Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues

Evan Krause[1] Michael Zillich[2] Tom Williams[1] Matthias Scheutz[1]
[1]Human-Robot Interaction Laboratory, Tufts University, 200 Boston Avenue, Medford, MA, USA
[2]Inst. For Automation and Control, Technical University Vienna, Gusshausstr 27-29/E376 1040 Vienna, Austria

Abstract

Being able to quickly and naturally teach robots new knowledge is critical for many open-world human-robot interaction scenarios. We present a novel approach to using natural language context for one-shot learning of visual objects, where the robot is immediately able to recognize described objects. We describe the architectural components involved and demonstrate the proposed approach on a robotic platform in a proof-of-concept evaluation.

Introduction

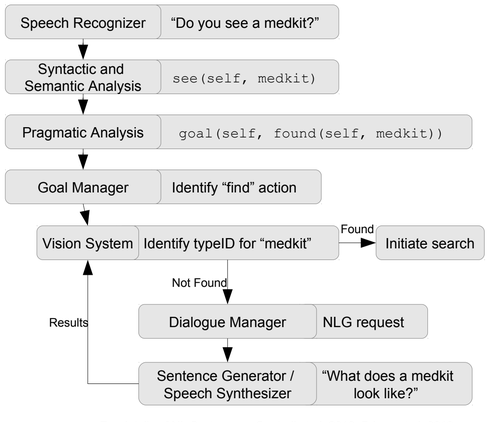
Two core assumptions of data-driven methods are that (1) data sets are available for training and (2) training is not time-critical.

Neither assumption is typically met in "open-world" scenarios where robots must quickly acquire new knowledge during task performance. Data-driven methods must thus be augmented with methods such as one-shot learning, that allow for on-line learning from only a few exemplars

Most approaches to one-short object learning are very limited with respect to allowable teaching inputs and often require multiple trials.

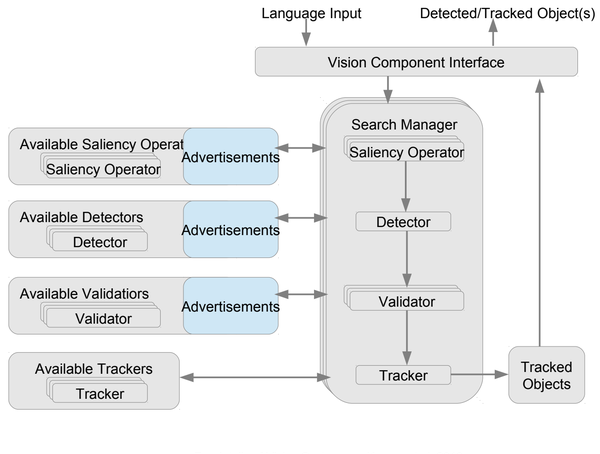
This research builds on the successes of recent integrated approaches and shows how (1) deep interactions between vision and natural language processing algorithms and (2) exploitation of structured representations can enable one-shot object learning from sufficiently specified natural language descriptions, thus allowing the robot to recognize the object in its environment in different contexts and poses.

NL Pipeline and Overall Architecture



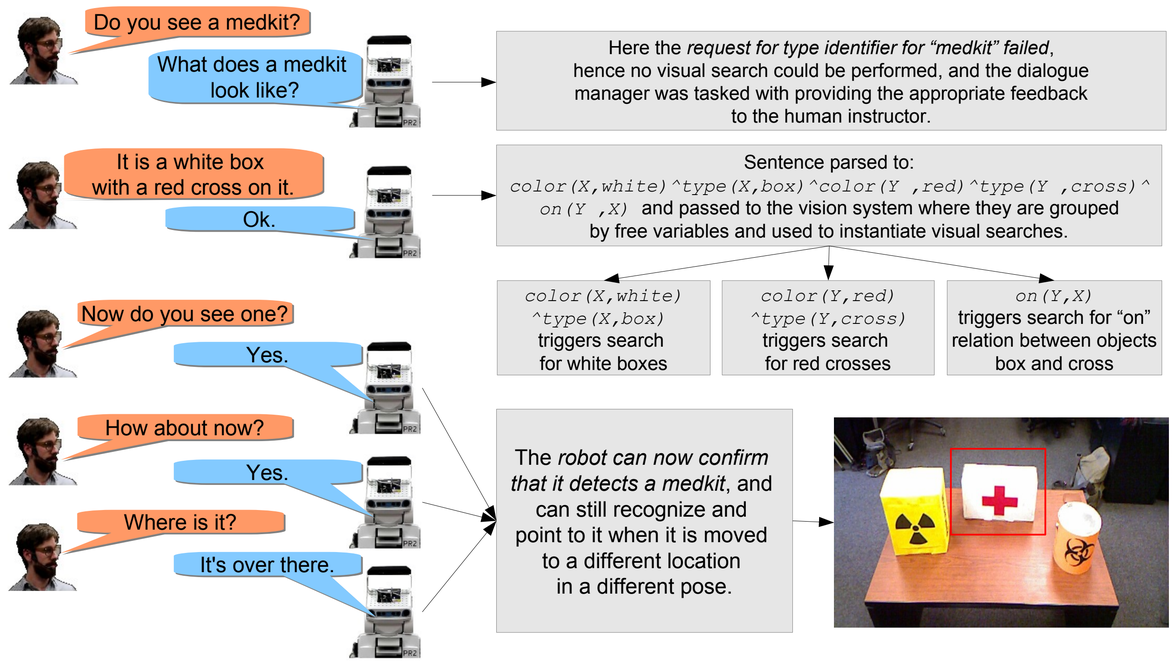
For details of NL System see Cantrell et al. 2010, Briggs et al. 2013

Vision System



For details of Vision System see Krause et al. 2013

Validation Scenario



Requirements for Language-Guided Visual One-Shot Learning

The vision system must be able to map structured linguistic descriptions to hierarchical object descriptions that capture the types and relationships of referenced object parts. These mappings can then be used to build representations of previously unknown objects which can then be recognized so long as the vision system is able to recognize its constituent parts and the described relations between those parts.

The vision system must be able to handle descriptions of 3-D properties of objects, 2-D surface patterns and textures, nouns referring to atomic and complex object types, adjective descriptions of various object characteristics, and spatial relations.

We ultimately aim to handle descriptions of objects that involve complex embedded clauses and descriptions that stretch over multiple sentences, but currently restrict ourselves to simple descriptions encapsulated in individual utterances.

Conclusion

Data-driven learning methods must be complemented by one-shot learning methods to meet the demands of future human-robot interaction scenarios. We introduced a method for one-shot language-guided visual object learning that requires deep integration of natural language and vision processing algorithms, and demonstrated the approach on a robot in a simple human-robot dialogue. Future work will extend the current system to allow the robot to maximally exploit the information present in both the linguistic and visual stimuli, in order to perform one-shot learning of shapes, textures, and spatial relations.

References

Briggs, G., and Scheutz, M. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. AAAI'13
Cantrell, R.; Scheutz, M.; Schermerhorn, P.; and Wu, X. 2010. Robust spoken instruction understanding for HRI. HRI '10.
Krause, E.; Cantrell, R.; Potapova, E.; Zillich, M.; and Scheutz, M. 2013. Incrementally Biasing Visual Search Using Natural Language Input. AAMAS'13.

Acknowledgments

This work was in part supported by US NSF grant IIS-1111323, ONR grants #N00014-11-1-0289 and #N00014-14-1-0149 to the last author and EU FP7 grants #600623 and #610532 and Austrian Science Foundation grant TRP 139-N23 to the second author.