Understanding Video and Text by Joint Spatial, Temporal, and Causal Inference

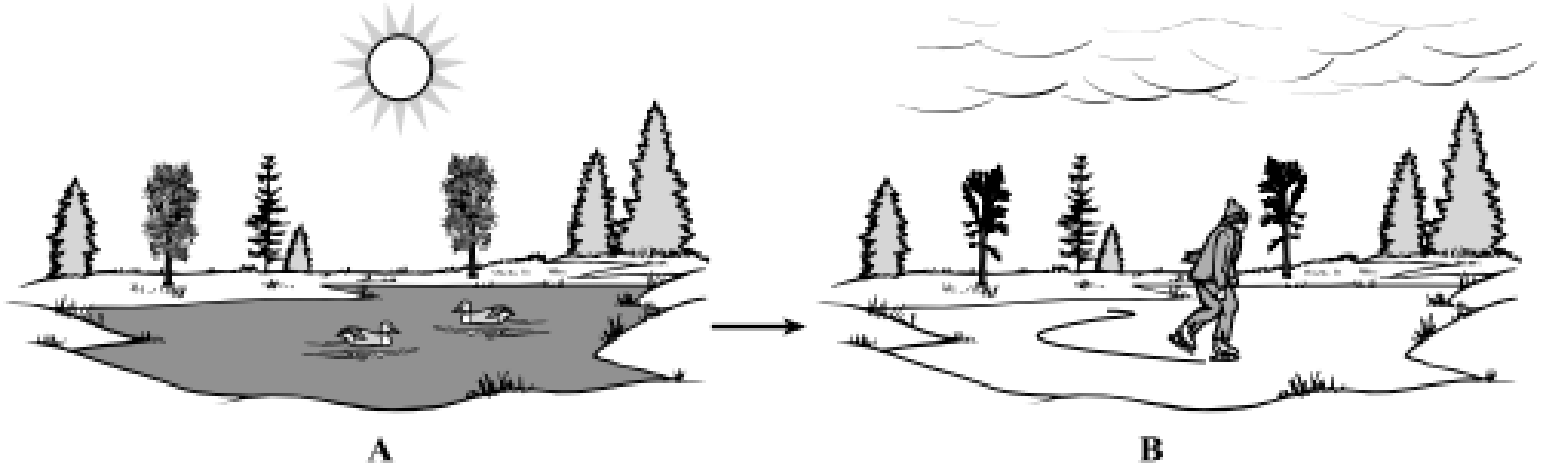
Song-Chun Zhu

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

Stanford Workshop on AI and Knowledge, April 16, 2014

Question in a 5th Grade Test

Need joint reasoning using Vision + Language + Cognition (physics, causality)



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.

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

Click this youtube video to watch the demo
https://www.youtube.com/watch?v=FmFK52WwSQg#t=89

Joint video-text parsing

(a) Input Video

Input Text

"A silver SUV stopped and took pictures of the federal building at 09:21."

Ref: K. Tu, et al. Joint Video and Text Parsing for Understanding Events and Answering Quires, IEEE Multimedia, 2014. [pdf]

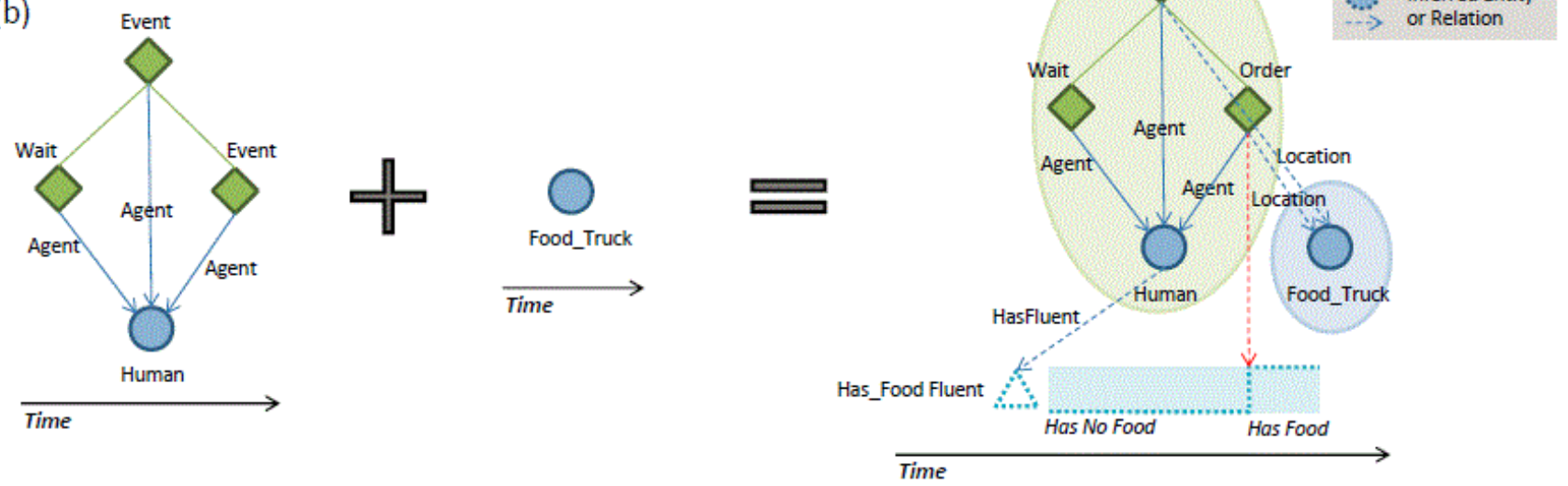
Joint video-text parsing

(a) Input Video

Input Text

"There is a food truck in the courtyard."

(b)



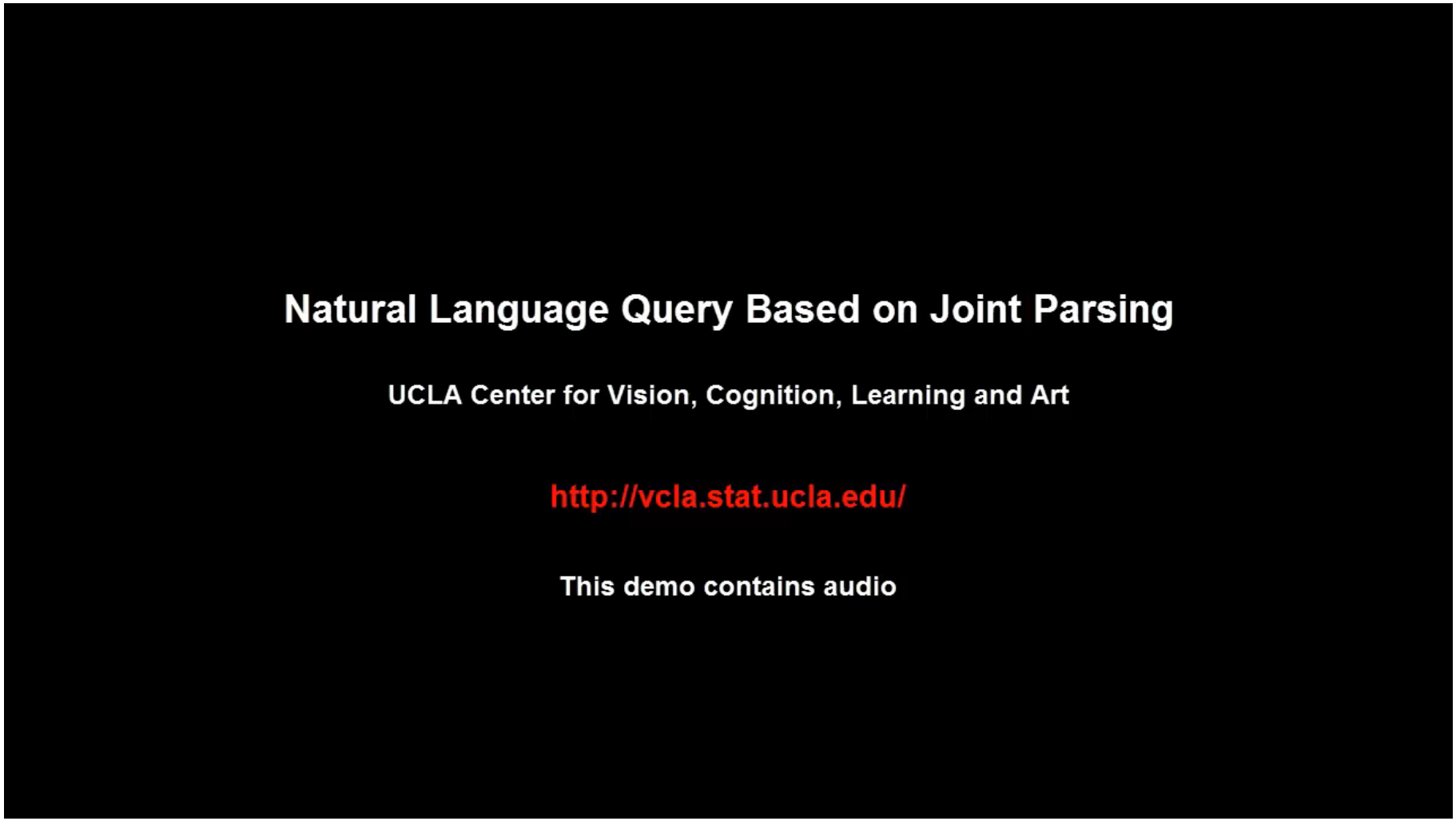
Video Parse

Text Parse

Joint Parse

2, 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.



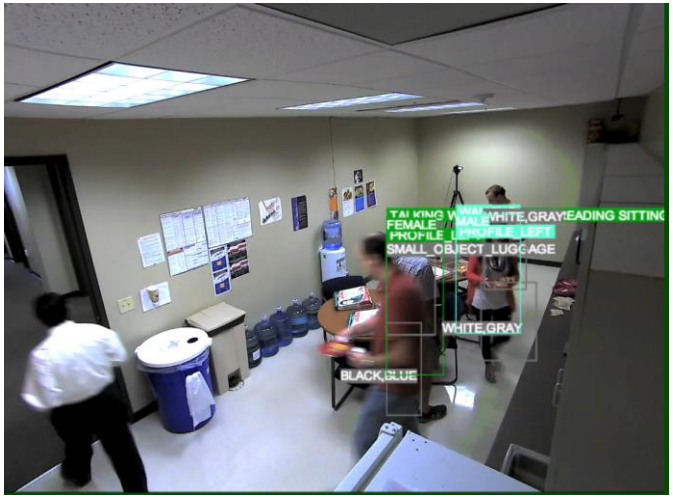
Click this youtube video to watch the demo

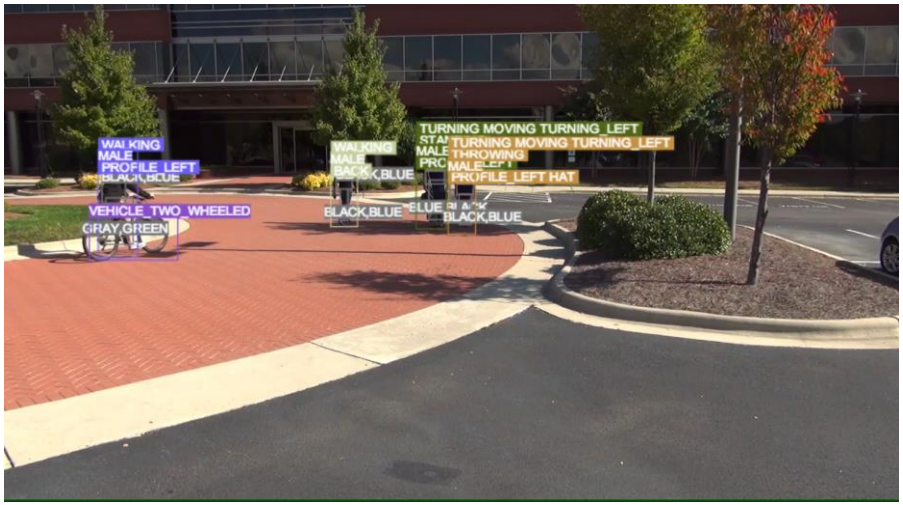
https://www.youtube.com/watch?v=FnbYODNEgM8

3, 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.

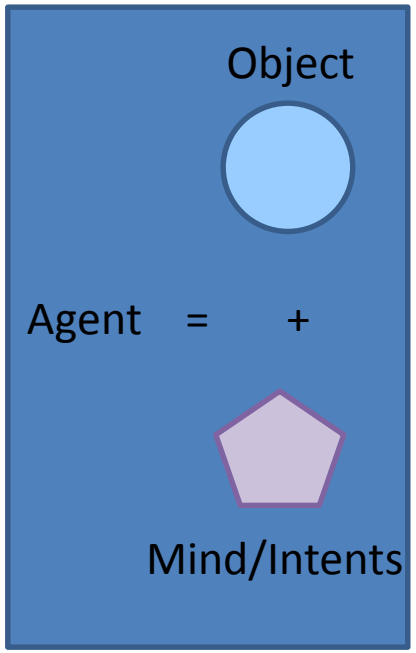




For example:
Location: Conference Room
Time: 15:47:00 - 16:19:00 [32 minute duration] [Video and Question are prepared by SIG]

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?
…

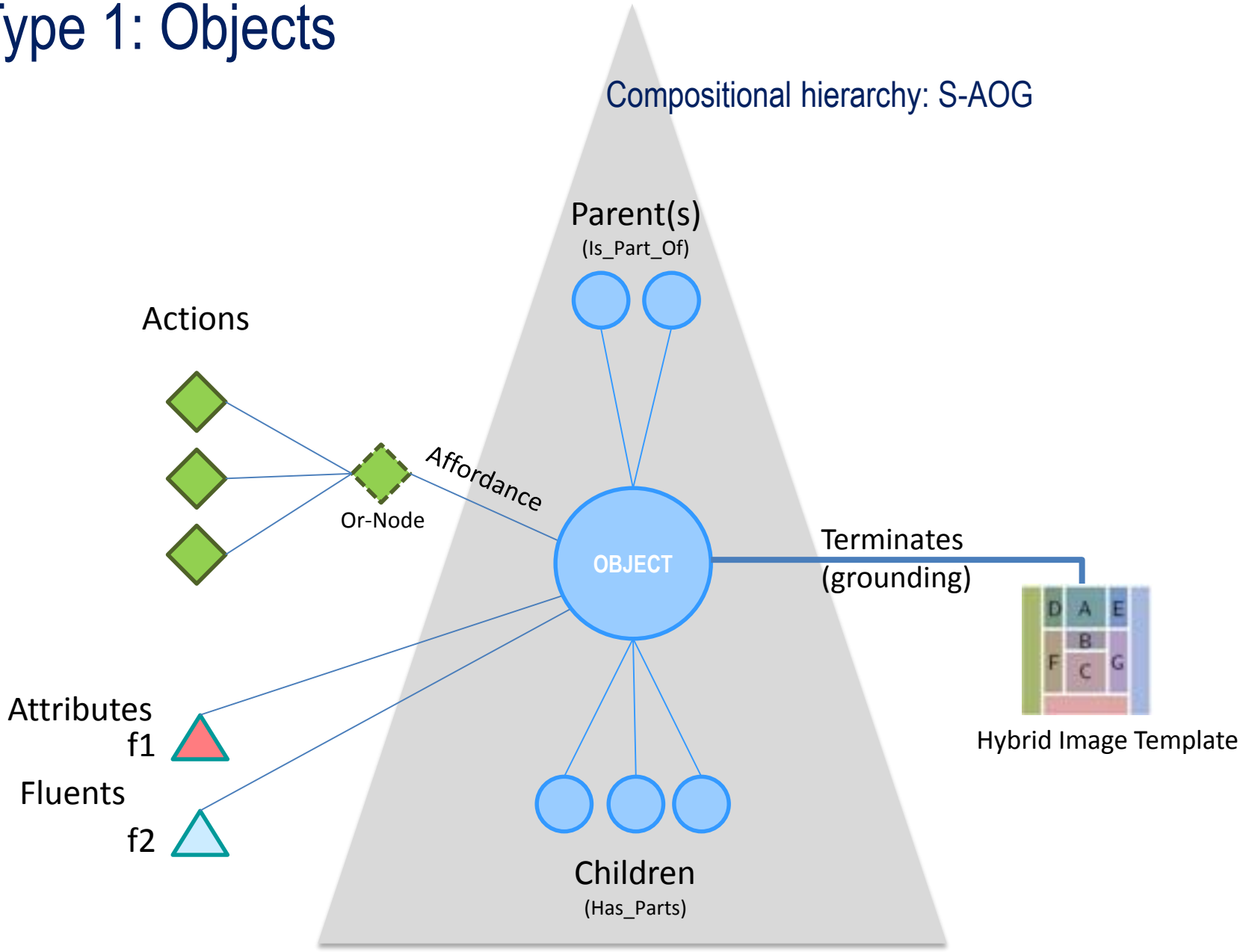
4, Knowledge representation:
the Spatial, Temporal, Causal And-Or Graph (STC-AOG)



Action

Attribute and Fluent

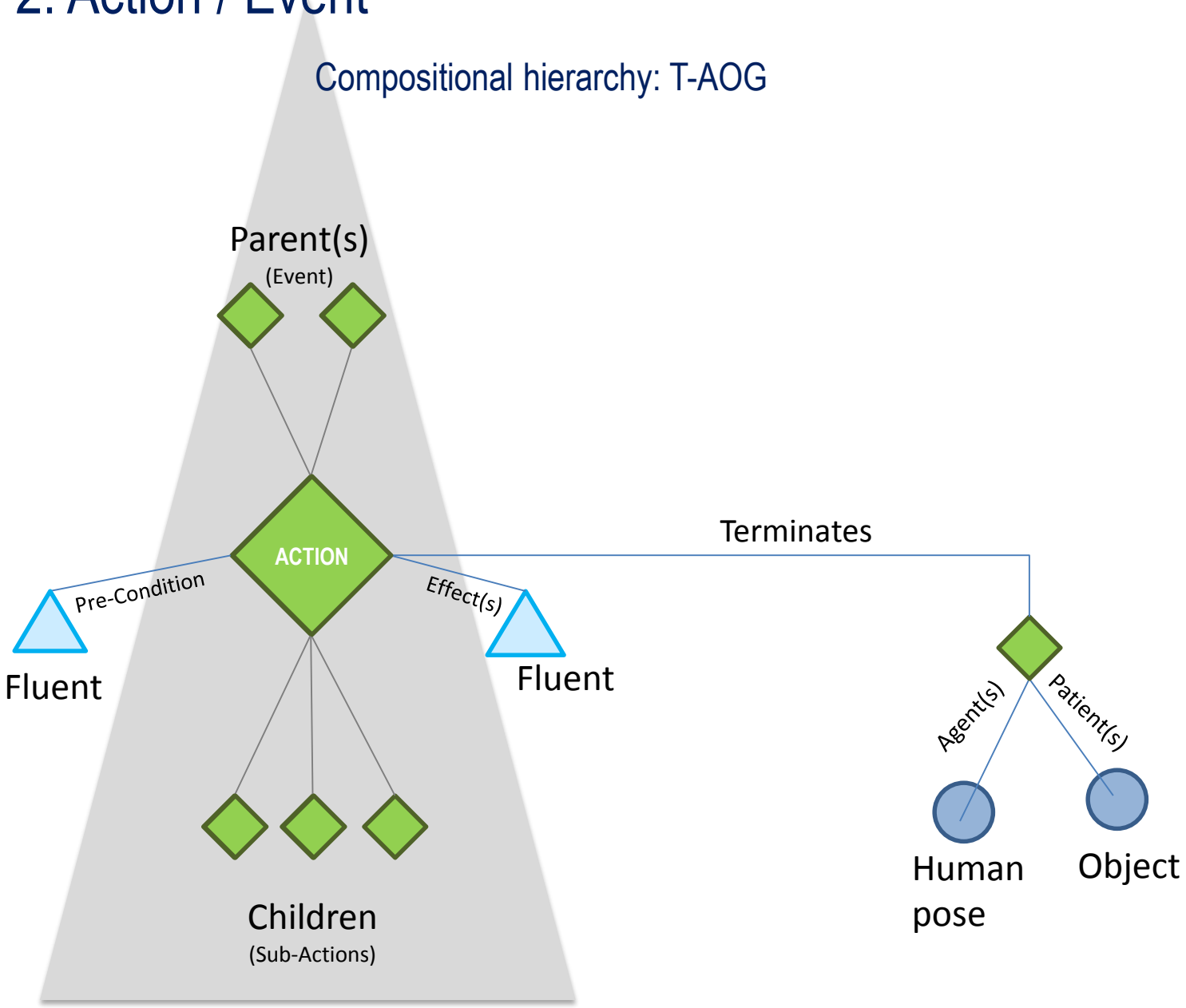
Type 1: Objects



S.C. Zhu and D. Mumford, A stochastic grammar of images, 2006 [pdf]

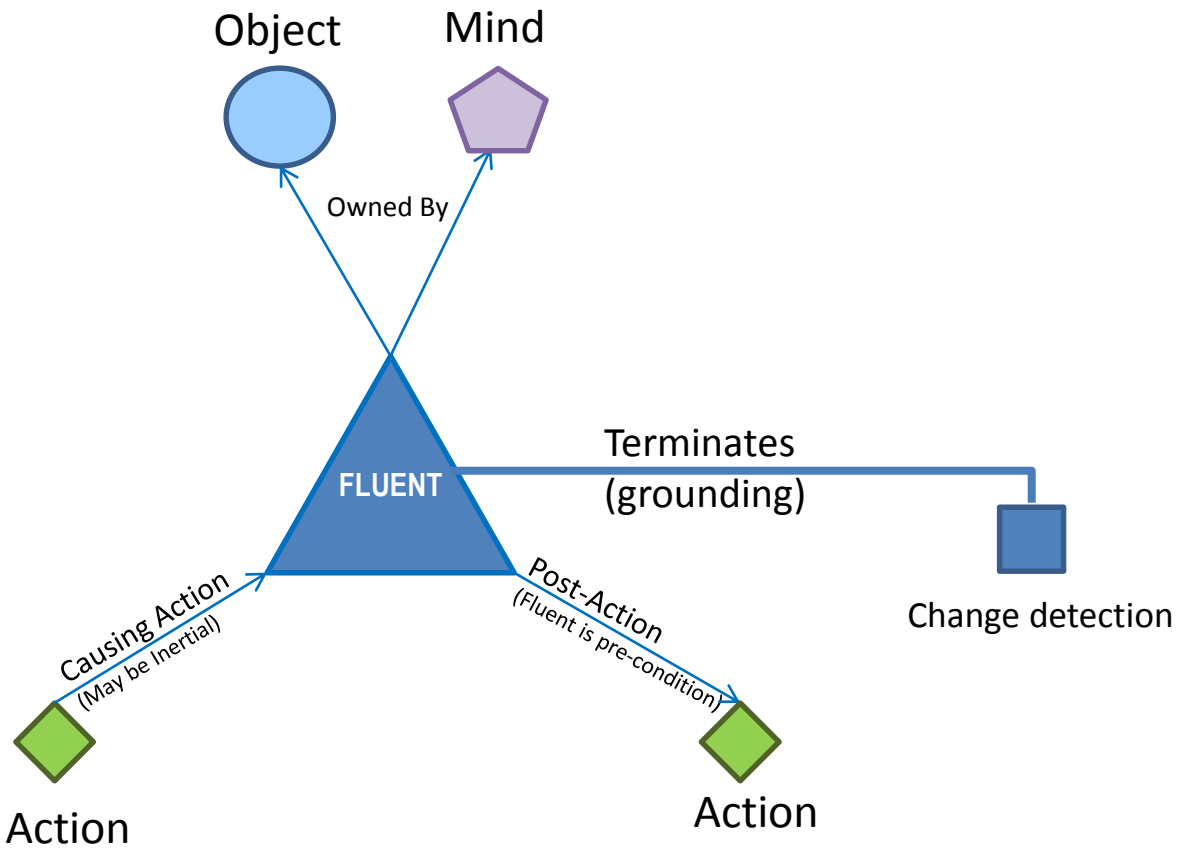
Z.Z. Si and S.C. Zhu, Learning And-or templates for object modeling and recognition, PAMI 2013, [pdf]

Type 2: Action / Event



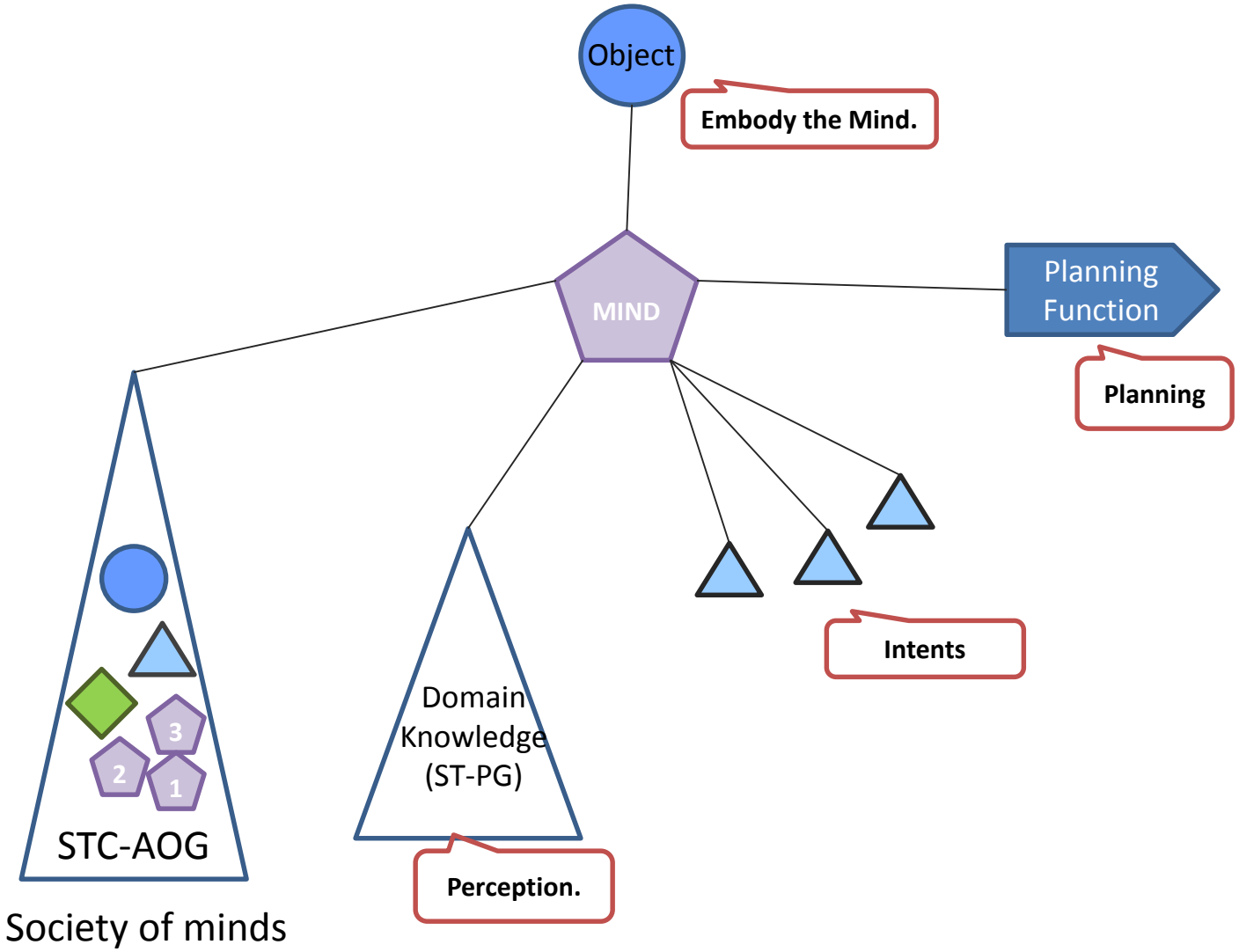
Ref: Pei et al, Video Event Parsing and Learning with Goal and Intent Prediction, *Computer Vision and Image Understanding*, vol. 117, no. 10, pp 1369-1383, 2013. [pdf]

Type 3: Fluent



A. Fire and S.C. Zhu, Using Causal Induction in Humans to Learn and Infer Causality from Video, CogSci, 2013 [pdf]

Type 4: Intents/Minds



4. Augmenting Visual Knowledge with Commonsense

To achieve deeper understanding of objects, scenes, and events, one need to consider many other aspects:

1, Function and affordance:

Scenes are often defined by activities in space;
Objects are often defined by how they can be used.

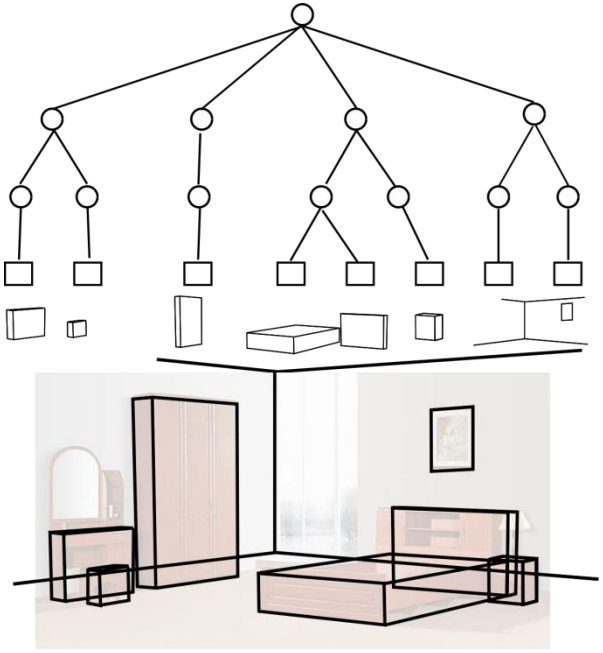
2, Physics: such as material, center of mass, velocity, force, torque, work, temperature, state (solid, liquid).

3, Intents: the intention and goals of agents in the scenes and events.

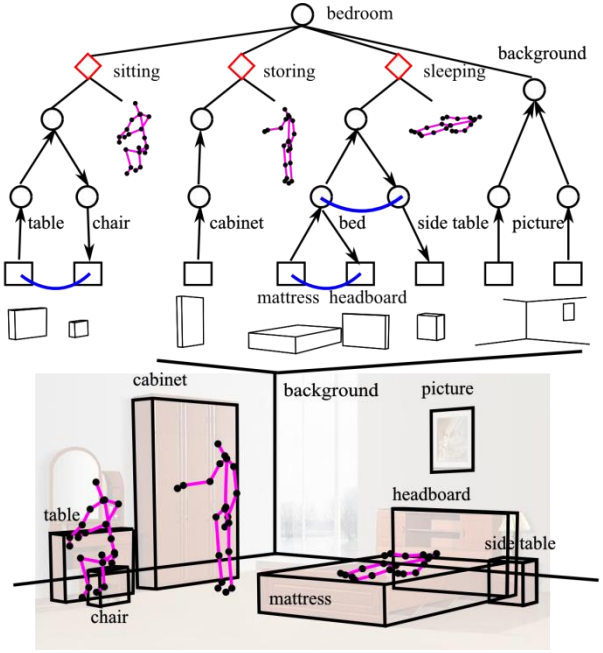
4, Causality: causal-effects, laws, and equations,

Example 1: Augmenting Traditional Image Parsing with Functionality

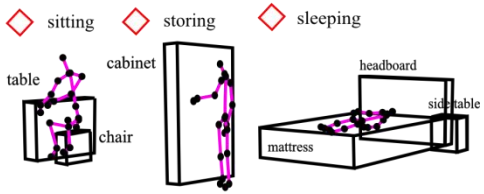
Traditional parse tree



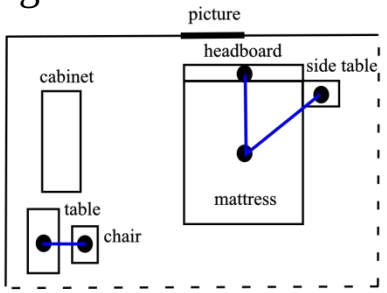
Augmented parse graph



Augmented object affordance



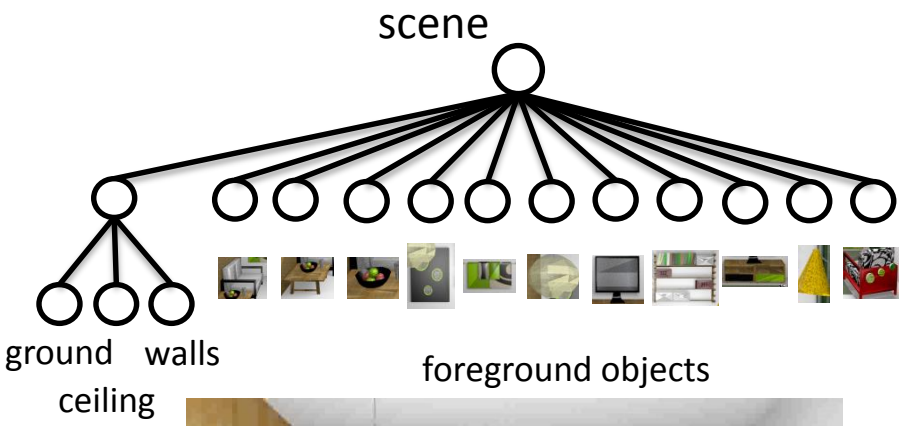
Augmented contextual relations



Y.B. Zhao and S.C. Zhu, NIPS 2011, CVPR2013. [pdf]

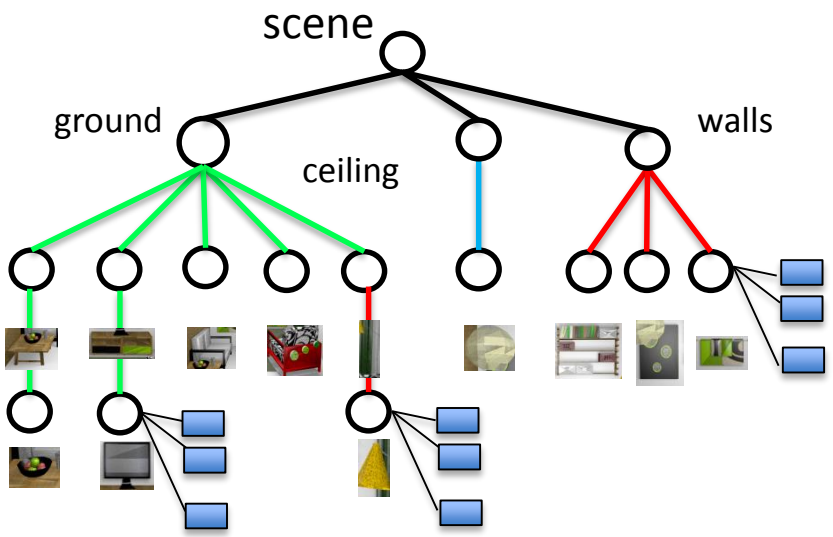
Example 2: Augmenting Traditional Image Parsing with Physics

Traditional parse tree



foreground objects

Augmented parse graph



Augmented physical properties:

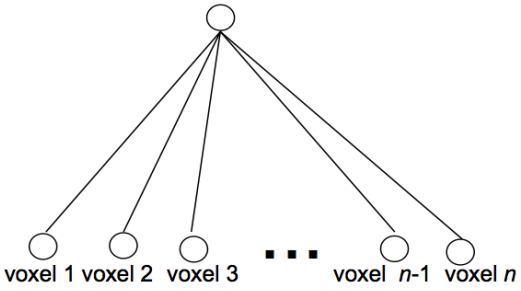
- material, friction, mass, velocity

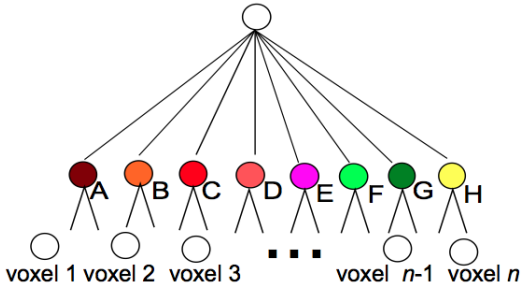
Augmented physical relations:

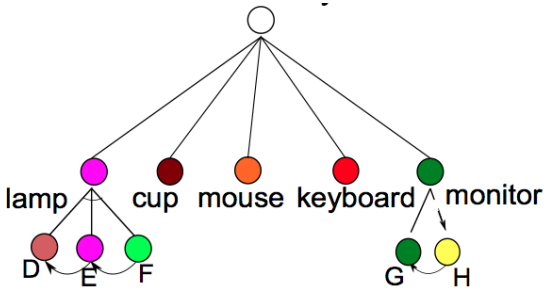
- supporting, attaching, hanging

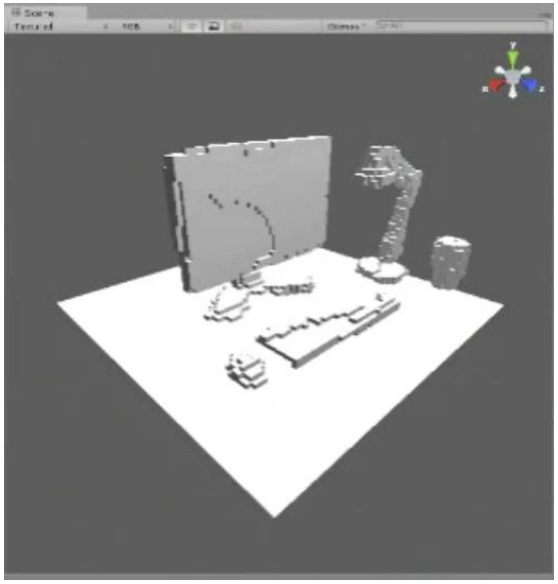
B. Zheng et al cvpr 2013. ICRA2014. [pdf]

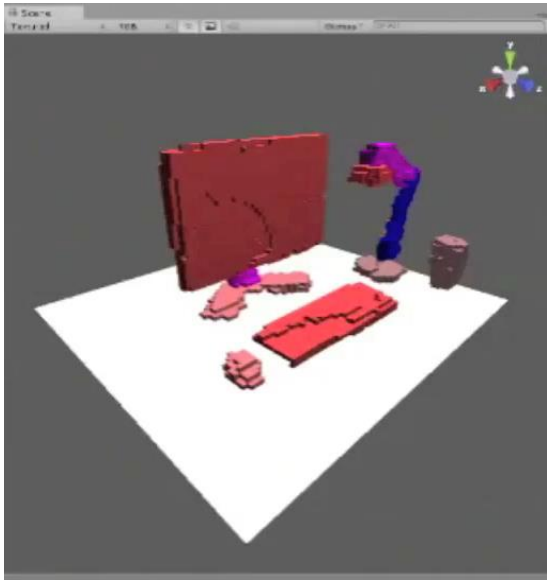
Below is an example that uses physical constraints to help scene parsing,
i.e. a valid parse (interpretation) must be physically plausible.

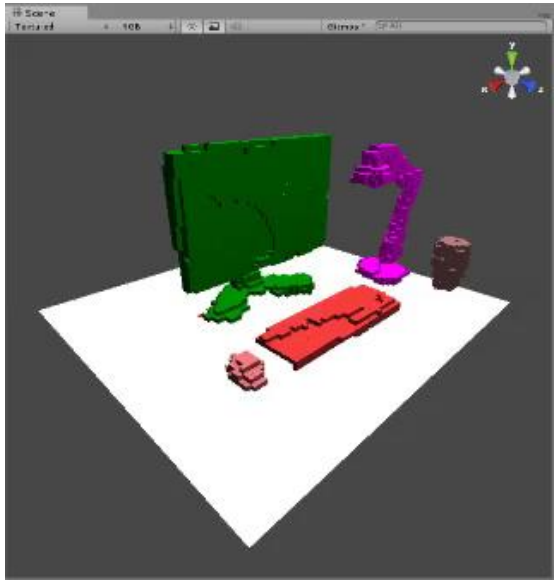












By grouping the voxels (captured by depth sensor) into geometric solids (parts) and then into Object (segmentation), so as to minimize physical instability, and maximize functionality to serve humans.

B. Zheng et al, CVPR2013.

What is commonsense ?

In nature, low rank animals, like crow, have astonishing commonsense knowledge which goes way beyond current computer intelligence.

Video example I: making tool to reach food.

In this process, the crow must understand the scene, know the material property of the metal stick, make the hook with torque, use the hook, lots of physics, ….

Click this youtube video https://www.youtube.com/watch?v=dbwRHIuXqMU

Video example II: Cracking nuts using vehicles at crosswalk

In this process, the crow demonstrates **profound scene understanding capabilities**: dynamics of human/vehicle, causality, physical properties of objects,…

https://www.youtube.com/watch?v=BGPGknpq3e0

The crow videos prove that there exists a solution

--- small volume

embedded in your smart phones, wearable devices;

--- low-power

< 0.1 Watt (human brain is upper bounded by ~10 Watt, crow's brain is ~ 100 times smaller.)

An unifying math foundation for visual knowledge