Affordance Based Disambiguation and Validation in Human-Robot Dialogue

Kim Wölfel and Dominik Henrich

University of Bayreuth,
Chair for Applied Computer Science III,
(Robotics and Embedded Systems)
Universitätsstraße 30, 95447 Bayreuth, Germany
kim.woelfel@uni-bayreuth.de,
dominik.henrich@uni-bayreuth.de,
<http://robotics.uni-bayreuth.de/>

Abstract. Speech-based robot instruction is a promising field in private households and in small and medium-sized enterprises, because it facilitates a comfortable way of communicating with robot systems, even while the users' hands are occupied. An essential problem in transforming the speech-based instructions into complex robot motions is the validation of the resulting motions regarding feasibility. This emerges from the fact that the user may not be fully aware of the capabilities of the robot including its gripper in a given working environment. We present an approach that tackles this problem by utilizing dynamic affordances for the disambiguation and elemental validation of speech-based instructions.

Keywords: Situated Dialogue, Affordances, Disambiguation

1 Introduction

One long term goal in current robotic research is the development of robot systems which have approximately the same cognitive, communicational, and handling abilities as humans. As part of this ongoing development, application domains for robot systems shall be expanded, from industrial settings with separated working cells, fixed object positions, and preprogrammed motions towards a flexible usage in small or medium-sized enterprises or private households.

An intuitive way of communicating with such a robot system is to instruct it via spoken commands. Advantages of this approach are that it facilitates the instruction of a robot without having to learn a robot programming language and an instruction while the users' hands are occupied. In user studies we noticed that ambiguities are a problem that occurs while instructing a robot and that in most cases ambiguities result from differences in the mental representation of the user and the robot system. For example, users omit parameters in an instruction, which seem *obvious* from their perspective. Informing the user about missing or incorrect parameters can only be performed well if a common ground is present. One way of doing this is by utilizing affordances which can be seen as

a functionality of objects, e.g. graspability or liftability. In the case of force based object manipulation this is a helpful tool, because affordances of the workpiece, the robot arm, and the tool in use can be evaluated for a given instruction.

The main contribution of this paper is a concept that is able to disambiguate and validate instructions based on dynamic affordances regarding the robot, its tool, and the workpiece. The scientific gain of this approach is bipartite: First, affordances are used in a preprocessing step to reduce the parameter space and even empowers a system to execute motions emerging from incomplete commands. Second, by using dynamic affordances the feasibility of instructions can be applied in changing working environments and for various hardware. All this is done to investigate the application of affordances for previously unknown tasks, instead of relying on plans or pre-trained models. Moreover, utilizing affordances facilitates a communication with the user on a common ground, which is helpful in the case of sensor-based robot motions.

In Section 2, an affordance overview is given, highlighting approaches regarding affordances, disambiguation, and a combination of these. In Section 3, the approach is described in detail. Section 4 contains the description and results from our user study verifying the validity and usefulness of our approach and in Section 5 a summary of the approach is given, as well as a glance at future work.

2 Related Work

In this chapter, approaches are presented that deal with the definition and use of affordances, disambiguation of spoken instructions in human-robot dialogue, and concepts combining these areas. We also describe in what way our approach differs from the state of the art.

Experts and non-experts may not always be aware of the capabilities of a robot system while instructing it. This holds true especially in force-based tasks, like peg-in-hole or cutting tasks, where force parameters have to be considered as well. A well studied approach for defining and evaluating these functionalities are affordances introduced by [6], defining what can or cannot be done with an object [20]. The concept of affordances has been part of the scientific community for several decades. It leads to a lot of publications surveyed regarding psychology, neuroscience and robotics [11], human-robot interaction [21], and developmental robotics [19]. Being dissatisfied from the variety of formalizations, a computational affordance model is given in [27] which classifies approaches regarding perception (selective attention, temporality, order, level, perspective), structure (chaining, competitive, abstraction), and development (exploitation, generalization, prediction, learning, acquisition). Affordance learning has also been studied for a lot of applications, for example grasping [22], [5], traversability [2], [12], [24], or tool-use [1], [28].

Disambiguation has also been discussed a lot along the last decades [15], [14], [17], [25]. Most approaches try to solve ambiguities by using clarification questions based on yes/no questions [4], [9], listing all options [16], or generic

Fig. 1. Concept overview: Blue arrows highlight the main information flow and grey arrows the control flow between the logics and the hardware/data base .

WH-questions [23]. In search for a general approach for disambiguation, [25] suggests three design recommendations. Robots should present all options only if the number of options is less than three, otherwise robot systems should generate yes/no or WH-questions (D1). Clarification requests should be formulated in a way, that the user thinks the robot understood the goal of the instruction (D2). Robots should use pragmatically appropriate phrasings (D3).

Investigating the combination of affordances and disambiguation in the field of speech-based interaction has been poorly investigated by just a few approaches [20], [8], [10], [3]. In [10] the application of affordances for co-reference resolution in speech based commands was mentioned, but not further investigated. In [8] ambiguous commands were solved utilizing the affordance concept together with a prediction based on previous commands. In [3] a neural network learned with a huge text corpora was used to disambiguate incomplete instructions.

In difference to the introduced approaches, we present a system which applies the concept of affordances in form of pre-conditions [13] for disambiguating and validating spoken instructions without relying on previous recorded data or text corpora, but solely on dynamic affordances, which means that their state can change over time or depending on the hardware set-up. For example, the *reachability* of an object can change during a manipulation task (hidden behind other work pieces) or not reachable by the robot arm currently installed.

3 System Overview

This section describes the main aspects of our framework. After giving a short framework overview (Section 3.1), we describe the modules in more detail, which are part of the main information flow. Besides explaining the dialogue module (Section 3.2), we present our disambiguation approach implemented in the identification and validation module (Section 3.3 and Section 3.4), and explain how and what kind of feedback is generated by the system (Section 3.5).

3.1 Framework Overview

Our framework for **SP**eech based **I**nstruction of **RO**bot systems (SPIRO) is divided in static modules: dialogue-manager (DM), identification module, validation module, feedback module, and changeable modules: The active robot (R), the attached tool (T), and the knowledge data base (DB) (see Figure 1).

The information flow starts by a spoken user instruction which is recognized and semantically enriched by the DM. This information is then passed to the identification module, which has the purpose to identify the motion contained in the instruction along with the necessary parameters based on attached affordances. The validation module then checks the given input regarding the affordances in the current world state. The validated information is then passed to the feedback module which passes the appropriate motion data to the hardware and the acknowledgement to the DM. If an ambiguous or, in terms of the domain, incorrect instruction is recognized by the DM, identification or validation module, the problematic affordance is tagged as non valid or ambiguous and an appropriate clarification question is generated by the DM. In this case, no robot motion is generated.

To allow a flexible use of our system we designed it in a way, that the robot, tool, and database can be exchanged easily. Being aware that switching the robot arm does not happen as often as switching the tool or underlying database we nevertheless included it to be also able to cover even rare cases.

3.2 Dialogue Manager

The DM is implemented using the ontology based dialogue platform *Semvox*¹, because it offers a fast design and adjustment of dialogue system, combined with the *Nuance*² framework, because of its valuable speech-to-text and speech-recognition features. The DM has four purposes: Automatic speech recognition (ASR), semantic role labeling (SRL) [18] of incoming user instructions based on pre-defined ontologies, i.e. labeling *red* as a colour or *shove* as a verb, generating verbal phrases or starting a dialogue based on data provided by the feedback module, and returning system answers in an acoustic way (TTS). Here, an ontology is an "is a"-hierarchy which allows the user to add semantic knowledge to the spoken instructions.

The ASR module converts the spoken user instruction into word set hypothesis. The hypothesis with the highest probability is then provided to the system for further processing. For example, an instruction like: "Place the cup on the plate!" is transformed into an instruction tuple

$$I = (\text{verb}(\text{Place}), \text{workpiece}(\text{cup}), \text{tool}(\text{gripper}), \text{destination}(\text{plate})).$$

This SRL is done automatically by Semvox as long as semantic roles like *verb*, *tool*, etc. are pre defined. The generated instruction tuple is then sent to the

¹ <https://www.semvox.de/en/technologies/odp-s3-the-basis-for-interactive-systems/>

² <https://www.semvox.de/en/technologies/nuance-speech-recognition-tts/>

Fig. 2. Validation Module validating incoming instructions based on affordances. If an ambiguity is perceived each task directly sends the information to the feedback module.

identification module and the validation module to compute the appropriate answer to the instruction.

The other purpose is to generate verbal feedback for the user in one of the four cases: Each parameter of the instruction was specified correctly (C1), the system estimated a disambiguation of an incomplete or ambiguous instruction (C2), the system was not able to disambiguate parameters of the instruction (C3), the system recognized incorrect instruction parameters (C4).

In case C1 an acknowledgement like "Yes!" or "Yes, of course!" is sent to the user. In case C2 the disambiguated instruction is sent to the user with the opportunity to interrupt the system by saying "No!" or "Stop!" and the motion is executed by the robot. In case C3, a clarification question based on the ambiguous parameters is generated and in case C4 the user is informed about the false parameters.

3.3 Identification Module

This module has the purpose to identify the motion, i.e. finding an appropriate motion template in the DB, and accompanied parameters intended by the user instruction. Given the instruction tuple $\mathcal{I}$ recognized by the DM, the suitable motion template is searched and filled using the *affordance filtering* approach [7] performed by the two sub modules: Motion identification (T1) and parameter identification (T2) (see Figure 2).

In T1 the intended motion has to be identified. Here, two cases arise. Case 1: The verb is mentioned. In this case, the system directly picks the appropriate motion templates for this verb from the DB and continues with T2. Case 2: No verb is mentioned. In this case, we consider two possibilities: The user implicitly uses the verb from the previous instruction or he explicitly did not mention the verb because its obvious for him. The implicit case can easily be solved with Semvox, which allows an auto-filling of the verb slot if the current instruction has the same structure as the previous instruction, which then automatically leads us to Case 1. The explicit case leads to a work around. The system then has a look at the other specified parameters and tries to find a verb which maps to these parameters, both syntactically and affordance-based. If this still

does not lead to a single verb candidate, the verb slot is filled with the possible candidates, tagged as ambiguous and the augmented instruction tuple is sent to the validation module. Imagine the user just saying: "The big screw!". Instead of not answering at all or responding: "What do you want me to do with it?", we have a look at the valid affordance at this time step and try to disambiguate the opportunities for the verb slot. If this is not possible, we tag the verb slot as ambiguous.

After the motion is identified, the remaining parameters are filled using the motion template provided by the DB. In the simple case all the parameters are defined and we send the augmented instruction tuple to the validation module. If one of the parameters is ambiguous, a disambiguation is necessary. Here again, we are using affordances to filter ambiguous or even generate missing parameters. Imagine the instruction: "Shove to the left!" Normally, we would have to ask the user which object he wants us to shove to the left, but maybe there is only one object in our workspace that is shove able. If there is more than one option left, we tag the affected parameter slots as ambiguous before sending the augmented instruction tuple to the validation module.

3.4 Validation Module

The purpose of the validation module is to evaluate the instruction parameters by evaluating the affordances for the given world state and tag them with the description: VALID, INVALID, AMBIGUOUS, and DISAMBIGUATED. Using dynamic affordances is necessary, because in some cases valid defined instructions may not be executable, because the object is out of reach for the robot arm or the gripper is not suitable for grasping the workpiece. A parameter is tagged as VALID, if it is distinct regarding the necessary affordance and has the correct type regarding the verb frame. A parameter is tagged INVALID, if it does not have the correct type regarding the verb frame. A parameter is tagged AMBIGUOUS if a set of possible real world objects is possible for it and they are of the right type. A parameter is tagged DISAMBIGUATED if the identification and validation module managed to disambiguate the user instruction. The tagged instructions are then sent to the feedback module for further investigations.

3.5 Feedback Module

The feedback module is responsible for transforming validated instructions into either a text-based response for the DM or a text-based response for the DM in combination with robot motion parameters for the motion generator. What kind of feedback is given depends on the validity of the parameters slots of the instruction. In the case that the validity slot of all parameters holds VALID the feedback is given in form of an acknowledgement phrase ("Ok!", "Sure!", "Yes!") sent to the DM together with motion parameters sent to the motion generator, which generates a hybrid motion based on the approach presented in [26]. If the validity slots hold DISAMBIGUATED as well as VALID, the instructions are sent to the DM for a spoken response before the motion is executed. If the validity slot

Fig. 3. 1: Example working environment for instructing a robot arm. 2 - 4: Intermediate state of the key holder. 5: Finalized key holder.

of at least one parameter of an instruction holds I, the user is told to correct the specific parameter and no robot motion is generated. If one of the given parameter holds AMBIGUOUS, a clarification question based on D1 - D3, defined in section 2, is generated and send to the DM. This process is repeated for one property at a time until the presented instruction is valid regarding all occurring parameters.

4 User Study

In this section, a user study is presented which was conducted to gain information on how users react to incomplete and ambiguous instructions and why they choose their reaction. After a presentation of the user study setup (Section 4.1), the results are evaluated as well as discussed (Section 4.2).

4.1 Setup

The study was set up using the online tool *Survey Monkey*³ and was divided in two main parts.

In the first part, we wanted to gain information about the necessity of a validation component and presented the participants the working environment shown in Figure 3, 1, together with the request to instruct the shown robot to place a motor with a given weight of 4 kg from location P1 to location P2. This was asked to see, if users are aware of the capabilities of lightweight robot arms and at the same time if a validation or correction of user instructions is necessary in cases where a robot is not able to perform an action.

In the second part, the participants were put in the position of an assistant robot, which helps its co-worker to build a key holder (see Figure 3, 2 - 5). The tasks for the participants where chosen in that manner, because they are either better executable by a robot due its precision or because the user typically does

³ <https://www.surveymonkey.de/>

Table 1. Votes for the different reaction types (RT1 - RT6) for the tasks marking, inserting, and cleaning in the second part of the user study.

task	RT1	RT2	RT3	RT4	RT5	RT6
marking	2	3	1	7	3	0
inserting	3	0	1	9	3	0
cleaning	2	1	1	4	7	1

not want to perform the task. For each instructed task, the participants had an image of the current state of the key holder, an instruction of his co-worker, and possible tools or objects to perform the tasks. They were then asked to choose between a set of reaction types: spoken disambiguation of three candidates and no motion generation (RT1), spoken disambiguation of two candidates and no motion execution (RT2), spoken query for the missing/ambiguous parameter and no motion execution (RT3), spoken feedback in form of the disambiguated instruction and motion execution (RT4), Spoken acknowledgement and motion execution (RT5), no spoken feedback and motion execution (RT6).

The first instruction is: "Mark the board with pattern M1!", the information what M1 looks like is given as well as the tools: pencil, text marker and an permanent marker. The participants are then told that their co-worker inserted holes at the marked spots and that the new instruction is: "Insert the hooks into the holes!", followed by a set of hooks with different lengths. After that the participants are told that their co-worker screwed in the hooks, took the key holder away, leaving a dirty workplace with the comment: "Clean the workplace!". They are also given a set of tools suitable for cleaning a workplace: sponge, cloth and hand brush.

4.2 Results

Overall 15 participants joined the first part of the study. Although experts as well as non-experts were among the participants, all of them instructed the robot to move the motor from place P1 to place P2 without considering if the robot arm is even capable of doing it. This result clearly shows, that a validation even in simple tasks, like moving an object, is necessary. Without a validation, this would normally lead to an incorrect execution of the motion resulting in frustration of the user or even damage. Using our affordance based validation, we are not only able to perceive this error, but also to formulate a spoken solution to this problem, in this case: "I can not lift the motor since it is too heavy!".

Overall 16 participants joined the second part of the study. The results are shown in Table 1. In the marking task most of the participants (43%) favoured RT4, especially the use of the pencil. One participant even wrote, that mentioning the instruction he understood, enables his co-worker to intervene, if he is not pleased with the disambiguation. In the inserting task, even more (56%) participants chose the reaction RT4. Here, two participants justified their reactions by

saying that the user is able to interrupt the motion, if he is not pleased with it. The remaining participants also mentioned that they made their decision based on the physical properties of the hooks and the board, thus, in a way our system would work as well. In the cleaning task, the majority of the participants chose RT5 as their favourite reaction and RT4 only got the second place. Based on the participants' justifications, the brush is in this case the best tool, or at least as good as the cloth, but no participant mentioned, that either the sponge or the cloth would be the best tool. This leads us to the conclusion that a listing of opportunities is only useful, if the evaluation of affordances leads to a similar rating.

5 Conclusion

We present a concept for an affordance based disambiguation and validation of incomplete, incorrect, and ambiguous spoken robot instructions. By utilizing dynamic affordances the system is able to reduce the amount of candidates in the case of ambiguous instructions or even solve the ambiguity. Moreover, using dynamic affordances allows not only validating semantically correct instructions, but also contains the information of the problems' source on a human understandable level. Limits of our approach are the common problem of not being able to reduce the solution space and the complexity of computing the dynamic affordances, while the latter one is still part of current research.

To evaluate our concept, we also conducted a user study where we collected information on how and why users react to incomplete and ambiguous reaction, and if they are aware of robot capabilities even in simple tasks. On the one hand, the results correspond to the presented concept. On the other hand, we also found a way to improve our concept by using acknowledgements if the disambiguated parameters are rated high enough.

In the future, we will implement the whole concept and perform further user studies to be able to improve our dialogue system.

Acknowledgements

This work has partly been supported by Deutsche Forschungsgemeinschaft (DFG) under grant agreement He2696-18.

References

1. Abelha, P., et al.: A model-based approach to finding substitute tools in 3d vision data. In: International Conference on Robotics and Automation (ICRA) (2016)
2. Carvalho, J.T., Nolfi, S.: Behavioural plasticity in evolving robots. *Theory in Biosciences* **135**(4) (2016)
3. Chen, H., et al.: Enabling robots to understand incomplete natural language instructions using commonsense reasoning. *arXiv:1904.12907* (2019)
4. Deits, R., et al.: Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction* **2**(2) (2013)

5. Detry, R., et al.: Learning grasp affordance densities. *Paladyn, Journal of Behavioral Robotics* **2**(1) (2011)
6. Gibson, J.J.: The theory of affordances. *Hilldale, USA* **1**(2) (1977)
7. Gorniak, P., Roy, D.: Situated language understanding as filtering perceived affordances. *Cognitive science* **31**(2) (2007)
8. Heikkilä, S.S., Halme, A., Schiele, A.: Affordance-based indirect task communication for astronaut-robot cooperation. *Journal of Field Robotics* **29**(4) (2012)
9. Hemachandra, S., Walter, M.R., Teller, S.J.: Information theoretic question asking to improve spatial semantic representations. In: *AAAI Fall Symposia* (2014)
10. Horton, T.E., Chakraborty, A., Amant, R.S.: Affordances for robots: a brief survey. *AVANT. Pismo Awangardny Filozoficzno-Naukowy* **2** (2012)
11. Jamone, L., et al.: Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems* **10**(1) (2016)
12. Kostavelis, I., et al.: Collision risk assessment for autonomous robots by offline traversability learning. *Robotics and Autonomous Systems* (2012)
13. Kroemer, O., Niekum, S., Konidaris, G.: A review of robot learning for manipulation: Challenges, representations, and algorithms. *arXiv:1907.03146* (2019)
14. Lemaignan, S., et al.: Artificial cognition for social human-robot interaction: An implementation. *Artificial Intelligence* (2017)
15. Liu, C., Walker, J., Chai, J.Y.: Ambiguities in spatial language understanding in situated human robot dialogue. In: *2010 AAAI Fall Symposium Series* (2010)
16. Marge, M., Rudnicky, A.: Miscommunication recovery in physically situated dialogue. In: *Meeting of the Special Interest Group on Discourse and Dialogue* (2015)
17. Marge, M., Rudnicky, A.I.: Miscommunication detection and recovery in situated human-robot dialogue. *Transactions on Interactive Intelligent Systems* (2019)
18. Márquez, L., Carreras, X., Litkowski, K.C., Stevenson, S.: Semantic role labeling: an introduction to the special issue (2008)
19. Min, H., Luo, R., Zhu, J., Bi, S., et al.: Affordance research in developmental robotics: A survey. *Transactions on Cognitive and Developmental Systems* (2016)
20. Min, H., et al.: Affordance learning and inference based on vision-speech association in human-robot interactions. *International Conference on Robotics and Biomimetics (ROBIO)* (2017)
21. Moratz, R., Tenbrink, T.: Affordance-based human-robot interaction. In: *Towards Affordance-Based Robot Control*. Springer (2008)
22. Song, H.O., Fritz, M., Goehring, D., Darrell, T.: Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering* **13**(2) (2015)
23. Tellex, S., Thakur, P., Deits, R., Simeonov, D., Kollar, T., Royl, N.: Toward information theoretic human-robot dialog. *Robotics* (2013)
24. Ugur, E., Oztog, E., Sahin, E.: Goal emulation and planning in perceptual space using learned affordances. *Robotics and Autonomous Systems* **59**(7-8) (2011)
25. Williams, T., Yazdani, F., Suresh, P., Scheutz, M., Beetz, M.: Dempster-shafer theoretic resolution of referential ambiguity. *Autonomous Robots* **43**(2) (2019)
26. Wölfel, K., Henrich, D.: Grounding verbs for tool-dependent, sensor-based robot tasks. In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE (2018)
27. Zech, P., et al.: Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior* (2017)
28. Zhu, Y., Zhao, Y., Chun Zhu, S.: Understanding tools: Task-oriented object modeling, learning and recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.