narratives and logics) are provided in the Appendix:

1. 2 different scenes in the home scenario including 2 objects: an office chair (oriented object) and a stool (non-oriented object, Table 6).
2. 2 different scenes in the home scenario with 3 objects: an armchair (oriented object) and a stool and a rubbish bin (two non-oriented objects, Table 7).
3. 3 different scenes in the office scenario including 2 objects: a non-oriented one-legged white table and a chair (oriented object, Table 8).
4. 3 different scenes in the office scenario including 3 objects: a non-oriented one-legged white table and two distinct chairs (oriented objects, Table 9).

7 Validation study

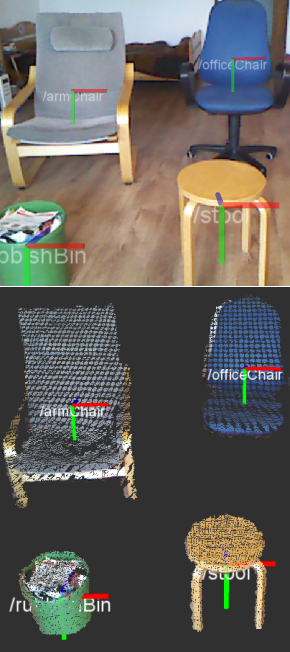
This section presents a validation study carried out to investigate whether the descriptions provided by $QSn3D$ approach were human understandable. This survey was carried out on-line using Google Forms platform. First, participants were explained the context of the study and were asked about their consent. Then, participants were asked to match the descriptions generated by $QSn3D$ approach with a scene. Each question presented both descriptions in a pseudo-random order: (A) description taking into account the biggest object as the most salient object, and (B) description taking into account the closest object as the most salient object. After choosing a scene, participants were asked which description they preferred: A, B, both or none of them. There was a total of 13 questions, 12 of them showing the descriptions and pictures presented in this paper (see Table 2 and Tables 6, 7, 8 and 9 in the appendix) and 1 control question whose description was not matching any scene. All scene pictures were randomized automatically by Google Forms. Fig. 9 shows an example question. At the end of the survey, participants were asked about their age, sex, nationality, level of english,

⁵ <http://www.ros.org>

⁶ <http://www.openni.org>

⁷ <http://www.pointclouds.org>

Table 2 QSn3D narratives and logics obtained in the home scenario using 4 pieces of furniture: 2 oriented and 2 non-oriented.

Photo	Language description	Logic description
	<p>Scene 1</p> <p>A. The biggest object (an armchair) as the most salient object: <i>In the background there are two chairs. There is an armchair (oriented to the front). The armchair has an office chair on the left, a rubbish bin in the front and a stool on the left.</i></p> <p>B. The closest object to the observer (a stool) as the most salient object: <i>In the foreground, there is a stool in the centre. There is a rubbish bin on the left. In the background there are two chairs. There is an office chair (oriented to the front). The office chair has an armchair on the right.</i></p>	<pre> is_categorized(object_0b, armChair). is_categorized(armChair, chair). location_wrt_observer(object_0b, left). distance_wrt_observer(object_0b, background). close_object(object_0b, object_1b). close_object(object_0b, object_2b). is_oriented(yes, object_0b, front). location_wrt_close_object(object_0b, object_1b, left). location_wrt_close_object(object_0b, object_2b, centre). is_categorized(object_1b, officeChair). is_categorized(officeChair, chair). location_wrt_observer(object_1b, centre). distance_wrt_observer(object_1b, background). close_object(object_1b, object_0b). close_object(object_1b, object_2b). close_object(object_1b, object_3b). is_oriented(yes, object_1b, front). location_wrt_close_object(object_1b, object_0b, right). location_wrt_close_object(object_1b, object_2b, right). location_wrt_close_object(object_1b, object_3b, centre). is_categorized(object_2b, rubbishBin). location_wrt_observer(object_2b, left). distance_wrt_observer(object_2b, foreground). close_object(object_2b, object_0b). close_object(object_2b, object_1b). close_object(object_2b, object_3b). is_oriented(no, object_2b, none). is_categorized(object_3b, stool). location_wrt_observer(object_3b, centre). distance_wrt_observer(object_3b, foreground). close_object(object_3b, object_1b). close_object(object_3b, object_2b). is_oriented(no, object_3b, none). </pre>
	<p>Scene 2</p> <p>A. The biggest object (an armchair) as the most salient object: <i>In the background there are two chairs. There is an armchair (oriented to the right). The armchair has an office chair on the left, a rubbish bin on the right and a stool on the right.</i></p> <p>B. The closest object to the observer (a rubbish bin) as the most salient object: <i>In the foreground, there is a rubbish bin in the centre. There is a stool on the left. In the background there are two chairs. There is an office chair (oriented to the left). The office chair has an armchair in the front.</i></p>	<pre> is_categorized(object_0c, armChair). is_categorized(armChair, chair). location_wrt_observer(object_0c, left). distance_wrt_observer(object_0c, background). close_object(object_0c, object_1c). close_object(object_0c, object_3c). is_oriented(yes, object_0c, right). location_wrt_close_object(object_0c, object_1c, left). location_wrt_close_object(object_0c, object_3c, right). is_categorized(object_1c, officeChair). is_categorized(officeChair, chair). location_wrt_observer(object_1c, centre). distance_wrt_observer(object_1c, background). close_object(object_1c, object_0c). close_object(object_1c, object_2c). close_object(object_1c, object_3c). is_oriented(yes, object_1c, left). location_wrt_close_object(object_1c, object_0c, centre). location_wrt_close_object(object_1c, object_2c, left). location_wrt_close_object(object_1c, object_3c, left). is_categorized(object_2c, rubbishBin). location_wrt_observer(object_2c, centre). distance_wrt_observer(object_2c, foreground). close_object(object_2c, object_1c). close_object(object_2c, object_3c). is_oriented(no, object_2c, none). is_categorized(object_3c, stool). location_wrt_observer(object_3c, left). distance_wrt_observer(object_3c, foreground). close_object(object_3c, object_0c). close_object(object_3c, object_1c). close_object(object_3c, object_2c). is_oriented(no, object_3c, none). </pre>

level and field of studies and their opinion regarding having a robot at home and teaching it to tidy furniture. The average time of filling in the survey was 15 minutes.

Participants. Using the Prolific platform⁸ 57 participants were recruited. They were paid 0.85 GBP per test. The answers from 7 participants were discarded because of their low quality (e.g., repeated or random answers or too little knowledge of English). The discarded participants received an e-mail explaining our reasons for not paying them and there was no disagreement. Moreover, 20 volunteers took part in the survey, who were not paid for their answers. Thus, a total of 70 answers were collected. The distribution of ages and level of English of our participants is shown in Fig. 10. Note that, although the majority of participants were between 21–40 years old, there were participants of all age ranges. Note also that 45% of the participants were English native speakers and that 34% reported to understand English language very well. This indicates that most of the participants have a high level of knowledge in English. The gender distribution is shown in Fig. 11: 40% of women vs. 60% of men. This Figure also shows that participants' nationalities were very diverse: mostly European citizens from 13 different countries (36% British) and there were also participants from North and South America from 5 different countries (6% from USA). Most of the participants studied at University (see Fig. 12) and half of them graduated in the field of engineering and technology. Moreover, there were also participants from the fields of natural and social science.

Regarding participants' acceptability of having a robot at home which could tidy their furniture, Fig. 13 shows that only 1/3 of the participants would like to have it. This indicates that most of our participants still do not believe in the *robot at home* idea. Moreover, 53% of our participants think *the robot must learn by itself*, and only 36% would agree to teach the robot *using language and gestures*. 17% of our participants would not like to teach the robot at all, whereas 13% propose their own ideas to do so, e.g., using a graphical user interface and some metadata, or tapping the correct locations on a touchscreen displaying each room.

Validation Results. This section presents an analysis of the answers given by the 70 participants (71% were paid participants and 29% were volunteers). Table 3 presents the confusion matrix obtained from their answers. Columns show the scenes described by *QSn3D*. In rows, participants' answers are shown. For example, when showing the *QSn3D* description corresponding to Scene 1 (see Fig. 2), 86% of the participants selected the picture corresponding to Scene 1, while 14% of the participants selected the picture of Scene 2. The diagonal of the matrix presents colored cells to show the percentage of matching between descriptions and scenes. Table 4

⁸ <https://www.prolific.ac/>

Explaining scenes in natural language

*Required

Your robot explains what it sees...

(It provides 2 descriptions of the same scene)

A:

In the background there are two chairs. There is an armchair (oriented to the front). The armchair has an office chair on the left, a rubbish bin in the front and a stool on the left.

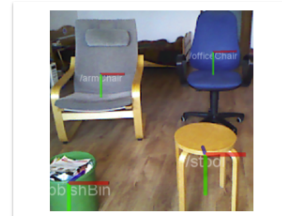
B:

In the foreground, there is a stool in the centre. There is a rubbish bin on the left. In the background there are two chairs. There is an office chair (oriented to the front). The office chair has an armchair on the right.

What do you think your robot sees? *



Scene 11



Scene 1

(...) the 12 scenes appearing in this paper are displayed

So, your robot said:

A: In the background there are two chairs. There is an armchair (oriented to the front). The armchair has an office chair on the left, a rubbish bin in the front and a stool on the left.

B: In the foreground, there is a stool in the centre. There is a rubbish bin on the left. In the background there are two chairs. There is an office chair (oriented to the front). The office chair has an armchair on the right.

Which description do you consider as more appropriate? *

A

Both

B

Other: _____

Fig. 9 Example of a question in the survey: participants were asked to match the descriptions provided by *QSn3D* with pictures of scenes. Note that this question presents *QSn3D* description for Scene 1.

shows the descriptions preferred by our participants. The results in Table 3 show that the acceptability of the descriptions provided by *QSn3D* approach is between 82%–95% — apart from considerable and systematic disagreements for two scenes (Scene 4, see Fig. 6, and Scene 12, see Fig. 9), as well as for a fake not-matching scene (control question). Next, we speculate about possible reasons for these disagreements.

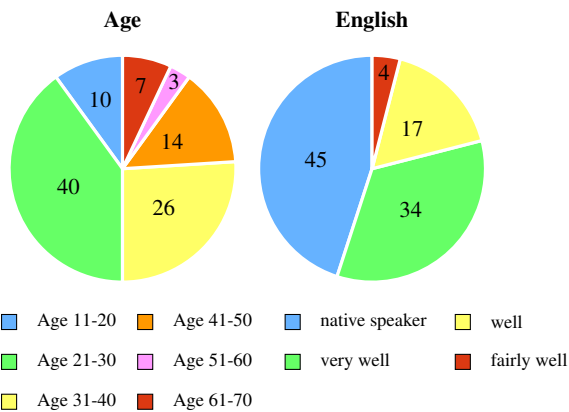


Fig. 10 Participation in the survey (in % and read anti-clockwise) regarding age and English language understanding.

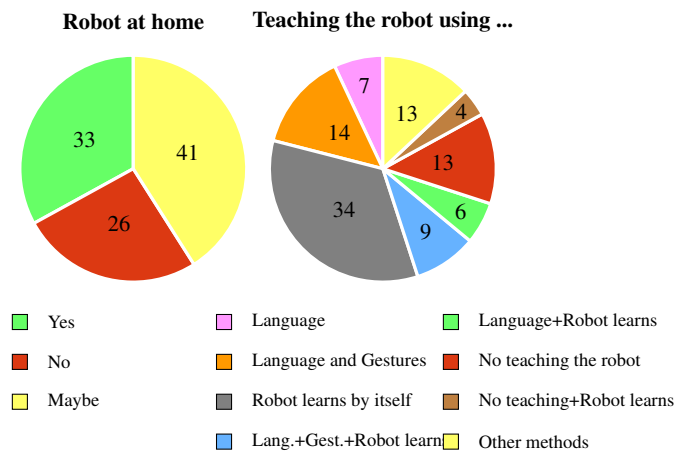


Fig. 13 Participation in the survey (in % and read anti-clockwise) regarding acceptance of the scenario provided regarding home robotics tidying furniture.

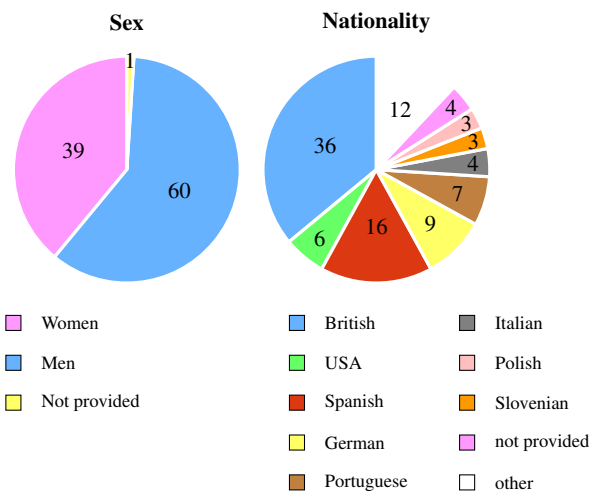


Fig. 11 Participation in the survey (in % and read anti-clockwise) regarding sex and nationalities. Other nationalities refer to participation of less than 1% from: Canadian, Uruguayan, Mexican, Venezuelan, Serbian, Austrian, Dutch, Latvian, Bulgarian, Hungarian, etc.

Scene	1	2	3	4	5	6	7	8	9	10	11	12	none
1	86	11	2	6		4							1
2	14	82			1	4							
3			90	51									
4			7	33		1							
5		3			95	3							
6		1	1	4	4	88							
7				1			87	7					1.5
8							4	90	1.5				1.5
9								93					31
10				1				4	87	3	9	3	
11							9		1.5	89	21		
12									1.5	10	7	69	4
none		3		4					1.5	1.5	1.5		59

Table 3 Confusion matrix, % of answers by 70 participants (in rows) for each scene (in columns).

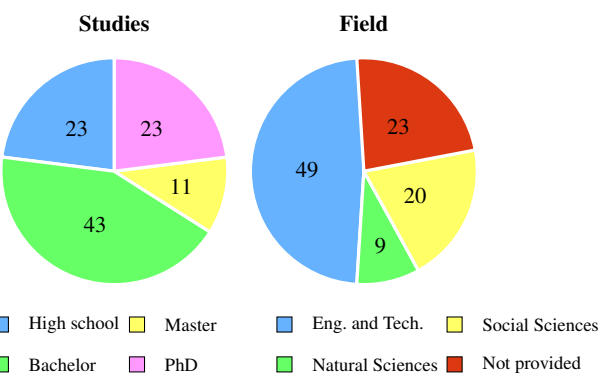


Fig. 12 Participation in the survey (in % and read anti-clockwise) regarding level and field of studies.

For Scene 4, the description B provided by *QSn3D* (see Table 4) was not considered adequate by the participants. Only 33% of them selected Scene 4. The rest selected: Scene 3 (51%), Scene 1 (6%), Scene 6 (4%), Scene 7 (1%), Scene 10 (1%) or none of them (4%). Description A was a bit more adequate since participants could discriminate between Scene 3 and 4 by noticing the location of the stool with respect to the office chair. This is reflected in the answers regarding the preferred description: 53% preferred description A vs. 21% preferred B (see Table 4). From our point of view, none of the pictures was perfectly matching the descriptions, since the office chair was not located on the *centre*, but mostly on the *right*, as seen from the picture taken by the Kinect. We think that this problem was generated because of the use of crisp boundaries in the reference systems. Most of the matching points in the point cloud were located in the centre and some on the right, so the description should have indicated *centre* and *right*, but using both descriptors at the same time is not possible in the defined reference systems. We will address

Table 4 Summary of % of preferred descriptions (in rows) for each scene (in columns) from 70 participants.

Scene	A	B	Both	None
1	48	30	19	3
2	30	49	14	7
3	56	20	24	0
4	53	21	21	5
5	47	27	21	4
6	44	30	22	4
7	16	31	53	0
8	40	23	36	1
9	26	37	36	1
10	50	19	30	1
11	29	61	10	0
12	23	59	16	2
None	17	14	17	52

this challenge in future work by considering vague locations (with a degree of certainty) instead of crisp.

For Scene 12, the description A provided by *QSn3D* (see Table 4) could be confused with Scene 11 (and the other way round). The reason for this could be that in these scenes there are 3 objects and in both A descriptions the wooden chair and the white table are on the right with respect to the white chair, without any information indicating which object is *in between*. The results from the survey for descriptions regarding Scene 12 indicated that: 69% agreed on Scene 12, but 21% selected Scene 11. Description B helps to discriminate between scenes, since it starts the description by selecting the wooden chair, which is located in different locations in both scenes. This may be the reason why our participants preferred description B for Scene 12 (see Table 4): 59% preferred B vs. 23% preferred A. However, the results of the survey show that participants matched Scene 11 with its description 98% of times, while only 2% confused it with Scene 12. The reason why Scene 12 was confused with Scene 11, but not the other way round may be because Description A was provided in the first place in the question regarding Scene 12, while Description B was provided in the first place for Scene 11. Again, note that description B is more discriminative, and that participants noticed that: 61% preferred B vs. 29% preferred A.

We did not find robust general preferences for any of the two strategies for generating descriptions (descriptions A or B, i.e. starting the description with the most salient or the closest object). Possibly, both strategies are well-suited for natural descriptions. However, more studies are needed in order to support this suggestion, which we will take into account in future work. However, we were mostly interested to validate the overall adequacy of the generated descriptions and indeed found that most descriptions were well understood by the participants. This provides overall trust in our proposed *QSn3D* approach.

Control Question. Participants may not always read or follow instructions carefully, even the most diligent participants sometimes get tired or distracted. In order to minimize noise, a straightforward and simple solution is to introduce attention checks or an Instructional Manipulation Check (IMC, [Oppenheimer et al, 2009](#)). In general, an IMC decreases noise in data and increases statistical power, while it also increases reliability and validity of our data. As IMC, a control question was introduced in the validation survey showing a description which did not match any of the scenes presented in this paper. This description was the following:

- A. In the background, there is a white-table on the left. There is a white-chair on the right (oriented to the back).
- B. In the background, there is a white-chair on the right (oriented to the back). The white-chair has a white-table on the left.

The participants who noticed that the description was not matching any scene selected the option *none* or provided a corrected description in the second question where they were asked about their preferences for the A or B description.

We thought that this control question was easy to identify since there is no white chair facing to the back in any of the scenes. Apparently, however, it was not so obvious since 31% of the participants selected Scene 9. This might be because both pieces of furniture mentioned in the fake description appeared in Scene 9. However, as some participants may have failed the question not on purpose, but because of their general poor understanding of the task, we carried out a second analysis. By discarding the responses of those participants who did not identify the control question, 40 participants remain (63% paid participants and 37% volunteers). Although the acceptability of the generated scene descriptions increases to 90%–98% (Table 5), the results and their analysis do not change qualitatively. This is why we can be confident in the overall goodness of the *QSn3D* approach.

8 Discussion

The *QSn3D* presented in this paper is a step towards establishing an interactive human-computer dialogue. In human-human communication, common ground in language is achieved as a joint activity through interaction ([Clark, 1996](#); see also language games for robot communication, e.g., [Steels, 2015](#)). However, when people talk to task-oriented robots, they do not necessarily speak to robots the same way as they speak to other people ([Carlson et al, 2014](#)).

This research work is part of a larger project whose main goal is to develop a system with *spatial* intelligence, that is, a system able to understand *what* has been changed in space and how to *reverse* this change, if necessary. For example, this would help to establish tidying tasks in household environments (first scenario presented in the Introduction).

Table 5 Confusion matrix, % of answers by 40 participants (in rows) for each scene (in columns).

Scene	1	2	3	4	5	6	7	8	9	10	11	12
1	93			7	2.5							
2	7	98			2							
3			95	39								
4			5	42								
5					98	2.5						
6				7		95						
7							90					
8							2.5	95	2.5			
9									95			
10										93		5
11							7.5				98	22
12								2.5	5	2		73

Moreover, natural language capabilities will allow the system to explain changes to the user.

The *QSn3D* could be combined with psycholinguistic research, for instance the Attentional Vector-Sum (AVS) model (Regier and Carlson, 2001). The AVS model is a cognitive model that tries to explain the mechanism underlying spatial term comprehension. It can be used to determine acceptable locations for a located object with respect to a reference object corresponding to a spatial term. There are extensions to the AVS model (Carlson et al, 2006; Kluth and Schultheis, 2014) that integrate the functionality of objects into these processes, as well as recent modifications to the AVS model linking it to new empirical findings (Kluth et al, 2017, submitted). Due to its cognitive nature it would support the *QSn3D* approach in generating more natural language.

Finally, the *QSn3D* approach obtains first-order logics which can be easily translated to description logics (Falomir et al, 2011b). We envision a future World Wide Web (WWW) which may include a Web for robots to share information regarding real objects and human environments, from which they will retrieve information to carry out their tasks (cloud robotics, Waibel et al, 2011; Tenorth and Beetz, 2013). The information provided by the *QSn3D* approach presented here might be part of that Web which other robots might use for detecting objects in their environment and for giving spatial descriptions taking them as a reference.

9 Conclusions and future work

This paper presents a qualitative spatial 3D scene descriptor (*QSn3D*) which detects objects in real scenes and describes their locations and orientations using deictic and intrinsic reference systems. As a result, absolute but also relative locations of objects with respect to other oriented objects are described. These qualitative descriptors are the input to a grammar which produces natural language narratives. Moreover, first-order logics are also obtained for further reasoning.

This paper also studies how to organize these qualitative spatial descriptors in order to produce a cognitive explanation. For that, a survey test was carried out with human participants which openly described a scene containing some pieces of furniture. The data obtained in this survey were analysed and the most common saliency strategies identified were the following: (1) naming objects by closeness to the observer (relative feature) and (2) naming objects by their size in the scene (absolute feature).

The *QSn3D* approach has been developed and tested using a Microsoft XBox 360 Kinect in combination with ROS and PCL to obtain 3D-data from the scene. Features are computed on 3D-data and used to generate a SVM-model for classifying different objects in the scene. Using the 3D-coordinates and the orientation of the objects, qualitative spatial relations between the objects are obtained to generate a natural-language and logic descriptions of the scenes.

In order to validate the obtained natural language descriptions, a validation survey was carried out. The conclusions extracted from this validation survey are the following. The obtained acceptability of the scene descriptions is between 82%–95% considering the responses provided by 70 participants, while it increases to 90%–98% when considering the responses given by the 40 participants who read more carefully and identified the control question (one description was not matching any scene). Moreover, the lessons learned from two not so well understood descriptions were that (1) scene descriptions could be improved avoiding crisp locations and using vague locations which include different alternatives for locations by giving them a degree of certainty, and (2) in scenes involving three objects, starting the description by the object located in the middle or by including the location relation *in between* may increase human readability.

As future work, we intend to: (1) use the logics obtained to reason about changes in the scenes; (2) integrate the *QSn3D* descriptor presented here with the the Qualitative 3D descriptor for reasoning about depth perspectives (*Q3D*) (Falomir, 2015) to infer views of the objects and to accelerate their object identification and learning process applied to the point clouds; (3) combine the *QSn3D* with psycholinguistic research, such as the AVS-model and its extensions (Regier and Carlson, 2001; Carlson et al, 2006; Kluth and Schultheis, 2014) and modifications (Kluth et al, 2017, submitted) and (4) extend the grammar presented here to describe in natural language not only static objects but also moving objects and their motion using qualitative movement descriptors (QMD) (Falomir and Rahman, 2015).

Acknowledgements This work was conducted on the scope of the project *Cognitive Qualitative Descriptions and Applications* (CogQDA: <https://sites.google.com/site/cogqda/>) (CogQDA) funded by the Central Research Development Fund (CRDF) at Universität Bremen through the *04-Independent Projects for Postdocs action*. The authors also thank

Niels Eicke, Susanne Knoop, Bengt Kohrt, Nico Lehmann and Mareike Picklum for helping with the implementation.

Appendix

More results obtained by QSn3D (narratives and logics) are shown in Table 6, 7, 8, and 9.

References

- Barclay M, Galton A (2013) Selection of reference objects for locative expressions: The importance of knowledge and perception. In: Tenbrink T, Wiener J, Claramunt C (eds) *Representing Space in Cognition: Interrelations of Behavior, Language, and Formal Models, Explorations in Language and Space*, Oxford University Press, pp 57–169
- Bo L, Lai K, Ren X, Fox D (2011a) Object recognition with hierarchical kernel descriptors. In: *In Proc. of Computer Vision and Pattern Recognition*
- Bo L, Ren X, Fox D (2011b) Depth kernel descriptors for object recognition. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, IEEE*, pp 821–826
- Carlson LA, Regier T, Lopez W, Corrigan B (2006) Attention Unites Form and Function in Spatial Language. *Spatial Cognition & Computation* 6(4):295–308
- Carlson LA, Skubic M, Miller J, Huo Z, Alexenko T (2014) Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task. *topiCS* 6(3):513–533, DOI 10.1111/tops.12101
- Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27, URL <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Clark HH (1996) *Using Language*. Cambridge University Press, Cambridge, UK
- Du H, Henry P, Ren X, Cheng M, Goldman D, Seitz SM, Fox D (2011) Interactive 3D modeling of indoor environments with a consumer depth camera. In: *Proc. 13th Int. Conf. on Ubiquitous computing, ACM, New York, NY, USA, UbiComp '11*, pp 75–84
- Falomir Z (2013) Towards cognitive image interpretation qualitative descriptors, domain knowledge and narrative generation. In: K Gibert VB, Reig-Bolao R (eds) *Artificial Intelligence Research and Development, Frontiers in Artificial Intelligence and Applications*, vol 256, IOS Press, pp 45–57
- Falomir Z (2015) A qualitative model for reasoning about 3D objects using depth and different perspectives. In: Lechowski T, Walega P, Zawidzki M (eds) *LQMR 2015 Workshop, PTI, Annals of Computer Science and Information Systems*, vol 7, pp 3–11, DOI 10.15439/2015F370
- Falomir Z, Rahman S (2015) From qualitative descriptors of movement towards spatial logics for videos. In: *Proc. 3rd Workshop on Recognition and Action for Scene Understanding (REACTS)*, co-located at 16th Int. Conf. of Computer Analysis of Images and Patterns (CAIP), Valletta, Malta, pp 119–128
- Falomir Z, Castelló V, Escrig MT, Peris JC (2011a) Fuzzy distance sensor data integration and interpretation. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems IJUFKS* 19(3):499–528, DOI 10.1142/S0218488511007106
- Falomir Z, Jiménez-Ruiz E, Escrig MT, Museros L (2011b) Describing images using qualitative models and description logics. *Spatial Cognition and Computation* 11(1):45–74, DOI 10.1080/13875868.2010.545611
- Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 24(6):381–395
- Genesereth MR, Nilsson NJ (1987) *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann Publishers
- Henry P, Krainin M, Herbst E, Ren X, Fox D (2010) RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In: *In RGB-D: Advanced Reasoning with Depth Cameras Workshop in conjunction with RSS*
- Herbst E, Henry P, Ren X, Fox D (2011a) Toward object discovery and modeling via 3-D scene comparison. In: *ICRA, IEEE*, pp 2623–2629
- Herbst E, Ren X, Fox D (2011b) RGB-D object discovery via multi-scene analysis. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, IEEE*, pp 4850–4856
- Hernández D, Clementini E, Di Felice P (1995) Qualitative distances. In: Frank AU, Kuhn W (eds) *Spatial Information Theory - A Theoretical Basis for GIS (COSIT'95)*, Springer, Berlin, Heidelberg, pp 45–57
- Huo Z, Skubic M (2016) Natural spatial description generation for human-robot interaction in indoor environments. In: *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp 1–3, DOI 10.1109/SMARTCOMP.2016.7501708
- Ji Z, Yu Y, Pang Y, Chen L, Zhang Z (2017) Zero-shot learning with multi-battery factor analysis. *Signal Processing*
- Kluth T, Schultheis H (2014) Attentional distribution and spatial language. In: Freksa C, Nebel B, Hegarty M, Barkowsky T (eds) *Spatial Cognition IX, Lecture Notes in Computer Science*, vol 8684, Springer International Publishing, pp 76–91, DOI 10.1007/978-3-319-11215-2_6
- Kluth T, Burigo M, Knoeferle P (2017) Modeling the directionality of attention during spatial language comprehension. In: Herik Jvd, Filipe J (eds) *Agents and Artificial Intelligence, Lecture Notes in Computer Science*, vol 10162, Springer International Publishing AG, chap 16, pp 283–301, DOI 10.1007/978-3-319-53354-4_16
- Kluth T, Burigo M, Schultheis H, Knoeferle P (submitted) Does direction matter? Linguistic asymmetries reflected in visual attention. *Cognition*
- Krainin M, Henry P, Ren X, Fox D (2011) Manipulator and object tracking for in-hand 3D object modeling. *Int J Rob Res* 30(11):1311–1327, DOI 10.1177/0278364911403178
- Lai K, Bo L, Ren X, DFox (2011a) Sparse distance learning for object recognition combining RGB and depth information. In: *IEEE International Conference on on Robotics and Automation*
- Lai K, Bo L, Ren X, Fox D (2011b) A scalable tree-based approach for joint object and pose recognition. In: *Twenty-Fifth Conference on Artificial Intelligence (AAAI)*
- Landau B (2016) Update on what and where in spatial language: A new division of labor for spatial terms. *Cognitive Science* pp 1–30, DOI 10.1111/cogs.12410
- Levinson S (2003) *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press
- Lison P (2010) *Robust Processing of Spoken Situated Dialogue*. Diplomica Verlag
- Lloyd JW (1987) *Foundations of logic programming. Symbolic computation: Artificial intelligence*. Springer-Verlag, 2nd, extended edition edition
- Marton ZC, Pangercic D, Rusu RB, Holzbach A, Beetz M (2010) Hierarchical object geometric categorization and appearance classification for mobile manipulation. In: *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on, IEEE*, pp 365–370
- Mast V, Falomir Z, Wolter D (2016) Probabilistic reference and grounding with PRAGR for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence* 28(5):889–911, DOI 10.1080/0952813X.2016.1154611
- Moratz R, Tenbrink T (2006) Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a

- model of projective relations. *Spatial Cognition and Computation* 6(1):63–106
- Moratz R, Tenbrink T (2008) Affordance-based human-robot interaction. In: Proc. of the 2006 International conference on Towards affordance-based robot control, Springer-Verlag, Berlin, Heidelberg, pp 63–76
- Museros L, Falomir Z, Sanz I, Gonzalez-Abril L (2014) Sketch retrieval based on qualitative shape similarity matching: Towards a tool for teaching geometry to children. *AI Communications* 28(1):73–86, DOI 10.3233/AIC-140614
- Olszewska JI (2015a) 3D spatial reasoning using the clock model. In: Bramer M, Petridis M (eds) *Research and Development in Intelligent Systems XXXII: Incorporating Applications and Innovations in Intelligent Systems XXIII*, Springer International Publishing, Cham, pp 147–154, DOI 10.1007/978-3-319-25032-8_10
- Olszewska JI (2015b) Where is my cup? - fully automatic detection and recognition of textureless objects in real-world images. In: Azopardi G, Petkov N (eds) *Computer Analysis of Images and Patterns: 16th Int. Conf., CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I*, Springer International Publishing, pp 501–512, DOI 10.1007/978-3-319-23192-1_42
- Olszewska JI (2016) Interest-point-based landmark computation for agents' spatial description coordination. In: van den Herik HJ, Filipe J (eds) *Proc. of the 8th Int. Conf. on Agents and Artificial Intelligence (ICAART 2016)*, Volume 2, Rome, Italy, February 24–26, 2016., SciTePress, pp 566–569, DOI 10.5220/0005847705660569
- Oppenheimer DM, Meyvis T, Davidenko N (2009) Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45(4):867–872, DOI <https://doi.org/10.1016/j.jesp.2009.03.009>
- Pederson E (2003) How many reference frames? In: Freksa C, Brauer W, Habel C, Wender KF (eds) *Spatial Cognition III: Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning*, Springer, Berlin, Heidelberg, pp 287–304, DOI 10.1007/3-540-45004-1_17
- Regier T, Carlson LA (2001) Grounding Spatial Language in Perception: An Empirical and Computational Investigation. *Journal of Experimental Psychology: General* 130(2):273–298, DOI 10.1037/0096-3445.130.2.273
- Ruiz-Sarmiento JR (2016) Probabilistic techniques in semantic mapping for mobile robotics. PhD thesis, Department of Systems Engineering and Automatics, University of Malaga, Spain
- Ruiz-Sarmiento JR, Galindo C, González-Jiménez J (2015) Olt: A toolkit for object labeling applied to robotic RGB-D datasets. In: *European Conference on Mobile Robots*
- Rusu RB, Bradski G, Thibaux R, Hsu J (2010) Fast 3D recognition and pose using the viewpoint feature histogram. In: *Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan
- Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: *Proc. of 2011 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Washington, DC, USA, CVPR '11, pp 1297–1304
- Skubic M, Blisard S, Bailey C, Adams J, Matsakis P (2004) Qualitative analysis of sketched route maps: Translating a sketch into linguistic descriptions. *IEEE Trans Syst, Man, Cyber B, Cybern* 34(2):1275–1282
- Socher R, Ganjoo M, Manning CD, Ng A (2013) Zero-shot learning through cross-modal transfer. In: *Advances in neural information processing systems*, pp 935–943
- Steels L (2015) The Talking Heads experiment: Origins of words and meanings. *Computational Models of Language Evolution*, Language Science Press, DOI 10.17169/langsci.b49.75, URL <http://langsci-press.org/catalog/book/49>
- Steinhauer HJ (2005) A qualitative model for natural language communication about vehicle traffic. In: *AAAI Spring Symposium: Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance*, AAAI, pp 52–57
- Tenbrink T, Fischer K, Moratz R (2002) Spatial strategies in linguistic human-robot communication. In: Freksa C (ed) *KI-Themenheft 4/02 Spatial Cognition*, arenDTaP Verlag, pp 19–23
- Tenbrink T, Maiseyenko V, Moratz R (2007) Spatial reference in simulated human-robot interaction involving intrinsically oriented objects. In: *Symposium Spatial Reasoning and Communication at AISB'07 Artificial and Ambient Intelligence*, vol 7
- Tenbrink T, Coventry KR, Andonova E (2011) Spatial strategies in the description of complex configurations. *Discourse Processes* 48(4):237–266
- Tenorth M, Beetz M (2013) Knowrob: A knowledge processing infrastructure for cognition-enabled robots. *The International Journal of Robotics Research* 32(5):566–590, DOI 10.1177/0278364913481635
- Waibel M, Beetz M, Civera J, D'Andrea R, Elfring J, Galvez-Lopez D, Haussermann K, Janssen R, Montiel J, Perzylo A, Schiessle B, Tenorth M, Zweigle O, van de Molengraft R (2011) Roboearth. *Robotics Automation Magazine, IEEE* 18(2):69–82, DOI 10.1109/MRA.2011.941632
- Zhang X, quan Li Q, xiang Fang Z, wei Lu S, lung Shaw S (2014) An assessment method for landmark recognition time in real scenes. *Journal of Environmental Psychology* 40:206–217, DOI <http://dx.doi.org/10.1016/j.jenvp.2014.06.008>

Table 6 QSn3D narratives and logics obtained in the home scenario using 2 pieces of furniture: 1 oriented and 1 non-oriented.

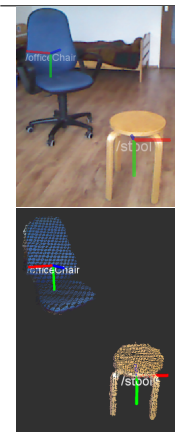

Photo	Language description	Logic description
	<p>Scene 3</p> <p>A. The biggest object (an office chair) as the most salient object: <i>In the background, there is an office chair on the left (oriented to the front right). The office chair has a stool in the front.</i></p> <p>B. The closest object to the observer (a stool) as the most salient object: <i>In the foreground, there is a stool in the centre. In the background, there is an office chair on the left (oriented to the front right).</i></p>	<pre>is_categorized(object_0z, stool). location_wrt_observer(object_0z, centre). distance_wrt_observer(object_0z, foreground). close_object(object_0z, object_1z). is_oriented(no, object_0z, none). is_categorized(object_1z, officeChair). is_categorized(officeChair, chair). location_wrt_observer(object_1z, left). distance_wrt_observer(object_1z, background). close_object(object_1z, object_0z). is_oriented(yes, object_1z, front_right). location_wrt_close_object(object_1z, object_0z, centre).</pre>
	<p>Scene 4</p> <p>A. The biggest object (an office chair) as the most salient object: <i>In the background, there is an office chair in the centre (oriented to the front). The office chair has a stool on the right.</i></p> <p>B. The closest object to the observer (a stool) as the most salient object: <i>In the foreground, there is a stool in the centre. In the background, there is an office chair in the centre (oriented to the front).</i></p>	<pre>is_categorized(object_0d, officeChair). is_categorized(officeChair, chair). location_wrt_observer(object_0d, centre). distance_wrt_observer(object_0d, background). close_object(object_0d, object_1d). is_oriented(yes, object_0d, front). location_wrt_close_object(object_0d, object_1d, right). is_categorized(object_1d, stool). location_wrt_observer(object_1d, centre). distance_wrt_observer(object_1d, foreground). close_object(object_1d, object_0d). is_oriented(no, object_1d, none).</pre>

Table 7 QSn3D narratives and logics obtained in the home scenario using 3 pieces of furniture: 1 oriented and 2 non-oriented.

Photo	Language description	Logic description
	<p>Scene 5</p> <p>A. The biggest object (an armchair) as the most salient object: <i>In the background, there is an armchair in the centre (oriented to the right). The armchair has a rubbish bin on the right and a stool behind.</i></p> <p>B. The closest object to the observer (a rubbish bin) as the most salient object: <i>In the foreground, there is a rubbish bin in the centre. In the background, there is an armchair in the centre (oriented to the right). The armchair has a stool behind.</i></p>	<pre>is_categorized(object_0e, rubbishBin). location_wrt_observer(object_0e, centre). distance_wrt_observer(object_0e, foreground). close_object(object_0e, object_2e). is_oriented(no, object_0e, none). is_categorized(object_1e, stool). location_wrt_observer(object_1e, left). distance_wrt_observer(object_1e, background). close_object(object_1e, object_2e). is_oriented(no, object_1e, none). is_categorized(object_2e, armChair). is_categorized(object_2e, chair). location_wrt_observer(object_2e, centre). distance_wrt_observer(object_2e, background). close_object(object_2e, object_0e). close_object(object_2e, object_1e). is_oriented(yes, object_2e, right). location_wrt_close_object(object_2e, object_0e, right). location_wrt_close_object(object_2e, object_1e, behind).</pre>
	<p>Scene 6</p> <p>A. The biggest object (an office chair) as the most salient object: <i>In the background, there is an office chair on the left (oriented to the front right). The office chair has a rubbish bin on the left and a stool on the left.</i></p> <p>B. The closest object to the observer (a rubbish bin) as the most salient object: <i>In the foreground, there is a rubbish bin in the centre. In the background, there is a stool in the centre. There is an office chair on the left (oriented to the front right).</i></p>	<pre>is_categorized(object_0w, rubbishBin). location_wrt_observer(object_0w, centre). distance_wrt_observer(object_0w, foreground). close_object(object_0w, object_1w). is_oriented(no, object_0w, none). is_categorized(object_1w, stool). location_wrt_observer(object_1w, centre). distance_wrt_observer(object_1w, background). close_object(object_1w, object_0w). close_object(object_1w, object_2w). is_oriented(no, object_1w, none). is_categorized(object_2w, officeChair). is_categorized(object_2w, chair). location_wrt_observer(object_2w, left). distance_wrt_observer(object_2w, background). close_object(object_2w, object_1w). is_oriented(yes, object_2w, front_right). location_wrt_close_object(object_2w, object_1w, left).</pre>

Table 8 QSn3D narratives and logics obtained in the office scenario using 2 pieces of furniture: 1 oriented and 1 non-oriented.

Photo	Language description	Logic description
 	Scene 7 A. The biggest object (a white-table) as the most salient object: <i>In the foreground, there is a white-table on the left. There is a wooden-chair in the centre (oriented to the back).</i>	<pre>is_categorized(object_0b, white-table). location_wrt_observer(object_0b, left). distance_wrt_observer(object_0b, foreground). close_object(object_0b, object_1b). is_oriented(no, object_0b, none). is_categorized(object_1b, wooden-chair). location_wrt_observer(object_1b, centre). distance_wrt_observer(object_1b, foreground). close_object(object_1b, object_0b). is_oriented(yes, object_1b, back). location_wrt_close_object(object_1b, object_0b, left).</pre>
	B. The closest object to the observer (a wooden-chair) as the most salient object: <i>In the foreground, there is a wooden-chair in the centre (oriented to the back). The wooden-chair has a white-table on the left.</i>	
 	Scene 8 A. The biggest object (a wooden-chair) as the most salient object: <i>In the background, there is a wooden-chair on the right (oriented to the left). The wooden-chair has a white-table in the front.</i>	<pre>is_categorized(object_0j, wooden-chair). location_wrt_observer(object_0j, right). distance_wrt_observer(object_0j, background). close_object(object_0j, object_1j). is_oriented(yes, object_0j, left). location_wrt_close_object(object_0j, object_1j, centre). is_categorized(object_1j, white-table). location_wrt_observer(object_1j, centre). distance_wrt_observer(object_1j, foreground). close_object(object_1j, object_0j). is_oriented(no, object_1j, none).</pre>
	B. The closest object to the observer (a white-table) as the most salient object: <i>In the foreground, there is a white-table in the centre. In the background, there is a wooden-chair on the right (oriented to the left).</i>	
 	Scene 9 A. The biggest object (a white-chair) as the most salient object: <i>In the background, there is a white-chair in the centre (oriented to the front right). The white-chair has a white-table in the front.</i>	<pre>is_categorized(object_0b, white-chair). location_wrt_observer(object_0b, centre). distance_wrt_observer(object_0b, background). close_object(object_0b, object_1b). is_oriented(yes, object_0b, front_right). location_wrt_close_object(object_0b, object_1b, centre). is_categorized(object_1b, white-table). location_wrt_observer(object_1b, right). distance_wrt_observer(object_1b, foreground). close_object(object_1b, object_0b). is_oriented(no, object_1b, none).</pre>
	B. The closest object to the observer (a white-table) as the most salient object: <i>In the foreground, there is a white-table on the right. In the background, there is a white-chair in the centre (oriented to the front right).</i>	

Table 9 QSn3D narratives and logics obtained in the office scenario using 3 pieces of furniture: 2 oriented and 1 non-oriented.

Photo	Language description	Logic description
 	<p>Scene 10</p> <p>A. The biggest object (a white-chair) as the most salient object: <i>In the background, there is a white-chair in the centre (oriented to the front). The white-chair has a white-table on the right and a wooden-chair on the left.</i></p> <p>B. The object closest to the observer (a wooden-chair) as the most salient object: <i>In the background, there is a wooden-chair on the right (oriented to the front). The wooden-chair has a white-table on the right and a white-chair on the left.</i></p>	<pre> is_categorized(object_01, white-chair). location_wrt_observer(object_01, centre). distance_wrt_observer(object_01, background). close_object(object_01, object_11). close_object(object_01, object_21). is_oriented(yes, object_01, front). location_wrt_close_object(object_01, object_11, left). location_wrt_close_object(object_01, object_21, right). is_categorized(object_11, wooden-chair). location_wrt_observer(object_11, right). distance_wrt_observer(object_11, background). close_object(object_11, object_01). close_object(object_11, object_21). is_oriented(yes, object_11, front). location_wrt_close_object(object_11, object_01, right). location_wrt_close_object(object_11, object_21, right). is_categorized(object_21, white-table). location_wrt_observer(object_21, left). distance_wrt_observer(object_21, foreground). close_object(object_21, object_01). close_object(object_21, object_11). is_oriented(no, object_21, none). </pre>
 	<p>Scene 11</p> <p>A. The biggest object (a white-chair) as the most salient object: <i>In the foreground, there is a white-chair on the right (oriented to the front left). The white-chair has a white-table on the right and a wooden-chair on the right.</i></p> <p>B. The object closest to the observer (a wooden-chair) as the most salient object: <i>In the foreground, there is a wooden-chair in the centre (oriented to the back). The wooden-chair has a white-chair on the right and a white-table on the left.</i></p>	<pre> is_categorized(object_0h, white-table). location_wrt_observer(object_0h, left). distance_wrt_observer(object_0h, foreground). close_object(object_0h, object_1h). close_object(object_0h, object_2h). is_oriented(no, object_0h, none). location_wrt_close_object(object_0h, object_1h, behind). location_wrt_close_object(object_0h, object_2h, right). is_categorized(object_1h, wooden-chair). location_wrt_observer(object_1h, centre). distance_wrt_observer(object_1h, foreground). close_object(object_1h, object_0h). close_object(object_1h, object_2h). is_oriented(yes, object_1h, back). location_wrt_close_object(object_1h, object_0h, left). location_wrt_close_object(object_1h, object_2h, right). is_categorized(object_2h, white-chair). location_wrt_observer(object_2h, right). distance_wrt_observer(object_2h, foreground). close_object(object_2h, object_0h). close_object(object_2h, object_1h). is_oriented(yes, object_2h, front_left). location_wrt_close_object(object_2h, object_0h, right). location_wrt_close_object(object_2h, object_1h, right). </pre>
 	<p>Scene 12</p> <p>A. The biggest object (a white-chair) as the most salient object: <i>In the background, there is a white-chair on the right (oriented to the front). The white-chair has a wooden-chair on the right and a white-table on the right.</i></p> <p>B. The object closest to the observer (a wooden-chair) as the most salient object: <i>In the foreground, there is a wooden-chair on the left (oriented to the back). The wooden-chair has a white-chair on the right and a white-table on the right.</i></p>	<pre> is_categorized(object_0i, white-chair). location_wrt_observer(object_0i, right). distance_wrt_observer(object_0i, background). close_object(object_0i, object_1i). close_object(object_0i, object_2i). is_oriented(yes, object_0i, front). location_wrt_close_object(object_0i, object_1i, right). location_wrt_close_object(object_0i, object_2i, right). is_categorized(object_1i, wooden-chair). location_wrt_observer(object_1i, left). distance_wrt_observer(object_1i, foreground). close_object(object_1i, object_0i). close_object(object_1i, object_2i). is_oriented(yes, object_1i, back). location_wrt_close_object(object_1i, object_0i, right). location_wrt_close_object(object_1i, object_2i, right). is_categorized(object_2i, white-table). location_wrt_observer(object_2i, centre). distance_wrt_observer(object_2i, background). close_object(object_2i, object_0i). close_object(object_2i, object_1i). is_oriented(no, object_2i, none). </pre>