A Corpus-Guided Framework for Robotic Visual Perception

Ching L. Teo, Yezhou Yang, Hal Daume III, Cornelia Fermuller and Yiannis Aloimonos
University of Maryland Institute for Advanced Computer Studies, College Park
Robot Perception Control Unit (RPCU)

Output and Command

Output and Command Generation

Information fusion

Language Inference and prediction

Sensors and processing inputs
The Functions of RPCU

- 1) fuse (noisy) information from various sensors and process inputs;
- 2) perform inference and predictions using language;
- 3) eventually generate a useful output or command that show that the robot has truly perceived the world with all its complexity and richness.
Our example of RPCU for Visual Perception

1) Using Language: We use language (large corpora) as a prior in guiding other modules;

2) Information Fusion: We use state-of-art object detectors to detect hands, tools and direct-objects, then predict actions using an EM framework;

3) Output (Command) Generation: We model the sentence generation process as a HMM;

Both 2) and 3) are language guided.
Robot Perception Control Unit (RPCU)

Output and Command

Output and Command Generation

Information fusion

Language Inference and prediction

Input of low level features and detections
Action Features
Robot Perception Control Unit (RPCU)

Output and Command Generation

Information fusion

Sensors and processing inputs

Co-occurrence from NY times Corpus

Language Inference and prediction
Robot Perception Control Unit (RPCU)

Output and Command

Output and Command Generation

Action Prediction

Language Inference and prediction

Sensors and processing inputs
RPCU: Predicting Actions
RPCU: Predicting Actions

- Define a latent assignment variable $A$:

$$A_{ijd} = \begin{cases} 
1 & j \text{ is performed using } i \text{ during } d \\
0 & \text{otherwise}
\end{cases}$$

- Expectation Step:

$$\mathcal{W} = \mathbb{E}_{P(A)}[A]$$

$$W_{ijd} \propto P_I(i)P_L(j|i)Pen(d|j)$$
RPCU: Predicting Actions

- **Maximization Step:**

\[
\hat{C} = \arg \max_C \mathbb{E}_{\mathcal{P}(A)} \left[ \log \mathcal{P}(A|\mathcal{D}, C) \mathcal{P}(\mathcal{D}|C) \right]
\]

\[
\hat{C}_j = \frac{\sum_{i \in N_1, j \in V, d \in M} W_{ijd} F_d}{\sum_{i \in N_1, j \in V, d \in M} W_{ijd}}
\]

- **Action Prediction:**

\[
Z = \sum_{j \in V} \sum_{i \in N_1} \left( \mathcal{P}_I(i|d) \mathcal{P}_L(j|i) \text{Pen}(F_t|C_j^*) \right)
\]

\[
\mathcal{P}_I(j|d) = \frac{\sum_{i \in N_1} \left( \mathcal{P}_I(i|d) \mathcal{P}_L(j|i) \text{Pen}(F_t|C_j^*) \right)}{Z}
\]
Robot Perception Control Unit (RPCU)

Output and Command

Language Inference and prediction

Information fusion

Sensors and processing inputs

Descriptive Sentence Generation
RPCU: Sentence Generation

Corpus-Guided Sentence Generation of Natural Images, EMNLP. 2011
Robot Perception Control Unit (RPCU)
Dataset and Results

{towel,clean,table}
The person is cleaning the table with the towel.

{knife,cut,cheese}
The person is cutting the cheese with the knife.

{knife,cut,tomato}
The person is cutting the tomato with the knife.

{spoon,toss,salad}
The person is tossing the salad with the spoon.
Telluride Experiments

- Kinect
A person is using ladle to pour water into the bowl.
Future Work

- Expand to more sensors input, such as Sound.
- Discover from language, the co-located set of such tools, objects and actions via attributes, rather than pre-defined sets.
- Extend the language generation module to generate even more complicated sentences that involves, for example, adjectives and adverbs.
- ...

...
Thank You!