# Studying the Role of Location in 3D Scene Description using Natural Language

**Thomas Kluth**
Universität Bremen

**Zoe Falomir***
Universität Bremen

## Abstract

In this paper the description of 3D indoor scenes in natural language is studied from the point of view of intrinsic and relative location of the objects. An approach has been developed for this purpose which uses a XBox 360 Kinect in combination with ROS and PCL to obtain 3D-data from the scene. Object features are computed on these 3D-data, which are used to generate a SVM-model which classifies the different objects in the scene. After detecting the objects in the scene, their orientation is obtained and qualitative spatial relations between the objects are computed to generate a natural-language description of the scene.

## 1 Introduction

Imagine you have a robot at home that every morning after you leave to work it arranges the furniture in your living room that you untidied the previous night. Or imagine that you move to a new house and a decorator tutor helps you to arrange new furniture in your rooms in a nicely way. Those situations would both involve ambient intelligence. In the first one, your robot at home would need scene understanding for identify the objects and their spatial locations in the living room, detect changes from the desired arrangement and then interact with the environment. In the second one, the decorator tutor would involve human machine interaction. It would need to produce natural language descriptions to give instructions to the users, then it would need scene understanding for interpreting if the user has done nicely the task described or not, and finally it would also interpret the changes to provide feedback to the users. These systems are still 'ideal', but there are a lot of research works in the literature that are focusing in solving those challenges. In this paper, we focus on scene understanding by detecting pieces of furniture in a 3D scene and describing its location using natural language descriptions based on qualitative models of space description. For human-robot/systems interaction qualitative models are really useful because they can deal with abstractions, uncertainty, etc. and they produce descriptions which can be understood by people.

*Correspondence to: Zoe Falomir, Cognitive Systems (CoSy), FB3 - Informatics, Universität Bremen, P.O. Box 330 440, 28334 Bremen, Germany. E-mail: zfalomir@informatik.uni-bremen.de

## 2 Related Work

Recognizing objects is currently one of the most challenging tasks in the field of 3D computer vision and robotics, because 3D data usually suffer from distortions due to noisy sensors, viewpoint changes and point density variations. 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 cameras) such as the Microsoft Kinect and the Asus Xtion [1].

There have been a large number of research efforts in applying RGB-Depth perception for enabling robots to operate in unstructured real-world environments. Some of the key challenges in this direction are understanding humans and their worlds, which is basic for robots to operate and perform various tasks in human environments. In the literature, interesting progresses in this direction can be found, most of them related to two serial of workshops happening yearly:

- *RSS Workshop on RGB-Depth: Advanced Reasoning with Depth Cameras*: 2010-2013[2].
- *ICRA Workshop on Semantic Perception Mapping and Exploration (SPME)*: 2011- 2013[3].

On the other hand, cognitive studies can be found in the literature which investigates how people describe object arrangements in the space [Tenbrink *et al.*, 2007, 2011]. Some of the results obtained were applied to improve human-robot interaction by Moratz and Tenbrink [2008, 2006], which used a robot incorporating a range laser sensor to extract information from the environment. Taking into account these previous works, here we will apply the results of cognitive studies to the description of a 3D scene captured by a Xbox Kinect device, which provides very rich spatial information about the

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

[2] `http://www.cs.washington.edu/ai/Mobile\_Robotics/rgbd-workshop-2013/`

[3] `http://www.spme.ws`

*Actas de las XV Jornadas de ARCA, JARCA 2013*
*Sistemas Cualitativos y sus Aplicaciones en*
*Diagnosis, Robótica e Inteligencia Ambiental*
*I. Sanz, Ll. Museros, J. A. Ortega (eds.)*

*ISBN: 978-84-616-7622-4 © 2013*

space (i.e. information about depth, for distinguishing foreground from background, textures of the objects for identifying them, etc.).

## 3 Describing 3D Object Location in Natural Language

A first step for describing a 3D scene involves detecting the objects within and then describing their locations. Here an approach is proposed for detecting objects in 3D pointclouds obtained by a Kinect RGB-Depth camera (Section 3.1) and for describing their locations from an intrinsic or relative point of view (Section 3.2).

### 3.1 3D-Object Recognition

Our approach obtains the 3D-pointclouds in the scene and it proceeds in the following way:

1. the floor in the scene is extracted by applying a RANSAC-based segmentation (RANdom Sample And Consensus) [Fischler and Bolles, 1981].
2. an Euclidean Cluster Extraction is carried out in order to distinguish different objects. For each extracted cluster, two geometrical 3D-features are calculated:
   (a) the Viewpoint Feature Histogram (VFH) which is described in more detail by Rusu *et al.* [2010]. It is scale invariant but viewpoint variant. The main idea of this feature 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) by Marton *et al.* [2010]. The basic idea of GRSD is to approximate 3D-objects by searching for best-fitting circles at each point.

For each type of object, a bunch of pointclouds is obtained, recorded and labeled with the name of the object. These pointclouds contain different orientations and scales of the objects. With these labeled feature vectors a SVM-model is trained, which is later used to classify extracted clusters by using LIBSVM by Chang and Lin [2011].

### 3.2 Spatial References in Natural Language

In spatial expressions, *projective terms* capture the idea that a spatial relationship is *projected* from an origin (position anchoring the view direction) to a *relatum* (a known object nearby) in order to specify the location of the intended object, called also the *locatum* [Tenbrink *et al.*, 2007]. This is done using lexical items such as *front, back, left, right*.

The employment of projective terms presupposes underlying conceptual reference systems, which were systematically categorized by Levinson [2003] as *relative* versus *intrinsic*. In relative reference, a viewer specifies the location of an object relative to a relatum, as in *The chair is in front of the table*. Here, the relatum does not necessarily possess intrinsic sides, and the reference system consists of three different positions. In intrinsic reference systems, the role of the relatum coincides with the role of origin, which therefore needs to possess intrinsic sides, which then serve as basis for reference. In *The table is in front of me*, the speaker serves both as relatum and as origin, and her/his view direction determines the direction of front. For example, in the scene in Figure 1 the narratives generated from the spatial relations between the two objects may be: *The rubbish bin is in front of the office chair.*

Figure 1: Two spatial configurations which can be described using the same natural language sentence using a relative reference system located at the office chair.

The approach presented here is intending to generate two types of natural language narratives describing location by taking into account:

1. an intrinsic reference system located at the RGB-Depth camera from which the objects in the scene are described and
2. a relative reference system between objects in the scene that have clear orientations, as for example, chairs, sofas, armchairs, etc.

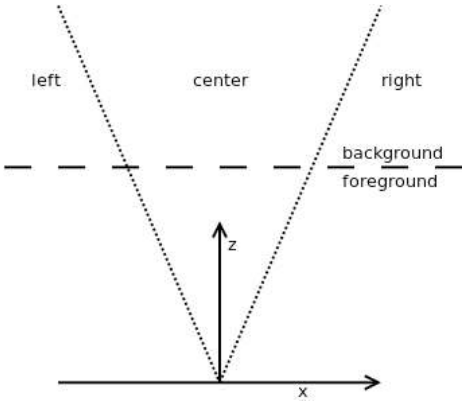


Figure 2: Model for dividing the space observed from a RGB-Depth camera.

To generate these type of natural language descriptions, the coordinates from the 3D-data obtained from the point clouds are used. As Figure 2 shows, the scene is divided into a 3D spatial model for distinguishing between *foreground* and

*background* and locations such as *center, front, back, left, right, back-left, back-right, front-left, front-right*.

Note that for spatial relations between two objects, the reference frame is located at the oriented object so that its front corresponds to the front of the reference system. For both configurations in Figure 1, although the objects are on different positions considering the global reference frame, the relative reference system will produce the same description because the chair has an oriented *front* side, which is taken as reference.

## 4 Experimentation and Results

This approach has been tested in an indoor environment using four pieces of furniture of different colours and sizes: an armchair, an office chair, a rubbish bin and a stool. As a result, a proof-of-concept is presented here. The sensor used for extracting the 3D-pointcloud from the given scene was a Microsoft XBox 360 Kinect. This RGB-depth sensor is based on a structured infrared-light system, which has a ranging limit of roughly 0.7 to 6 m distance, and is applicable in most indoor environments. Experimental results by Khoshelham and Elberink [2012] showed that the random error of depth measurement increases with increasing distance to the sensor, and ranges from a few millimeters up to about 4 cm at the maximum range of the sensor. Ruther *et al.* [2011] investigated how this sensor can be modied to work at much higher accuracy, on a limited but scalable measurement range by altering the sensor baseline and depth of eld, and as a result, they reliably retrieved depth elds of objects at an accuracy in the sub-millimeter range. However, in the work presented, the original depth accuracy provided by the Kinect is enough, since the information extracted here is qualitative and it does not requires exact accurate values and it can also deal with range of depth measurements. The system presented is written in C++ and build upon the Robot Operating System (ROS) framework[4]. To receive the 3D-data from the Kinect device we have used the openNI-driver[5], included in ROS. To process the obtained point clouds the Point Cloud Libray (PCL) framework[6] is used, which is also included in ROS. And for training the SVM-model with the labeled 3D-feature vectors extracted from the clusters, the SVM library is applied (LIBSVM [Chang and Lin, 2011]).

Figure 3 shows the example scenario used in our proof-of-concept. Figure 4 shows the point clouds obtained by the RGB-Depth sensor in the scenario in Figure 3. Figure 5 shows the results after applying our 3D object recognition process. Note the display of probabilities, which can be used to process the cluster further, if it is reliable classified.

Examples of the description of this scene in natural language using an intrinsic reference system placed in the RGB-Depth camera are the following ones:

`There is an armchair and a rubbish bin on the left and a stool and an office chair on the right` or

Figure 3: Scenario for testing.

Figure 4: Point clouds of the scene extracted by the RGB-Depth sensor.

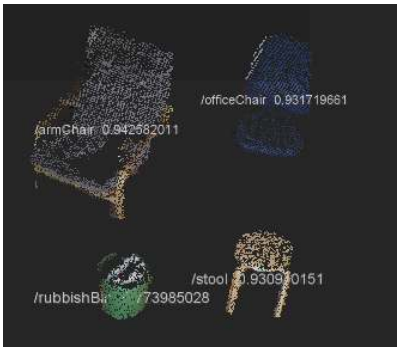


Figure 5: Object recognition in the scene: output of the SVM classification system.

[4] `http://www.ros.org`
[5] `http://www.openni.org`
[6] `http://www.pointclouds.org`

`There is an armchair and an office chair at the back and a rubbish bin and a stool at the front.`

Note that there are two oriented objects in the scene: the armchair and the office chair. Both they have a *front* and a *back* which people use when describing orientation. Examples of the description of this scene in natural language using a relative reference are:

`There is stool in front of the office chair and a rubbish bin in front of the armchair` and

`The office chair is to the left of the armchair.`

## 5 Discussion

In this paper the description of 3D scenes in natural language is studied from the point of view of intrinsic and relative location of the objects. Moreover, the role of depth is also studied to point out objects in the foreground or in the background.

The presented system is able to generate a natural language description of a indoor scene captured by a Microsoft XBox 360 Kinect in combination with ROS and PCL to obtain 3D-data from the scene. Features are computed on this 3D-data, and then they are 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 description of the scene.

As future work, we intend to: (i) further test our approach in the office indoor environment at Cognitive Systems department at Universität Bremen; and (ii) carry out a cognitive test in which we compare the descriptions made by people to the descriptions provided by our system.

## Acknowledgments

This work has been partially supported by Universität Bremen under the project named *"Providing human-understandable qualitative and semantic descriptions"*, the Deutscher Akademischer Austausch Dienst (DAAD), the interdisciplinary Transregional Collaborative Research Center *Spatial Cognition* SFB/TR 8 under Project R3-[Q-Shape], and the Deutsche Forschungsgemeinschaft (DFG).

## References

C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

S.C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, 2003.

Z-C Marton, Dejan Pangercic, Radu Bogdan Rusu, A. Holzbach, and M. Beetz. Hierarchical object geometric categorization and appearance classification for mobile manipulation. In *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*, pages 365–370. IEEE, 2010.

R. Moratz and T. Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, 6(1):63–106, 2006.

R. Moratz and T. Tenbrink. Affordance-based human-robot interaction. In *Proc. of the 2006 International conference on Towards affordance-based robot control*, pages 63–76, Berlin, Heidelberg, 2008. Springer-Verlag.

Radu Bogdan Rusu, G. Bradski, R. Thibaux, and J. Hsu. 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, October 2010.

M. Ruther, M. Lenz, and H. Bischof. nect: On using a gaming rgbd camera in micro-metrology applications. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 52–59, 2011.

T. Tenbrink, V. Maiseyenka, and R. Moratz. Spatial reference in simulated human-robot interaction involving intrinsically oriented objects. In *Symposium Spatial Reasoning and Communication at at AISB'07 Artificial and Ambient Intelligence*, volume 7, 2007.

Thora Tenbrink, Kenny R. Coventry, and Elena Andonova. Spatial strategies in the description of complex configurations. *Discourse Processes*, 48(4):237–266, 2011.