# Inferring Human Intent from Video by Sampling Hierarchical Plans

Shaorong Yan, Weilun Ding

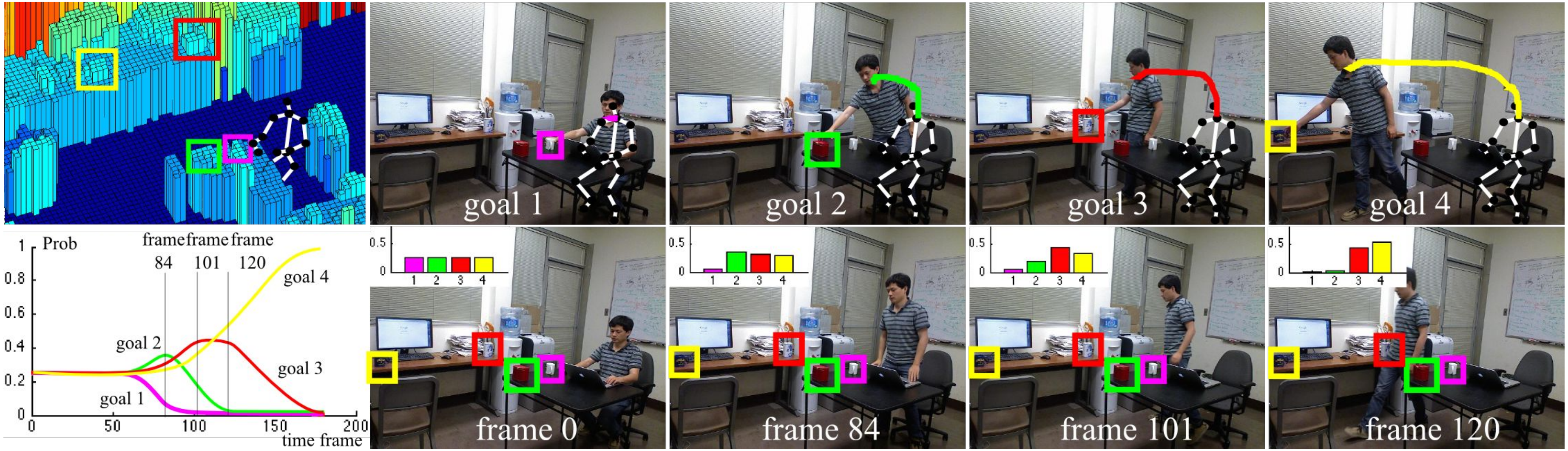
# Reasoning about intentions

- Hard for robots
- Hard for us
- ? for non-human primates
- Important for observational learning

# How to teach robot to do this?

- Assumption
  - Human planning is optimal
  - The agent (human) has perfect knowledge about the scene
- Key issue
  - Infer the agent’s intent
  - Represent the state of the scene
- Approach
  - Co-infer intent and scene representation
  - And-Or graph (AoG): Hierarchical, Compositional, Probabilistic
  - Particle filtering-like algorithm: only tracking the most likely explanation over time

# Goal of the model



# Basic steps

- Define the posterior distribution over plans;
- Compute probabilities over the And-Or graph and specific parse graphs;
- Simulate trajectories for a given parse graph;
- Compare simulated and observed trajectories;
- Update the distribution of plans.

# Renovation

- Generative hierarchical, compositional, and probabilistic And-Or graph.
- Infer long-term planning dependencies  and context-sensitive policies.
- Jointly infer object recognition, action detection, and intent.

# Temporal And-Or Graph (T-AoG)

- Grammar

root node

production rules

$$S = < \quad S; \quad V_n; \quad T; \quad R; \quad P_i >$$

non-terminal nodes

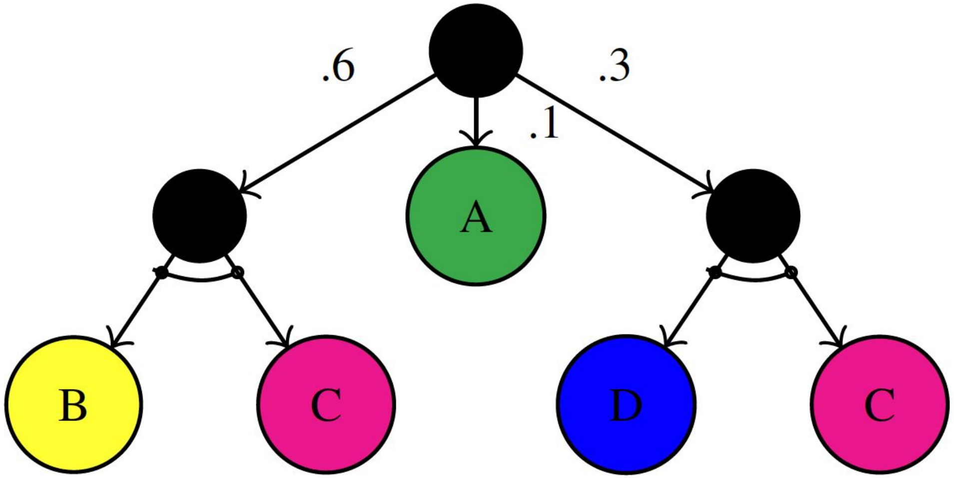
terminal nodes

probabilities on production rules

- AND nodes:
  - Constrain their children to be executed in sequence (temporal).
  - Production probability of 1
- Or Nodes:
  - Associated with a probability $w_i$

# Parse Graph (pg)

- A valid sequence generated by the grammar
- Corresponding to one plan



# Calculate posterior

$$
\begin{aligned}
P(pg \mid X_{\text{obs}}) &\propto P(pg)P(X_{\text{obs}} \mid pg) \\
&\propto \sum_{X_{\text{pred}}} P(pg)P(X_{\text{pred}} \mid pg)P(X_{\text{obs}} \mid X_{\text{pred}}) \\
&\propto \sum_{X_{\text{pred}}} P(pg)\delta_{f(pg,X)}(X_{\text{pred}})P(X_{\text{obs}} \mid X_{\text{pred}}),
\end{aligned}
$$

$\boldsymbol{\delta_{f(pg;X)}(X_{pred})}$ : whether current $X_{pred}$ can be generated from the pg

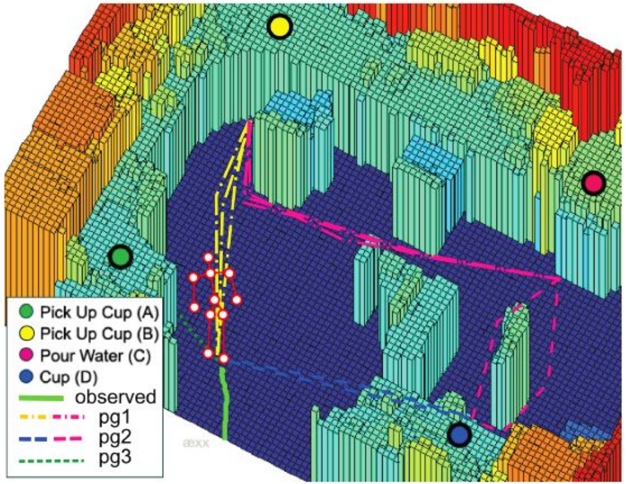
**f(pg, X)**: hierarchical planner

$$
pg^{\star} = \arg\max_{pg} P(pg \mid X_{\text{obs}})
$$

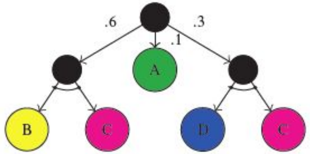
# Modeling Intent

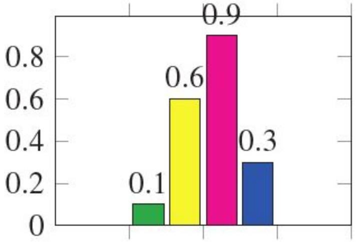
- Represent intent as a temporal And-Or graph (T-AoG)

$$q(n) = \begin{cases} \omega_i \times q(\mathrm{child}(n)) & \text{If } n \text{ is an } \textsc{Or}\text{-node} \\ \prod_i q(\mathrm{children}(n)) & \text{If } n \text{ is an } \textsc{And}\text{-node} \\ T_i & \text{If } n \text{ is a terminal node} \end{cases}$$



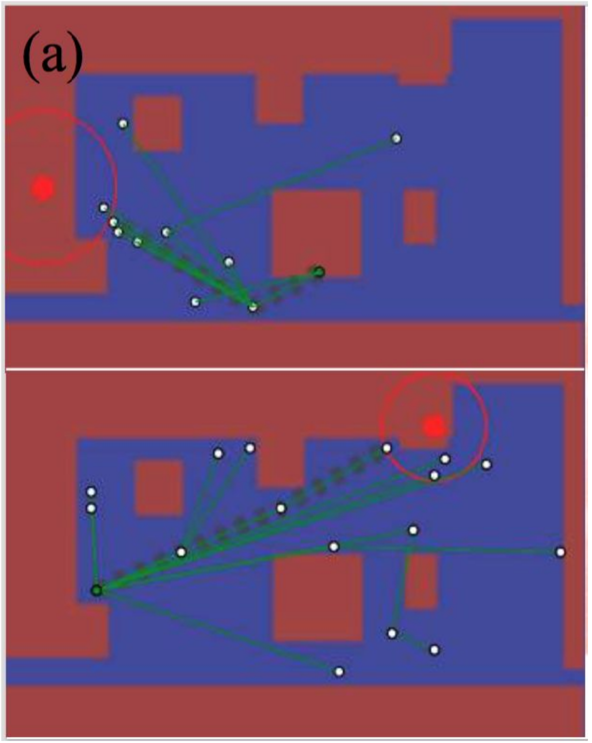

(a) Time $t_1$



# Rapidly-Exploring Random Tree* (RRT*)

- Generate terminal nodes
- Finding minimal cost path from one location to another:
  - $f(\ pg\ ;\ B) \rightarrow X_{pred}$
  - B:  background collision map
- $P(\ X_{obs} \mid X_{pred})$



# Dynamic Time Warping (DTW)

- Measure similarity between two temporal sequences varying in time or speed.

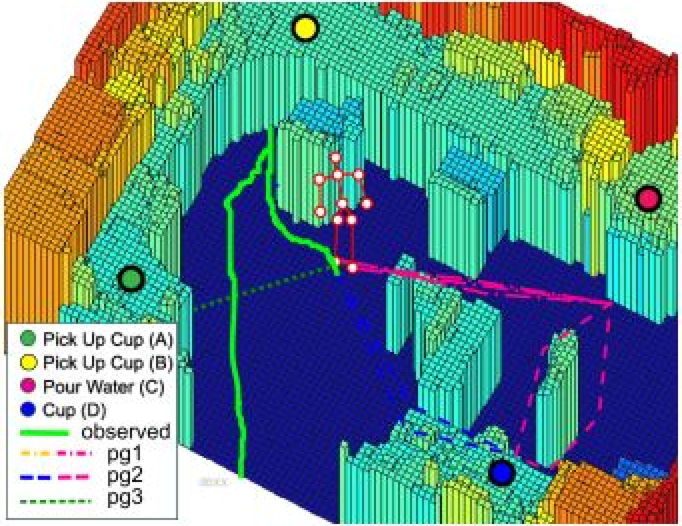
- Loss: the Euclidean distance between the observed trajectories and the complete predicted/simulated trajectories, fed into a stochastic likelihood function

- P(X_obs | X_pred): feed the loss into a stochastic likelihood function, and larger loss leads to lower probability.

# Stochastic Inference

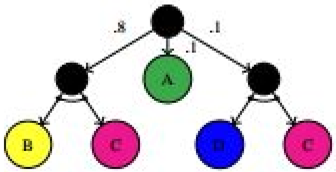
$$\lim_{N \to \infty} \frac{\#(O_i^j)}{\#(O^j)} = \omega_i .$$

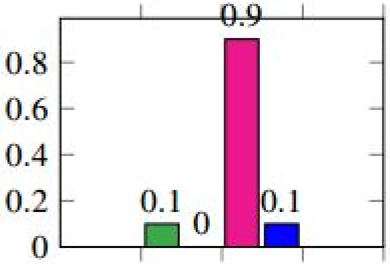
$$\omega_i^{t+1} \propto \frac{\#(O_i^j)}{\#(O^j)} \times \prod_{k \in A} P_{\omega_i^t}(X_{\text{obs}} \mid X_{\text{pred}}^k)$$




(b) Time $t_2$

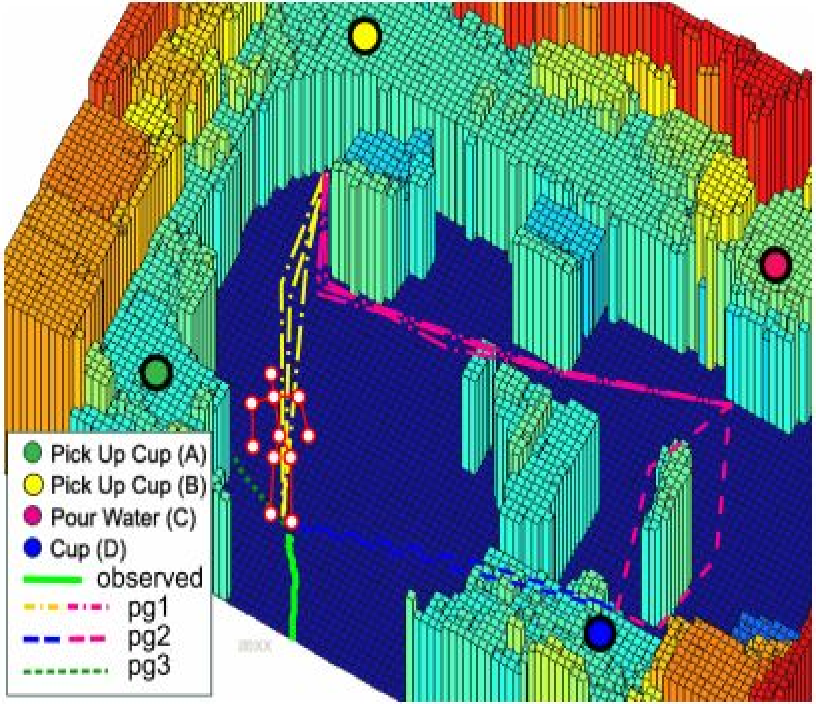
The worse a path/particle performs, the more penalized the corresponding weight is by the rule above.



# Algorithm: Intent Prediction and Reweighting

**Data:** 3D Scene, Video Frames ($V$), Dictionary ($\Delta$)
**Result:** $\omega$ (Parameterized T-AoG)
CollisionMap = SceneReconstruction($V$);
RRT* = RRT*Planner(CollisionMap);
P = Planner($\Delta$);
Particles = [];
**for** *Frame* $v_t$ *in* $V$ **do**
    ObservedTrajectory = ObjectTracking($v_1, \cdots, v_t$);
    PredictedPlan = Planner.sample(Particles);
    PredictedTrajectories = RRT*.search(PredictedPlan);
    Loss = DTW(ObservedTrajectory,
      PredictedTrajectory);
    Planner.reweight(Loss);
    Particles = PredictedTrajectories;
**end**
**return** Planner.weights



# Intent Prediction (30 candidate Intents)

- Euclidean Distance(ED): predict the nearest goal in the scene
- ED + Grammar: ED + prior weights for knowledge of actions
- ? Accuracy decreases with more observation

TABLE II

INTENT PREDICTION ACCURACY

| % of observation | 90% | 70% | 30% | 10% |
|---|---|---|---|---|
| Euclidean Distance (E.D.) | 5% | 0% | 0% | 0% |
| E.D. w/ Grammar Prior | 13.1% | 10.5% | 6% | 3% |
| Inverse Planning w/o Hierarchy | 15.8% | 13.2% | 10.5% | 10.5% |
| Ours | 28.9% | 13.3% | 18.4% | 15.8% |

# Action Recognition

- Input: -3 sec ~ +3 sec surrounding the action time
- ICCV13: Use wavelet features representing action sequences together with temporal logic describing the actions relations
- “Better at recognizing the action when an object is involved”

TABLE III

ACTION RECOGNITION ACCURACY

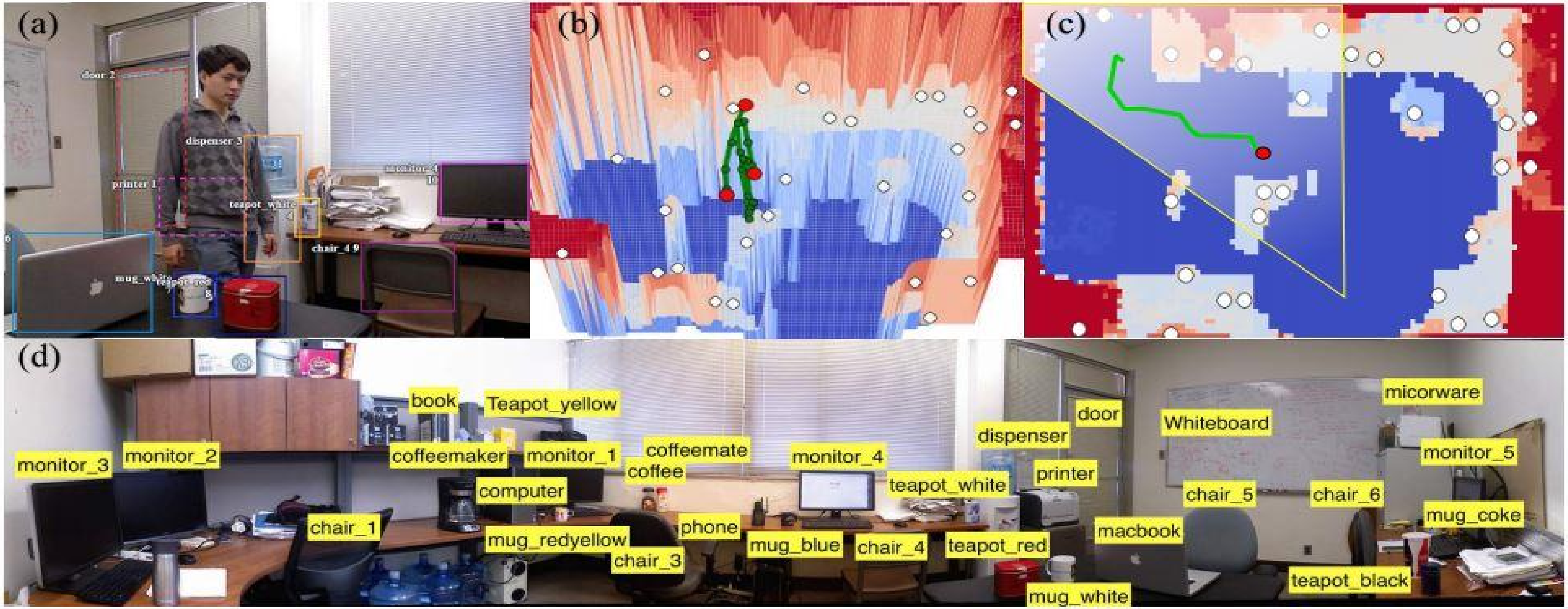
| | SVM | ICCV13 [36] | Ours |
|---|---|---|---|
| walk | 88% | 98% | 91% |
| stand_up | 68% | 94% | 92% |
| sit_down | 65% | 92% | 92% |
| grasp | 43% | 64% | 59% |
| put | 25% | 44% | 53% |
| fetch | 33% | 54% | 83% |
| touch | 35% | 41% | 54% |
| drink | 70% | 91% | 91% |
| call | 65% | 89% | 94% |
| eat | 22% | 54% | 73% |

# Object Tracking

TABLE IV

OBJECT TRACKING ACCURACY

| | ICCV11[15] | ICCV13[30] | Ours |
|---|---|---|---|
| No Occlusion | 34% | 72% | 74% |
| With Occlusion | - | - | 35% |
| All Frames | - | - | 65% |



# Conclusions

- To some extent realizing inference of human hierarchical plans through robotic imagination with input of observed actions and rationality assumption.

- Advantage: Unlimited by the hierarchical depth of the plan or the time length

- Disadvantage: Planning dictionary should be provided a-priori : learn weights for different plans with the tree structure given

# Discussion

- This paper: Model human intent through observed motion patterns
- How to relax the rationality hypothesis
- What if make use of the speed variation in movement? Or other forms of information can be integrated?
- Further improvement might be gained through brain imaging data (e.g. motor area) to infer human intents?