Chapter 5

Object Detection and Pose Estimation for Warehouse Automation

The 1st Amazon Picking Challenge (APC) was held in parallel with 2015 IEEE International Conference on Robotics and Automation (ICRA) in Seattle, Washington, 26−27 May. The objective of the competition was to provide a challenging problem to the robotics research community that involves integrating the state-of-the-art in object perception, motion and grasp planning to manipulate real-world items in industrial settings. To that end, Amazon Robotics posed a simplified version of the task that many humans face in warehouses all over the world, i.e., picking items from shelves and putting them into containers. In this case, the shelves were prototypical pods from Kiva Systems, and the picker had to be a fully autonomous robot.

In APC 2015, our joint team, Z.U.N., with Zhejiang University and Nanjiang Robotics Co. Ltd is ranked No. 5 in 28 participated teams including top universities around the world such as MIT and UC Berkeley. We designed a dual arm robot with suction gripper. The perception systems have two sets of an RGB-D sensor and a monocular camera in order to fully-cover the width of the shelf. This chapter illustrates the perception module we designed for warehouse automation as a case study and demonstrates the effectiveness of the proposed system under realistic environments.

5.1 Environment Setup for APC 2015

24 items were selected by the organiser which were commonly sold on Amazon.com and pose various degrees of difficulties in terms of both recognition, estimation and grasping, as shown in Fig. 5.1. Rigid and textured objects such as a box of straws and a box of pencils are examples of trivial cases. Some of items are difficult because the non-rigidity thus they are easy to be reshaped or damaged such as books and soft package cat food. Transparent coverage items pose difficulties in the recognition and pose estimation procedures because of the reflective surfaces and the lack of depth information. Cheez-it box, even though being textured and rigid, because of its oversize, it needs to be twisted to be taken out from the bin thus pose extra challenge in grasping planning. Some other items such as the plush toys bring difficulties in the grasping period. One extreme case is the meshed pen holder which is difficult in both detection and grasping, therefore, this item is ignored when met during in the competition.

FIGURE 5.1: Targeting objects for recognition and pose estimation in APC 2015

The 24 items are placed on a pod with 12 bins which is shown in Fig. 5.2. A bin may consists of a single item or multiple items which can be identical or different and only one of them items is target item which needs to be grasped safely from the bin. The score starts from 10 for one successful pick-up and the score increases if additional items are distributed in the same bin. Based on the specific characteristic of each items, some items were given 1 to 3 extra scores such as the plush toys and books. Damaging an item incurred a five-point penalty, while picking the wrong item incurred a 12-point penalty. Each competitor had 20 minutes to pick as many

of the 12 target items as possible and could score as many as 190 points. All the items with additional scores are also highlighted in Fig. 5.1 as well.

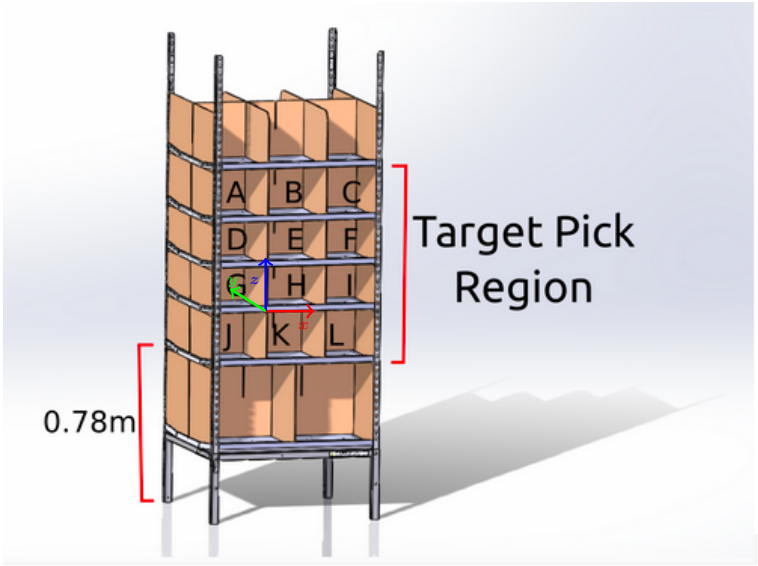


FIGURE 5.2: The shelf which contains the objects used in APC 2015.

Here, we would like to highlight some difficulties in detecting and grasping the object reliably from the pod:

1. The walls and shelves are not equi-distributed. This introduces differences in the nominal size of the openings of each individual bin, with height ranging between 19 and 22cm, and width between 25 and 30 cm;

2. Each bin has a lip on the bottom and top edges, as shown in Fig. 5.3, which impedes exposing an object by sliding it;

3. The lateral bins have a lip of the exterior edge, as shown in Fig. 5.3, which impedes exposing an object by pulling on it;

4. Finally, also worth noting, is the metallic bottom of the structure, which produced bad reflections from depth sensors and proved to be an impediment for accurate estimation of the location of the shelf by model fitting to point cloud data.

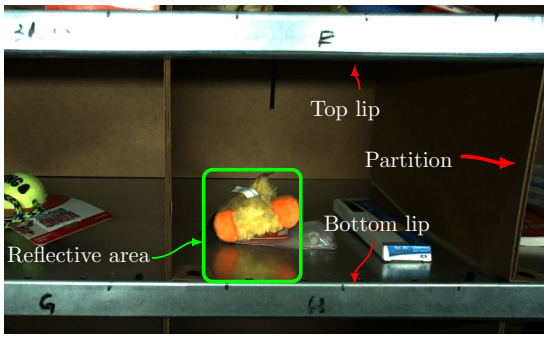


FIGURE 5.3: Difficulties in recognition and grasping.

5.2 Analysis of the Particularities of the Warehouse Pick-and-Place Problem

In APC 2015, compared with the our previous work in Chapter. 3 and Chapter. 4, the environment set-up is significantly different from our previous work. Some of the differences make the problem more difficult and some of them provide extra information which make the case easier. Here we address the key issues as below:

1. In both Chapter. 3 and Chapter. 4, the proposed algorithms are designed to recognise all target items from an unstructured environment given seldom prior knowledge. However, in APC 2015, we are provided with the prior information of the environment and also the objects:

   (a) The size of the shelf and the dimension of each bin are given;

   (b) The relative transformation between the "shelf" frame (highlighted in Fig. 5.2) and the robotic body frame can be pre-calibrated;

   (c) The targeting object in each bin is given and the other objects which are placed together with the targeting object in the same bin are also given as additional information;

2. In the previous two chapters, the full 6 DoFs relative pose $[\mathbf{R}, \mathbf{t}]$ of the detected object is required to be estimated. However, in APC 2015, successful grasping and manipulation of the object do not require the estimation of 6 DoFs pose. In fact, we only need to detect the region on the object which is suitable to be grasped in 3D space. Using the pre-calibrated transformation between the robot body frame and bin frame, we are only required to estimated the position targeting region w.r.t the bin frame.

Note: Compared with our previous work, it is fair to say that the environment set-up in APC 2015 is more constrained and trivial compared with the proposed framework in Chapter. 3 thus makes the object detection and pose estimation problem easier simpler.

3. In Chapter. 3 and Chapter. 4, most of objects which can be handled by the proposed approaches are non-reflective and more importantly rigid. In this perspective, the objects included in APC 2015 is much more challenging such as:

   (a) Soft-cover books and objects with crushable package;

   (b) Objects in plastic and transparent package which shows invalid information on depth sensor;

   (c) Non-rigid objects which can be easily deformed by external force;

   (d) Meshed objects;

   (e) Plush toys;

   (f) Tiny objects sized within 3 cm;

   All the candidature objects have already been presented in Fig.5.1 and bonus points are available for difficult objects. In Fig.5.1, one coin is placed beside the object to roughly tell the size of the object.

   Note: The targeting objects in APC 2015 is more difficult compared with the objects from our previous approaches.

5.3 Hardware Design of the Robotic Platform

In this section, we present the hardware system of the robotic platform including its main body, gripper , the equipped sensors and more importantly explain why we designed it in such way. As shown in Fig.5.4, the designed platform has dual arms and can be lifted up and down using the belt transportation system mounted on the back. Because of the height of the shelf is above $1.7m$, in order to reach the items on different levels, the lift system is a essential part in order to allow the sensor to observe the bins and the gripper to grasp the target items. The lift system allows the robot body (arm and sensor) to 4 levels of different heights, corresponding to 4 rows on the shelf. Considering the length of the robotic arm, the width of the shelf and the Field-of-View (FoV) of the sensor (xtion RGB-D camera),

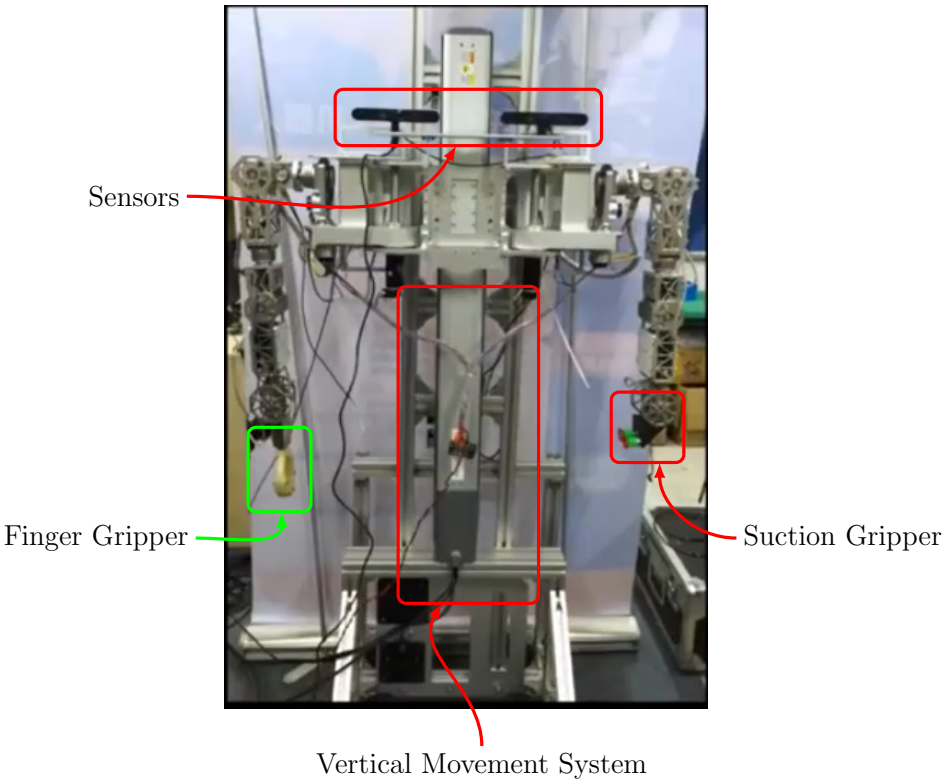


FIGURE 5.4: Robotic platform of APC 2015.

2 sets of the perception system are equipped on the platform, responsible for the left and right column of the bins respectively and the middle column is covered by both sets. Each perception module shown in the top red box in Fig. 5.4 consists:

1. An *Asus xtion* RGB-D camera which provides the depth information of the bin and this is the key sensor to provide the pose estimation results;

2. A *pointgrey* high resolution RGB-D camera which aims at recognising the targeting objects. The extrinsic parameter between the *xtion* and *pointgrey* are calibrated in advance;

3. A LED light which is used to control the illumination condition of the bin;

The following Fig. 5.5 demonstrates the configuration of the sensor w.r.t to the shelf. Some basic and critical parameters for the robotic platform are: 1) the horizontal FoV of xtion sensor is $59°$; 2) The object has to be placed at least $0.5 \sim 0.6m$ in front of the sensor to provide depth information[1]; 3) the length of the arm is $0.68m$ and the length of the gripper is approximately $0.2m$ thus the robot platform cannot be placed too far from the shelf. In order make both the sensor and the arm work properly, the distance between the sensor (robotic platform) and the shelf is

[1]Distance within $0.5m$ is the "blind-zone" for xtion sensor

set to be $0.6m$. Therefore, as shown in Fig. 5.5(a) is not capable to cover the whole shelf. Fig. 5.5(b) provides a simplified presentation of the final configuration map and two sets of sensors are introduced as we explained before.

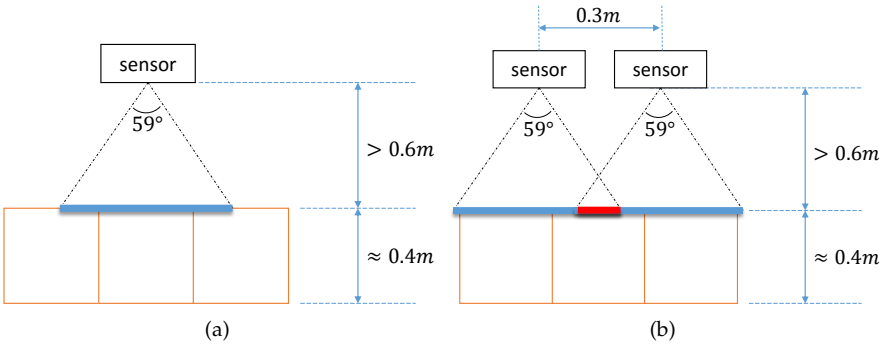


FIGURE 5.5: Sensor configuration for Amazon Picking Challenge.

In Fig. 5.4, the robotic platform is equipped with 2 different kinds of gripper: 1) a two-finger gripper on the right arm (left green box in Fig. 5.4) and 2) a vacuum gripper on the left arm (left green box in Fig. 5.4). The original idea is to grasp the rigid objects using the two-figure gripper and the objects with plastic cover using vacuum gripper. However, due to the oversize two-figure gripper, in the actual competition, both armed is equipped with vacuum gripper because of the reliability issue. During the experiments, except for the mesh pencil cup and one tiny object, the vaccum gripped is able to handle all other objects well enough. We will not provide the detailed specifications of the hardware system here and we will only focus on the perception system which detects the targeting object and estimates the relative pose of the surface (where the gripper is placed on) in the next section.

5.4 Perception Module

5.4.1 Pre-Processed Prior Knowledge

Before introducing the perception module, we present how to utilise the prior knowledge of the environment in order to provide more accurate and faster object detection and pose estimation performances. Besides the pre-built object model and trained classifier, we have:

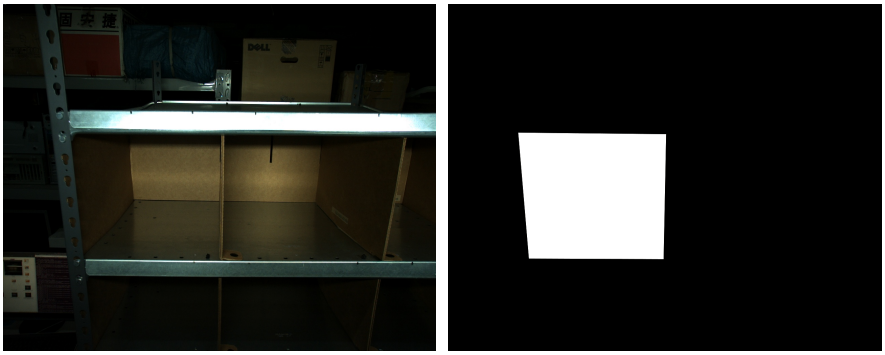
1. RGB mask image of each bin;

2. RGB image of each empty bin;

3. Depth mask image of each bin;

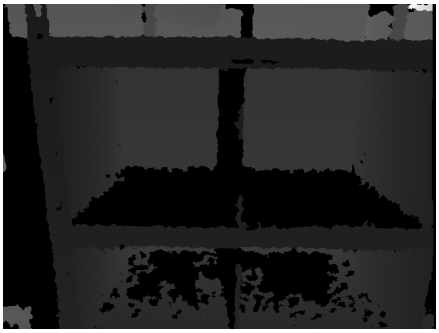
4. Depth image of each empty bin;

A set of examples images from bin *A* is presented in Fig. 5.6 as below. Given the mask images (RGB and depth), we are able to quickly identify the correct region which is corresponded to the bin where the target object is located in. By doing the subtraction w.r.t the empty images, we can locate the placed objects in the bin roughly. This allows us to avoid the time consuming sliding window step in object detection especially for the *Kernel Descriptor* recogniser and *EBlearn recogniser* which we will explain later.

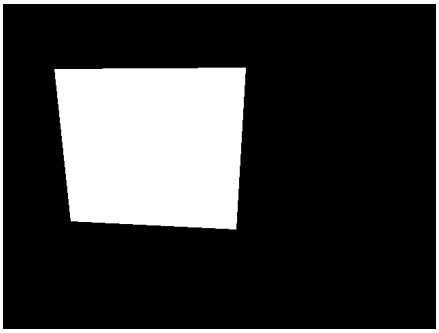


(a) RGB image of the empty bin

(b) Mask image of the empty bin(RGB)



(c) Depth image of the empty bin



(d) Mask image of the empty bin(Depth)

FIGURE 5.6: Pre-processed image for object detection.

5.4.2 Perception System Pipeline

The designed perception module is shown in Fig.5.7. The system accepts inputs from 2 sensor: a Xtion RGB-D camera with lower RGB resolution and a PointGrey high resolution RGB camera. The extrinsic parameters between the depth sensor on Xtion and the PointGrey camera are calibrated[2].

[2]Because the different resolution on depth sensor and PointGrey camera, we can not guarantee that there will be depth value for each pixel on RGB image.

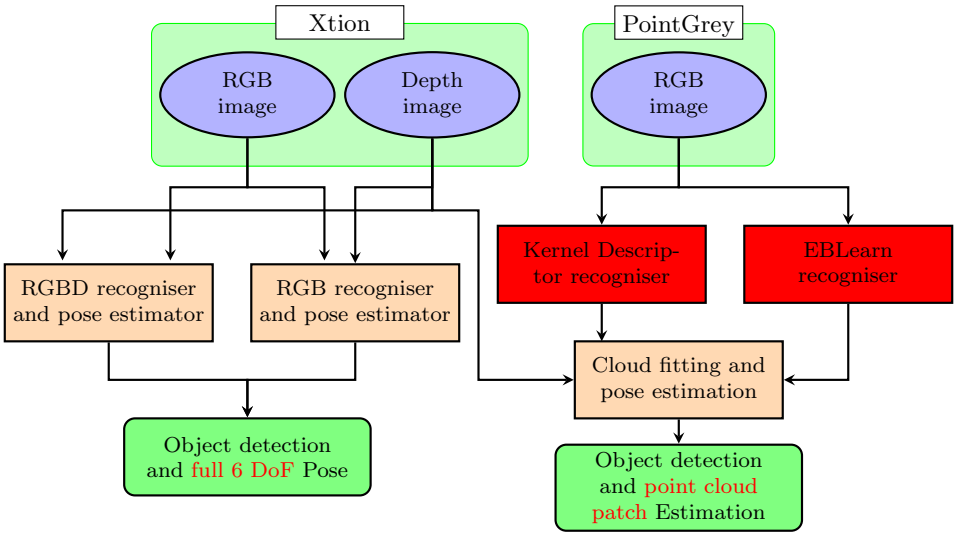


FIGURE 5.7: The competition shelf in APC 2015.

As we explained in section.5.1, the objects in APC 2015 are more challenging due their texture and deformation. We separate the 24 objects into two categories:

1. Textured objects with enough size;

2. Objects with plastic cover, small objects and deformable objects;

As labelled in Fig. 5.8 where yellow items belong to category 1 and the red items belong to category 2.

oreo mega stuf	champion copper plus spark plug	expo dry erase board eraser	genuine joe plastic stir sticks	munchkin white hot duck bath toy	
crayola 64 ct	mommys helper outlet plug	sharpie accent tank style highliters	stanley 66 052	safety works safety glasses	
cheezit big original	papermate 12 count mirado black warrior	feline greenies detal treats	elmers wahsable no run school glue	mead index cards	rolodex jumbo pencil cup*
first year take and toss straw cup	highland 6539 self stick notes	mark twain huckelberry finn	kyjen squeakin eggs plush puppies	kong sitting frog dog toy	kong air god squekair tennis ball
kong duck dog toy	laugh out loud joke book				

FIGURE 5.8: Perception module for APC 2015.

For objects of category 1, we modified and implemented the proposed object recognition and pose estimation in Chapter. 3 and also MOPED. Some of the key modifications on the original systems are:

1. Mask image: in order to extract the features only in the current bin, thus achieving better feature matching results, mask image is applied in feature extraction and selection step;

2. Dynamic kd-tree construction: rather than built a SIFT descriptor KD-tree of all 24 objects, we built the kd-tree using only the objects in the current bin in runtime;

3. Single object detection and pose estimation: In the proposed object detection system and MOPED, multiple existed objects are detected, however, in this work, we only interest in the targeting object. Please notice that the poses of the other objects are not considered in the perception module and collision avoidance between the gripper and other objects are *not* considered;

For objects of category 2, we adopted kernel descriptor[24] and EBLearn[123] into the system. Different from RGB-Recogniser and RGBD-Recogniser, both KD-Recogniser and EBLearn-Recogniser can only provide the bounding-box of the detected objects and this is the reason for another plane fitting and estimation after the recognition step. We list some of the key points that worth sharing as below:

1. Instead of using sliding window detector with KD-Recogniser and EBLearn-Recogniser, we use the given mask image, RGB and depth image of the empty bin. By doing image subtraction, the image patch of the targeting objects are captured;

2. The classifier for KD-Recogniser for object A is trained using the images of object A and the background. Compared with training the classifier of object A against all other objects, the benefits are:

   (a) Since there are only limited number of objects in the bin, when using classifier trained from all other unnecessary objects will produce inaccurate classification results;

   (b) If we want to use the classifier trained from only the objects in the current bin, since there is no prior knowledge of the objects in the bin, we need either train the classifier online or provide all possible combinations of objects and none of them seem to be realistic;

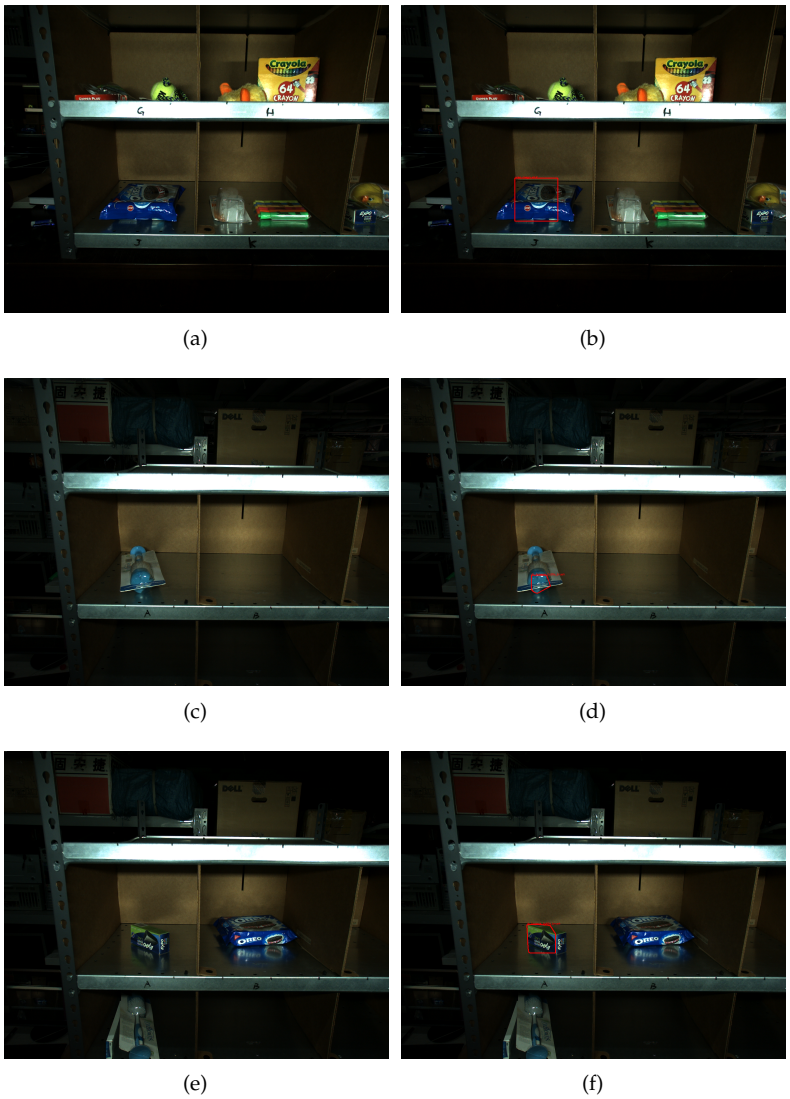
3. For the all candidature image patches in the bin, we test them using trained classifier and select the image patch which shows better score from the Lib-SVM classifier as the correct one;

4. For the EBLearn-Recogniser, we use the similar idea and trained the model for each object individually;

5. Without knowing the full 6 DoFs pose of the object, in order to grasp the object stably using the vacuum gripper, we need to estimate the position of the surface patch where the vacuum gripper can be placed on.

The pose estimation for objects in category 1 is trivial since the 6-DoFs relative pose can be estimated from the RGB-D and RGB recogniser. For the challenging objects in category 2, Via registrating the RGB image with the depth image captured from xtion sensor, we were able to obtain the point cloud of the detected objects. We adopted surface fitting method available in PCL to select the correct suction point.

5.5 Experimental Results



(a) (b) (c) (d) (e) (f)

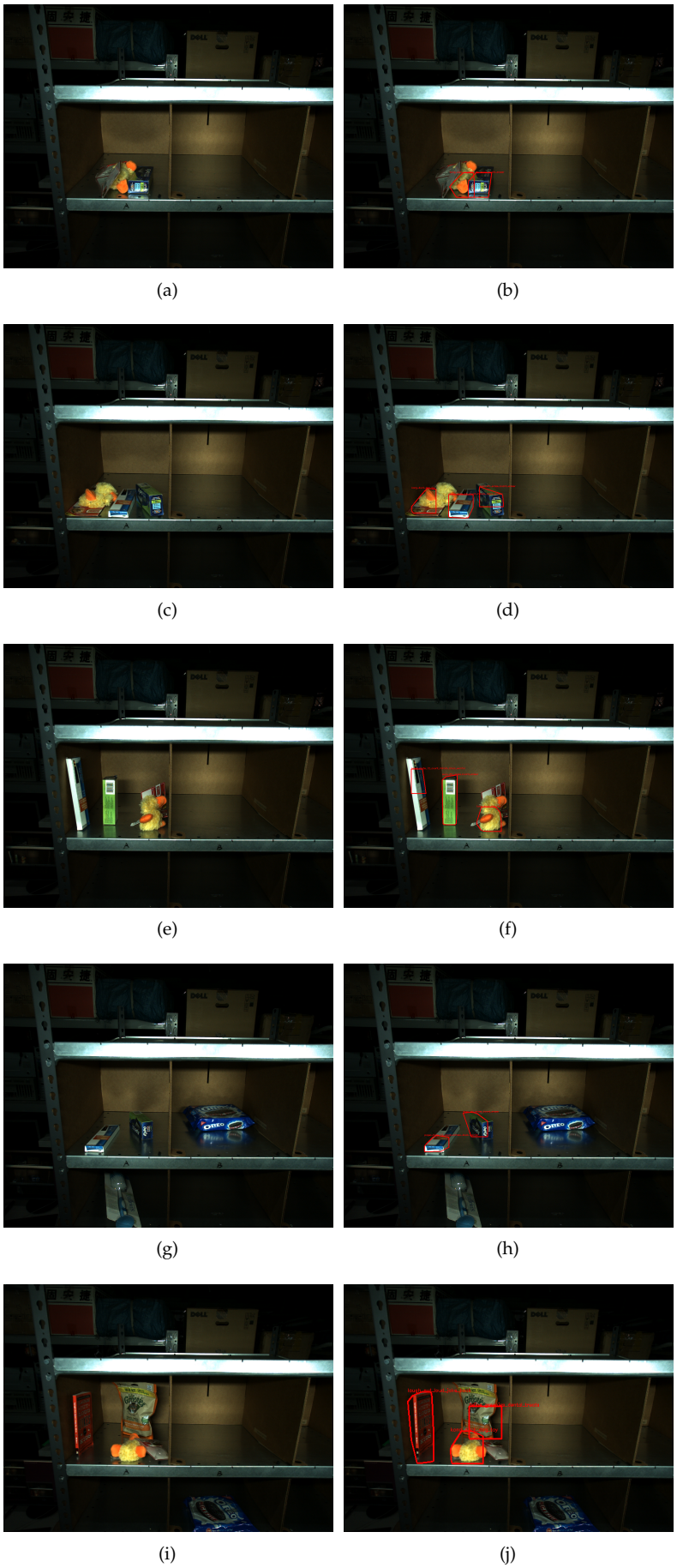
FIGURE 5.9: Single object detection results.

In this section, we present some of the object detection results in the period of system testing. First of all, Fig. 5.9 demonstrates the detection results when only 1 targeting object is placed in the bin. Even this single object case may sound to be trivial, however, as shown in Fig. 5.9, it is difficult to identify the whole object given the imperfect depth information and reflective surface. For example, in Fig. 5.9(c) and Fig. 5.9(d), the cup brush can be captured only in its bottom part. In Fig. 5.9(e) and Fig. 5.9(f), there exists a reverted reflection of the box due to the material of the surface.

Fig. 5.10 demonstrates more examples of multiple object detection results. The designed perception module is capable of providing reliable and robust object detection results even under occluded environment by only capturing parts of the object. In Fig. 5.10(a) and Fig. 5.10(b), parts of the duck toy and the box were under the shaded area, our framework is capable of finding the object given limited observable information. Our method can also detect the targeting objects under the side-view where only very limited information are provided, e.g., the vertically placed books presented in Fig. 5.10(e), Fig. 5.10(f), Fig. 5.10(i) and Fig. 5.10(j).

5.6 Summary

In this chapter, using the proposed work in Chapter. 3 and Chapter. 4 with the additional Kernel Descriptor recogniser and EBlearn recogniser, we designed a robotic platform with a reliable and robust perception module under warehouse environment. The proposed framework fully utilises the provided prior knowledge of the environments to achieve better detection results. Using the suction gripper on both arms, the robot can grasp the targeting objects from the bin given the detected boundingboxed. We were also one of top-5 teams in all of the participants in Amazon Picking Challenge.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

(i) (j)

FIGURE 5.10: Single object detection results.