Detecting social information in a dense database of infants' natural visual experience

Bria L. Long[1], George Kachergis[1], Ketan Agrawal[1], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University

Author Note

Correspondence concerning this article should be addressed to Bria L. Long, 450 Serra Mall, Stanford CA 94305. E-mail: bria@stanford.edu

Abstract

The faces and hands of caregivers and other social partners offer a rich source of social and causal information that may be critical for infants' cognitive and linguistic development. Previous work using manual annotation strategies and cross-sectional data has found systematic changes in the proportion of faces and hands in the egocentric perspective of young infants. Here, we examine the prevalence of faces and hands in a longitudinal collection of nearly 1700 headcam videos collected from three children along a span of 6 to 32 months of age—the SAYCam dataset (Sullivan, Mei, Perfors, Wojcik, & Frank, under review). To analyze these naturalistic infant egocentric videos, we first validated the use of a modern convolutional neural network of pose detection (OpenPose) for the detection of faces and hands. We then applied this model to the entire dataset, and found a higher proportion of hands in view than previous reported and a moderate decrease the proportion of faces in children's view across age. In addition, we found variability in the proportion of faces/hands viewed by different children in different locations (e.g., living room vs. kitchen), suggesting that individual activity contexts may shape the social information that infants experience.

Keywords: social cognition; face perception; infancy; head cameras; deep learning

Word count: 3444

Detecting social information in a dense database of infants' natural visual experience

Introduction

Infants are confronted by a blooming, buzzing onslaught of stimuli (James, 1890) which they must learn to parse to make sense of the world around them. Yet they do not embark on this learning process alone: From as early as 3 months of age, young infants follow overt gaze shifts (Gredeback, Theuring, Hauf, & Kenward, 2008), and even newborns prefer to look at faces with direct vs. averted gaze (Farroni, Csibra, Simion, & Johnson, 2002), despite their limited acuity. As faces are likely to be an important conduit of social information that scaffolds cognitive development, psychologists have long hypothesized that faces are prevalent in the visual experience of young infants.

Yet until recently most hypotheses about infants' visual experience have gone untested. Though parents and scientists alike have strong intuitions about what infants see, even the viewpoint of a walking child is not easily predicted by these intuitions (Clerkin, Hart, Rehg, Yu, & Smith, 2017; Franchak, Kretch, Soska, & Adolph, 2011). By equipping infants and toddlers with head-mounted cameras, researchers have begun to document the infant's egocentric perspective on the world. Using these methods, a growing body of work now demonstrates that the viewpoints of very young infants (less than 4 months of age) are indeed dominated by frequent, persistent views of the faces of their caregivers (Jayaraman & Smith, 2018; Jayaraman, Fausey, & Smith, 2015; Sugden, Mohamed-Ali, & Moulson, 2014).

Beyond these early months, infants' motor and cognitive abilities mature, leading to vastly different perspectives on the world. For example, crawlers see fewer faces and hands than do walking children (Franchak, Kretch, & Adolph, 2017; Kretch, Franchak, & Adolph, 2014; Sanchez, Long, Kraus, & Frank, 2018) as well as different views of objects (Smith, Yu, & Pereira, 2011). Further, as infants learn to use their own hands to act on the world, they seem to focus on manual actions taken by their social partners, and their perspective starts to capture views of hands manipulating objects (Fausey, Jayaraman, & Smith, 2016). In turn, caregivers may also start to use their hands with more communicative intent, directing infants' attention by pointing and gesturing to different events and objects during play (Yu & Smith, 2013).

Here, we examine the social information present in the infant visual perspective—the presence of faces and hands—by analyzing a longitudinal collection of nearly 1700 headcam videos collected from three children along a span of 6 to 32 months of age—the SAYCam dataset (Sullivan et al., under review). In addition to its size and longitudinal nature, this dataset is more naturalistic than those previously used in two key ways. First, recordings were taken under a large variety of activity contexts (Bruner, 1985; B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015) encompassing infants' viewpoints during both activities outside and inside the home. Even in other naturalistic datasets, the incredible variety in a typical infant's experience has been largely underrepresented (see examples in Figure 1; e.g., riding in the car, gardening, watching chickens during a walk, browsing magazines, nursing, brushing teeth). Second, the head-mounted cameras used in the SAYCam dataset captured a larger field of view than those typically used, allowing a more complete picture of the infant perspective. While head-mounted cameras with a more restricted field of view do represent where infants are foveating most of the time (Smith, Yu, Yoshida, & Fausey, 2015; Yoshida & Smith, 2008), they may fail to capture short saccades to either faces or hands in the periphery, as the timescale of head movements is much longer.

With hundreds of hours of footage (>40M frames), however, this large dataset necessitates a shift to an automated annotation strategy. Indeed, annotation of the frames extracted from egocentric videos has been prohibitively time-consuming, meaning that most frames are typically not inspected, even in the most comprehensive studies. For example, Fausey et al. (2016) collected a total of 143 hours of head-mounted camera footage (15.5 million frames), of which one frame every five seconds was hand-annotated (by four coders), totalling 103,383 frames (per coder)—an impressive number of annotations but nonetheless only 0.67% of the collected footage. To address this challenge, we use a modern computer vision model of pose detection to automatically detect the presence of hands and faces from the infant egocentric viewpoint. Specifically, we use OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018), a model optimized for jointly detecting human face, body, hand, and foot keypoints that operates well on scenes including multiple people, even if they are partially-occluded (see Figure 1). In prior work examining egocentric videos, OpenPose performed comparably to other modern face detection models (Sanchez et al., 2018).

In this paper, we first describe the dataset and validate the use of this model by comparing face and hand detections to a human-annotated set of 24,000 frames. Next, we report how the proportion of faces and hands changes with age in each of the three children in the dataset. We then investigate sources of variability in our more naturalistic dataset that may explain differences from prior work, including both the field-of-view of the head cameras as well as a diversity of locations in which videos were recorded.

Method

Dataset

The dataset is described in detail in Sullivan et al. (under review); we summarize these details here. Children wore Veho Muvi miniature cameras mounted on a custom camping headlamp harness ("headcams") at least twice weekly, for approximately one hour per recording session. One weekly session was on the same day each week at a roughly constant time of day, while the other(s) were chosen arbitrarily at the participating family's discretion. At the time of the recording, all three children were in single-child households. Videos captured by the headcam were 640x480 pixels, and a fisheye lens was attached to the camera to increase the field of view to approximately 109 degrees horizontal x 70 degrees vertical.

Videos[1] with technical errors or that were not taken from the egocentric perspective were excluded from the dataset. While data collection for the third child (Y) is still ongoing, here we analyze 1698 videos, with a total duration of 374.88 hours (>40 million frames).

Detection Method

To automatically annotate the millions of frames in SAYCam, we used a pose detector, OpenPose[2] (Cao et al., 2018; Simon, Joo, Matthews, & Sheikh, 2017), which provided the locations of 18 body parts (ears, nose, wrists, etc.). To do so, a convolutional neural network was used for initial anatomical detection, and part affinity fields were subsequently applied for part association to produce a series of body part candidates. Once these body part candidates were matched to a single individual in the frame, they were finally assembled into a pose. Thus, while we only made use of the outputs of the face and hand detections, the entire set of pose information from an individual was used to determine the presence of a face/hand, making the process more robust to occlusion than methods optimized to detect only faces or hands. Note, however, that these face/hand detections are reliant on the detection of at least a partial pose, so some very up-close views of faces/hands may go undetected.

Detection Validation

To test the validity of OpenPose's hand and face detections, we compared the accuracy of these detections relative to human annotations of 24,000 frames selected uniformly at random from videos of two children (S and A). Frames were jointly annotated for the presence of faces and hands by one author. A second set of coders recruited via AMT

[1] All videos are available at https://nyu.databrary.org/volume/564
[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

(Amazon Mechanical Turk) additionally annotated 3150 frames; agreement with the primary coder was >95%.

As has been observed in other studies on automated annotation of headcam data (Bambach, Lee, Crandall, & Yu, 2015; e.g. Frank, Simmons, Yurovsky, & Pusiol, 2013; Sanchez et al., 2018), detection tasks that are easy in third-person video can be quite challenging in egocentric videos, due to difficult angles and sizes as well as substantial occlusion. For example, the infant perspective often contains non-canonical viewpoints of faces (e.g., looking up at a caregiver's chin) as well as partially-occluded or oblique viewpoints of both faces and hands. Further, hand detection tends to be a harder computational problem than face detection (Bambach et al., 2015; Simon et al., 2017). We thus expected overall performance to be lower in these naturalistic videos than on either photos taken from the adult perspective or in egocentric videos in controlled, laboratory settings (e.g., Sanchez et al., 2018).

To evaluate OpenPose's performance, we compared its detections to the manually-annotated gold set of frames, calculating precision (hits / hits + false alarms), recall (hits / hits + misses), and F-score (the harmonic mean of precision and recall). In our data, for faces, the F-score was 0.59, with a precision of 0.65 and recall of 0.54. For hands, the F-score was 0.49, with a precision of 0.70 and recall of 0.38. While face and hand detections showed moderately good precision, face detections were overall slightly more accurate than hand detections. In general, hand detections suffered from fairly low recall, indicating that OpenPose likely underestimated the proportion of hands in the dataset.

We suspected that this was in part because children's own hands were often in view of the camera and unconnected to a pose—-a notoriously challenging detection problem (Bambach et al., 2015). To assess this possibility, we obtained human annotations for the entire subsample of 9051 frames in which a hand was detected; participants (recruited via AMT) were asked to draw bounding boxes around children's and adult's hands. Overall, we found that 42% of missed hand detections were of child hands. When frames with children's hands were removed from the gold set, recall did improve somewhat to 0.54. We also observed that children's hands tended to appear in the lower half of the frames; heatmaps of the bounding boxes obtained from these annotations can be seen in Figure 2.

Finally, we examined whether the precision, recall, and F-score for hands and faces varied with age or child, and did not find substantial variation. Thus, while OpenPose was trained on photographs from the adult perspective, this model still generalized relatively well to the egocentric infant viewpoint with no fine-tuning or post-processing of the detections. As these detections were imperfect compared to human annotators, fine-tuning these models to better optimize for the infant viewpoint remains an open avenue for future work. Standard computer vision models are rarely trained on the egocentric viewpoint, and we suspect that training these models on more naturalistic data may lead to more robust, generalizable detectors.

Results

Access to social information across age

We analyzed the social information in view across the entire dataset, looking specifically at the proportions of faces and hands that were in view for each child.[3] Data from videos were binned according to the age of the child (in weeks). First, we saw that the proportion of faces in view showed a moderate decrease across this age range (see Figure 3); in contrast, we did not observe an increase in the proportion of hands in view, but rather a slight decrease. These effects were quantified with two separate linear mixed-models (see

[3] All analyses and preprocessed data files for this paper are available at https://tinyurl.com/detecting-social-info

Tables 1 & 2).[4]

However, the most striking result from these analyses is a much overall greater proportion of hands in view than have previously been reported (Fausey et al., 2016). We found this to be true across all ages, in all three children, and regardless of whether we analyzed human annotations (on the 24K random subset, see dotted lines in Figure 3) or OpenPose annotations on the entire dataset (see solid lines in Figure 3). This is notable especially given that OpenPose showed relatively low recall for hands, indicating that this may be an underestimate of the proportion of hands in view. Nonetheless, one reason this could be the case is the much larger field of view that was captured by the cameras used in this study: These cameras were outfitted with a fish-eye lens in an attempt to capture as much of the children's field of view as possible, leading to a larger field of view (109 degrees horizontal x 70 degrees vertical) than in many previous studies; for example, in Fausey et al. (2016) the FOV was 69 x 41 degrees. This larger FOV may have allowed the SAYCam cameras to capture not only the presence of a social partner's hands interacting with objects or gestures, but also the children's own hands, leading to more frequent hand detections.

As children's hands tended to occur in the lower visual field (see Figure 2), we thus re-analyzed the entire dataset while restricting our analysis to the center field of view, decreasing the proportion of hand detections from 24% to 16%, but only decreased face detections from 18% to 10%. This cropping likely removed both the majority of detections of children's own hands but also some detections of adult hands (see Figure 2), especially as OpenPose was biased to miss children's hands when they were in view. Nonetheless, within this modified field of view, we still observed more hand detections than face detections (see dashed lines in Figure 3).

[4] Face/hand detections were binned across each week of filming. Participant's age was converted into months and centered for these analyses. Random slopes for the effect of age by child led to a singular fit and were removed from both analyses; see full model specification in accompanying codebase.

Access to social information in different locations

How does variability across different contexts influence the social information in the infant view? Intuitively, some activities in different contexts may be characterized by a much higher proportion of faces (e.g., diaper changes in bedrooms) than others (e.g., playtime in the living room). We thus next examined variation in presence of hands and faces across different locations. Of the 1698 videos, 639 were annotated (Sullivan et al., under review) for the location they were filmed in. Of these, 296 videos were filmed in single location, representing 17 percent of the dataset and over 5 million frames. Activities varied somewhat predictably by these contexts: for example, eating tended to occur in the kitchen, whereas playtime was the dominant activity in the living room. Overall, we found that the proportion of faces vs. hands varied across filming locations, and, to some extent, across children; separate chi-squared tests for each child and detection type revealed significant variability in detections by location in each case.[5] For example, while both A and S saw a relatively similar proportion of faces vs. hands in the bedroom, they saw quite different amounts of faces vs. hands in kitchens (see Figure 4).

General Discussion

Here, we analyzed the social information in view in a dense, longitudinal dataset, applying a modern computer-vision model to quantify the proportion of hands and faces seen from each of three children's egocentric perspective from 6 to 32 months of age. This analysis has yielded a better understanding of infants' evolving access to social information. We found a moderate decrease across age in the proportion of faces in view in the videos, in keeping with previous work (Fausey et al., 2016). This finding is particularly notable given that, in previous cross-sectional data, this effect seems to be most strongly driven by infants

[5] all p<.001, see accompanying codebase

younger than 4 months of age (e.g., Fausey et al., 2016; Jayaraman et al., 2015; Sugden et al., 2014) who see both more frequent and more persistent faces (Jayaraman & Smith, 2018).

We also found an unexpectedly high proportion of hands in the view of infants, even when restricting the field-of-view to the center field of view the videos to make the viewpoints comparable to those of headcams used in previous work (Fausey et al., 2016). Why might this be the case? One idea is that these videos contain the viewpoints of children not only during structured interactions (e.g., play sessions at home or in the lab) but during everyday activities when children may be playing by themselves or simply observing the actions of caregivers and other people in their environment. During these less structured times, caregivers may move about in the vicinity of the child but not interact with them as directly—leading to views where a person and their hands are visible from a distance, but this person's face may be turned away from the infant or occluded (see examples in Figure 1). Indeed, using the same pose detector on videos from in-lab play sessions, Sanchez et al. (2018) found the opposite trend: slightly fewer hand detections than face detections from 8-16 months of age. Work that directly examines the variability in the social information in view across more vs. less structured activity contexts could further test this idea.

A coarse analysis based on the location the videos were filmed in further highlights the variability of the social information in view during different activities, showing differences across locations and between individual children. Within a given, well-defined context—e.g., mealtime in kitchens—S saw more faces than A, and S saw more faces in the kitchen than in other locations. This variability likely stems from the fact that there are at least three ways to feed a young child: 1) sitting in front of the child, facing them as they sit in a high chair; 2) sitting behind the child, holding them as they face outward, and 3) sitting side by side. Each of these positions offer the child differing degrees of visual access to faces and hands. While the social information in view may be variable across children in different activity contexts, these analyses suggest they could be stable within a given child's day-to-day experience.

Overall, these analyses underscore the importance of how, when, from whom, and what data we sample; these choices become central when we attempt to draw conclusions about the regularities of experience. Indeed, while unprecedented in size, this dataset still has many limitations. These videos only represent a small portion of the everyday experience of these three children, all of whom come from relatively privileged households in western societies and thus are not representative of the global population. Any idiosyncrasies in how and when these particular families chose to film these videos also undoubtedly influences the variability seen here. And without eye-tracking data, we do not know if children are attending to the social information in their visual field.

Nonetheless, we believe that these advances in datasets and methodologies represent a step in the right direction. The present paper demonstrates the feasibility of using a modern computer vision model to annotate the entirety of a very large dataset (here, >40M million frames) for the presence and size of people, hands, and faces, representing orders of magnitude more data relative to prior work. We propose that the large-scale analysis of dense datasets, collected with different fields of view, cameras and from many different laboratories, will lead to generalizable conclusions about the regularities of infant experience that scaffold learning.

Acknowledgements

Thanks to the creators of the SAYCam dataset who made this work possible and to Alessandro Sanchez for his contributions to the codebase. This work was funded by a Jacobs Foundation Fellowship to MCF, a John Merck Scholars award to MCF, and NSF #1714726 to BLL.

References

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proc. of the IEEE international conference on computer vision* (pp. 1949–1957).

Bruner, J. (1985). The role of interaction formats in language acquisition. In *Language and social situations* (pp. 31–46). Springer.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. In *ArXiv preprint arXiv:1812.08008*.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. (2017). Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, *372*(1711), 20160055.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, *99*(14), 9602–9605.

Fausey, C. M., Jayaraman, S., & Smith, L. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107.

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2017). See and be seen: Infant–caregiver social looking during locomotor free play. *Developmental Science*.

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.

Frank, M. C., Simmons, K., Yurovsky, D., & Pusiol, G. (2013). Developmental and postural changes in children's visual access to faces. In *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 454–459).

Gredeback, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of infants' gaze as they view adult shifts in overt attention. *Infancy*, *13*(5), 533–543.

James, W. (1890). *The principles of psychology* (Vol. 1). Macmillan London.

Jayaraman, S., & Smith, L. (2018). Faces in early visual environments are persistent not just frequent. *Vision Research*.

Jayaraman, S., Fausey, C. M., & Smith, L. (2015). The faces in infant-perspective scenes change over the first year of life. *PLoS One*. https://doi.org/10.1371/journal.pone.0123780

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, *85*(4), 1503–1518.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proc. of the National Academy of Sciences*, *112*(41), 12663–12668.

Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments modulate children's visual access to social information. In *Proceedings of the 40th annual conference of the cognitive science society*.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Smith, L., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*, *14*(1), 9–17.

Smith, L., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407–419.

Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye: Typical, daily exposure to faces documented from a first-person infant perspective. *Developmental Psychobiology*, *56*(2), 249–261.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. (under review). Head cameras on children aged 6 months through 31 months.

Yoshida, H., & Smith, L. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy*, *13*, 229–248.

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

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.100	0.014	1.955	7.049	0.021
Age	-0.002	0.001	386.972	-4.461	0.000

Table 1

Model coefficients from a linear mixed model predicting the proportion of faces seen by infants in the center FOV.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.166	0.011	2.082	15.340	0.004
Age	-0.002	0.001	380.308	-3.114	0.002

Table 2

Model coefficients from a linear mixed model predicting the proportion of hands seen by infants in the center FOV.

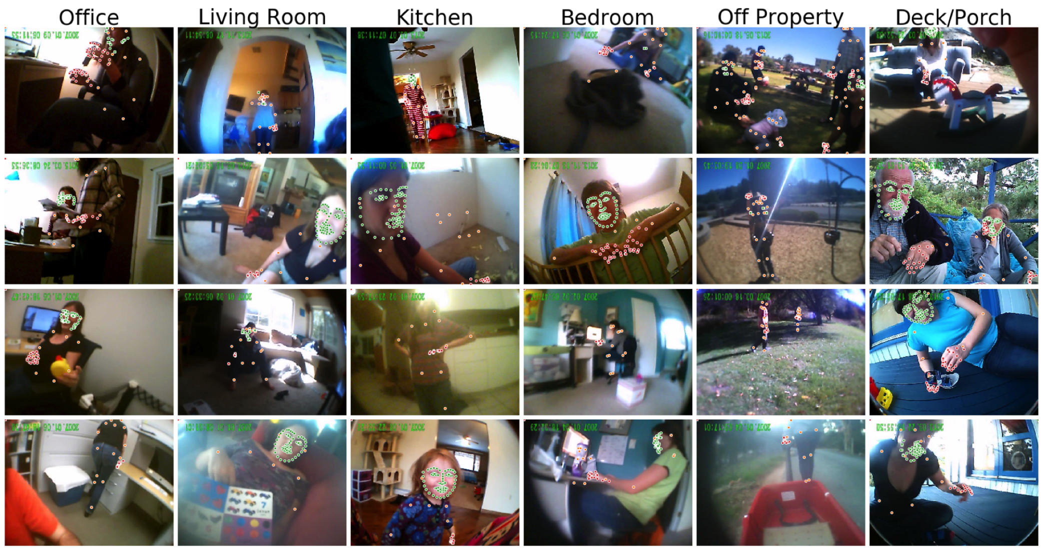

Figure 1. Example frames taken from the dataset, illustrating variability in the infant perspective across different locations. OpenPose detections are shown overlaid on these images (green dots = face, red dots = hands, orange dots = pose).

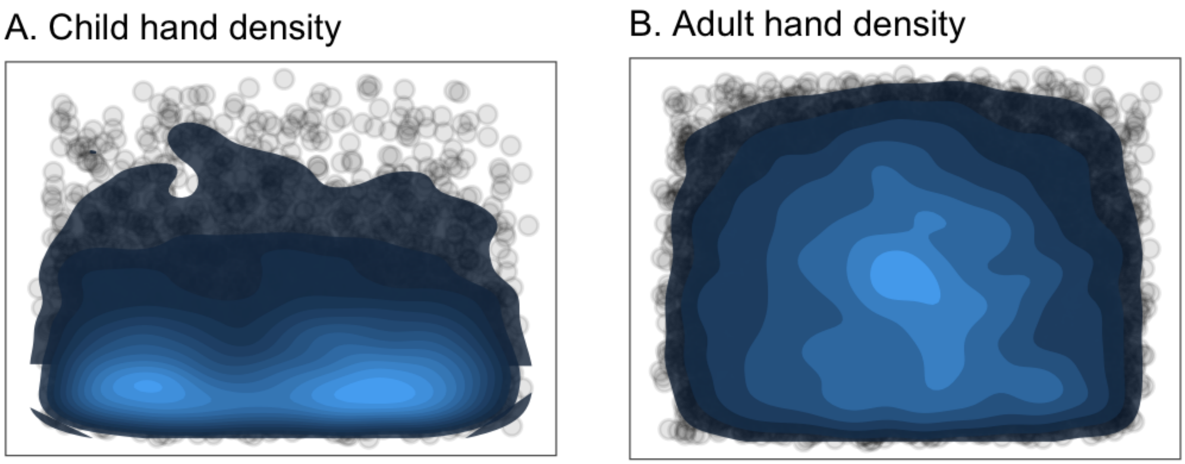

Figure 2. Density estimates for the child (left) and adult (right) hands that were detected in the 24K frame random gold set; each dot represents the center of a bounding box made by an adult participant. Brighter values indicate more detections.

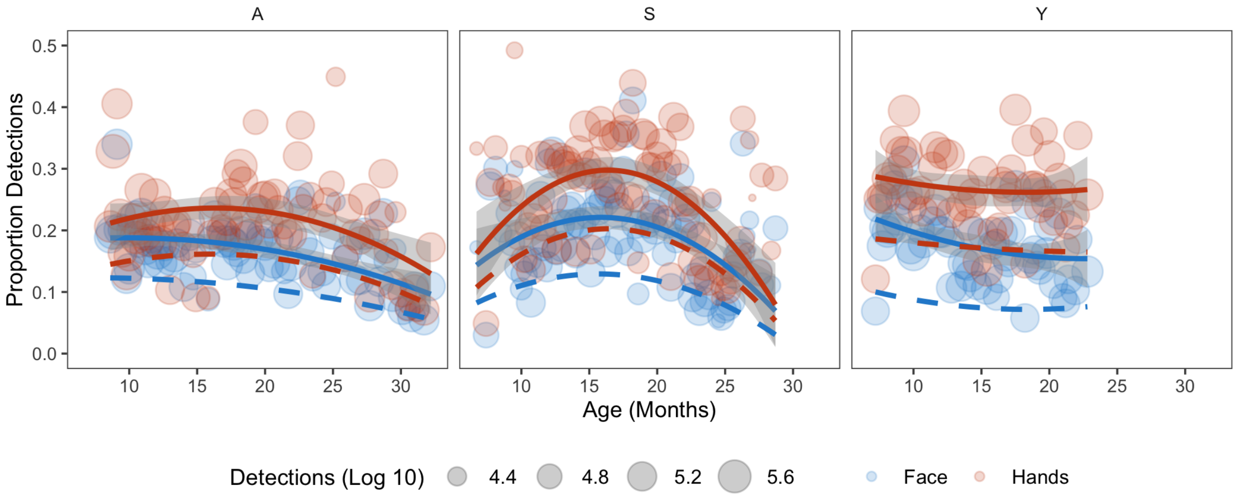

Figure 3. Proportion of faces and hands seen as a function of age for each child in the dataset. Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range. Dashed lines show estimated trend lines from proportion of faces/hands in view when detections are restricted to the center FOV, reducing the contribution of children's own hands.

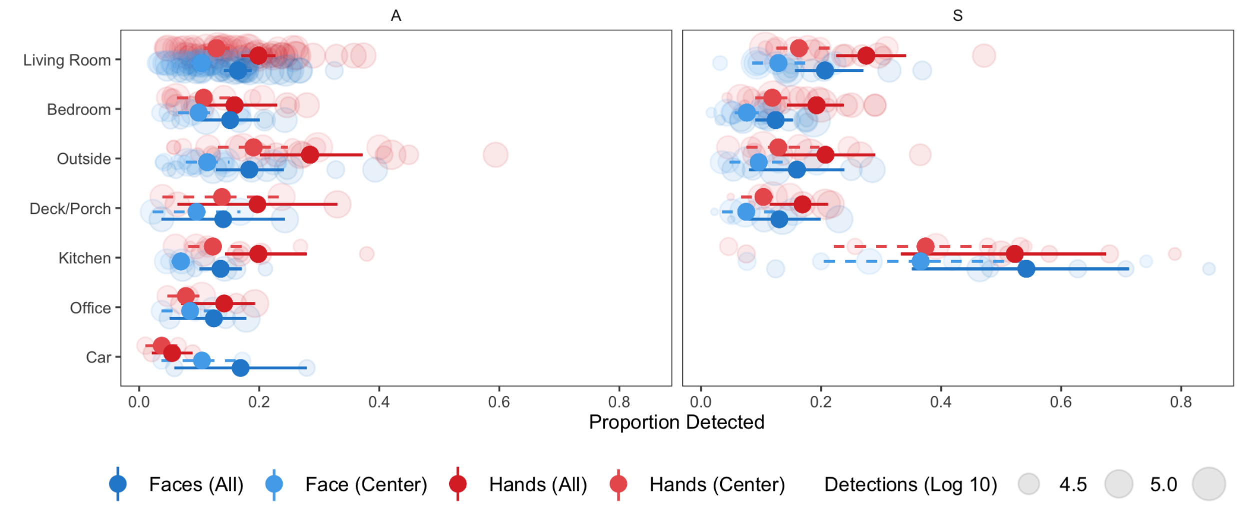

Figure 4. Proportion of faces and hands by location in which egocentric videos were filmed; each panel represents data from an individual child (location annotations were not yet available for Y). Each dot represents data from a week in which videos were filmed and are scaled by the number of frames.