Robotics and Autonomous Systems 62 (2014) 632–645

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Grounding semantic categories in behavioral interactions: Experiments with 100 objects

Jivko Sinapov *, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, Alexander Stoytchev

Developmental Robotics Laboratory, Iowa State University, Ames, IA, United States

ARTICLE INFO

Article history:
Available online 9 November 2012

Keywords:
Semantic perception
Active and interactive perception
Category recognition
Learning and adaptive system
Behavior-based robotics
Developmental robotics

ABSTRACT

From an early stage in their development, human infants show a profound drive to explore the objects around them. Research in psychology has shown that this exploration is fundamental for learning the names of objects and object categories. To address this problem in robotics, this paper presents a behavior-grounded approach that enables a robot to recognize the semantic labels of objects using its own behavioral interaction with them. To test this method, our robot interacted with 100 different objects grouped according to 20 different object categories. The robot performed 10 different behaviors on them, while using three sensory modalities (vision, proprioception and audio) to detect any perceptual changes. The results show that the robot was able to use multiple sensorimotor contexts in order to recognize a large number of object categories. Furthermore, the category recognition model presented in this paper was able to identify sensorimotor contexts that can be used to detect specific categories. Most importantly, the robot's model was able to reduce exploration time by half by dynamically selecting which exploratory behavior should be applied next when classifying a novel object.

1. Introduction

Object categories are all around us—our homes and offices contain a vast multitude of objects that can be organized according to a diverse set of criteria ranging from form to function. A robot operating in human environments would undoubtedly have to assign category labels to novel objects because it is simply infeasible to preprogram it with knowledge about every individual object that it might encounter. For example, to clean a kitchen table, a robot has to recognize semantic object category labels such as silverware, dish, or trash before performing an appropriate action.

The ability to learn and utilize object category memberships is an important aspect of human intelligence and has been extensively studied in psychology [1]. A large number of experimental and observational studies have revealed that object category learning is also linked to our ability to acquire words [2,3]. Researchers have postulated that, with a few labeled examples, humans at various stages of development are able to identify common features that define category memberships as well as distinctive features that relate members and non-members of a target category [4,5]. Other lines of research have highlighted the importance of object exploration [6,7], which is important for learning object categories since many object properties cannot always be detected by passive observation [8,9].

Recently, several research groups have started to explore how robots can learn object category labels that can be generalized to novel objects [10–14]. Most studies have examined the problem exclusively in the visual domain or have used a relatively small number of objects and categories. To address these limitations, this paper proposes an approach to object categorization that enables a robot to acquire a large number of category labels from a large set of objects. This is achieved with the use of multiple behavioral interactions and multiple sensory modalities. To test our method, the robot in our experiment (see Fig. 1) explored 100 different objects classified into 20 distinct object categories using 10 different interactions (e.g., grasp, lift, tap, etc.) making this one of the largest object sets that a robot has physically interacted with.

Using features extracted from the visual, auditory, and proprioceptive sensory modalities, coupled with a machine learning classifier, the robot was able to achieve high recognition rates on a variety of household object categories (e.g., balls, cups, pop cans, etc.). The robot's model was also able to identify which sensory modalities and behaviors are best for recognizing each category label. In addition, the robot was able to actively select the exploratory behavior that it should try next when classifying an object, which resulted in faster convergence of the model's accuracy rates when compared to random behavior selection. Finally, the model was evaluated on whether it can detect if a novel object does not belong to any of the categories present in the robot's training set.

* Corresponding author.
E-mail addresses: jsinapov@iastate.edu (J. Sinapov), cschenck@iastate.edu (C. Schenck), kerrick@iastate.edu (K. Staley), sukhoy@iastate.edu (V. Sukhoy), alexs@iastate.edu (A. Stoytchev).

0921-8890/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2012.10.007

Fig. 1. The humanoid robot used in our experiments, along with the 100 objects that it explored.

2. Related work

Most object categorization methods in robotics fall into one of two broad categories: (1) unsupervised methods, in which objects are categorized using unsupervised machine learning algorithms (e.g., *k*-Means, Hierarchical Clustering, etc.) and (2) supervised methods, in which a labeled set of objects is used to train a recognition model that can label new data points. Several lines of research have demonstrated methods that enable robots to autonomously form internal object categories based on direct interaction with objects [15,11,16,17]. For example, Griffith et al. [11] showed how a robot can use the frequencies with which certain events occur in order to distinguish between container and non-container objects in an unsupervised manner. Dag et al. [16] and Sinapov and Stoytchev [18] have also shown that robots can categorize and relate objects based on the type of effects that they produce when an action is performed on them.

In contrast, the focus of this paper is on supervised methods for object categorization, which attempt to establish a direct mapping between the robot's object representation and human-provided semantic category labels. A wide variety of computer vision methods have been developed that attempt to solve the problem using visual image features coupled with machine learning classifiers [19–21]. Several such methods have been developed for use by robots, almost all exclusively working in the visual domain [22,23,12,24,14,25]. One advantage of visual object classifiers is that they can often be trained offline on large image datasets. Nevertheless, they cannot capture object properties that cannot always be perceived through vision alone (e.g., object compliance, object material, etc.). In other words, disembodied object category representations that are grounded solely in visual input cannot be used to capture object properties that require active interaction with an object. Thus, even the best visual classifier is guaranteed to fail on certain object classification tasks. For example, Lai et al. [26] report that using state-of-the-art RGB and depth features for classifying 300 objects into 51 categories results in 85.4% accuracy, which demonstrates that there is still a lot of information about object categories that cannot be captured using disembodied vision-based systems. Furthermore, it has been argued that embodied perception is not only desirable, but also required for achieving intelligent autonomous behavior by a robotic system [27]. Therefore, to address the limitation of disembodied systems, our robot grounded the semantic category labels of objects in its own sensorimotor experience with them, which is in stark contrast with approaches that rely purely on computer vision datasets.

The importance of non-visual sensory modalities for robotic object perception has been recognized by several lines of research, which have shown that robots can recognize objects using auditory [28–30], tactile [31,32], and proprioceptive [33,34] sensory modalities. For example, Natale et al. [33] showed that proprioceptive information obtained from the robot's hand when grasping an object can be used to successfully recognize the identity of the object. Similarly, Bergquist et al. [34] performed an experiment in which a robot was able to recognize a large number of objects using proprioceptive feedback from the robot's arm as it manipulated them. Other research has also shown that auditory features (e.g., sounds generated as the robot's end effector makes contact with an object) can also be useful for recognizing a previously explored object [28,29]. Most recently, a study by Sinapov et al. [35] demonstrated that a robot can achieve high object recognition rates when tested on a large set of 50 objects by integrating auditory and proprioceptive feedback detected over the course of exploring the objects. In contrast to this previous work, the study in this paper demonstrates that behavior-grounded object perception can also be used by a robot to both learn and recognize human-provided semantic category labels for novel objects.

Several studies have already demonstrated some ability of robots to assign category labels to objects based on interaction with them. For example, Takamuku et al. [36] demonstrated that a robot can classify 9 different objects as either a rigid object, a paper object, or a plastic bottle using auditory and joint angle data obtained when the robot shakes the objects. An experiment by Chitta et al. [37] has shown that tactile feedback produced during grasping can be useful for categorizing cans and bottles as either full or empty. In another study, Sinapov and Stoytchev [38] showed that by applying five different exploratory behaviors on 36 objects, a robot may learn to recognize their material type and whether they are full or empty, based on the auditory feedback produced by the objects.

In previous work, we proposed a graph-based learning method that allows a robot to estimate the category label of an object based on pairwise object similarity relations estimated from different couplings of five exploratory behaviors and two sensory modalities [13]. In that experiment, the robot was able to classify 25 objects according to object categories such as plastic bottles, objects with contents, pop cans, etc. The accuracy was substantially better than chance, despite the fact that visual feedback was not used.

To further improve category recognition rates, the study presented in this paper describes a method that scales to a much larger number of exploratory behaviors, sensory modalities, and objects than any previously published experiments in which robots have perceived objects by interacting with them. More specifically, in addition to doubling the number of objects, this paper also doubles the number of behaviors and more than triples the number of sensorimotor contexts as compared to our previous work [35] (which only focused on object recognition rather than category recognition). In addition, we also show that by using prior information in the form of confusion rates for all categories, the robot can actively select which behavior to apply next when classifying a novel object.

3. Experimental platform

3.1. Robot and sensors

The experiments were performed with the upper-torso humanoid robot shown in Fig. 1. The robot has as its actuators two 7-DOF Barrett Whole Arm Manipulators (WAMs), each with an attached 3-finger Barrett Hand. Each WAM has built-in sensors that measure joint angles and torques at 500 Hz. An Audio-Technica U853AW cardioid microphone mounted in the robot's head was used to capture auditory feedback at the standard 16-bit/44.1 kHz resolution and rate over a single channel. The robot's right eye