Design and Development of an Automated System for creation of Image Datasets intended to Allow Object Identification and Grasping by Service Robots

Marlon de Oliveira Vaz[1], João Alberto Fabro[2] and André Schneider de Oliveira[3]

Abstract— **This work presents the design and construction of an operational prototype that allows the creation of datasets that allows service robots to identify and grasp household objects. The prototype was designed and built with several movable and removable parts. Unlike similar systems described in the literature, this structure can be disassembled and transported. Due to the lack of specific datasets designed to allow detection, recognition and grasping tasks for service robots, it was decided to build a structure for data collection in competitive environments, such as RoboCup@Home. With the prototype ready, a first image capture test was performed and 251 images were obtained for a single object. After the necessary adjustments, the prototype is expected to produce automatically 3600 images from various points of view for each object of interest without human interference.**

I. INTRODUCTION

Competitions such as RoboCup@Home [1] and Amazon Picking Challenge (APC) [2] were created with the premise of improving the performance of service robots in daily tasks, by providing a set of features, such as automating the storage process, searching for products in a warehouse, finding a certain product on the environment, or picking up a certain product from a table or shelf inside a house [3]. This type of action requires the integration of the robot with the environment through a predetermined set of skills and technologies with a high level of abstraction. This makes the robot act in the real world and be prepared to identify and recognize both the region in which it will operate, and the objects it must manipulate in this kind of environment [1].

Object manipulation is not a trivial activity, since in a residential environment there are several elements, often grouped and with different characteristics that interfere with the recognition process [4]. Although the human action of manipulating a specific object is very simple, for a robot it is a difficult task because it depends on several components. In this case, accuracy in localizing is a critical factor for the success or failure of the robotic arm in picking-up a specific object [5]. The large number of variables found in these environments, such as variations in object position, lighting, and distance, encourage research in this area, thus searching for a higher level of autonomy for the robot [6].

For an autonomous robot to grasp an arbitrarily positioned object, it requires a great deal of learning capability, coupled with high quality object models so that they can be detected, identified, have their position estimated, and only then manipulated [7]. Sets of object models are called datasets, and are powerful tools when applied in conjunction with recognition algorithms (such as neural networks) for the recognition task. Usually this datasets consist of images or videos, accompanied by information about the exact position of objects inside the images (ground truth information). In recent years several works have been presented with the theme of building datasets [2][5][6][7].

Detection and recognition usually involves determining the position of the object in the scene and categorizing it by a name. Object recognition is the task of identifying objects in an image. This information has become crucial with the advent of robotic systems in picking up objects in unstructured environments [8].

A good image dataset should have a large variety of information about the experimental conditions, such as the objects present in the images, their exact positions, lighting, and perspective. These informations are important regarding the accuracy of the data collected. Most datasets are built to solve specific problems [5][6][7][9]. When it comes to dataset for use in competitions like RoboCup@Home, especially aimed at grasping and picking-up challenges, most of the images in the cited datasets are not well suited, because they were not built specificaly for this purpose.

The main focus of this paper is to present an automated prototype to capture a set of images of household object, from different angles and in several positions and lighting conditions . The images collected will be used to define and structure an image dataset and will later be made available to the scientific community for use in competition such as RoboCup@Home. This dataset will be used in the future by the robot to perform tasks such as grasping objects. The grasping task is the action taken by a robotic hand to grip/lift any object.

The next section present related work that guided the development of this project. Section 3 describes the 3D modeling of the proposed project and explains how the prototype was built. Section 4 shows the first capture of images of a single object, that validates the prototype, while

[1]M. O. Vaz is a professor at Federal Institute of Parana -IFPR. marlon.vaz@ifpr.edu.br

[2]J. A. Fabro is a professor with the Graduate Program on Applied Computing - PPGCA, at the Informatics Department - DAINF - at UTFPR - Federal University of Technology, Parana, Brazil - Campus Curitiba. fabro@utfpr.edu.br

[3]A. S. de Oliveira is a professor with CPGEI - Graduate Program of Electrical Engineering and Industrial Informatics, UTFPR, Brazil - Campus Curitiba. andreoliveira@utfpr.edu.br

section 5 presents the conclusions and future work.

II. Related Works

In recent decades, with the evolution of machine learning and especially with convolutional neural networks, several researchers have focused their efforts on building datasets of 3D and 2D images with large amounts of data [10]. By incorporating this data into learning methods, the automatic classification of objects in images becomes possible.

Researchers such as Garcia-Garcia et al. [8] listed 11 (eleven) datasets with different formats to prove the quality and performance of their dataset, while Ruiz-Sacramento, Alcindor and Gonzalez-Jimenez [9] listed another 13 (thirteen). Similarly, Georgakis et al. [4] did a more focused research and used only two datasets in their project. In these works, the line of action was to generate a large amount of data automatically or semi-automatically. However, each dataset has different characteristics that may be useful for performing a particular task but not for another. This need leads researchers to build new datasets with different information and formats for each specific application [11].

When creating a data set, especially with images, whether 3D or 2D, it is useful to have an environment that enables data collection with high quality. In this sense, Jasper, Xe and Mailman [7] describe a semi-automatic system for capturing 3D and 2D images that was created to enable object tracking and detection, and to plan a way to pick up objects by service robots. For the capture of the images, a sled base was developed where a pair of cameras was fixed. Object capture by different positions occurs by moving the sled over a rail, starting from the front of the object and ending directly over the object. A 3D digitizer, a turnable plate and a dimmable lighting system were employed in the system. With this system, images from 134 objects of general use were captured. Of these 134, 21 objects where selected specifically for the robot to be able to pick up. For each object, 360 images were generated, taking into consideration the lighting and the position of the camera.

Garcia-Garcia et al. [8] present another system for image collection. In this case, the structure used is simpler, having a table as a base structure and in its center a swivel base controlled by an Arduino device. The camera is positioned using a tripod in front of the object at an angle of 30 degrees. In this case, the position of the camera relative to the object is fixed. To control and spread the luminosity over the object, light projection techniques were used, with white tissues, which were positioned around the structure.

In the work of Martinez-Martin and Pobil [6] a set of techniques for creating a dataset with articulated objects is proposed. Articulated objects have movement, such as a drawer that opens and closes. The idea of this dataset came from the need to study interaction with moving objects. Data such as force and torque were captured using a sensor tool to ensure that the robot can pick up the object. In that work, 14 object models with articulated mechanisms were used. To define the movement of an object, 380 images were used.

The difference between the presented methods and the method proposed in this paper is in the positioning of the camera in relation to the object. The proposed method allows the collection of images with different distances from the camera to the object, since mobile robots such as those used in Robocup@Home, approach the objects from different distances, unlike the proposals [7] and [8], where the cameras have a fixed distance from the object.

III. The Prototype

Based on the information obtained from several of the works in the previous section, we started the procedures for the definition of a low cost system that will allow to capture 2D images and map 3D objects. This project is being developed in the Laboratory of Embedded Systems and Robotics (LASER) of UTFPR-Federal University of Technology-Parana - Campus Curitiba. The materials used in this project include: 4 servo motors, 1 linear actuator and the Arduino^{TM}board and shields. Based on these equipment, it was possible to start the modeling part of the Prototype.

A. The 3D Model

The development of a 3D model with a computer aided design (CAD) tool allows the creation of virtual models to guide the construction of a real prototype. In this step, the SketchUp^{TM}tool [12] was used.

The model designed in this work used as reference the systems developed by Kasper, Xue and Dillmann [7], Garcia-Garcia et al. [8], Martinez-Martin and Pobil [6] and Hodan et al. [13]. However, all these works consist of fixed structures, which makes the transportation impossible, thus preventing data collection in competitions. This information guided the creation of a more compact system that enables its transportation.

Figure 1 shows the final 3D model that was designed. The model has three motors: one servo motor to rotate the upper base, one servo to move the table (adjusting the distance from the camera to the object) and a linear actuator to move the camera base up and down, adjusting the viewing angle of the object. A more detailed explanation of the positions and function of each motor is present in the next sub-sections.

B. Movable Parts of the 3D Model

In the upper base (or turnable base) is where the object of interest is positioned, in order to start the capture of images. Figure 2 shows the turnable base of the 3D model. This element was projected to rotate 360 degrees, controlled by a stepper motor. Each step displace the base 10 degrees, requiring 36 steps to complete a full spin, thus allowing to capture 36 images by object. This amount of images are important due to the necessity to be able to identify/grasp the object that is arbitrarily positioned in the environment, according to Robocup@Home rules.

The table was designed to be the support structure of the turnable base and to move linearly (Figure 3 shows the possible movements - red arrows). In this context, the model was idealized so that the object can be moved away from

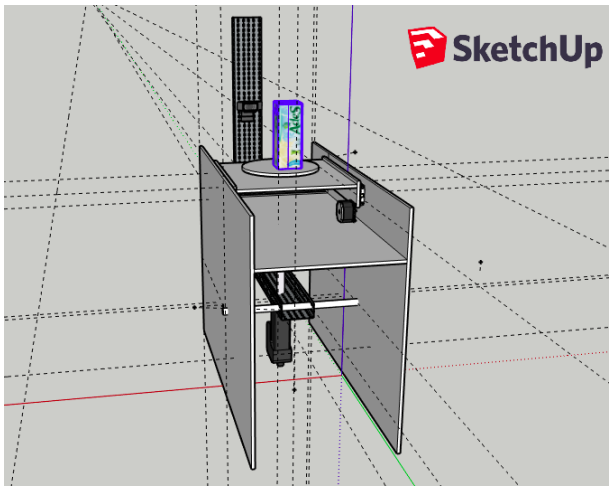


Fig. 1. Prototype drawn with SketchUp

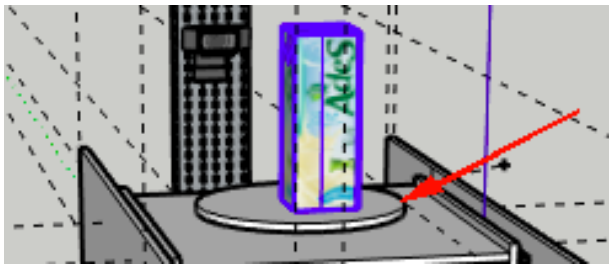


Fig. 2. Turnable base

the camera. This forward and backward motion of the table is achieved by a stepper motor. Two telescopic slides were used to fix the table in the structure and enable its horizontal movement. The maximum extension of the telescopic slides is 25 cm.

The systems presented in the articles [7][8][6][13], don't have the linear movement for approaching or moving away from the object. The movement of the table was designed to take images at different distances, emulating the approximation of the robot to the object in the real world. Since different robots have diverse manipulator arm sizes, it would be important to be able to identify the position of objects from different distances and points-of-view.

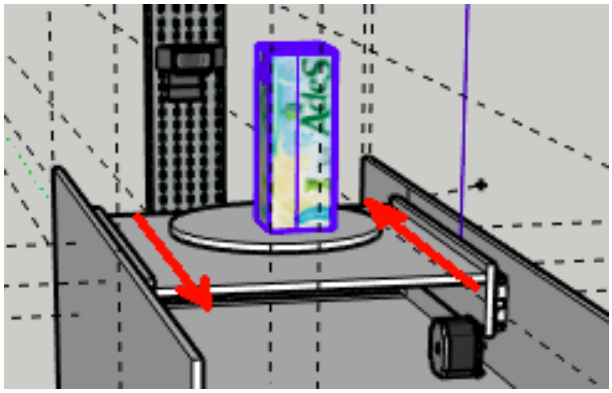


Fig. 3. Table with forward and backward motion

Some of the cited authors propose camera movements, or from the bottom up, or in angular movements [7][13], others keep the camera fixed [8][13][6]. In this work, the camera has an angular movement, allowed by a linear actuator. The proposed structure produces an angular movement of the camera, which allows the identification of the object from different angles, emulating different camera angles that are possible during the approximation of the service robot to the object to be identified/manipulated.

Figure 4 shows the possible movements generated by the camera base. The camera is positioned in the top of the blue structure, directed towards the rotating base over the moving table.

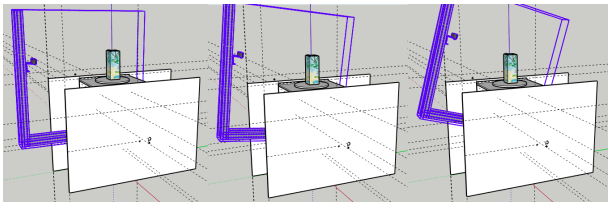


Fig. 4. Camera base movement prediction

C. Building the Prototype

After the definition of the 3D model, the construction of the prototype began. As pointed out earlier, in order to reduce costs and also enable the easy transportation of the structure, the entire structure is built in plywood. All moving parts are easily removable, making it possible the transport of the entire system simplified (in order to execute capture of images of objects during the robotics events). This equipment can be used to extend the dataset with new objects presented during the competitions.

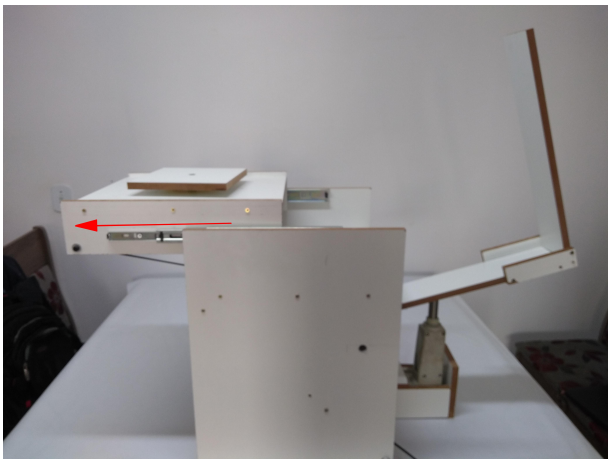


Fig. 5. Maximum forward table

Figures 5 and 6 show the completely built prototype. In Figure 5 the table is fully extended, while in Figure 6 the table is retracted to its point of origin. In both figures, red arrows indicates the table displacement like in the Figure 3. Figure 7 show other view the prototype, in (a) show

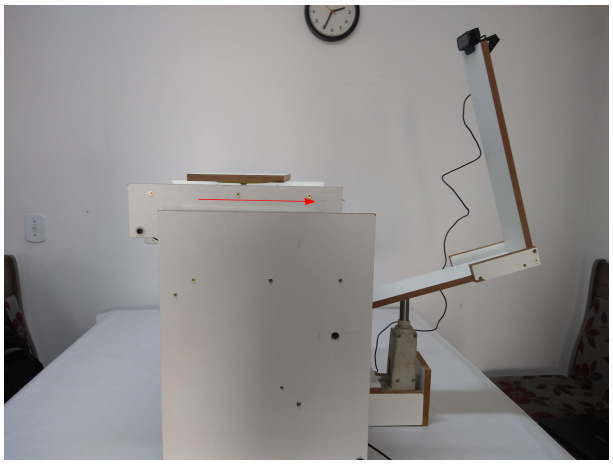
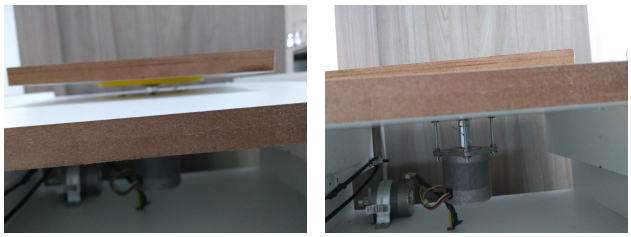


Fig. 6. Maximum backward table

(a) Prototype frontal view (b) Prototype orthogonal view

Fig. 7. Complete Prototype

the frontal view and (b) the orthogonal view. The turnable base it's on the table and this is connected in stepper motor. This motor is fixed in the table, which give stability in the movement of the turnable base. This can be seen in Figure 8.



(a) Turnable base (b) Stepper motor fixed in the table

Fig. 8. Union of the turnable base to the table.

Similarly, the table is connected to the main structure by pair of telescopic slides. Table movement is controlled by a second stepper motor that is fixed to the base of the structure. The movement follows the principle of 3D printers with the use of 6mm GT2 belt (Figure 9).

A linear actuator is fixed in structure by a circular aluminum profile and and is used to move the camera base. Your maximum extension is 12 cm, but is only 10 cm are used, because the actuator need 2 cm to keep the camera aligned with the object. Like the other elements of this system, the actuator can be easily removed and replaced and a box was additionally built to maintain its stability. This box was not

Fig. 9. Steeper motor fixed in the structure.

previewed in the 3D model. Figure 10 shows the actuator connected to the structure by a circular aluminum profile.

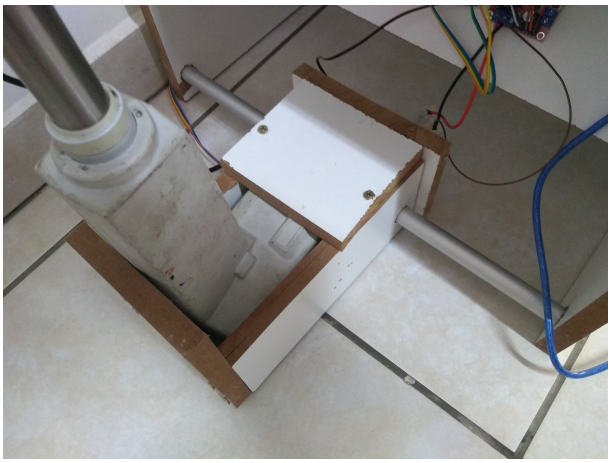


Fig. 10. Linear actuator

D. Motors controller

All motors are controlled by an Arduino Mega 2560, which is fixed to the structure, as shown in Figure 11.

The programming consists the three steps: the first step controls the rotation of the turnable base. This action rotates the object 36 times, 10 degrees in each step. Initially, the rotation was executed one degree at a time, in 360 steps, but many images had almost no difference, and thus the capture was simplified to just 36 images.

After completing the 36-step cycle of the turnable base, the control is transferred to the table. The table is initially fully retracted (see Figure 5(b)). The telescopic slide has a maximum extension of up to 25 cm, but to allow for a more secure control, a maximum extension of 20 cm is used, with a step of 1 cm for each cycle of the turnable base. For each position of the table base, a complete rotation of the object is executed, collecting 36 images in each of the 20 different distances.

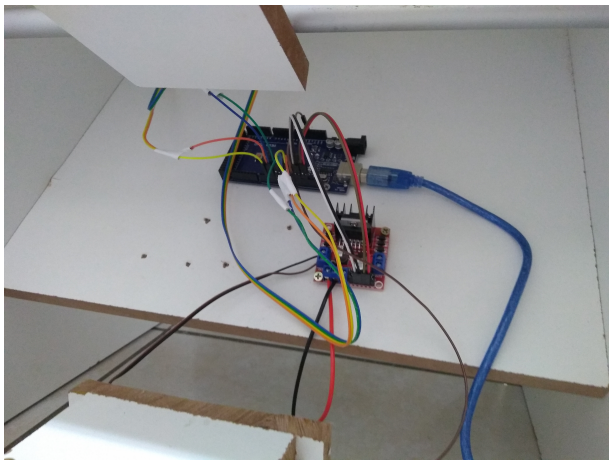


Fig. 11. Arduino Mega 2560 fixed into the structure.

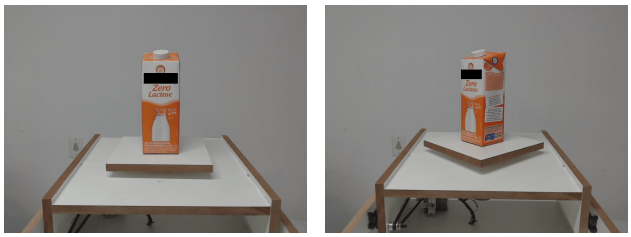
After completing the 20 steps of the table, the linear actuator is used to move the camera base in a circular motion. The actuator in its original state keeps the base upright with the frame. This enables the camera view to be directed towards the center of the object, as seen in Figure 1. To keep the camera in the initial position, the actuator was extended in 2 cm. At each step the actuator lifts the camera 1 cm, with maximum height of 10 cm. So, at the end, 36*20*10(7200) images of the same object are obtained.

IV. Image Capture

To capture the images, a Logitech C920 webcam was used. This camera accepts a maximum resolution of 1920x1080 pixels, however a resolution of 1024x768 was used in the tests. This resolution reduced the image size, but also reduced image quality. Figure 5(b) shows the camera positioning in the proposed system.

The capture of the images in this first phase was made using the Cheese tool, available on Ubuntu 18.04 operating system. This feature was used to validate the functionality of the prototype in relation to the movements of all moving parts. In this first evaluation, only 251 images of a single object were collected. However, the system allows to collect up to 7,200 images of each object. Based on the images collected, the developed prototype allows the creation of a dataset for use in robotics competitions. Each of the 7,200 different images reproduce various points of view for each object of interest. This entire procedure is done without human interference.

Figure 12 and Figure 13 presents 8 example images collected with this system.

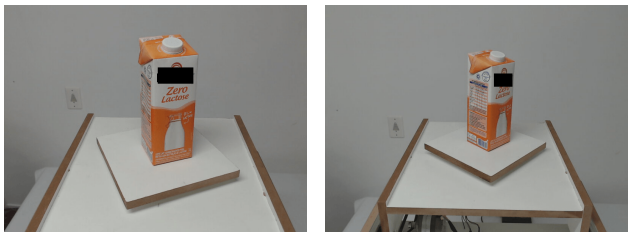


(a) Initial position to capture first image (b) Forward table offset of 1 cm

(c) Forward table offset of 2 cm (d) Forward table offset of 3 cm

Fig. 12. Images captured by forward table displacement

(a) Camera base original position (b) Camera base displacement 2 cm



(c) Camera base displacement 3 cm (d) Camera base displacement 4 cm

Fig. 13. Images captured by camera base displacement

V. Conclusions

This work presented the design and construction of an operational prototype that allows the creation of datasets for service robots to identify and grasp household objects. Multiple images of objects, from different points of view, can be automatically captured by a webcam. This information permit the building of an initial dataset, that can be used to allow service robots to recognize and estimate positions of objects, aiming at the grasping/picking-up of these objects, as proposed by the robotics competitions such as RoboCup@Home.

The construction of this plywood prototype was designed to be of easy assembly/disassembly, thus allowing it to be easily transported. The prototype was split in three parts, being: a main structure, the table where the turnable base is fixed, and the base of the camera.

Some problems were encountered during this construction, such as the table's moving structure. In the first phase, a system with rails and gears was used. However, with heavy objects (over 1 kg) on the desk, pressing it down, the table bended. The movement of rails and gear could not achieve the complete length of the telescopic slide, due to increase in pressure over the equipment. After half of the distance were extended, the table goes down approximately 1mm forcing the stepper motor to stop. This system was replaced with a toothed belt system as shown in section III.c, improving the table movement.

Another problem occuring during the image capture phase: variation in environment brightness. This problem interfered with the image quality, since variations in luminosity due to solar reflections made some of the images generated by the webcam to seem like a flash. It was necessary to decrease the brightness of the environment in order to capture the images presented in section IV.

Initially 251 images of a single object were collected, but the system allows to collect 7,200 images by object. A large number of images is ideal for establishing a more robust and structured dataset.

The next step is to change the camera base to support two cameras and an RGBD sensor (such as Microsoft Kinect ^{TM}). This will enable the capture of images through stereo vision and to establish the different depths of the object with each movement of the table.

Another possibility of enhancement, that is very common in Robocup@Home competitions, is to cope with different brightness conditions, that could be easily included in the prototype with some kind of variable illumination system.

REFERENCES

[1] 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.

[2] C. Rennie, R. Shome, K. E. Bekris, and A. F. D. Souza, "A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place," *CoRR*, vol. abs/1509.01277, 2015. [Online]. Available: http://arxiv.org/abs/1509.01277

[3] F. Jumel, J. Saraydaryan, R. Leber, L. Matignon, E. Lombardi, C. Wolf, and O. Simonin, "Context aware robot architecture, application to the robocup@home challenge," 2018.

[4] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing training data for object detection in indoor scenes," *CoRR*, vol. abs/1702.07836, 2017. [Online]. Available: http://arxiv.org/abs/1702.07836

[5] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," *CoRR*, vol. abs/1803.11469, 2018. [Online]. Available: http://arxiv.org/abs/1803.11469

[6] E. Martinez-Martin and A. P. d. Pobil, "Vision for robust robot manipulation," 2019-04-06.

[7] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.

[8] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, J. Garcia-Rodriguez, J. Azorin-Lopez, M. Saval-Calvo, and M. Cazorla, "Multi-sensor 3d object dataset for object recognition with full pose estimation," *Neural Computing and Applications*, vol. 28, no. 5, pp. 941–952, May 2017. [Online]. Available: https://doi.org/10.1007/s00521-016-2224-9

[9] J. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez, "Robot@home, a robotic dataset for semantic mapping of home environments," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 131–141, 2017.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: https://www.nature.com/articles/nature14539

[11] A. Aldoma, T. Fulhammer, and M. Vincze, "Automation of ground truth annotation for multi-view rgb-d object instance recognition datasets," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2014, pp. 5016–5023.

[12] SketchUp. (2019) Sketchup, ferramenta gratuira de modelagem web. [Online]. Available: https://www.sketchup.com/pt-BR/plans-and-pricing/sketchup-free

[13] T. Hodan, P. Haluza, S. Obdrzálek, J. Matas, M. I. A. Lourakis, and X. Zabulis, "T-LESS: an RGB-D dataset for 6d pose estimation of texture-less objects," *CoRR*, vol. abs/1701.05498, 2017. [Online]. Available: http://arxiv.org/abs/1701.05498