

¹ Detecting social information in a dense database of infants' natural visual experience

² Bria L. Long¹, George Kachergis¹, Ketan Agrawal¹, & Michael C. Frank¹

³ ¹ Department of Psychology, Stanford University

⁴ Author Note

⁵ Correspondence concerning this article should be addressed to Bria L. Long, 450 Serra
⁶ Mall, Stanford CA 94305. E-mail: bria@stanford.edu

7 Abstract

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

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

23 Word count: 3444

²⁴ Detecting social information in a dense database of infants' natural visual experience

²⁵ Introduction

²⁶ Infants are confronted by a blooming, buzzing onslaught of stimuli (James, 1890) which
²⁷ they must learn to parse to make sense of the world around them. Yet they do not embark
²⁸ on this learning process alone: From as early as 3 months of age, young infants follow overt
²⁹ gaze shifts (Gredeback, Theuring, Hauf, & Kenward, 2008), and even newborns prefer to
³⁰ look at faces with direct vs. averted gaze (Farroni, Csibra, Simion, & Johnson, 2002), despite
³¹ their limited acuity. As faces are likely to be an important conduit of social information that
³² scaffolds cognitive development, psychologists have long hypothesized that faces are
³³ prevalent in the visual experience of young infants.

³⁴ Yet until recently most hypotheses about infants' visual experience have gone untested.
³⁵ Though parents and scientists alike have strong intuitions about what infants see, even the
³⁶ viewpoint of a walking child is not easily predicted by these intuitions (Clerkin, Hart, Rehg,
³⁷ Yu, & Smith, 2017; Franchak, Kretch, Soska, & Adolph, 2011). By equipping infants and
³⁸ toddlers with head-mounted cameras, researchers have begun to document the infant's
³⁹ egocentric perspective on the world. Using these methods, a growing body of work now
⁴⁰ demonstrates that the viewpoints of very young infants (less than 4 months of age) are
⁴¹ indeed dominated by frequent, persistent views of the faces of their caregivers (Jayaraman &
⁴² Smith, 2018; Jayaraman, Fausey, & Smith, 2015; Sugden, Mohamed-Ali, & Moulson, 2014).

⁴³ Beyond these early months, infants' motor and cognitive abilities mature, leading to
⁴⁴ vastly different perspectives on the world. For example, crawlers see fewer faces and hands
⁴⁵ than do walking children (Franchak, Kretch, & Adolph, 2017; Kretch, Franchak, & Adolph,
⁴⁶ 2014; Sanchez, Long, Kraus, & Frank, 2018) as well as different views of objects (Smith, Yu,
⁴⁷ & Pereira, 2011). Further, as infants learn to use their own hands to act on the world, they
⁴⁸ seem to focus on manual actions taken by their social partners, and their perspective starts

49 to capture views of hands manipulating objects (Fausey, Jayaraman, & Smith, 2016). In
50 turn, caregivers may also start to use their hands with more communicative intent, directing
51 infants' attention by pointing and gesturing to different events and objects during play (Yu
52 & Smith, 2013).

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

69 With hundreds of hours of footage (>40M frames), however, this large dataset
70 necessitates a shift to an automated annotation strategy. Indeed, annotation of the frames
71 extracted from egocentric videos has been prohibitively time-consuming, meaning that most
72 frames are typically not inspected, even in the most comprehensive studies. For example,
73 Fausey et al. (2016) collected a total of 143 hours of head-mounted camera footage (15.5
74 million frames), of which one frame every five seconds was hand-annotated (by four coders),

75 totalling 103,383 frames (per coder)—an impressive number of annotations but nonetheless
76 only 0.67% of the collected footage. To address this challenge, we use a modern computer
77 vision model of pose detection to automatically detect the presence of hands and faces from
78 the infant egocentric viewpoint. Specifically, we use OpenPose (Cao, Hidalgo, Simon, Wei, &
79 Sheikh, 2018), a model optimized for jointly detecting human face, body, hand, and foot
80 keypoints that operates well on scenes including multiple people, even if they are
81 partially-occluded (see Figure 1). In prior work examining egocentric videos, OpenPose
82 performed comparably to other modern face detection models (Sanchez et al., 2018).

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

89 Method

90 Dataset

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

99 Videos¹ with technical errors or that were not taken from the egocentric perspective were
100 excluded from the dataset. While data collection for the third child (Y) is still ongoing, here
101 we analyze 1726 videos, with a total duration of 404.13 hours (>40 million frames).

102 Detection Method

103 To automatically annotate the millions of frames in SAYCam, we used a pose detector,
104 OpenPose² (Cao et al., 2018; Simon, Joo, Matthews, & Sheikh, 2017), which provided the
105 locations of 18 body parts (ears, nose, wrists, etc.). To do so, a convolutional neural network
106 was used for initial anatomical detection, and part affinity fields were subsequently applied
107 for part association to produce a series of body part candidates. Once these body part
108 candidates were matched to a single individual in the frame, they were finally assembled into
109 a pose. Thus, while we only made use of the outputs of the face and hand detections, the
110 entire set of pose information from an individual was used to determine the presence of a
111 face/hand, making the process more robust to occlusion than methods optimized to detect
112 only faces or hands. Note, however, that these face/hand detections are reliant on the
113 detection of at least a partial pose, so some very up-close views of faces/hands may go
114 undetected.

115 Detection Validation

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

¹All videos are available at <https://nyu.databrary.org/volume/564>

²<https://github.com/CMU-Perceptual-Computing-Lab/openpose>

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

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

133 To evaluate OpenPose’s performance, we compared its detections to the
134 manually-annotated gold set of frames, calculating precision (hits / hits + false alarms),
135 recall (hits / hits + misses), and F-score (the harmonic mean of precision and recall). In our
136 data, for faces, the F-score was 0.59, with a precision of 0.65 and recall of 0.54. For hands,
137 the F-score was 0.50, with a precision of 0.70 and recall of 0.38. While face and hand
138 detections showed moderately good precision, face detections were overall slightly more
139 accurate than hand detections. In general, hand detections suffered from fairly low recall,
140 indicating that OpenPose likely underestimated the proportion of hands in the dataset.

141 We suspected that this was in part because children’s own hands were often in view of
142 the camera and unconnected to a pose—a notoriously challenging detection problem
143 (Bambach et al., 2015). To assess this possibility, we obtained human annotations for the
144 entire subsample of 9051 frames in which a hand was detected; participants (recruited via
145 AMT) were asked to draw bounding boxes around children’s and adult’s hands. Overall, we

¹⁴⁶ found that 42% of missed hand detections were of child hands. When frames with children's
¹⁴⁷ hands were removed from the gold set, recall did improve somewhat to 0.54. We also
¹⁴⁸ observed that children's hands tended to appear in the lower half of the frames; heatmaps of
¹⁴⁹ the bounding boxes obtained from these annotations can be seen in Figure 2.

¹⁵⁰ Finally, we examined whether the precision, recall, and F-score for hands and faces
¹⁵¹ varied with age or child, and did not find substantial variation. Thus, while OpenPose was
¹⁵² trained on photographs from the adult perspective, this model still generalized relatively well
¹⁵³ to the egocentric infant viewpoint with no fine-tuning or post-processing of the detections.
¹⁵⁴ As these detections were imperfect compared to human annotators, fine-tuning these models
¹⁵⁵ to better optimize for the infant viewpoint remains an open avenue for future work.
¹⁵⁶ Standard computer vision models are rarely trained on the egocentric viewpoint, and we
¹⁵⁷ suspect that training these models on more naturalistic data may lead to more robust,
¹⁵⁸ generalizable detectors.

Results

¹⁶⁰ Access to social information across age

¹⁶¹ We analyzed the social information in view across the entire dataset, looking
¹⁶² specifically at the proportions of faces and hands that were in view for each child.³ Data
¹⁶³ from videos were binned according to the age of the child (in weeks). First, we saw that the
¹⁶⁴ proportion of faces in view showed a moderate decrease across this age range (see Figure 3);
¹⁶⁵ in contrast, we did not observe an increase in the proportion of hands in view, but rather a
¹⁶⁶ slight decrease. These effects were quantified with two separate linear mixed-models (see

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

167 Tables 1 & 2).⁴

168 However, the most striking result from these analyses is a much overall greater
169 proportion of hands in view than have previously been reported (Fausey et al., 2016). We
170 found this to be true across all ages, in all three children, and regardless of whether we
171 analyzed human annotations (on the 24K random subset, see dotted lines in Figure 3) or
172 OpenPose annotations on the entire dataset (see solid lines in Figure 3). This is notable
173 especially given that OpenPose showed relatively low recall for hands, indicating that this
174 may be an underestimate of the proportion of hands in view. Nonetheless, one reason this
175 could be the case is the much larger field of view that was captured by the cameras used in
176 this study: These cameras were outfitted with a fish-eye lens in an attempt to capture as
177 much of the children's field of view as possible, leading to a larger field of view (109 degrees
178 horizontal x 70 degrees vertical) than in many previous studies; for example, in Fausey et al.
179 (2016) the FOV was 69 x 41 degrees. This larger FOV may have allowed the SAYCam
180 cameras to capture not only the presence of a social partner's hands interacting with objects
181 or gestures, but also the children's own hands, leading to more frequent hand detections.

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

⁴Face/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.

190 Access to social information in different locations

191 How does variability across different contexts influence the social information in the
192 infant view? Intuitively, some activities in different contexts may be characterized by a much
193 higher proportion of faces (e.g., diaper changes in bedrooms) than others (e.g., playtime in
194 the living room). We thus next examined variation in presence of hands and faces across
195 different locations. Of the 1726 videos, 639 were annotated (Sullivan et al., under review) for
196 the location they were filmed in. Of these, 296 videos were filmed in single location,
197 representing 17 percent of the dataset and over 5 million frames. Activities varied somewhat
198 predictably by these contexts: for example, eating tended to occur in the kitchen, whereas
199 playtime was the dominant activity in the living room. Overall, we found that the
200 proportion of faces vs. hands varied across filming locations, and, to some extent, across
201 children; separate chi-squared tests for each child and detection type revealed significant
202 variability in detections by location in each case.⁵ For example, while both A and S saw a
203 relatively similar proportion of faces vs. hands in the bedroom, they saw quite different
204 amounts of faces vs. hands in kitchens (see Figure 4).

205 General Discussion

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

⁵all $p < .001$, see accompanying codebase

²¹³ younger than 4 months of age (e.g., Fausey et al., 2016; Jayaraman et al., 2015; Sugden et
²¹⁴ al., 2014) who see both more frequent and more persistent faces (Jayaraman & Smith, 2018).

²¹⁵ We also found an unexpectedly high proportion of hands in the view of infants, even
²¹⁶ when restricting the field-of-view to the center field of view the videos to make the
²¹⁷ viewpoints comparable to those of headcams used in previous work (Fausey et al., 2016).
²¹⁸ Why might this be the case? One idea is that these videos contain the viewpoints of children
²¹⁹ not only during structured interactions (e.g., play sessions at home or in the lab) but during
²²⁰ everyday activities when children may be playing by themselves or simply observing the
²²¹ actions of caregivers and other people in their environment. During these less structured
²²² times, caregivers may move about in the vicinity of the child but not interact with them as
²²³ directly—leading to views where a person and their hands are visible from a distance, but
²²⁴ this person’s face may be turned away from the infant or occluded (see examples in Figure
²²⁵ 1). Indeed, using the same pose detector on videos from in-lab play sessions, Sanchez et al.
²²⁶ (2018) found the opposite trend: slightly fewer hand detections than face detections from
²²⁷ 8-16 months of age. Work that directly examines the variability in the social information in
²²⁸ view across more vs. less structured activity contexts could further test this idea.

²²⁹ A coarse analysis based on the location the videos were filmed in further highlights the
²³⁰ variability of the social information in view during different activities, showing differences
²³¹ across locations and between individual children. Within a given, well-defined context—e.g.,
²³² mealtime in kitchens—S saw more faces than A, and S saw more faces in the kitchen than in
²³³ other locations. This variability likely stems from the fact that there are at least three ways
²³⁴ to feed a young child: 1) sitting in front of the child, facing them as they sit in a high chair;
²³⁵ 2) sitting behind the child, holding them as they face outward, and 3) sitting side by side.
²³⁶ Each of these positions offer the child differing degrees of visual access to faces and hands.
²³⁷ While the social information in view may be variable across children in different activity
²³⁸ contexts, these analyses suggest they could be stable within a given child’s day-to-day

239 experience.

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

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

257 Acknowledgements

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

262

References

- 263 Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands
264 and recognizing activities in complex egocentric interactions. In *Proc. of the IEEE*
265 *international conference on computer vision* (pp. 1949–1957).
- 266 Bruner, J. (1985). The role of interaction formats in language acquisition. In *Language and*
267 *social situations* (pp. 31–46). Springer.
- 268 Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime
269 multi-person 2D pose estimation using Part Affinity Fields. In *ArXiv preprint*
270 *arXiv:1812.08008*.
- 271 Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. (2017). Real-world visual statistics
272 and infants' first-learned object names. *Phil. Trans. R. Soc. B*, 372(1711), 20160055.
- 273 Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in
274 humans from birth. *Proceedings of the National Academy of Sciences*, 99(14),
275 9602–9605.
- 276 Fausey, C. M., Jayaraman, S., & Smith, L. (2016). From faces to hands: Changing visual
277 input in the first two years. *Cognition*, 152, 101–107.
- 278 Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2017). See and be seen: Infant–caregiver
279 social looking during locomotor free play. *Developmental Science*.
- 280 Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye
281 tracking: A new method to describe infant looking. *Child Development*, 82(6),
282 1738–1750.
- 283 Frank, M. C., Simmons, K., Yurovsky, D., & Pusiol, G. (2013). Developmental and postural

- 284 changes in children's visual access to faces. In *Proceedings of the 35th annual meeting*
285 *of the cognitive science society* (pp. 454–459).
- 286 Gredeback, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of infants'
287 gaze as they view adult shifts in overt attention. *Infancy*, 13(5), 533–543.
- 288 James, W. (1890). *The principles of psychology* (Vol. 1). Macmillan London.
- 289 Jayaraman, S., & Smith, L. (2018). Faces in early visual environments are persistent not just
290 frequent. *Vision Research*.
- 291 Jayaraman, S., Fausey, C. M., & Smith, L. (2015). The faces in infant-perspective scenes
292 change over the first year of life. *PLoS One*.
293 <https://doi.org/10.1371/journal.pone.0123780>
- 294 Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see
295 the world differently. *Child Development*, 85(4), 1503–1518.
- 296 Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of
297 a spoken word. *Proc. of the National Academy of Sciences*, 112(41), 12663–12668.
- 298 Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments
299 modulate children's visual access to social information. In *Proceedings of the 40th*
300 *annual conference of the cognitive science society*.
- 301 Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single
302 images using multiview bootstrapping. In *CVPR*.
- 303 Smith, L., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler
304 visual experience. *Developmental Science*, 14(1), 9–17.
- 305 Smith, L., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted

306 cameras to studying the visual environments of infants and young children. *Journal*
307 *of Cognition and Development*, 16(3), 407–419.

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

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

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

315 Yu, C., & Smith, L. (2013). Joint attention without gaze following: Human infants and their
316 parents coordinate visual attention to objects through eye-hand coordination. *PloS*
317 *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.

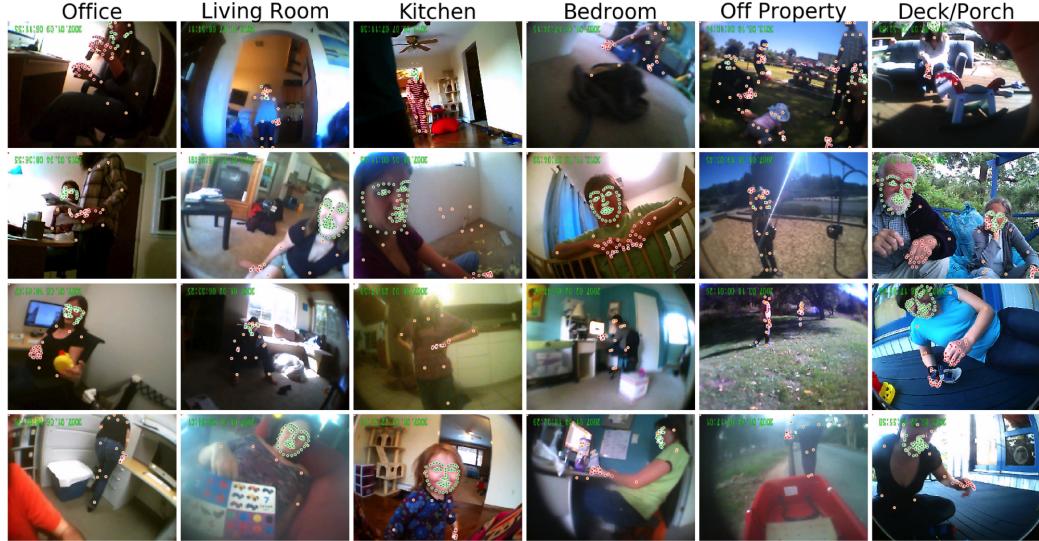


Figure 1. Example frames taken from the dataset, illustrating variability in the infant perspective across different locations. OpenPose detections are shown overlaid on these images (green dots = face, red dots = hands, orange dots = pose).

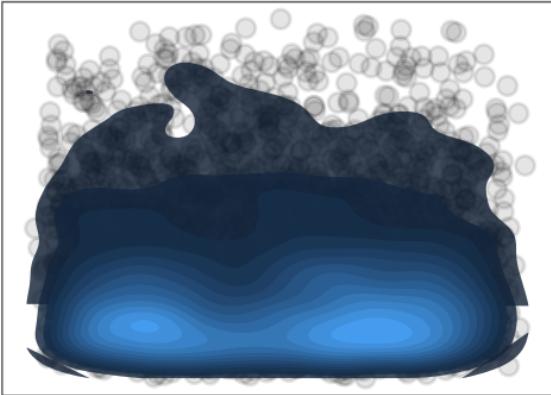
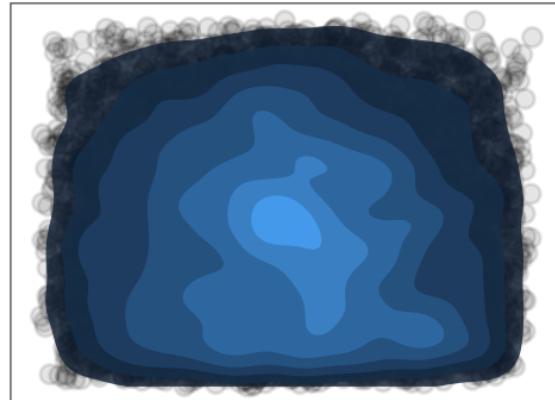
A. Child hand density**B. Adult hand density**

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.

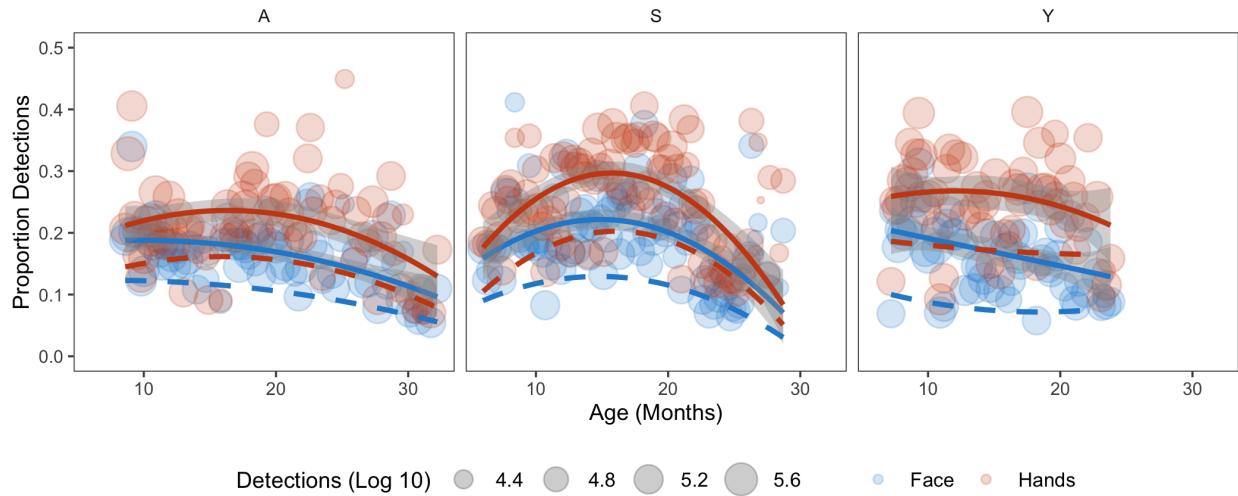


Figure 3. Proportion of faces and hands seen as a function of age for each child in the dataset. Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range. Dashed lines show estimated trend lines from proportion of faces/hands in view when detections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

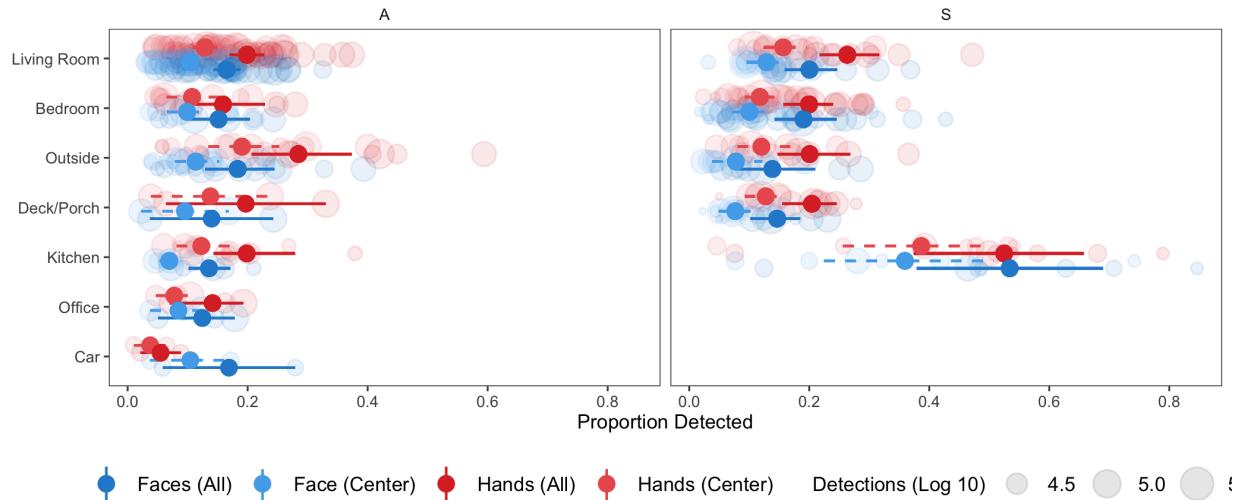


Figure 4. Proportion of faces and hands by location in which egocentric videos were filmed; each panel represents data from an individual child (location annotations were not yet available for Y). Each dot represents data from a week in which videos were filmed and are scaled by the number of frames.