Review:
Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions

About the paper

- Authors: Preferred Networks, Inc.
- Presented at ICRA 2018
- Best paper award on Human-Robot Interaction

Motivation

- We want robots to understand us

- How?
  - Traditional UI
  - Gestures
  - Imitation
  - Verbal instructions

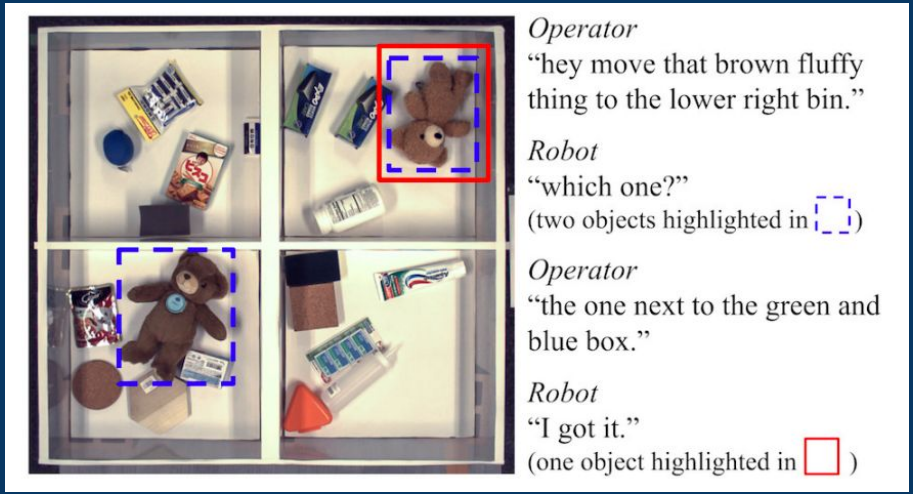
- Emphasis on interaction

Challenges in NL comprehension

- Complex structures
- Wide variety of expressions
- Ambiguity: how to resolve conflicts?

Proposal

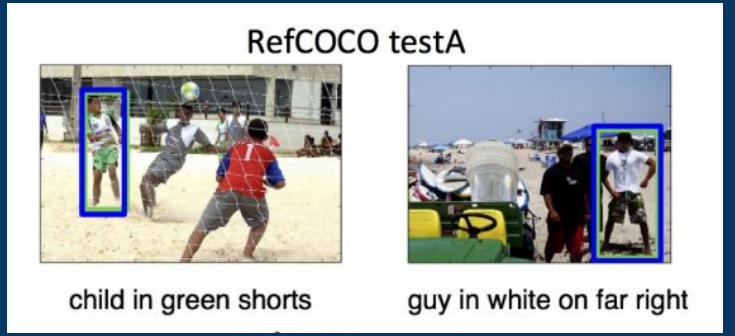
- Framework for controlling robots via unconstrained natural language
- Task: moving real world objects (zero-shot)
- Can resolve ambiguity through dialogue, using visual and verbal feedback

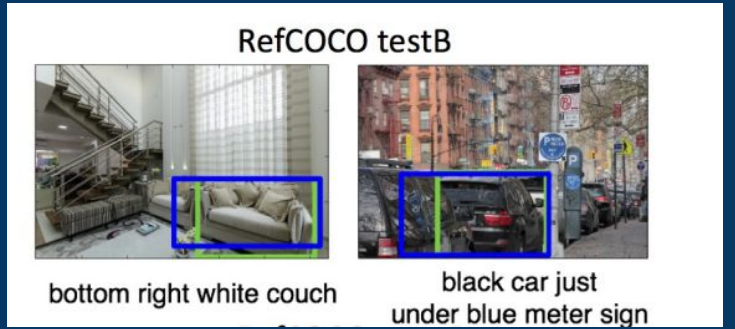


Enabling technologies

- State-of-the-art object detection models
  - Deep Learning approaches
    - Objectness detectors (Faster R-CNN)
    - Single-shot multibox detectors (SSD)

- Object-referring expression models
  - Context modeling
  - Speaker-listener-reinforcer models





Dataset (PFC-PIC)

- Highly cluttered environments
  - Challenge: occlusion
  - Many copies force the use of relative directions

- Organized environments
  - Challenge: indirect references

- Wide variety of objects (> 100)
  - Some are abstract
  - 22 unseen in test set

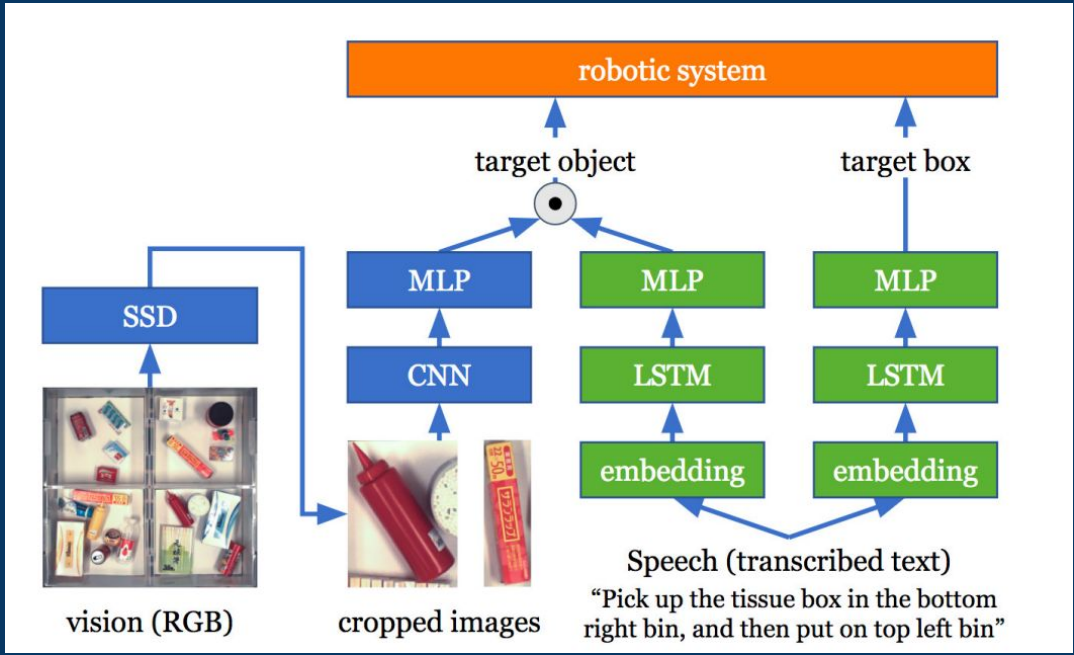
- Training set:
  - 1k images
  - 71k instructions
  - 25k bounding boxes

"Move the tissue box to the top right bin"

Proposed method

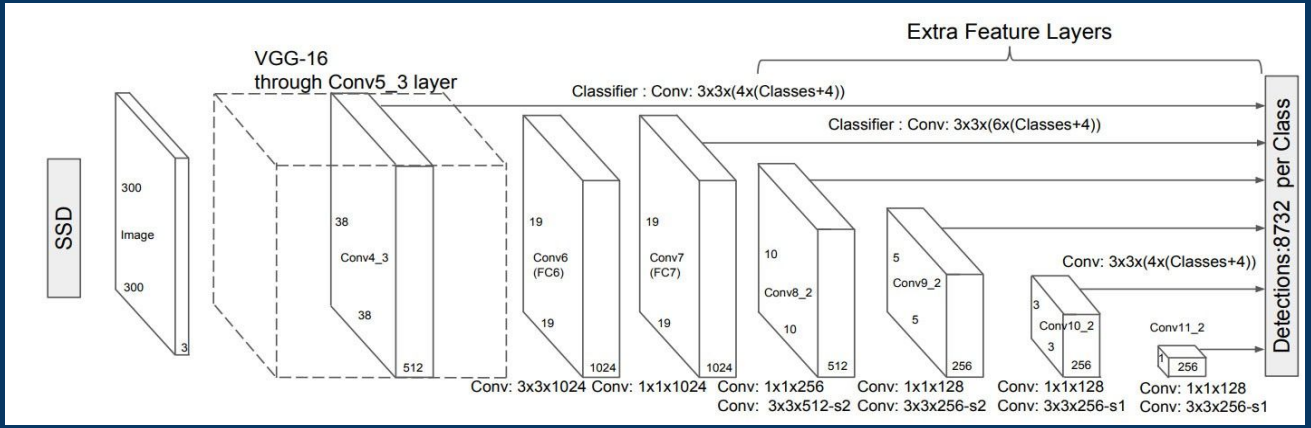
- 2 modules
  - Object recognition
  - Language understanding

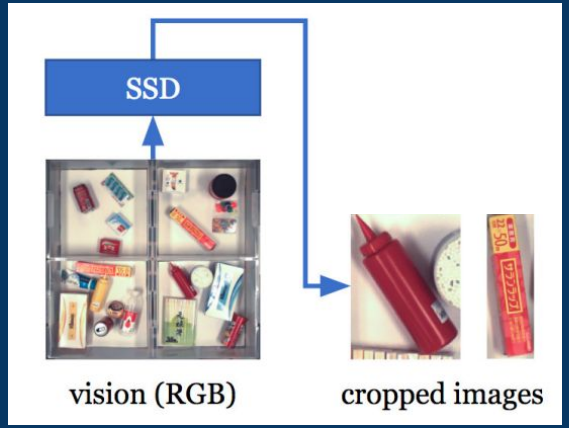
- Joint training



Object detection: Single-shot multibox detector

- No region proposal network (default boxes) ⟶ speed!
- Classifies each area
- New: each candidate is either “foreground object” or “background”

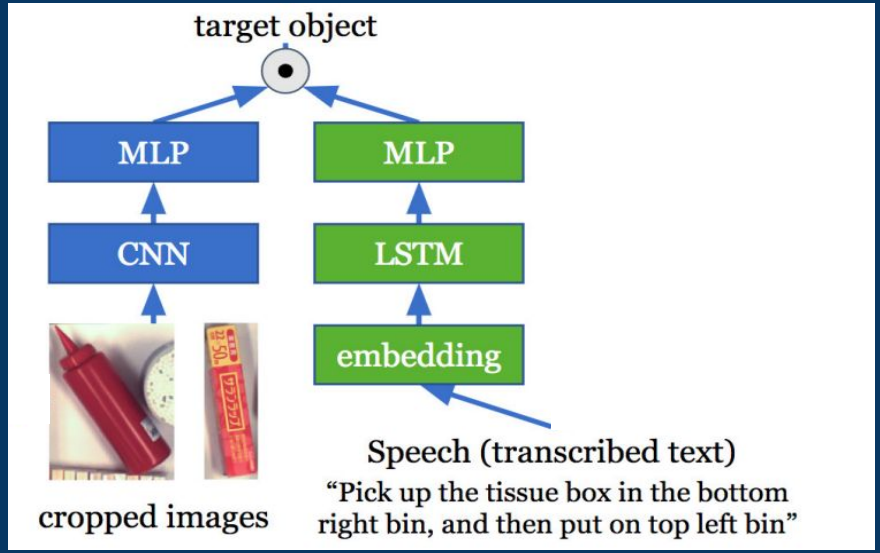




Target object selection

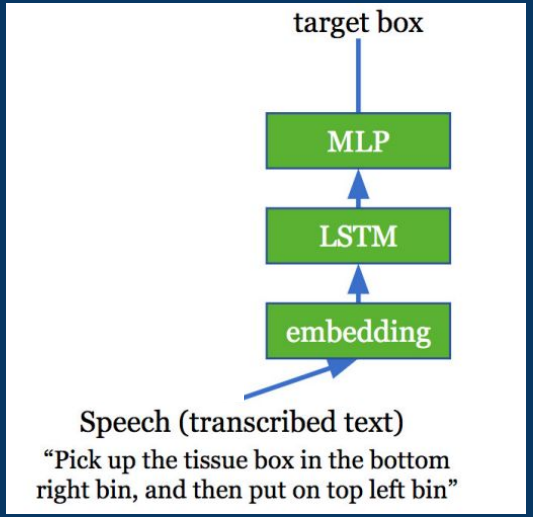
Task definition: $B = \{b\_1, ..., b\_n\}$, find b^* given q and I such that $b^* = b\_true$

- Referring expression listener model
- Modified for zero-shot recognition of unseen objects
- Scores: $\{ s\_i \mid s\_i \in [-1, 1] \}$
  where $s\_i = cos\_dist(feat(I), feat(q))$



Target box selection

- Same NN architecture as previous' step instruction processing works nicely



Handling ambiguity

- Margin-based approach
- Unambiguous instruction only if **(object, box)** has score **> m_obj** and **> m_box**
- Formally

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{q,o}[\max\{0, m - f_\theta(q,o) + f_\theta(q,\hat{o})\}+ \max\{0, m - f_\theta(q,o) + f_\theta(\hat{q},o)\}],$$

- f_θ is a pre-trained ResNet-50 CNN
- Ironically, this is the most ambiguous part of the paper
  - Ratio of the negative undersampling?
  - If ambiguous, what objects does the clarification process reason over?

Training details

- Candidate object detection:
  - ImageNet-trained VGG16
  - Fine-tuning with data augmentation

- 512 hidden units for MLPs and LSTMs

- Loss functions:
  - SSD: IoU over real bounding box
  - Target object: cross-entropy over correct bounding box
  - Target box: cross-entropy over boxes
  - Ambiguity: margin maximization

Robotic system setup

- FANUC M10iA industrial robot arm
- Vacuum gripper as end-effector
- Grasp validation: PPG-CV pressure sensor
- Ensenso N35 stereo camera (point-cloud)
- IDS uEye RGB camera
- PC specs: GTX 1070, i7 6700K
- Arm planning: RRT
- Grasp planning: IK engine

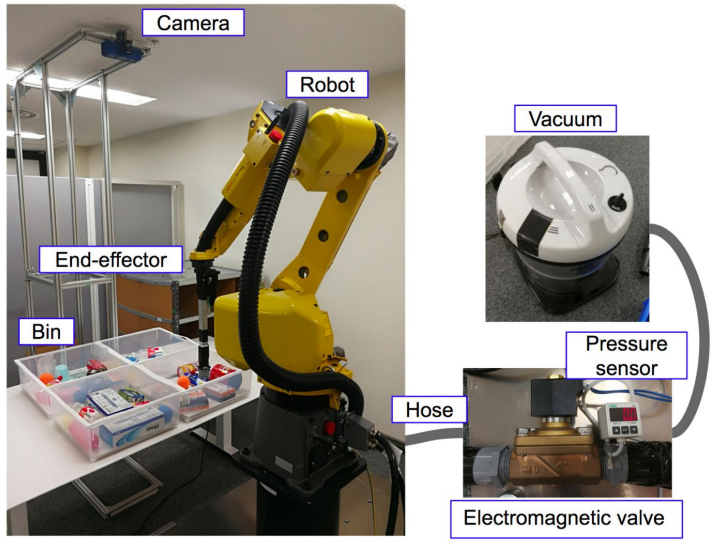


Fig. 5: Robot setup for experiments

Results

- Each module solves its own task successfully
-
- Top-k accuracy rapidly approaches 99%
-
- Clarification is validated as a useful tool to achieve a better performance

	Target Object Selection
Unambiguous cases only	94.9%
Ambiguous cases only	63.6%
Total (without clarification)	88.0%
Total (with clarification)	92.7%

TABLE II: Comparison of the top-1 target object selection accuracies for unambiguous/ambiguous cases, and the total accuracies with and without the interactive clarification process. The accuracy for ambiguous cases was calculated for the top-ranked object output by the system.

Top-1 accuracy per module

Candidate object detection	Destination box selection	Target object selection
98.6%	95.5%	88.0%

Top-k accuracy for target object selection

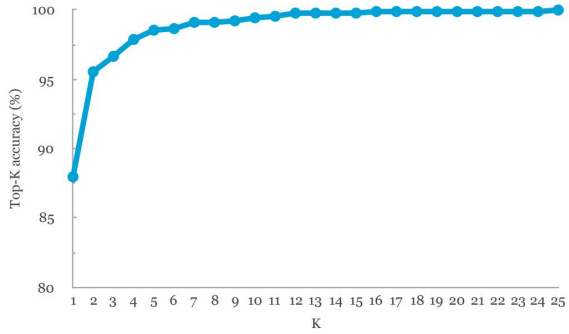
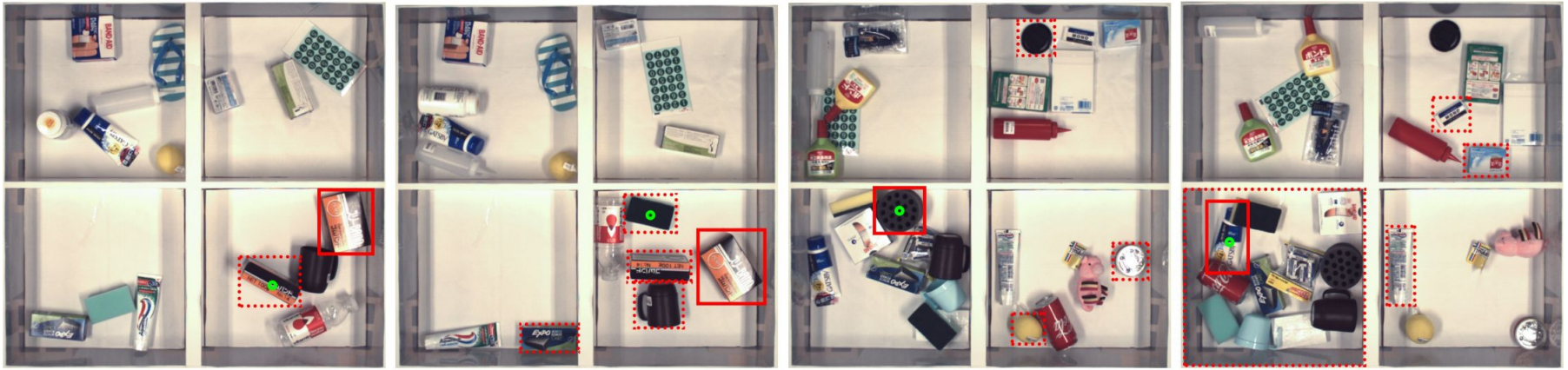


Fig. 6: Top-k accuracies of target object selection given a spoken instruction.

Results



(a) "grab the thin orange and black box and put it in the left lower box" (failure)

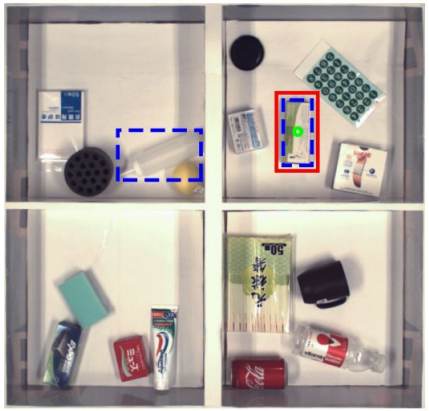
(b) "move the lower right side *black box* to the upper left hand box" (failure)

(c) "move the round object with multiple holes to upper right box" (success)

(d) "grab the blue and white tube under coke can and move to the right bottom box." (success)

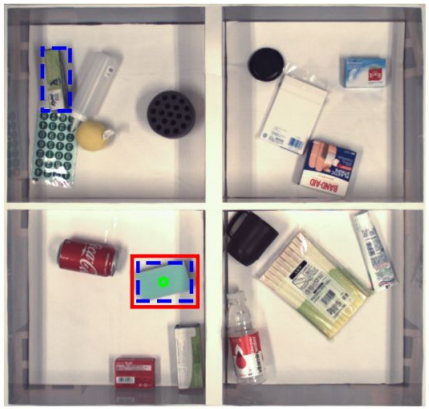
Fig. 7: Examples of success and failure cases with input images and corresponding text instructions. The green dot indicates the correct target object, and the red rectangle with a solid line represents the object that the system predicted. Some regions are also enclosed by a dashed line rectangle to highlight challenges in each instance. Note that these are not actually predicted bounding boxes.

Results



1. “pick the white packet in center and put it into the upper left box”

2. “move the rectangular object, with a green and white label, located in the middle of the top right box, to the top left box.”



1. “move the blue rectangle the top left box.”

2. “pick green sponge and put it in the upper box”

Fig. 8: Examples of success cases which were judged as ambiguous by the first instructions, but our system could correctly identify the correct object after a clarifying instruction. Blue rectangles with a dashed line represent ambiguous objects for the first (ambiguous) instruction, and red rectangle with a solid line represents the final (correct) prediction after clarification.

Results

	Destination Box Selection	Target Object Selection	Pick and Place (only)	Pick and Place (end-to-end)	Avg. Number of Feedback
Without unknown objects	88.9% (56/63)	77.8% (49/63)	98.0% (48/49)	76.2% (48/63)	0.41 (26/63)
With unknown objects	91.2% (31/34)	70.6% (24/34)	95.8% (23/24)	67.6% (23/34)	0.53 (18/34)
Total	89.7% (87/97)	75.3% (73/97)	97.3% (71/73)	73.1% (71/97)	0.45 (44/97)

TABLE III: Experimental results with a physical robot arm. *Destination Selection* and *Target Object Selection* correspond to our destination box and target object selection accuracies. *Pick and Place (only)* and *Pick and Place (end-to-end)* respectively correspond to the success rate of our object picking and placing task calculated only for successfully-detected instances (*only*) and that for all instances (*end-to-end*), including those in which the target box or object detection has failed. *Avg. Number of Feedback* indicates the average number of per-session clarification questions asked by the robot.

Conclusions

- Successfully introduces a robotic system that handles unconstrained spoken language instructions and clarifies ambiguity through interactive dialogue
- Achieved a high end-to-end picking accuracy of 73.1% with an industrial robot
- Demonstrated that an interactive clarification process is effective for disambiguation of a human operator's intention

Demo