Tracking Hands of Interacting People in Egocentric Video

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

John M. Franchak Chen Yu
Psychological and Brain Sciences
Indiana University
{jmfranch,chenyu}@indiana.edu

Abstract

Wearable cameras are becoming practical and inexpensive, creating new applications and opportunities including novel studies of social interaction and human development. Recent work has focused on identifying the camera wearer's hands in egocentric video, as a first step towards more complex analysis. Here we study how to disambiguate and track not only the observer's hands but also those of social partners. We present a probabilistic framework for modeling paired interactions that incorporates the spatial, temporal, and appearance constraints inherent in egocentric video. We test our approach on a dataset taken from a psychology study, consisting of 30 minutes of first-person video from five different child-parent subject pairs.

1. Introduction

Head-mounted cameras capture video that is fundamentally different from static cameras, recording an approximation of a person's field of view during everyday life. This technology is creating new applications in areas like lifelogging [13], healthcare [3], and security [15]. In addition, head-mounted cameras are being used for psychology research [1, 7] by recording fine-grained information about people's activities and interactions.

Hands are perhaps the most frequent objects in egocentric video, and are arguably also the most important, since they are the primary way that humans physically interact with the world. In fact, most work in egocentric activity recognition assumes that activities can be characterized by the in-hand manipulation of certain objects [4, 11]. Other work on egocentric hand detection is motivated by the idea that hands are important for understanding complex object manipulations, gestures, and motor skills [9, 10, 14].

Most of this work assumes that only the camera wearer's hands are visible in the scene, even though real-world egocentric video includes frequent interactions with other people [5]. Recognizing gestures, handled objects, and activities in practice will thus require distinguishing the camera owner's hands from others that occur in the scene. We are especially motivated by psychology experiments that use head-mounted cameras to study how young children and adults interact with one another, and how children coordinate their hands and gaze in order to manipulate objects [1, 6, 7, 12]. In our experiments, for example, a parent and child play with toys on a table and frequently point to, reach for, and exchange toys with their hands (Figure 2). The child's view is extremely dynamic: the hands of both the child and parent frequently disappear and reappear or are partly occluded. Manually labeling hand positions in these large-scale datasets is slow and tedious, so a main motivation of our work is to develop a technique that can perform the labeling automatically.

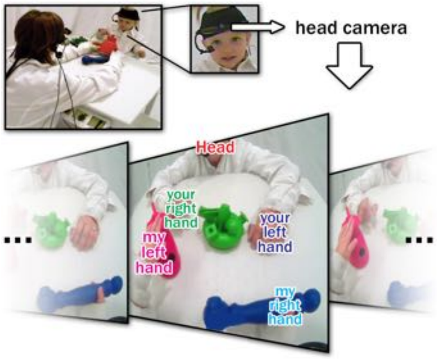

Figure 1. A mother and child play while both wear head-mounted cameras. To help investigate how different hands help capture the child's attention, we study hand tracking and disambiguation in egocentric video, and propose a probabilistic model that incorporates appearance, spatial, and temporal cues.

In this extended abstract, we describe an approach for detecting and distinguishing hands in egocentric videos of interacting people, and apply it to differentiating between the left and right hands of the child (the camera wearer), as well as the left and right hands of the parent (Figure 1) in each frame of a video. This work was detailed in two recent publications at egocentric vision [8] and cognitive science [2] venues, but we believe it may be interesting to the HANDS community as well.

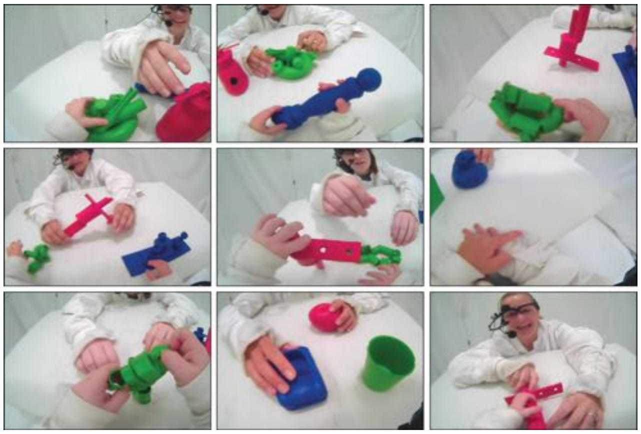

Figure 2. Some sample images from the egocentric videos that were used in our experiments. The child's view is very dynamic; hands come in and out of view or overlap very frequently.

2. Methodology

Idea. We propose a graphical model framework that encodes key spatiotemporal constraints inherent to the egocentric perspective. These constraints are quite intuitive; for example, given that we see the scene through the eyes of the child, we expect the child's left hand to enter the child's view predominantly from the lower left and the right hand to enter from the lower right. In addition to these absolute spatial assumptions, one can also use relative spatial statistics: we generally expect the child's left hand to be to the left of the right hand, and vice-versa. Similar assumptions can be made for the parent. Since the parent usually faces the egocentric observer, the parent's right hand usually occupies the left side of the child's field of view and the parent's left hand is on the right.

Formal Overview. We give the details of our formulation in [8], but sketch out the major details here. Given an egocentric video sequence with n frames, $I = \{I^1, ..., I^n\}$, of an interaction between two subjects, we would like to estimate the locations of the observer's hands (my left, my right), the other person's hands (your left, your right), and the other person's head (your head) throughout the video. We encode these parts $P = \{ml, mr, yl, yr, yh\}$ in each frame as latent variables $\{L_p^i\}_{p \in P}^{1 \leq i \leq n}$. We also explicitly model the global shift (caused by head motion) between to frames as the latent variables G^i.

A graphical depiction of our model for a two frame video example is shown in Figure 3. We model spatial constraints between hand positions with a fully-connected graph within each frame (black edges). We further model temporal smoothness with (green) edges between corresponding parts in adjacent frames. Blue edges that connect parts in adjacent frames through the global motion variable G^i allow for more rapid, global motion between frames.

We use weak (but fast) skin, head and arm appearance models to generate noisy likelihood maps for each part within each frame. All spatial constraints are learned from our training data in the form of absolute and relative spatial "priors." We model these as isotropic normal distributions for efficient inference. We also take advantage of this formulation to explicitly 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. To solve the tracking problem, we perform inference by approximately maximizing $P(L, G|I)$ for the whole video using Gibbs sampling.

3. Results

To evaluate our approach, we manually annotated 2,400 randomly sampled frames with bounding boxes from 30 minutes of video and five different child-parent pairs. Depending on which body parts are in view, each frame has up to five bounding boxes: two child hands, two parent hands, and one parent face.

For each frame, our system predicts the location of each of the five body parts, by either providing a coordinate or indicating that it is outside the frame. Figure 4 shows some sample frames, where rectangles depict the ground-truth bounding boxes, and dots mark our predicted position. Part identities are represented by color, so that dots inside boxes of the same color indicate correct estimates. The first two rows show perfect frames, while the last row shows errors.

We also quantitatively evaluate our system using three different metrics: The overall prediction accuracy, the percentage of frames in which all predictions were correct, and the disambiguation error rate. The last metric captures the percentage of hands for which we fail to disambiguate between the observer's hands and the partner's hands.

Table 1 summarizes our results and compares them to

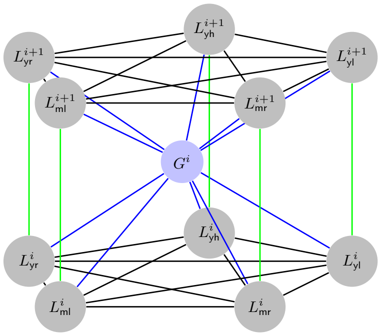

Figure 3. *Graphical depiction of our model for 2 frames,* where the bottom 5 nodes represent the locations of the head and hands in one frame, and the top 5 nodes represent the locations in the next frame. **Between-frame links** enforce temporal smoothness, **shift links** model global shifts in the field of view, and **in-frame links** constrain the spatial configuration of the body parts.

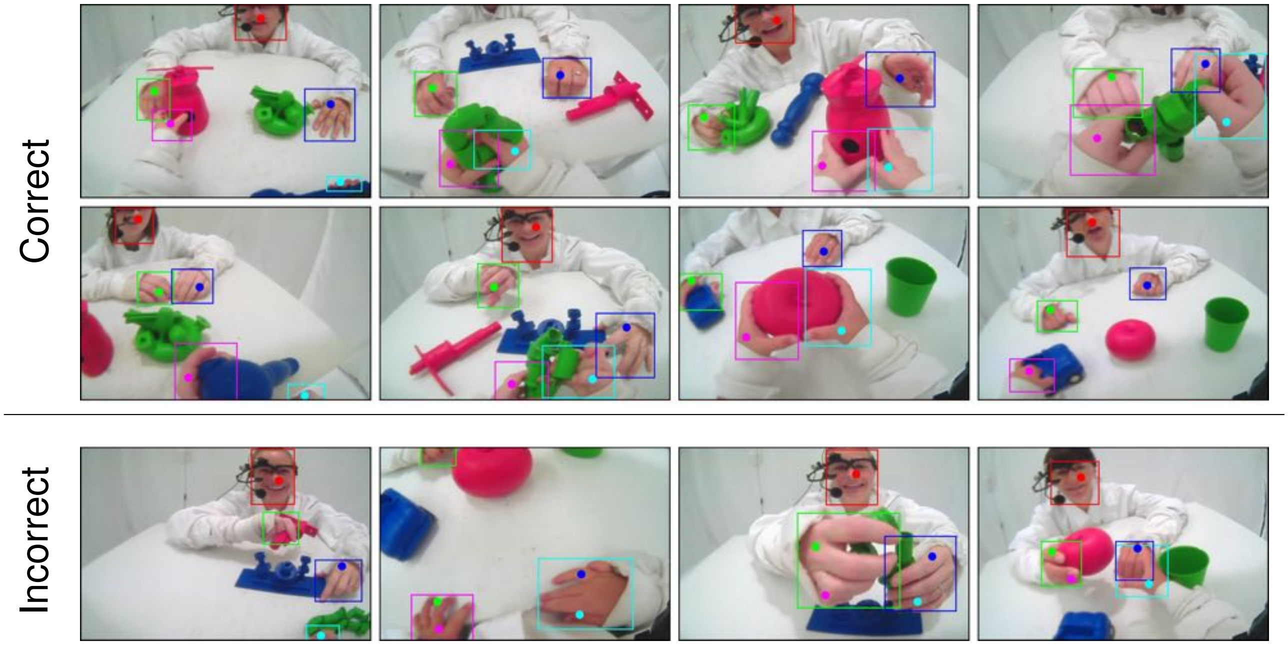

Figure 4. *Sample frames from our results,* with rectangles showing ground truth bounding boxes and dots showing predicted part positions (red = your head, blue = your left hand, green = your right hand, magenta = my left hand, cyan = my right hand). The first two rows show our robustness with respect to partial occlusions and changes in hand configurations, while the bottom row shows failure cases.

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**

Table 1. Comparison of our model's results to baselines, in terms of overall accuracy, percentage of perfect frames, and hand disambiguation error rate. For more details, please refer to [8].

baselines of increasing complexity. Our full model outperforms all the baseline methods by more then 10 percentage points for accuracy and also performs best in terms of perfect frames and hand disambiguation error.

References

[1] S. Bambach, D. Crandall, and C. Yu. Understanding embodied visual attention in child-parent interaction. In *ICDL-EPIROB*, 2013. 1

[2] S. Bambach, J. M. Franchak, D. J. Crandall, and C. Yu. Detecting hands in childrens egocentric views to understand embodied attention during social interaction. *CogSci*, pages 134–139, 2014. 1

[3] A. Doherty, S. Hodges, A. King, A. Smeaton, E. Berry, C. Moulin, A. Lindley, P. Kelly, and C. Foster. Wearable cameras in health: The state of the art and future possibilities. *Am J Prev Med*, 44, 2013. 1

[4] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *CVPR*, 2011. 1

[5] A. Fathi, J. Hodgins, and J. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 1

[6] J. Franchak, K. Kretch, K. Soska, and K. Adolph. Head-mounted eye tracking: A new method to describe infant looking. *Child development*, 82(6):1738–1750, 2011. 1

[7] M. C. Frank. Measuring children's visual access to social information using face detection. In *CogSci*, 2012. 1

[8] S. Lee, S. Bambach, D. Crandall, J. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *CVPR Workshops*, 2014. 1, 2, 3

[9] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *ICCV*, 2013. 1

[10] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *CVPR*, 2013. 1

[11] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1

[12] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, and et al. Decoding children's social behavior. In *CVPR*, 2013. 1

[13] A. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood. Do life-logging technologies support memory for the past? An experimental study using SenseCam. In *CHI*, 2007. 1

[14] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara. Hand segmentation for gesture recognition in egovision. In *Workshop on Interactive Multimedia on Mobile & Portable Devices*, 2013. 1

[15] R. Stross. Wearing a badge, and a video camera. *The New York Times*, April 6 2013. 1