Detecting Hands in Children's Egocentric Views to Understand Embodied Attention during Social Interaction

Sven Bambach†, John M. Franchak, David J. Crandall†, Chen Yu
{sbambach, jmfranch, djcran, chenyu}@indiana.edu
† School of Informatics and Computing, Indiana University
Department of Psychological and Brain Sciences, Indiana University
Bloomington, IN, 47405 USA

Abstract

Understanding visual attention in children could yield insight into how the visual system develops during formative years and how children's overt attention plays a role in development and learning. We are particularly interested in the role of hands and hand activities in children's visual attention. We use head-mounted cameras to collect egocentric video and eye gaze data of toddlers during playful social interaction with their parents, and developed a computer vision system to track and label different hands within the child's field of view. We report detailed results on appearance frequencies and spatial distributions of parents' and children's hands both in the child's field of view and as the target of the child's attentional fixation.

Keywords: Attention; Development; Eye tracking; Vision

Introduction

The visual world is cluttered with objects and events generated by oneself and others. To efficiently process a cluttered and complex visual world, perceptual and cognitive systems must selectively attend to a subset of this information. Attention can be viewed as a spatial spotlight (Posner, 1980) that can be implemented both internally and externally. Although adults can attend to a location outside the area targeted by eye gaze (Shepherd, Findlay, & Hockey, 1986), attention is often tied to the body and sensory-motor behaviors — adults typically orient gaze direction to coincide with the focus of the attentional spotlight. Studies of adults engaged in complex tasks from making sandwiches to copying block patterns (Ballard, Hayhoe, Pook, & Rao, 1997; Hayhoe & Ballard, 2005) suggest that the momentary disposition of the body in space serves as a deictic (pointing) reference for binding sensory objects to internal computations (Ballard et al., 1997; Spivey, Tyler, Richardson, & Young, 2000). These studies analyzed the coordination of eye, head, and hands by measuring multiple streams of behavior in free-flowing tasks with multiple goals and targets for attention.

Attention and information selection are critical in early development and learning (Mundy & Newell, 2007) as early attention is predictive of later developmental outcomes (Ruff & Rothbart, 1996). Most studies of the development of attention employ highly-controlled experimental tasks in the laboratory. Many studies use remote eye tracking systems to measure looking behaviors, revealing much about the visual attention of toddlers as they passively examine visual stimuli displayed on a computer screen. However, more recent studies using head-mounted eye tracking have addressed visual selection in freely-moving toddlers when they are engaged in everyday tasks (Franchak, Kretch, Soska, & Adolph, 2011). In more natural interactions, there are multiple objects competing for attention, various manual actions toward those toy objects, and spontaneous goals. Visual attention changes from moment to moment according to the child's own actions and the parent's actions toward the child and objects. Though complex, these are the contexts in which real-world learning occurs. Compared with adults, young children's attentional systems may be even more tied to bodily actions.

The goal of the present study is to understand how sensory-motor behavior supports effective visual attention in toddlers. Towards this goal, we developed a more naturalistic experimental paradigm in which a child and parent wear head-mounted eye trackers while freely engaged with a set of toys. Each eye tracking system captures egocentric video from a first person perspective as well as gaze direction in the first-person view. In this way, we precisely measure the visual attention of both the parent and child, and also their manual actions. Recent findings using the same paradigm show that in toy play, both children and parents visually attend to not only the objects held by oneself but also the objects held by the social partner (Yu & Smith, 2013); in doing so, they create and maintain coordinated visual attention by looking at the same object at the same time. The target object is likely to be held by child or parent. Similarly, other work has shown that by holding objects, parents increase the likelihood that infants will look at parents' hands (Franchak et al., 2011). These results suggest the important role of hands and hand activities (of both children and parents) in toddlers' visual attention.

Given previous findings, the present study focuses on providing new evidence on how eye and hand actions interact to support effective visual attention to objects in toddlers. We first describe a new method to automatically detect hands and faces in egocentric video, allowing us to locate (at a pixel level) both one's own hands and the social partner's hands in the first person view. Next, we report a series of results that link hands and hand actions with visual attention, to show how the child's and parent's hands contribute to visual information selection in the child's view.

Experiment

To realize our overall goal of measuring visual attention in natural interactions, we developed a multi-modal sensing system that allows us to capture a wide variety of video and sensing data from participants in our lab.

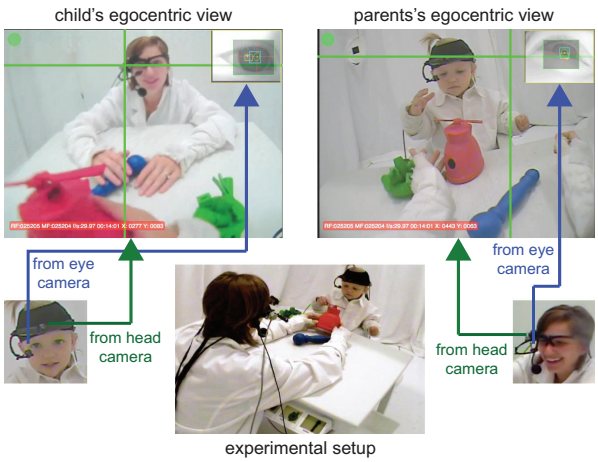

Figure 1: Experimental setup. We use 4 cameras to record joint play between a child and parent. The head-mounted eye tracking systems (worn by both) each consist of a head camera to capture its wearer's egocentric view and an eye camera that tracks the eye's pupil. All cameras work with a temporal resolution of 30Hz and a spatial resolution of 480×720px.

Muti-modal Sensing System

Our sensing environment allows us to monitor parents and children as they engage in free-playing interaction with toy objects, as shown in Figure 1. A child and parent sit at a table in the lab and face one another. Each wears a lightweight, head-mounted eye tracking system consisting of two cameras (Franchak et al., 2011): a wide-angle outward-facing camera (100° diagonal) capturing the egocentric field of view of the participant, and an inward-facing infrared camera pointed at the participant's left eye, which tracks the pupil in order to measure eye gaze position (shown by green cross-hairs in Figure 1). The eye tracker was calibrated by encouraging participants to look at known points in the environment; once calibrated, the accuracy of the eye tracker is about 3°. In addition, two scene cameras, two microphones and two head-mounted motion sensors with 6 degree-of-freedom tracking allowed for a variety of multi-modal coding. As the purpose of this study is to investigate the role of hands in a toddler's visual attention, the focus of this paper will be on the child's egocentric video and eye gaze data.

Subjects

For the study, we considered 6 child-parent dyads. The children's mean age was 19 months (SD = 2.56 months). Dyads were chosen based on hand tracking performance (see next section) among a pool of 14 candidates to ensure the greatest accuracy in the reported results. Although the sample size was small, analyzing high-density data with more than 10,000 frames per child yielded highly consistent results.

Procedure

Parents were told to engage their child with toys (three possible toys were on the table) and otherwise interact as naturally as possible, leading to a free-flowing interaction with no constraints on where parents or children looked or what they should do or say. Each experiment consisted of four trials and each trial lasted about 1 to 2 minutes. In between trials, the toy sets were replaced to keep the children interested.

Data

We collected a total of 67,913 frames (about 38 minutes) of video data from the 6 children. Of those frames, 54,367 had valid gaze data (i.e. located within the camera's field of view) in the form of an x-y coordinate, indicating the gaze center. To detect, track, and distinguish all hands that appear in our video data (including the child's left and right hands, and the parent's left and right hands), we developed a special hand tracking algorithm that is described in the following section.

Hand Tracking

Given the large amount of video data collected in our experiments, we needed automated techniques to track and label the positions of the hands in each video frame. Tracking is a well-studied problem in the computer vision literature, and some work has specifically studied hand tracking (Chen, Fu, & Huang, 2003) in the context of gesture recognition. However, most of that work studies video from stationary cameras. The fact that our video comes from head-mounted cameras introduces significant new challenges because observers' heads are free to move, continually changing the locations of hands in the field of view. In fact, we are not aware of any work that has studied multiple hand tracking in egocentric video; the closest is that of Ren and Gu (2010), who propose a system for recognizing objects held by the camera wearer.

Fortunately, the constraints of our lab environment help to ease our tracking problem: we know there are at most two people in each frame, that the child's hands are closer to the camera than the adult's hands, that in general the children and parents are facing one another, and that the participants' clothing is white. Our goal is to identify which of the four hands (child's hands and parent's hands) are visible in each frame, and then to identify the position of the visible ones. Our approach consists of four major steps: (1) identifying potential skin pixels based on color; (2) clustering these pixels into candidate hand and face regions; (3) tracking these regions over time; and (4) labeling each region with its body type (face, child left or right hand, parent left or right hand).

Step 1: Skin Detection

To look for faces and hands, we first identify pixels having skin-like colors. Although human skin colors are surprisingly consistent across people when represented in an appropriate color space (we use YUV here), pixel-level skin classification is difficult because illumination can dramatically alter skin appearance and because many common objects (like walls) often have skin tones. We thus tuned our skin classifier for each individual subject, by sampling 20 frames at random and having a human label the skin regions in each frame. We then used these labeled pixels as training exemplars to learn a simple Gaussian classifier, in which each pixel is encoded as

a 2d feature vector consisting of the two color dimensions (U and V). To detect skin in unlabeled images, we evaluate the likelihood of each pixel under this model, threshold to find candidate skin pixels, and use an erosion filter to eliminate isolated pixels.

Step 2: Skin Grouping

Given the skin detection results from Step 1, we apply Mean Shift clustering (Comaniciu & Meer, 2002) to each frame to group skin pixels into candidate skin regions. Mean Shift does not require knowing the number of clusters ahead of time (as K-means does), but instead requires an estimate of the size and shape of the clusters; we use circular disks of radius 75 pixels in our implementation.

Step 3: Tracking

We next attempt to find correspondences between the skin blobs estimated in temporally-adjacent frames, in order to create *tracks* of skin regions over time. To do this, we scan the frames of the video in sequence. For each frame i, we assign each skin region to the same track as the closest region in frame $i-1$ as long as the Euclidean distance between the region centroids is below a threshold (we use 50 pixels), and otherwise we start a new track. Each track thus consists of a starting frame number indicating when the region appears, an ending frame number indicating when it disappears, and an (x,y) position of the region within each intervening frame.

Step 4: Labeling Skin Regions

Finally, we need to identify which tracks from Step 3 correspond to actual skin regions, and then to label each track with one of five possible body parts (parent's head, child left or right hand, parent left or right hand). We experimented with various strategies and settled on a relatively simple approach that uses the relative spatial location of regions within the frame (and in particular the observation that the parent's head is usually above and between the parent's hands, which are in turn above the child's hands). We thus first try to find tracks corresponding to the face, and then check the relative position of other regions to find and label the hands.

Detecting Face Tracks. We tried off-the-shelf face detectors, but they are not reliable in our context because the parent's head is often not fully visible (e.g. in left frame of Figure 2). Instead we built a very simple face detector that uses the fact that the parents in our experiments wear a black head-mounted camera. In particular, we trained a linear Support Vector Machine classifier (Burges, 1998) on manually-labeled head regions (using the same 20 frames that we used to learn skin color, with the remaining skin regions serving as negative exemplars), where the features consist of a 256-bin grayscale histogram over the pixels in the track region. We then identify faces by finding tracks for which the trained SVM classifies over half of the regions in the track as faces.

Labeling Hands. Once face tracks have been found, we mark potential hand tracks based on their relative position with respect to the face. Anchoring the expected spatial locations of hands to the parent's head helps compensate for view changes due to the child's head motion. In particular, we create a configuration of five points ("hotspots") that roughly correspond to the expected (mean) position of the four hands relative to the face, illustrated as red circles in Figure 2. For each non-face candidate track generated by Step 3, we compute the centroid of its location across the frames in which it is visible, find the hotspot closest to the centroid, and assign the track to the corresponding body part. When no face is detected, the hotspots take a default position that assumes the face is in the top-center (Figure 2, left pane).

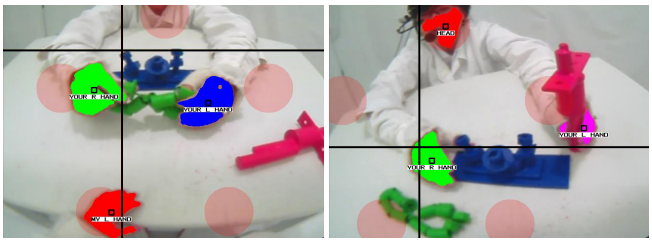

Figure 2: Two sample frames showing the results of our egocentric hand-tracking algorithm. The red circles denote the "hotspots" used to label skin blobs. The black crosshair indicates eye gaze center. *Left:* Hotspots in their default center location. *Right:* Hotspots aligned based on detected face.

Evaluation

We manually tested the accuracy of our hand tracking algorithm on 600 randomly-selected frames (100 frames for each of 6 subjects), and counted the proportion of correctly-labeled regions. We found that the overall accuracy was 71%, ranging from 67% to 75% across the subjects. In comparison, a baseline method that randomly assigns labels to skin regions (and assuming that the skin segmentation and clustering perform correctly) achieves 20% accuracy. Labeling errors are caused by a variety of factors, but the two most common are: (1) when hands are close together and the clustering algorithm incorrectly combines them into a single body part, and (2) when hands spend a significant amount of time away from their expected location relative to the head.

Results and Discussion

The hand tracking algorithm provides frame-by-frame data about the position, size, and shape of each hand in the child's field of view. We analyzed this data in terms of spatial distributions of the different hand classes and in terms of how often each class appears over time. Further, we used children's eye gaze data to investigate moments where hands were the target of the child's overt visual attention.

Hands in the Child's Field of View

To determine how often children had the opportunity to view their own hands and their parents' hands, we calculated how often hands are present in the field of view.

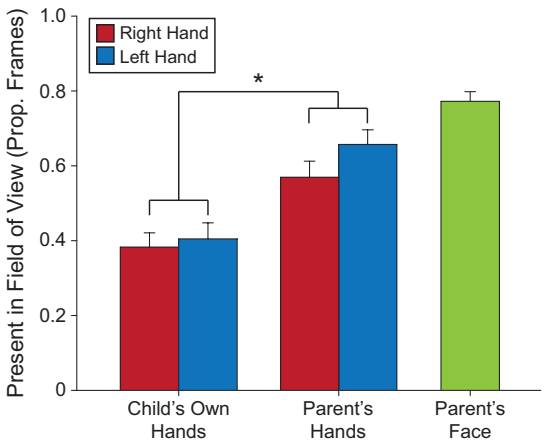

Figure 3: Bar graphs showing the proportions of frames in which each class of hands was detected (error bars show 1 SE). For comparison, the value for the parent's face is shown.

Frequency of Hands in View. Figure 3 shows the proportion of frames in which each hand class was detected. As the hand tracking algorithm needs to distinguish hands from faces and thus implicitly tracks the parent's face as well, we include results for the face for reference. Overall, hands were frequently in view, although the child's own hands (right hand = 38% and left hand = 40%) are in view less frequently than the parent's hands (right hand = 57% and left hand = 66%). A 2 (agent: child, parent) $\times$ 2 (hand: left, right) repeated-measures ANOVA confirmed a main effect of agent, $F(1,5) = 21.74$, $p = .006$. The main effect of agent $\times$ hand interaction did not reach significance.

The parent's face also appeared frequently in the child's view and was detected in 77% of the frames. We note here that the set of subjects we chose (based on hand detection accuracy) might be slightly biased to have parent's faces in view more often than others as our algorithm uses face information to improve its prediction and thus tends to perform better for subjects where the face is in view more frequently.

Spatial Distribution of Hands in View. Spatial asymmetries might account for the different frequencies with which children's and parents' hands were visible. Next, we present spatial distributions of hands in the children's field of view in the form of heat maps. The first row of Figure 4B-C shows the distributions of the child's left hand, the child's right hand, the parent's right hand and the parent's left hand, respectively. Each data point in the heat map corresponds to the centroid (mean of the hand blob) of the detected hand. The distributions are accumulated over all 6 subjects where N depicts the number of frames with the hand in view. To allow quantitative comparison, we calculated robust (60% trimmed) statistics in the form of horizontal and vertical mean (μ) as well as horizontal and vertical standard deviation (σ) of the distributions (off-diagonal co-variances are not shown).

Children's left and right hands had very similar distributions in terms of variance (Figure 4B) with distributions that expanded more horizontally than vertically: σ_x was roughly twice as much as σ_y for each hand. Parents' left and right hands also have similar distributions in terms of variance (Figure 4C). A 2 (agent: child, parent) $\times$ 2 (hand: left, right) $\times$ 2 (direction: horizontal, vertical) ANOVA confirmed the main effect of direction, $F(1,5) = 36.4$, $p = .002$. However, a significant agent $\times$ direction interaction, $F(1,5) = 10.5$, $p = .023$ and follow-up pairwise comparisons show that parents' hands occupy a larger vertical space (right hand $\sigma_y = 59$, left hand $\sigma_y = 60$) compared to children's hands (right hand $\sigma_y = 37$, left hand $\sigma_y = 43$, $p = .009$). Horizontal variance terms did not differ between the hands of children and parents, and no other effects approached significance.

Children's and parents' hands were spatially segregated in visual space. Overall, children's hands were lower in the visual field compared to parents' hands and were often seen towards the lower boundary of the field of view ($\mu_y = -172$ for the left hand and $\mu_y = -178$ for the right hand). A 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on μ_y revealed that parents' hands were significantly higher than children's hands (main effect of agent, $F(1,5) = 184.6$, $p < .001$). In the horizontal dimension, the child's right hand and parents' left hand tended to reside in the right half of the visual field, while the child's left hand and parents' right hand tended to reside in the left half of the visual field. A 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on μ_x confirmed a significant agent $\times$ hand interaction, $F(1,5) = 1377.7$, $p < .001$.

Since our automatic hand labeling is not perfect and makes spatial assumptions, these results could potentially be biased by our algorithm. We manually labeled the location of hands in 2,800 randomly sampled frames and repeated our analyses. A 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on the μ_y's of manually labeled frames confirmed that parents' hands were located higher than those of children (main effect of agent, $F(1,5) = 111.1$, $p < .001$). In addition, a 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on the μ_x's in hand labeled frames showed a significant agent $\times$ hand interaction as in frames labeled by our algorithm, $F(1,5) = 529.4$, $p < .001$). We conclude that our results on spatial locations of hands in the field of view are not an artifact of our algorithm.

Different spatial distributions of hands may account for different frequencies of hands being visible. Most likely, children's hands were not as frequent as parents' hands because they occupied locations towards the lower boundary of the field of view. If children moved their hands down or tilted their heads up, their own hands would leave the field of view.

Hands as Target of the Child's Overt Attention

Next we examined how often and where hands were targeted by children's gaze. We counted a gaze fixation on the hand whenever a 10° hot spot (corresponding to a circle with radius of 32 pixels) around the gaze center overlapped with the area of a detected hand.

Frequency of Hands Being Targeted by Gaze. Figure 5 (left) shows mean values for the overall proportion of frames

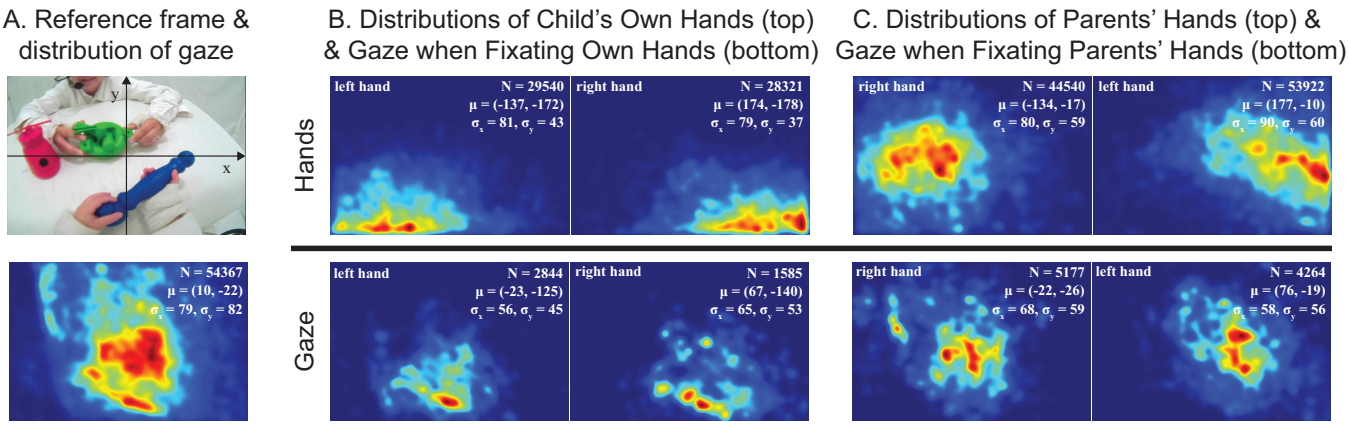

Figure 4: Spatial distributions of hands and eye gaze. *Column A:* The top image shows a sample frame from the child's view, while the bottom image shows the spatial distribution of the children's eye gaze across all valid frames. *Column B:* The top row shows the distributions of children's own hands (based on hand centroids) within their field of view. The bottom row shows the distributions of children's eye gaze while looking at their own hands. *Column C:* The top row shows the distributions of parent's hands within the children's field of view. The bottom row shows the distributions of children's eye gaze while looking at their parent's hands. Also shown are robust (60% trimmed) estimates of mean (μ) and standard deviation (σ) of the distributions as well as the number of data points (N). Heat maps are 480×720px and a small Gaussian blur ($\sigma_G = 10$px) was applied.

in which children's gaze overlapped with each hand. Children spent about twice as long looking at parent's hands (about 9.5% for the right hand and 7.8% for the left hand) than they did looking at their own hands (3.0% right hand and 5.3% left hand). A 2 (agent: child, parent) × 2 (hand: left, right) on proportion of frames targeting hands confirmed a main effect of agent, $F(1,5) = 8.52$, $p = .03$, and found no other significant effects.

Higher rates of looking to parents' hands may be the result of parents' hands being in view more often. Thus, we recalculated the proportion of looking to hands based on the number of frames where each hand was present in the field of view (right side of Figure 5). This normalization increased the proportion of looking for both the child's own hands and the parent's hands. Furthermore, the difference between the time spent looking at parent's hands and looking at their own hands is no longer significant when taking the availability of hands into account (no effects found in a 2 (agent: child, parent) × 2 (hand: left, right) on normalized proportions of frames targeting hands).

Spatial Distribution of Gaze when Targeting Hands. Finally, we present the spatial distributions of children's eye gaze (bottom row of Figure 4) when viewing hands. The gaze heat maps are composed similarly to the hand heat maps, except that each data point now corresponds to the eye gaze center as opposed to a hand centroid. Prior work has shown that gaze tends to be biased towards the center of the field of view (Foulsham, Walker, & Kingstone, 2011). Figure 4A shows the overall distribution of eye gaze during the experiment, accumulated over all 6 subjects. Indeed, the distribution has this center bias with a mean close to the center of field of view and similar variances in horizontal and vertical direction.

The bottom row of Figure 4B-C shows subsets of the eye

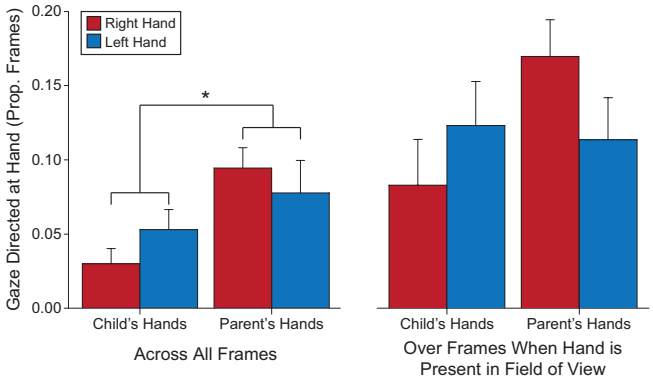

Figure 5: Bar graphs showing the proportions of frames in which each class of hands was looked at (based on a 10° gaze hot spot). *Left:* Fractions based on all frames with valid eye gaze ($N = 54367$). *Right:* Fractions based on all frames where the corresponding hand was in the field of view.

gaze data, taking only into account moments when gaze was fixated on one of the hands. Across children's and parents' hands, we observed that distributions when gaze targeted hands were more centrally located compared to the overall distributions of hands in the field of view. To verify this statistically, we calculated the distance from the center of the field of view for the means of the distributions of the hands and the gaze locations when hands were fixated (top and bottom rows of Figure 4B-C). A 2 (agent: child, parent) × 2 (hand: left, right) × 2 (distribution: hands overall, gaze-targeted) revealed a main effect of agent, $F(1,5) = 66.1$, $p < .001$, distribution, $F(1,5) = 13.7$, $p = .014$, and a significant 3-way interaction, $F(1,5) = 17.7$, $p = .008$. Overall, parents' hands ($M = 131.4$ pixels) were closer to the center of the field of view compared to children's hands ($M = 195.5$ pixels). Follow-up tests on the 3-way interaction showed that

child's left hand, child's right hand, and parent's left hand were more centrally located when targeted by gaze compared to their overall distributions ($p < .05$), while the parent's right hand location did not change when targeted by gaze ($p = .48$).

Discussion

Hands are an important visual stimulus. One's own hands are relevant for guiding reaching actions and manipulating objects (Hayhoe & Ballard, 2005; Franchak et al., 2011), while the hands of others can convey information about the attention and goals of social partners (Olofson & Baldwin, 2011; Ullman, Harari, & Dorfman, 2012). But for toddlers to learn from hands, they must be able to see them. Here, we demonstrate that for toddlers playing with adults, hands are frequently in view. However, what infants see depends on where they actively point their heads: the resulting spatial constraints (e.g., child's hands being low in the field of view) mean that children's own hands are in view less often than their parents' hands. Consequently, children overtly attend to parents' hands more often than their own hands. Moreover, we show that when children fixate on hands, they do so more often when hands are centrally located in their fields of view, suggesting that children move their heads to bring visual targets into the center of their visual fields. Most likely, children coordinate their eyes and heads to focus on areas relevant to the task at hand, looking down towards their own hands when reaching and looking up towards their parents' hands when parents present objects (Yu & Smith, 2013).

Future Directions

There are two major directions that we are exploring in future work. First, we want to improve the performance of our hand tracking algorithm. Towards this goal, we are working on probabilistic frameworks that will allow us to jointly take the spatial configurations of all hands into account when deciding on a hand label. Better performance will allow us to evaluate more participants in the future to further validate these results in a larger sample. Second, we want to take advantage of the idea that hands can be useful clues towards predicting overt visual attention in children by building models that attempt to predict children's eye gaze. We briefly experimented with very simple models that predict the child's gaze location based on the most visualy dominant hand in view and achieved accuracies that could compete with saliency-based predictions on our data (Bambach, Crandall, & Yu, 2013).

Acknowledgments

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

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(04), 723–742.

Bambach, S., Crandall, D. J., & Yu, C. (2013). Understanding embodied visual attention in child-parent interaction. In *IEEE Joint International Conference on Development and Learning and Epigenetic Robotics,* (pp. 1–6).

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167.

Chen, F.-S., Fu, C.-M., & Huang, C.-L. (2003). Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Vision Computing*, *21*(8), 745–758.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920 - 1931.

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

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current Directions in Psychological Science*, *16*(5), 269–274.

Olofson, E. L., & Baldwin, D. (2011). Infants recognize similar goals across dissimilar actions involving object manipulation. *Cognition*, *118*(2), 258–264.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Ren, X., & Gu, C. (2010). Figure-ground segmentation improves handled object recognition in egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition.*

Ruff, H. A., & Rothbart, M. K. (1996). *Attention in early development: Themes and variations*. Oxford University Press.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology*, *38*(3), 475–491.

Spivey, M., Tyler, M., Richardson, D., & Young, E. (2000). Eye movements during comprehension of spoken scene descriptions. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 487–492).

Ullman, S., Harari, D., & Dorfman, N. (2012). From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, *109*(44), 18215-18220.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11), e79659.