Robotics and Autonomous Systems 61 (2013) 1665–1679

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Structure-based object representation and classification in mobile robotics through a Microsoft Kinect

Antonio Sgorbissa *, Damiano Verda

DIBRIS - University of Genova, Via Opera Pia 13, 16145 Genova, Italy

HIGHLIGHTS

- A new approach for object representation and classification is proposed.
- We rely on considerations about the structure of furniture-sized objects.
- The 3D point cloud returned by the Kinect is segmented into a set of clusters.
- Objects are represented by expressing mutual relationships between clusters.
- The approach is validated through experiments with real data.

ARTICLE INFO

Article history:
Received 8 January 2013
Received in revised form
6 June 2013
Accepted 21 June 2013
Available online 9 July 2013

Keywords:
Object classification
Structure-based modelling and recognition
Microsoft Kinect
Mobile robotics

ABSTRACT

A new approach enabling a mobile robot to recognize and classify furniture-like objects composed of assembled parts using a Microsoft Kinect is presented. Starting from considerations about the structure of furniture-like objects, i.e., objects which can play a role in the course of a mobile robot mission, the 3D point cloud returned by the Kinect is first segmented into a set of "almost convex" clusters. Objects are then represented by means of a graph expressing mutual relationships between such clusters. Off-line, snapshots of the same object taken from different positions are processed and merged, in order to produce multiple-view models that are used to populate a database. On-line, as soon as a new object is observed, a run-time window of subsequent snapshots is used to search for a correspondence in the database.

Experiments validating the approach with a set of objects (i.e., chairs, tables, but also other robots) are reported and discussed in detail.

1. Introduction

The article describes a system for recognizing and classifying furniture-like objects composed of assembled parts on the basis of a sequence of snapshots taken with a Microsoft Kinect.[1] The system is meant to provide a mobile robot with advanced perceptual capabilities, with the final aim of enabling it to interact with the environment in different ways depending on the affordances offered by different classes of objects.

The concept of affordance initially proposed by the psychologist J. Gibson [1] has been widely adopted in robotics, mainly in the context of grasping and manipulation. Even if the present work does not deal explicitly with object affordances, the concept helps us to motivate the development of a system for classifying furniture-like objects in the context of mobile robotics, and therefore deserves a deeper discussion. "Affordance" is a term used to describe a possibility for actions: this possibility varies depending both on the object and on the agent performing the action, and therefore cannot be uniquely expressed as an intrinsic characteristic of a given object or environmental feature. For example a chair *affords pushing* both to humans and mobile robots, but it *affords sitting* only to humans. A table *affords pushing* to humans but usually not to mobile robots. A door *affords passing through* to robots and humans, but not to elephants or cars.

Suppose now that, in the course of a mission, the robot encounters something that blocks its path. The robot can try to push it away, or ask the object to move on: this corresponds to exploring and learning the affordances of the new object. However, if the robot is able to use the perceptual data to recognize the object and to classify it as belonging to a given category (e.g., a chair, a table, or a human), it can immediately infer what the object affords or not, and act accordingly. As the reader can imagine, what the robot needs is not the capability to recognize a particular chair or table, but to be able to abstract from the peculiar characteristics of individual objects through a conceptualization process whose final output is a model of a given class of objects. The robot needs

* Corresponding author. Tel.: +39 010 3532706; fax: +39 010 3532154.
E-mail addresses: antonio.sgorbissa@unige.it (A. Sgorbissa), damiano.verda@unige.it (D. Verda).

[1] The Microsoft Kinect is an RGB-D camera: in addition to standard RGB information it returns, for every pixel, the distance to the closest object, thus effectively providing depth information that can be used to build a volumetric representation of the scene.

0921-8890/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.robot.2013.06.006

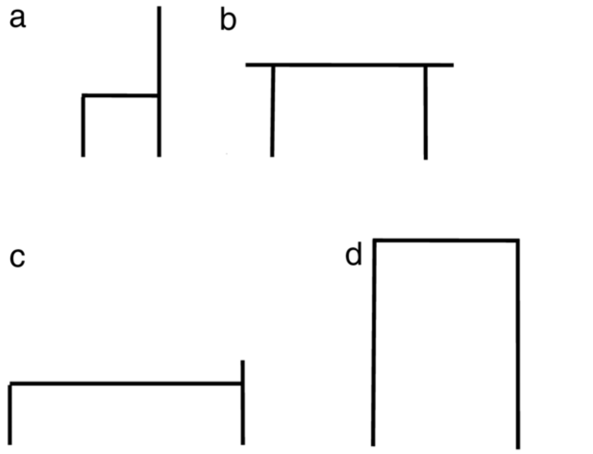

Fig. 1. Pictographs representing a chair (a), a table (b), a bed (c), and a door (d).

to be able to distinguish chairs from tables (since the former can be pushed away, whereas the latter cannot), and not to recognize a particular chair or table.

In the recent robotic literature, object recognition and classification has received great attention [2–13]. The major contribution of this work is to propose a novel approach that relies on the depth data returned by a Microsoft Kinect, and focuses on recognizing and classifying only those objects that offer possibilities of actions to a mobile robot: for instance chairs, tables, other robots, *but not* cups, pens, or wall clocks, which appear not to play a relevant role in the context of mobile robot navigation. Using the Kinect, objects are initially perceived as point clouds, possibly through different snapshots taken from different perspectives: the general idea is to merge the information acquired in subsequent snapshots to model objects as composed of "almost convex" geometrical primitives, and finally to match the data acquired in run-time with precomputed models stored in a database.

The approach is motivated by the consideration that a very small set of geometric primitives and their mutual geometrical relationships appear to be sufficient for humans to distinguish among common use objects. Consider the pictographs in Fig. 1: by submitting the picture to volunteers and by asking them which objects are represented in it, almost 100% of the interviewed volunteers have no difficulties in recognizing (a) as a chair, (b) as a table, (c) as a bed, and (d) as a door (the test has open answers, i.e., volunteers are not provided with a list of objects to choose among). In a similar spirit, we conjecture that modelling an object in the real world through a limited set of geometrical primitives (labelled with their geometrical properties, as well as their mutual geometrical and topological relationships) should be sufficient to characterize it as a member of the class it belongs to, thus being recognizable by the system. For instance, chairs have a seat orthogonal to legs, a back orthogonal to the seat, and legs parallel to each others.

This conjecture is enforced by the fact that, in most cases, furniture-like objects in human environments are not only *modelable* as composed of "almost convex" parts. As a matter of fact, pieces of furnitures are actually *built* by assembling smaller parts, for obvious constraints puts by the manufacturing process. Interestingly enough, notice also that the design of many pieces of furniture has not substantially changed since centuries: the way of designing and building a modern chair or table has inherited the constraints put by woodworking centuries ago, even if some of these constraints do not hold any more after the introduction of plastic. Then, even when a plastic chair is actually composed of a single piece, the latter is still modelable as if it were composed of a seat, a back, and legs.

Section 2 describes related work. Section 3 introduces the system's architecture. The processing phases which are required to build the model of an object starting from depth data are described in Sections 4–6, whereas the classification process is described in Section 7. Section 8 describes the experimental results. The conclusions follow.

2. Related work

Object recognition and classification is a widely investigated topic in the scientific literature: among the others, it plays an important role in robotics, for navigation as well as for grasping and manipulation, in order to enable the robot to behave differently when dealing with different objects.

Most approaches in the recent literature rely on appearance-based techniques, and perform object recognition and classification on the basis of a set of invariant descriptors. In a few words, the basic idea is that of taking snapshots of the objects, computing a set of global and local descriptors which summarize data, and then comparing objects by computing the "distance" between vectors of such descriptors. This is different from past research (i.e., up to twenty years ago), when holistic approaches based on global features were intensively investigated in computer vision.

Among the others, many works assume that the robot has 2D vision capabilities, and rely on the local descriptors such as SIFT features [14] or similar. As an example, the authors of [15] focus on robots controlled via the Internet, and deal with object recognition as a prerequisite to allow the user to issue voice commands to refer to the objects by their names (e.g., "grasp the cube"): a neural-network is proposed, which is able to classify the objects depicted in low-resolution and noisy images on the basis of a set of invariant descriptors. In [2] an agent-based architecture is presented which is able of continuous, supervised learning through the feedback that the robot receives from the user. The approach relies on a sequence of processing phases involving multiple object representations (i.e., different orientation-independent features are extracted from the same image) as well as multiple classifiers (i.e., different similarity and membership measures are adopted), which are then combined in a probabilistic framework. The work described in [8] focuses on the problem of learning affordances of objects which are relevant for navigation and manipulation tasks, and presents a probabilistic model that describes the relationships between object categories, affordances, and their visual appearance.

Some authors argue that object recognition can be improved by considering multiple perceptual and motor modalities. In this spirit, [3] propose a statistical technique for multimodal object categorization based on audio–visual and haptic information, by allowing the robot to use its physical embodiment to grasp and observe an object from various view points, as well as to listen to the sound during the observation. In [16] a method is proposed to learn a motor/perceptual model of objects belonging to different categories (e.g., container/non-container objects), which can be generalized to classify novel objects. In [13] a dataset of 100 objects divided in 20 categories is considered to test the ability of a humanoid robotic torso to interact in the most appropriate way with different categories of objects: to this purpose, the authors propose a supervised recognition method that does not rely on vision alone, but on the integration of different exploration behaviours taking into account visual, proprioceptive and acoustic information: look, grasp, lift, hold, shake, drop, tap, poke, push, and press. A feature vector is properly recorded for each combination of perceptual modality/behaviour/object, which is then used to train a Support Vector Machine for classification. Similarly, [17] start from the consideration that the simple interactions with objects in the environment lead to a manifestation of the perceptual