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

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

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

Abstract

Egocentric cameras are becoming more popular, introducing increasing volumes of video in which the biases and framing of traditional photography are replaced with those of natural viewing tendencies. This paradigm enables new applications, including novel studies of social interaction and human development. Recent work has focused on identifying the camera wearer's hands as a first step towards more complex analysis. In this paper, 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 of over 30 minutes of video from six pairs of subjects.

1. Introduction

Head-mounted cameras capture video that is fundamentally different from hand-held cameras, recording an approximation of a person's field of view during everyday life. This technology is creating new applications in areas like life-logging [23], healthcare [5], and security [27]. In addition, head-mounted cameras are being used for psychology research [1, 12] by recording fine-grained information about people's activities and interactions. However, these applications create huge amounts of video, so automated techniques are needed to process and understand them.

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 [7, 19]. Other work on egocentric hand detection is motivated by the idea that hands are important for understanding complex object manipulations, gestures, and motor skills [16, 17, 24, 32].

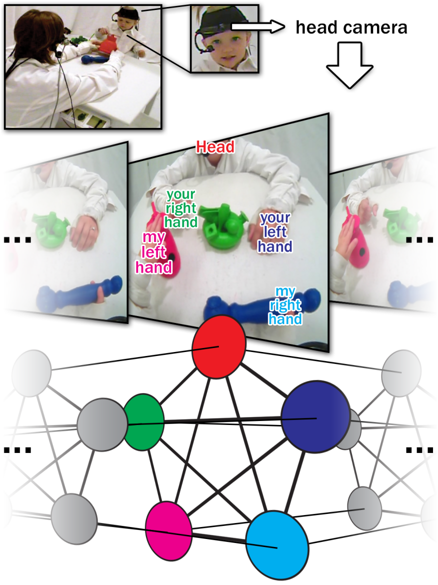

Figure 1. *Overview of our paper.* A mother and child play while both wear head-mounted cameras. To help investigate how different hands help capture the child's attention, we study hand tracking and disambiguation in egocentric video, and propose a probabilistic model that incorporates appearance, spatial,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

This existing hand detection 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 [8]. 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 recent psychology experiments that use head-mounted cameras to study how young children and adults interact with one another, and how children coordinate their hands, head turns, and gaze in order to manipulate objects [1, 11, 12, 20]. In these experiments, 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.

In particular, we would like to detect and distinguish 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). Due to the hands' dynamic appearance, frequent occlusions, rapid and unpredictable camera motion (from head turns), and frequent entering and exiting of the frame, disambiguating based on appearance alone is difficult. Distinguishing one's own hands from others is in itself an interesting cognitive challenge, and how humans accomplish this rapidly even in complex social interactions is unknown.

To overcome these challenges and accurately detect and disambiguate hands, 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 social partner (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.

We evaluate our framework on a collection of 20 parent-child interaction videos constituting over 31 minutes of video and thousands of labeled frames. Results show the unified framework outperforms sensible baselines and achieves about 70% performance overall (compared to about a 17% baseline). We also tested a more consumer-oriented application, with videos of interacting adults in a naturalistic environment taken by Google Glass, and achieve about 51% accuracy (versus about 15% baseline).

2. Related Work

Egocentric video is becoming a popular research topic in computer vision, with papers recognizing activities [7, 8, 19, 22], events [15, 18, 25], and objects [9, 21]. Detecting and tracking hands of the camera wearer has received special attention. Ren and Gu [21] pose this as a figure-ground segmentation problem, analyzing dense optical flow to partition frames into hands (or held objects) with irregular flow patterns, and background with coherent flow. Fathi *et al.* [9] segment between hand and object areas based on color features. Since the primary goal of these papers is to recognize held objects, they assume a static and rigid scene (where optical motion in the background can only be caused by head movements). In contrast, our videos include a second person who moves independently from the camera, such that motion features alone cannot isolate hands.

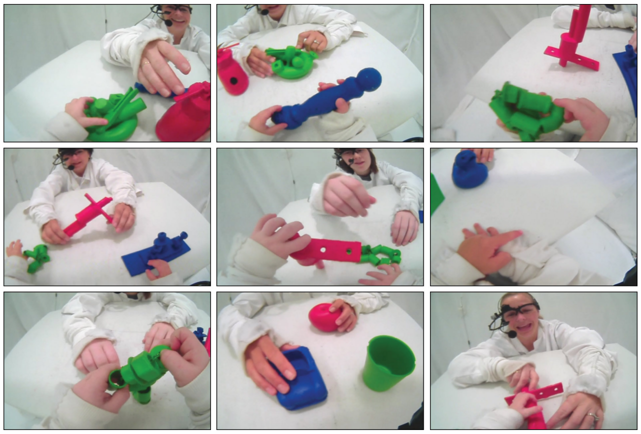

Figure 2. Some sample images from the egocentric videos that were used in our experiments. The child's view is very dynamic; hands come in and out of view or overlap very frequently.

Other work relies on color and shape features. Zariffa and Popvic [32] perform hand segmentation in egocentric video using pixelwise skin classifiers followed by shape-based post-processing, while Serra *et al.* [24] find skin superpixels using random forests. Li and Kitani [17] study changing illumination in hand segmentation, and find that a combination of color, texture and gradient features performs best. Follow-up work by the same authors [16] proposes model recommendation to pick the best detector for each environment using scene-level features. Our videos have relatively controlled lighting so we use a simple color-based skin detector, although our framework can easily incorporate more advanced techniques.

Many papers have studied hand tracking in static cameras [6]. Perhaps the papers most related to ours pose hand tracking in probabilistic frameworks to model constraints between frames of a video [26, 31], or between parts of the hand within a single frame [13]. Sudderth *et al.* [28] track hand poses using a Markov Random Field that models both temporal constraints and spatial constraints. Our approach is similar in that we use graphical models that combine evidence within and across frames, but the details are quite different: we study a different problem (hand disambiguation) with a completely different graphical model and a different inference technique (sampling instead of BP), and we explicitly address the challenges of egocentric video by modeling head motion (instead of assuming static cameras).

The above work either detects one or two hands in video from a static camera, or detects the hands of the owner of a first-person camera. Most of this work does not track or disambiguate the hands. In contrast, we study hand tracking of interacting people captured from head-mounted egocentric video, in which we must disambiguate between up to four moving hands that are regularly entering and exiting the frame, with extreme, erratic camera motion.

3. Modeling Egocentric Interactions

Given an egocentric video (from a head-mounted camera) of the interaction between two people, we would like to estimate the locations of the observer's hands, the other person's hands, and the other person's head throughout the video. This task is difficult because these parts frequently enter and leave the frame, and there is erratic camera motion caused by head motion of the person wearing the camera.

More formally, given a video sequence of n frames, each with $r \times c$ pixels, our goal is to estimate the position of each of a set of parts $\mathcal{P}$ in each frame. In this paper, we consider five parts in particular, $\mathcal{P} = \{\mathsf{yh}, \mathsf{yl}, \mathsf{yr}, \mathsf{ml}, \mathsf{mr}\}$, corresponding to the other person's head, hands ('your left/right') and the camera wearer's hands ('my left/right'), respectively. We denote the latent 2-D position of part $p \in \mathcal{P}$ in frame i as L_p^i and define L^i to be the full configuration of parts within the frame, $L^i = \{L_p^i\}_{p \in \mathcal{P}}$. Because of the dynamic nature of egocentric video, hands often enter and exit the frame, due both to motion of the hands and motion of the head-mounted camera. To address the possible absence of any given part, we augment the domain of L_p^i with an additional state $\emptyset$ indicating that the part is not visible in the frame, i.e. $L_p^i \in \{\emptyset\} \cup ([1, r] \times [1, c])$.

In addition to part position, we also explicitly model global motion caused by head movements by introducing random variables $G = (G^1, \dots, G^{n-1})$, where G^i is an estimate of the two dimensional global coordinate shift between frame i and frame $i + 1$. In this way, we assume the world has uniform depth such that a change of viewing angle would have the same effect on all points in the 2-D projection of the environment. This assumption simplifies the model and is reasonable given that the distances involved in a paired interaction are relatively constrained.

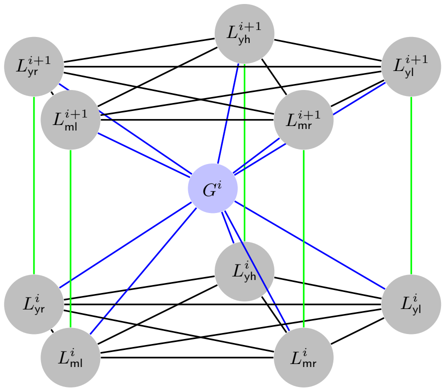

Figure 3. *Graphical depiction of our model for 2 frames,* where the bottom 5 nodes represent the locations of the head and hands in one frame, and the top 5 nodes represent the locations in the next frame. **Between-frame links** enforce temporal smoothness, **shift links** model global shifts in the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

We use a graphical model framework to model and estimate the locations of parts across all video frames jointly. The model uses head and hand appearance models to help identify parts within an individual frame. It also incorporates two types of constraints: (1) intra-frame constraints on spatial relationships between body parts, and (2) inter-frame constraints between body parts, which enforce temporal smoothness on part positions. A visualization of the five-part model for a two-frame video is presented in Figure 3. The connections within a frame (in black) form a complete graph over the five part nodes and capture the pairwise correlations between spatial locations of the parts. The green edges between each part and its corresponding variable in the next frame enforce the temporal smoothness constraint. Finally, the global shift variable is influenced by all pairs of corresponding parts such that a similar motion in all part pairs is likely to indicate a global shift. Placing these (soft) constraints into an undirected graphical model yields a joint distribution over all the latent variables $L = (L^1, ..., L^n)$ and G, conditioned on the image sequence $I = (I^1, ..., I^n)$,

$$P(L, G|I) \propto \prod_{i=1}^{n} \Bigg[P(I^i, I^{i+1}|G^i) \prod_{(p,q) \in \mathcal{E}} P(L_p^i|L_q^i) \prod_{p \in \mathcal{P}} P(I^i|L_p^i) P(L_p^{i+1}|L_p^i, G^i) P(L_p^i) \Bigg] \quad (1)$$

where $I = (I^1, ..., I^n)$ is the image sequence and $\mathcal{E} \subset \mathcal{P}^2$ is the set of undirected edges in the complete graph over $\mathcal{P}$.

We can solve the part-tracking problem for an entire video I by maximizing equation (1). Unfortunately, finding the global maximum is intractable. We thus settle for approximate inference using Gibbs sampling. As we discuss in Section 3.4, this avoids the need to compute or store the full joint distribution because the sampling involves only small neighborhoods of the graph. We now describe the components of the model in more detail.

3.1. Pairwise Spatial Priors

In egocentric videos of interacting people, body parts tend to have a distinct spatial relationship: the camera owner's hands are closer and thus lower in the frame, the other person's hands are further away and higher in the frame, and the head tends to be in between and above the hands. As is common in part-based models [4], we assume the relative spatial distribution between each pair of parts is Gaussian. The main complication in our problem is that we need to model explicitly the possibility of a body part being out of the field of view. To show how to do this, we start by considering a 2-D isotropic Gaussian function,

$$f_{\mu,\Sigma}(x, y) = \mathcal{N}(x; \mu_1, \Sigma_{11})\mathcal{N}(y; \mu_2, \Sigma_{22}),$$

parameterized by $\mu = [\mu_1 \; \mu_2]^T$ and $\Sigma = \text{diag}(\Sigma_{11}, \Sigma_{22})$. If this function represents a probability distribution over the location of a given part, then calculating the probability that the part is 'out' of a frame is equal to one minus the probability of being within the frame, $1 - F_{\mu,\Sigma}([1,c],[1,r])$, with

$$\begin{aligned} F_{\mu,\Sigma}\left([x_1,x_2],[y_1,y_2]\right) &= \int_{x_1}^{x_2}\int_{y_1}^{y_2} f_{\mu,\Sigma}(x,y)\,dy\,dx \qquad (2) \\ &= [\Phi(x_2;\mu_1,\Sigma_{11}) - \Phi(x_1;\mu_1,\Sigma_{11})] * \\ &\quad [\Phi(y_2;\mu_2,\Sigma_{22}) - \Phi(y_1;\mu_2,\Sigma_{22})]\,, \end{aligned}$$

where $\Phi(\cdot)$ is the normal cumulative density function and can be precomputed for efficient computation.

In-Frame Conditionals. Consider a pair of parts $p, q \in \mathcal{P}$ $(p \neq q)$ in frame i having positions L_p^i and L_q^i, respectively. Based on training data, suppose we have an estimate of the relative spatial relationship between these parts such that $L_p^i - L_q^i \sim \mathcal{N}(\mu_{qp}, \Sigma_{qp})$ for diagonal Σ_{qp}. We define the conditional probability distribution between L_p^i and L_q^i as,

$$P(L_p^i|L_q^i) = \begin{cases} \beta : L_q^i = \emptyset \\ 1 - F_{\mu_{qp}+L_q^i,\Sigma_{qp}}([1,c], \\ \qquad [1,r]) : L_p^i = \emptyset, L_q^i \neq \emptyset \\ F_{\mu_{qp}+L_q^i,\Sigma_{qp}}([L_{p,x}^i, L_{p,x}^i + 1], \\ \qquad [L_{p,y}^i, L_{p,y}^i + 1]) : L_p^i, L_q^i \neq \emptyset, \end{cases}$$

where β is a constant. Intuitively, this means that if part q is outside the frame, then it does not constrain part p's location (the conditional probability distribution is uniform), whereas if q is inside the frame, then p is either outside (and the conditional probability is given by one minus the probability of being inside the frame), or it is inside the frame with a probability given by a Gaussian distribution. If it were not for state $\emptyset$, our in-frame model would be similar to models from part-based recognition [4, 10].

Between-Frame Conditionals. The inter-frame conditionals impose temporal smoothness on part locations, connecting together part p's location L_p^{i+1} in frame $i{+}1$, its location L_p^i in frame i, and the latent global shift G^i between frames i and $i+1$ (due to head motion). We assume that if the part is within the image in both frames i and $i+1$, then L_p^i and L_p^{i+1} are related by a Gaussian distribution with diagonal Σ_p around the location predicted by the global shift, $L_p^{i+1} - L_p^i \sim \mathcal{N}(G^i, \Sigma_p)$. Including the possibility of parts entering or leaving the frame, the full conditional probability is similar to the above in-frame distribution,

$$P(L_p^{i+1}, G^i|L_p^i) = \begin{cases} \alpha : L_p^i, L_p^{i+1} = \emptyset \\ \frac{1-\alpha}{rc} : L_p^i = \emptyset, L_p^{i+1} \neq \emptyset \\ 1 - F_{\mu_i,\Sigma_p}([1,c],[1,r]) : L_p^i \neq \emptyset, L_p^{i+1} = \emptyset \\ F_{\mu_i,\Sigma_p}([L_{p,x}^{i+1}, L_{p,x}^{i+1}+1], \\ \qquad [L_{p,y}^{i+1}, L_{p,y}^{i+1}+1]) : L_p^i, L_p^{i+1} \neq \emptyset, \end{cases}$$

where $\mu^i = L_p^i + G^i$ and α is a constant. This conditional encodes the intuition that if a part is outside the image in one frame, it is outside the next frame with probability α or is uniformly distributed at a pixel in the frame; on the other hand, if a part is in the image in one frame, its probability distribution over pixels in the next frame is Gaussian, or it is outside the frame with probability one minus the integral over all pixel locations. This formulation encourages parts to stay at roughly the same position from one frame to the next, but allows for large jumps due to global motion if the jump is observed for many of the parts.

3.2. Absolute Spatial Priors

It is also helpful to add absolute spatial priors to encode implicit geometric biases of the egocentric viewpoint, like the fact that humans' tendency to shift their gaze to important objects in a scene means that faces are often in the upper-center part of the view. We model these biases as Gaussian distributions on absolute part location,

$$P(L_p^i) = \begin{cases} 1 - F_{\mu_{pp},\Sigma_{pp}}([1,c],[1,r]) : L_p^i = \emptyset \\ F_{\mu_{pp},\Sigma_{pp}}([L_{p,x}^i, L_{p,x}^i + 1],[L_{p,y}^i, L_{p,y}^i + 1]) : L_p^i \neq \emptyset, \end{cases}$$

with mean absolute position μ_{pp} and diagonal covariance matrix Σ_{pp} for each part p.

3.3. Full Conditionals

We use Gibbs sampling to perform inference, as described in the next section. To do this, we need to sample each random variable from its full conditional. Fortunately, because of the independence assumptions of our model, the full conditionals can be written and computed easily.

Part Nodes. We begin by deriving the conditional distribution of a part node given the rest of the variables in the graph. From Equation (1), we can compute the full conditional up to a proportionality constant,

$$\begin{aligned} P(L_p^i|G,L,I) &\propto P(L_p^{i+1}, G^i|L_{\mathsf{yl}}^i) P(L_p^{i-1}, G^{i-1}|L_p^i) \\ &\quad * P(I^i|L_p^i) P(L_p^i) \prod_{q \in \mathcal{P}-\{p\}} P(L_q^i|L_p^i), \qquad (3) \end{aligned}$$

where $P(I^i|L_p^i)$ is produced by an appearance model for p, which we define in Section 4. For instance, in the 5-part model we study here, the Markov blanket for L_{yl}^i is shown in the left panel of Figure 4, and the conditional for L_{yl}^i is

$$\begin{aligned} P(L_{\mathsf{yl}}^i|G,L,I) &\propto P(L_{\mathsf{yl}}^{i+1}, G^i|L_{\mathsf{yl}}^i) P(L_{\mathsf{yl}}^{i-1}, G^{i-1}|L_{\mathsf{yl}}^i) \\ &\quad * P(L_{\mathsf{yh}}^i|L_{\mathsf{yl}}^i) P(L_{\mathsf{yr}}^i|L_{\mathsf{yl}}^i) P(L_{\mathsf{ml}}^i|L_{\mathsf{yl}}^i) \\ &\quad * P(L_{\mathsf{mr}}^i|L_{\mathsf{yl}}^i) P(I^i|L_{\mathsf{yl}}^i) P(L_{\mathsf{yl}}^i). \qquad (4) \end{aligned}$$

Since the state space is discrete, the normalization constant is not needed for sampling, as it can be computed at runtime.

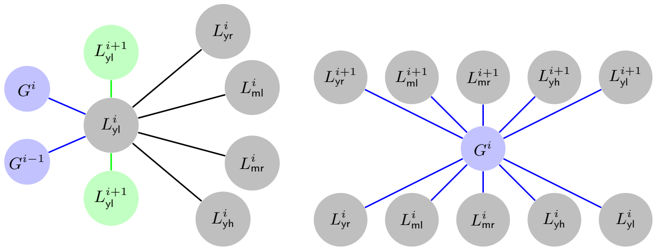

Figure 4. *Components of the full conditional* in our 5-part case, for *(left)* part node L^i_{yl}, and *(right)* shift node G^i.

Shift Nodes. The full conditional of a shift node G^i can also be written as a product of its neighbors in the graph,

$$P(G^i|G,L,I) \propto P(I^i,I^{i+1}|G^i) \prod_{p\in\mathcal{P}} P(L_p^{i+1},G^i|L_p^i), \quad (5)$$

which in the 5-part case we consider here (and as illustrated in the right panel of Figure 4) becomes

$$\begin{aligned} P(G^i|G,L,I) \propto\ & P(I^i,I^{i+1}|G^i)P(L^{i+1}_{\mathsf{yh}},G^i|L^i_{\mathsf{yh}}) \\ &* P(L^{i+1}_{\mathsf{yl}},G^i|L^i_{\mathsf{yl}})P(L^{i+1}_{\mathsf{yr}},G^i|L^i_{\mathsf{yr}}) \\ &* P(L^{i+1}_{\mathsf{mr}},G^i|L^i_{\mathsf{mr}})P(L^{i+1}_{\mathsf{ml}},G^i|L^i_{\mathsf{ml}}). \quad (6) \end{aligned}$$

This product has several intuitive properties. If there is disagreement between the relative movements of parts, then the overall distribution is diffuse and the likelihood term dominates. If parts are in agreement, there is a high peak. We require that the domain of G^i be a finite subset of the reals so we can sample without the normalization constant, avoiding a costly integral. The domain could be further constrained for specific applications (e.g. based on expected rate of movement). The likelihood $P(I^i,I^{i+1}|G^i)$ could also be application-dependent; here we use a normal distribution fit to the optical flow between frames i and $i+1$ [29].

3.4. Inference

We use Gibbs sampling [2] to perform inference on our model. Gibbs is a Markov-Chain Monte-Carlo method that generates samples from a full joint distribution over multiple random variables without a representation of the distribution (parametric or otherwise). In the limit, these samples form an accurate representation of the true distribution. We obtain a solution from these samples as follows. If for any given frame, the majority of samples for a given part are in the $\emptyset$ state, we label the part as "out" of the frame. Otherwise, we take the median position over the in-frame samples. In our experiments, just 50 samples provided good solutions.

4. Specializing to Child-Parent Joint Activity

Our hand-tracking approach could in principle be applied to any egocentric video data, with the various parameters and distributions set to customize it to a specific application. As mentioned in Section 3.3, one can apply any object model to generate the distributions for the per-part image likelihood terms $P(I^i|L^i_p)$ for each part location p in frame i. Since our context is rather controlled, we use features to demonstrate the effectiveness of our methodology. In particular, we first detect skin pixel regions and use a distribution of $P(I^i|L^i_p)$ that is zero unless L^i_p is on a skin pixel, and otherwise is proportional to the likelihood that an image patch around L^i_p 'looks like' part p, described below.

Skin Model. As our data only contains indoor footage with controlled lighting, we found that a color-based approach was sufficient for pixel-level skin detection [14]. To suppress occasional false detections around the red (skin-like) toys, we tuned our skin classifier for each individual subject. A human labeled the skin regions of 20 random frames from each subject's video. We then learn non-parametric skin and background models in YUV color space (discarding the luminance plane Y). To detect skin in unlabeled images, we compute the log odds of each pixel under these models as,

$$\log \frac{P(U,V|skin)}{P(U,V|background)},$$

and threshold the output value to create a binary skin mask. We then apply a median filter to suppress noise.

Face Model. We used the Viola & Jones [30] face detector to compute the face likelihood distribution, $P(I^i|L^i_{\mathsf{yh}})$. We used a simple formulation in which pixels inside a detected face box are assigned high likelihoods and pixels outside are assigned a low (non-zero) likelihood. We trained the detector on a small set of hand-labeled faces from our data.

'Your Hand' Model. Distinguishing hands is difficult due to their appearance variance caused by hands' deformable nature as well as scale and lighting. We implemented a simple model that takes advantage of our lab setting in which the edge density of the other person's sleeves is high compared to the background. We apply an edge detector to each image, blur the output, apply a threshold to detect arm regions, and find skin patches that are adjacent to these regions. Suppose a skin patch and arm region intersect at a point u. We calculate the longest possible straight line through u intersecting the set of candidate arm pixels (i.e. the diameter of the arm pixel region through u). The direction and length of this line are a measure of the arm direction and length, so we use them to set the 'your hand' likelihoods, $P(I^i|L^i_{\mathsf{yl}})$ and $P(I^i|L^i_{\mathsf{yr}})$, based on thresholding the line length and direction. For instance, a skin patch with a long, upwards line is likely to belong to the partner's hand.

Spatial Priors. Finally, we learn the Gaussian parameters of the relative and absolute spatial priors, $P(L^i_p)$ and $P(L^i_p|L^i_q)$, from a small set of labeled training frames.

5. Experiments

We collected two datasets: video from a lab setting with interacting children and parents, and from a naturalistic setting with two adults. Our main motivation in this paper is to label the lab data automatically, which will be eventually used to study how interaction affects children's learning. In these experiments, a child and parent sit at a table and face one another with each wearing head-mounted cameras (top of Figure 1). Parents were told to engage their child with the three toys on the table and interact as naturally as possible. To try to limit distractions, the walls of the lab are colored white, and participants wear white coats.

We use the video from the child's camera, so that the other person in view is always the adult. The video is captured at 30Hz with 480×720 pixel resolution. Some sample frames are shown in Figure 2. We use video data from 5 different parent-child dyads (where toddlers' mean age was 19 months). Each of the 5 play sessions consist of 4 trials, with toys replaced between trials to keep the children interested. The trials had an average length of 1.5 minutes, leading to a total of 20 videos containing 56,535 frames (about 31 minutes) of social interaction from the children's perspective.

Our second dataset was designed to test our model in more naturalistic settings. We used Google Glass to record a small set of egocentric videos containing two adults engaged in three kinds of social interactions: playing cards, playing tic-tac-toe, and solving a 3-D puzzle. Each video is 90 seconds long, for a total of 4.5 minutes (8,100 frames), and was captured at 30Hz with a resolution of 1280×720.

5.1. Evaluation

To evaluate our approach, we manually annotated 2,400 random frames (around 120 per trial) from the lab dataset, and 300 frames (100 per video) from our Google Glass dataset, with bounding boxes. This is about one frame for every second of video. Depending on which body parts are in view, each frame has up to 5 bounding boxes: two observer's hands, two partner's hands, and one partner face.

Detection Accuracy. For each frame, our system estimates the location of each of the 5 body parts, by either providing a coordinate or indicating that it is outside the frame. We evaluate the accuracy of our method as the fraction of true positives (i.e. cases where we correctly estimate a position inside the ground-truth bounding box) and true negatives (i.e. where we correctly predict the part to be outside the frame). We also evaluate the percentage of "perfect" frames, those in which all five parts are predicted correctly.

Hand Disambiguation Error Rate. We are particularly interested in errors made when disambiguating the observer's hands from the partner's hands, so we measure this explicitly. We consider a ground-truth hand to be a disambiguation error if it is either unlabeled, labeled as the wrong person's hand, or is marked with multiple labels of different people (falsely estimating that hands overlap). The disambiguation error rate is the total number of incorrectly disambiguated hands over the total number of hands in all frames.

5.2. Results

We first present qualitative results on the lab dataset. Figure 5 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. Common failures include incorrectly estimating a hand to be out of frame (e.g. leftmost image) or falsely estimating overlapping hands. This can be caused by hands that are closer to the observer than expected and thus too big (e.g. in the middle two images), or because one hand is farther away from the other than usual (e.g. wrong prediction for 'my left hand' in the right image). We also show results for the naturalistic videos in Figure 6.

Quantitative Evaluation. We present detailed quantitative results in Table 1. Our overall detection accuracy across the five subjects of the lab data set is 68.4%. The technique generalized well between different subjects, as evidenced by a low standard deviation across videos (σ = 3.0). Accuracies between different hands are also fairly stable, ranging from 61.2% for 'my left hand' to 70.7% for 'my right hand'. Overall, our approach perfectly predicted 19.1% of frames, and for Subject 3 achieved a 24.7% perfect detection rate.

As expected, accuracy was lower for the naturalistic videos, at 50.7% overall. This drop in accuracy is caused by two factors. First, we do not use a model for the partner's hand (the edge based method described in Section 4 does not work well here). Second, the simple pixel-wise color-based skin detection suffers from illumination changes in the natural environment, as evidenced by our near-perfect skin detection accuracy on the lab videos (with 97% of detected skin pixels located inside ground-truth bounding boxes), but only 70% for the natural videos. Interestingly, we can still retain a relatively low disambiguation error rate in the naturalistic videos (35.6% versus 32.7%), showing that our model can compensate for noisy likelihoods.

Although our main purpose is to detect hands, the temporal and spatial constraints in our model also improve face detection. Table 1 compares the head-detection accuracy of our model to that of the raw Viola-Jones detector (column head$^{\mathsf{VJ}}$). We achieve about a 10-percentage-point increase for the lab dataset, and an over 17-percentage-point improvement on the Google Glass videos.

Comparing to Baselines. We compared our model to three baselines of increasing complexity. First, we tried a sim-

	Overall	Observer		Partner				% Perfect	Disambiguation
	Accuracy	right hand	left hand	right hand	left hand	head	head$^{\mathrm{VJ}}$	Frames	Error Rate
Subject 1	64.1	50.3	60.2	68.0	54.2	87.7	86.2	14.8	37.8
Subject 2	72.6	78.5	63.3	63.8	79.7	77.5	55.5	22.8	27.4
Subject 3	70.1	64.2	66.7	60.5	68.8	90.0	85.5	24.7	34.5
Subject 4	67.3	88.0	54.7	59.5	59.3	75.2	66.0	15.5	33.1
Subject 5	68.1	72.5	61.0	66.2	60.5	80.2	69.0	17.7	30.5
Average	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

Table 1. *Detection accuracies of our approach,* as well as a breakdown into different hands. We also compare our head-detection accuracy with the accuracy of the raw Viola-Jones detector (head$^{\mathrm{VJ}}$). The second to last column shows the percentage of frames in which all five predictions were correct and the last column shows the error we made when differentiating the observer's hands and the partner's hands.

ple random predictor: for every part in every frame, we first flip a coin to decide whether it is in the frame or not, and if it is in the frame, we assign it a random position. Second, we added the skin likelihood by repeating the same process but limiting the space of possible positions to be in skin patches. Finally, we build a more sensible baseline, clustering the detected skin pixels into hand-sized patches using Mean Shift [3]. Then, we greedily assign each part the position of the closest cluster centroid based on distance between centroid and part-wise absolute spatial priors.

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
ours (likelihood + spatial prior)	59.1	9.2	44.5
our method (full)	**68.4**	**19.1**	**32.7**
Naturalistic Videos			
skin clusters	39.2	0.0	65.4
our method	**50.7**	**9.0**	**35.6**

Table 2. *Comparison of our model's results to baselines,* in terms of overall accuracy, percentage of perfect frames, and hand disambiguation error rate (see text).

The results of these baselines and our method are compared in Table 2. The two random baselines perform poorly, with accuracies of 17.0% and 27.3%, respectively. The third method using clustering and distances to centroids performs better at 58.1%, but our approach still beats it with 68.4% accuracy. We also tested a simplified version of our model in which the in-frame and between-frame links were removed, so that only absolute spatial priors and likelihoods are used. This achieved 59.1% accuracy, comparable to the third baseline (which similarly does not incorporate temporal or relative spatial constraints). 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.

Finally, we compare the performance of our method and the third baseline on our naturalistic videos. The baseline method suffers drastically from the noisy skin detections and does not predict a single frame perfectly. Our method does much better at overcoming weak object models, perfectly predicting almost 10% of frames, showing that it has the potential to work well in less constrained scenarios.

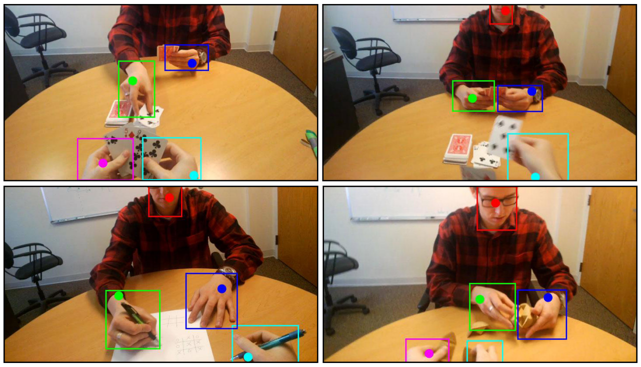

Figure 6. *Sample results for naturalistic video,* in which two people played cards (top) and tic-tac-toe and puzzles (bottom), while one wore Google Glass. (See Fig. 5 caption for color legend.)

6. Summary and Conclusion

We presented a probabilistic model for tracking hands in paired interaction and demonstrated its effectiveness using simple features on egocentric video. In future work, we would like to extend our experiments to more natural settings by collecting a more extensive paired-interaction dataset and by exploring stronger deformable object models (possibly solving jointly for the appropriate appearance model and part position). We also note that our work could inform and be informed by cues from action-recognition: a hierarchical framework in which actions are treated as latent variables is worth exploring.

Acknowledgments. This work was funded in part by NSF (CAREER IIS-1253549), NIH (R01 HD074601 and R21 EY017843), and the Indiana University Vice President for Research through an IU Collaborative Research Grant. JMF was supported by NICHD Training Grant 5T32HD7475-17.

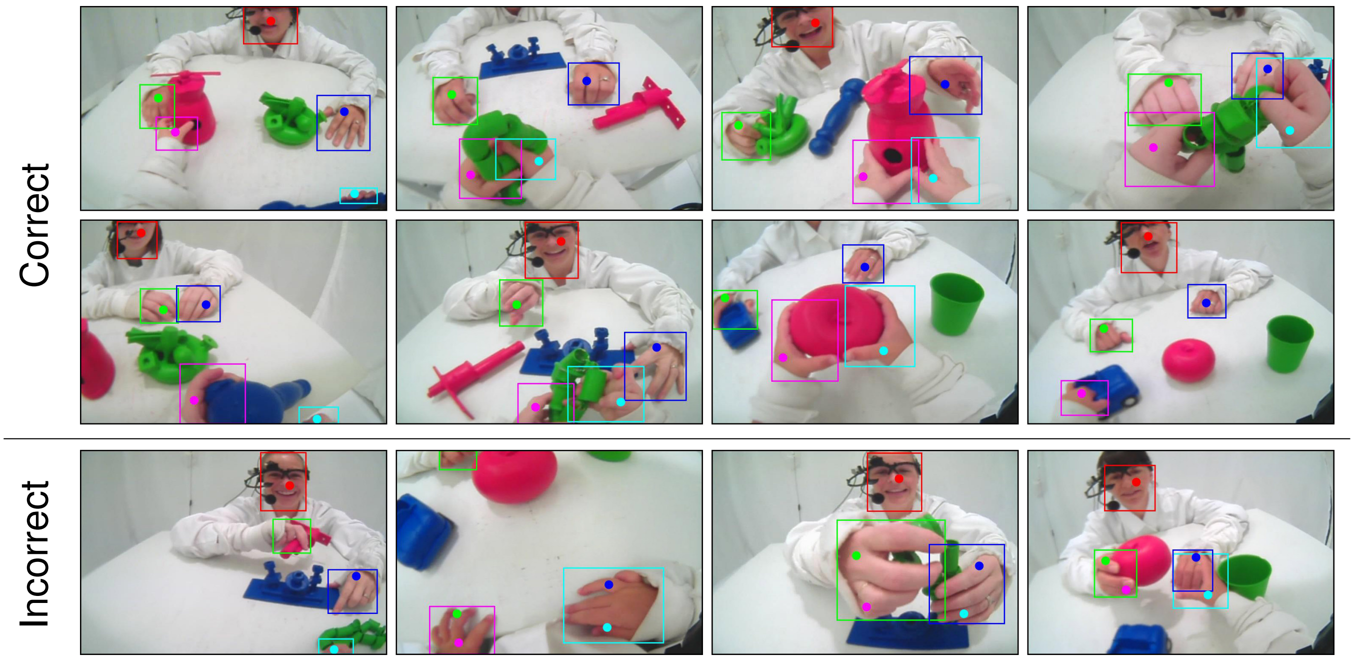

Figure 5. *Sample frames from our results,* with rectangles showing ground truth bounding boxes and dots showing predicted part positions (red = your head, blue = your left hand, green = your right hand, magenta = my left hand, cyan = my right hand). The first two rows show our robustness with respect to partial occlusions and changes in hand configurations, while the bottom row shows failure cases.

References

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

[2] G. Casella and E. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992. 5

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002. 7

[4] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, 2005. 3, 4

[5] 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

[6] A. Erol, G. Bebis, M. Nicolescu, R. Boyle, and X. Twombly. Vision-based hand pose estimation: a review. *CVIU*, 108:52–73, 2007. 2

[7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *CVPR*, 2011. 1, 2

[8] A. Fathi, J. Hodgins, and J. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 1, 2

[9] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2

[10] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 4

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

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

[13] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a hand manipulating an object. In *CVPR*, 2009. 2

[14] M. Jones and J. Rehg. Statistical color models with application to skin detection. *IJCV*, 45:81–96, 2002. 5

[15] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[16] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *ICCV*, 2013. 1, 2

[17] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, 2013. 1, 2

[18] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[19] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1, 2

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

[21] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 2

[22] M. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2

[23] 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

[24] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Chucchiara. Hand segmentation for gesture recognition in egovision. In *Workshop on Interactive Multimedia on Mobile & Portable Devices*, 2013. 1, 2

[25] E. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshop on Egocentric Vision*, 2009. 2

[26] T. Starner, J. Weaver, and A. Pentland. Real-time ASL recognition using desk and wearable computer based video. *PAMI*, 20, 1998. 2

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

[28] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Visual hand tracking using nonparametric belief propagation. In *CVPR*, 2004. 2

[29] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439. IEEE, 2010. 5

[30] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 5

[31] Q. Yuan, S. Scarloff, and V. Athitsos. Automatic 2d hand tracking in video sequences. In *WACV*, 2005. 2

[32] J. Zariffa and M. Popovic. Hand contour detection in wearable camera video using an adaptive histogram region of interest. *J. Neuroeng. Rehabil.*, 10(114), 2013. 1, 2