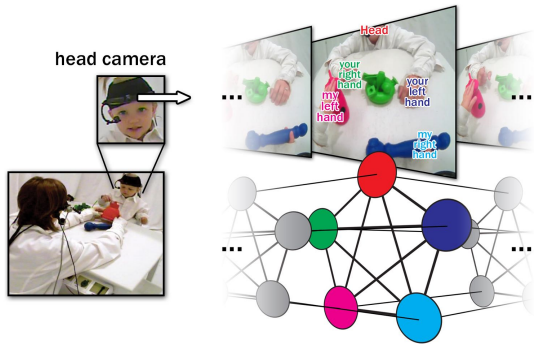
This Hand Is My Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric Video

Stefan Lee, Sven Bambach, David Crandall, School of Informatics and Computing, Indiana University
John Franchak, Chen Yu, Department of Psychological and Brain Sciences, Indiana University

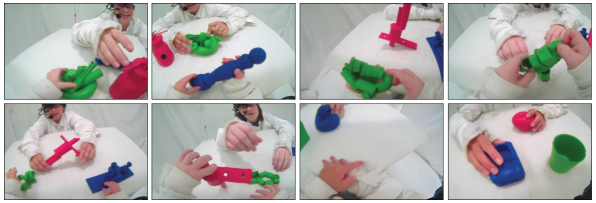
1. Motivation

- We use head-mounted cameras to **study how toddlers interact with parents**, including how they coordinate hands, head turns, and gaze
- To study how different hands help capture the toddler's attention, we need to **detect, disambiguate, and track all hands** in the video
- We use a **probabilistic framework to jointly model head motion and hand position** of interacting people in egocentric video



Our probabilistic framework models paired interaction, incorporating hands' spatial, temporal and appearance constraints in egocentric video.

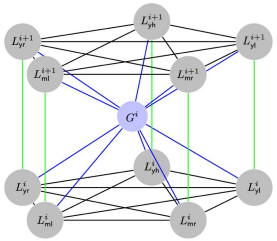
2. Challenges



The child's view is extremely dynamic: hands vary drastically in size, shape, and orientation, and hands come in and out of view and overlap frequently.

3. Modeling Egocentric Interactions

- **Given** an egocentric video sequence $I = \{I^1, \ldots, I^n\}$
- **Estimate** location of parts $P = \{yr, yh, yl, mr, ml\}$ in each frame as latent variables $\{L_p^i\}_{p \in P}^{1 \le i \le n}$
- Also jointly estimate **global shift** G^i between pairs of consecutive frames (caused by head motion)



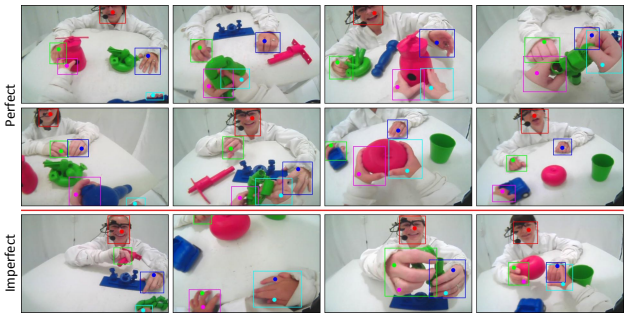
Graphical depiction of our model for a two frame video.

- Use weak **skin, head and arm appearance models** to generate (noisy) likelihood maps within each frame
- **Model spatial constraints** on hand position with a fully-connected graph within each frame
- **Model temporal constraints** with edges between corresponding parts in adjacent frames, and the global shift variables
- **Handle out-of-view parts** with a special $\emptyset$ state, estimating its probability as an integral over the portion of spatial constraints outside the frame
- **Solve using Gibbs sampling,** modeling priors as isotropic normal distributions for efficient inference

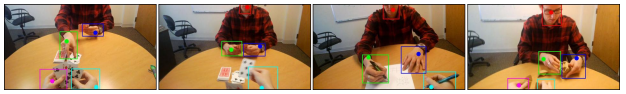
4. Experiments

- We tested our model on **5 different parent-child subject pairs** in a lab setting (31 min of video)
- We additionally captured **naturalistic videos** of two adults using **Google Glass** (4.5 min)
- To evaluate our results, we manually **annotated 2,700 frames** (about 1 frame/second) with ground-truth bounding boxes of head and hands

5. Results



Frames from our lab videos, where rectangles are ground truth bounding boxes and dots are estimated positions. Key: red: your head, blue: your left hand, green: your right hand, magenta: my left hand, cyan = my right hand.



Frames from our naturalistic videos, with participants playing cards, tic-tac-toe, and solving a 3-d puzzle, while one wore Google Glass.

	Overall Accuracy	Observer		Partner				% Perfect Frames	Disambiguation Error Rate
		R. Hand	L. Hand	R. Hand	L. Hand	Head	Head (V-J)		
Lab	68.4	70.7	61.2	63.6	64.5	82.1	72.4	19.1	32.7
Natural	50.7	54.3	18.7	73.3	49.3	57.7	40.3	9.0	35.6

Above: *Detection accuracies for hands and head (compared to a raw Viola-Jones detector).*
Right: *Comparison to various baselines in terms of overall accuracy, percentage of perfect frames and error in disambiguating child hands from parent hands.*

Method	Overall Accuracy	% Perfect Frames	Disambiguation Error Rate
Lab Videos			
random	17.0	0.1	95.1
random (skin)	27.3	4.3	72.0
skin clusters	58.1	14.4	36.0
our method	**68.4**	**19.1**	**32.7**
Naturalistic Videos			
skin clusters	39.2	0.0	65.4
our method	**50.7**	**9.0**	**35.6**

6. Future Work

- More complex, naturalistic video data
- Stronger appearance models
- Joint models of attention and hand/head motion

Acknowledgments: *This work was funded in part by the National Science Foundation (CAREER IIS-1253549), the National Institutes of Health (R01 HD074601 and R21 EY017843), the National Institute of Child Health and Human Development (5T32HD7475-17), and the Indiana University Vice President for Research.*