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**
  
  \[ S = \langle S; V_n; T; R; P_i \rangle \]

- **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

\[
P(pg \mid X_{obs}) \propto P(pg)P(X_{obs} \mid pg)
\]

\[
\propto \sum_{X_{pred}} P(pg)P(X_{pred} \mid pg)P(X_{obs} \mid X_{pred})
\]

\[
\propto \sum_{X_{pred}} P(pg)\delta_f(pg; X)(X_{pred})P(X_{obs} \mid X_{pred}),
\]

\(\delta_f(pg; X)(X_{pred})\): whether current \(X_{pred}\) can be generated from the \(pg\)

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

\[
pg^* = \arg \max_{pg} P(pg \mid X_{obs})
\]
Modeling Intent

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

\[ q(n) = \begin{cases} 
\omega_i \times q(\text{child}(n)) & \text{If } n \text{ is an OR-node} \\
\prod_i q(\text{children}(n)) & \text{If } n \text{ is an AND-node} \\
T_i & \text{If } n \text{ is a terminal node}
\end{cases} \]
Rapidly-Exploring Random Tree* (RRT*)

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

- \( P(X_{\text{obs}} | X_{\text{pred}}) \): feed the loss into a stochastic likelihood function, and larger loss leads to lower probability.
Stochastic Inference

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

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

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 = \text{SCENE}\text{RECONSTRUCTION}(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**

<table>
<thead>
<tr>
<th>% of observation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Euclidean Distance (E.D.)</td>
</tr>
<tr>
<td>E.D. w/ Grammar Prior</td>
</tr>
<tr>
<td>Inverse Planning w/o Hierarchy</td>
</tr>
<tr>
<td>Ours</td>
</tr>
<tr>
<td>90%</td>
</tr>
<tr>
<td>5%</td>
</tr>
<tr>
<td>13.1%</td>
</tr>
<tr>
<td>15.8%</td>
</tr>
<tr>
<td>28.9%</td>
</tr>
</tbody>
</table>
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”
Object Tracking

TABLE IV

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No Occlusion</td>
<td>34%</td>
<td>72%</td>
<td>74%</td>
</tr>
<tr>
<td>With Occlusion</td>
<td>-</td>
<td>-</td>
<td>35%</td>
</tr>
<tr>
<td>All Frames</td>
<td>-</td>
<td>-</td>
<td>65%</td>
</tr>
</tbody>
</table>
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?