A Dynamic Programming Approach for Fast and Robust Object Pose Recognition from Range Images

Xuewen Yang

August 5 2018

Abstract

Joint object recognition and pose estimation solely from range images is an important task e.g. in robotics applications and in automated manufacturing environments. The lack of color information and limitations of current commodity depth sensors make this task a challenging computer vision problem, and a standard random sampling based approach is prohibitively time-consuming.The author propose to address this difficult problem by generating promising inlier sets for pose estimation by early rejection of clear outliers with the help of local belief propagation (or dynamic programming). By exploiting data-parallelism this method is fast, and they also do not rely on a computationally expensive training phase.They demonstrate state-of-the art performance on a standard dataset and illustrate our approach on challenging real sequences.

1. Introduction

Object detection from 3D inputs has been widely researched during the past decade. Initially many solutions focused on trying to solve object detection from laser scans or even from synthetically generated meshes [1] [2]. However, with the popularization of RGB-D sensors since 2010 there has been an increasing demand of algorithms that operate at interactive frame rates and that are able to cope with inputs that are less reliable than laser scans. Most of the latter algorithms rely heavily on RGB data to perform detection, which prohibits the application of these methods on 3D only inputs. Several approaches use a global description of the object (using RGB edges and depth normals), hence these methods have difficulties in handling occlusions.Methods that utilize only 3D data as input can be based on either global or local object representations. Several proposed methods based on a global object representation employ the Hough transform. These approaches create a set of features that are accumulated in a Hough voting space and then select the pose which gathered the largest number of votes.Mian *et al.* [4] also use the normal to obtain an invariant descriptor: they fix the local coordinate frame by using two points on the model (in additional to the normal), and fill an occupancy grid given the local coordinate frame.

2. Approach

2.1. Descriptor Computation

A natural choice for a descriptor to represent (local) geometry is based on an implicit volumetric representation of range images and 3D surface meshes.The author employ a binary occupancy grid to compute descriptors. A slightly more discriminative volumetric data structure would be a (truncated) signed distance function (TSDF), but they discard this option for efficiency reasons (proper TSDF computation is costly, and the descriptors would use several bits per voxel).they believe that using generalizations of successful gradient-based image descriptors to 3D shapes (such as 3D-SURF [3]) is not necessary, since the intensity values of the (3D) image are known to be only 0 and 1 for occupancy grids (and therefore invariance to intensity transformations is unnecessary). Consequently, this descriptor is a bit string of occupancies in the vicinity of a surface point.

2.2. Matching

At test time descriptors are computed for each pixel with valid depth and estimated surface normal in the (sub-sampled) depth image, and the task is to efficiently determine the set of object coordinates with similar local shape appearance. The natural choice to quantify similarity of binary strings is the Hamming distance. The author experimented with approximated nearest neighbours implementation for binary data in FLANN [5] and with a hashing based indexing data structure using orthonormal projections. Since in experience the performance is roughly similar for both acceleration strategies,they only report the results using FLANN below.

3. Conclusions

The author have addressed the problem of 3D object detection and corresponding pose estimation, and they discussed a more efficient paradigm to solve this task while still obtaining state of the art detection rates.The author believe that this work creates a new and robust framework from which to build new 3D object detection approaches. In this current work they left out basically any learning-based technique to boost the detection performance or run-time behavior. While they argue that computationally expensive learning techniques will limit the general applicability of 3D object recognition (since adding new objects requires time-consuming retraining), they foresee that more sophisticated processing of training objects than this current one will lead to more discriminative descriptors, and therefore will be highly beneficial for this task.

References

[1] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *CVPR*, 2010. 1

[2] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE TPAMI*, 2002. 1

[3] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. V. Gool. Hough transform and 3d surf for robust three dimensional classification. In *ECCV*, 2010. 1

[4] A. S. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE TPAMI*, 2006. 1

[5] M. Muja and D. G. Lowe. Fast matching of binary features. In *Computer and Robot Vision*, 2012. 1