A Dataset for Improved RGBD-based Object Detection and Pose Estimation for Warehouse Pick-and-Place

Colin Rennie[1], Rahul Shome[1], Kostas E. Bekris[1] and Alberto F. De Souza[2]

Abstract— An important logistics application of robotics involves manipulators that pick-and-place objects placed in warehouse shelves. A critical aspect of this task corresponds to detecting the pose of a known object in the shelf using visual data. Solving this problem can be assisted by the use of an RGBD sensor, which also provides depth information beyond visual data. Nevertheless, it remains a challenging problem since multiple issues need to be addressed, such as low illumination inside shelves, clutter, texture-less and reflective objects as well as the limitations of depth sensors. This paper provides a new rich dataset for advancing the state-of-the-art in RGBD-based 3D object pose estimation, which is focused on the challenges that arise when solving warehouse pick-and-place tasks. The publicly available dataset includes thousands of images and corresponding ground truth data for the objects used during the first Amazon Picking Challenge at different poses and clutter conditions. Each image is accompanied with ground truth information to assist in the evaluation of algorithms for object detection. To show the utility of the dataset, a recent algorithm for RGBD-based pose estimation is evaluated in this paper. Given the measured performance of the algorithm on the dataset, this paper shows how it is possible to devise modifications and improvements to increase the accuracy of pose estimation algorithms. This process can be easily applied to a variety of different methodologies for object pose detection and improve performance in the domain of warehouse pick-and-place.

I. INTRODUCTION

There is significant interest in warehouse automation, which frequently involves pick-and-place tasks for products located in shelving units. This interest is exemplified by competitions such as the first Amazon Picking Challenge (APC) [1], which brought together multiple academic and industrial teams from around the world, as well as similar competitions, like the Robocup@Home

Fig. 1. An example frame from the Rutgers dataset, where a pose estimate generated by the test algorithm is superimposed.

Challenge [2]. One way to approach the APC involved perception, motion planning and grasping of 25 different objects, which were placed in a semi-structured way inside the bins of an Amazon-Kiva Pod. Solving such problems reliably can significantly alter the logistics of distributing products. Frequently, manipulation research on pick-and-place tasks has focused on flat surfaces, such as tabletops. These are relatively simpler problems, which do not involve many of the issues that often arise in warehouse automation, where the presence of tight spaces, such as shelves, plays a critical role.

Accurate pose estimation is crucial for successfully picking an object inside a shelf. In flexible warehouses, this pose will not be a priori known but must be detected from sensors, especially visual ones. The increasing availability of RGB-D sensors, which can simultaneously sense color and depth, brings the hope that such problems can be eventually solved reliably. But warehouse shelves have narrow, dark and obscuring bins that complicate object detection. Clutter can further challenge detection through the presence of multiple objects. A variety of object types need to be dealt with, some of which may be texture-less and not easily identifiable from color, and others reflective and virtually undetectable by a depth sensor. Furthermore, some popular depth sensors exhibit limits in terms of the smallest

and largest sensing radius that make it harder for a manipulator to utilize them. Thus, RGBD-based object detection and pose estimation is an active research area and a critical capability for warehouse automation.

This paper provides tools that help in improving the performance of object detection solutions for such challenges. In particular, it describes a new rich dataset and software for utilizing it. The motivation is to better equip the research community in evaluating and improving robotic perception solutions for warehouse picking. The dataset contains over 10,000 depth and RGB registered images, complete with hand-annotated 6DOF poses for 24 of the APC objects (for details, see Section 3). Also provided are 3D mesh models of the APC objects, which may be used for training of recognition algorithms. The code for utilizing and integrating the dataset with different algorithms is also publicly available.

The dataset includes images of warehouse objects in a shelf environment. The objects are placed in different poses in various bins of warehouse shelves, so as to allow a variety of experimental conditions (Figure 6). Multiple camera perspectives and frames account for rich information, as well as spatial and temporal variation in data. The effect of clutter is evaluated by controlling the presence of additional objects in a scene.

The dataset is compared against the one available as part of the LINEMOD framework for object detection [3], to highlight the need for additional varying conditions, such as clutter, camera perspective and noise, which affect pose detection. This is the chief contribution of the dataset, the utility of which is further evaluated by using the open-source implementation of the LINEMOD framework [3]) easily accessible via OpenCV [4]. This paper does not argue that this algorithm is the best solution for pose estimation in shelves. The method is used as an example of a modern, accessible algorithm for object detection, which at least performs effectively in tabletop setups.

The dataset reveals that the considered algorithm faces significant difficulties when used in a warehouse scenario. This allows to appreciate the features of warehouse picking, which complicate pose estimation. With the aid of the dataset, it was also possible to identify algorithmic and engineering adaptations to increase performance in warehouse pick-and-place.

Overall, the proposed dataset emphasizes the need for the development of pose detection algorithms that can operate robustly for a wide variety of objects and conditions, especially in narrow, dark and cluttered spaces. Such algorithms need to optimally utilize all available sources of sensing data and prior information.

II. RELATED WORK

Datasets for the task of object recognition have rapidly grown in recent years both in terms of number as well as size and scale. The applications of such datasets include industrial warehouse applications like the APC [1], and domestic applications like in the RoboCup@home [5]. Some standard RGB benchmarks for the task include CIFAR-10/100 [6], ImageNet [7], and PASCAL VOC [8]. Some use bounding boxes as ground truth and others use image segmentation with inliers/outliers for accuracy metrics. While useful for 2D image object recognition, RGB datasets are not ideal in manipulation applications, which rely not only on segmenting the object of interest but also on accurate pose estimation.

Up until the last decade, the problem of 3D recognition was often addressed using a stereo camera. More recently, RGB-D cameras' availability and widespread use have increased interest in solutions to common 2.5D[1] problems, such as face, object, and gesture recognition. Such technology has allowed researchers to begin to build "modern-scale" datasets, which help evaluating performance and identifying challenges. Several such datasets are described below[2].

Segmented Scenes Datasets

B3DO [9]: A project by UC Berkeley, the dataset contains >3,000 2.5D crowd-sourced images. The images primarily focus on indoor scenes, where ground truth bounding boxes have been annotated for more than 50 object categories. The dataset has also been augmented to include (x, y, z) Cartesian coordinates for many object centroids.

NYU Depth Dataset v2 [10]: The NYU dataset also focuses on indoor scenes, but ground truth labels are presented as full image segmentation. The dataset includes around 500k 2.5D images, with approximately 1,500 fully labeled ground truth images.

Manipulation Datasets

YCB Objects & Models [11]: A collaboration between several robotics labs, the YCB dataset provides object models in a variety of formats for common household objects. The focus of this project is to create common metrics for the growing interest in robotic manipulation research by providing reliable benchmarks for several common manipulation tasks.

---

[1]2.5D refers here to the projection of a 2D image to 3D space, which results in a sparse 3D image.

[2]A more complete list of available RGB-D datasets can be found at: `http://www0.cs.ucl.ac.uk/staff/M.Firman/RGBDdatasets/`

Fig. 2. (Left) Items used in the Amazon Picking Challenge 2015 and featured in the dataset. Three groups of objects are identified based on their effects on pose estimation from RGB-D data: a) cuboid and non-transparent, b) non-cuboid and non-transparent, c) transparent. (Right) An arrangement of the shelf with the APC objects.

Object Datasets

Table-Top Object Dataset [12]: A collaboration between Willow Garage and the Univ. of Michigan, this dataset consists of ∼1,000 2.5D images with ground truth labels for 480 frames. The objects presented belong to 3 different classes, each class consisting of approximately 10 different instances. Objects are shown on table tops in clutter of between 2-6 items per image. The images were collected using a structured light stereo camera.

Solutions in Perception Dataset [13]: This dataset by Willow Garage contains 35 objects in ∼1,000 3D training images and 120 test images. In training, objects were presented in clutter with 6DOF ground truth for each item. The scenes were captured using RGB-D cameras with objects on a turn-table to capture and reconstruct the scenes from multiple viewpoints. All images were captured using a consistent azimuth angle between the camera and the turntable.

UW Dataset [14]: This large dataset consists of over 50 object categories and 300 distinct instances. It features objects from multiple viewpoints, and is presented with ground truth pose for one axis [0,2π].

LINEMOD Dataset [15]: As part of the body of work detailing the LINEMOD framework, the authors released a dataset of 18 object models and over 15,000 6D ground truth annotated RGB-D images. Objects in these images are shown in clutter from a variety of viewpoints. Because of the size, setting, and focus on 6D pose estimation, this dataset is the most closely related to the current paper.

The dataset proposed here presents more than 10,000 ground truth annotated RGB-D images of 24 objects of different types. As opposed to prior datasets [14], [10], [9], this new dataset is specifically aimed at perception for robotic grasping and hence features full 6DOF ground truth poses for all 2.5D images. While some existing datasets [12], [15] provide ground truth poses for objects in cluttered space, the new one additionally controls for clutter by presenting poses of the objects both with and without clutter. Other controls employed in data collection correspond to multiple viewpoints and collection of additional frames for control of slight noise in sensors. Additionally, scenes are not reconstructed as in alternatives [13], but the dataset includes the transformation matrices between the camera location, stationary robotic base, and object location. This allows users of the dataset to reconstruct the scene to suit their own methods. Lastly, this new dataset is specifically designed for warehouse perception task and is focused on the placement of objects in narrow spaces, such as shelf bins. To the best of the authors' knowledge, this is the first attempt to generate a real-world dataset for this important application.

III. RUTGERS APC RGB-D DATASET

This paper presents a large 2.5D dataset consisting of 10k+ images and corresponding ground truth 6DOF poses for all these images, which is made available to the research community [3]. The focus is on supporting warehouse pick-and-place tasks. The accompanying software allows the easy evaluation of object detection and pose estimation algorithms in this context.

A. Objects and 3D-mesh Models

The selected objects correspond to those that were used during the first Amazon Picking Challenge (APC) [1], which took place in Seattle during May 2015.

The provided dataset comes together with 3D-mesh object models for each of the APC competition objects. For most objects, the CAD 3D scale models of the objects were textured using the open-source MeshLab software. For simple geometric shapes, such as cuboids, this simple combination of CAD modeling and texturing is sufficient and can yield results of similar quality to more involved techniques [16]. For objects with non-uniform geometries, models were produced using 3D photogrammetric reconstruction from multiple views of a monocular camera.

[3]It can be accessed online at the following url: http://www.pracsyslab.org/rutgers_apc_rgbd_dataset

Fig. 3. The data collection setup for the warehouse pick-and-place dataset: A Motoman SDA10F robot and an Amazon-Kiva Pod stocked with objects. At this configuration, the Kinect sensor mounted on the arm is used to detect an object at the bottom row of shelf bins.

B. Dataset Design

The intention with the dataset was to provide to the community a large scale representation of the problem of 6DOF pose estimation using current 2.5D RGB-D technology in a cluttered warehouse shelf environment. The set of 25 APC objects that were part of the competition exhibit a variety of features in the problem domain, including different size, shape, texture and transparency.

C. Extent of the Dataset

Data collection was performed using a Microsoft Kinect v1 2.5D RGB-D camera mounted to the end joint of a Motoman Dual-arm SDA10F robot (Figure 3). Changing the intensity of the structured light in the Kinect driver allowed operation at closer distances.

To provide better coverage of the scene and the ability to perform pose estimation from multiple vantage points, data from 3 separate positions (referred to, here, as "mapping" positions) were collected: i) One directly in front of the center of a bin at a distance of 48cm, ii) a second roughly 10cm to the left of the first position, and iii) a third with the same distance to the right of the first position. Four 2.5D images were collected at each mapping position to account for noise.

To measure the effects of clutter, for each object-pose combination, images were collected: (1) with only the object of interest occupying the bin, (2) with a single additional item of clutter within the bin, and (3) with two additional items of clutter. In all, the dataset can be broken down into the following parameters:

- 24 Objects of interest[4]
- 12 Bin locations per object
- 3 Clutter states per bin

- 3 Mapping positions per clutter state
- 4 Frames per mapping position

Considering all these parameters, the dataset is composed of a total of 10,368 2.5D images. For each image, there is a YAML file available containing the transformation matrices (rotation, translation) between: (1) the base of the robot and the camera, (2) the camera and the ground truth pose of the object, and (3) the base of the robot and ground truth pose of the object.

The process for generating the ground truth data involved iterating over all the frames of the Rutgers APC dataset in a semi-manual manner. A human annotator translated and rotated the 3D model of the object in the corresponding RGB-D point cloud scene using RViz. Every annotation superimposes the model to the corresponding portion of the point cloud.[5]

IV. Opportunities for Pose Estimation Improvements Through the Dataset

This paper employs a setup similar to that of the Amazon Picking Challenge (APC) to evaluate the proposed dataset, which is a helpful testing ground for robotic perception algorithms in a relatively controlled but realistic warehouse environment. The available software infrastructure for using the dataset allows the incorporation of different algorithms for this problem, given the rich literature on the subject [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. The current paper utilizes one such approach that is easily accessible to the robotics community and corresponds to the LINEMOD algorithm [3], for which an implementation based on the OpenCV library [4] is available.

LINEMOD is an object detection and pose estimation pipeline [3], which received as input a 3D mesh object model. From the model, various viewpoints and features from multiple modalities (RGB gradients, surface normals) are sampled. The features are filtered to a robust set and stored as a template for the object and the given viewpoint. This process is repeated until sufficient coverage of the object is reached from different viewpoints. The detection process implements a template matching algorithm followed by several post-processing steps to refine the pose estimate. The approach was designed specifically for texture-less objects, which are notoriously challenging for pose estimation methods based on color and texture. LINEMOD uses surface normals in the template matching algorithm and limits RGB gradient features to the object's silhouette.

---

[4]The "mead_index_cards" item from the APC list is not included as this simplified the experimental process for collecting the data and it was the item that exhibited the most redundant qualities.

[5]Additional details regarding naming conventions for the dataset and instructions for download and use can be found at the project's website: `http://www.pracsyslab.org/rutgers_apc_rgbd_dataset`.

Starting with the *baseline* open source implementation of LINEMOD, the paper shows the incremental performance improvements achieved over the basic implementation algorithm through the use of the Rutgers APC dataset. Most of the improvements are algorithm-agnostic and can be useful in general to warehouse detection and pose estimation tasks.

A. Masking

In the context of the APC, the bin of the shelf from which the object is detected and grasped is specified. In order to take advantage of such information, precise calibration of the shelf's location with respect to the robot is performed prior to detection. Using ROS' TF functionality [30] it is possible to compute the boundary of the current bin of interest: $([x_{min}, x_{max}], [y_{min}, y_{max}], [z_{min}, z_{max}])$. Then, all points $p_i = (p_i^x, p_i^y, p_i^z)$ are masked if

$$\prod_{j \in (x,y,z)} (j_{max} > p_i^j > j_{min}) == 0$$

B. Post-processing

Dynamically selecting a viable detection threshold value allows the algorithm to always allow a fixed number of detections through and increases the positive detection rate in the context of warehouse pick-and-place since it is known that the object is present in the scene. Pose detection is made more accurate by dividing the RGB image into four quadrants and doing a hue-saturation histogram comparison for individual quadrants. The method is similar to an existing one [31] with the addition of quadrant processing, which aids in predicting the correct orientation of the object.

C. Temporal smoothing

A single query for object detection operates over a single frame of RGB-D data from a single perspective. Capturing multiple frames from the same perspective helps in mitigating effects of noisy sensor data or inconsistent pose estimates. In the implementation of the temporal smoothing enhancement, 12 frames of RGB-D images are aggregated, with the most frequently reported pose estimate reported on the final frame. Effectively, for all $P_i \in P_{pose\_estimations}$:

$$\underset{P_i}{\mathrm{argmax}}(Q(P_i) + \sum_{\forall\ P_j \in neigh(P_i)} Q(P_j)/dist(P_i, P_j))$$

where $Q(P_i)$ is a quality measure of pose estimation P_i, $dist(P_i, P_j)$ is a distance function between poses, and $neigh(P_i)$ returns all other pose estimates within a small neighborhood of P_i.
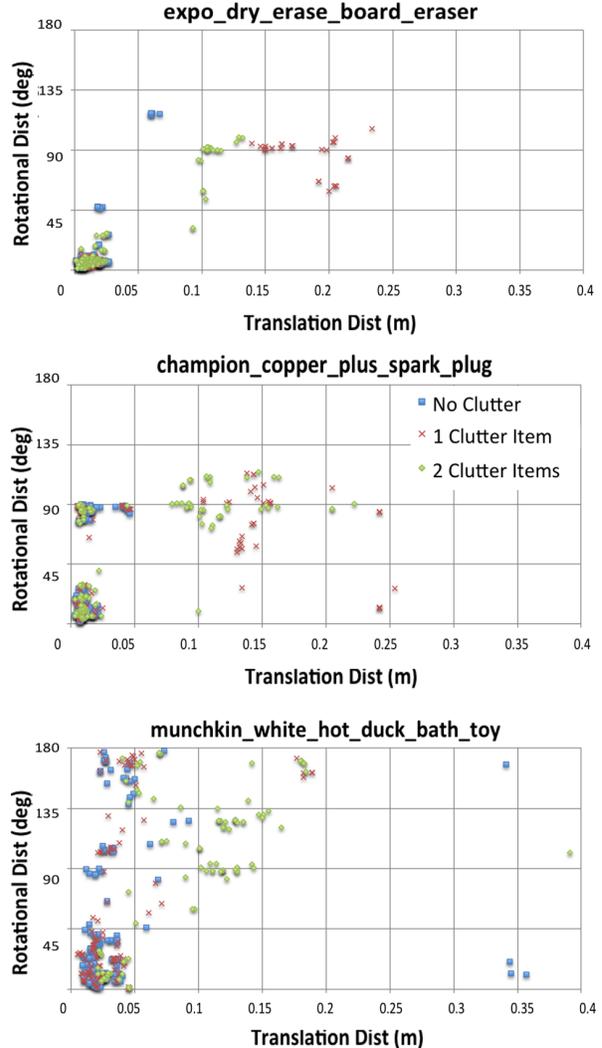

Fig. 4. Scatter plots of raw pose estimation accuracy results for three example objects from the APC dataset. X-axis is translational error (L2 dist) in meters, Y-axis is rotational error in degrees.

For objects where the likelihood of getting a good detection is low, temporal smoothing might bias the final detection towards bad pose estimations. Nevertheless, in this environment the positive effects of smoothing pose estimates outweigh the negatives on average.

V. FEATURES AND COMPARISON

One of the defining characteristics of this dataset is the amount of control put into isolating certain environmental factors in its collection. To exemplify the importance of these controls, an analysis over the example LINEMOD algorithm is performed over both the proposed dataset and the original LINEMOD dataset.
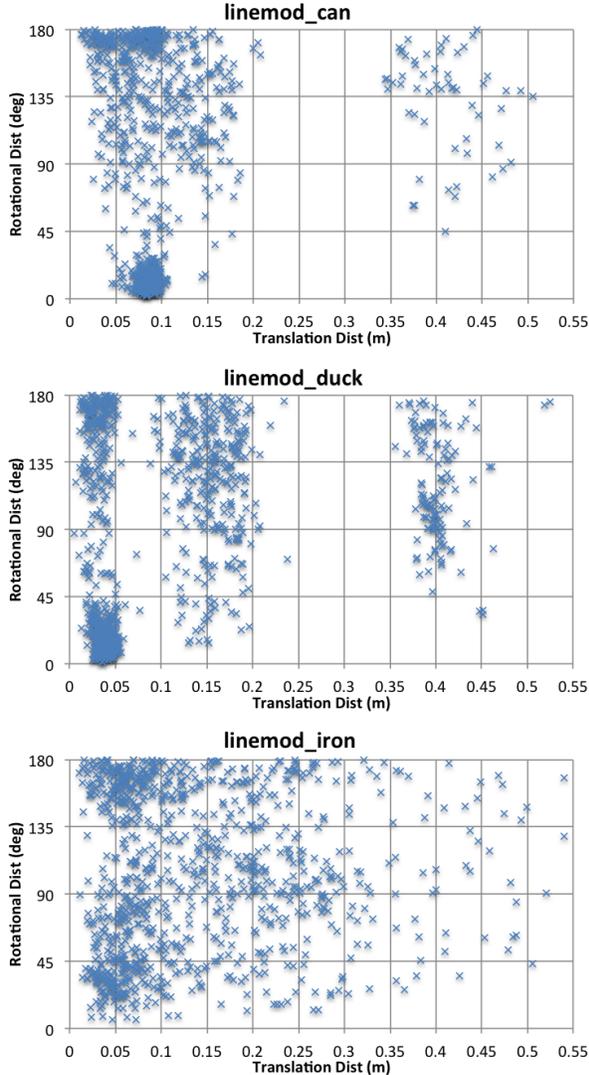
Fig. 5. Scatter plots of raw pose estimation accuracy results for three example objects from the LINEMOD dataset. Axes measure the same dimensions as the plots to the left.

A. Effects of Clutter

A situation known to cause difficulty for pose estimation algorithms, including state-of-the-art solutions, corresponds to the presence of significant clutter present in the scene containing the target object. For example, in a scene containing only the target object, simple segmentation techniques may provide reliable results. Adding, however, even a small number of other objects with vaguely similar colors or other visual features can easily cause simple approaches to fail. As such, it is a high priority goal for current solutions to 6D pose estimation problems to be as robust as possible to the presence of clutter.

To elaborate on the description of the control for clutter, for every distinct target object pose across each of the 12 bins, the dataset provides frames (i) with the object alone in the bin, (ii) accompanied by a single clutter item, and (iii) accompanied by two clutter items. By doing so, users of the dataset can directly compare the accuracy of different algorithms under these three different conditions.

Though the LINEMOD dataset shares similarities with the one proposed here (e.g., in terms of providing 6D ground truth, a variety of scenes and poses, and environments containing lots of clutter), it does not provide insights regarding the effects of clutter. In the comparison provided in this paper on Figures 4 and 5, these insights are exemplified. In the graph from Figure 4, which corresponds to the proposed dataset, a majority of the inaccurate pose estimates arise from scenes containing more clutter, but the overall effect is relatively small. In the plot from Figure 5, corresponding to LINEMOD, inaccurate pose estimates occur across all variations in clutter and are dominated by the translation error in cluttered scenes. This is a likely indication that the detection algorithm is confusing other objects for the object of interest and thus indicates a weaker detection strength for this target object. Specifically regarding the middle graph of Figure 5, there is a cluster of inaccurate pose estimates at the 90-degree rotational error rate. Due to the cuboid shape of the target object, it is likely that the algorithm often estimates an incorrect orientation of the object. Made possible by the controls placed in the data collection of the proposed dataset, these types of observations are valuable when making improvements to pose estimation algorithms or when comparing different approaches to suit a specific task. Additionally, given the inconsistent accuracy observed when evaluating the LINEMOD dataset, these insights would be extremely useful in this case.

B. Coverage and Variety of Poses

Another control prioritized when assembling the proposed dataset was to ensure a variety of ground truth poses for each object. At the same time, there was also the objective to choose poses such that they were representative of probable placements of objects in the APC competition. As such, all target placements are located several centimeters away from the front edge of the bin. The utility of the coverage characteristic is in allowing users to test their solutions when each of the major faces of the object is the primary viewable face. This allows to immediately identify troublesome surfaces and, given these insights, to design more robust solutions. Figure 6 illustrates this control for one example object.

Fig. 6. Simulated scene showing the variety of ground truth poses for one example object from the proposed dataset, as it is rotated through the 12 bins of the shelf.

C. Viewpoint Variety

Among the additional features of the proposed dataset is the accumulation of samples from several vantage points for each target object pose and clutter combination. Specifically, samples were collected for each configuration from three different viewpoints in front of the shelf: (i) left of, (ii) centered in front of, and (iii) right of the bin. Since the left and right positions may incur some level of occlusion of the target object by parts of the shelving unit, one of the applications of this feature is in the determination of the effects of these partial occlusions. Additionally, these samples can be used for pose hypothesis aggregation and smoothing, or for 3D reconstruction approaches.

D. Noisy Sensing

While Kinect v1 is an inexpensive and widely available sensor, a major detriment in its use is the noise inherent in RGBD samples produced using this equipment. To counteract this, for each configuration and camera position, the dataset provides four samples taken over a period of several seconds with all objects and hardware stationary. Similar to the above, this feature will allow users to easily determine which situations and target objects are robust to this noise and which are not.

E. Extensions

In addition to the above controls, an inherent feature of the dataset is that it can be used not only for single-object pose estimation, but also for multi-object. Because all transforms from object to robotic base are

stored in the ground truth pose files, users may easily extend this dataset to the multi-object case simply by reading in all ground truth poses of neighboring bins within the same "run", or configuration, of data collection. Since within a single "run", no item's placement is changed, this is straightforward to do. And because the dataset is organized by these runs, the implementation is rather easy. This feature makes the dataset a good candidate for testing 3D reconstruction techniques.

VI. DISCUSSION

This work contributes a large hand-annotated RGB-D dataset with 6DOF ground truth poses. The dataset is specifically designed to support advancing solutions for the problem of pose estimation in tight environments that appear in warehouse picking problems. The extent and structure of the dataset provides flexibility to researchers and allows them to use the data to apply and evaluate pose estimation methods using a variety of different techniques. The dataset is not only large relative to alternatives but is also designed to allow evaluation of several additional factors that can affect pose estimation accuracy. The accompanying software allows for improvements that are agnostic to the pose detection algorithm.

The evaluation of an easily available pose estimation algorithm to the robotics community over the proposed dataset emphasizes the difficulties that RGBD-based solutions face when dealing with transparent and reflective surfaces [32]. Cuboid objects also pose some difficulties for algorithms that are based primarily on RBG-D data but it was possible to deal with these issues through the improvements described in this work, which were tailored to fit the context of a warehouse environment and provide robustness.

The current dataset does not focus on the case of partially occluded objects, where a pose estimation process may be used to evaluate the pose of both the occluding and the occluded objects so as to assist rearrangement manipulation algorithms [33], [34]. Such problems can be potentially benefited by the utilization of cloud computation in order to improve performance and deal with the inferent uncertainty in the pose estimation and manipulation processes [35].

There is also an influx of new results in the area of machine learning that can potentially be applied for the problem of pose estimation for warehouse picking and it would be interesting to see the quality of such solutions given the available dataset. Similarly, more classical methods developed for monocular cameras that primarily depend upon color and texture may exhibit complementary behavior to the one displayed by the

considered RGB-D approach. Fusing such methods can also be another way of achieving solutions that operate robustly over a wide variety of object classes and environmental conditions. The dataset can be useful in the evaluation of solutions in the context of related applications, such as directly detecting a handle [36] or a grasp [37] from point cloud data.

REFERENCES

[1] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Lessons from the Amazon Picking Challenge," *http://arxiv.org/abs/1601.05484 - Submitted to IEEE Transactions on Automation*, 2016.

[2] T. Wisspeintner, T. Van Der Zant, L. Iocchi, and S. Schiffer, "RoboCup@ Home: Scientific competition and benchmarking for domestic service robots," *Interaction Studies*, vol. 10, no. 3, pp. 392–426, 2009.

[3] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lapetit, "Gradient Response Maps for Real-time Detection of Texture-Less Objects," *IEEE TPAMI*, 2012.

[4] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, vol. 25, no. 11, pp. 120–126, 2000.

[5] J. Stuckler, D. Holz, and S. Behnke, "Demonstrating Everyday Manipulation Skills in RoboCup@ Home," *IEEE Robotics and Automation Magazine*, pp. 34–42, 2012.

[6] A. Krizhevsky and G. Hinton, "CIFAR Dataset - Learning multiple layers of features from tiny images," 2009.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, pp. 1–42, April 2015.

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[9] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3D object dataset: Putting the kinect to work," in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 141–165.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *ECCV*, 2012.

[11] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols," *arXiv preprint arXiv:1502.03143*, 2015.

[12] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese, "Depth-encoded hough voting for joint object detection and shape recovery," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., 2010, pp. 658–671.

[13] N. Vaskevicius, K. Pathak, A. Ichim, and A. Birk, "The Jacobs Robotics Approach to Object Recognition and Localization in the context of the ICRA'11 Solutions in Perception Challenge," in *IEEE ICRA*, 2012, pp. 3475–3481.

[14] K. Lai, L. Bo, X. Ren, and D. Fox, "A Large-Scale Hierarchical Multi-View RGB-D Object Dataset," in *IEEE ICRA*, 2011, pp. 1817–1824.

[15] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model-based Training, Detection and Pose Estimation of Texture-less 3D Objects in Heavily Cluttered Scenes," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 548–562.

[16] K. Narayan, J. Sha, A. Singh, and P. Abbeel, "Range Sensor and Silhouette Fusion for High-Quality 3D Scanning," in *ICRA*, 2015.

[17] L. Ma, M. Ghafarianzadeh, D. Coleman, N. Correll, and G. Sibley, "Simultaneous Localization, Mapping and Manipulation for Unsupervised Object Discovery," in *IEEE ICRA*, 2015.

[18] L. Ma and G. Sibley, "Unsupervised Dense Object Discovery, Detection, Tracking and Reconstruction," in *ECCV*, 2014, pp. 80–95.

[19] C. Choi and H. I. Cristensen, "3D Pose Estimation of Daily Objects Using an RGB-D Camera," in *IEEE/RSJ IROS*, Moura, Algarve, Portugal, 2012.

[20] C. Choi, Y. Taguchi, O. Tuzel, M. Liu, and S. Ramalingam, "Voting-based Pose Estimation for Robotic Assembly using a 3D Sensor," in *IEEE ICRA*, 2012.

[21] K. Lai, L. Bo, X. Ren, and D. Fox, "A Scalable Tree-based Approach for Joint Object and Pose Recognition," in *AAAI Conference on Artificial Intelligence*, 2011.

[22] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, and G. Bradski, "CAD-model recognition and 6DOF pose estimation using 3D cues," in *ICCV Workshops*, 2011, pp. 585–592.

[23] M. Pham, O. Woodford, F. Perbet, A. Maki, B. Stenger, and R. Cipolla, "A New Distance for Scale-invariant 3D Shape Recognition and Registration," in *IEEE ICCV*, 2011.

[24] M. Muja, R. B. Rusu, G. Bradski, and D. G. Lowe, "REIN - A fast, robust, scalable reconigtion infrastructure," in *IEEE ICRA*, Shanghai, China, 2011.

[25] E. Kim and G. Medioni, "3D Object Recognition in Range Images using Visibility Context," in *IEEE/RSJ IROS*, 2011, pp. 3800–2807.

[26] K. K. B. Steder, R. B. Rusu, and W. Burgard, "NARF: 3D Range Image Features for Object Recognition," in *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics (IEEE/RSJ IROS)*, 2010.

[27] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model Globally, Match Locally: Efficient and Robust 3D Object Recognition," in *IEEE CVPR*, 2010.

[28] R. Triebel, J. Shin, and R. Siegwart, "Segmentation and Unsupervised Part-based Discovery of Repetive Objects," in *Robotics: Science and Systems*, 2010.

[29] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson, "Object Recognition and Full Pose Registration from a Single Image for Robotic Manipulation," in *IEEE ICRA*, 2009, pp. 48–55.

[30] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "ROS: an open-source Robot Operating System," in *ICRA Workshop on Open Source Software*, vol. 3, no. 3.2, 2009, p. 5.

[31] M. D. Ozturk, M. Ersen, M. Kapotoglu, C. Koc, S. Sariel-Talay, and H. Yalcin, "Scene Interpretation for Self-Aware Cognitive robots," in *AAAI-14 Spring Symposium on Qualitative Representations for Robots*, 2014.

[32] I. Lysenkov, V. Eruhimov, and G. Bradski, "Recognition and Pose Estimation of Rigid Transparent Objects with a Kinect Sensor," in *Robotics: Science and Systems*, 2012.

[33] A. Krontiris and K. E. Bekris, "Dealing with Difficult Instances of Object Rearrangement," in *Robotics: Science and Systems (RSS)*, 2015.

[34] ——, "A Hierarchical Scheme with Efficient Primitives for Hard Rearrangement Problems," *International Journal of Robotics Research (IJRR)*, 2016 (invited).

[35] K. E. Bekris, R. Shome, A. Krontiris, and A. Dobson, "Cloud Automation: Precomputing Roadmaps for Flexible Manipulation," *IEEE Robotics and Automation Magazine*, 2015.

[36] A. Ten Pas and R. Platt, "Localizing Handle-like Grasp Affordances in 3D Point Clouds," in *International Symposium on Experimental Robotics (ISER)*, 2014.

[37] ——, "Using Geometry to Detect Grasps in 3D Point Clouds," arXiv:1501.03100v3, Tech. Rep., January 15 2015.