Beyond What and Where:
Reasoning Function, Physics, Intent and Causality

Song-Chun Zhu

Center for Vision, Cognition, Learning and Arts
University of California, Los Angeles

Outline

1. Motivation: AI, Commonsense and Robotics
2. Introduction: Function, Physics, Intents, and Causality
3, Unified Representation: STC-And-Or Graph
4, Joint Inference: spatial, temporal and causal parsing.
5, Discussion: challenges and future work

1. Motivation

Motivation 1: Beyond What and Where

Vision is to find what are where by looking

--- David Marr's book 1982.

Elementary school experiments on dogs




Motivation 1: Beyond What and Where

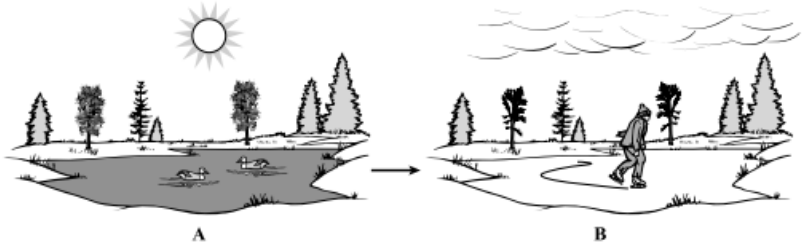
Know how

The crow must understand the scene with geometric, functional and physical relations.
Knowing the material property of the metal stick, making the hook, using the hook, ….

Motivation 2: Reasoning How, Why, What If

Question in a 5th Grade Test

Need joint reasoning using Vision + Language + Cognition (commonsense)



Which of the following has caused the changes in the pond from A to B?

A. The pond water has lost heat energy.
B. The pond water temperature has increased.
C. Warm water has risen to the top of the pond.
D. All of the water has evaporated from the pond.

Motivation 3: Filling the ROC gaps and Generalizing to reasoning unseen cases

"Dark Matter" and "Dark Energy"

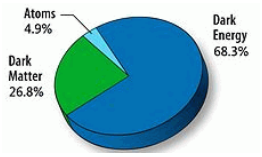
Function, Physics, Intents and Causality (FPIC)

↓

Geometry, 1970-1990

↓

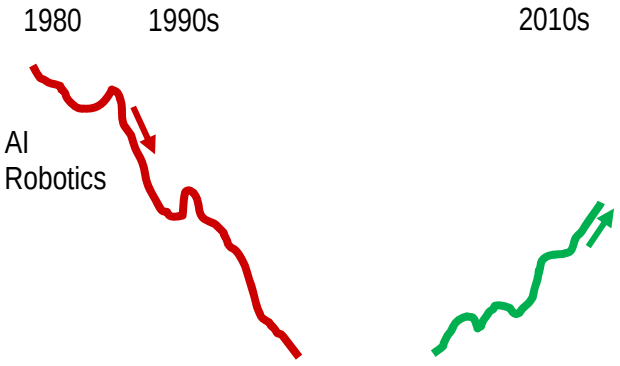
Appearance, 1990-2010



Vision must compute the visible and the dark jointly.

Motivation 4: AI and Robotics are Bouncing Back

In the future, vision will increasingly interact with AI and robotics.



2. Introduction to
Function, Physics, Intent, and Causality

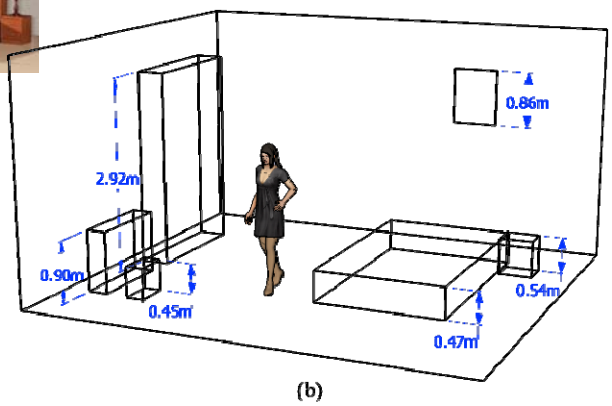
1, Functionality: Re-defining Scenes and Objects

Most scenes are functional spaces that serve human activities.

Most objects are functional entities that assist human actions.

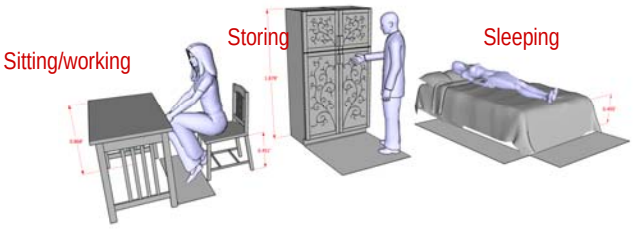
Reasoning scene functionality

Functionality = imagined human actions in the dark !

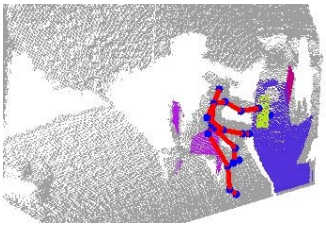


Y. Zhao and S.C. Zhu, "Scene Parsing by Integrating Function, Geometry and Appearance Models," CVPR, 2013.

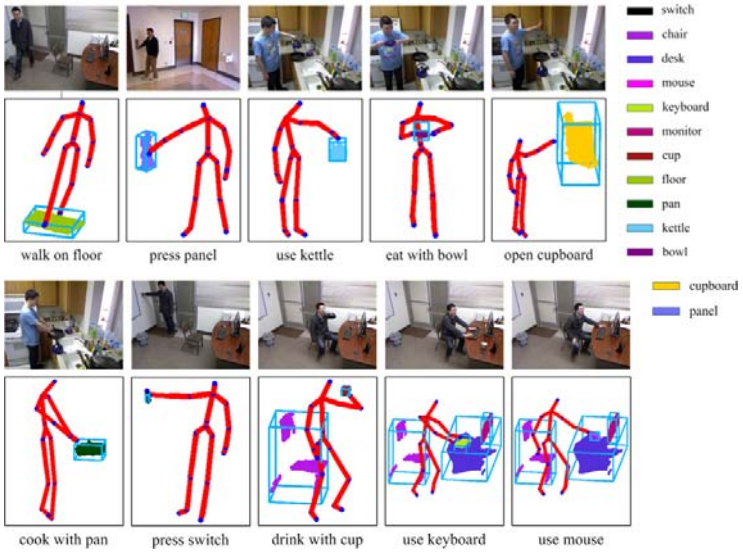
Functionality = imagined human actions in the dark



One can learn these relations from Kinect RGBD data and use them for reasoning.



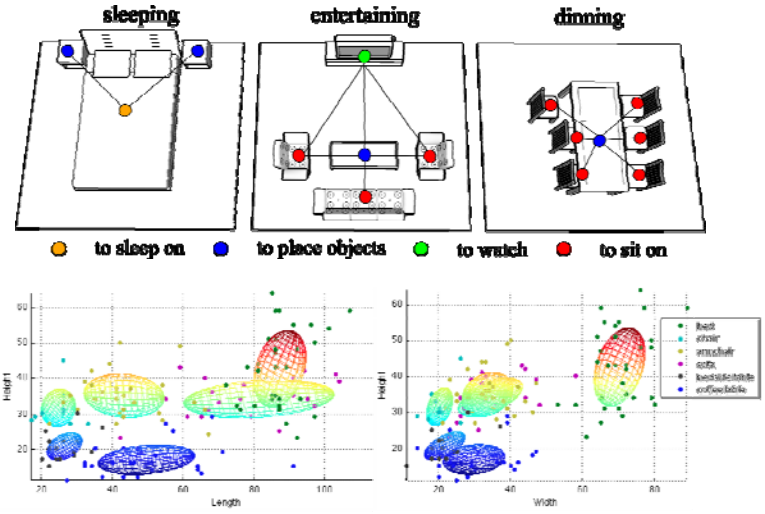
Learning the function and affordance in RGBD video



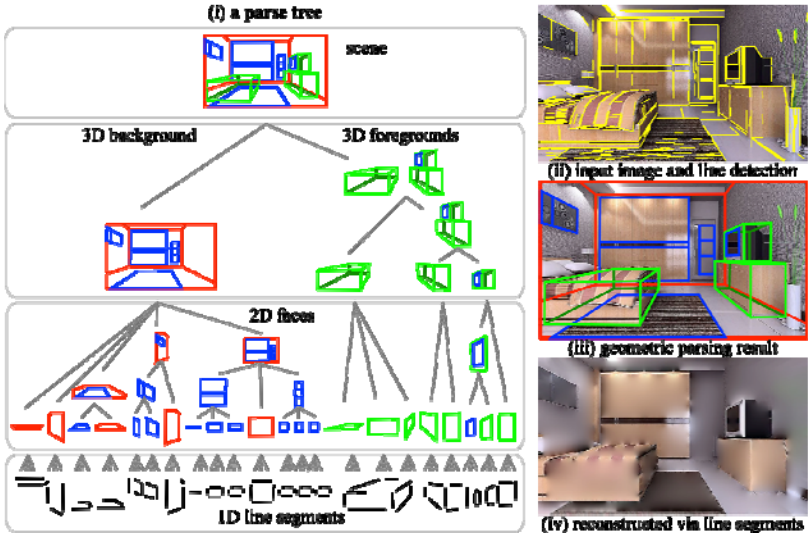
P. Wei et al. "Modeling 4D human object Interactions for event and object recognition ," ICCV, 2013.
Other groups: A. Gupta, M. Hebert, A. Saxena, Y. Wu et al.

Representing human-furniture relations in simulated actions

These relations are the grouping "forces" for the layout of the scene. (C. Yu et al Siggraph 2012)

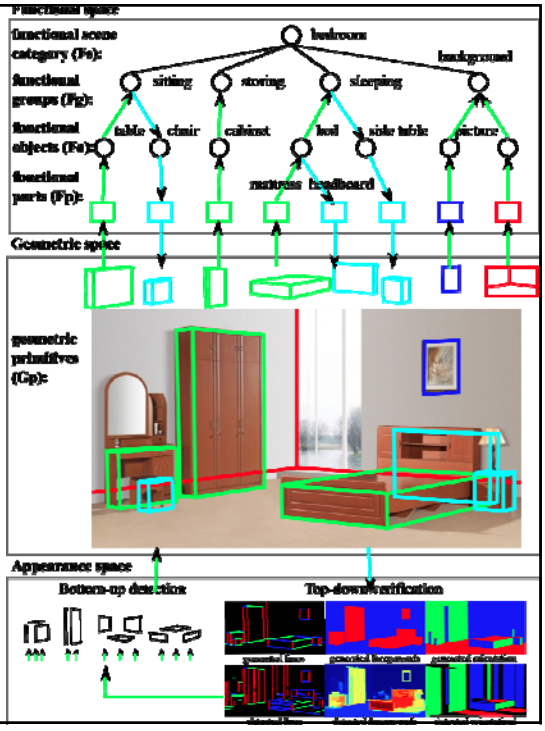


Syntactic 3D scene parsing from a single image



Y. Zhao and S.C. Zhu, "Image Parsing via Stochastic Scene Grammar" NIPS, 2011.

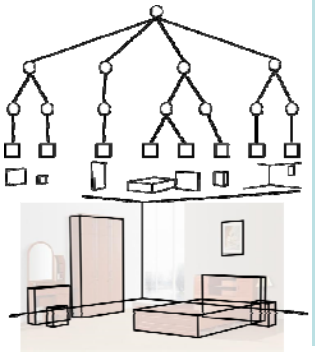
Bottom-up / Top-down inference by MCMC

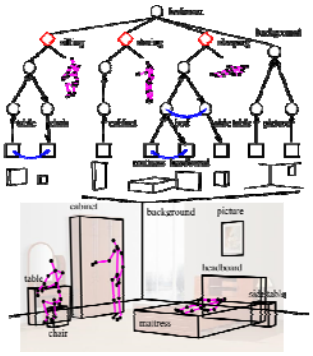


Augmenting Syntactic Image Parsing with Functionality

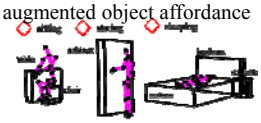
syntactic parse tree

augmented parse graph

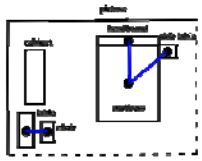




augmented object affordance



augmented contextual relations

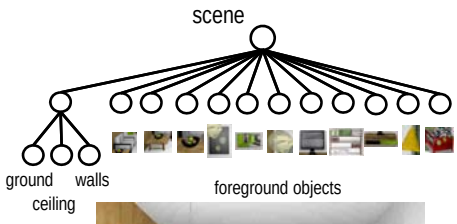


syntactic scene parsing can be dated to K.S. Fu in 1978.

Y. Zhao and S.C. Zhu, "Scene Parsing by Integrating Function, Geometry and Appearance Models," CVPR, 2013.

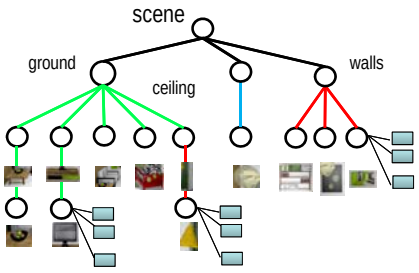
2, Physics plays a key role in image/scene understanding

parse tree



foreground objects

augmented parse graph



Augmented physical properties:

- material, friction, mass, velocity

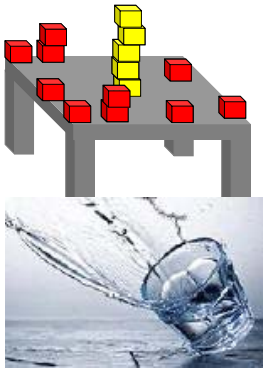
Augmented physical relations:

- supporting, attaching, hanging

B. Zheng et al CVPR 2013. ICRA2014. [pdf]

Intuitive Physics in Cognitive Modeling:

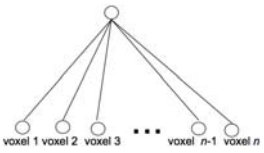
How stable are the objects?
What will happen next?

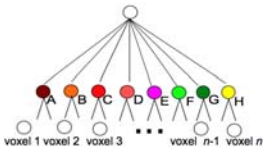


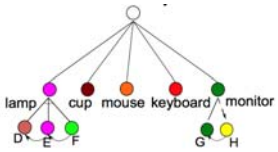
From Josh Tenenbaum

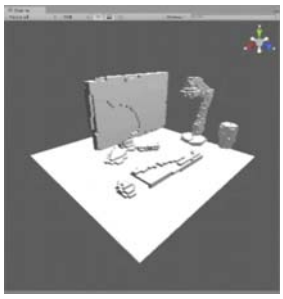
Physical constraints help scene parsing,

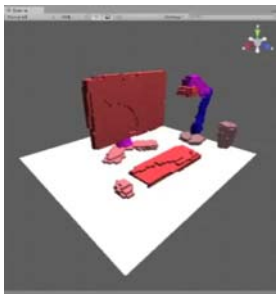
i.e. a valid parse (interpretation) must be physically plausible.

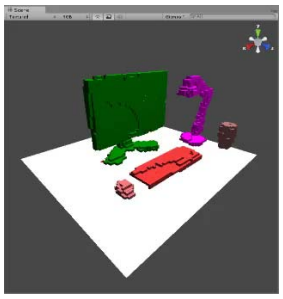






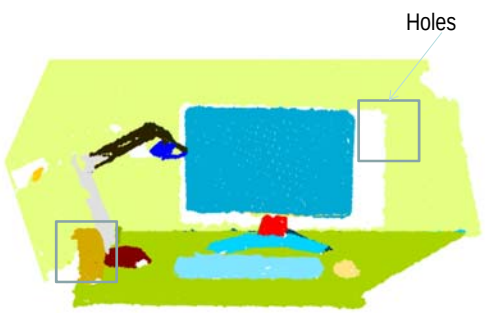


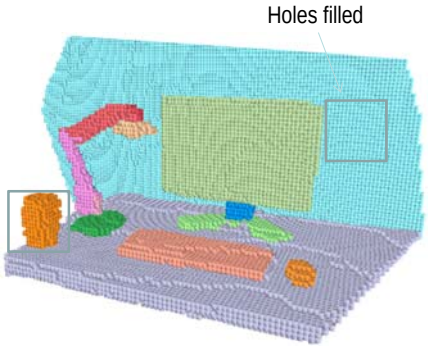




By grouping the voxels (captured by depth sensor) into geometric solids (parts) and then into Object (segmentation), so as to minimize physical instability, and maximizing functionality to serve humans.

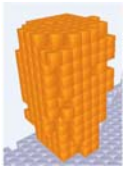
Parsing and grouping from point clouds



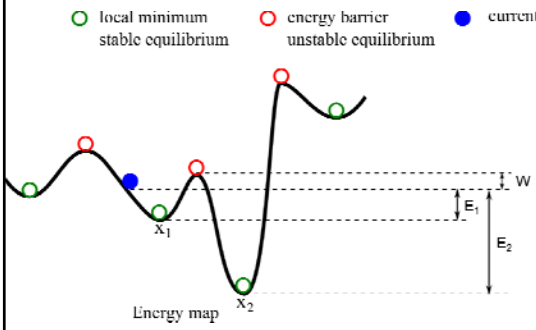


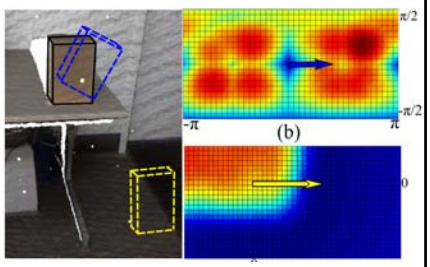
Segmentation result

Result of volumetric completion

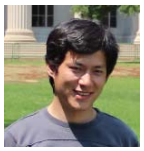


Defining instability in gravity field



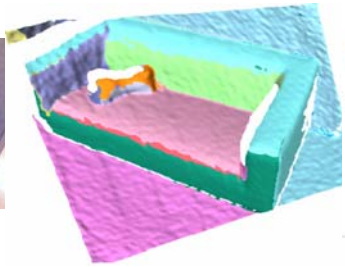
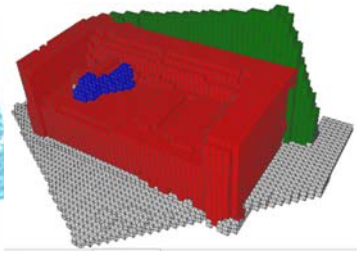


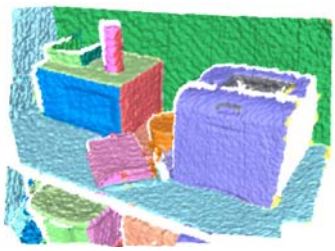
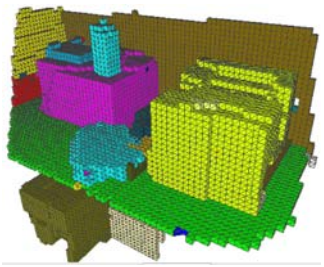
$$\mathrm{S}(a, \mathbf{x}_0, W)$$
$$= \max_{\mathbf{x}'_0} \triangle\mathcal{E}(\tilde{\mathbf{x}} \rightarrow \mathbf{x}'_0)\delta([\min_{\tilde{\mathbf{x}}} \triangle\mathcal{E}(\mathbf{x}_0 \rightarrow \tilde{\mathbf{x}})] \leq W)$$



B. Zheng, Y. B. Zhao et al. “Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics,” CVPR 2013.

Physical reasoning:
minimizing instability and maximizing functionality

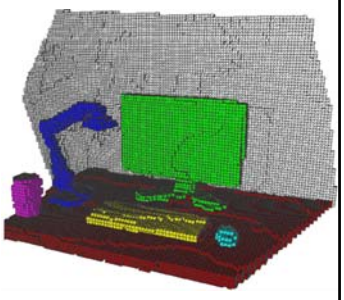



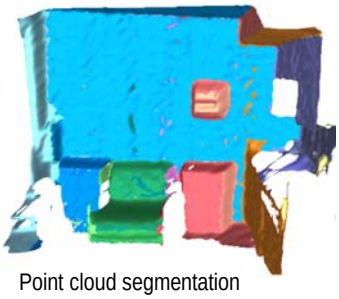
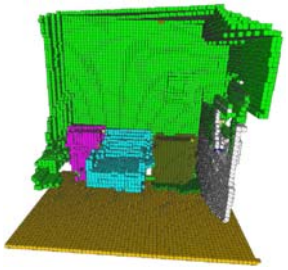
Point cloud segmentation

results

Results of physical reasoning



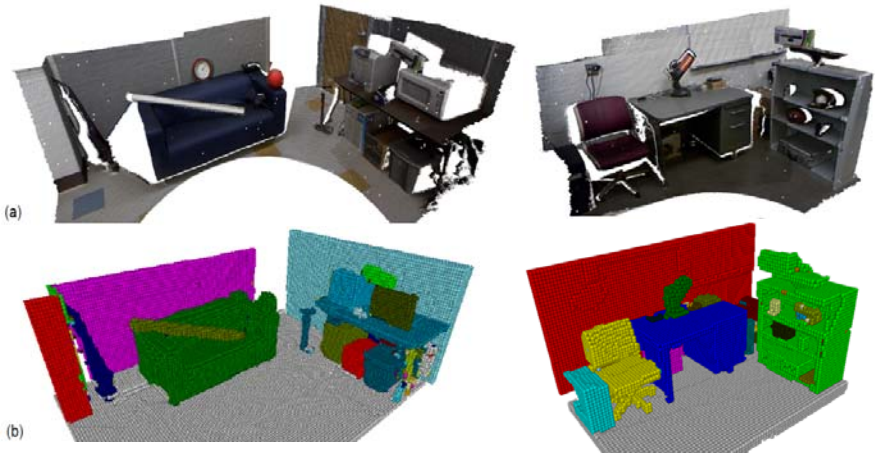


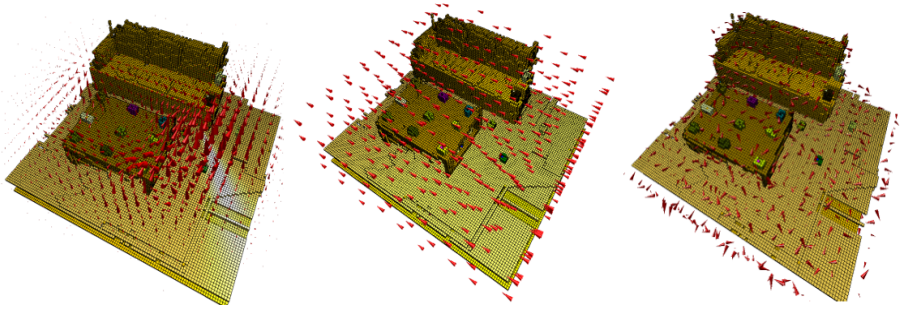
Point cloud segmentation

results

More results of physical reasoning

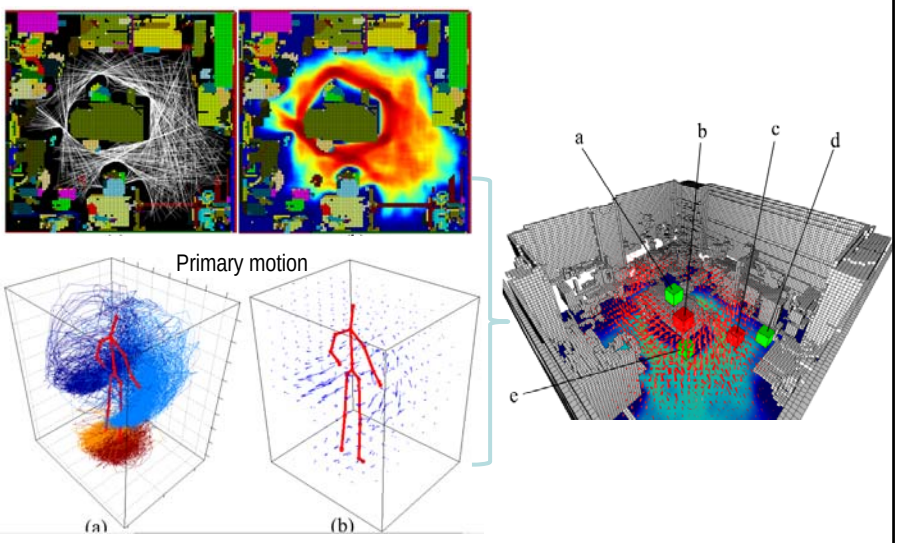


Other physics in scenes: earthquake, gust, human activities

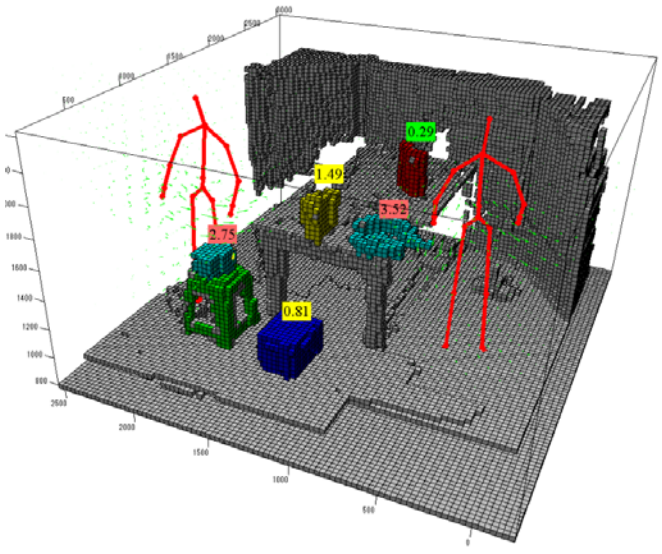


human activity wind earthquake

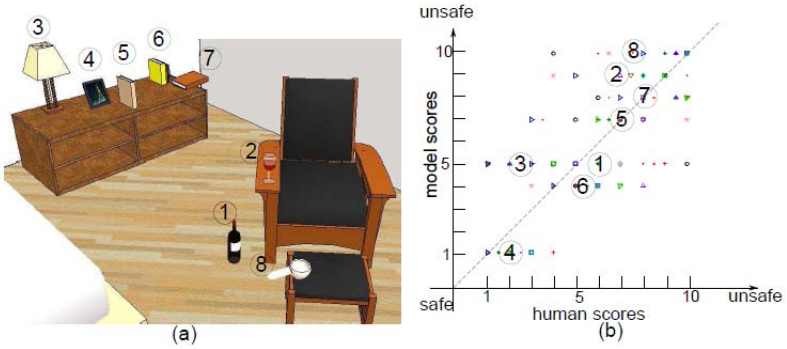
Modeling the disturbance field by human activities



Evaluating the risks



Which objects are most risky and unsafe?



B. Zheng, Y.B. Zhao et al. ICRA 2014.

Applications and future work

Robot rescue

Baby-proof scene

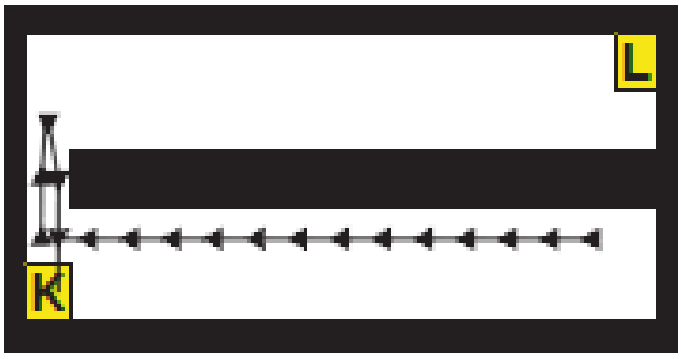
3, Intents: Reasoning intents of agents and predicting their actions

What is the pig doing? How did you figure out?

A teleological stance for scene understanding (many work in cognitive psychology).

Intents reasoning:

which food truck is the most favorite, K, L, or M?



From Baker, Saxe & Tenenbaum.

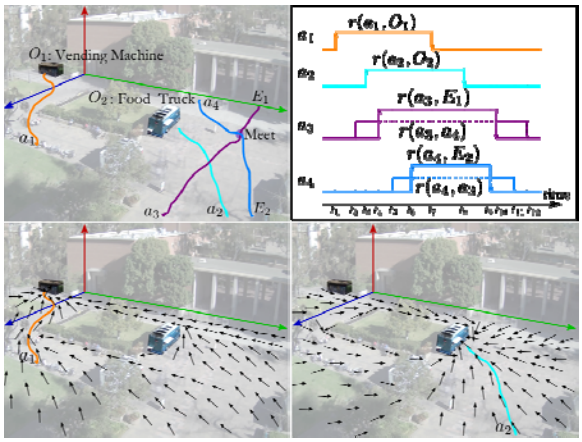
Scene understanding by

Inferring the "dark matters" --- hidden objects

"dark energy" --- hidden relations.

"dark" means not directly detectable by appearance in a bounding box.

By analogy of cosmology:
5% observables
23% dark matters and
72% dark energy



D. Xie, S. Todorovic and S.C. Zhu, "Inferring 'Dark Matter' and 'Dark Energy' from Videos," ICCV 2013.

A scene has multi-layered "fields" generated by "dark matters"

Given a surveillance video, we infer

- functional objects ("dark matter"), S
- Attractive or repulsive fields ("dark energy"), F
- Obstacles in the scene, C
- Intents of people, R
- People's trajectories prediction, Γ

Lagrangian Mechanics

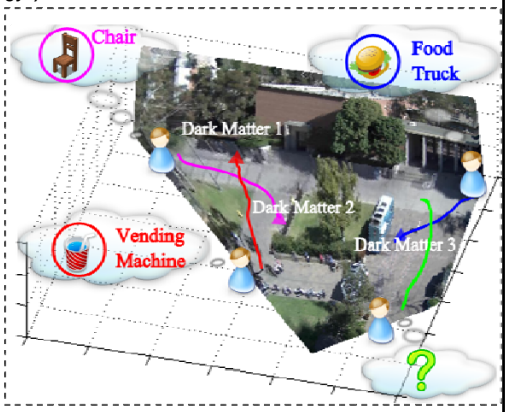
$$L(\mathbf{q}, \dot{\mathbf{q}}, t) = \frac{1}{2} m\dot{\mathbf{q}}^2 - U(\mathbf{q})$$

- $U(\mathbf{q})$: Potential Function involving human property and human need
- $\mathbf{q}$: Generalized Position
- $\dot{\mathbf{q}}$: Generalized Velocity

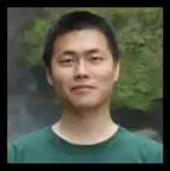
Action S, defined as the time integral of the Lagrangian.

$$S = \int_{t_1}^{t_2} L(\mathbf{q}, \dot{\mathbf{q}}, t) dt$$

Discovering "Dark" Objects by Functionality and Trajectories



Inferring “Dark Matter” and “Dark Energy” from Videos


Dan Xie


Sinisa Todorovic


Song-Chun Zhu

(Accepted in ICCV 2013)

Multi-level intention prediction by event and-or grammar (Pei et al 2011-13)

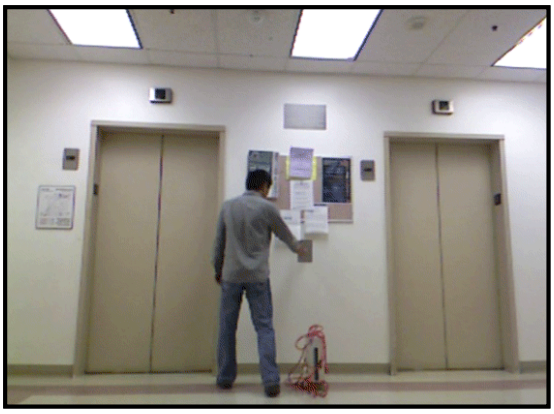
4, Causality: reasoning why, why not, how, what if

Cognitive studies:

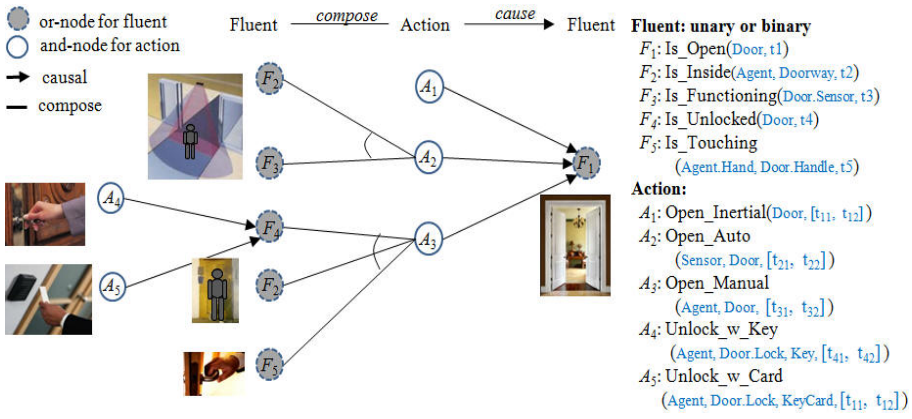
- Infants (12 months) can perceive visual causality;
- Perceived causality can influence perception of velocity (K. Nakayama).

Vision tasks:

- Inferring fluent --- time varying object status, unlike attributes.
- Re-defining action by their effects.



Representing causality by a Causal And-Or Graph



Fluent: unary or binary

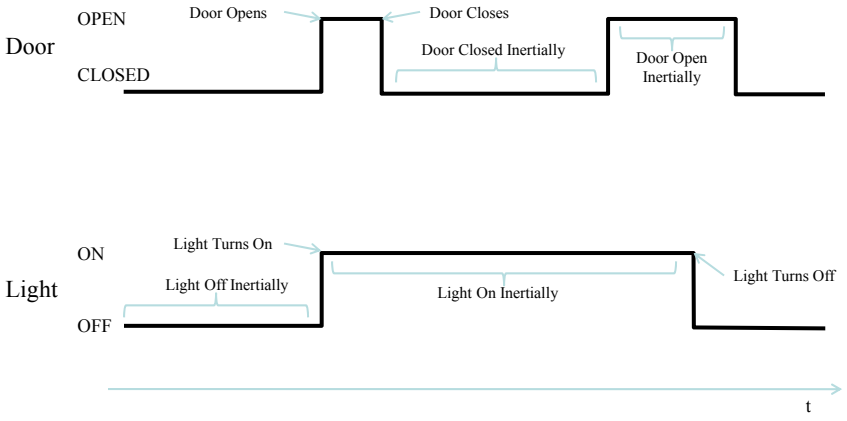
F_1: Is_Open(Door, t1)
F_2: Is_Inside(Agent, Doorway, t2)
F_3: Is_Functioning(Door.Sensor, t3)
F_4: Is_Unlocked(Door, t4)
F_5: Is_Touching
(Agent.Hand, Door.Handle, t5)

Action:

A_1: Open_Inertial(Door, $[t_{11}, t_{12}]$)
A_2: Open_Auto
(Sensor, Door, $[t_{21}, t_{22}]$)
A_3: Open_Manual
(Agent, Door, $[t_{31}, t_{32}]$)
A_4: Unlock_w_Key
(Agent, Door.Lock, Key, $[t_{41}, t_{42}]$)
A_5: Unlock_w_Card
(Agent, Door.Lock, KeyCard, $[t_{11}, t_{12}]$)

Amy Fire and S.C. Zhu, "Using Causal Induction in Humans to Learn and Infer Causality from Video," 35th Annual Cognitive Science Conference (CogSci), 2013.

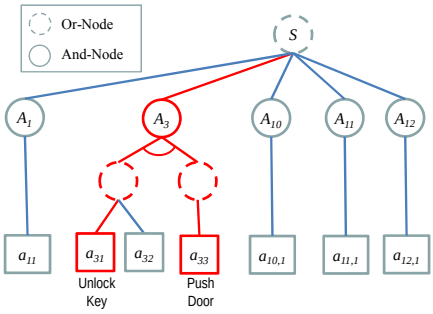
Common fluents in scenes



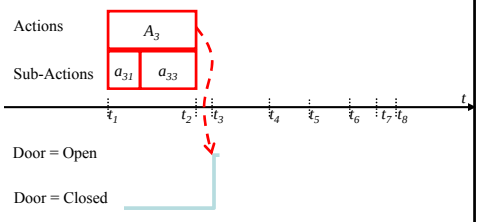
39

Causal relations: linking actions to fluent changes

ST-AOG

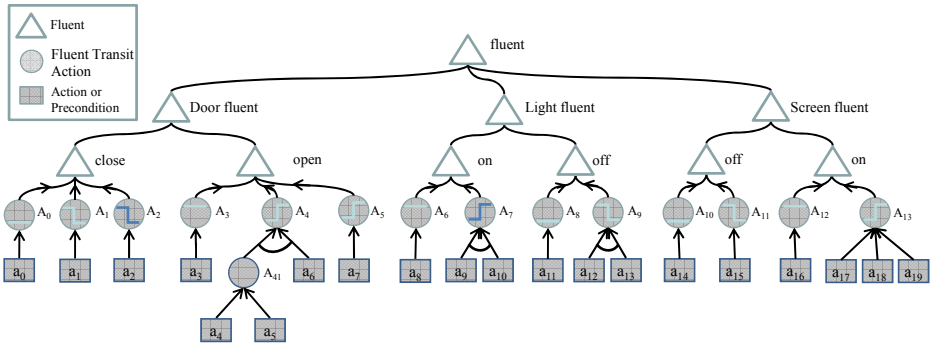


PG



40

Unsupervised Learning of Causal-AOG



A_0: inertial action
a_0: precondition (door closed)

A_1: close door
a_1: pull/push

A_2: door closes inertially
a_2: leave door

A_3: inertial action
a_3: precondition (door open)

A_4: open door
A_{41}: unlock door
a_4: unlock by key
a_5: unlock by passcode
a_6: pull/push

A_5: open door from inside
a_7: person exits room

A_6: inertial action
a_8: precondition (light on)

A_7: turn on light
a_9: touch switch
a_{10}: precondition (light off)

A_8: inertial action
a_{11}: precondition (light off)

A_9: turn off light
a_{12}: touch switch
a_{13}: precondition (light on)

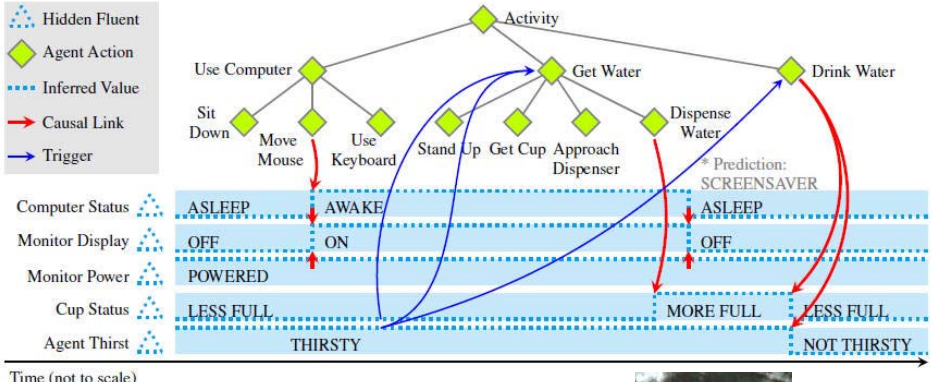
A_{10}: inertial action
a_{14}: precondition (screen off)

A_{11}: turn off screen
a_{15}: push power button

A_{12}: inertial action
a_{16}: precondition (screen on)

A_{13}: turn on screen
a_{17}: touch mouse
a_{18}: touch keyboard
a_{19}: push power button

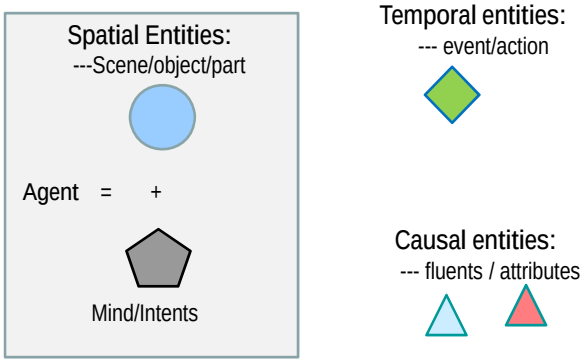
Reasoning hidden fluents in scene by causality



Amy Fire

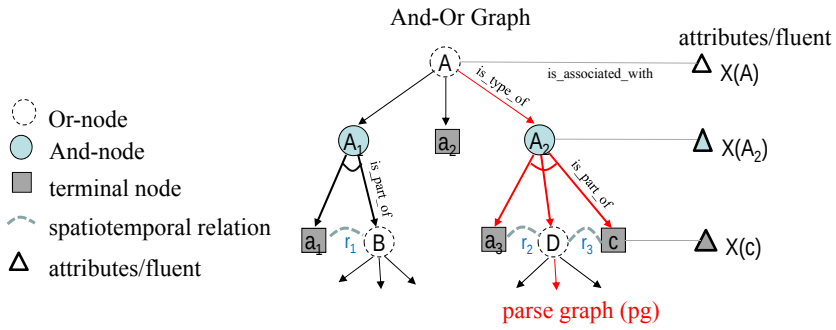
3. Unified Representation:
Spatial, Temporal, Causal And-Or Graph

Knowledge Representation



And-Or Graph

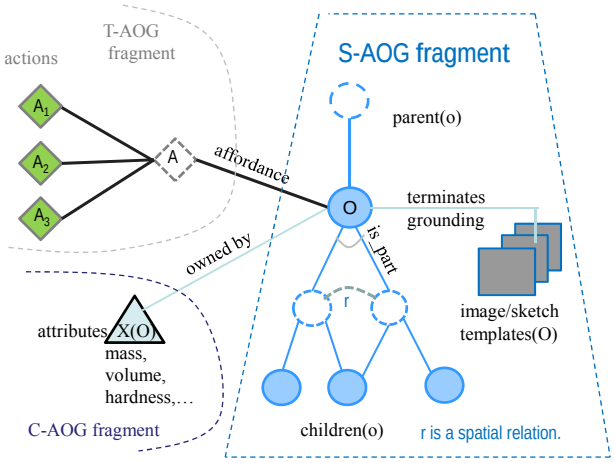
Production rule A ::= aB | a | aDc
Logic formula A = (a ∧ B) ∨ a ∨ (a ∧ D ∧ c)



AOG embodies a stochastic, attributed, context-sensitive grammar.

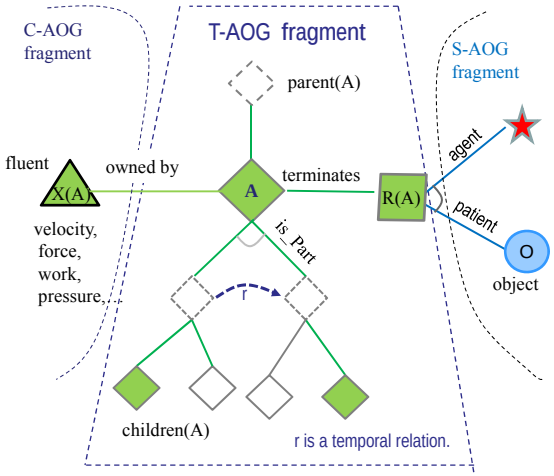
Zhu and Mumford, "A Stochastic Grammar of Images", 2006.

Interweaving concepts in the STC And-Or Graph

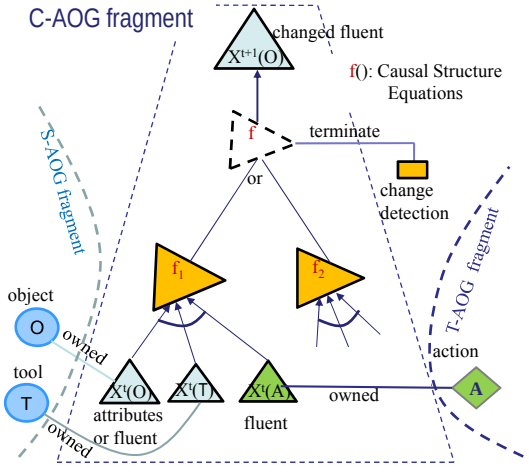


Unsupervised/weakly supervised/supervised learning of the AOG
(Z. Si,et al ICCV 2009, PAMI 2011, PAMI 2013, Y. N. Wu, 2007-14, K.Tu NIPS 2013)

Interweaving concepts in the STC And-Or Graph

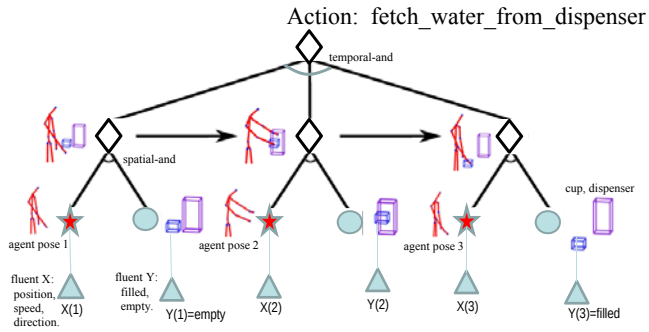


Interweaving concepts in the STC And-Or Graph



A visual concept is a sub-graph in the STC-AOG

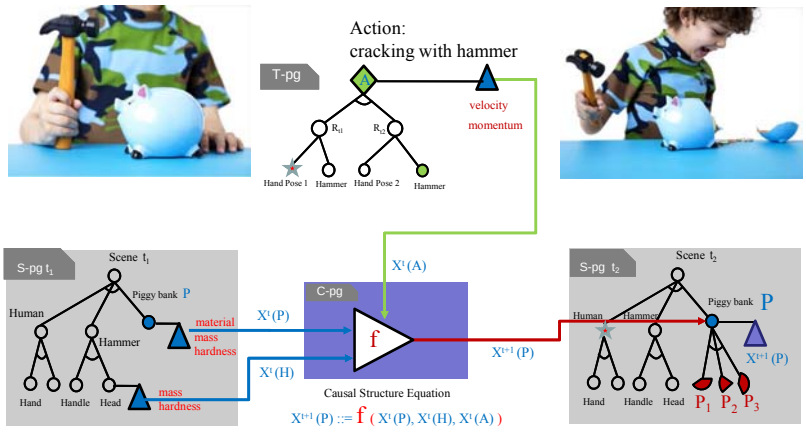
A concept, like this action, is a graph spanned
In the spatial, temporal, and causal joint space.



P. Wei et al. ICCV 2013.

4. Joint Inference:
Spatial, Temporal, Causal Parsing

Scene understanding needs joint STC-parsing



Group of Structure Equations for causality

Understanding Scene by Joint Spatial, Temporal, Causal and Text Parsing

Joint Spatial, Temporal, Causal and Text Parsing

UCLA Center for Vision, Cognition, Learning and Art

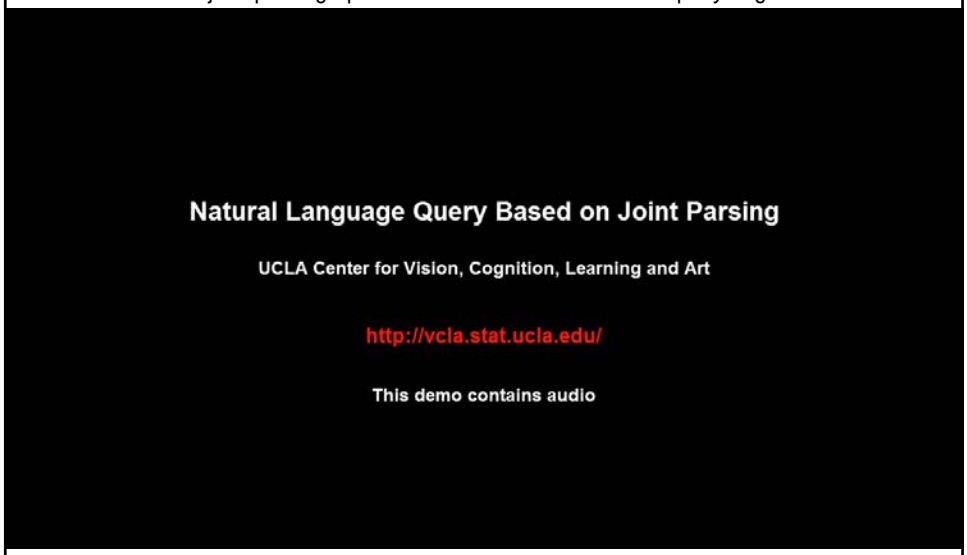
University of California, Los Angeles

Aprial.2014

This demo contains audio

Demo: Answering User Queries on What, Who, Where, When and Why

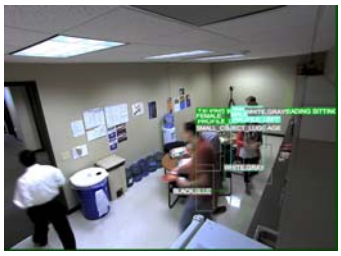
We transfer the joint parse graph in RDF format and feed into a query engine.



A Restricted Turing Test on Understanding Object, Scene & Event

3 Areas, 30+ cameras (ground, tower, mobile), 3,000,000 frames (1 TB).

Ontology: objects, attributes, scenes, actions, group activities, spatial-temporal relations.



Humans/vehicles are tracking across cameras.
Trajectories are mapped to google map for outdoor and floor maps for indoor.

Location: Conference Room
Time: 15:47:00 - 16:19:00 [32 minute duration]

Q: Is there at least one chair in the conference room that no one ever sits in?

Q: Is there a person putting food into the mouth?

Q: Is the upper and lower leg of a person in a white shirt occluded from the view of camera by a table?
…

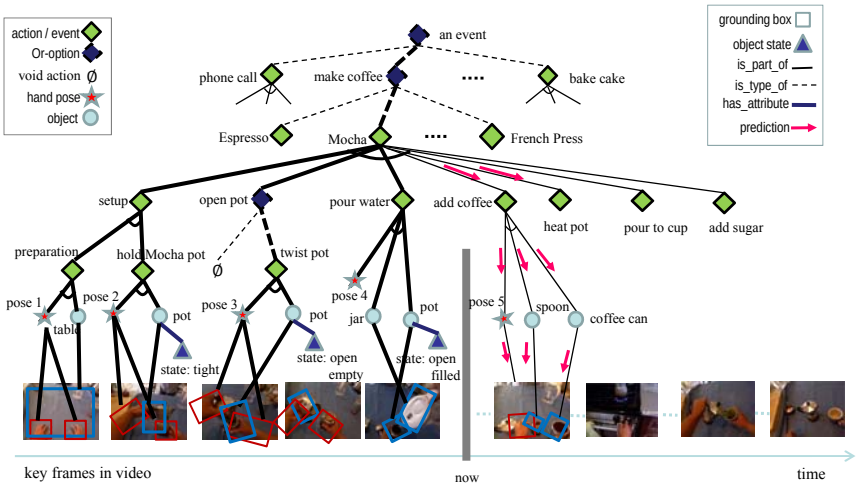
Q: Are there more than 2 people in the geo-coordinate bounding box? (TRUE)

Q: Is there a person wearing a yellow shirt in a car within the AOR after the abduction event? (TRUE)

5. Challenges and future work

Human-computer dialogue and lifelong learning
will be essential for knowledge acquisition:
--- Shared Knowledge, Shared Situation Shared Intention, Shared Attention

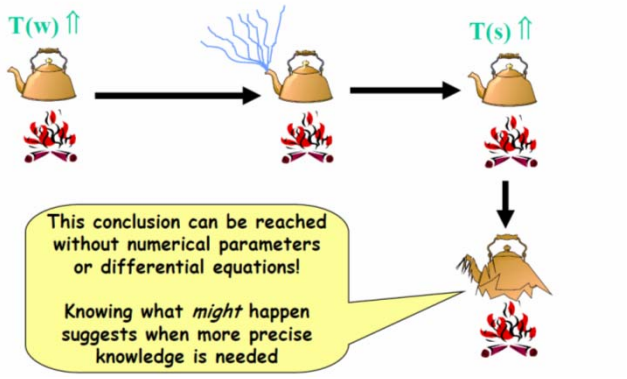
A STC-pg for making coffee with a Mocha pot.



How do we represent stuff like water, phase transition

Qualitative reasoning

- What can happen when you leave a tea kettle on a stove unattended for an hour?



Slide from: Ken Forbus, Northwestern University

Our famous crow again: Cracking nuts by vehicle at crosswalk

In this process, the crow must have a deep profound understanding the scene, dynamics of human/vehicle, causality, timing of actions, physical properties of objects,…

Taken Home message: the Crow Inspiration

The crow videos prove that there exists a solution for deep scene understanding:

--- small volume
    embedded in your smart phones, wearable devices;
--- low-power
    < .1 Watt (human brain is about 10 watt, crow brain is 100 times smaller)