A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections

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 is likely 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 more than 1700 headcam videos from three children ages 6 to 32 months. To analyze these naturalistic infant egocentric videos, we validated the use of a modern convolutional neural network (OpenPose) for the detection of faces and hands and then applied this model to the entire dataset. First, we found a higher proportion of hands in view than previously reported and a moderate decrease in the proportion of faces in children's view across age. Second, we found substantial variability in the proportion of faces and 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. Third, we found evidence that children may see closer, larger views of people, hands, and faces earlier in development. These analyses provide new insight into the changes in the social information in view across the first few years of life and call for further work that examines their generalizability across populations and their relationship to learning outcomes.

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

Word count: 4681

A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections

Introduction

Infants are confronted by a blooming, buzzing onslaught of stimuli (James, 1890) that 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 hard to intuit (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 (Franchak et al., 2011; Smith, Jayaraman, Clerkin, & Yu, 2018; Smith, Yu, Yoshida, & Fausey, 2015) and the consequences of this changing view for early learning. 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, Fausey, & Smith, 2013, 2015, 2017; Jayaraman & Smith, 2018; Sugden, Mohamed-Ali, & Moulson, 2014).

Beyond these early months, infants' motor and cognitive abilities mature, leading to vastly different perspectives on the world (Iverson, 2010). For example, children see fewer faces and hands when crawling than walking or sitting (Franchak, 2019; Franchak, Kretch, &

Adolph, 2017; Kretch, Franchak, & Adolph, 2014; Luo & Franchak, 2020; Sanchez, Long, Kraus, & Frank, 2018; Yamamoto, Sato, & Itakura, 2020) as well as different views of objects (Luo & Franchak, 2020; 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 et al., 2016a). 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 more than 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, 2021). 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; 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 et al., 2015; Yoshida & Smith, 2008), they may fail when faces or hands appear in children's peripheral vision but are still part of a joint interaction.

With hundreds of hours of footage (>42M 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. (2016a) 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. Finally, making use of automated annotation of pose bounding boxes, we analyze the size, location, and variability of detected faces and poses across development.

Method

Dataset

The dataset is described in detail in Sullivan et al. (2021); 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. We analyze 1745 videos, with a total duration of 391.11 hours (>42 million frames).

Detection Method

To annotate the millions of frames in SAYCam automatically, we used a pose detector, OpenPose[2] (Cao et al., 2018; Simon, Joo, Matthews, & Sheikh, 2017). The OpenPose system provides the locations of up to 18 body parts (ears, nose, wrists, etc.) from individual frames. OpenPose relies on a convolutional neural network for initial anatomical detection. It then uses part affinity fields for part association to produce a series of body part candidates. Once these body part candidates are matched to a single individual in the frame, they are finally assembled into a pose. While in this study we only measured face and hand presence, the entire set of pose information from an individual was used to determine the presence of a face/hand, making the process much more robust to occlusion than methods optimized to detect *only* faces or hands. Of course, these face/hand detections are nevertheless reliant on the detection of at least a partial pose, so some very up-close views of faces/hands may still go undetected.

[1] All videos are available at https://nyu.databrary.org/volume/564

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

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 (Amazon Mechanical Turk) additionally annotated 3150 frames; agreement with the primary coder was >95%. Upon manually inspecting these frames, we noticed that 1642 were sampled from videos taken from the allocentric perspective (i.e., not from the infant viewpoint); these frames and videos containing these frames were subsequently excluded from all other analyses.

As has been observed in other studies on automated annotation of headcam data (e.g. Frank, Simmons, Yurovsky, & Pusiol, 2013; Bambach, Lee, Crandall, & Yu, 2015; Long, Sanchez, Agrawal, Kraus, & Frank, n.d.; 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., Long et al., n.d.).

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 0.64, with a precision of 0.70 and recall of 0.58. For hands, the F-score was 0.51, with a precision of 0.73 and recall of 0.40. 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 also found that restricting our detections to high-confidence face/hand detections (>.5 confidence, default threshold for visualization in OpenPose) was not beneficial – improving precision but dramatically impairing recall and thus overall performance: the F-score for high-confidence face detections was 0.41, with a precision of 0.95 and recall of 0.26; for high-confidence hand detections, the F-score was 0.18, with a precision of 0.97 and recall of 0.10).

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 additional human annotations for the subsample of 9051 frames in the gold set frames where which a hand was present; participants (recruited via Amazon Mechanical Turk) were asked to draw bounding boxes around children's and adult's hands. Overall, we found that 43% 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 0.57. 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 Appendix Figure B1.

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 and Discussion

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 detected for each child. All analyses and preprocessed data files for this paper are available at tinyurl.com/longitudinal-social-info. 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 2), in keeping with prior findings (Fausey et al., 2016a); in contrast, we did not observe an increase in the proportion of hands in view. These effects were quantified with two separate linear mixed-effect models (see Tables 1 & 2).[3] After visualizing the data (see Figure 2A), we examined whether the addition of quadratic terms relating children's age to the proportion of faces/hands detected would provide better fit to the data than linear terms alone, and found that this was true in both cases (see Tables 1 & 2), though the linear term was also significant for faces. Thus, these exploratory results point towards the idea that some children may experience overall more social information in view in the second year of life.

However, the most striking result from these analyses is a much greater overall proportion of hands in view than has previously been reported (Fausey et al., 2016a). We found this observation to be true across all ages, in all three children, and regardless of

[3]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.

whether we analyzed human annotations (on the 24K random subset, see dotted lines in Appendix Figure A1) or OpenPose annotations on the entire dataset (see Figure 2A). This finding is notable especially given that OpenPose showed relatively low recall for hands, indicating that our measurements may in fact be an underestimate of the proportion of hands in view. In fact, analysis of the human gold standard annotations revealed a much higher proportion of hands relative to faces than the automated annotations.

One reason we could have observed more hands in view than previous studies 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. (2016a) 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 we found that children's hands tended to occur in the lower visual field (see Figure B1), 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%, and decreasing face detections from 20% to 9.90%. This cropping likely removed both the majority of detections of children's own hands but also some detections of adult hands (see Figure B1), 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 2). We also still found a higher proportion of hands in view relative to faces when excluding any frames containing child hand's from the human annotated gold sample (see Appendix Figure A1).

As we found that children's hands tended to occur in the lower visual field (see Figure Appendix B1), 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%, and face detections from 20% to 9.90%. This cropping likely removed both the majority of detections of children's own hands but also some detections of adult hands (see Figure B1), 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 2B). We also still found a higher proportion of hands in view relative to faces when excluding any frames containing child hand's from the human annotated gold sample (see Appendix Figure A1).

Finally, we analyzed how these two sources of social information co-occurred. To do so, we calculated the number of frames in which infants saw faces and hands together relative to overall proportions of faces/hands that were detected for each child and age range. Faces and hands were jointly present in 11.50 percent of frames (see face hand-occurrences across age in Figure 2C). As shown in Figure 3, all three infants were more likely to see hands independently – without the presence of a face – than they were likely to see faces independently. That is, generally speaking when a face was present, a hand also tended to be present.

Variability in social information across learning contexts

How does the child's context influence the social information in view? Bruner (1985) discussed the role of children's activities in shaping the information present for learning. Following this idea, we investigated whether there were differences in access to faces by the activity that the child was engaged in. This hypothesis seems intuitively appealing. Some activities seem likely to be characterized by a much higher proportion of faces (e.g., diaper changes) than others (e.g., a car trip). Following this same idea, perhaps other activities involve the presence of more hands in the field of view (e.g., playtime).

We did not have access to annotations of activity. Thus, following Roy et al. (2015), we used spatial location as a proxy for activity context, taking advantage of the presence of these annotations for a subset of the SAYCam videos. Of the 1745 videos in the dataset, 639 were annotated for the location or locations they were filmed in. These location annotations were only available for two children, S and A. Annotated locations mostly consisted of rooms of the house (e.g., "living room") but also included some other locations (e.g., "car," "outside"). Of this set, 296 videos were filmed in only a single location (e.g., the location label did not change within the video), representing 17 percent of the dataset and over 5 million frames. In our viewing of the SAYCam videos and in other annotations available with the dataset, activities varied somewhat predictably by location: for example, eating tended to occur in the kitchen, whereas playtime was the dominant activity in the living room.

Figure 4 shows the proportion of faces vs. hands across locations. We found substantial variation across 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, with all ps $< .001$. For example, while both A and S saw a relatively similar proportion of faces and hands in the bedroom, the two children saw quite different amounts of faces and hands from one another in the kitchen. This difference is likely explained by differences in arrangement of the kitchen in the two children's households (Sullivan, personal communication), such that mealtimes in one kitchen resulted in a face-to-face orientation while they did not in the other. This example illustrates how specifics of the geometry of a particular context can play an outsize role in the child's access to social information during that context.

Fine-grained changes in the social information in view

In a third set of analyses, we explored fine-grained changes in the SAYCam infants' access to social information across development. In these analyses, we capitalize on the fact that OpenPose provides not only face and hand detections but also positional keypoints. In particular, we explored this keypoint dataset with the idea that greater mobility allows older children to be further from their caregivers on average. Thus, younger, less mobile children may tend to see larger faces towards the center of their visual field while older, more mobile children may experience more smaller, more variable views of faces. The same dynamic would be predicted hold for hands as well, as it would be driven by overall differences in distance.

Supporting this idea, we found that the averages sizes of the people, faces, and hands in the infant view became smaller over development (Figure 5). This effect was relatively consistent across the three children in the dataset, despite the fact that the three children showed sometimes disparate overall proportions of faces/hands in view. Thus, children may see closer, larger views of people, hands, and faces earlier in development.

In keeping with this hypothesis, we also found evidence that faces tended to be farther away from older children. We restricted our analysis here to faces where both eyes were detected and computed interpupillary distance as a rough metric of distance, since eyes should be closer together on average when a face is further from the camera. Figure 6A shows the average interpupillary distance on faces as a function of each child's age at the time of recording. There was a trend from larger, closer faces (with a larger interpupillary distance) to smaller faces that were farther away (with a smaller interpupillary distance).

Finally, we also examined whether there were changes in where faces tended to appear in the camera's (and hence, by proxy, the child's) field of view. As expected, faces tended to be located towards the upper field of view, while views of hands were more centrally distributed (see Appendix, Figure C1 for average density distributions). However, we also found evidence that older children tended to see more faces in more variable positions than younger children. Specifically, we examined how variable the horizontal and vertical coordinates were of the faces in the infant view. To do so, we calculated the coefficient of variation of the horizontal (x) and vertical (y) positions of centers of the faces detected by

OpenPose (see Figure 6B), and examined changes across age. Faces tended to be more variable in the vertical than their horizontal position (see Figure 6B). We also found that as children got older, they tended to see faces that varied more in their horizontal – but not their vertical position – suggesting that older children might be more likely to see more smaller faces in their periphery (see Figure 6B).

General Discussion

Here, we analyzed the social information in view in a dense, longitudinal dataset, applying a modern computer-vision model to quantify the hands and faces seen from each of three children's egocentric perspective from 6 to 32 months of age.

First, 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., 2016a; Jayaraman et al., 2015). This finding is particularly notable given that, in previous cross-sectional data, this effect seems to be most strongly driven by infants younger than 4 months of age (e.g., Fausey et al., 2016a; Jayaraman et al., 2015; Sugden et al., 2014) who see both more frequent and more persistent faces (Jayaraman & Smith, 2018). We also found this to be true when restricting our analyses to full-field faces, suggesting this effect is not driven by a concurrent shift from more full-view to partial-views of faces.

We also found an unexpectedly high proportion of hands in infants' view, even when restricting the field-of-view to the center field-of-view to make the viewpoints comparable to those of headcams used in prior work. 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.

We also used these detailed pose annotations to explore finer-grained changes in how children experience the faces and hands of their caregivers over development. We found that the faces, hands, and people in the infant view tended to become smaller and that faces tended to be farther away and in more variable horizontal positions, in keeping with prior work examining the sizes of faces in the infant view during the first year of life (Jayaraman et al., 2015). Overall, these data support the idea that the social information in view changes across development as infants become increasingly mobile and independent (Fausey et al., 2016b; Franchak et al., 2017). As children explore the world on their own (Xu, 2019), they may experience fewer close-up interactions with their caregivers and more bouts of play where they are exploring the objects in their environment.

More broadly, however, 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 in many ways of the global population (Henrich, Heine, & Norenzayan, 2010; Karasik, Tamis-LeMonda, Ossmy, & Adolph, 2018). Any idiosyncrasies in how and when these particular families chose to film these videos also undoubtedly influenced the variability seen here, and may contribute to the individual differences between the three children in this dataset. And without eye-tracking data, we do not know the extent to which children are attending to the social information in view.

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, >42M million frames) for the presence and size of people, hands, and faces, representing orders of magnitude more data relative to human annotations in prior work. While OpenPose did not provide annotations that were as accurate as those provided by human annotators, we found relatively consistent results with prior literature, suggesting that the sheer scale and density of the annotations provided by this method may overcome some of its limitations.

In future work, the adaptation of deep neural networks for the infant egocentric view remains a promising avenue for collaboration between computer vision experts and developmental psychologists. Indeed, this combination has already yielded new insights about the learning mechanisms needed to build visual representations (Orhan, Gupta, &

Lake, 2020; Tsutsui, Chandrasekaran, Reza, Crandall, & Yu, 2020; Zhuang, She, Andonian, Mark, & Yamins, 2020). We propose that the use of novel algorithms with 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

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. B. (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. B. (2016a). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107.

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

Franchak, J. M. (2019). Changing opportunities for learning in everyday life: Infant body position over the first year. *Infancy*, *24*(2), 187–209.

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.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29.

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language*, *37*(2), 229–261.

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

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2013). Visual statistics of infants' ordered experiences. *Journal of Vision*, *13*(9), 735–735.

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

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2017). Why are faces denser in the visual experiences of younger than older infants? *Developmental Psychology*, *53*(1), 38.

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

Karasik, L. B., Tamis-LeMonda, C. S., Ossmy, O., & Adolph, K. E. (2018). The ties that bind: Cradling in tajikistan. *PloS One*, *13*(10), e0204428.

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

Long, B., Sanchez, A., Agrawal, K., Kraus, A. M., & Frank, M. C. (n.d.). Automated detections reveal the social information in the changing infant view.

Luo, C., & Franchak, J. M. (2020). Head and body structure infants' visual experiences during mobile, naturalistic play. *Plos One*, *15*(11), e0242009.

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *arXiv Preprint arXiv:2007.16189*.

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. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, *22*(4), 325–336.

Smith, L. B., 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. B., 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. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*.

Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A computational model of early word learning from the infant's point of view. *arXiv Preprint arXiv:2006.02802*.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, *126*(6), 841.

Yamamoto, H., Sato, A., & Itakura, S. (2020). Transition from crawling to walking changes gaze communication space in everyday infant-parent interaction. *Frontiers in Psychology*, *10*, 2987.

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

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).

Zhuang, C., She, T., Andonian, A., Mark, M. S., & Yamins, D. (2020). Unsupervised learning from video with deep neural embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9563–9572).

Table 1

Coefficients from a mixed-effects regression predicting the proportion of faces seen by infants in the center FOV.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.098	0.011	1.953	8.850	0.013
Age	-0.195	0.060	429.926	-3.257	0.001
Age**2	-0.160	0.059	429.032	-2.708	0.007

Table 2

Coefficients from a mixed-effects regression predicting the proportion of hands seen by infants in the center FOV.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.161	0.007	1.828	21.906	0.003
Age	-0.145	0.078	422.334	-1.855	0.064
Age**2	-0.319	0.077	429.968	-4.134	<.001

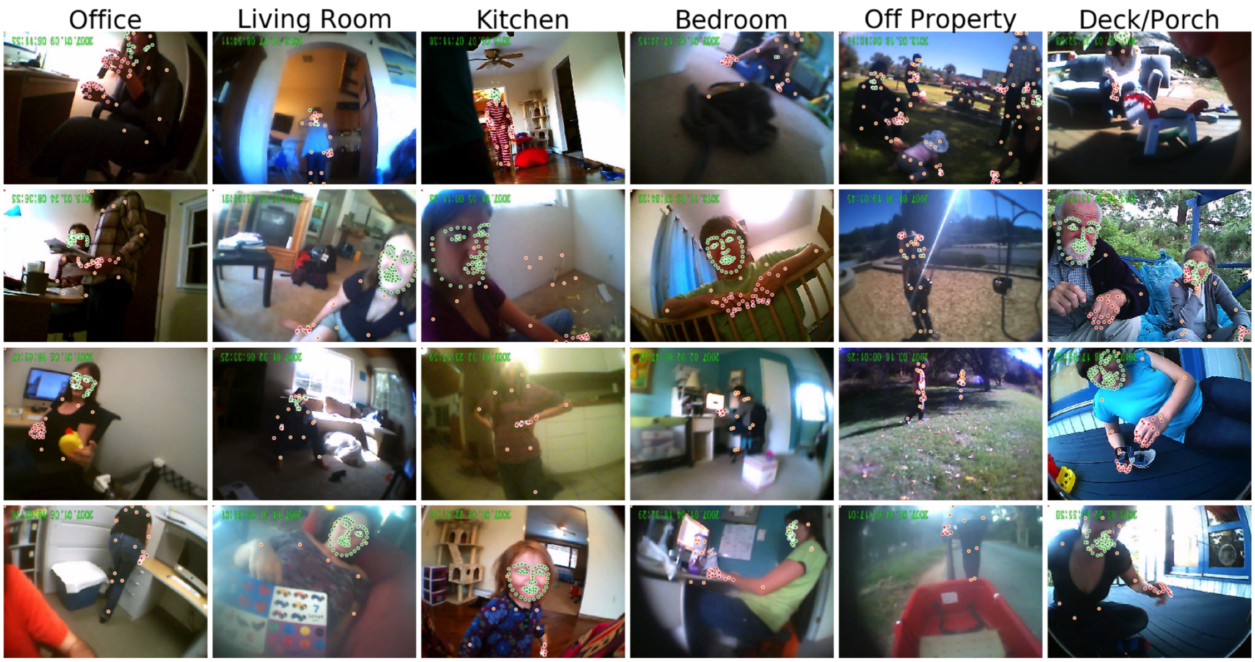

Figure 1. Example frames taken from the dataset, illustrating variability in the infant perspective across different locations. OpenPose detections are shown overlaid on these images (green dots = face, red dots = hands, orange dots = pose).

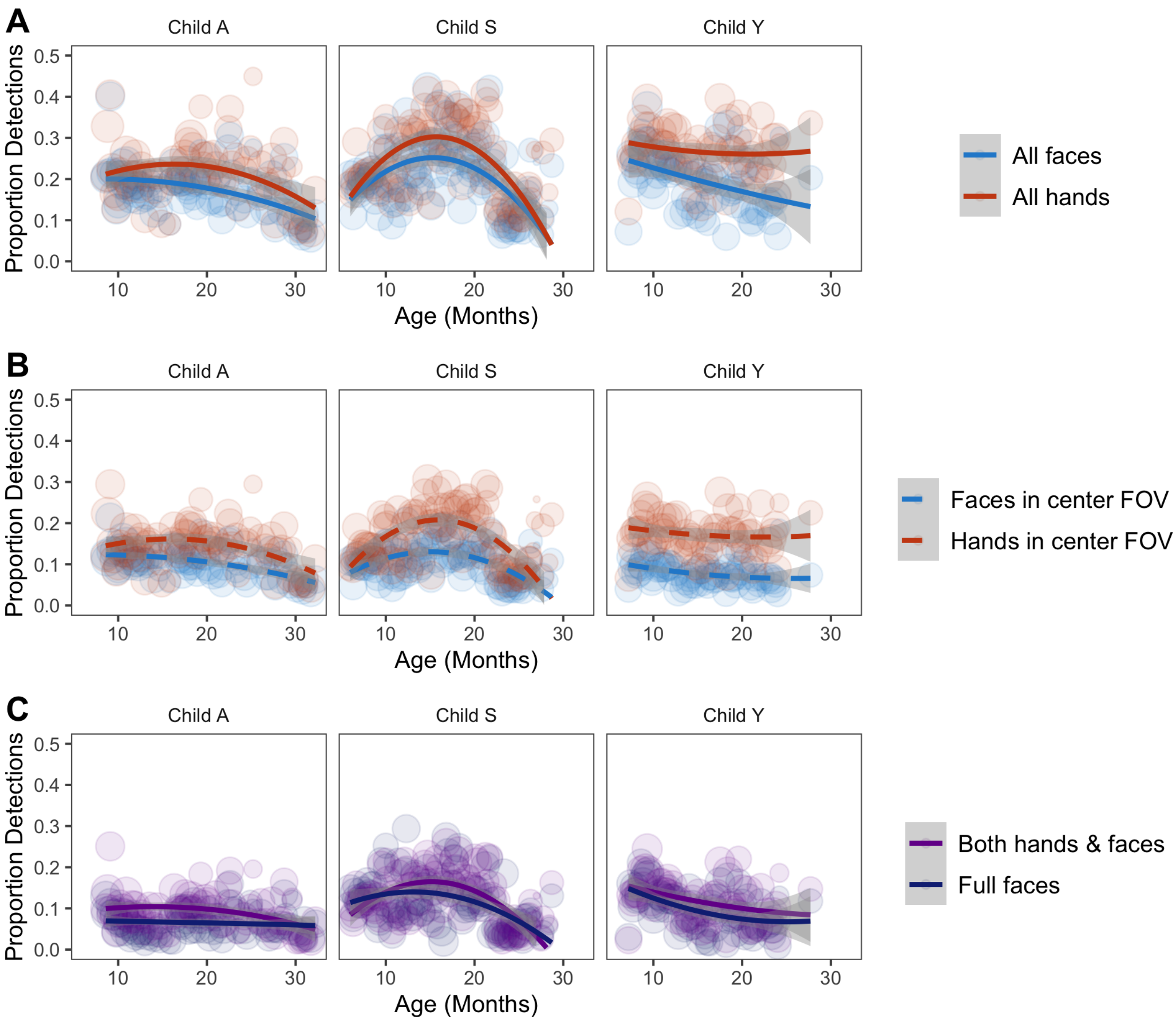

Figure 2. Proportion of frames with (A) All face and hand detections, (B) Face/hand detections that fell within the center field-of-view (reducing the contribution of children's own hands) and (C) Face detections that were full faces (e.g., eyes, nose, and mouth all visible) and that co-occurred with hands, plotted as a function of age for each child (A, S, and Y). Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range.

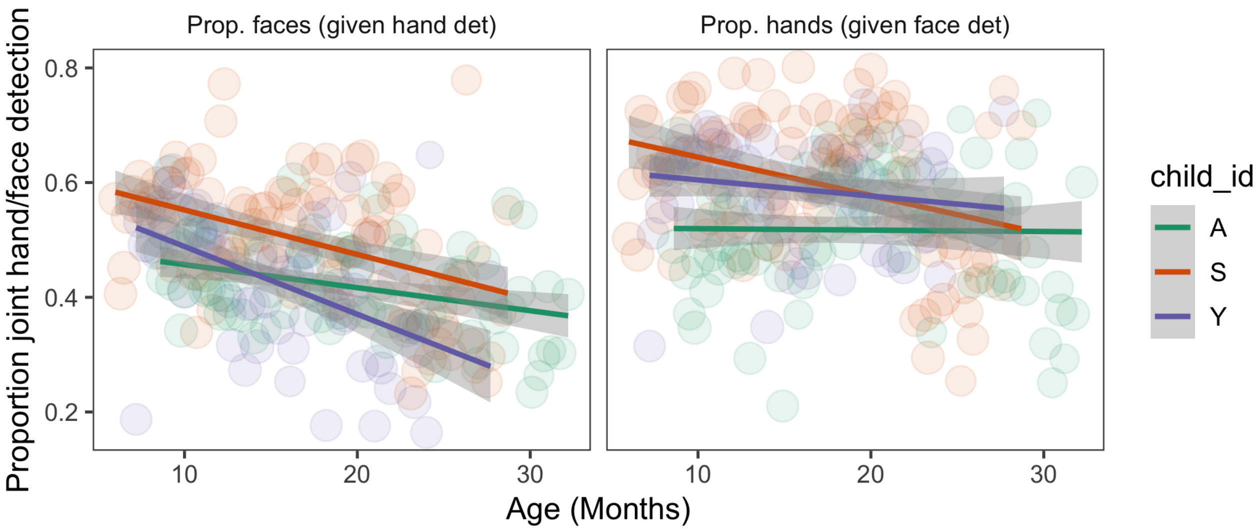

Figure 3. Proportion of joint face and hands detection within frames where hands (left) or faces (right) were detected.

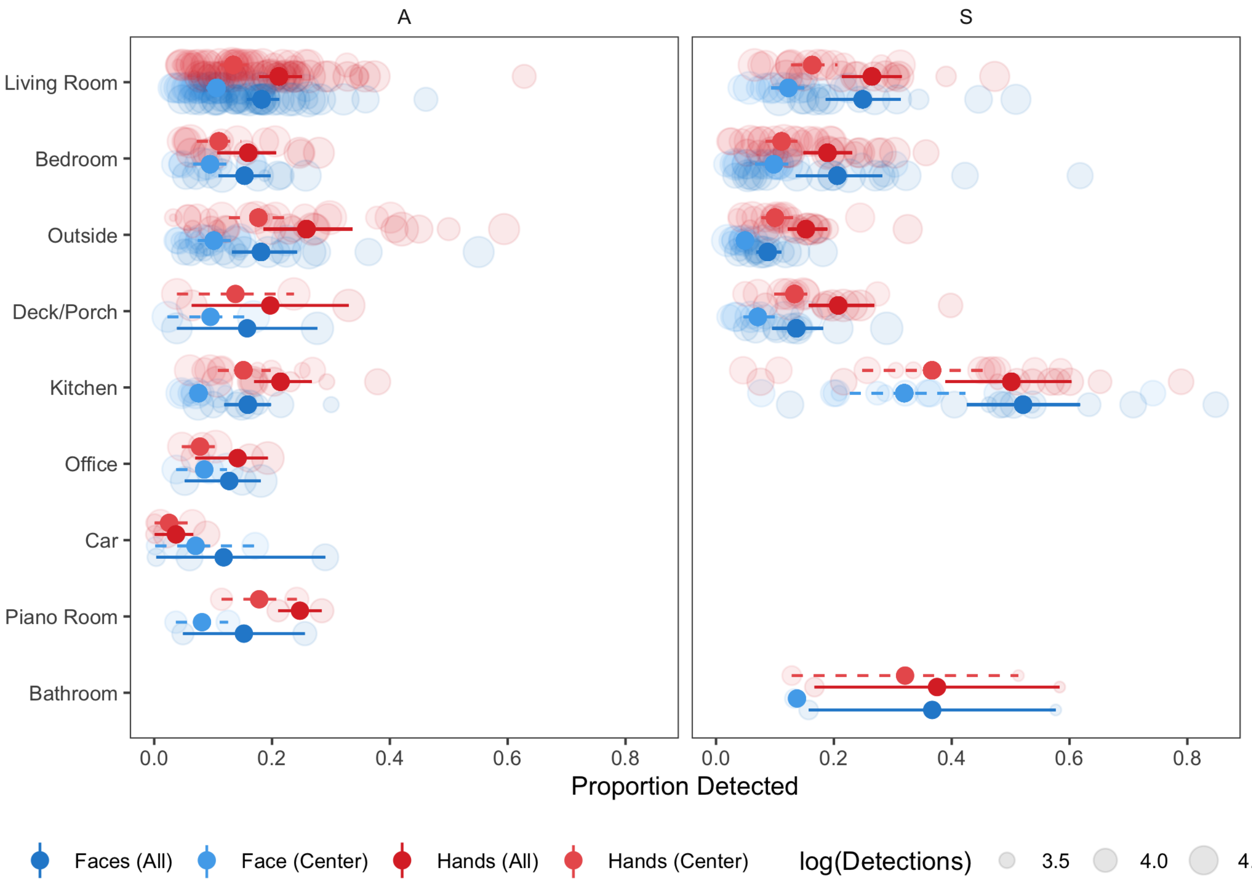

Figure 4. Proportion of faces and hands by location in which egocentric videos were filmed; each panel represents data from an individual child (location annotations were not yet available for Y). Each dot represents data from a week in which videos were filmed and are scaled by the number of frames.

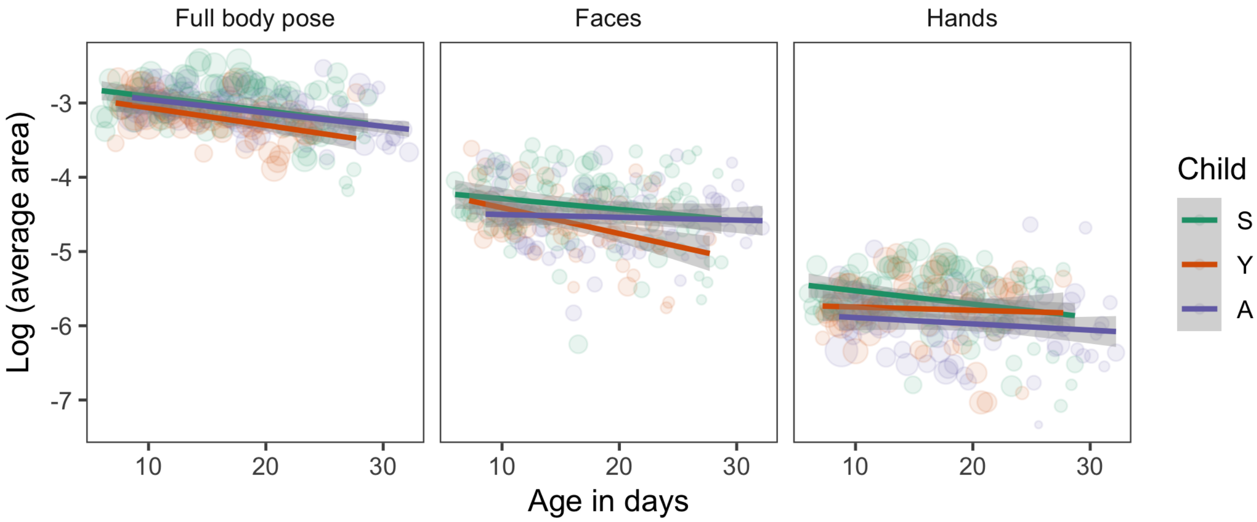

Figure 5. Average size of poses, faces, and hands detected in the dataset between eyes in faces detected as a function of age for each child in the dataset (each color = different child). Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range.

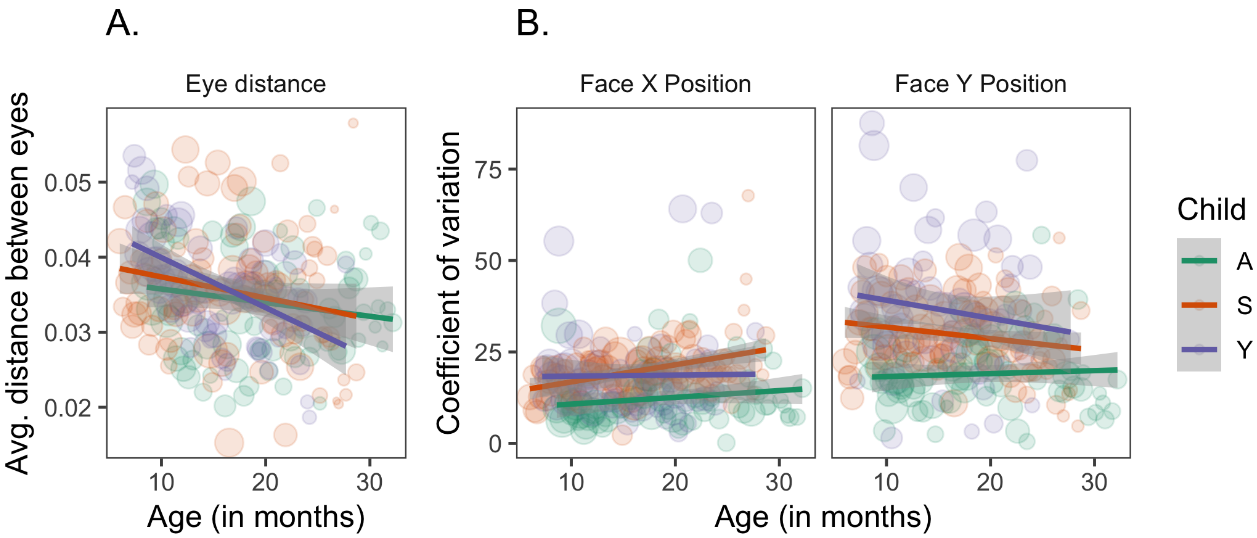

Figure 6. (A) Average distance between eyes and (B) average coefficient of variation for the x and y position of faces detected by OpenPose as a function of each child's age at the time of filming. Data in (A) are restricted to faces where both eyes were detected. Data are binned by each week that the videos were filmed and scaled by the number of face detections in that age range.

Appendix A

Face/hand detections relative to human annotations

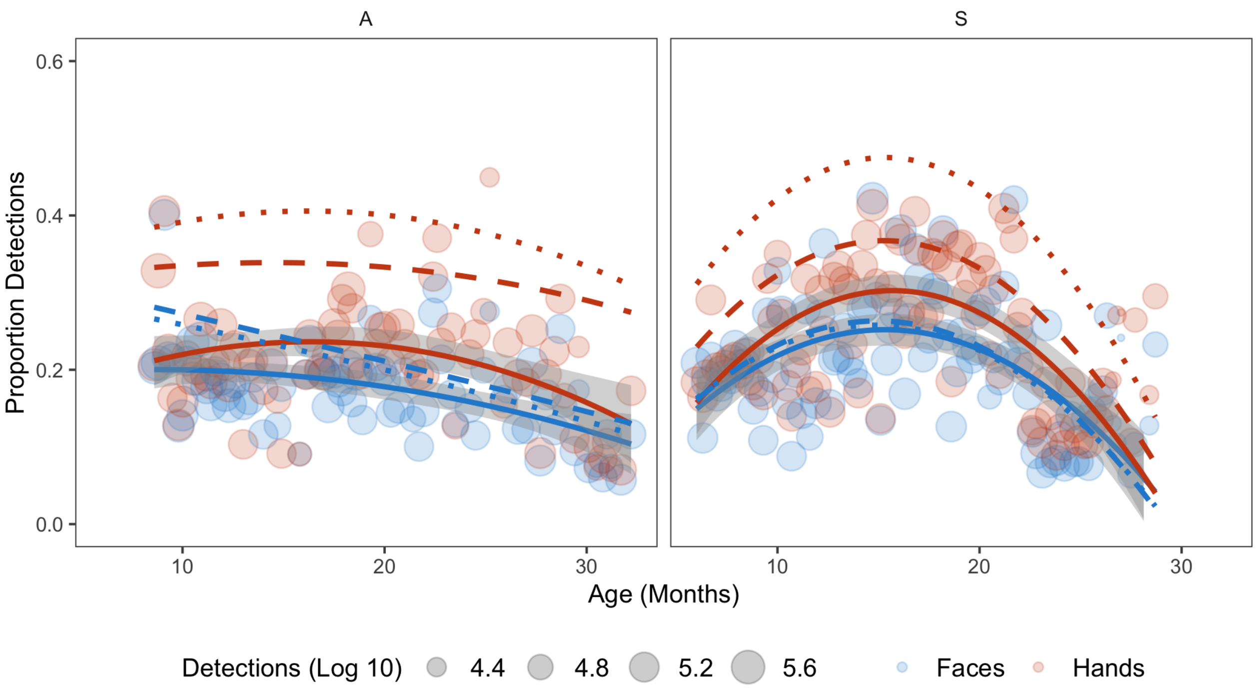

Figure A1. Proportion of faces and hands seen as a function of age for each child in the dataset. Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range. Dashed lines show estimated trend lines from proportion of faces/hands in view when analyzing the gold set of frames made by human annotators. Dotted lines show trend lines from the goldset when frames when children's own hand were detected.

Appendix B

Density of child vs. adults hands in the visual field

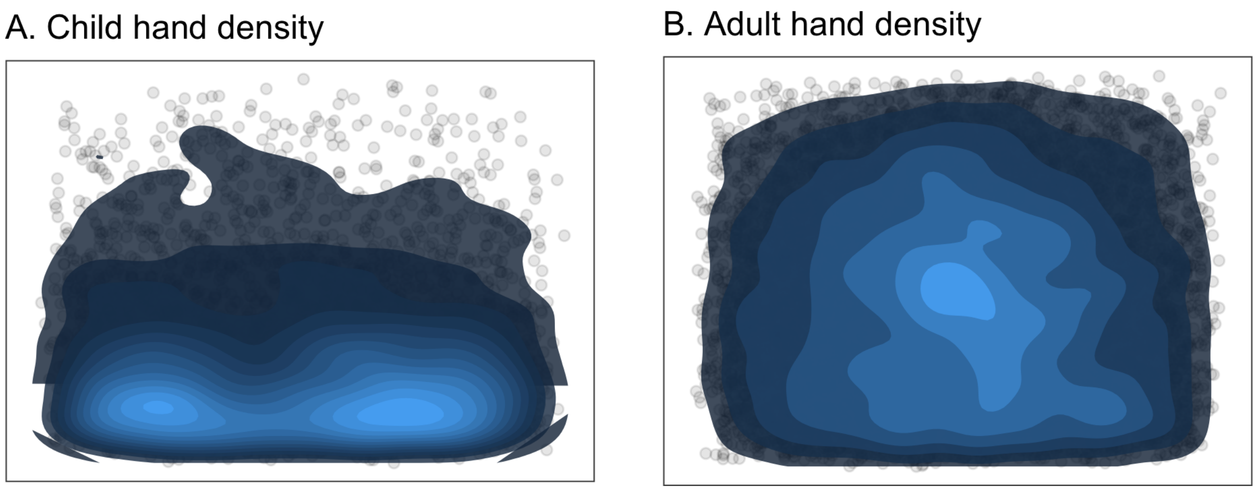

Figure B1. Density estimates for the child (left) and adult (right) hands that were detected in the 24K frame random gold set; each dot represents the center of a bounding box made by an adult participant. Brighter values indicate more detections.

Appendix C

Distribution of faces and hands in the visual field

We explored where in the visual field children tended to see faces and hands, suspecting that these distributions might become wider as children grow older and learn to locomote on their own, following preliminary analyses from Frank (2012). As expected, faces tended to appear in the upper visual field in contrast to hands, which tended to be more centrally located (see Figure C1). However, we found little evidence for any changes in the positions of faces and hands across age, suggesting that this is a relatively stable property of infants' visual environment from 6 months of age.

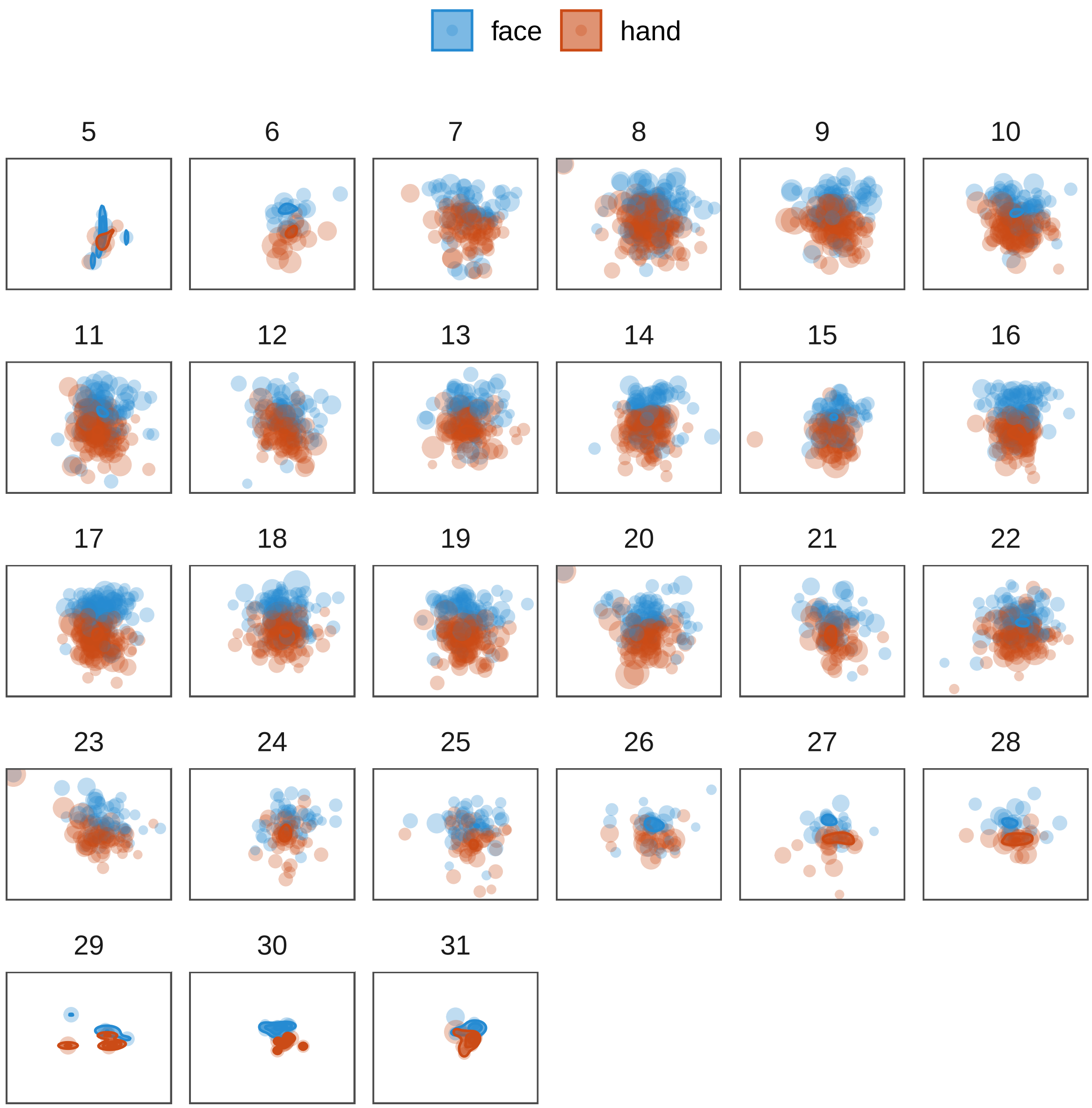

Figure C1. Each panel shows the average position of faces and hands in the visual field; each dot represents the average position from one video within a given age range.