SRIP2019 drone proposal

Po Hsiang Huang, Sean Kamano, Brandon Leung, John Ho
UCSD

{phh003, skamano, b7leung, chh279}@.ucsd.edu

1. Goals and plans

The drone group consists of 5 undergraduate students (2 SRIP and 3 non-SRIP), 2 high school students and 2 mentors. With such a large group, the goal for our group has been divided into several partitions. The main goal for this summer is to accomplish the OOWL3D dataset collection. The detail of this dataset can be found in Sec. 2. In parallel to the dataset collection, drone group has a weekly paper reading group[1] focusing on 3D recognition, reconstruction and registration. We hope that by organizing the reading group, all the group members will have a deeper understanding about the 3D related literature. Meanwhile, some of the members who are more familiar with machine learning package (e.g. Pytorch) have started reproducing the code available online. Hopefully, when the dataset is collected, we can test those off-the-shelf 3D models on our dataset.

Aside from the dataset collection, some members are also working on the Intel drone. Since the Intel drone have not been fully understood after it was purchased, this is critical for the future development of the drone project. More details about the drone development can be found in Sec. 3.

Overall, by the end of this summer, there are 3 goals, 1) OOWL3D collection, 2) drone preliminary development, 3) designing tasks for OOWL3D by paper reading. We hope that OOWL3D dataset and drone preliminary development could help this project in the long run.

2. OOWL3D Dataset collection

In this section we discuss the importance of the proposed large scale 3D real world object centric dataset, namely OOWL3D. The current design of the dataset collection are also presented.

2.1. Motivation

The development of deep learning on 3D world has progressed rapidly in recent years and particularly due to the availability of commercial RGB-D and LIDAR sensors, novel architectures and the collection of 3D datasets. Similar to the importance of ImageNet [8] to deep learning in 2D cases, several important 3D datasets have contributed significantly to the evolution of 3D deep learning. ModelNet [24] and ShapeNet [5] are one of the large scale 3D synthetic datasets that are widely used in the study of 3D representations, including voxel, point cloud and multiview. Similar to the organization if ImageNet, both ModelNet and ShapeNet models are categorized into 40 or more classes. The introduction of those 2 datasets raise the interests of 3D object classification and retrieval tasks and the development in 3D deep learning architectures.

However, models in either ModelNet or ShapeNet are textureless and synthetic generated, which is far from the realist models obtained in the real world. Several datasets are aggregated to overcome this issue by either scanning objects in the real world[20, 2, 17, 14, 16, 4, 15, 23, 6, 19] or extracting 3d partial models from the scanned scenes[7, 1, 12, 21, 3]. Unfortunately, both solution have its own drawback. For the dataset composed of real world object scan, the quality of the dataset [6] are usually poor and small[20, 15], compared to the size of synthetic generated datasets. Those dataset are usually designed for specific task[16, 19] and organized in instance level [23, 20, 14] instead of class level. Moreover, while both ModelNet and ShapeNet allows the study of different 3D representations (voxel, point cloud and multiview), these datasets [6, 16] are usually presented in one of the modalities, which hampers the study of 3D representation.

Another solution is to present datasets [7, 1, 12, 21, 3] leveraging the 3D scanning of numerous scenes. As each scene contains several objects, part of the 3D scene scanning are cropped and used for 3D object dataset, which is a byproduct of the scene dataset. Similar to the drawback above, it is difficult to understand the object itself by looking at the cropped part as not all view points of the object

[1] The schedule for the reading group and be found in https://docs.google.com/document/d/1fMN2NtKWkncfYBYu1Y1DpF4QAmSKj-WSsSdj71Klfek/edit

are covered during scanning. Moreover, the cropped object models are usually paired with an existing 3D synthetic CAD model from either ModelNet and ShapeNet, as no ground truth of the specific object is provided in this type of datasets.

We think this is an incorrect approach to further understand the robust representation of a 3D real world object and a large scale real world object centric dataset is needed. Several properties are needed in this dataset. First, the dataset has to be organized into different categories with several objects in each category. This is different from those instance level datasets, which does not allow the study of 3D object classification and class retrieval. Second, dense view point coverage of the object is required. We are the first to proposed a real world large scale object-centric dataset that contains several modalities. While graphically generated CAD models can be converted into different representations, most of the real world scanned models are presented in the final 3D CAD model. Of course, one can convert a CAD model in to point cloud, voxel or mutliview representations, but these converted representations are based on the CAD model that might be computed from sparse view points, indicating the converted representations have already lose the information of the object. As a result, our dataset contains dense view point coverage of an object and the camera parameters are provided as the ground truth for where the image is taken. Finally, our dataset is object-centric, which is more similar to the style ModelNet and ShapeNet. These type of dataset allows the study of object itself, including the analysis of the effect of different poses and different representations. The fundamental study of different type of object allows the development of task such as robot grasping [9], object retrieval and feature disentanglement. Although similar to the style of ModelNet and ShapeNet and other 3D synthetic dataset [25, 22, 11], our dataset are all composed of real world object scan, including texture and appearance. The dataset is also designed to have high overlapping classes between ShapeNet, ModelNet and widely used 2D dataset ImageNet.

2.2. Setup

We have been collaborating with the UCSD Digital Media Lab (DML) and a PhD student in Professor Nguyen's lab. Currently, we prefer to work with DML as their setup, as shown in Figure 1, has perfectly fit our requirement. The preliminary reconstruction results from DML and the PhD student are illustrated in Figure 2. It can be observed that the reconstruction quality of the CAD model from DML is better then that from the PhD student.

3. Drone development

To facilitate efficient collection of data, we intend to use an Intel Ready to Fly Drone in order to circle around an object and incrementally take photographs or record a video of the object. The drone is equipped with an Intel Aero Compute Board running ROS and Ubuntu 18. A ground control station running the software QGroundControl will handle RSTP/UDP communications with the drone; the plan is to implement this communication logic with ROS, since ROS has extensive support for communication with the PX4 flight controller through MAVROS.

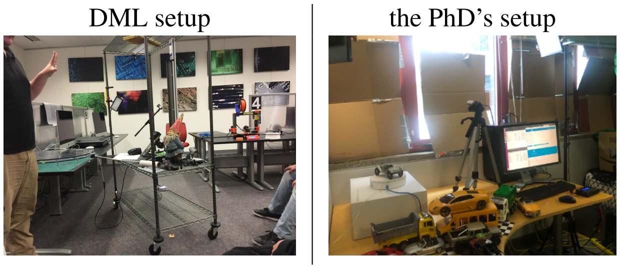

Figure 1. Different setups for dataset collection

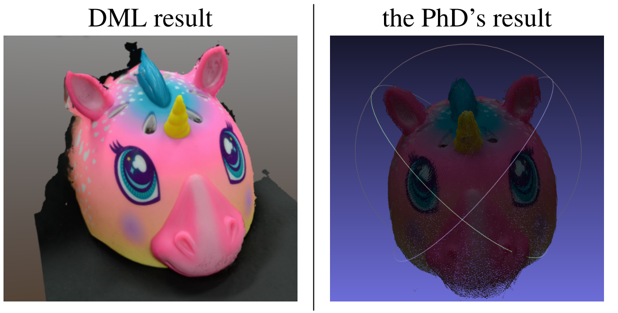

Figure 2. Reconstructed results for different setup. The CAD model provided by the PhD has only 961,923 vertices, while the CAD model from DML contains 106.3k. The reconstructed result from DML can be found in http://www.svcl.ucsd.edu/projects/OOWL/oowl3d/

Prior to conducting tests on the drone itself, we intend to run tests for our control algorithms using Gazebo, which interfaces with both ROS and the PX4 flight controller. Gazebo is a 3D simulation environment in which we can safely test more experimental control schemes.

3.1. Navigation & Object Detection

In order to achieve a good representation of how the coordinate frame of the drone is oriented and positioned with respect to the coordinate frame centered at the origin, we will need to use PX4's support for visual inertial odometery, which will give us the intuition shown in Figure 3. We intend to use this information to help the drone adjust its position and orientation so that it can reliably take pictures with the object centered and stabilized in the frame.

The drone will record video of the desired object following a polygonal instead of circular path. This is due to the observation that the roll and yaw of the quadcopter is dependent on interaction of the different strength of each propeller. We are unable to control the drone to consistently stay focused on the object while orbiting without tilting. Therefore we will control the roll and yaw of the drone

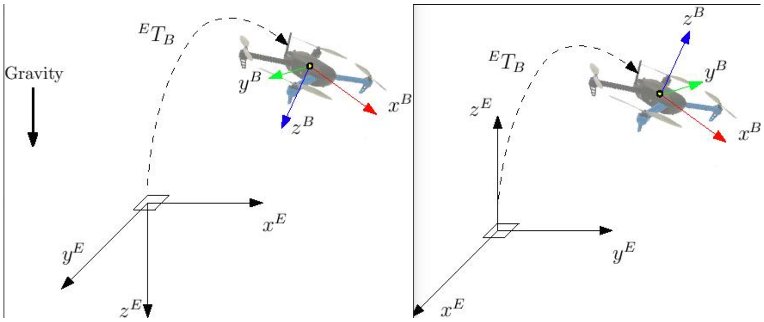

Figure 3. The origin frame to the left corresponds to the PX4 representation which uses the NED (x North, y East, z Down) convention, and the origin frame to the right corresponds to the ROS representation which uses the ENU (x East, y North, z Up) convention

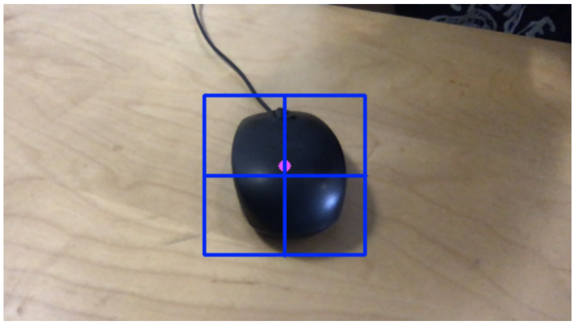

Figure 4. The blue bounding box and the pink center point are used as reference for the drone to navigate around the object.

separately and thus, the drone will fly along a polygonal pathway.

We are using object trackers to manually set a ROI bounding box around the object. Out of the 8 built-in object trackers in OpenCV, we are using the CSRT tracker (Discriminative Correlation Filter with Channel and Spatial Reliability)[18] for better video accuracy at the cost of lower frame rate (30 FPS). The bounding box follows the object as either the object or the camera moves. We use this bounding box and the center point of the video stream as seen in Figure 4as references for the drone to navigate. If the center point is within the selected ROI, the drone will constantly move along the tangent line with respect to the circle formed by the object and the starting point; if the center point moves outside of the selected ROI, the drone will then turn towards the object to recenter the region of interest, and the process repeats.

An alternative object tracker we will look into is the GOTURN tracker (Generic Object Tracking Using Regression Networks)[10] which utilizes CNN to be much more accurate in regards to ”viewpoint changes, lighting changes, and deformation”[2]. While other available object trackers including the CSRT tracker learn the objects’ appearance at runtime, the GOTURN tracker is pre-trained and does not need any learning at runtime. It allows the bounding box for the ROI to be accurate and adjustable to a high extent, closely adhering to the size of the object as well as the distance between the object and the camera. However, for simplicity reasons, we are working with CSRT for now. We will possibly switch to using the GOTURN tracker for objects with impeding factors (irregular shape, texture shine, deformation... etc) or future outdoor collections where the environment is a lot more variant, after we finish our initial indoor tests.

Furthermore, we would like to capture view angles of objects. This will be calculated from the coordinates of the drone relative to its starting point in space given by the IMU of the drone[13]. The result can then be used for further experiments involving image registration, shape recognition and partial reconstruction.

[2] OpenCV GOTURN Documentation `https://docs.opencv.org/3.3.1/d7/d4c/classcv_1_1TrackerGOTURN.html`

References

[1] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, Feb. 2017. 1

[2] Stylianos Asteriadis, Petros Daras, Alexandros Doumanoglou, Dimitrios Alexiadis, and Dimitrios Zarpalas. A dataset of kinect-based 3d scans. 06 2013. 1

[3] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Niessner. Scan2cad: Learning cad model alignment in rgb-d scans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[4] B. Browatzki, J. Fischer, B. Graf, H. H. Blthoff, and C. Wallraven. Going into depth: Evaluating 2d and 3d cues for object classification on a new, large-scale object dataset. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1189–1195, Nov 2011. 1

[5] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 1

[6] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *CoRR*, abs/1602.02481, 2016. 1

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[9] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. *CoRR*, abs/1803.11469, 2018. 2

[10] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 FPS with deep regression networks. *CoRR*, abs/1604.01802, 2016. 3

[11] Tomas Hodan, Pavel Haluza, Stepán Obdrzálek, Jiri Matas, Manolis I. A. Lourakis, and Xenophon Zabulis. T-LESS: an RGB-D dataset for 6d pose estimation of texture-less objects. *CoRR*, abs/1701.05498, 2017. 2

[12] Binh-Son Hua, Quang-Trung Truong, Minh-Khoi Tran, Quang-Hieu Pham, Asako Kanezaki, Tang Lee, HungYueh Chiang, Winston Hsu, Bo Li, Yijuan Lu, Henry Johan, Shoki Tashiro, Masaki Aono, Minh-Triet Tran, Viet-Khoi Pham, Hai-Dang Nguyen, Vinh-Tiep Nguyen, Quang-Thang Tran, Thuyen V. Phan, Bao Truong, Minh N. Do, Anh-Duc Duong, Lap-Fai Yu, Duc Thanh Nguyen, and Sai-Kit Yeung. Shrec'17: Rgb-d to cad retrieval with objectnn dataset, 2017. 1

[13] Manon Kok, Jeroen D. Hol, and Thomas B. Schn. Using inertial sensors for position and orientation estimation. *Foundations and Trends in Signal Processing*, 11(1-2):1–153, 2017. 3

[14] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824, May 2011. 1

[15] B Leibe and B Schiele. Analyzing appearance and contour based methods for object categorization. volume 2, pages II–409, 07 2003. 1

[16] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing IKEA Objects: Fine Pose Estimation. *ICCV*, 2013. 1

[17] Anan Liu, Zhongyang Wang, Weizhi Nie, and Yuting Su. Graph-based characteristic view set extraction and matching for 3d model retrieval. *Information Sciences*, 320:429 – 442, 2015. 1

[18] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[19] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 778–785, June 2009. 1

[20] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 509–516, May 2014. 1

[21] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, June 2015. 1

[22] Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B. Smith, and James M. Rehg. Incremental object learning from contiguous views. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[23] F. Viksten, P. Forssen, B. Johansson, and A. Moe. Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. In *2009 IEEE International Conference on Robotics and Automation*, pages 2779–2786, May 2009. 1

[24] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920. IEEE Computer Society, 2015. 1

[25] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10, 000 3d-printing models. *CoRR*, abs/1605.04797, 2016. 2