**Cartman: The low-cost Cartesian Manipulator that won the Amazon Robotics Challenge**

D. Morrison1,2, A.W. Tow1,2, M. McTaggart1,2, R. Smith1,2, N. Kelly-Boxall1,2, S. Wade-McCue1,2, J. Erskine1,2, R. Grinover1,2, A. Gurman1,2, T. Hunn1,2, D. Lee1,2, A. Milan5, G. Rallos1,2, C. Lehnert1,2, I. Reid1,3, P. Corke1,2, and J. Leitner1,2

1Australian Centre for Robotic Vision (ACRV), 2Queensland University of Technology, Brisbane Australia, 3University of Adelaide, Adelaide Australia, 4Australian National University, Canberra Australia, 5Amazon, Berlin Germany

### System Design

- **Cartesian Axes**
- **Cameras**
- **Vacuum Pumps & Power Supply**
- **Storage System**
- **Scales**

Custom-built Cartesian Manipulator for the 2017 Amazon Robotics Challenge.

Cartesian design simplifies planning and executing motions in the confines of the storage system compared to an articulated arm.

Stowed 14 out of 16 and picked all 9 items in 27 minutes, scoring 272 points to win the challenge finals held in Nagoya, Japan.

### Perception

A wrist-mounted RGB-D camera can be positioned anywhere within the workspace.

A RefineNet network is used for pixelwise semantic segmentation of the RGB image.

Scales mounted underneath the storage system and stow tote are used to confirm the picked item’s weight.

Secondary camera mounted to the side of the robot allows for classification of isolated objects if the initial classification and weight do not match.

### Fast Item Learning

- **RGB Centroid**
  - No valid point cloud (e.g. clear or black objects)
    - Object position is estimated from the RGB segment using camera parameters, and the centroid is chosen.

- **Surface Normals**
  - Good quality point clouds (e.g. regular, matt objects)
    - Grasp candidates (arrows) are generated in a grid and ranked heuristically before being sparsified.

- **Incomplete point clouds (e.g. reflective objects)**
  - Point cloud is filtered for noise and the centroid is chosen as the grasp point.

- **RGB Centroid**
  - No high quality grasps
    - Not enough valid points

- **Point Cloud Centroid**
  - No valid point cloud (e.g. clear or black objects)
    - Object position is estimated from the RGB segment using camera parameters, and the centroid is chosen.

- **RGB Image**
  - Item Segment
  - Grasp Ranking
  - Grasp Output

Half of the items are only seen 45 minutes before each competition run.

We collect 7 images of each new item, and fine-tune the RefineNet network for 30 minutes.

There is only a marginal performance gain by collecting more than 7 images.

### System Robustness

#### Active and Interactive Perception

Multiple viewpoints are used to help locate partially occluded items. If no wanted items are visible, the system will move objects within the storage system based on the likelihood that they are obscuring wanted items.

#### Item Reclassification

Items can be reclassified to correct errors, based on consensus from two sensors (primary/secondary visual classification and weight).

#### Error Detection and Recovery

A number of sensors are used to detect failed grasps and dropped items. Toward the end of a task, visual classification is used to double-check that the location of items matches the internal state.

### Multi-modal End Effector

- **Suction End**
- **Camera**
- **Tool Change Mechanism**
- **Gripper End**

Cartesian robot design allows for dual-ended tool, as opposed to hybrid.

Custom end-effector designed to complement the grasp detection system.

6-degrees-of-freedom at each end.

Uses a parallel plate gripper and suction gripper to handle a wide variety of items (rigid, semi-rigid, hinged, deformable, porous).

Tool selection via a 180 degree tool change mechanism.

### Hierarchical Grasp Detection for Visually Challenging Environments

**Visual Quality Grasp Precision**

- **Surface Normals**
  - Good quality point clouds (e.g. regular, matt objects)
  - Grasp candidates (arrows) are generated in a grid and ranked heuristically before being sparsified.

- **Point Cloud Centroid**
  - Incomplete point clouds (e.g. reflective objects)
  - Point cloud is filtered for noise and the centroid is chosen as the grasp point.

- **RGB Centroid**
  - No valid point cloud (e.g. clear or black objects)
  - Object position is estimated from the RGB segment using camera parameters, and the centroid is chosen.