Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions

Authors: Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, Jethro Tan

Publisher: Proceedings of International Conference on Robotics and Automation
Year: 2018

Keywords: Computer Vision, NLP, Robot grasping

1. Abstract:

"Comprehension of spoken natural language is an essential skill for robots to communicate with humans effectively. However, handling unconstrained spoken instructions is challenging due to (1) complex structures and the wide variety of expressions used in spoken language, and (2) inherent ambiguity of human instructions. In this paper, we propose the first comprehensive system for controlling robots with unconstrained spoken language, which is able to effectively resolve ambiguity in spoken instructions. Specifically, we integrate deep learning-based object detection together with natural language processing technologies to handle unconstrained spoken instructions, and propose a method for robots to resolve instruction ambiguity through dialogue. Through our experiments on both a simulated environment as well as a physical industrial robot arm, we demonstrate the ability of our system to understand natural instructions from human operators effectively, and show how higher success rates of the object picking task can be achieved through an interactive clarification process."

2. Introduction:

- For the Human-Robot interaction, the normal communication would be very assistive instead of using a device and controlling with it.
- So, an idea of using the normal language to control the robot had been interesting topic where many of the works were not focused.
- One of the obstacle to be faced is that the natural language used by humans is not limited or simple. However, the authors determined to achieve this.
- In addition, in the architecture the natural language has to map the corresponding visual image in the real world.
- The authors also concentrated on narrowing down the objects to be chosen by further clarification by the robot.

3. Related works:

- Authors claim that it had been the first approach which integrated the natural language for object picking with non-constrained instructions.
- However, some of the previous works had been concentrated on natural language interaction with robots. But these works had limited to the precise speeches far from ambiguity.
- Some of the works had focused only on the finite objects with NLP and had not concentrated on the ambiguity of the robot.
- Appreciable work had been going in improving the image captioning with the help of neural networks. These advancements can be used to improve the state of the art.

4. Approach and Experiments:

- Around 100 objects were selected for the experiment and some of them are placed in the four boxes attached together. These objects were placed randomly and scattered.
- Objects were also selected that they had a wide variety and also around 20 items were selected such that they are not trained into the network. Some of the object were covered by others to make the task more challenging.
- As the authors don’t want to rely on any of the predefined vocabulary sets, accents and grammar structures, and they cannot find the dataset which can understand the texture, size and shape, they ought to create dataset.
- The main tasks that the authors defined as the solution are:
    - Converting the speech into textual information through Web Speech API.
    - Recognizing all the objects by SSD-based detectors and applying bounding boxes to the objects.
    - To detect the target object based on the input text instruction.
    - To detect the target box based on the input textual instruction.
    - If the instruction is enigmatic, then ask for the further elaboration by providing the feedback.
- The network architecture is created such that the target object is selected by combining the object recognition and the Speech Recognition.
- Multi-layer perceptron layer combines both the features of the object in bounding box and relative features of all other objects, which results in similar objects.
- As in skip gram model, each word is represented as the vector succeeded by MLP and LSTM layers.
- Although the object detection and destination box selection can be solved by single architecture, the authors used separate architectures as they have not known before.
- A threshold is created for the confidence of object detection or else it will create a state of ambiguity, which it ask the instructors for the further clarification.

- The tuning of this margin is crucial, as too much threshold may increase the confidence level or the low threshold will neglect the ambiguity and may result in dangerous actions.
- The dataset is also created known as *Picking Instructions for Commodities Dataset(PFN-PIC).*
- Each object was labelled by three annotators through crowdsourcing called as Mechanical Trunk by Amazon.
- The ambiguity expressions in the dataset were not accurately concentrated to improve, as their focus of work had been one of them.
- The training had been given on the CNNs with pretrained ImageNet dataset on VGG16 model.
- To improve the training, they even had adapted the data augmentation technique, which helped them in training the network with different orientations of the same object.
- The total number of MLP and LSTM layers are 3 with 512 hidden units. The Adam optimization was adapted and operated with around 120000 iterations.
- The words which had not appeared in the data training were given the "UNK" token as normally done in NLP.
- The robotics system was operated on ubuntu 16.04 with internal ROS kinetic, which consists of motion planner to plan the random movements of the motion planner.

5. Results:

- The results had been categorized into the object detection accuracy, Destination box accuracy and pick and place accuracy.
- Further the results were also compared based on known and unknown objects which had given similar results with around 10% threshold either sides.
- The destination box selection has around 89.7% of accuracy and target object selection had only 75.3%. These are based on the neural networks.
- The pick and place accuracy was 97.3% related to the robotic manipulation activities.
- The former results refers to the average of the results with known and unknown objects. As expected, the known objects were resulted in better accuracy than the unknown objects in the task.
- A simulation is also being done before experiment practically which had given better results than the practise. This is defended by the authors by reasoning a fact of color temperature and contrast differences which they ought to improve in the future.

6. Conclusion and Future works:

- Despite of being pioneers to the approach, they have achieved satisfiable results with variable terminologies and accents.
- It is also appreciable that the human-robot and robot-environment interaction had been improved with the new dataset.
- It is believed that when the objects were limited as in industrial use, it can increase the accuracies than the former experiment.
- The further works can be continued on improving the performances in pick and place by changing the grippers suitable for most applications.
- Different languages can also be applied on the model for the convenience of the regional workers. Moreover the dataset the authors had introduced can be further improved and used for the future prospects.