Scene Representation for Robots: From Elementary Behaviors to Grasping

Stephan Zibner
Outline

• Representing Scenes
  – In Humans
  – Robotic Scenario

• Building Blocks

• Elementary Behaviors
  – Exploration
  – Maintenance
  – Query

• Reaching and Grasping
Representing Scenes
Scene Representation

- internal representation of environment
- foundation for every higher cognitive operation and action
- stable despite eye and body movements
- limited capacity, link to long-term memory
Scene Representation in Humans

- human eye movement is not random (saliency)

Bruce, Tsotsos (2009)
Scene Representation in Humans

- Human eye movement is not random (saliency)
- Visual search highlights target objects

Bottom: Treisman, Gelade (1980)
Scene Representation in Humans

- Human eye movement is not random (saliency)
- Visual search highlights target objects
- Details of a scene are kept in memory

Hollingworth (2005)
Scene Representation in Humans

- Human eye movement is not random (saliency)
- Visual search highlights target objects
- Details of a scene are kept in memory
- No full representation (change blindness)
Scene Representation in Humans

- human eye movement is not random (saliency)
- visual search highlights target objects
- details of a scene are kept in memory
- no full representation (change blindness)

attention is a key theme in visual processing
DFT Model

- model of human visual working memory based on dynamic neural fields
- representation of low-level features
- decomposition of features, binding through shared space

Johnson, Spencer, Schöner (2008)
Robotic Scenario

- apply to table-top scenario and human interaction
- use the internal representation for behavior generation (e.g., grasping)
- interact with humans (“hand me the red screwdriver”, “what's to the left of the pliers?”)
Robotic Scenario

- explore the environment and store objects and their features internally
- maintain the internal representation
- query the representation to create autonomous action based on the representation
Robotic Scenario: Challenges

- real sensory input
- moving sensors
- limited field of view
- 3D space
- dynamic scenes
- multiple behaviors
- computational constraints
- and many more ...
Building Blocks
Saliency

Itti, Koch, Niebur (1998)
Saliency

- on-/off-center responses
- uniform regions result in zero responses
- objects fitting into on-center region produce non-zero responses
- these lead to detection decision in fields
Feature estimates are linked to spatial positions.
Autonomy of Behaviors

- elementary cognitive units (ECU)
- intention node boosts fields
- CoS node detects completion of behavior in fields
- sequences or exclusions through precondition and suppression nodes
Change Detection with Fields

Excitatory input

"no change"

Inhibitory input

"change"
Elementary Behaviors of Scene Representation
Architecture Overview
Exploration Behavior

- **Input**
- **Saliency**
- **Attention**
- **Feature Extraction**
- **Location**
- **Representation**
Autonomy of Exploration

- Input
- Saliency
- Attention
- Feature extraction
- Location
- Representation
- To ECU
Maintenance Behavior

working memory vs. updating

tracking
Video
Query Behavior

- **Input**
- **Bottom-up saliency**
- **Attention**
- **Top-down saliency**
- **Feature cue**
- **Representation**
Autonomy of Query

bottom-up saliency

attention

top-down saliency

input

feature cue

representation

to ECU
Reaching and Grasping
Challenges

- focus attention
- estimate shape
- multiple behaviors
- arm movement
Behaviors Involved in Grasping

Perception
- Query
- Exploration
- Scene Representation
- Object Recognition
- Classification
- Pose Estimation

Action
- Approach
- Open Hand
- Grasp
- Lift
From Camera to a Height Map
From Height Map to Grasp Parameters
Grasp Execution

Legend:
- ○ node
- connections:
  - → excitatory
  - • inhibitory

1. **open hand**
   - int
   - CoS
   - mem
   - hand open?

2. **approach**
   - int
   - CoS
   - mem
   - wrist at target?

3. **grasp**
   - int
   - CoS
   - mem
   - object grasped?

4. **lift**
   - int
   - CoS
   - mem
Grasp Execution

![Graph showing the execution of grasp with intention nodes: approach, open hand, grasp, lift. The x-axis represents time in seconds (0 to 20) and the y-axis represents intention nodes. The graph illustrates the duration of each step.]
Video
Take-home Message

- exploration, maintenance and query are the core behaviors of scene representation
- change detection is a driving force for autonomy
- integration with other DFT architectures yields complex behaviors such as grasping
- integration is facilitated by DFT framework
Thanks for your attention!