Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions

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

INDIANA UNIVERSITY BLOOMINGTON

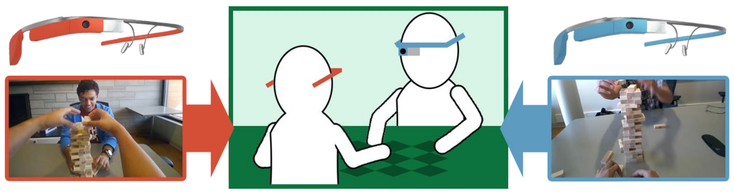
Introduction

- **Hands** appear very often in **egocentric video** and can give important cues about what people are doing and what they are paying attention to
- We aim to **robustly detect** and **segment hands** in dynamic first-person interactions, **distinguish hands** from another, and use hands as cues for **activity recognition**

Our contributions include:

1. Deep models for **hand detection/classification** in first person video, including **fast region proposals**
2. Pixelwise **hand segmentations** from detections
3. A quantitative analysis of the power of hand location and pose in **recognizing activities**
4. A **large dataset of egocentric interactions**, including fine-grained ground truth for hands

EgoHands: A Large-scale Hand Dataset

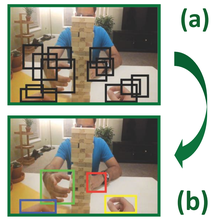


- We recorded first-person video from two interacting subjects, using two **Google Glass** devices
- 4 actors x 4 activities x 3 locations = **48 unique videos**
- Pixel-level ground truth annotations for **15,053 hands** in **4,800 frames** allow training of data-driven models

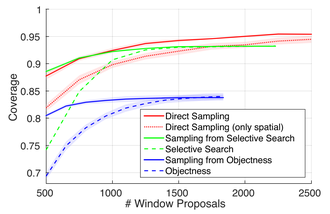
Left) Some sample frames from the data with ground truth hand masks superimposed in different colors (indicating different hand types). Each column shows one activity: *Jenga*, jigsaw puzzle, cards, and chess.
Right) Random samples of ground truth segmented hands.
The *EgoHands* dataset is publicly available: **http://vision.indiana.edu/egohands/**

Hand Type Detection

- Lightweight sampling approach proposes a set of image regions that are likely to contain hands **(a)**
- CNN classifies each region for the presence of a specific hand type **(b)**



(a) Generating Region Proposals Efficiently

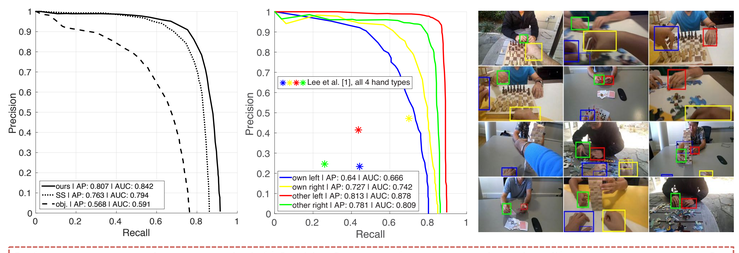


Hand coverage versus number of proposals per frame for various proposal methods.

- Regions are **sampled** from **learned distributions** over region size and position, biased by **skin color** probabilities
- Yields **higher coverage** than standard methods at a fraction of the computational cost

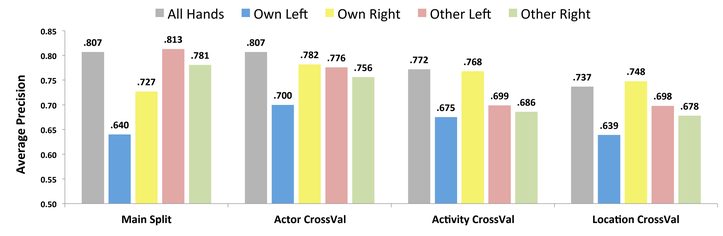
(b) Hand Type Classification

- A **CNN** is trained to distinguish regions between hand types and background (defined based on IoU overlap with ground truth), then non-max suppression produces detections



Performance on the main dataset split. ***Left)*** Precision-Recall plot for hand detection using different proposal methods. Our proposed approach outperforms the baselines. ***Center)*** Precision-Recall for hand detection and disambiguation with comparison to Lee et al. [1]. "Own hand" detection is more difficult due to extreme occlusions. ***Right)*** Some detection examples at 0.7 recall.

- We see **similar performance** and trends for **cross-validation** experiments with respect to actors, activities, and locations

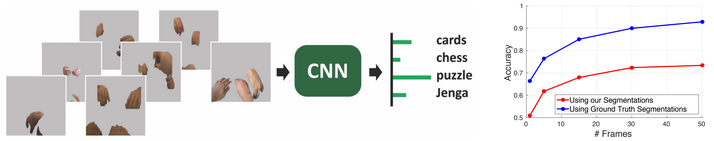


Hand Segmentation

- We use our strong detections to initialize **GrabCut**, modified to use **local color models** for hands and background, yielding state of the art results for hand segmentation (**Ours:** 0.556 average IoU, **Li et al. [2]:** 0.478 average IoU)

Hand-based Activity Recognition

- We hypothesize that first-person **hand pose** can reveal significant information about the **interaction**
- To test this, we build CNNs to classify activities based on frames with non-hands masked out:



Left) Examples of masked hand frames that are used as input to the CNN. ***Right)*** *Classification accuracy increases as more frames are considered. Here, we sample ~1 frame per second such that 10 frames span roughly 10 seconds of video.*

- With ground truth hand segmentations, **66.4% accuracy**
- With our hand segmentations, **50.9% accuracy**
- Aggregating frames **across time** futher increases accuracy

More activity recognition evaluation can be found in our *ACM ICMI 2015* paper:
http://vision.soic.indiana.edu/hand-interactions/

Conclusion and Future Work

- We showed how to accurately detect and distinguish hands in first person video and explored the potential of segmented hand poses as cues for activity recognition
- We plan to extend our activity recognition approach to **finer-grained actions** and more **diverse social interactions**

[1] S. Lee, S. Bambach, D. Crandall, J. Franchak, C. Yu, "This hand is my hand. A probabilitic approach to hand disambiguation in egocentric video", CVPR Workshops 2014
[2] C. Li, K. Kitani, "Model recommendation with virtual probes for egocentric hand detection", ICCV 2013

Acknowledgments: *This work was supported by the National Science Foundation (CAREER IIS-1253549, CNS-0521433), the National Institutes of Health (R01 HD074601, R21 EY017843), Google, and the Indiana University Office of the Vice President for Research through the Indiana University Collaborative Research Grant, and Faculty Research Support Program. It used compute facilities provided by NVidia, the Lilly Endowment through its support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. SB was supported by a Paul Purdom Fellowship.*