How Shall We Learn How to Learn How to Grasp?

Benjamin Kuipers
University of Michigan

Grasping Research

- Traditional robot grasping research focuses on geometric reasoning.
  - Geometric object shape, grasp points and forces, zero net force and torque, etc.

- Current directions (all very exciting):
  - Dexterous, compliant hands
  - Integrated haptic sensing
  - Learning to grasp

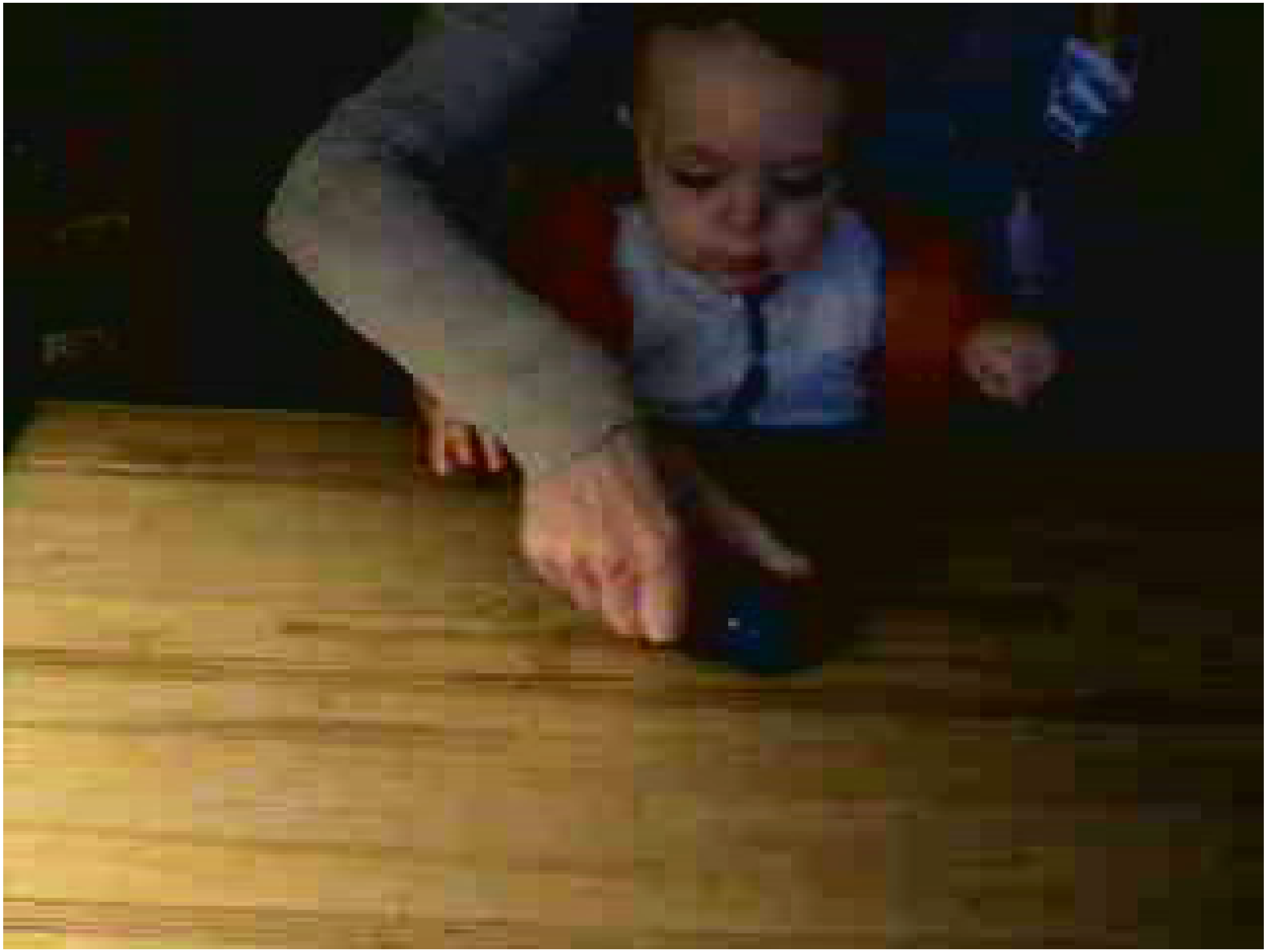
Learning How to Grasp

- It's a nice supervised learning problem.
  - The world provides supervision:
    - each attempt to grasp succeeds or not.
  - We still need problems easy enough to learn.

- Learning about objects, and learning about actions (including grasping), happen together.

- For clues on how to learn how to grasp
  - observe **the most powerful learning agent** the world has ever known.

The Most Powerful Learning Agent



What do we see here?

- The baby is eight months old
  - Quite competent, but still has a lot to learn.

- Compliant, whole-hand grasping
  - Trial and error, not careful planning
- Failure and recovery
- Hand-to-hand passing
- Exploration of object properties
- Exploration of actions

- Let’s explore a more complex object.

Exploring a more complex object

What do we see here?

- More trial-and-error grasping
  - Opportunistic capture of the object
  - Has he learned a strategy? Hard to say.
- More hand-to-hand passing
- Exploring the object in pose space

- Onward to tool use

Using an object as a tool

Exploring actions systematically

Exploring un-grasping

Exploring a new object

What can we learn from this?

- A short segment in a learning sequence:
  – the child is 8 months old;
  – the palmar reflex is long gone.
- Learning about objects and learning about actions are closely intertwined.
- Focused, attentive, autonomous learning
  – Exploration, not goal-oriented.
- Q: What is the intrinsic motivation that drives this exploration?
  – Leads to increasing competence.

What does the baby learn?

- What is the content of the knowledge that the baby acquires from its learning?
- Claim: this knowledge can be separated into distinct aspects.
  - A "*semantic hierarchy*" provides a finer-grained description of the types of knowledge involved.
  - Learning each level of the hierarchy is easier.

The Spatial Semantic Hierarchy

- Distinguish *scales of behavioral space*.
  - **Small-scale space**
    - Within the agent's sensory horizon
  - **Large-scale space**
    - Beyond the agent's sensory horizon
- Distinguish *ontologies for spatial maps*.
  - **Metrical mapping**:
    - Within a single frame of reference, define location, heading, pose, distance, and shape.
  - **Topological mapping**:
    - Places, paths, and regions are related by connectivity, order, and containment.

Local Metrical Mapping Works

- In small-scale space, modern SLAM methods work extremely well with lasers.
  - Great progress with visual SLAM.

	Metrical Mapping	Topological Mapping
Small-scale space	Local SLAM	
Large-scale space		

Global Metrical Mapping Is Hard

- Within a single global frame of reference over large-scale space, errors accumulate.
  - Sufficiently large loops are always a problem.

	Metrical Mapping	Topological Mapping
Small-scale space	Local SLAM	
Large-scale space	Cumulative errors Scalability	

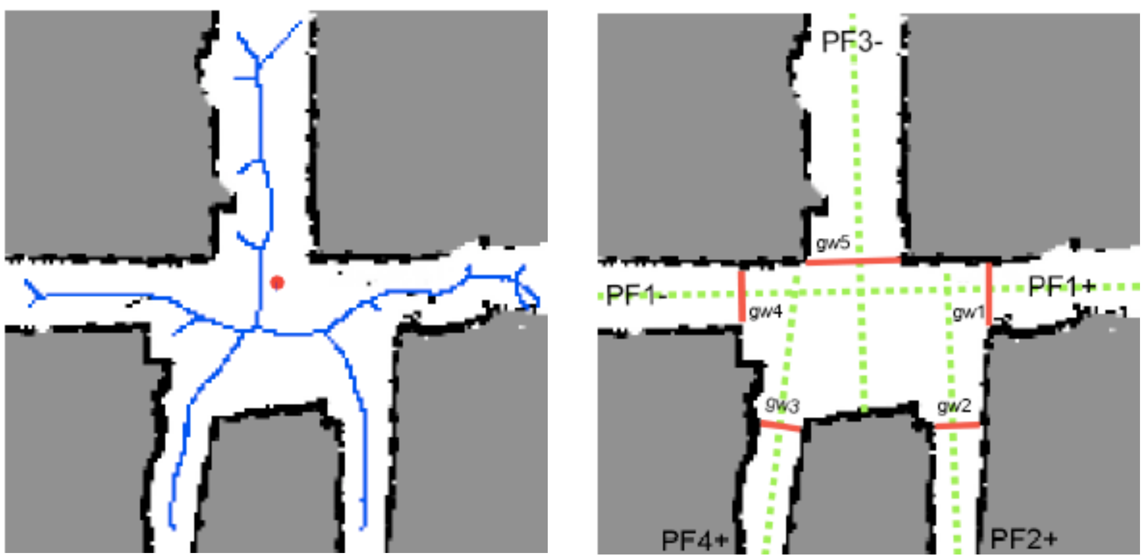
Identify the Local Topology

- Identify the local decision structure of each place neighborhood.
  – Travel experience as graph exploration

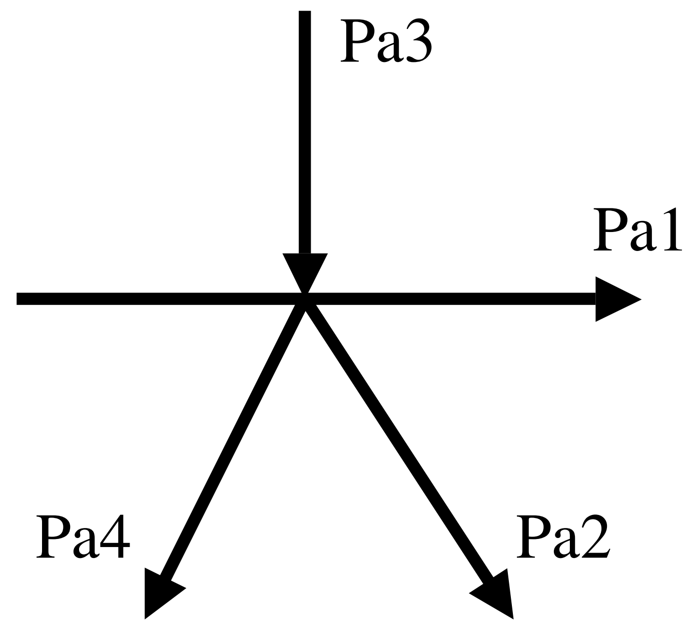
	Metrical Mapping	Topological Mapping
Small-scale space	Local SLAM →	Local decision structure
Large-scale space		

Local Decision Structure

- Identify *gateways* and *path fragments*
  - 2 gateways & 1 path fragment ⇒ on a path
  - Otherwise ⇒ at a place neighborhood



in small-scale space



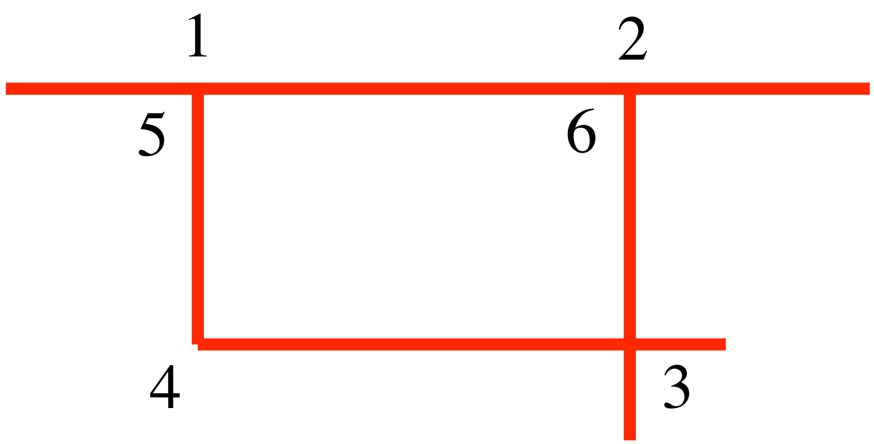
in large-scale space

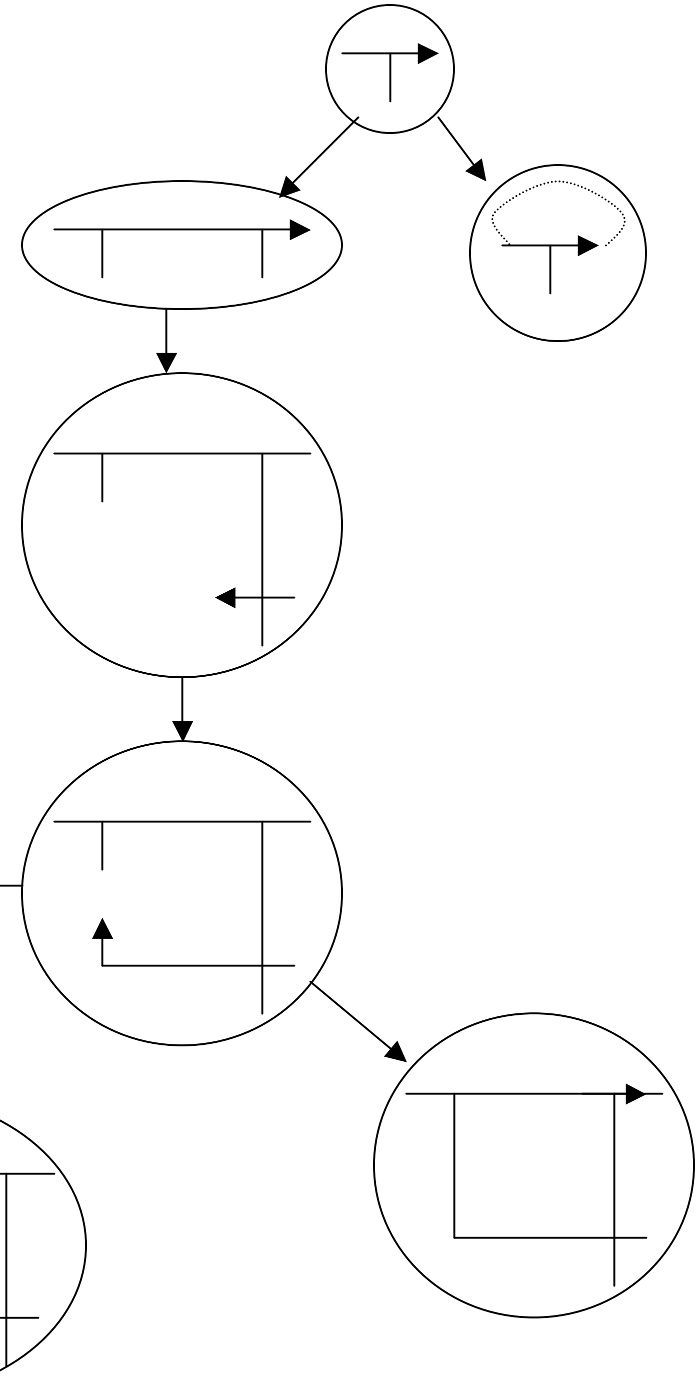
Build the Global Topological Map

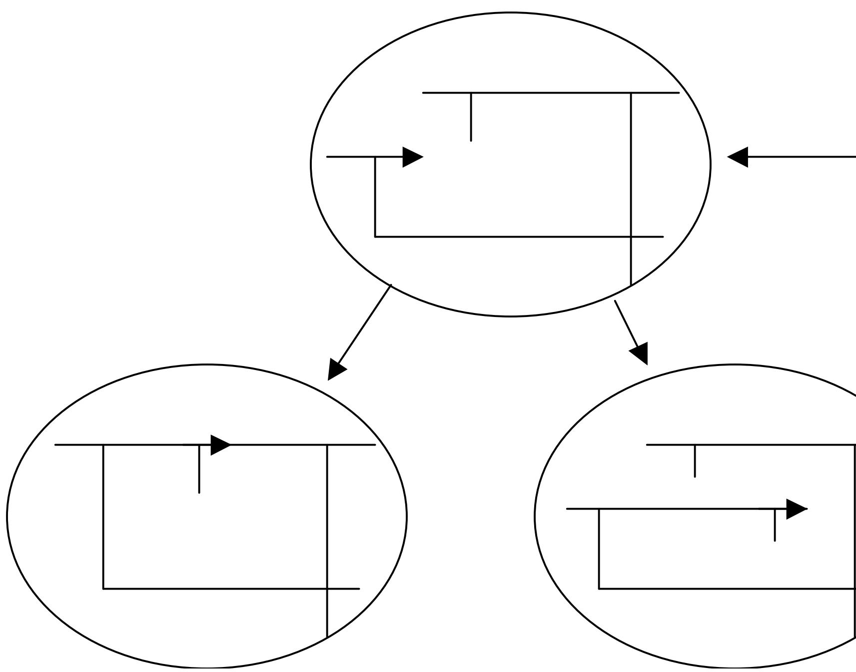
- Decide when and how loops are closed
  - When does the next place match a previous place?
- Build a tree of all possible topologies

	Metrical Mapping	Topological Mapping
Small-scale space	Local SLAM	Local decision structure
Large-scale space		Global topological map

Tree of Maps





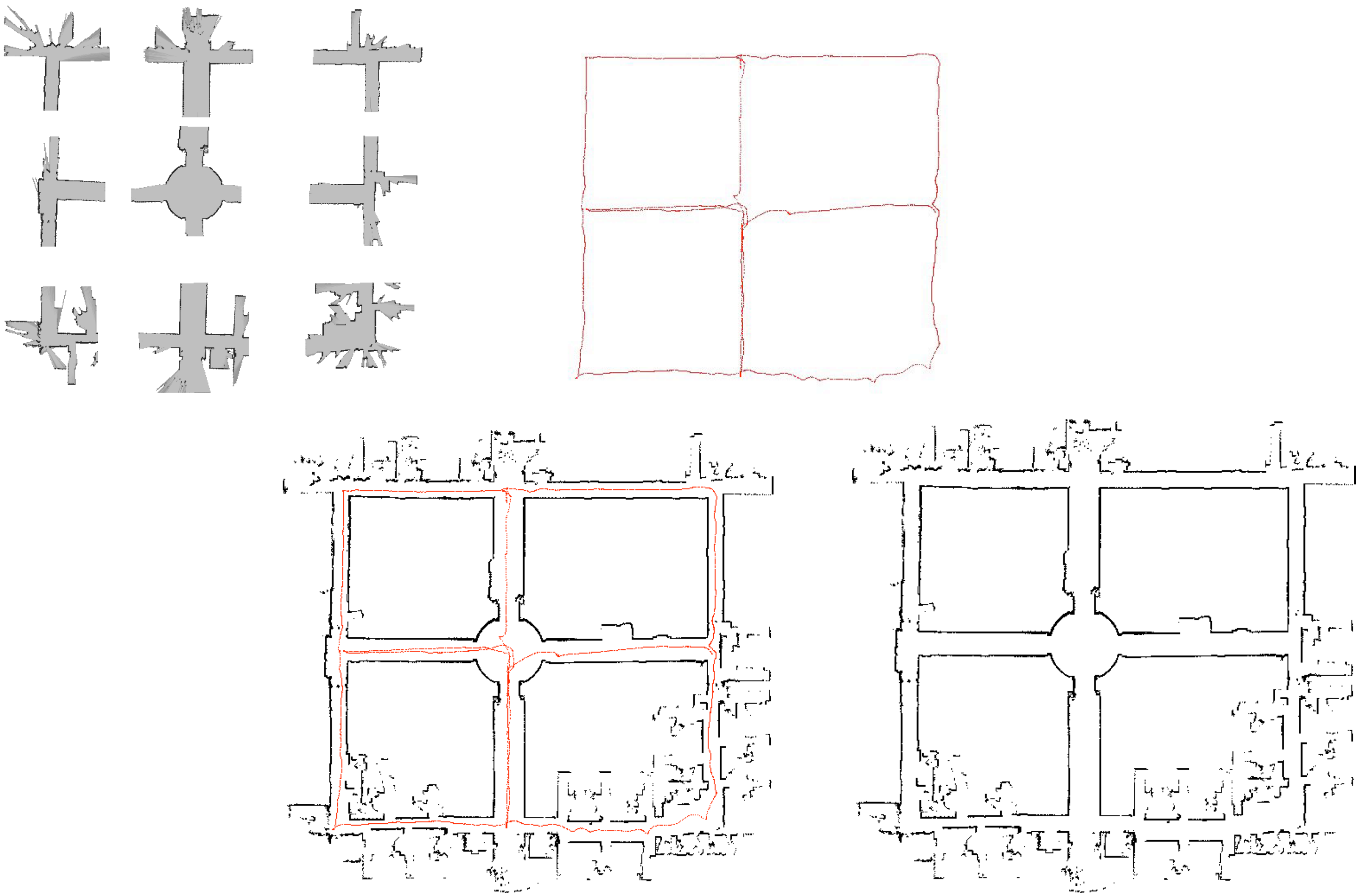


Global Metrical Map

- Use the topological map as a skeleton.
  - Lay out places in a single global frame of reference.
  - Fill in the details from local places and segments.

	Metrical Mapping	Topological Mapping
Small-scale space	Local SLAM	Local decision structure
Large-scale space	Global metrical map	Global topological map

Build the Global Metrical Map on the Topological Skeleton



What have we got?

- Four representations for navigable space
  - Agent can learn them, or be told

	Metrical Mapping	Topological Mapping
Small-scale space	Local map for safe motion	Well separated decision points
Large-scale space	Heuristics to guide planning	Scalable map for route planning

Human-Robot Interaction

- Different kinds of human instructions map to different spatial knowledge representations

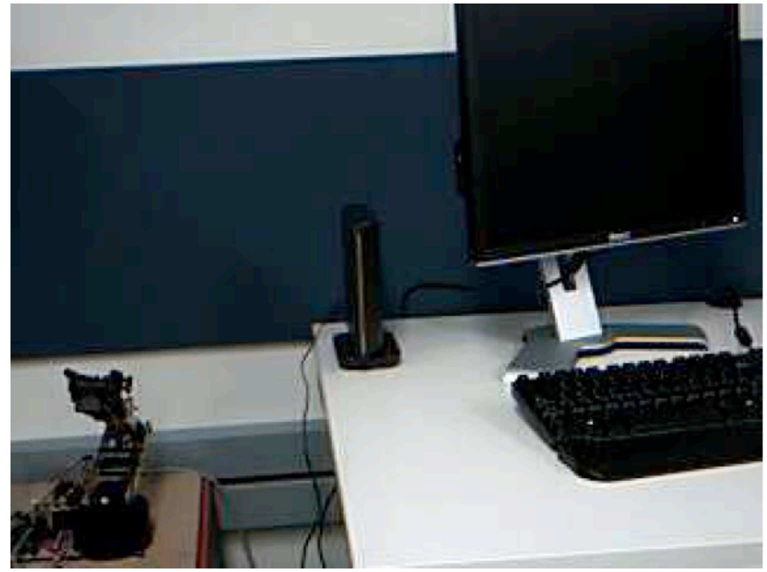
	Metrical Mapping	Topological Mapping
Small-scale space	*"Go there"* *"To my desk"*	*"Turn right"* *"Second left"*
Large-scale space	Select map point	*"To the kitchen"* *"Doctor's office"*

Hierarchies of Representations

- Spatial Semantic Hierarchy (SSH)
    - Local Metrical Mapping
    - Local Topological Maps
    - Global Topological Map
    - Global Metrical Map
  - [Beeson, Modayil & Kuipers, IJRR, 2010]

- Object Semantic Hierarchy (OSH)
    - Static 2D background model
    - 2D foreground object in 2D image space (2D2D)
    - 2D foreground face with pose in 3D space (2D3D)
    - 3D object with 2D faces in 3D space (3D3D)
  - [Xu & Kuipers, 2009; Xu, Kuipers & Murarka, 2009]

Build a Static Background Model

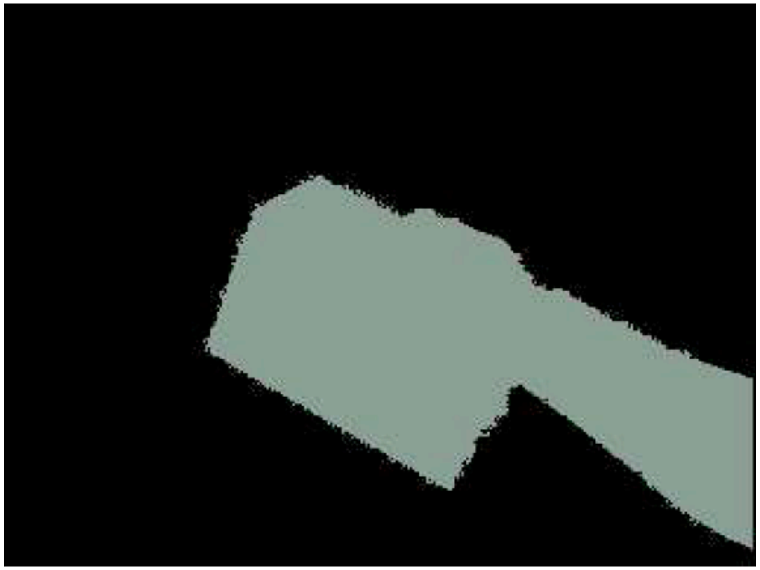


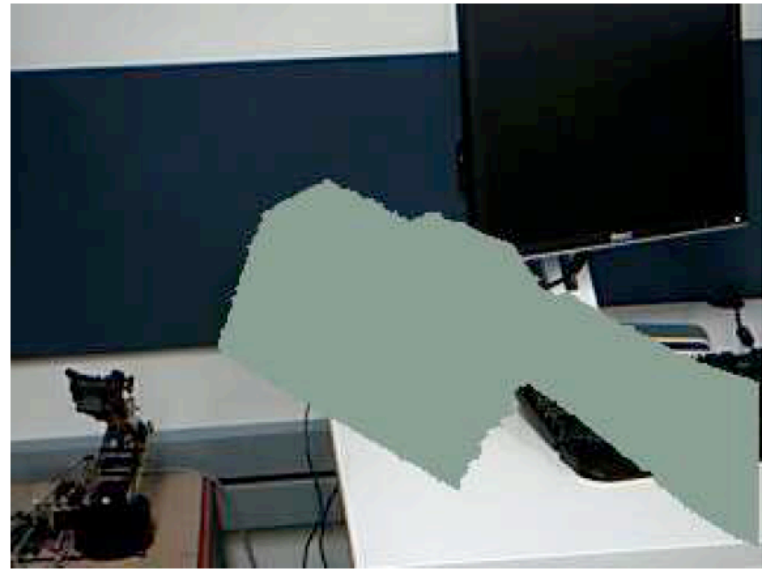
- Identify background model b. x_t is constant.

$$z_t = G_1(b, x_t) + \epsilon_1$$

- Foreground is treated as noise: part of ε_1.

Identify Foreground Object





- Cluster and track in ε_1 to define foreground.
  - Identify constant property (e.g., average color)
    - Deviations from average color are part of ε_2.
  - Shape and pose are time-varying properties.

$$z_t = G_2(b, o, x_t, s_t) + \epsilon_2$$

Identify Planar Facets on Objects

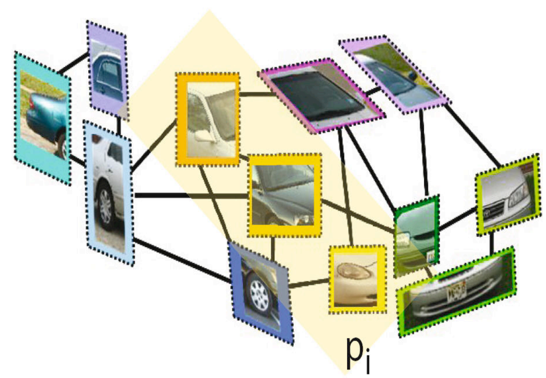
- Many objects have near-planar facets.
  - Identify c: constant surface texture and boundary, and time-varying surface normal (pose q_t).

$$z_t = G_3(b, c, x_t, q_t) + \epsilon_3$$

3D Constellations of 2D Facets

- 2D object facets with constant relative pose can represent 3D object shape
  - Constant 3D object model m.
  - Only 3D object pose x_{ot} is dynamic.

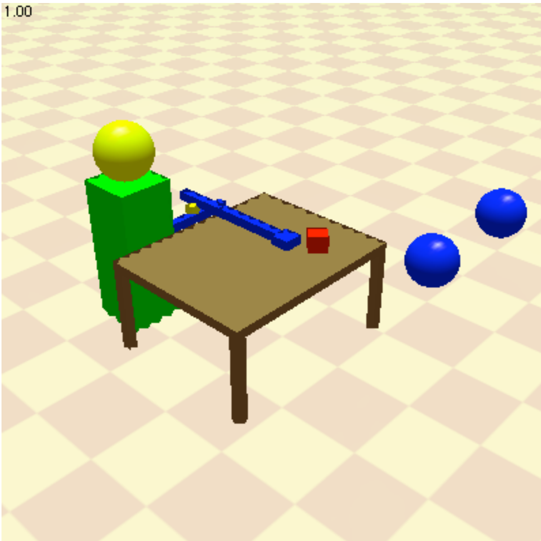
$$z_t = G_4(b, m, x_t, x_{ot}) + \epsilon_4$$

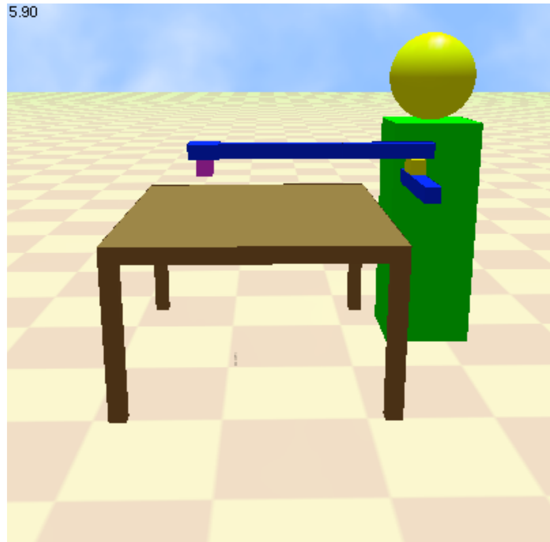


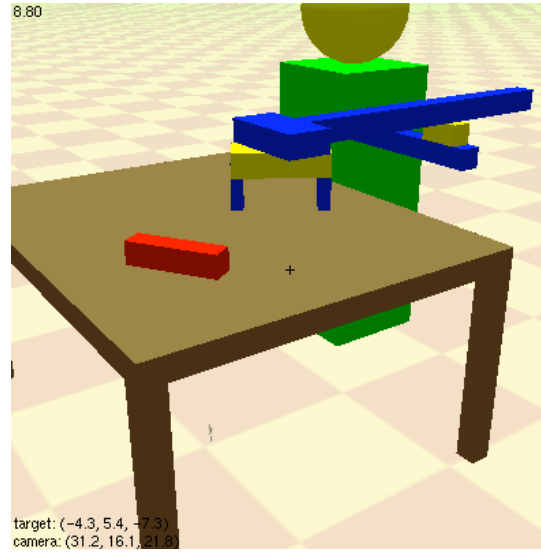
An *Action* Semantic Hierarchy?

- Identify *contingencies*: $\langle C \ : \ E_1 \Rightarrow E_2 \rangle$
  - *Events* E_k are defined by qualitative state changes.
  - Learn new landmarks for better qualitative states.
- Learn *contexts* C to increase reliability.
  - Express as dynamic Bayes nets (DBNs).
- Define *options* for reinforcement learning:
  - how to achieve a reliable context,
  - then act to make the antecedent event happen,
  - to cause the consequent event: an *action*.
- QLAP: [Mugan & Kuipers, 2007, 2008, 2009]

Actions in Visual Space

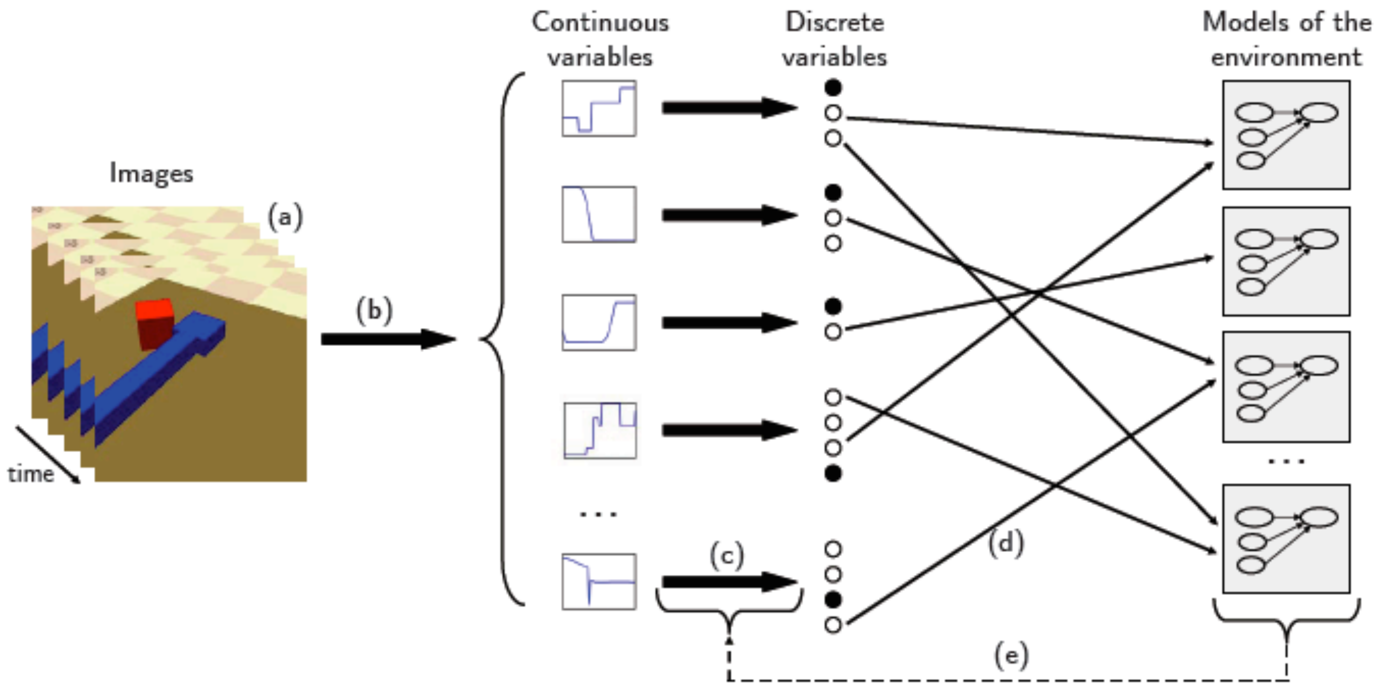






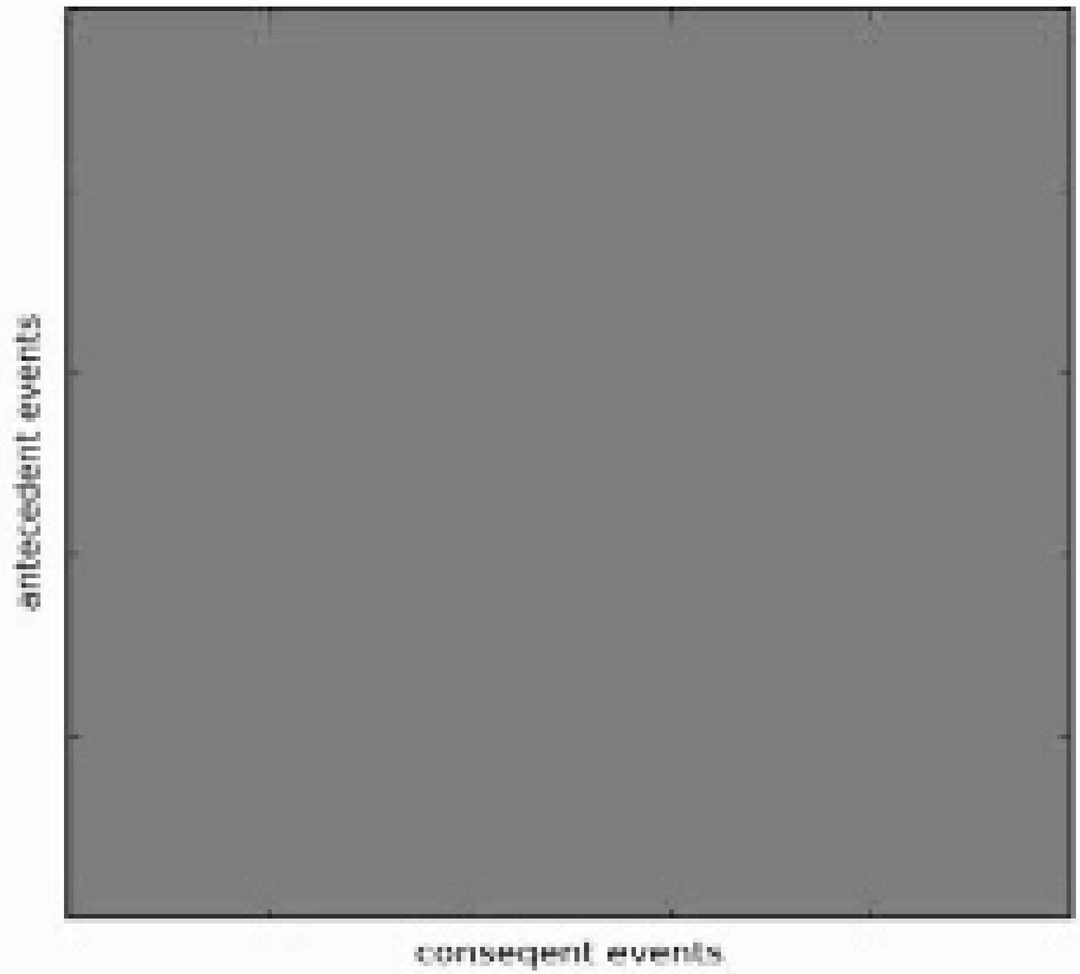
- Agent needs to learn how to assemble low-level motor signals into high-level actions.
  - Starting from "blooming, buzzing confusion"
- Objects are identified based on sensory changes.
  - Actions are what make changes happen.
- Work with Jonathan Mugan (U. Texas, Austin)

Learn Action Models



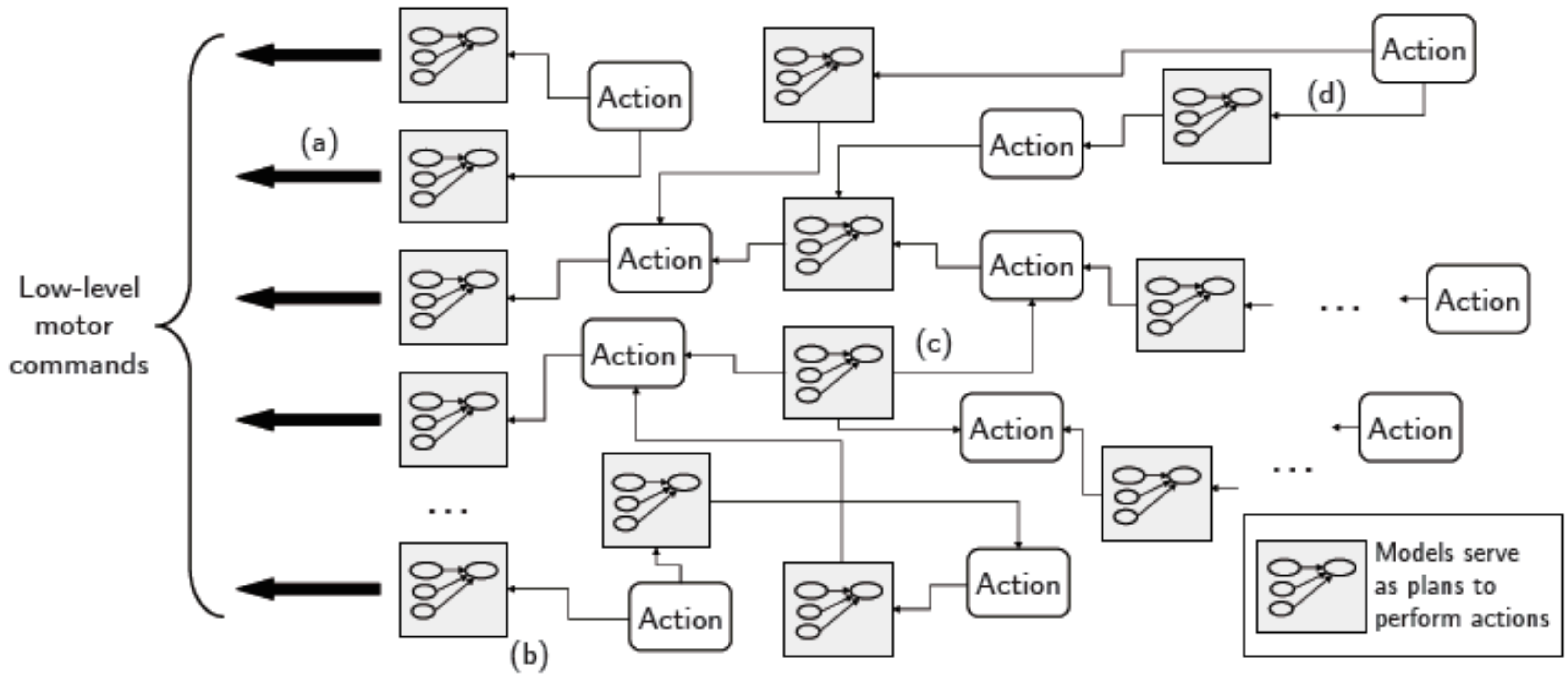
- Learn dynamic Bayesian network (DBN) models for the effects of actions.
- Once reliable (i.e., low entropy), define *options* to achieve effect of DBN models.
  - These options define the next higher level of actions.

Learning Contingencies $\langle C \ : \ E_1 \Rightarrow E_2 \rangle$



- Grey: reliability; Yellow: contingency; Green: deterministic

Learn to Make Hierarchical Plans



- Lowest-level actions defined directly with motor signals.
- Higher-level actions defined as options over lower-level actions.
- And so on.

Growing graph of goals and actions

- The graph grows with experience.

Grasping is a Special Action

- We are only just beginning to learn to grasp.
- But this is how to learn actions, we think, including grasping.
  - Learn like a (stereotyped) baby.
  - Start without semantics, even of body space.
  - Learn contingencies from motor babbling.
    - The *palmar reflex* helps constrain the search space.
  - Search for near-deterministic contingencies.
  - Learn a lattice of actions and events.
- A semantic hierarchy for actions.

References

- Beeson, Modayil & Kuipers, Factoring the mapping problem: Mobile robot map-building in the Hybrid Spatial Semantic Hierarchy. *IJRR*, 2010.
- Xu & Kuipers, Construction of the Object Semantic Hierarchy. *International Cognitive Vision Workshop (ICVW),* 2009.
- Xu, Kuipers & Murarka, 3D pose estimation for planes. *ICCV Workshop on 3D Representation for Recognition (3dRR),* 2009.
- Mugan & Kuipers, Autonomously learning an action hierarchy using a learned qualitative state representation. *IJCAI*, 2009.
- Mugan & Kuipers, Toward the application of reinforcement learning to undirected developmental learning. *EpiRob*, 2008.
- Modayil & Kuipers, The initial development of object knowledge by a learning robot. *Robotics & Autonomous Systems*, 2008.
- Mugan & Kuipers, Learning to predict the effects of actions: Synergy between rules and landmarks. *ICDL*, 2007.
- Savarese & Fei-Fei, 3D generic object categorization, localization and pose estimation. *ICCV*, 2007.
- Kuipers, The Spatial Semantic Hierarchy. *AIJ*, 2000.
- Pierce & Kuipers, Map learning with uninterpreted sensors and effectors. *AIJ*, 1997.

Learning to Grasp is Hard

- Infants take months to learn to grasp well.
  - Evolution provides the “palmar reflex”
    - Tickling the palm causes flexion of the fingers.
    - Present at birth. Disappears by 5-6 months.
  - Reduces the “motor babbling” search space.
    - Random motion causes contact with object.
    - Palmar reflex causes finger flexion: grasp.
    - Learn effect of grasp action.
    - Learn to grasp as intentional action.

Levels of object representation

- Multiple representations provide robustness.
  - "Blooming, buzzing confusion"
    - $z_t = \epsilon_0$
  - Static background model
    - dynamic change is treated as noise.
  - 2D foreground objects against 2D background
    - "blobs" are easy to track and characterize
  - each 2D object face has a pose in 3D space
    - learn constant surface properties of each face
  - 3D objects are constellations of 2D faces
    - learn constant relations between face poses

    - Work with Changhai Xu (U. Texas Austin)