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 (1) exploitation of language descriptions, thus allowing the robot to object learning from sufficiently specified natural processing algorithms and (2) exploitation of knowledge during task performance. Data-driven scenarios where robots must quickly acquire new inputs and often require multiple trials.

Most approaches to one-shot object learning are from only a few exemplars methods must thus be augmented with methods such as (1) data sets are available for training and (2) training is not time-critical.

This research builds on the successes of recent 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. 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.

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.

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


Acknowledgments

This work was in part supported by US NSF grant 11111323, ONR grants N00014-11-1-0289 and N00014-14-1-0149 to the last author and EU FP7 grants 600623 and 601052 and Austrian Science Foundation grant TRP 139-N23 to the second author.