Analyzing Hands to Recognize Social Interactions with a Large-scale Egocentric Hands Dataset

Sven Bambach Stefan Lee David J. Crandall
School of Informatics and Computing
Indiana University
{sbambach,steflee,djcran}@indiana.edu

Chen Yu
Psychological and Brain Sciences
Indiana University
chenyu@indiana.edu

Abstract

Hands appear very often in egocentric video, and their appearance and pose give important cues about what people are doing and what they are paying attention to. But existing work in hand detection has made strong assumptions that work well only in simple scenarios, such as with limited interaction with other people. We develop methods to locate hands and distinguish between hand-types in egocentric video using an efficient bounding-box proposal method and strong appearance models based on Convolutional Neural Networks. We show how these high-quality bounding boxes can be used to create accurate pixelwise hand regions, and as an application, we investigate the extent to which hand pose segmentation alone can distinguish between different social interactions from an egocentric view. We evaluate these techniques on a new dataset of 48 first-person videos of people interacting in realistic environments, with pixel-level ground truth for over 15,000 hand instances..

1. Introduction

Hands are almost omnipresent in a person's field of view and the first-person perspective creates a very functional and embodied perspective of one's own hands. It is therefore perhaps somewhat surprising that there has been relatively little work on analyzing hands in the context of wearable, first-person cameras. In this extended abstract, we describe an approach for detecting and distinguishing hands in egocentric videos of interacting people, and also demonstrate that we can use the extracted hand poses (without any other cue) to recognize the kind of interaction (e.g. playing chess). This work was detailed in two recent publications in computer vision [2] and multimodal interaction [1] venues, but we believe it will be interesting to the HANDS community as well.

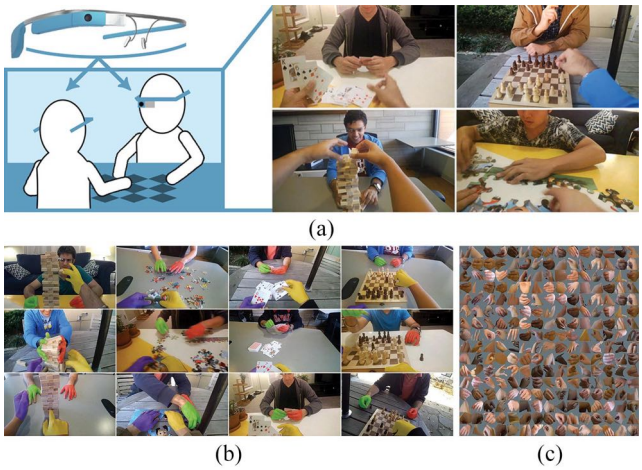


Figure 1: *Data collection and dataset.* (a) Using Google Glass, we collect 48 videos of different actors performing four different interactions at various locations. (b) We provide ground-truth hand(-type) segmentations of over 15,000 hands in 4,800 frames. (c) Random samples of ground-truth segmented hands.

2. The EgoHands Dataset

We start by presenting a new, large-scale dataset of first-person videos collected using Google Glass on pairs of interacting people. Many existing first-person video datasets (e.g. [3, 7]) are designed to recognize activities or handled objects, but do not include annotations of hands. The few existing datasets that do include hand annotations (e.g. [6]) are rather small, and contain no other people (i.e. only the observer's hands) in view. In contrast, our dataset was explicitly designed to capture naturalistic social interactions between people. This introduces the problem of *hand-type* classification and detection, where hands can be semantically distinguished not only between left and right, but also between the observer's hands and other hands in view.

We collected data from different pairs of four partici-

pants who sat facing each other while engaged in different interactions. We chose four activities that encourage interaction and hand motion: (1) playing cards; (2) playing chess, where for efficiency we encouraged participants to focus on speed rather than strategy; (3) solving a small jigsaw puzzle; and (4) playing Jenga. We also varied context by collecting videos in three different locations. The recording setup, as well as example frames for each of the four activities, are shown in Figure 1(a).

We systematically collected data from four actors performing all four activities at all three locations, resulting in $4 \times 4 \times 3 = 48$ unique combinations of videos. Each participant wore a Google Glass device, which recorded 720×1280 video at 30 Hz. In total, our dataset contains around 130,000 frames of video, of which 4,800 frames have pixel-level ground truth consisting of 15,053 hands, which we manually annotated. Figures 1(b) and 1(c) show different examples of ground-truth annotated hand-types.

To our knowledge, this is the largest dataset of hands in egocentric video, and we have released it online[1] to further encourage work in this domain.

3. Hand(-Type) Detection & Segmentation

Convolutional neural networks (CNNs) offer state-of-the-art performance for visual classification tasks [4]. We use CNNs as part of our hand-type detection framework to distinguish between four classes of hands: *own left*, *own right*, *other left* and *other right*. We first devise an efficient method to propose a set of bounding boxe candidates in a frame, and then evaluate each box with a CNN trained on hands, producing bounding box detections of hands. We take advantage of those high-quality detections to further segment hand poses.

Generating Proposals Efficiently with Spatial Sampling. Hands tend to have spatial biases with respect to where they appear in a first-person view [5]. We include these biases by sampling proposal windows accordingly. Our primary motivation is to model the probability that an object O appears in a region R of image I,

$$P(O|R,I) \propto P(I|R,O)P(R|O)P(O)$$

where $P(O)$ is the object occurrence probability, $P(R|O)$ is the prior distribution over the size, shape, and position of regions containing O, and $P(I|R,O)$ is an appearance model evaluated at R for O. Given a parameterization that allows for sampling, high quality regions can then be drawn from this distribution directly. We learn $P(O)$ and $P(R|O)$ from training data and use a simple (but fast) skin color heuristic to estimate $P(I|R,O)$.

We find that this technique generates higher coverage than standard window proposal techniques at a fraction of the computational costs.

[1] http://vision.soic.indiana.edu/egohands/

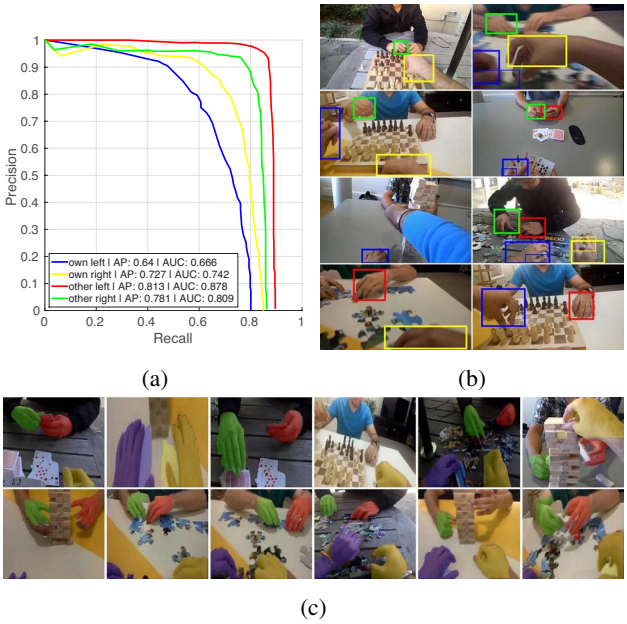


Figure 2: *Hand-type detection and segmentation results.* (a) Precision-recall curves for bounding box based detection of hands. (b) Randomly chosen detection examples. (c) Random segmentation examples.

Window Classification using CNNs. Given our accurate, efficient window proposal technique, we can now use a standard CNN classification framework to classify each proposal. We use CaffeNet from the Caffe software package, which is a slightly modified form of the well-known AlexNet [4] CNN architecture. The full detection pipeline consists of generating spatially sampled window proposals, classifying the window crops with the fine-tuned CNN, and performing per-class non-maximum suppression for each test frame.

Figure 2(a) shows the precision-recall curves for detecting the four types of hands while Figure 2(b) shows detected bounding boxes on a set of example frames. We overall achieve quite high detection scores (mean AUC = 0.723), with little confusion between hand-types.

Segmenting Hand Poses. We build upon our bounding-box detection both to focus segmentations on local image regions, and to provide semantic labels for the segments. We assume most pixels inside a box correspond with a hand, albeit with a significant number of background pixels caused both by detector error and because hands rarely fill a bounding rectangle. This assumption allows us to apply a variation of the well-known semi-supervised segmentation algorithm, GrabCut [8], to our problem. Some example segmentations of different hands are shown in Figure 2(c).

4. Recognizing Interactions via Hands

To explore if hand poses can uniquely identify interactions, we create masked frames in which all content except hands is replaced by a gray color. Some examples of masked frames are shown on in Figure 3(a). We train a CNN with the architecture of [4] on a four-way classification task by feeding it the masked hand frames from a training set of 24 videos. Each frame was labeled with one of the four activities (cards, chess, puzzle, Jenga). In this training phase, we used ground-truth hand segmentations to prevent the classifier from learning any visual bias not related to hands (e.g. portions of other objects that could be visible due to imperfect hand extraction).

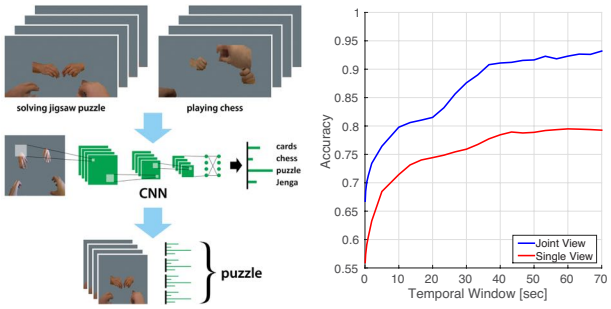
To test the performance of the trained CNN, we first apply the hand detection and segmentation approach from Section 3 to each frame of a test dataset (16 videos), resulting in $16 \times 2{,}700 = 43{,}200$ test frames. Classifying each frame individually gave 53.6% accuracy on the four-way classification problem, nearly twice the random baseline (25.0%). This promising result suggests a strong relationship between hand poses and interactions.

Temporal integration. We integrate evidence across time, given the evidence in individual frames across a time window from t_i to t_j, as

$$\hat{H}_p^{t_i,t_j} = \arg\max_{H \in \mathcal{H}} \prod_{k=t_i}^{t_j} P(H|F_k),$$

where $\mathcal{H}$ is the set of possible interactions and F_k is the observed frame at time k. The latter equation follows from assumptions that frames are conditionally independent given activity, that activities are equally likely *a priori*, and from Bayes' Law. We evaluated this approach by repeatedly testing classification performance on our videos over many different time windows of different lengths (different values of $|t_j - t_i|$). The red line in Figure 3(b) shows that accuracy increases with the number of frames considered. For instance, when observing 20 seconds of interacting hands from a single viewpoint, the system predicts the interaction with 74% accuracy.

Viewpoint integration. Since we captured synchronized videos from both participants of our interactions, we could also investigate if considering both viewpoints jointly further improved accuracy. We integrate the frame-based evidence across viewpoints as above, making the additional assumption that the viewpoints are independent conditioned on activity. We again test over many different temporal windows of different sizes in our videos, but now using frames from both viewpoints. The results are plotted in blue in Figure 3(b) and clearly outperform the single view approach, showing that the two views are indeed complementary.



(a) (b)

Figure 3: *Social interaction recognition.* (a) Workflow for hand-based interaction recognition. (b) Recognition accuracy using a sliding temporal window across each video.

5. Summary

We introduce a new, large-scale dataset of hands in egocentric interactions that allows training of data-driven recognition models such as CNNs. We also devise a CNN-based method to detect all hands in view and distinguish between semantically different types of hands. Finally, we demonstrate the expressive potential of hands in the first-person view by automatically recognizing the type of interaction based on hand-poses only.

Acknowledgments. This work was supported by NSF (CAREER IIS-1253549, CNS-0521433), NIH (R01 HD074601, R21 EY017843), Google, and IU OVPR through IUCRG and FRSP grants. It used compute facilities provided by NVidia, the Lilly Endowment through support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. SB was supported by a Paul Purdom Fellowship. We thank Benjamin Newman, Brenda Peters, Harsh Seth, and Tingyi Wanyan for helping with dataset collection.

References

[1] S. Bambach, D. J. Crandall, and C. Yu. Viewpoint integration for hand-based recognition of social interactions from a first-person view. In *ACM ICMI*, 2015. 1

[2] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 1

[3] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 1

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3

[5] S. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *CVPR Workshops*, 2014. 2

[6] C. Li and K. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, 2013. 1

[7] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPR Workshops*, 2009. 1

[8] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM TOG*, 2004. 2