Robotics and Autonomous Systems 62 (2014) 25–37

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Localization of furniture parts by integrating range and intensity data robust against depths with low signal-to-noise ratio

Pascal Meißner [a,*], Sven R. Schmidt-Rohr [b], Martin Lösch [b], Rainer Jäkel [b], Rüdiger Dillmann [b]

[a] FZI—Research Center for Information Technologies, Germany

[b] Institute of Anthropomatics, Karlsruhe Institute of Technology, Germany

ARTICLE INFO

Article history:
Available online 8 August 2012

Keywords:
3D computer vision
Localization
Range and intensity data
Furniture
Kinect
Time-of-flight camera

ABSTRACT

In this article we present an approach for localizing planar parts of furniture in depth data from range cameras. It estimates both their six-degree-of-freedom poses and their dimensions. The system has been designed for enabling robots to autonomously manipulate furniture. Range cameras are a promising sensor category for this application. As many of them provide data with considerable noise and distortions, detecting objects, for example, using canonical methods for range data segmentation or feature extraction, is complicated. In contrast, our approach is able to overcome these issues. This is done by combining concepts of 2D and 3D computer vision as well as integrating intensity and range information in multiple steps of our processing chain. Therefore it can be employed on range sensors with both low and high signal-to-noise ratios and in particular on time-of-flight cameras. This concept can be adapted to various object shapes. It has been implemented for object parts with shapes similar to ellipses as a proof-of-concept. For this, a state-of-the-art ellipse detection method has been enhanced regarding our application.

1. Introduction

In Programming by Demonstration (PbD), robots should be able to learn by observing interactions between humans and their environment. As learning is based on perception in PbD, solving problems like learning planning models for probabilistic decision making [1] or manipulation tasks [2] highly depends on powerful systems to understand human activities [3], as well as on approaches to detect objects in the real world. To enable robots to manipulate and navigate, it is necessary to localize objects with respect to their six-degree-of-freedom (6-DoF) poses. Not only are poses for common household objects like cups needed, but also the pose and dimensions of furniture must not be assumed as fixed or known in advance.

For being able to estimate precise 6-DoF poses (and dimensions) of objects, data about 3D shapes in the environment has to be acquired. For example, a cup and dust bin can have the same shape, but they usually differ in their sizes. Therefore they cannot be distinguished in general using just conventional image data. In PbD, 3D data is captured using sensors like the ToF (time-of-flight) camera SwissRanger 4000 [4] or the structured light scanner Kinect [5] as they are small, safe and real-time capable. These sensors have specific strengths and weaknesses depending on the measuring principle they are based on. Thus the right sensor has to be chosen depending on the application scenario. In particular, they significantly differ in the signal-to-noise ratio of the data they provide. One of our main concerns is to develop localization methods that can cope with varying image data qualities.

In general, range sensors provide distance information that is less dense and accurate than intensity images from high resolution cameras being state-of-the-art in robotics. In contrast, shapes in conventional images do not necessarily originate from objects in the real world. They can for example easily be mixed up with illustrations on posters or shadows. On the account that range and intensity data complement one another, combining them is a way to compensate the restrictions of both. For example, in the approach presented here, segmenting range information directly is avoided. Instead, edges [6] from intensity and range images are combined. The enhanced result is then used to segment range data. Furthermore, instead of solely using distances to compute the shape of an object, we just estimate the surface of a segment in range data and calculate the contours of the segment by projecting information from intensity data on the surface from range data. The resulting object part models are sufficiently accurate for the purpose of autonomous robot manipulation. With respect to the varying data qualities of the employed range sensors, this approach should resolve the following problems. While for the SwissRanger

* Corresponding author. Tel.: +49 72160845942; fax: +49 72160845959.
E-mail addresses: meissner@fzi.de (P. Meißner), sven.schmidt-rohr@kit.edu (S.R. Schmidt-Rohr), martin.loesch@kit.edu (M. Lösch), rainer.jaekel@kit.edu (R. Jäkel), ruediger.dillmann@kit.edu (R. Dillmann).

0921-8890/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2012.07.018

Fig. 1. Stereotypical setup with the furniture used in our laboratory.

with low signal-to-noise ratio, it has to make localization in our scenario possible in the first place, the quality of the results as well as the robustness of the system should be maximized by combining range and intensity data when using the high quality Kinect.

Furniture like the table and the chair depicted in Fig. 1 is often poorly textured and, unlike the red cup on the table which is quite outstanding in its environment, its colors are nondescript. Its geometry is more significant but plain. The big parts of both objects are planar and their shapes in $\mathbb{R}^3$ are similar to simple geometrical objects like circles or rectangles. Consequently, detecting these parts should be feasible if we analyze their surfaces and their overall shapes; concerning shape, a holistic feature extraction like a (generalized) Hough transform is suggested. In the approach presented here, it is used to precisely segment 3D point clouds based on intensity information. Due to its size and function, furniture is often occluded. Using holistic methods in this context is useful, because they are highly robust against occlusions and can cope with incomplete object contours. Besides pure localization, it is also necessary to identify what has been localized. In our system this is achieved with a part-based object recognition [7] method that has been extended using fuzzy set theory [8], so that it can work with fuzzy object definitions given by humans. This is particularly suitable for capturing furniture concepts, so we preferred it to a learning-based recognition. In order to simplify implementation and evaluation, our system only detects shapes similar to ellipses so far. Nevertheless, it could be extended to other shapes as described in Section 10. As we rather want to localize types of objects than identifying a certain object instance, another concern for our part-based recognition is that it is variable concerning the dimensions of the parts. Choosing the Hough transform for extracting the shape corresponds to this concept as it templates the form leaving its dimensions variable.

2. Related work

In recent years the emerging range camera technology has gotten a lot of attention in the robotics community. However, various types of such sensors like the ToF cameras deliver data that is highly affected by noise and distortions as depicted in Fig. 2. Canonical range image processing methods are designed for data with much higher signal-to-noise ratio. For example a state-of-the-art region-based method [7] for range image segmentation fails on data from a SwissRanger ToF camera. The same applies to feature extraction, e.g. with spin images [9]. Many current publications in the field still do not consider such data; see, e.g., [10]. The fact that numerous publications that deal with getting segmentation algorithms more robust for ToF camera data [11–13] were written in recent years also suggests that researchers have work to do on elementary computer vision problems when they use data from this sensor type.

The low resolution of images from range sensors is another issue that is currently being dealt with. For example, high resolution pictures taken by conventional cameras and depicting the same scenery as range images can be used to interpolate depth information. In [14] this is done this by applying generic modeling frameworks such as "Markov Random Fields" on laser scans. Such frameworks generally consume too much processing time to be applicable on mobile robots. These approaches assume that changes in distance and color values co-occur in the involved range and intensity images. But distortions in data from ToF cameras cause noticeable displacements of image structures. Consequently, the co-occurrence requirement of these methods valid for laser range finder data is not met by data from ToF cameras. In contrast, the approach presented here relies neither on expensive methods nor on precise co-occurrence of discontinuities in images.

Instead of dealing with a particular computer vision issue like segmentation, the author of [15] presents an entire process chain combining methods from different research fields. It is able to classify chairs composed of parts with simple geometries in images from ToF cameras. It does not aim to improve segmentation, but avoids problems when dealing with a range image of poor quality by processing image sequences in a tracking algorithm.

In our approach a complete computer vision processing chain is built up as well and its application area is similar. However, instead of fusing series of images from the same range camera, we combine data from different sources. The resulting object representations are also different. While the author of [15] attaches great importance to model the overall structure of detected objects, details of the shape of the objects' parts are not taken into account. For example, object parts are approximated by bounding boxes. In comparison, our method is designed to estimate precise models of the geometry of object parts. Alenya et al. [16] is a recent publication which takes advance of ToF camera data by fusing depth information with color segmentation, and to some extent this has a similar philosophy to our publication.

Radu et al. [17] presents a state-of-the-art processing chain that acquires object maps based on high quality laser range finder data. It focuses on furniture like tables or cupboards that are represented as rectangular planes. To localize them, planes are segmented in point clouds employing region growing and rectangles are detected based on planar subsets of the cloud. Parameters for the algorithms seem to be defined by hand using prior knowledge. Thanks to the provided data quality it can use different algorithms than our method, but the approach to use strong modeling combined with prior knowledge is similar to ours.

Our method combines 2D and 3D information focusing on the shapes of objects and their parts. Other approaches that integrate 2D and 3D data like [18] use local features extracted from color images. For our application, extracting local features is not as suitable as a holistic shape-based method like a Hough transform. Such local features are intended for objects with significant appearances, whereas furniture is better described by its shape, as already stated in Section 1. In [19], a new type of local feature extraction is presented that is solely based on 3D range images and evaluates information about borders of shapes.[1] Object models used for feature matching with a given scenery are views on instances of an object type. In contrast, we use object category models. Furthermore, it is stated that performance on objects like those depicted in Fig. 1 is suboptimal. We evaluated this method in Section 9.5 for the sake of a comparison with our approach.

[1] It is employed on data from laser range finders and stereo camera setups.