Guest lecture:
Visually Interpreting Human Actions for Intelligent Agent

Yezhou Yang
yz.yang@asu.edu
Prof. Kanade from CMU on Robotics System
• Robotics/AI research is fun to do!
• You can always find something interesting to you (hopefully...) in Robotics.
• Roller Coaster
• Perception is (one of) the bottlenecks...
• … … … …
Assistant Prof.  
2016-  

Active Perception Group
Human Manipulation Action Understanding for Cognitive Robots

Illustration by Austin Myers
Manipulation actions
Theoretical Foundations

Cognitive Penetration

Vision shall be studied with cognitive context, and the top-down influence can even penetrate to the early vision stages.


Action Grammar

X-bar theory was first proposed by Noam Chomsky (1970), building on Zellig Harris's 1951 approach to categories and further developed by Ray Jackendoff (1977).


Degrees of freedom problem (motor control)

Perception-Action integration (Visual Grounding)

Action Sequencing

Learning
1. SLICE

Slice → (Sliceable) Object

**Preconditions:**
- Objects:
  - X = (Slicing) (Graspable) (Tool) Object
  - Y = (Sliceable) (Graspable) Object
- Locations:
  - X at Loc1 = (p1, q1, r1)
  - Y at Loc2 = (p2, q2, r2)

**Postconditions:**
- Objects:
  - Y → Consequence DIVIDE
- Locations:
  - X at Loc2

--

Move X
(repeat action n times counter i)

**Preconditions:**
- Objects:
  - X = (Slicing) (Graspable) (Tool) Object
  - Locations:
    - X at Loc3
  - Preconditions:
    - (if i<n)
    - X moves Loc3 → Loc2 → Loc3
    - (else)
    - X moves Loc3 → Loc2

Grasp Y

**Preconditions:**
- Objects:
  - Y = (Sliceable) (Graspable) Object
  - Locations:
    - Y at Loc2

**Postconditions:**
- NULL

Grasp X

**Preconditions:**
- Objects:
  - X = (Slicing) (Graspable) (Tool) Object
  - Locations:
    - X at Loc1

**Postconditions:**
- NULL

Move X

**Preconditions:**
- Objects:
  - X = (Slicing) (Graspable) (Tool) Object
  - Locations:
    - X at Loc1

**Postconditions:**
- X moves (p1, q1, r1) → Loc3 = (p2, q2, r2+H)
Robotic Execution

Action Sequencing

Visual Grounding

Sensory Input

CVPR 15, ongoing

CVPR 13

ICRA 15

CogSys 14, AAAI 15

ACL 15, ongoing

CVPR 13
“...An agent is an active perceiver if it knows why it wishes to sense, and then chooses what to perceive, and determines how, when and where to achieve that perception.”

Vision + Language

Two cows in a field grazing near a gate.
The large cows hover over the young calf.
Three adult cows and one baby cow stand on the grass.
Three brown cows and a small calf in a field.
Three cows in a green pasture surrounding a baby cow.

Speech/Text Generation

EMNLP 11

ICRA 12

ICRA 13

ICCV 11
Image Captioning
Image Question Answering
Cognitive Dialogue

[Parikh et. al.]
DeepIU: Deep Vision + Commonsense reasoning

Somak Aditya, Chitta Baral, Yezhou Yang etc.
Manipulation actions
Robot Procedural Learning by Observation

Voice by Stratis Aloimonos
The Robot Visual Learner
by
UMD's
Robot Training Academy
Active Vision/Perception
Recognition or Prediction?
Proactive Vision

Grasp preparation ... Touching Action begins Action ends

Time

$T_0$ $T_0 + \tau_{\text{predict}}$ $T_0 + \tau_{\text{classification}}$

Prediction Delay: $\tau_{\text{predict}}$

Classication Delay: $\tau_{\text{classification}}$

PROACTIVE!

C. Fermuller, F. Wang, Y. Yang et. al. Under Review
Manipulation Action Prediction

The following video shows the results of the RNN for continuous classification (over 25 object-action pairs)
Recognition or Sensation?
Hand Force Estimation

The following video shows the results of the RNN as regressor (of the six forces on the palm and finger tips)
yz.yang@asu.edu