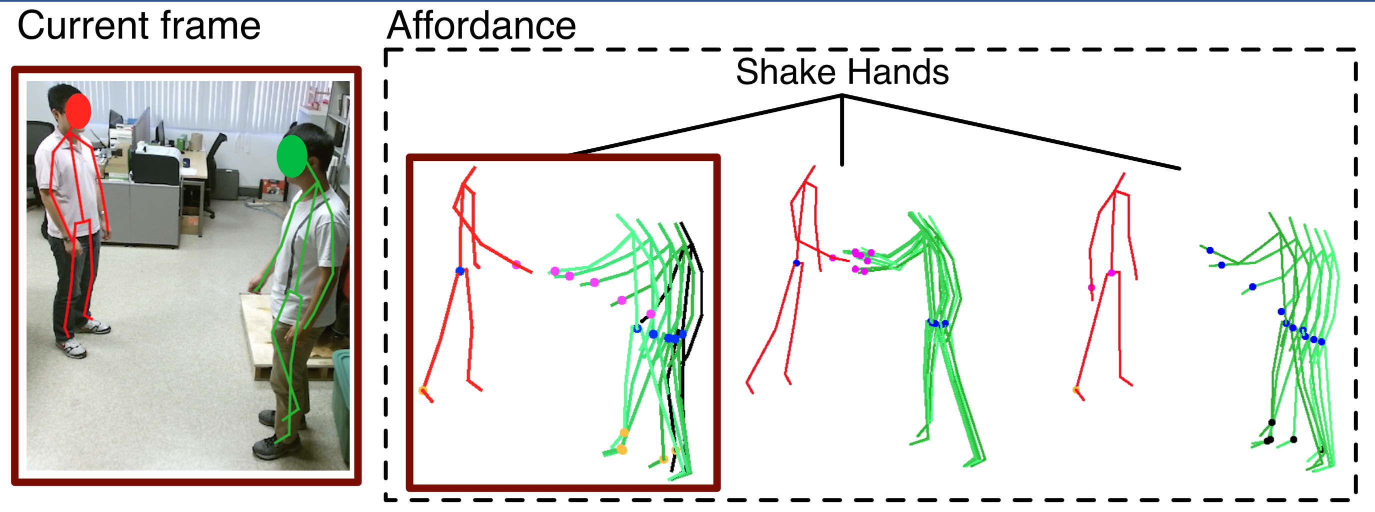
Learning Social Affordance for Human-Robot Interaction

Tianmin Shu[1] M. S. Ryoo[2] Song-Chun Zhu[1]

[1]Center for Vision, Cognition, Learning, and Autonomy, UCLA [2]Indiana University, Bloomington

Introduction



Objective:

Learning explainable knowledge from the noisy observation of human interactions in RGB-D videos to enable human-robot interactions.

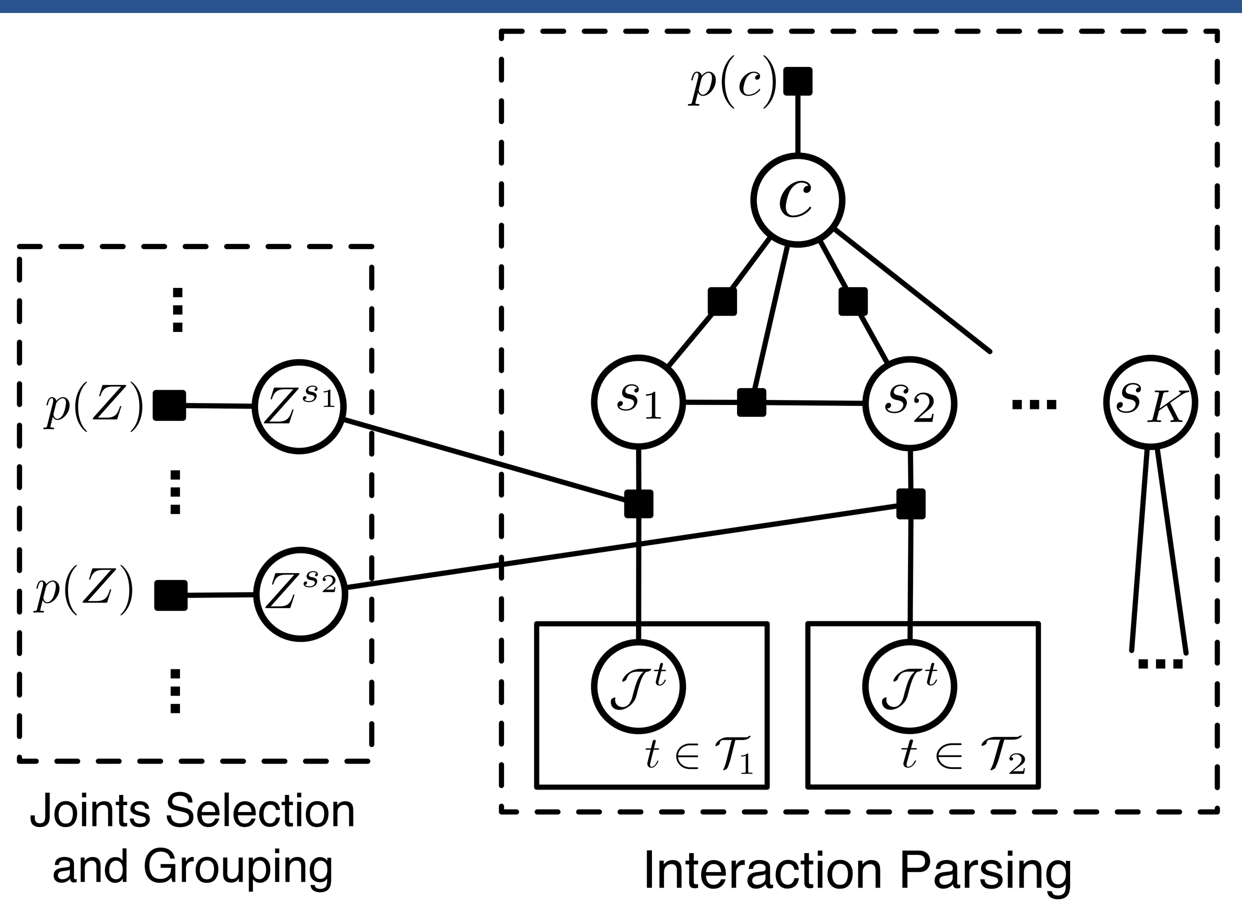
Key idea:

Beyond traditional object and scene affordances, we propose a weakly supervised learning of social affordances for HRI.

Contributions:

- First formulation and hierarchical representation of social affordance
- Weakly supervised learning from noisy skeleton input
- Efficient motion synthesis based on learned hierarchical affordances

Model



$$p(G|Z_c) \propto \underbrace{\prod_k p(\{J^t\}_{t\in\mathcal{T}_k}|Z^{s_k}, s_k, c)}_{\text{likelihood}} \cdot \underbrace{p(c)}_{\text{interaction prior}}$$

$$\cdot \underbrace{\prod_{k=2}^{K} p(s_k|s_{k-1}, c)}_{\text{sub-event transition}} \cdot \underbrace{\prod_{k=1}^{K} p(s_k|c)}_{\text{sub-event prior}}$$

$$p(\{J^t\}_{t\in\mathcal{T}}|Z^s, s, c) = \Psi_g(\{J^t\}_{t\in\mathcal{T}}, Z^s, s)\Psi_m(\{J^t\}_{t\in\mathcal{T}}, Z^s, s)$$

$$p(Z_c) = \prod_{s\in\mathcal{S}} p(Z^s|c)$$

For one instance of category *c*: $p(G, Z_c) = p(G|Z_c)p(Z_c)$

For *N* training examples of category *c* ($\mathcal{G} = \{G^n\}_{n=1,\dots,N}$):

$$p(\mathcal{G}, Z_c) = p(Z_c)\prod_n^N p(G^n|Z_c)$$

Learning

Goal:

Obtain the optimal joint selection and grouping Z_c and interaction parsing results $\mathcal{G} = \{G^n\}_{n=1,\dots,N}$ by maximizing the joint probability.

Algorithm:

Initialization Skeleton clustering for initial sub-event parsing

Outer loop

A Metropolis-Hasting algorithm for latent sub-event parsing

Dynamics: splitting/merging/relabeling

Inner loop A Gibbs sampling for our modified CRP

$$z_{ai}^s \sim p(\mathcal{G}|Z_c')p(z_{ai}^s|Z_{-ai}^s)$$

$$p(z_{ai}^s|Z_{-ai}^s) = \begin{cases} \beta\dfrac{\gamma}{M-1+\gamma} & \text{if } z_{ai}^s > 0, M_{z_{ai}^s} = 0 \\ \beta\dfrac{M_{z_ai}^s}{M-1+\gamma} & \text{if } z_{ai}^s > 0, M_{z_{ai}^s} > 0 \\ 1-\beta & \text{if } z_{ai}^s = 0 \end{cases}$$

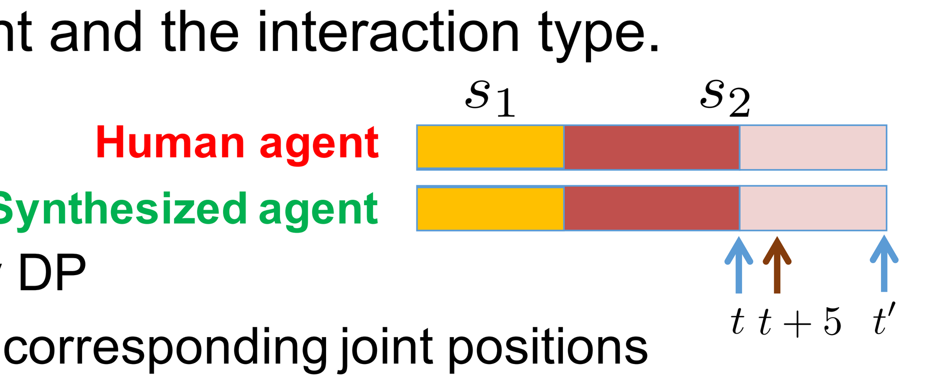
Motion Synthesis

Goal: Given the initial 10 frames (25 fps), synthesize the motion of an agent given the motion of the other agent and the interaction type.

Algorithm:

At time t,

1) Estimate the current sub-event by DP
2) Predict the ending time t' and the corresponding joint positions
3) Obtain the joint positions at $t+5$ through interpolation

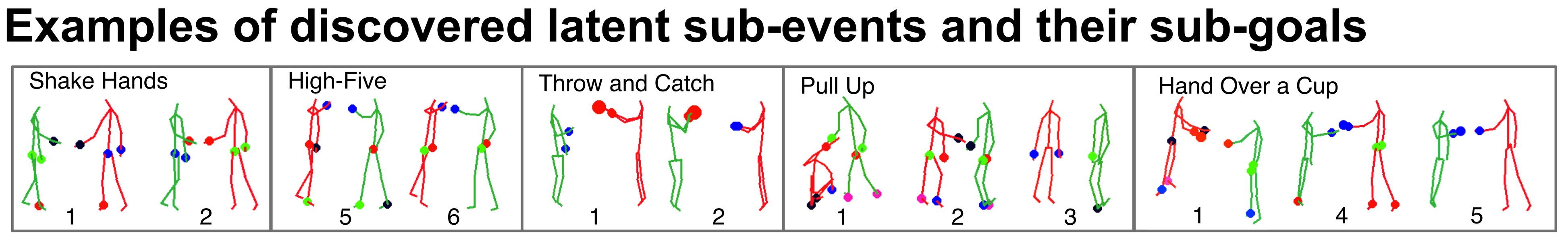


Experiment

UCLA Human-Human-Object Interaction Dataset

- Five types of interactions; on average, 23.6 instances per interaction performed by totally 8 actors. Each lasts 2-7 s presented at 10-15 fps.
- RGB-D videos, skeletons and annotations are available: http://www.stat.ucla.edu/~tianmin.shu/SocialAffordance

Examples of discovered latent sub-events and their sub-goals



Exp 1: Average joint distance in meters (compared with GT skeletons)

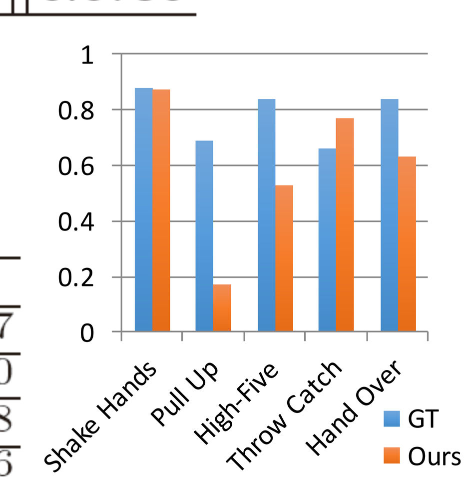
Method	Shake Hands	Pull Up	High-Five	Throw Catch	Hand Over	Average
HMM	0.362	0.344	0.284	0.189	0.229	0.2816
V1	0.061	0.144	0.079	0.091	0.074	0.0899
V2	0.066	0.231	0.090	0.109	0.070	0.1132
Ours	**0.054**	**0.109**	**0.058**	**0.076**	**0.068**	**0.0730**

Exp 2: User study (14 subjects)

Q1: Successful? Q2: Natural? Q3: Human vs. robot?

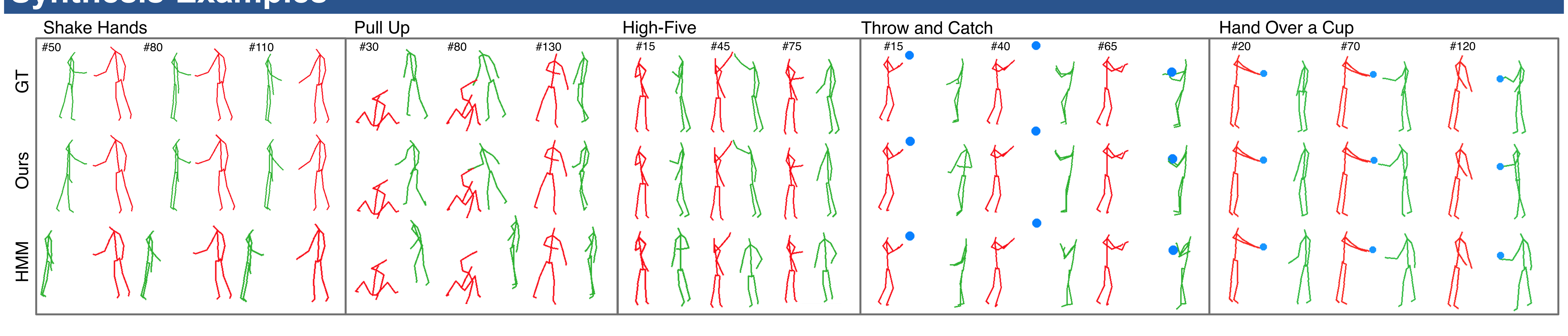
From 1 (worst) to 5 (best)

	Source	Shake Hands	Pull Up	High-Five	Throw & Catch	Hand Over
Q1	Ours	$\mathbf{4.60} \pm 0.69$	3.90 ± 0.70	4.53 ± 0.30	$\mathbf{4.31} \pm 0.89$	4.40 ± 0.37
	GT	4.50 ± 0.82	4.29 ± 0.58	4.64 ± 0.33	4.20 ± 0.76	4.64 ± 0.30
Q2	Ours	$\mathbf{4.23} \pm 0.34$	2.80 ± 0.75	3.70 ± 0.47	$\mathbf{4.06} \pm 0.83$	3.89 ± 0.38
	GT	4.20 ± 0.47	4.23 ± 0.48	4.64 ± 0.17	3.86 ± 0.53	4.24 ± 0.46
Q3	Ours	4.23 ± 0.50	2.63 ± 0.60	3.57 ± 0.73	$\mathbf{4.03} \pm 0.88$	3.69 ± 0.64
	GT	4.30 ± 0.60	3.71 ± 1.15	4.40 ± 0.63	3.97 ± 0.74	4.40 ± 0.24



Frequencies of high scores (4 or 5) for Q3

Synthesis Examples



Acknowledgment

This research has been sponsored by grants DARPA SIMPLEX project N66001-15-C-4035 and ONR MURI project N00014-16-1-2007.

Scan to visit our project website