Tracking Hands of Interacting People in Egocentric Video

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

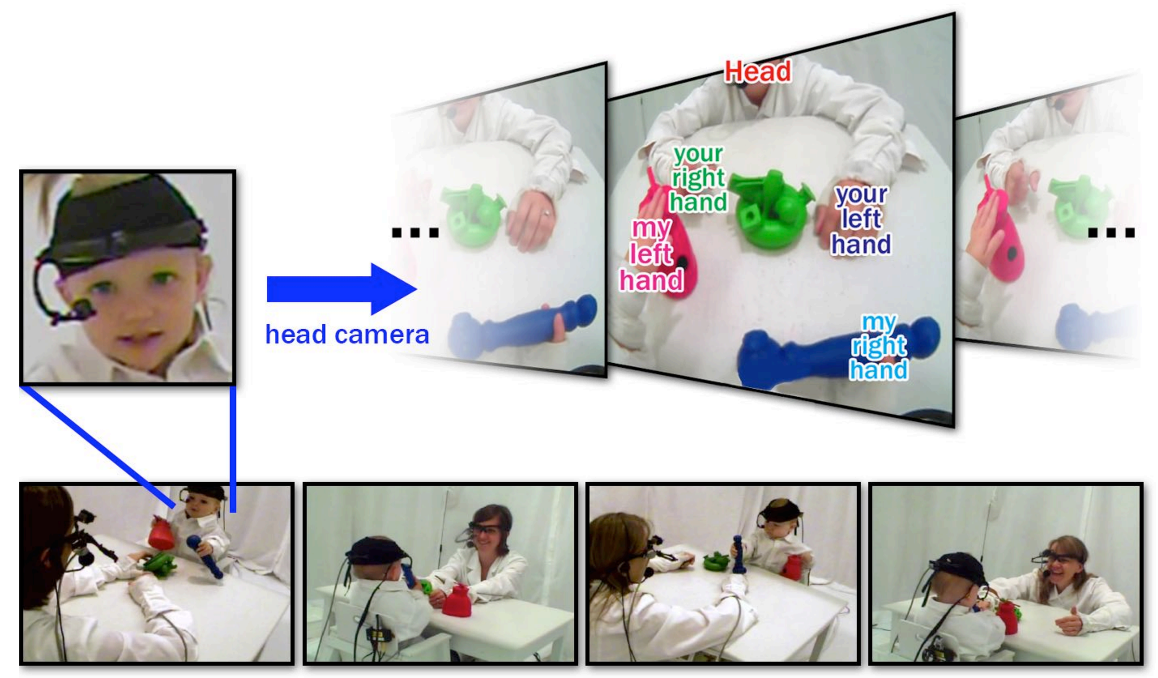
- We are interested in automatically **analyzing complex and dynamic interactions** from **first-person views**.
- To do this, we need to **robustly track hands** and **distinguish hand types** (my hands vs. your hands or left vs. right hands).
- We present two projects related to analyzing hands in first-person video. One considers "clean" video from lab settings, using weak (but fast) appearance models with **spatial constraints** of first-person views to distinguish hands. The second **detects, distinguishes and segments hands** in real-world interactions with **strong (deep) appearance models** that explicitly capture hand types.

Why Egocentric?
Wearable cameras are catching on, with many new consumer devices on the market. Hands appear often and prominently in first-person video, and their pose gives important cues about the camera wearer.

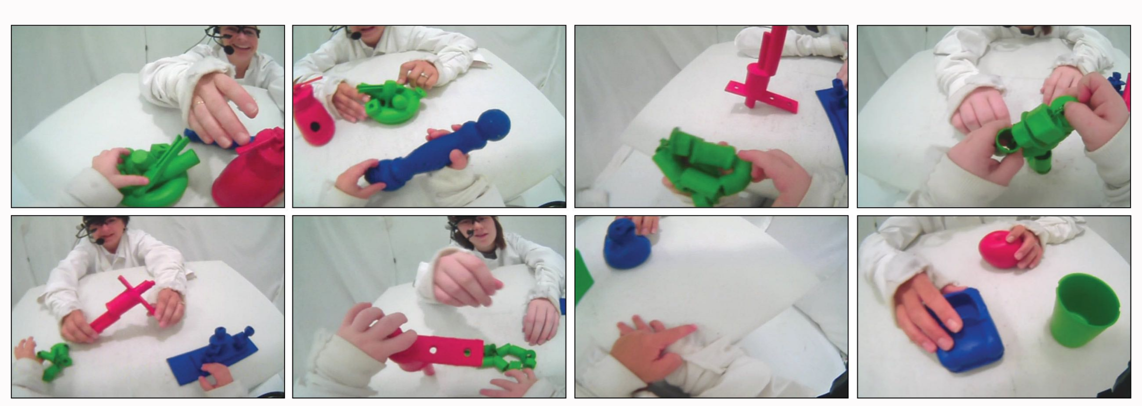
Lab-based Attention Project

1. Motivation

- We use head-mounted cameras to **study how toddlers interact with parents**, including how they coordinate hands and head turns.
- We need to **detect, disambiguate, and track all hands** in the toddler's view.
- We apply **probabilistic models of joint head and hand motion** in egocentric video.



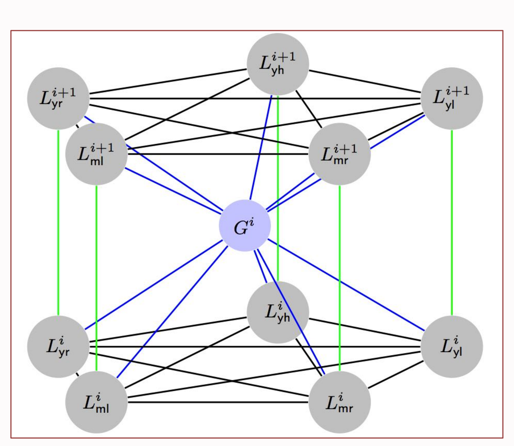
2. Challenges



Head motion makes the child's view 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 $\{L_p^i\}_{p \in P}^{1 \le i \le n}$, and **global shift** G^i between consecutive frames caused by head motion.
- Use weak **skin, head, arm appearance models** to generate (noisy) likelihood maps in each frame.
- **Model spatial constraints** on hand position with a fully-connected graphical model.
- **Model temporal constraints** with edges between parts in adjacent frames and global shift variables.
- **Model out-of-view parts** with a special state whose probability is integrated over off-frame spatial constraints.
- **Solve using Gibbs sampling.**

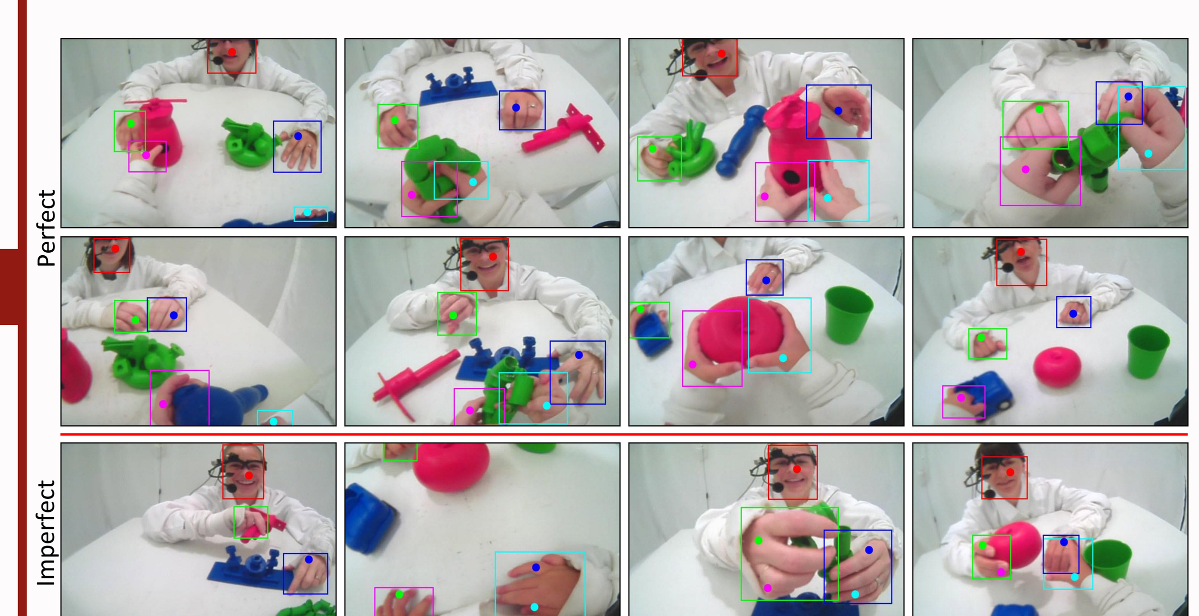


Graphical model for a 2-frame video.

4. Experiments

- We tested on **5 parent-child pairs** (31 min of video).
- We evaluated against **2,400 manually-annotated frames** (~1 frame/second).

5. Results



Results with estimated positions (dots) and ground truth boxes. Red: your head, blue/green: your left/right hand, magenta/cyan: my left/right hand.

Overall Accuracy	Observer		Partner				% Perfect Frames	Disambiguation Error Rate
	R. Hand	L. Hand	R. Hand	L. Hand	Head	Head (V-J)		
68.4	70.7	61.2	63.6	64.5	82.1	72.4	19.1	32.7

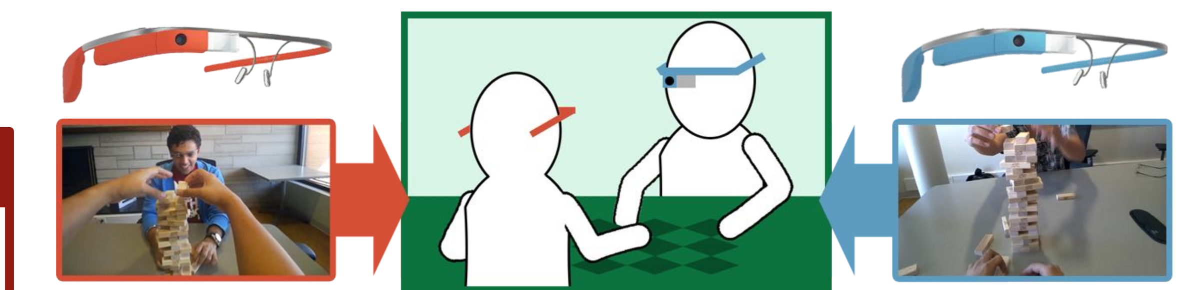
Top: Detection rates for hands and head (compared to Viola-Jones). Right: Various baselines.

Method	Overall Accuracy	% Perfect Frames	Disambiguation Error Rate
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**

See Full Papers for More!

- *This Hand Is My Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric Video*, **CVPR Workshops 2014.**
- *Detecting Hands in Children's Egocentric Views to Understand Embodied Attention during Social Interaction*, **CogSci 2014.**

Naturalistic Activities Project

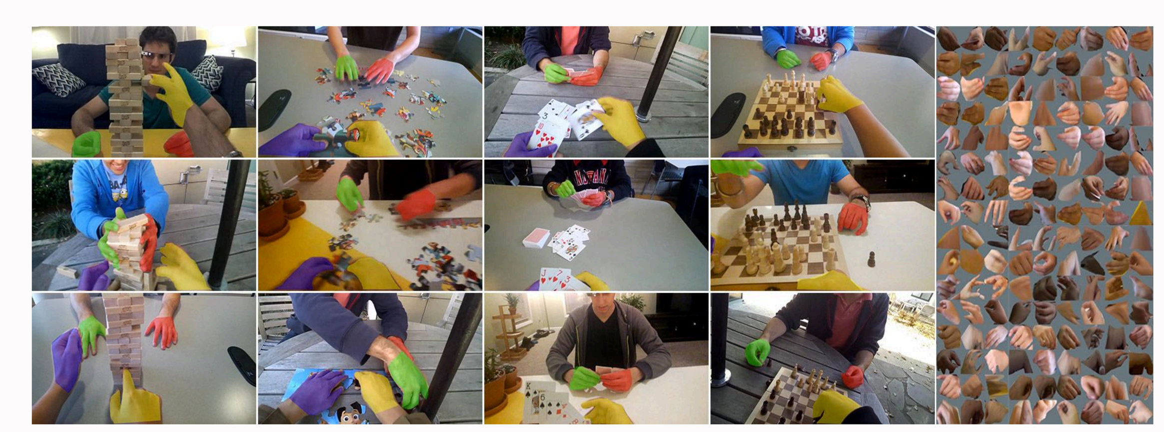


1. Motivation

- We study egocentric hand detection, identification, and segmentation of interacting people in **realistic settings**.
- Evaluate the potential of **deep hand appearance models** to detect different hand poses and types.
- Analyze how informative hand pose and location can be for **first-person activity recognition.**

2. Data Collection

- Recorded synchronized first-person video from interacting subjects, using two **Google Glasses.**
- Four different actors, four activities, at three locations, for 4x4x3 = **48 unique videos.**
- Annotated 4,800 random frames with **pixel-level ground truth for 15,053 hands.**

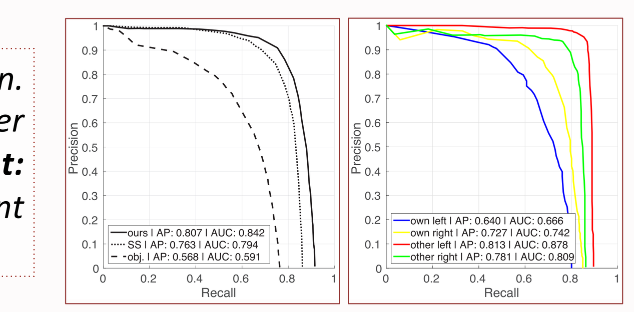


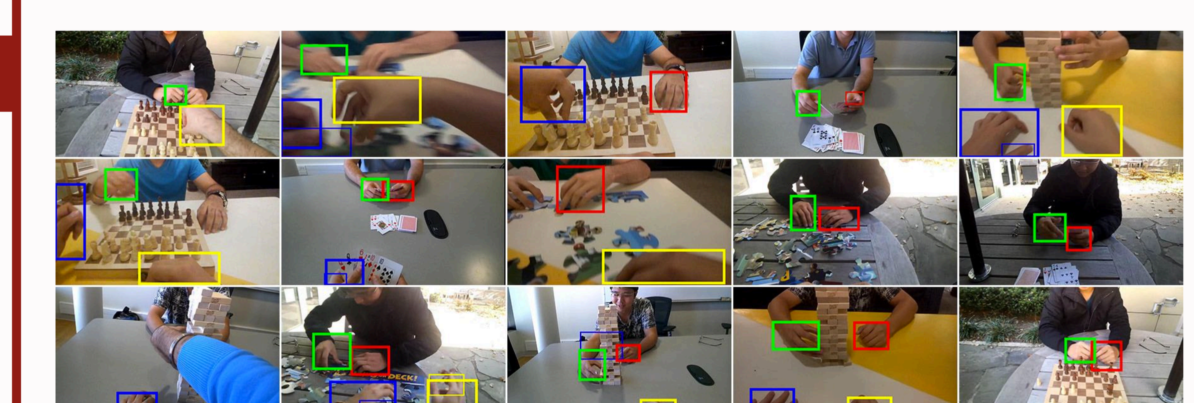
*Sample frames from our dataset. **Left**: Ground truth hand masks superimposed on sample frames, where colors indicate hand types. **Right:** Random subset of cropped hands according to ground truth segmentations.*

3. Hand Detection

- We apply **convolutional neural networks**, using a **lightweight region proposal** technique that samples based on skin color and spatial location.
- Our region proposals yielded **better coverage** than other methods like "selective search" or "objectness."
- CNN is trained for a 5-way classification task between own left hand, own right hand, other left hand, other right hand, and background.
- Different dataset splits show that performance generalizes across actors/activities/locations.

*Precision-recall for hand detection. **Left:** Results compared with other region-proposal methods. **Right:** Results for detecting four different hand types.*

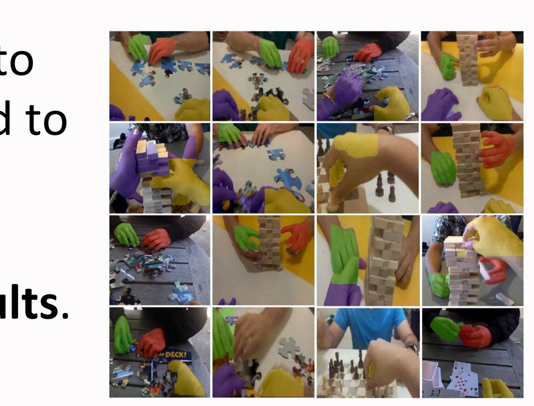




Random detection results of own left, own right, other left and other right.

4. Segmenting Hands

- Use our strong detections to initialize **GrabCut,** modified to use **local color models** for hands and background.
- Yields **state-of-the-art results.**

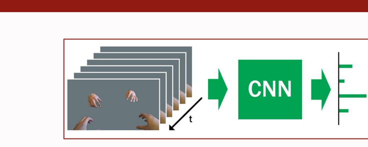


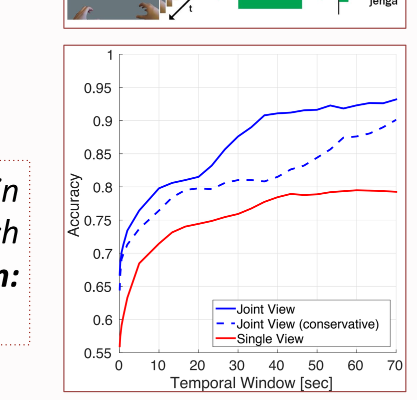
***Top:** Segmentation examples on random frames. **Bottom:** Intersection/union results.*

Method	Own Hands		Other Hands		Average
	Left	Right	Left	Right	
Li et al.	0.395	0.478	0.534	0.505	0.478
Ours	**0.515**	**0.579**	**0.560**	**0.569**	**0.556**

5. Activity Recognition

- Activities can be successfully estimated using hand pose and location alone.





***Top:** To predict activities based on hands, we train and test a CNN with video frames in which everything but hands is masked out. **Bottom:** Accuracy versus temporal window of video.*

More Information:

- Dataset will be published online: **vision.soic.indiana.edu/egohands**

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), Google, and the Indiana University Vice President for Research.*