From sensors to human spatial concepts: an annotated data set

Zoran Zivkovic, Olaf Booij, Ben Kröse, Elin A. Topp and Henrik I. Christensen

Abstract— **An annotated data set is presented meant to help researchers in developing, evaluating and comparing various approaches in robotics for building space representations appropriate for communicating with humans. The data consists of omnidirectional images, laser range scans, sonar readings and robot odometry. A set of base-level human spatial concepts is used to annotate the data.**

Index Terms— **Robot space representation, map building, human-robot interaction.**

I. INTRODUCTION

MOBILE robots are expected to become part of our daily life in the near future. Numerous studies show that people tend to perceive robots as social actors and not just tools, and therefore expect to communicate with them in a natural way [22], [1], [6]. One of the basic skills of the future home robot is goal-directed navigation, including localization, path planning and path following. This requires an internal model of the environment. Traditionally mapping and localization has focussed on metric properties and the feature model have typically been point and line based. In general metric place specification is a poor match to human instructions, it is more natural to use cognitive concepts used by humans when communicating about the space, e.g. rooms. In addition, object type entities, a common human concept related to space, should be included in the robot space representation.

In this sense, the problem is basically a pattern recognition problem. Most pattern recognition approaches use large data sets to *learn* to recognize concepts. In many application fields (object recognition, speech understanding), annotated databases of sensory data and corresponding labels (concepts) are available [21]. The databases currently available in the robotics community are mainly focused on the geometrical representations [8]. In order to make progress on learning conceptual representations of space, it is essential to have access to appropriate annotated data-sets that enable supervised learning and performance benchmarking against ground truth. It would be ideal to have huge data sets from extensive user studies for various situations. However, it is difficult to provide such a general data set and it is realistic to start with some specific scenarios that might be of broad interest. In this paper we

Z. Zivkovic, O. Booij and B. Kröse are with University of Amsterdam, Kruislaan 403, 1098SJ Amsterdam, The Netherlands, email:{zivkovic,obooij,krose}@science.uva.nl

E.A.Topp and H.I.Christensen are with Royal Institute of Technology (KTH), Stockholm, Kungl Tekniska Högskolan SE-100 44 Stockholm, Sweden, email:{topp,hic}@csc.kth.se

H.I.Christensen is also with the Georgia Institute of Technology.

present a data set that is closely related to the so called "home tour" scenario. This scenario describes a hypothetical situation in which a person receives a new robot and shows the robot how the home looks like. The data set contains sensor readings from some typical sensors used in robotics. Furthermore we define a set of base-level human spatial concepts and annotate the data. The annotation contains the room where each sensor reading was recorded and the objects visible in the current omnicam image. A flexible annotation tool is provided to allow for other annotations.

In Section 2 of this paper we give a short overview of the current research in robotics that address the problem of building spatial representations that are suited for communication with humans. In addition, references to the more widely studied domain of localization and mapping are presented. Some links are also made to the related research activities in the cognitive science and computer vision. In Section 3 we describe the structure of the provided data set, the design principles and some related practical issues. Section 4 identifies the type of evaluation schemes that can be performed using the presented data set. The conclusions and our final remarks are listed in Section 5.

II. RELATED WORK

Most traditional map building methods in robotics represent geometric properties of the environment, such as occupancy grids or polygonal representations of free space [3]. Another common space representation in robotics is the topological map. A topological map [14], [2], [9] describes the environment as a graph structure with nodes representing distinctive places and edges representing possible transitions.

The traditional maps in robotics are mainly related to the robot navigation task. In order to design space representations appropriate for communicating with humans extensive user studies and results from cognitive psychology are often considered. In "The intelligent use of space" [11] Kirsh stated that to understand complex (human) models of an environment, we have to observe the interaction of the (human) agent with and within the environment. An example of a robot map representation based on findings from cognitive psychology research is the "Spatial Semantic Hierarchy" by Kuipers [13] enabling a robot to explore an environment autonomously along the lines of human exploration strategies. Furthermore, a number of researchers report human-robot interaction studies in a guided tour scenario in which a person is guiding a robot [24], [7], [26], [15]. Kruijff *et al.* [12] consider the issue of clarification dialogues in ambiguous situations of a guided tour. Another
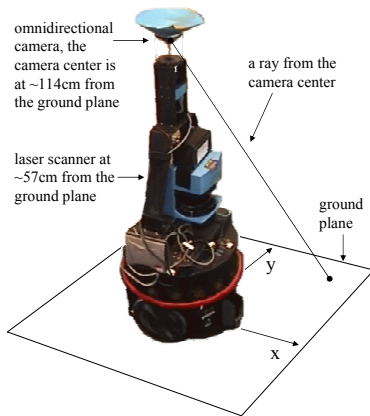
Fig. 1. One of the robots we used for collecting data. The camera axis is approximately aligned with the robot center of rotation. The scanner was mounted so that the origin of the scans had an offset of $\approx$ 70mm in the direction of the x axis of the robot.

Fig. 2. 2D maps of the home environments where the data was recorded. The location of the furniture in the drawing is approximate.

important issue is the adaptation of the environment model to the current situation, i.e. personalization. A number of observations from human-robot interaction user studies [24], [7] support the assumption that personal preferences result in quite large variations in what the persons actually presented to the robot during a guided tour.

The reported human-robot interaction studies indicate that hierarchical (or partially hierarchical) models, as could be confirmed with psychological studies [17], become explicit in the interaction of a human with a robot and form a useful base for communication. An indoor environment is typically divided into "delimited regions" (e.g., rooms). The second common concept is that of 'object'. Particular positions are often described by the relationship to large objects. Our annotation contains objects and rooms in indoor environments.

III. ANNOTATED DATA SET

We describe here how we collected the data and the base-level human spatial concepts we used to annotate the data. The data set with annotations is available from: www2.science.uva.nl/sites/cogniron.

A. Data gathering

The acquisition of the data set took place in three home environments, see Figure 2. The mobile robot was driven around by tele-operation to collect the data, see Figure 1. The following sensors were used:

- Omnidirectional camera - On average 7.5 omnidirectional images per second were taken by a camera with a hyperbolic mirror. The 1024×768 pixel images are in YUV422 color format. The camera is calibrated and images for calibration are available. More information on the omnidirectional vision sensor is presented in the related technical report [28]. The pose of the camera with respect to the robot is known and a Matlab toolbox is provided for performing basic geometric transformations.
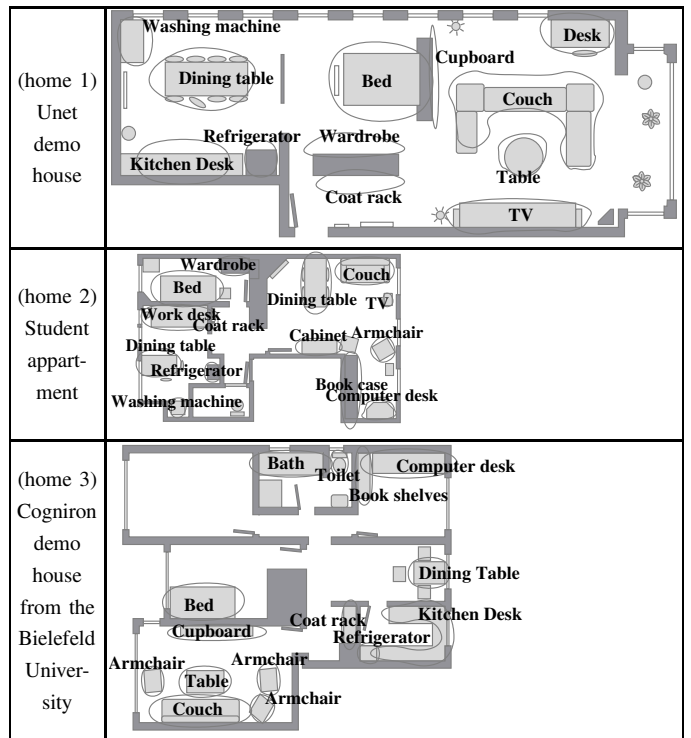- Laser scanner - A SICK-laser (LMS-200) was used to record range scans at the front of the robot. The scanner was running in millimeter precision, 0.5 degree angular resolution over 180 degrees and had approximately 8 meter maximum range. On average 3.5 scans were conducted per second.
- Odometry + sonar - On average 12 odometry measurements per second were taken. Because the robot has solid wheels the odometry is quite accurate. At the same time the current values of the 16 ultrasonic sonar sensors were recorded giving a 360 degrees range scan.

These are the sensors typically used for map building in robotics. An omnidirectional camera was chosen to provide the maximum field of view for navigation. In addition, "low-resolution" rectified images can be resampled from the omnidirectional camera.

The robot was driven through the environment under three different types of conditions:

- Clean data - We performed two tours by driving the robot at a more or less constant speed, without many people around and with constant lighting.
- Noisy data - Two tours were performed by driving the robot with people walking around and with more difficult lighting. Furthermore, a number of objects were moved or changed in appearance. These tours were intended to generate more challenging data.
- Home tour data - Finally we simulated the so called home tour scenario where a person is leading the robot around an environment. Again two tours are performed with two different persons leading the robot. These data are different from the previous runs since there is always a person close to the robot and the persons tend to stop the robot at certain places in the environment to give

explanations about the environment. These data were not meant as a user study but just to simulate the type of sensory inputs the robot would receive in such a situation.

For our small home environments each run took just a few minutes. The robot followed a different path in each run. The laser data and the odometry from one of the runs are shown in Figure 3. For each type of conditions above we recorded two runs assuming that typically there will be one training run used to build and learn the environment representation and the second run can be used for evaluation. It is more challenging to use training and testing runs with different conditions.

B. Base level concepts and data annotation

We constrain ourselves here to a simple but still rich set of spatial concepts. An indoor environment is typically divided into rooms. The second common concept is that of 'object'. We selected a number of prominent objects from the environment. The person who annotated the data was supplied with a list of objects and the task was to segment the objects in the omnidirectional images taken from the robot. The task was also to decide when the robot entered each of the rooms based on the images. The structure of the XML annotation is given in Figure 4 and Figure 5 illustrates the annotation.

C. Toolbox

In order to allow the researchers a quick start we provide a Matlab toolbox with a set of functions for accessing the data; the data annotations; and the information about sensor calibration and their positioning on the robot. The provided functions can be divided into following groups:

Geometrical - a set of functions that can be used to perform various geometric transformations on the sensory data. For example a laser scan can be transformed to 3D world coordinates with respect to the robot and then these points can be projected to the omnidirectional camera image.

Annotation - a data annotation tool is provided visualizing and generating new annotations.

Demonstrations - a set of demonstration scripts that illustrate usage of the functions from the toolbox. The images from Figure 3 are the result of a demo script.

Finally the provided XML annotation can be transformed to the 'Label Me' object database format [21].

IV. Evaluation methods

Proper evaluation of a robot space representation that contains human cognitive spatial concepts would consist of extensive user studies where users would interact with the robot in various scenarios. However, it is realistic to start with some specific scenarios and evaluation criteria that might be of broad interest. We concentrate on the so called "home tour" scenario and identify the following type of evaluation criteria that can be performed using the presented data set.

### A. Object and location instance recognition

One could simulate a realistic learning situation by using the annotation from one run to learn the representation. The annotation from another run from the same environment can be used for the evaluation. The evaluation would consist of testing if the robot can recognize if it revisits a certain location, e.g. room. Recognition of location was reported on the basis of range data [20], [23] or on the basis of visual information [27]. A related topic is the 'scene recognition' in image retrieval applications [4]. The robot should also be able to recognize and localize the objects when it observes them again. Object recognition is well studied in the computer vision area and robust algorithms are available [16]. While the annotation can be used for evaluating standard algorithms for either recognizing locations or recognizing objects, we encourage algorithms combining these two [25] and also combining the information from different sensors.

### B. Object and location category recognition

The evaluation in the previous section assumes that the robot needs to consider all possible objects and locations in a new environment. On the other hand from a more advanced home robot we would expect that it already has knowledge about some concepts, for example recognizing a TV, a chair, a kitchen, a living room etc. Evaluating the object (location) category recognition involves learning the representations on one set of home environments and testing it on other previously unseen environments. Currently the data set contained just three home environments but this will be extended in the future. The object (location) category recognition algorithms can also be trained on other data sources such as other object databases [18], internet [5], etc, and tested on the presented data set.
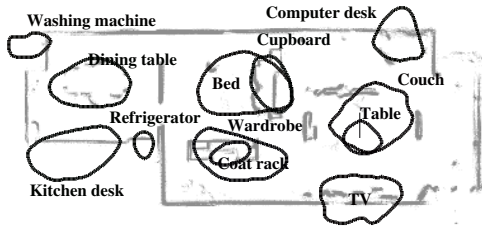
C. Object and location geometric properties

Traditional map building methods in robotics, i.e. SLAM, concentrate on geometric reconstruction which is important for robot navigation and/or the task of object grasping. The estimated geometric properties are useful for communicating with humans, e.g. the robot can decide if a room is elongated, one object is larger than the other object etc. We provide a base line geometric reconstruction as a starting point. First, the data set contains results of a SLAM algorithm based on laser range finder data and line features [10], see Figure 3b. Geometric reconstruction of the objects is more difficult. As a base line approach we provide a simple algorithm for estimating the 2D space occupied by the objects. The annotation from the omnidirectional images is back-projected onto a 2D grid and intersected, the resulting grid cells are the inferred visual hull of the object [19], see Figure 3c. We also provide the hand made drawings made by measuring various distances in the environment, see Figure 2. Note that the focus of data set are the semantic concepts and the data set is not intended to provide very precise geometric ground truth for evaluating SLAM approaches. Furthermore, home environments are usually not large and do not present a real challenge for current SLAM approaches.
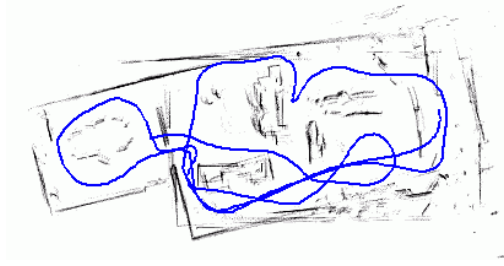
a) range scans and odometry
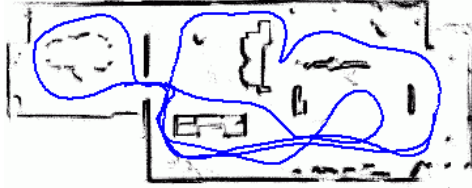
b) range scans and odometry aligned using SLAM procedure [10]

c) object visual hull reconstructed from annotations

Fig. 3.   Laser scan data and the extracted geometric properties, home 1.

```
<frame>
 <iFrame></iFrame><!--frame number-->
 <locationdescription><!--e.g.room name-->
 </locationdescription>
 <!--list of visible object-->
 <object>
   <name></name>
   <!--object segmentation by a polygon-->
     <polygon>
        <point2D><x></x><y></y></point2D>
     </polygon>
 </object>
</frame>
```

Fig. 4.   The basic structure of the XML annotation.

V. CONCLUSIONS

We described in this paper an annotated data set, its design principles and related practical issues. We also propose a set of evaluation criteria. We hope that the annotated data set will be useful for developing, testing and comparing algorithms for inferring human spatial concepts from sensory data. The data set simulates the home tour scenario currently in three different home environments. In our future work we aim to extend our data set to more different home environments as well to other types of environments, for example office environments. Furthermore, additional effort should be made in studying how to compare and evaluate different conceptual representations
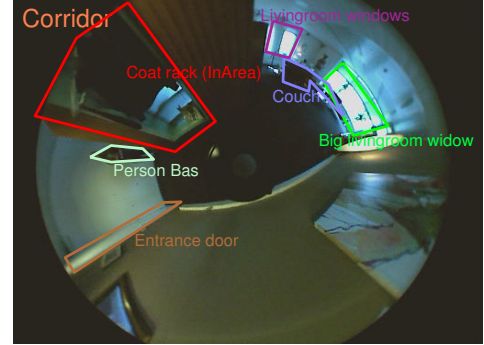

Fig. 5.   Example annotated omnidirectional image. The visible objects and persons are segmented using the polygonal lines. The current robot position is denoted as being in the corridor area.

through extensive users studies. Recently, some initial efforts have been reported in the EU FP6-002020 project Cogniron.

REFERENCES

[1] C. Breazeal. *Designing sociable robots*. MIT Press, Cambridge, 2002.
[2] H. Choset and K. Nagatani. Topological simultaneous localisation and mapping: Towards exact localisation without explicit localisation. *IEEE Trans. on Robotics and Automation*, 17(2):125–137, 2001.
[3] H. Durrant-Whyte and T. Bailey. Simultaneous localisation and mapping (SLAM): Part I - The Essential Algorithms. *Robotics and Automation Magazine*, June 2006.
[4] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.
[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proc. of the International Conference on Computer Vision*, 2005.
[6] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robots and Autonomous Systems*, 42:143–166, 2003.
[7] A. Green, H. Hüttenrauch, and E.A. Topp. Measuring up as an intelligent robot: On the use of high-fidelity simulations for human-robot interaction research. *In Proc. of the Workshop on Performance Metrics for Intelligent Systems (PerMIS)*, 2006.
[8] A. Howard and N. Roy. The robotics data set repository (Radish). *http://radish.sourceforge.net/*, 2003.
[9] Ulrich I. and Nourbakhsh I. Appearance-based place recognition for topological localization. *In Proc. of the IEEE Int. Conf. on Robotics and Automation*, pages 1023–1029, 2000.
[10] J.Folkesson, P.Jensfelt, and H.I.Christensen. Vision SLAM in the measurement subspace. *In Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2005.
[11] D. Kirsh. The intelligent use of space. *Artificial Intel.*, 73:31–68, 1995.
[12] G.-J.M. Kruijff, H. Zender, P. Jensfelt, and H.I. Christensen. Clarification dialogues in human-augmented mapping. *In Proc. of the ACM Conf. on Human-Robot Interaction*, 2006.
[13] B. Kuipers. The spatial semantic hierarchy. *Artificial Intel.*, 119:191–233, 2000.
[14] B.J. Kuipers and Y.T. Byun. A qualitative approach to robot exploration and map-learning. *In Proc. of the Workshop on Spatial Reasoning and Multi-Sensor Fusion*, 1987.

[15] T. Kyriakou, G. Bugmann, and S. Lauria. Vision-based urban navigation procedures for verbally instructed robots. *Robotics and Autonomous Systems*, 51:69–80, 2005.

[16] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004.

[17] T.P. McNamara. Mental representations of spatial relations. *Cognitive Psychology*, 18:87–121, 1986.

[18] J. Ponce and et. al. Dataset issues in object recognition. In *Toward Category-Level Object Recognition, LNCS, vol 4170*, pages 29–48. 2006.

[19] M. Potmesil. Generating octree models of 3d objects from their silhouettes in a sequence of images. *Comput. Vision Graph. Image Process.*, 40(1):1–29, 1987.

[20] A. Rottmann, O. Martinez Mozos, C. Stachniss, and W. Burgard. Semantic place classification of indoor environments with mobile robots using boosting. *In Proc. of the Nat. Conf. on Artificial Intelligence*, 2005.

[21] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *Submitted to Intl. J. Computer Vision*, 2005.

[22] B. J. Scholl and P. Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Science*, 4:299309, 2000.

[23] E.A. Topp and H.I. Christensen. Topological modelling for human augmented mapping. *In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2006.

[24] E.A. Topp, H. Huettenrauch, H.I. Christensen, and K. Severinson Eklundh. Bringing together human and robotic environment representations - a pilot study. *In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2006.

[25] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. *In Proc. of the Intl. Conf. on Computer Vision*, 2003.

[26] S. Vasudevan, S. Gachter, V.T. Nguyen, and R. Siegwart. Cognitive maps for mobile robots - an object based approach. *Robotics and Autonomous Systems Journal*, 55(3), 2007.

[27] Z. Zivkovic, B. Bakker, and B. Kröse. Hierarchical map building using visual landmarks and geometric constraints. *In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pages 7–12, 2005.

[28] Z. Zivkovic and O. Booij. How did we built our hyperbolic mirror omnidirectional camera - practical issues and basic geometry. *UvA technical report*, 2005.