Jointly Learning Grounded Task Structures from Language Instruction and Visual Demonstration

Changsong Liu, Shaohua Yang, Sari Saba-Sadiya, Joyce Y. Chai
Nishant Shukla, Yunzhong He, Song-Chun Zhu

Language And Interaction Research Lab, MSU
Centre for Vision, Cognition, Learning and Autonomy, UCLA

Motivation

• Research objective
  Enable AI agents to learn real-world tasks through natural dialogue with humans
  Learning hierarchical & grounded task representations from human demonstration and language instruction

- Data
  - Visual state
    - A 18-dimensional numeric vector
  - Alignment between vision & language
    - Based on timing information (i.e., co-occurrence)
  - Primitive action
    - One-step change-of-state, e.g., $x_2 \rightarrow x_1$ (fold left sleeve)
  - Complex action
    - Multiple step change-of-state, e.g., $x_1 \rightarrow x_2 \rightarrow x_3$ (fold two sleeves)

Evaluation

- Data
  - 45 t-shirt folding demonstrations from 6 people
  - 5-fold cross-evaluation

- Recognizing low-level (primitive) actions
  - Baseline
  - Map each visual state to the nearest cluster
  - Search for the same “change of state” in the training set
  - AOG joint inference
  - Each visual state has k cluster mapping hypotheses
  - Apply a parsing algorithm to find the best parse and state sequence

- Recognizing high-level (complex) actions
  - The same baseline
  - AOG joint inference
  - Nonterminal nodes in the parse tree correspond to complex actions

Conclusion & Future Work

- **Conclusion**
  - Automatic learning of hierarchical, grounded, and language-oriented task model
  - Learning from realistic demonstrations and natural instructions provided by humans
  - Learned model enables robust recognition of primitive and complex actions from noisy video inputs — an important capability for human-agent collaboration in the real-world
- **Future Work**
  - Extend to more complex tasks
  - Interactive learning using multimodal dialogue

This work was supported in part by N66001-15-C-4035 from the DARPA SIMPLEX program, and IIS-1208390 and IIS-1617862 from the National Science Foundation.