Toward Computer Vision Systems that Understand Real-World Assembly Processes

Jonathan D. Jones, Gregory D. Hager, Sanjeev Khudanpur

JOHNS HOPKINS UNIVERSITY

IEEE 2019 Winter Conference on Applications of Computer Vision

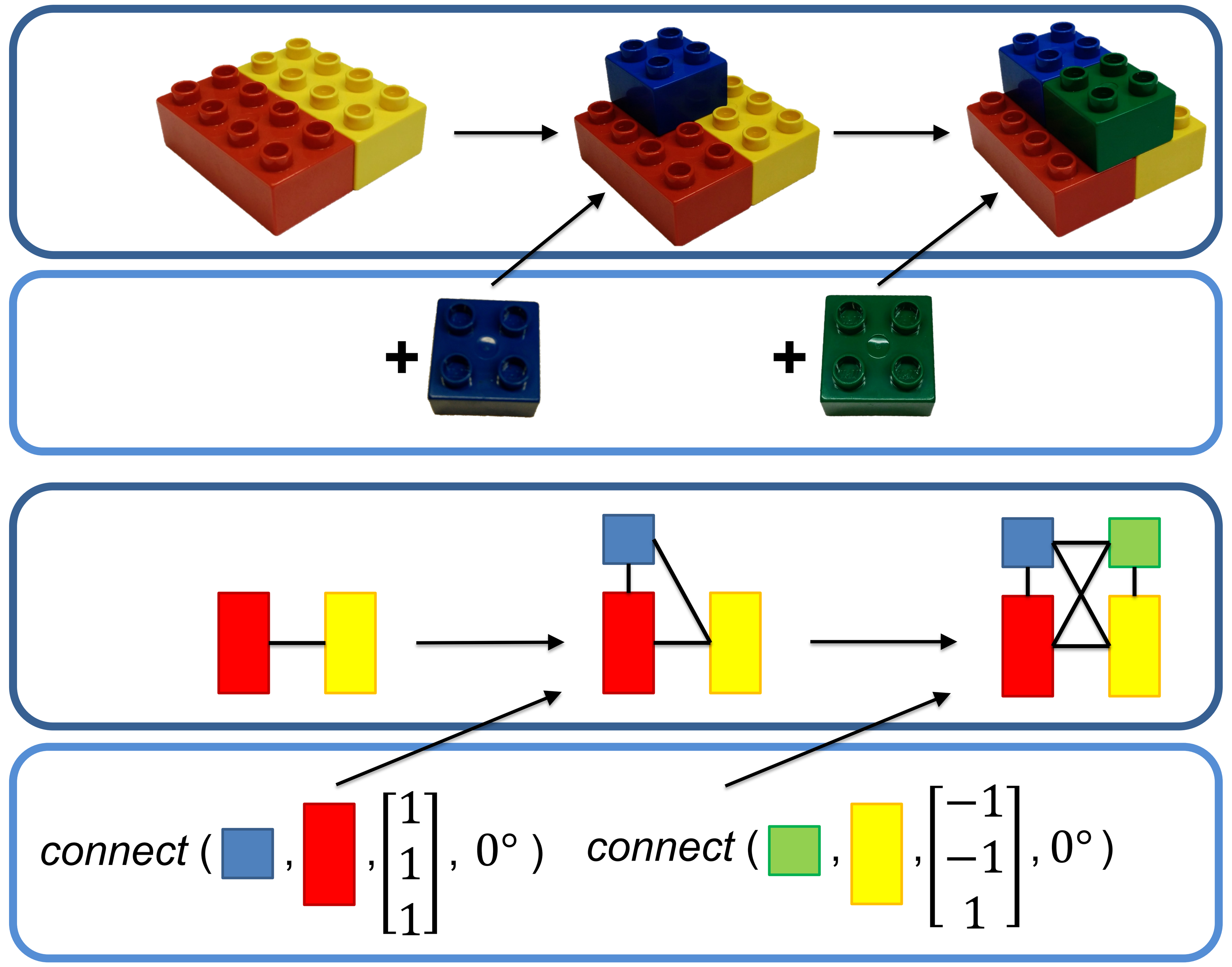
Introduction

GOAL: Build systems that understand how objects can be assembled to form larger parts, or how parts can be disassembled into constituent objects

USES: Collaborative robotics, industrial monitoring, information retrieval

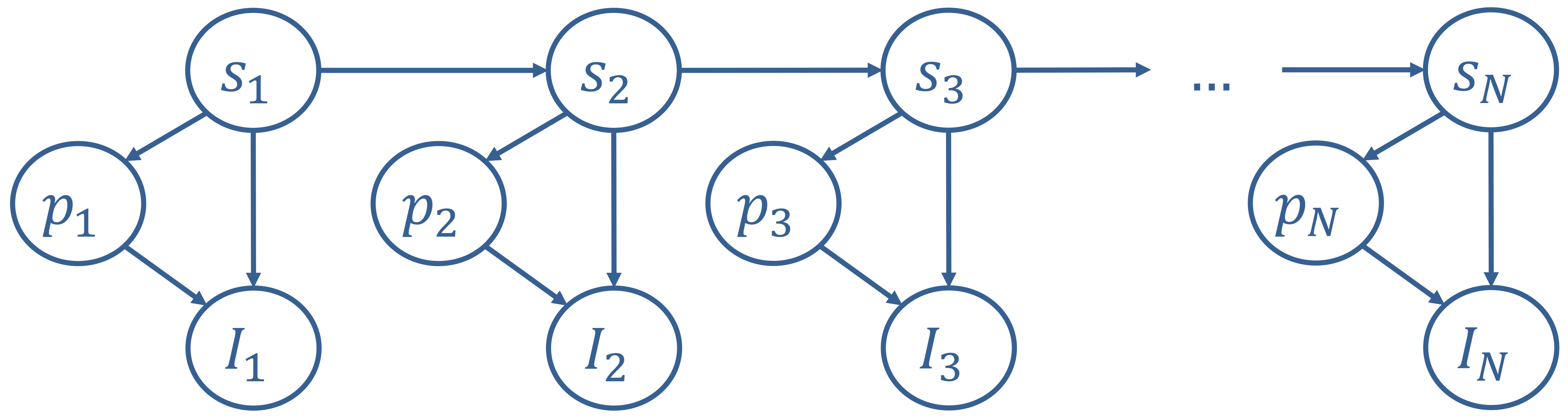
APPLICATION: Parsing DUPLO block structures in videos from child behavioral experiments

Representing spatial assemblies



We represent a spatial assembly as an edge-labeled graph (labels not shown). Vertices are objects, edges are connections, and edge labels are object relative poses. The state of an assembly can be changed by actions, which add or remove edges.

Method



We derive a time-series graphical model for parsing assembly states $s_{1:N}$ and poses $p_{1:N}$ from a sequence of video keyframes $I_{1:N}$.

Algorithm: Hypothesize and test

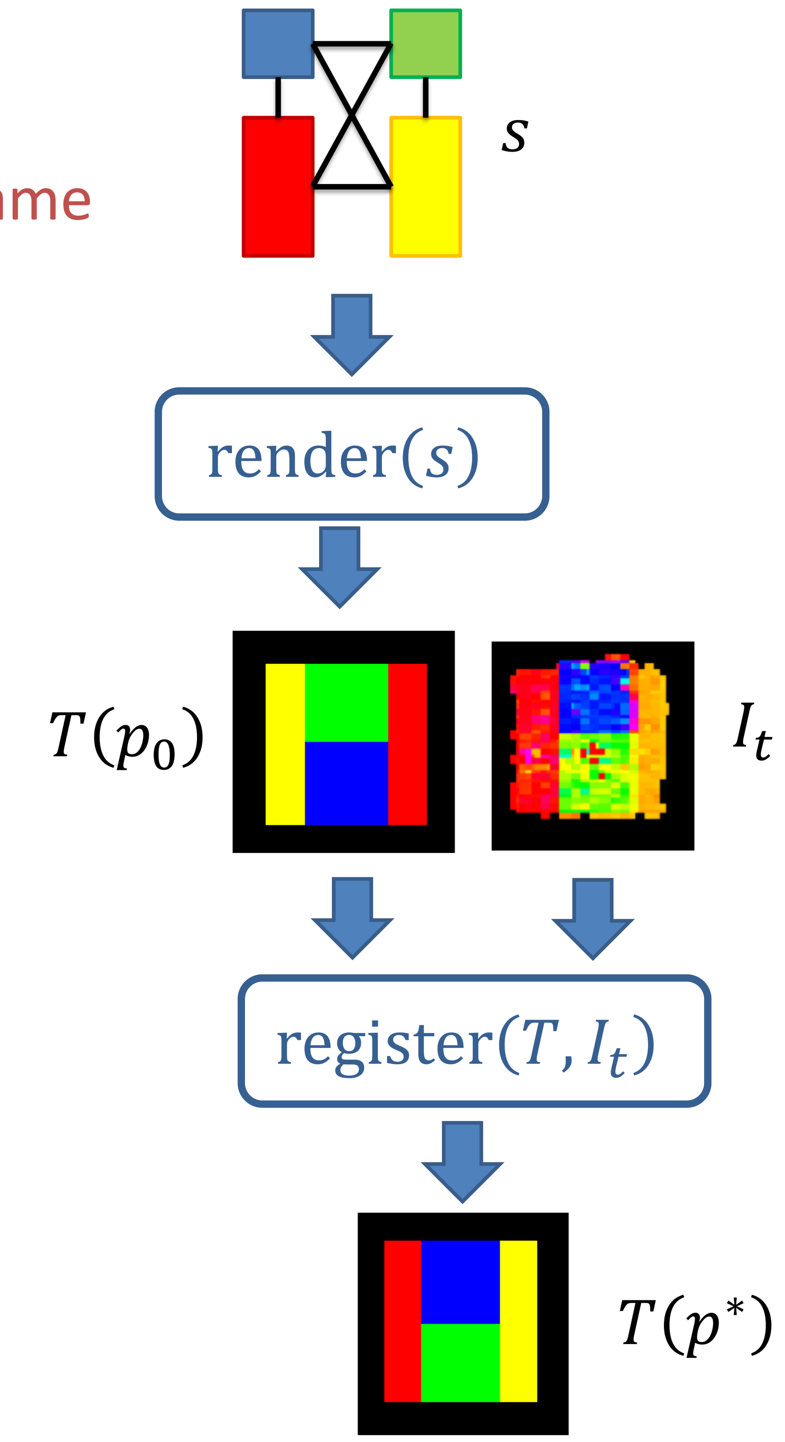
1. Generate a set of state hypotheses for each keyframe
2. Test each hypothesis locally
3. Decode state sequence globally

for $t \in \{1, \dots, N\}$:

  // new hypotheses: any possible transition from
  // previous best hypotheses
  $\mathcal{H}_{t-1} \leftarrow \mathrm{prune}(\mathcal{H}_{t-1})$
  $\mathcal{H}_t \leftarrow \mathrm{advance}(\mathcal{H}_{t-1})$

  // test each hypothesis and store resulting
  // probabilities in δ
  for $s \in \mathcal{H}_t$:
    $T \leftarrow \mathrm{render}(s)$
    $p_t^* \leftarrow \mathrm{register}(T, I_t)$
    $\delta[s,t] \leftarrow \log \mathrm{P}(s_{1:t}, p_{1:t}^*, I_{1:t})$

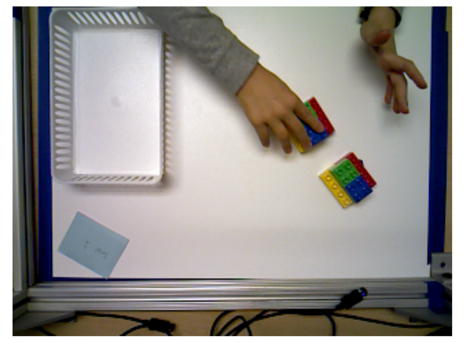
// find most probable state sequence (Viterbi)
$s_{1:N}^* \leftarrow \mathrm{decode}(\delta)$



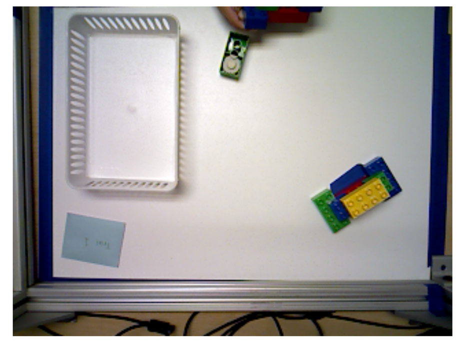
To test a hypothesis s, we render a template T in initial pose p_0, then register to find the best pose p^*.

Experiments

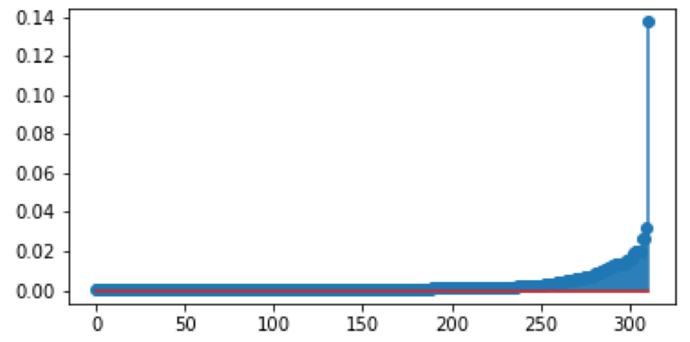
"Child's play" dataset



occlusion



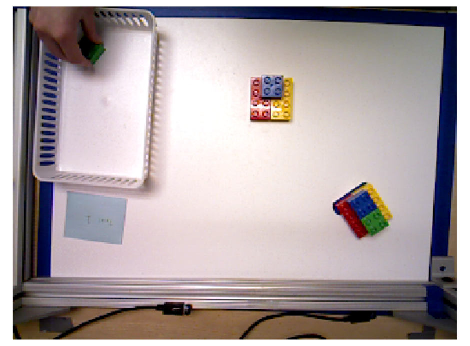
out-of-view state changes



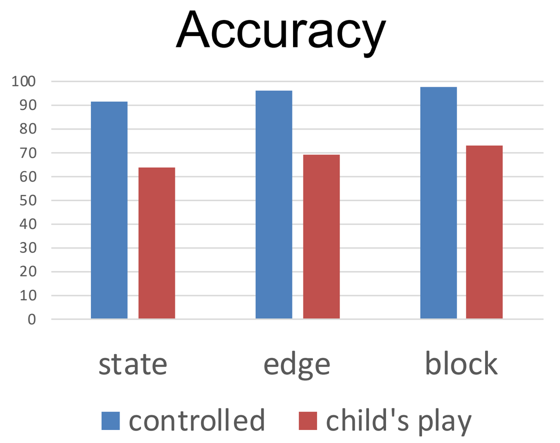
unbalanced state distribution

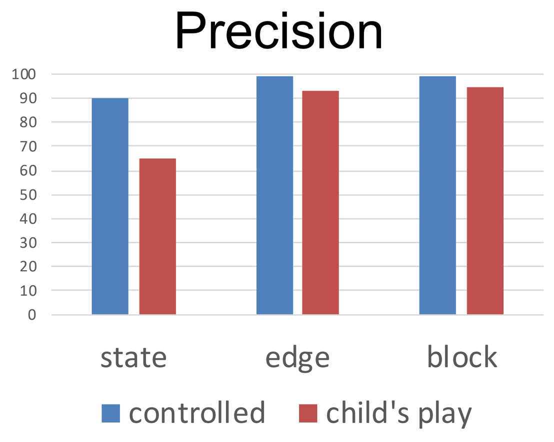
Controlled dataset

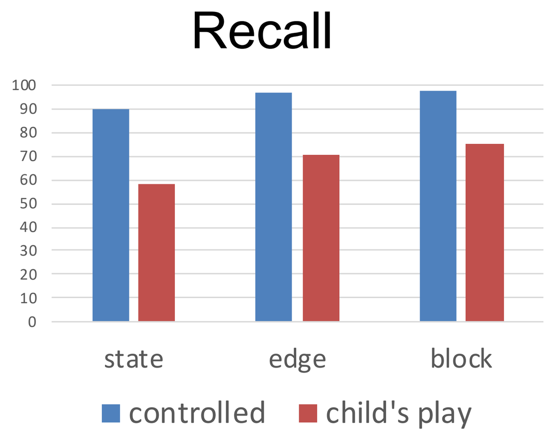
- Camera has a clear view of every state
- Tests algorithm performance under conditions that match the model

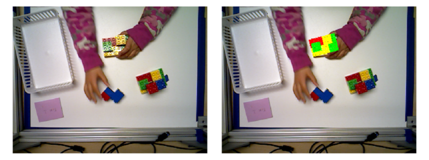


Results

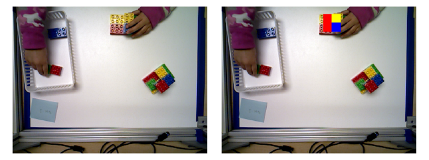




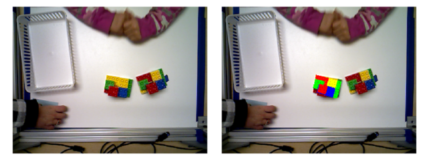




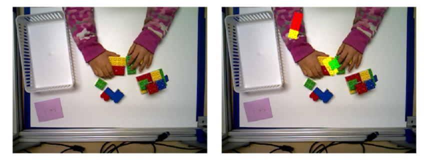
partly-occluded state recovered accurately



sub-part of wrong state matched accurately



yellow rectangle confused for red; true state OOV



scene clutter and weak prior lead to incorrect inference

Acknowledgements: This work was supported by NSF award No. 1561278. We also gratefully acknowledge our collaborators Cathryn Cortesa, Barbara Landau, and Amy Shelton for their leading role in the underlying behavioral experiments.