Collaborative robots learning spatial language for picking and placing objects on a table

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT
A shared understanding of language will assist natural interactions between humans and artificial agents or robots undertaking collaborative tasks. An important domain for collaborative armed robots is interacting with humans and objects on a table, for example, picking, placing, or handing over a variety of objects. Such tasks combine object representation and movement planning in the geometric domain with abstract reasoning about symbolic spatial representations. This paper presents an initial study in which a human partner teaches the robot words for spatial relationships by providing exemplars and indicating where words may be used over the surface. This study demonstrates how robots can be taught the words required for these tasks in a quick and simple manner that allows the concepts to be generalizable over different surfaces, objects, and object placements.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous;

Author Keywords
Robot; Language; Symbol grounding; Tabletop

INTRODUCTION

Collaborative robots that take part in human-robot interaction need to be able to communicate with their human partners. What it means for people to interact in a natural way with robots is still an open question. There are many different ways that humans and robots can communicate and a variety of skills are required: both language and gesture are important for communication; words must be both understood and produced correctly; and new words must be able to be learned for new situations.

For collaborative armed robots, interacting with humans and objects on a table is an important domain. The skills required include being able to interact appropriately with the objects, recognising objects indicated by people, and understanding what should be done with the objects. For example, given a set of objects on a table, the robot should be able to place the objects in the locations indicated by their human partner. Such object manipulation tasks combine object representation and movement planning in the geometric domain with abstract reasoning about symbolic spatial representations.

Robots communicating and interacting with humans and objects has been the focus of research since the 1990s [7], and parts of this problem have been solved in previous work. Robots have interacted with other robots or humans to learn words for objects [13, 12, 1, 21, 15] and to ground spatial relations [8, 5, 9, 18]. While robots have grounded spatial concepts for object placement in previous work, limitations typically include elaborate concept representations together with either pre-built concepts [9] or extended learning [5]. The focus of the work presented in this paper is simple and minimal interactions together with simple internal representations for learning spatial relationships for objects on a table.

The goal of this work is for robots to learn concepts for spatial relationships from simple interactions, that are then able to be used for object selection and placement. The concepts should be generalisable over different surfaces, different objects, and different object placements. The study presented in this paper compares two learning styles, with the human teacher providing exemplars of words or indicating where words may be used over the surface. The spatial relationships used in this paper are to the left, right, front, back, and centre of the table, all from the robot's perspective. The results indicate that the robot can learn a set of spatial relationships quickly from exemplars and generalize their use to different surfaces. In future work, we will expand the set of terms and the robot will use these terms in more interactive studies.

BACKGROUND

The symbol grounding problem [6] has been widely discussed for many years in the area of artificial intelligence, with much focus on the philosophical nature of the problem. We consider the symbol grounding problem to be a practical problem of whether robots can use language appropriately, rather than a philosophical problem about whether robots can actually understand langauge. The current state of the problem is that it has been solved for many situations [20], but there are still open research question for other situations, for example, actual practical applications with robots using and learning new words.

ACM ISBN .
DOI:

Challenges for human-robot collaborative tasks are communication, joint action, and human-aware execution. The main drive behind developing robots with natural language capabilities is for robots to be able to interact with people who do not have special training. In other words, people should be able to interact with a robot in the same way that they are able to interact with another human. For successful communication, a common vocabulary between robot and human is required, as well as the ability to extend this vocabulary through learning [7]. Key requirements for a cognitive robot to collaborate with a human include geometric reasoning, situation assessment, knowledge models, dialog, task planning, and task execution [9]. The challenge for communications between humans and robots is that their internal representations may be different, so that different concepts may be chosen to refer to an object. Shared grounding needs to be a collaborative process, in which both gestures and feedback can improve true understanding of how successful communication has been [2]. In our work, we focus on situated language and the practical symbol grounding problem.

Robots have been learning language for many years, either from each other or from humans. Language games are structured communication that allows a group of agents to create their own set of symbols to communicate with each other [19]. One of the first instances of robots learning words socially from humans was AIBO learning words for a few objects [21]. Objects on a table may have different names for each unique object, or they may be referred to by different features, including colour, size, shape, and relative location. Robots have successfully been able to select one of a set of objects using such descriptions [15, 13]. Several studies have involved robots learning words to describe both the objects and the relationships between the objects [18, 3, 5]. Spatial concepts can be divided into within sight and out of sight, within reach and out of reach. Robots learning spatial concepts for object manipulation tasks can focus on the area that is within reach, for example, inside or outside of a container [8], or generalising about the spatial relationships between two objects of various shapes and sizes [10]. In contrast to these studies, we focus on non-contact spatial relationships as a means for identifying which object.

Within the tabletop domain, different perspectives can be used, typically the perspectives of the different agents of the conversation, for example from the perspective of the robot or the human partner [18, 14]. In our current study, we are only interested in the robot's perspective, although an extension to the model would allow for different perspectives to be considered. More complicated spatial relationships are also possible, including the relationship to abstract concepts such as a 'row of blocks' [11], or specifying an object through collaborative referring expressions, where one partner first refers to a group of objects and receives confirmation before making further specifications [4].

Representations and methods for learning concepts can be quite elaborate, for example predicting which object using multi-class logistic regression trained on extensive crowdsourced data [5] and naming and selecting objects with extensive built-in knowledge about objects and spatial relationships [9]. While more complicated representations are required for more complicated concepts, for example abstract concepts [11], in this paper we aim to show that they are not required for the set of basic tabletop spatial relationships. The representations and language learning interactions used in this paper have been based on simple representations used for toponyms, distances, and directions [16, 17] that allow one-shot learning, dynamic concept creation, and multi-modal concepts.

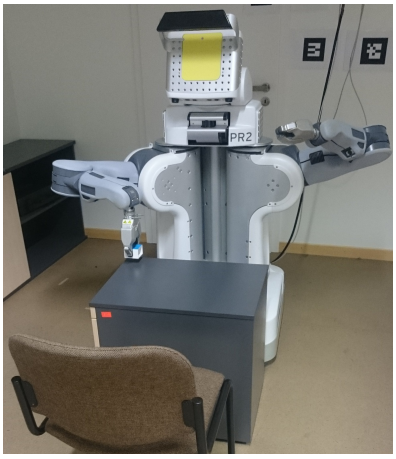

Figure 1. The environment consisting of robot, table, human, and objects

SYSTEM

This section provides a description of the system used, including the environment, the robot, the interactions, language representations, and lexicon evaluation.

Environment

The environment consists of the robot, a table, a human on the other side of the table, and a set of objects on the table (see Fig. 1). The objects used in this study are coloured rectangular lengths of plastic that are 40mm in height. The table used for training is a set of drawers 0.54m high, 0.44m deep, and 0.59m wide. The same table is used for testing, as well as a table that is 0.70m high, 0.8m deep, and 0.8m wide.

Robot system

A PR2 is used for the current study. A world model is built using the kinect and the head camera representing the colour, location, and dimensions of the objects and the table top. The robot can pick up the objects and place them at any position on the table that is within reach.

Interactions

The concepts learnt by the robot in this study are spatial relations relative to an area defined by a flat surface in front of the robot: front, back, left, right, and centre. The language learning interactions between the human and robot are based on exemplar or complete grid specification of word meanings. For an examplar specification, the human partner is asked to place the object at the best example of the training phrase. For example, "can you place the object at the left of the table?" For a grid specification, the human partner is asked to specify

which terms are appropriate for the current location of the object, with the set of locations being a grid covering the tabletop. That is, the human partner answers the question "where is the object?"

Lexicon evaluation

To test the lexicon of a robot, both word production and comprehension can be tested. In this domain, the tasks that can be used to test word production and comprehension include:

Object selection: Several objects are placed on the tabletop, the human uses language to specify an object, and the robot picks up the specified object.

Object placement: The human specifies a location using language, and the robot places an object at that location.

Word selection: Several objects are placed on the tabletop, the robot chooses an object and uses language to specify the object, and the human picks up the specified object.

The language of the robot can be tested on the surface used for training the robot as well as on tables with different dimensions.

Language representations

Words are learned using a distributed lexicon table with concept elements for the spatial relations specified on a 2-dimensional frame corresponding to the table surface (see [16] and [17] for a description of the distributed lexicon table for toponyms, distances, and directions). The 2-dimensional frame is projected onto the current tabletop, and can be stretched and skewed to match the surface. In a language interaction, the association between the concept element at the current location of the object in the 2-dimensional frame and the word used is increased.

The robot performs object selection, object placement, and word selection by calculating the confidence, c_{ij}, that a word, j, refers to a location, i:

$$c_{ij} = \frac{\sum_{k=1}^{X} \frac{a_{kj}(D-d_{ki})}{D}}{\sum_{m=1}^{X} a_{mi}} \quad (1)$$

where X is the number of concept elements associated with word j, D is the neighbourhood size, defined by the largest diagonal of the current surface, d_{ki} is the distance between concept element k and location i, and a_{kj} is the stored association between the concept element k and the word, j. The confidence is a value between 0.0 and 1.0, with 0.0 indicating that the word has not yet been used, and 1.0 indicating that the word has only been used for the current object location.

For word selection, the robot chooses the word for which the confidence at the object's location is highest. For object placement, the robot chooses the location on the table for which the confidence for the word is highest. For object selection, the robot chooses the object at the location for which the confidence for the word is highest. The result of the language representation used in this paper is a set of spatial relations for which the robot can determine areas for which the words are produced and comprehended.

EXPERIMENT

We conducted this experiment to investigate whether it is possible to quickly and easily teach a robot words for tabletop spatial relationships that can be used for object selection and placement, and that can be generalised across different situations.

Procedure

Prior to the experiment we explain to the participant that they will be teaching the robot the meaning of a set of phrases describing the spatial relationship of an object on the table in front of the robot, from the perspective of the robot.

In the learning phase of the experiment the participant teaches the robot what the words mean. There is one object on the table and the human is seated directly opposite from the robot. There are two conditions for the learning phase: exemplar and grid. Initially the participant provides the robot with exemplars of each of the phrases by placing the object in the location they believe best represents the phrase. In the second part of the learning phase the robot moves the object to 25 different locations and the participant indicates which phrases are relevant at each location. The information provided by the participant is used by the robot to construct two separate lexicons: one from the exemplars and one from the responses over the grid.

Two tasks are performed in the testing phase of the experiment: object selection and object placement. The tasks are undertaken using both lexicons first on the original table, and then on a table with different dimensions. For the object selection task, nine objects are placed on the table in a 3x3 grid pattern, and the robot is asked to point to the object specified by each of the phrases in turn. For the object placement task, one object is placed on the table, and the robot is asked to move the object to the location specified by each phrase in turn. The lexicons of the robot are also visually inspected, identifying areas in the framework associated with each phrase.

Participants

Five people (four male, one female, average age 32.2 years) participated in the experiment.

Results

The learning phase in the exemplar condition took an average of 29 seconds to complete, compared to an average of 3 minutes and 30 seconds for the grid condition.

The consistency across the lexicons is high for both object selection and object placement for all surfaces tested. A visual comparison of the lexicons indicates that all lexicons are very similar, and the combined lexicons for all participants is similar to each individual lexicon. As expected, the grid lexicons are smoother where the exemplar lexicons have a peak position (see Fig. 2 and 3).

Regarding the object selection task, the grid lexicons resulted in consistent object selection across all participants. The exemplar lexicons resulted in 17 out of 25 selections consistent

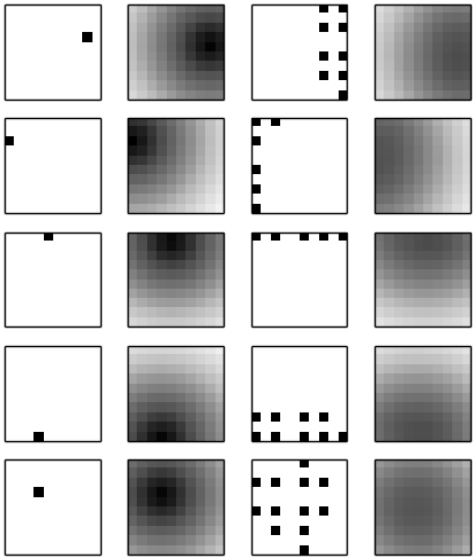

Figure 2. The robot's exemplar (left two columns) and grid lexicon (right two columns) after interacting with one of the participants, showing the raw lexicon and the confidence that each word is associated with each location within the 10 by 10 grid of the lexicon framework (black=1.0, white=0.0). The five rows show each word: left, right, front, back, centre. Note that the robot's position is at the top centre of the square facing the bottom of the page.

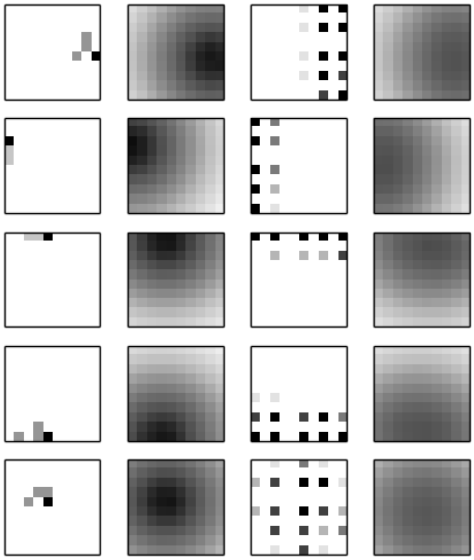

Figure 3. The robot's combined exemplar (left two columns) and grid lexicon (right two columns) for all participants, showing the raw lexicon and the confidence that each word is associated with each location within the 10 by 10 grid of the lexicon framework.

with the grid selection, with the other 8 selections being a neighboring object, for example, the right front object instead of the right centre object for 'right'.

Using the combined grid lexicon as the average placement for each of the five terms, we compared the actual placements for each of the words using each lexicon on the initial surface. The grid lexicons were again more consistent than the exemplar lexicons, with the average distance between the actual placement and the average being 0.119m for the exemplar lexicons (maximum 0.236m), and 0.045m for the grid lexicons (maximum 0.147m).

The results for other surfaces were qualitatively the same as for the original table. For object selection, the grid lexicons were consistent for 24 selections for the rotated table and for 23 selections for the larger square table, while the exemplar lexicons were consistent for 18 selections for the rotated table and 19 selections for the square table. For the placement task, the difference in position was the same relative to percentage of table size, as the framework is scaled to the size of the table.

DISCUSSION

The robot learnt the five terms quickly and successfully executed object selection and placement using the same and different tables. However, the current model is limited to robot's perspective, the surface is restricted to simple shapes within arms reach, and the number of objects that can be selected between is small. There are several extensions that could be made to address these limitations.

The 2-dimensional frame used in this study could be extended to have an orientation allowing the frame to be rotated prior to being used for language learning and use. The same frame could then be used to apply to 'your' perspective or 'my' perspective, and possibly even for certain intrinsic perspectives for objects

While stretching to a different shape is easy for rectangles of different dimensions, it is not as easy to stretch to tables that are different shapes. One approximation is to simply stretch to maximum width and length, and only consider positions on the table for object placement, but there may be other options that work better in certain situations.

The framework could be extended to apply to the areas surrounding objects, such that the spatial relations refer to the front, back, left, or right of the object instead of the surface. A further extension would allow spatial relations on the z axis, for example above and below an object. The model could then be combined with object descriptions to allow more complete description of more complicated sets of objects.

CONCLUSIONS

The study presented in this paper has shown that a robot can learn tabletop spatial relationships quickly, and that the words learned can be generalised to tables of different dimensions. While the grid lexicons have higher consistency, the exemplar lexicons are much quicker for the robot to learn and are potentially usable in many circumstances. We will undertake further studies to expand the set of terms and then use these terms in studies investigating interactions between humans, robots, and table-top objects.

ACKNOWLEDGMENTS

We thank the volunteers who participated in the study. We thank all current and previous members of the lab the development of the code base that made the study presented in this paper possible.

REFERENCES

1. Henny Admoni, Thomas Weng, and Brian Scassellati. 2016. Modeling communicative behaviors for object references in human-robot interaction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3352–3359.
2. Joyce Y Chai, Rui Fang, Changsong Liu, and Lanbo She. 2016. Collaborative Language Grounding Toward Situated Human-Robot Dialogue. *AI Magazine* 37, 4 (2016).
3. Haris Dindo and Daniele Zambuto. 2010. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 790–796.
4. Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 271–278.
5. Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, Trevor Darrell, and others. 2013. Grounding spatial relations for human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1640–1647.
6. Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3 (1990), 335–346.
7. Volker Klingspor, John Demiris, and Michael Kaiser. 1997. Human-robot communication and machine learning. *Applied Artificial Intelligence* 11, 7 (1997), 719–746.
8. Johannes Kulick, Marc Toussaint, Tobias Lang, and Manuel Lopes. 2013. Active Learning for Teaching a Robot Grounded Relational Symbols.. In *IJCAI*.
9. Séverin Lemaignan, Mathieu Warnier, E Akin Sisbot, Aurélie Clodic, and Rachid Alami. 2016. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence* (2016).
10. Oier Mees, Nichola Abdo, Mladen Mazuran, and Wolfram Burgard. 2017. Metric Learning for Generalizing Spatial Relations to New Objects. *arXiv preprint arXiv:1703.01946* (2017).
11. R Paul, J Arkin, N Roy, and TM Howard. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. *Proceedings of Robotics: Science and Systems (RSS), Ann Arbor, Michigan, USA* (2016).
12. Camilo Perez Quintero, Romeo Tatsambon, Mona Gridseth, and Martin Jägersand. 2015. Visual pointing gestures for bi-directional human robot interaction in a pick-and-place task. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 349–354.
13. Raquel Ros, Séverin Lemaignan, E Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann, and Felix Warneken. 2010. Which one? grounding the referent based on efficient human-robot interaction. In *RO-MAN, 2010 IEEE*. IEEE, 570–575.
14. Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. 2004. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34, 3 (2004), 1374–1383.
15. Deb K Roy. 2002. Learning visually grounded words and syntax for a scene description task. *Computer speech & language* 16, 3 (2002), 353–385.
16. Ruth Schulz, Gordon Wyeth, and Janet Wiles. 2011. Lingodroids: Socially grounding place names in privately grounded cognitive maps. *Adaptive Behavior* (2011), 1059712311421437.
17. Ruth Schulz, Gordon Wyeth, and Janet Wiles. 2012. Beyond here-and-now: extending shared physical experiences to shared conceptual experiences. *Adaptive Behavior* 20, 5 (2012), 360–387. DOI: **http://dx.doi.org/10.1177/1059712312449546**
18. Michael Spranger. 2016. The evolution of grounded spatial language. (2016).
19. Luc Steels. 2001. Language games for autonomous robots. *IEEE Intelligent systems* 16, 5 (2001), 16–22.
20. Luc Steels. 2008. The symbol grounding problem has been solved. so what's next. *Symbols and embodiment: Debates on meaning and cognition* (2008), 223–244.
21. Luc Steels and Frederic Kaplan. 2000. AIBO's first words: The social learning of language and meaning. *Evolution of communication* 4, 1 (2000), 3–32.