Learning to Sense Robustly and Act Effectively

Benjamin Kuipers, P.I. and Silvio Savarese, co-P.I.
EECS, University of Michigan, Ann Arbor, Michigan

Introduction

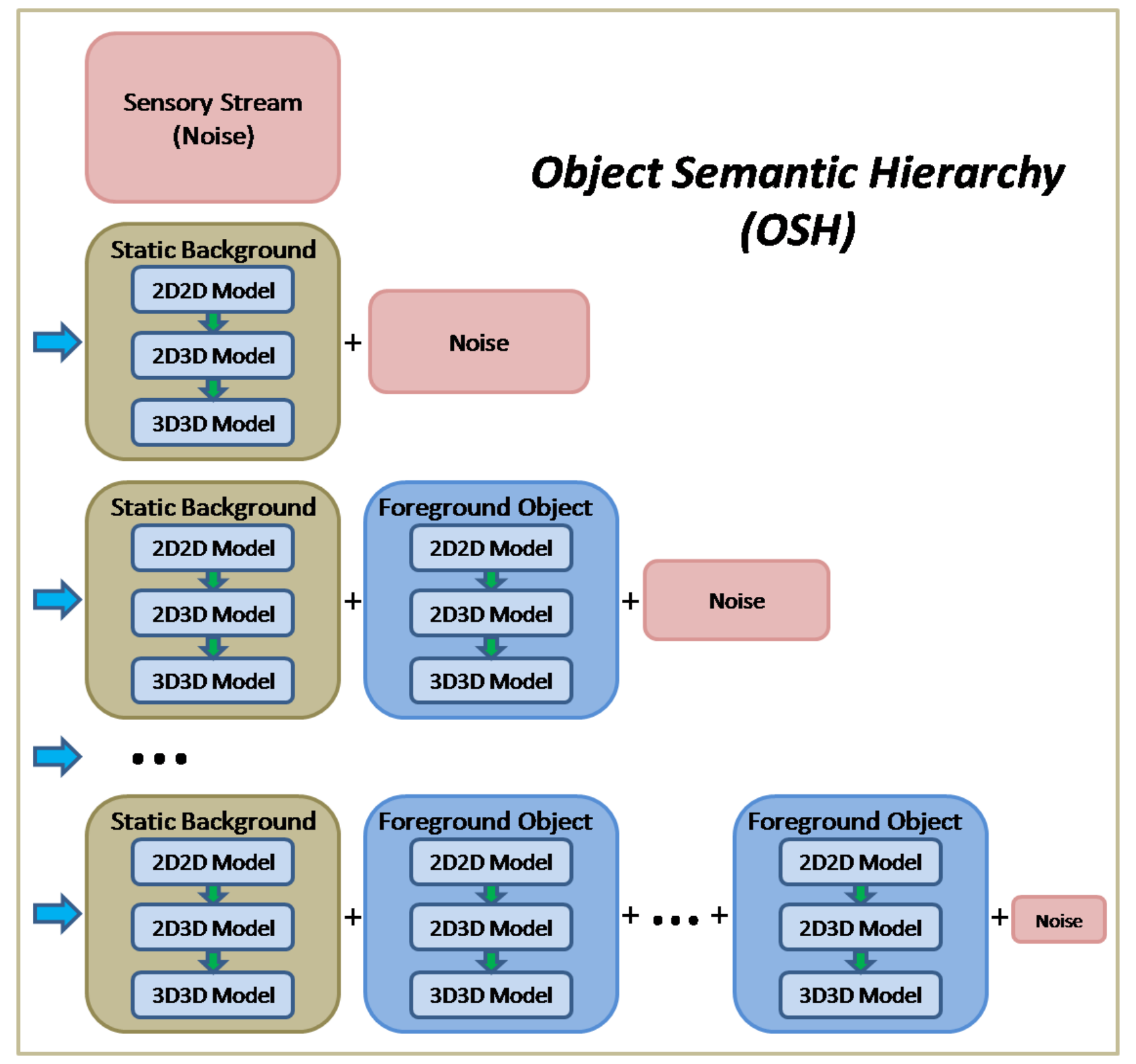
The physical environment of a cyber-physical system is unboundedly complex, changing continuously in time and space. An embodied cyber-physical system, embedded in the physical world, receives a high bandwidth stream of sensory information, and sends continuous control signals. Traditional embedded systems restrict the environment or the attributes considered relevant, and depend on human supervision.

To handle the complexity of unrestricted environments, future cyber-physical systems will need to be learning agents, learning the properties of sensors, effectors, and environment from their own experience, and adapting over time. Foundational concepts such as **Space**, **Object**, **Action**, etc., will be essential for abstracting and controlling the complexity of its world.

Our previous work on the **Spatial Semantic Hierarchy (SSH)** [Kuipers, AIJ, 2000; Beeson, et al, IJRR, 2010] shows how multiple representations of space can bridge the gap between continuous interaction with the physical environment, and discrete symbolic descriptions that support effective planning.

We are developing robot agents that use vision and manipulation to learn models of objects and actions at multiple levels of representation:
(1) learning to perceive objects in its environment;
(2) joint optimization of semantic constraints in vision;
(3) learning a hierarchy of increasingly skilled actions.

The **Object Semantic Hierarchy (OSH)** [Xu & Kuipers, ICDL, 2010] shows how a learning agent can create a hierarchy of representations for visual perception of objects it interacts with. The OSH "object abstraction" factors uncertainty in the sensor stream into object models and object trajectories.



Object Layers

Layer 0: $z_t = \varepsilon_0$

Layer 1: $z_t = G_1(M^b, x_t) + \varepsilon_1$

Layer 2: $z_t = G_1(M^b, x_t) + G_2(M_1^o, y_{1t}) + \varepsilon_2$

…

Layer *n*: $z_t = G_1(M^b, x_t) + \sum_{i=1}^{n} G_{i+1}(M_i^o, y_{it}) + \varepsilon_n$

Model Layers

Layer 2D2D: $z_t = G_v(v, y_t^v) + \varepsilon_v$

Layer 2D3D: $z_t = G_c(c, y_t^c) + \varepsilon_c$

Layer 3D3D: $z_t = G_o(o, y_t^o) + \varepsilon_o$

The static background is just another object!

z -- sensory stream
ε -- noise
G -- function that maps background and/or object models to a 2D image given background and/or object poses
M^b -- static background model
x -- robot pose
M^o -- object model
y -- object pose
v -- 2D view model
c -- 2D component model
o -- 3D object model

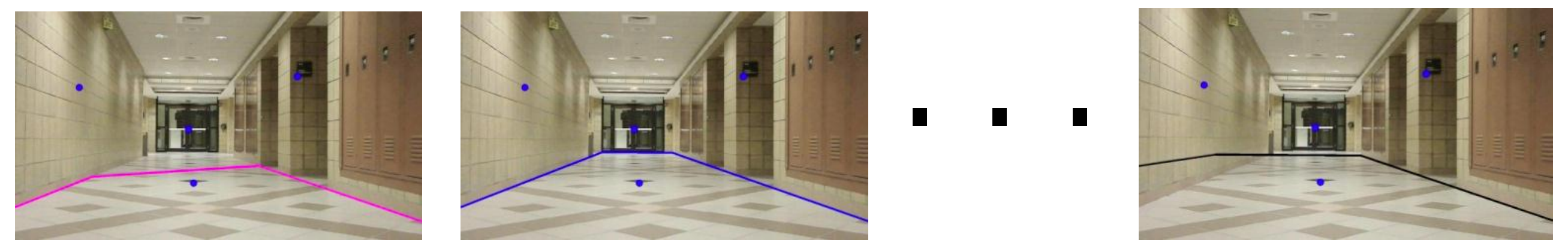
The uncertainty in the agent's sensory stream is factored into a collection of relatively compact representations:
- static background model
- pose trajectory of the agent
- constant foreground object models
- pose trajectories of foreground objects
- any remaining noise

Indoor Scene Understanding

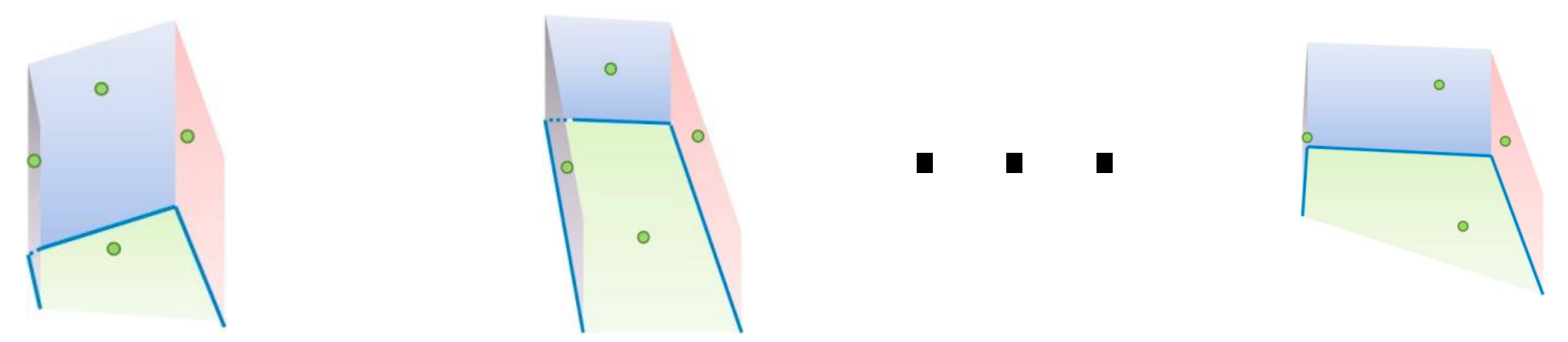
Building on the OSH, and treating the surrounding environment as an "object", Tsai, Xu, Liu & Kuipers [ICCV, 2011] present a new method whereby an embodied agent using visual perception can efficiently create a model of a local indoor environment from its experience moving within it.

Our method uses a single-image analysis, not to attempt to identify a single accurate model, but to propose a set of plausible hypotheses about the structure of the environment from an initial frame. We then use data from subsequent frames to update a Bayesian posterior probability distribution over the set of hypotheses. The likelihood function is efficiently computable by comparing the predicted location of point features on the environment model to their actual tracked locations in the image stream.

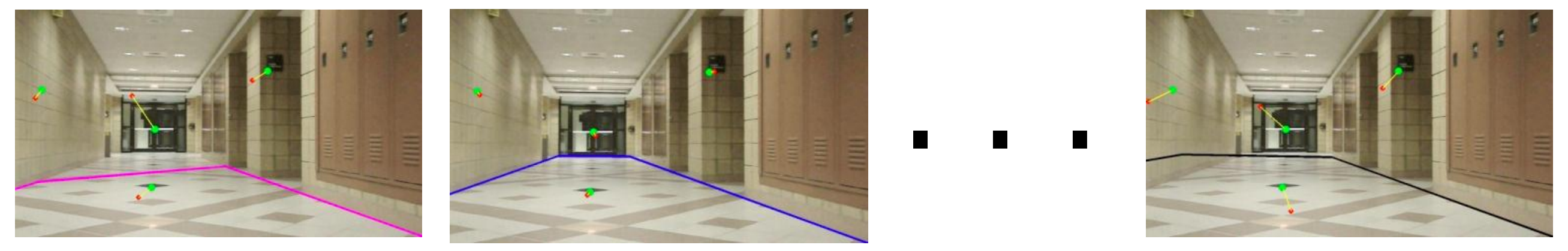
Generate hypotheses



Reconstruct 3D planar model



Estimate camera pose

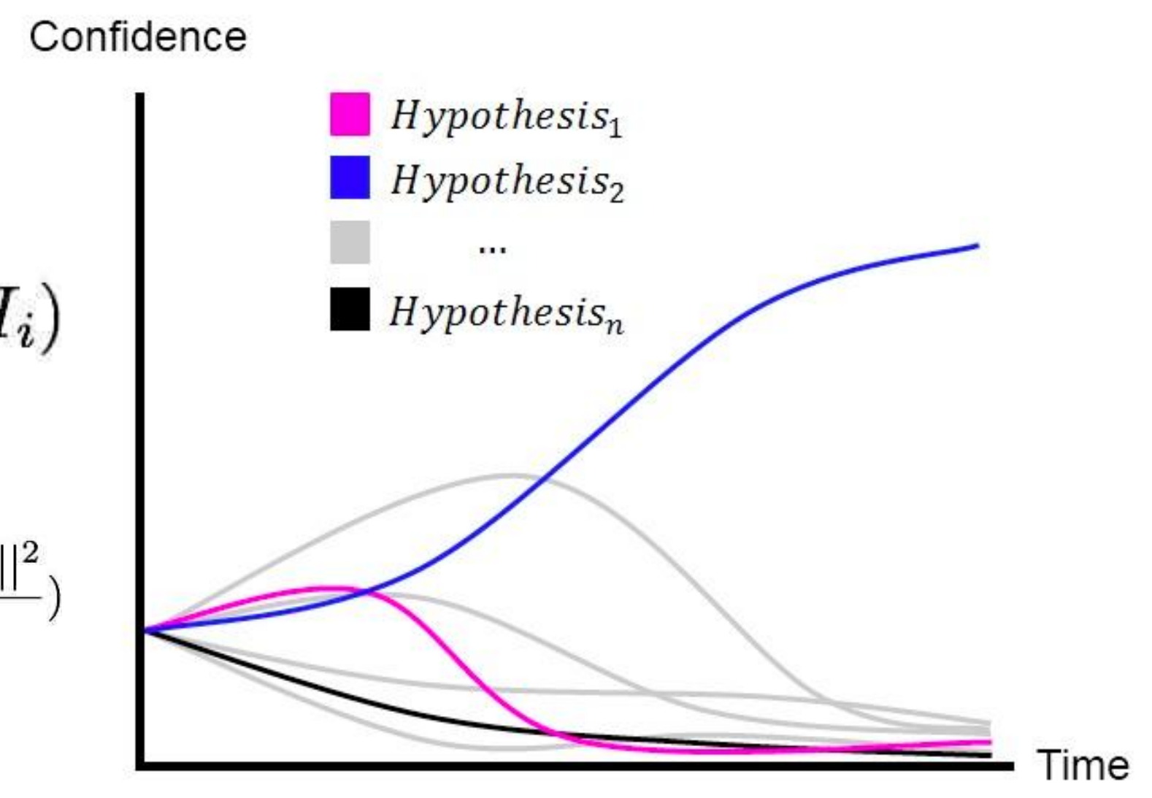


Bayesian filtering

$$p(H_i|\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_m)$$

$$\propto p(H_i) \prod_{j=1\ldots m} p(\mathbf{O}_j|H_i)$$

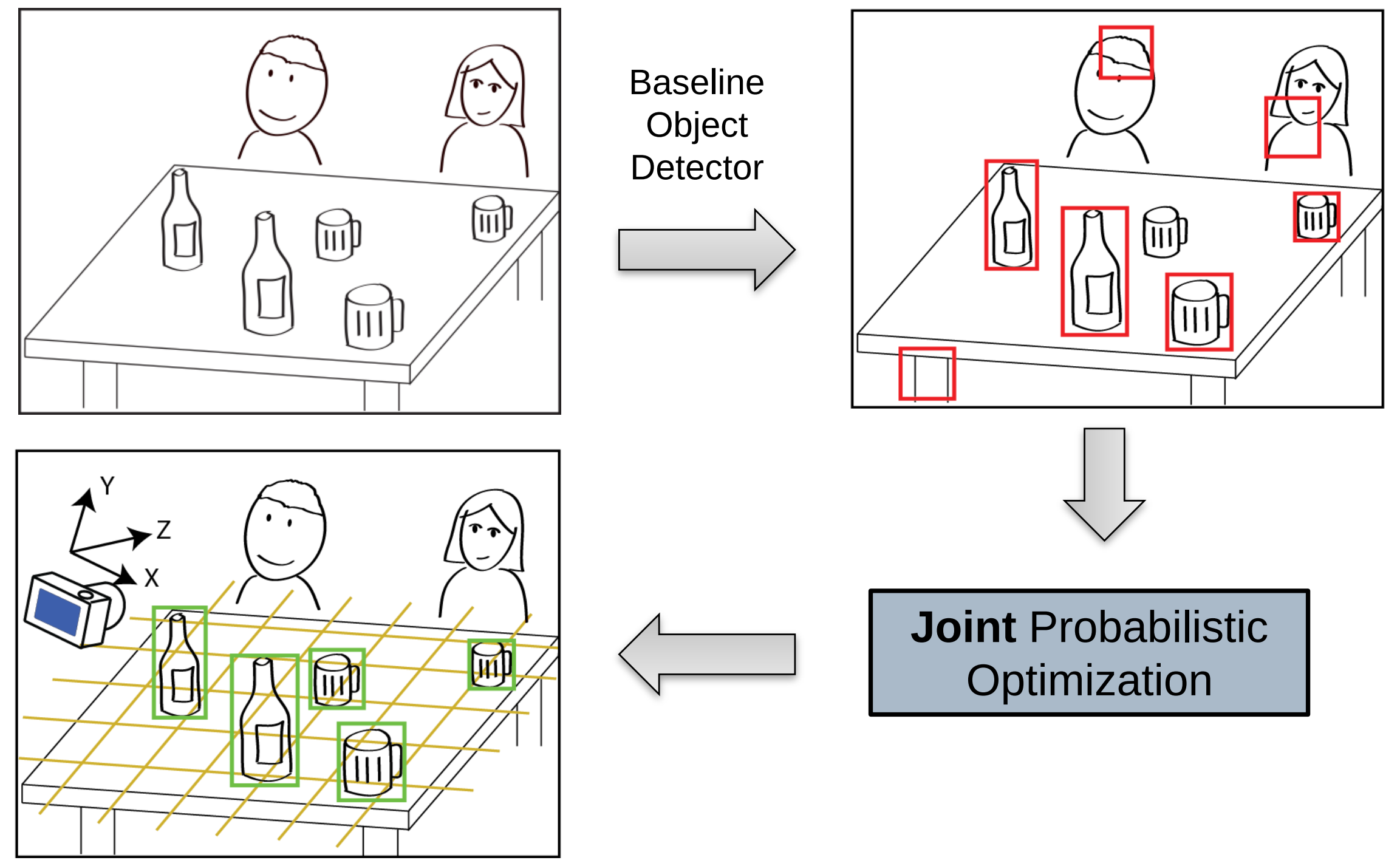
$$p(\mathbf{O}_j|H_i) \propto \prod_{o_k \in \mathbf{O}_j} \exp\left(\frac{-||\hat{\mathbf{L}}(o_k) - \mathbf{L}(o_k)||^2}{2\sigma^2}\right)$$



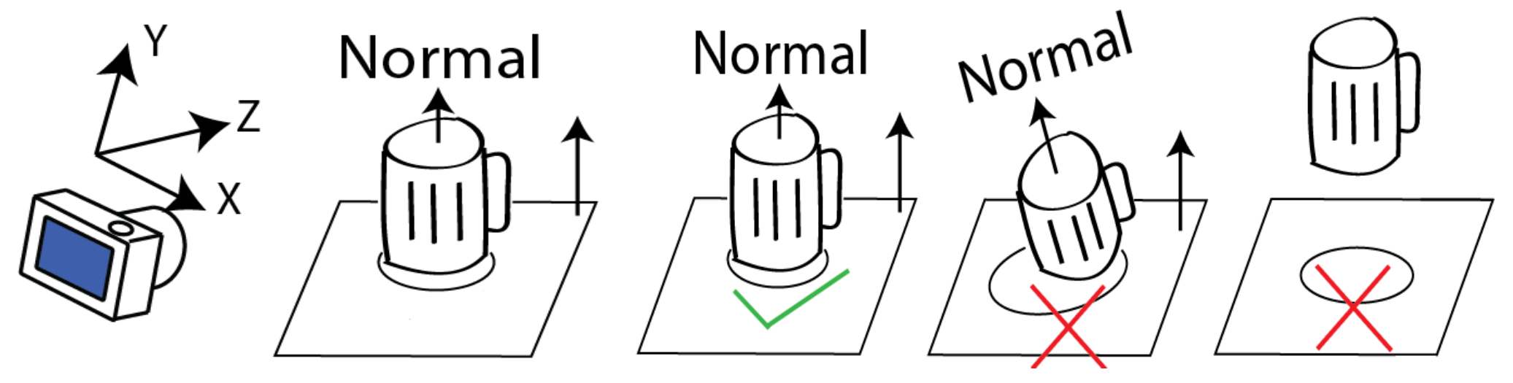
Our method runs in real time, and avoids the need for extensive prior training and the Manhattan-world assumption, which makes it more practical and efficient for an intelligent robot to understand its surroundings compared to most previous scene understanding methods. Experimental results on a collection of indoor videos suggest that our method is capable of an unprecedented combination of accuracy and efficiency.

Semantic Constraints in Vision

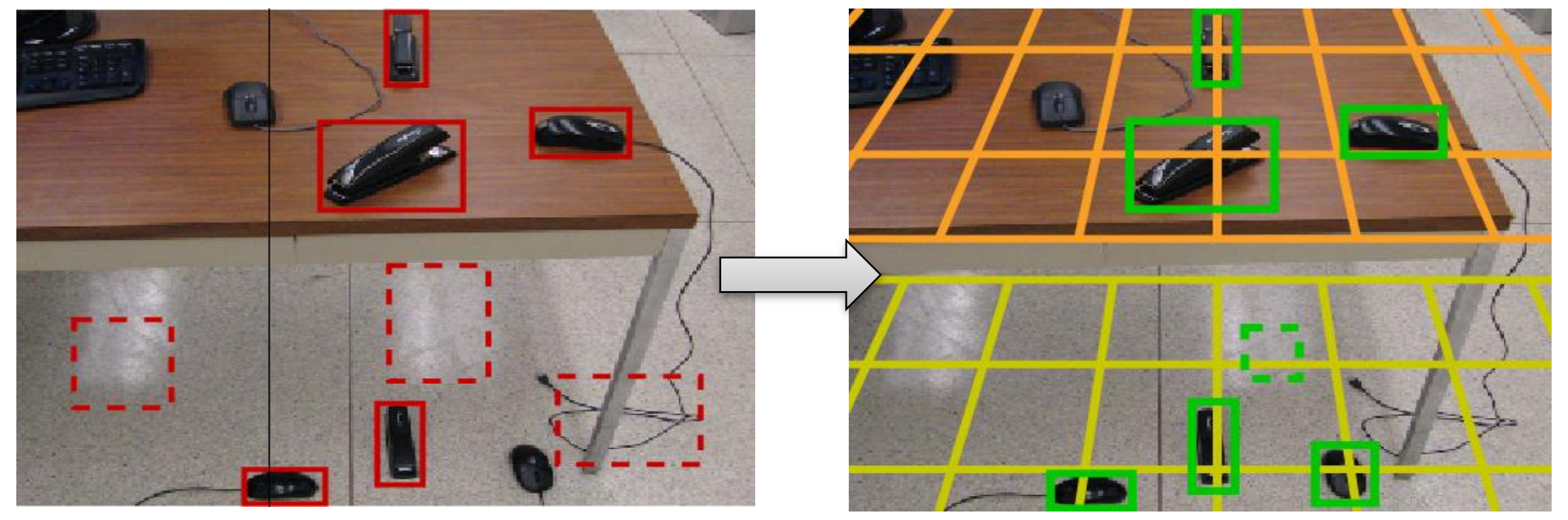
- **Overview**
  - Single un-calibrated image
  - Improve object detection's accuracy
  - Estimate camera pose and focal length
  - Recover 3D supporting planes
  - Locate object in 3D space



- **Tools**
  - Novel relationship between object's pose & location, and supporting plane
  - Layout priors



- Experimental Results

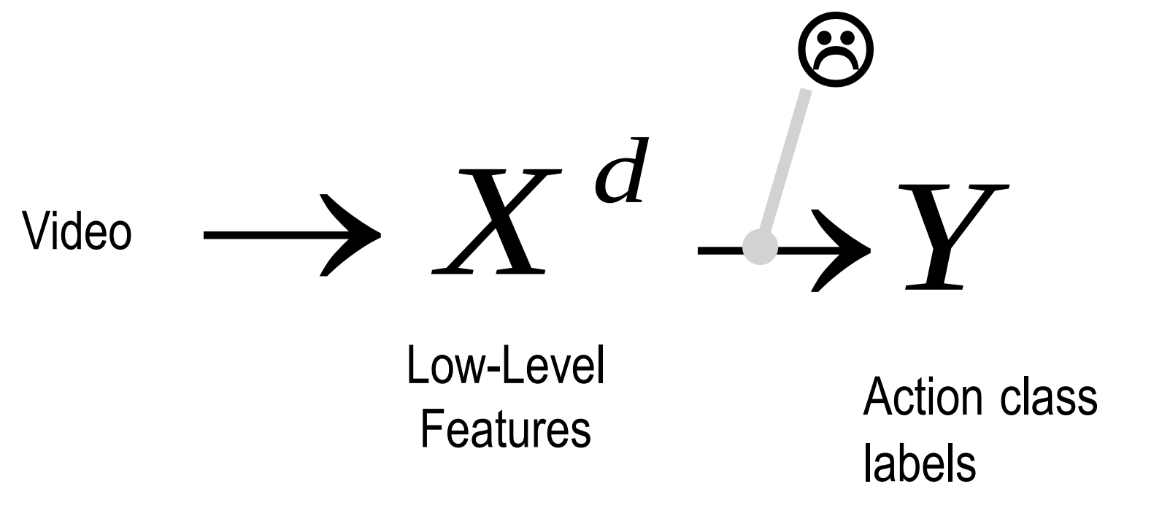


References

- J. Liu, B. Kuipers, S. Savarese, *Recognizing Human Actions by Attributes*, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011
- J. Liu, M. Shah, B. Kuipers, S. Savarese, *Cross-View Action Recognition via View Knowledge Transfer*, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
- Y. Bao, M. Sun & S. Savarese, Toward coherent object detection and scene layout understanding. CVPR, 2010.
- M. Sun, G. Bradsky, B. Xu & S. Savarese, Depth-encoded Hough voting for joint object detection and shape recovery. ECCV, 2010.
- G. Tsai, C. Xu, J. Liu & B. Kuipers, Real-time indoor scene understanding using Bayesian filtering with motion cues. ICCV, 2011.
- C. Xu & B. Kuipers, Towards the Object Semantic Hierarchy. ICDL, 2010.

Learning Human Actions by Attributes

Traditional paradigm

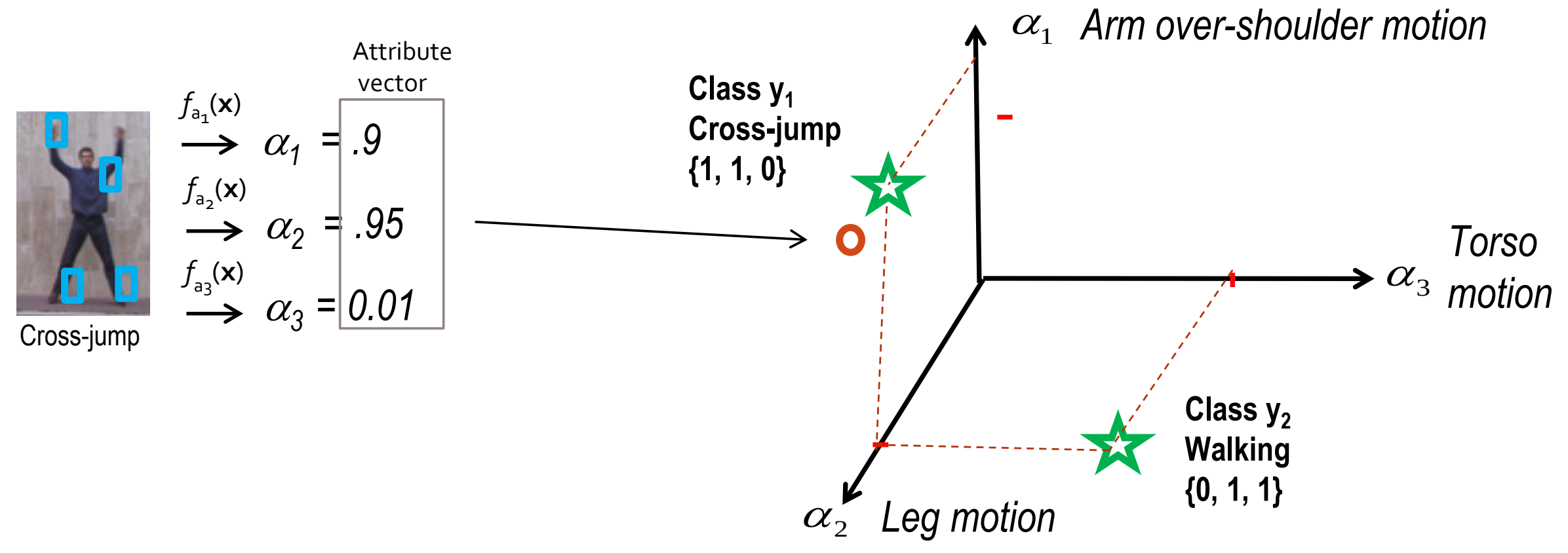


- Rich visual temporal-spatial structures cannot be well characterized by a single class label
- For complex activities this process is too restrictive and reductive

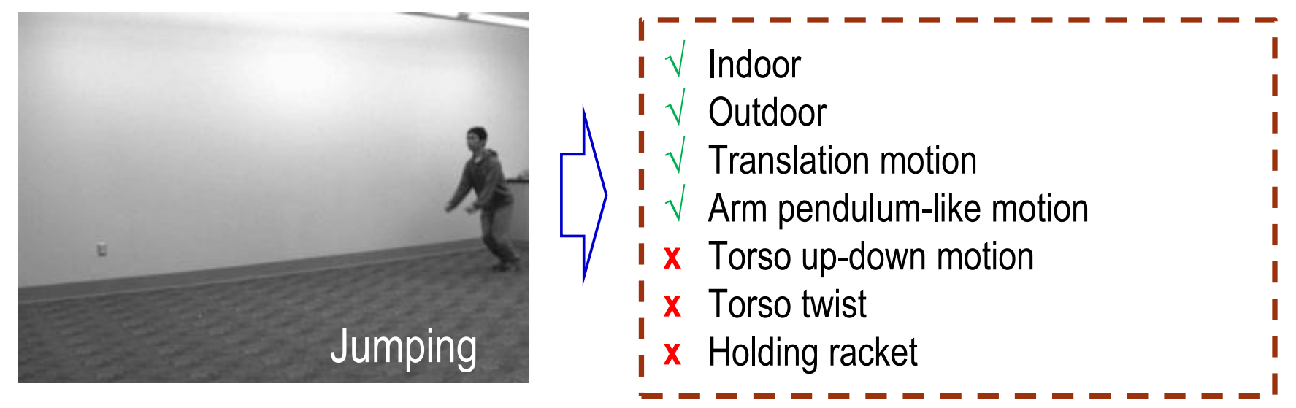
Proposed paradigm

- The action classifier $F: X^d \to Y$ can be decomposed into:

$$\boxed{S: X^d \to A^m} \quad \boxed{L: A^m \to Y}$$



- **Experimental Results**



√ Indoor
√ Outdoor
√ Translation motion
√ Arm pendulum-like motion
x Torso up-down motion
x Torso twist
x Holding racket

		Average Accuracy (%)
Activity data set Niebles et al, 2010	raw-feature	51.83
	specified attributes	60.48
	raw-feature + specified attributes	63.60
	data-driven attributes	45.31
	raw-feature + all attributes	65.09

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. CPS-0931474 and IIS-0713150. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.