

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, n.d.). To
15 analyze these naturalistic infant egocentric videos, we first validated the use of a modern
16 convolutional neural network of pose detection (OpenPose) for the detection of faces and
17 hands. We then applied this model to the entire dataset, and found a higher proportion of
18 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 Fausey, & Smith, 2015; Jayaraman & Smith, 2018; 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., n.d.). In addition to its size and longitudinal nature, this dataset is
57 more naturalistic than those previously used in two key ways. First, recordings were taken
58 under a large variety of activity contexts (Bruner, 1985; Roy, Frank, DeCamp, Miller, & Roy,
59 2015) encompassing infants' viewpoints during both activities outside and inside the home.
60 Even in other naturalistic datasets, the incredible variety in a typical infant's experience has
61 been largely underrepresented (see examples in Figure 1; e.g., riding in the car, gardening,
62 watching chickens during a walk, browsing magazines, nursing, brushing teeth). Second, the
63 head-mounted cameras used in the SAYCam dataset captured a larger field of view than
64 those typically used, allowing a more complete picture of the infant perspective. While
65 head-mounted cameras with a more restricted field of view do represent where infants are
66 foveating most of the time (Smith, Yu, Yoshida, & Fausey, 2015; Yoshida & Smith, 2008),
67 they may fail to capture short saccades to either faces or hands in the periphery, as the
68 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.

Method

90 Dataset

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

98 field of view to approximately 109 degrees horizontal x 70 degrees vertical. Videos¹ with
99 technical errors or that were not taken from the egocentric perspective were excluded from
100 the dataset. We analyze 1745 videos, with a total duration of 391.11 hours (>40 million
101 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); XX frames with allocentric videos were
119 excluded, as were these videos from the rest of analyses. Frames were jointly annotated for

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

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

120 the presence of faces and hands by one author. A second set of coders recruited via AMT
121 (Amazon Mechanical Turk) additionally annotated 3150 frames; agreement with the primary
122 coder was >95%.

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

134 To evaluate OpenPose's performance, we compared its detections to the
135 manually-annotated gold set of frames, calculating precision (hits / hits + false alarms),
136 recall (hits / hits + misses), and F-score (the harmonic mean of precision and recall). In our
137 data, for faces, the F-score was 0.64, with a precision of 0.70 and recall of 0.58. For hands,
138 the F-score was 0.51, with a precision of 0.73 and recall of 0.40. While face and hand
139 detections showed moderately good precision, face detections were overall slightly more
140 accurate than hand detections. In general, hand detections suffered from fairly low recall,
141 indicating that OpenPose likely underestimated the proportion of hands in the dataset.

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

¹⁴⁶ AMT) were asked to draw bounding boxes around children’s and adult’s hands. Overall, we
¹⁴⁷ found that 43% of missed hand detections were of child hands. When frames with children’s
¹⁴⁸ hands were removed from the gold set, recall did improve somewhat to 0.57. We also
¹⁴⁹ observed that children’s hands tended to appear in the lower half of the frames; heatmaps of
¹⁵⁰ the bounding boxes obtained from these annotations can be seen in Figure 2.

¹⁵¹ Finally, we examined whether the precision, recall, and F-score for hands and faces
¹⁵² varied with age or child, and did not find substantial variation. Thus, while OpenPose was
¹⁵³ trained on photographs from the adult perspective, this model still generalized relatively well
¹⁵⁴ to the egocentric infant viewpoint with no fine-tuning or post-processing of the detections.
¹⁵⁵ As these detections were imperfect compared to human annotators, fine-tuning these models
¹⁵⁶ to better optimize for the infant viewpoint remains an open avenue for future work.
¹⁵⁷ Standard computer vision models are rarely trained on the egocentric viewpoint, and we
¹⁵⁸ suspect that training these models on more naturalistic data may lead to more robust,
¹⁵⁹ generalizable detectors.

¹⁶⁰ Results

¹⁶¹ Access to social information across age

¹⁶² We analyzed the social information in view across the entire dataset, looking
¹⁶³ specifically at the proportions of faces and hands detected for each child.³ Data from videos
¹⁶⁴ were binned according to the age of the child (in weeks). First, we saw that the proportion
¹⁶⁵ of faces in view showed a moderate decrease across this age range (see Figure 3), in keeping
¹⁶⁶ with prior findings (Fausey et al., 2016); in contrast, we did not observe an increase in the
¹⁶⁷ proportion of hands in view. These effects were quantified with two separate linear

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

¹⁶⁸ mixed-models (see Tables 1 & 2).⁴

¹⁶⁹ However, the most striking result from these analyses is a much overall greater
¹⁷⁰ proportion of hands in view than have previously been reported (Fausey et al., 2016). We
¹⁷¹ found this to be true across all ages, in all three children, and regardless of whether we
¹⁷² analyzed human annotations (on the 24K random subset, see dotted lines in Appendix
¹⁷³ Figure ??) or OpenPose annotations on the entire dataset (see solid lines in Figure 3). This
¹⁷⁴ is notable especially given that OpenPose showed relatively low recall for hands, indicating
¹⁷⁵ that this may be an underestimate of the proportion of hands in view. In fact, analysis the
¹⁷⁶ human annotations underscores revealed a much higher proportion of hands relative to faces
¹⁷⁷ than the automated annotations.

¹⁷⁸ One reason this could be the case is the much larger field of view that was captured by
¹⁷⁹ the cameras used in this study: These cameras were outfitted with a fish-eye lens in an
¹⁸⁰ attempt to capture as much of the children's field of view as possible, leading to a larger field
¹⁸¹ of view (109 degrees horizontal x 70 degrees vertical) than in many previous studies; for
¹⁸² example, in Fausey et al. (2016) the FOV was 69 x 41 degrees. This larger FOV may have
¹⁸³ allowed the SAYCam cameras to capture not only the presence of a social partner's hands
¹⁸⁴ interacting with objects or gestures, but also the children's own hands, leading to more
¹⁸⁵ frequent hand detections.

¹⁸⁶ This larger field of view may have also led to the inclusion of children's own hands in
¹⁸⁷ the frames. As we found that children's hands tended to occur in the lower visual field (see
¹⁸⁸ Figure 2), we thus re-analyzed the entire dataset while restricting our analysis to the center
¹⁸⁹ field of view, decreasing the proportion of hand detections from 24% to NA%, but only
¹⁹⁰ decreased face detections from 20% to NA%. This cropping likely removed both the majority

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.

191 of detections of children's own hands but also some detections of adult hands (see Figure 2),
192 especially as OpenPose was biased to miss children's hands when they were in view.
193 Nonetheless, within this modified field of view, we still observed more hand detections than
194 face detections (see dashed lines in Figure 3). We also still found a higher proportion of
195 hands in view relative to faces when excluding any frames containing child hand's from the
196 human annotated gold sample (see Appendix Figure??).

197 Finally, we analyzed how these two sources of social information co-occurred, finding
198 that faces/hands were jointly present in 11.50 percent of frames (see face hand-occurrences
199 across age in Figure 4). To do so, we calculated the number of frames in which infants saw
200 faces and hands together relative to overall proportions of faces/hands that were detected for
201 each child and age range. We found that all three infants than were more likely see hands
202 independently – without the presence of a face – that they were likely to see faces
203 independently. That is, generally speaking when a face was present, a hand also tended to be
204 present.

205 Variability in social information across learning contexts

206 How does variability across different learning contexts influence the social information
207 in the infant view? Intuitively, some activities in different contexts may be characterized by
208 a much higher proportion of faces (e.g., diaper changes in bedrooms) than others (e.g.,
209 playtime in the living room). We thus next examined variation in presence of hands and
210 faces across different locations. Of the 1745 videos, 639 were annotated (Sullivan et al., n.d.)
211 for the location they were filmed in. Of these, 296 videos were filmed in single location,
212 representing 17 percent of the dataset and over 5 million frames. Activities varied somewhat
213 predictably by these contexts: for example, eating tended to occur in the kitchen, whereas
214 playtime was the dominant activity in the living room. Overall, we found that the
215 proportion of faces vs. hands varied across filming locations, and, to some extent, across

216 children; separate chi-squared tests for each child and detection type revealed significant
217 variability in detections by location in each case.⁵ For example, while both A and S saw a
218 relatively similar proportion of faces vs. hands in the bedroom, they saw quite different
219 amounts of faces vs. hands in kitchens (see Figure 5).

220 **Fine-grained changes in the social information in view**

221 In a second set of analysis, we explored fine-grained changes in how these infants
222 experienced social information across development, capitalizing on the fact that OpenPose
223 provides not only face and hand detections but also positional keypoints. In particular, we
224 explored this dataset with the idea that younger, less mobile children may tend to see larger
225 faces towards the center of their visual field while older, more mobile children may
226 experience more smaller, more variable views of faces.

227 Supporting this idea, we found that the averages sizes of the people, faces, and hands
228 in the infant view became smaller over development (Figure 6). This effect was relatively
229 consistent across the three children in the dataset, despite the fact that the three children
230 showed sometimes disparate overall proportions of faces/hands in view. Thus, children may
231 see closer, larger views of people, hands, and faces earlier in development. In keeping with
232 this hypothesis, we also found evidence that faces tended to be become farther away from
233 children across development age (restricting our analysis to faces where both eyes were
234 detected): Figure 7A shows the average interpupillary distance on faces as a function of each
235 child's age at the time of recording, showing a trend from larger, closer faces (with a larger
236 interpupillary distance) to smaller faces that were farther away (with a smaller interpupillary
237 distance). However, we found relatively equal proportions of complete, full-view faces across
238 development (see Figure ??: in other words, these very close views of eyes experienced by
239 younger infants were sometimes partial views.

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

Finally, we also examined whether there were changes in where faces tended to appear in the infant 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 XX 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 x and vertical y positions of centers of the faces detected by OpenPose (see Figure 7B), and examined changes across age. First, we found that faces tended to be more variable in the vertical than their horizontal position (see Figure 7B); evidenced by a much higher coefficients of variation in the y-direction. 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 7B).

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., 2016). 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., 2016; 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 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.

We also found changes in how children experienced faces and hands over development: we found that the faces, hands, and people in the infant view tended to become smaller and

291 that faces tended to be farther away and in more variable horizontal positions. Overall, these
292 data support the idea that the social information in view changes across development as
293 infants become increasingly mobile and independent. As children can navigate the world on
294 their own, they may experience fewer close-up interactions with the caregivers and more
295 bouts of play where they are exploring the objects and things in the environment around
296 them.

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

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

314

Acknowledgements

315

Thanks to the creators of the SAYCam dataset who made this work possible and to

316

Alessandro Sanchez for his contributions to the codebase. This work was funded by a Jacobs

317

Foundation Fellowship to MCF, a John Mereck Scholars award to MCF, and NSF #1714726

318

to BLL.

319

References

- 320 Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands
321 and recognizing activities in complex egocentric interactions. In *Proc. Of the IEEE*
322 *international conference on computer vision* (pp. 1949–1957).
- 323 Bruner, J. (1985). The role of interaction formats in language acquisition. In *Language and*
324 *social situations* (pp. 31–46). Springer.
- 325 Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime
326 multi-person 2D pose estimation using Part Affinity Fields. In *ArXiv preprint*
327 *arXiv:1812.08008*.
- 328 Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual
329 statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, 372(1711),
330 20160055.
- 331 Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in
332 humans from birth. *Proceedings of the National Academy of Sciences*, 99(14),
333 9602–9605.
- 334 Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual
335 input in the first two years. *Cognition*, 152, 101–107.
- 336 Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2017). See and be seen: Infant–caregiver
337 social looking during locomotor free play. *Developmental Science*.
- 338 Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye
339 tracking: A new method to describe infant looking. *Child Development*, 82(6),
340 1738–1750.

- 341 Frank, M. C., Simmons, K., Yurovsky, D., & Pusiol, G. (2013). Developmental and postural
342 changes in children's visual access to faces. In *Proceedings of the 35th annual meeting*
343 of the cognitive science society (pp. 454–459).
- 344 Gredeback, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of infants'
345 gaze as they view adult shifts in overt attention. *Infancy*, 13(5), 533–543.
- 346 James, W. (1890). *The principles of psychology* (Vol. 1). Macmillan London.
- 347 Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes
348 change over the first year of life. *PLoS One*.
349 <https://doi.org/10.1371/journal.pone.0123780>
- 350 Jayaraman, S., & Smith, L. B. (2018). Faces in early visual environments are persistent not
351 just frequent. *Vision Research*.
- 352 Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see
353 the world differently. *Child Development*, 85(4), 1503–1518.
- 354 Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of
355 a spoken word. *Proc. Of the National Academy of Sciences*, 112(41), 12663–12668.
- 356 Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments
357 modulate children's visual access to social information. In *Proceedings of the 40th*
358 *annual conference of the cognitive science society*.
- 359 Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single
360 images using multiview bootstrapping. In *CVPR*.
- 361 Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of
362 toddler visual experience. *Developmental Science*, 14(1), 9–17.

- 363 Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted
364 cameras to studying the visual environments of infants and young children. *Journal
365 of Cognition and Development*, 16(3), 407–419.
- 366 Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye:
367 Typical, daily exposure to faces documented from a first-person infant perspective.
368 *Developmental Psychobiology*, 56(2), 249–261.
- 369 Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (n.d.). Head cameras on
370 children aged 6 months through 31 months.
- 371 Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to
372 study visual experience. *Infancy*, 13, 229–248.
- 373 Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and
374 their parents coordinate visual attention to objects through eye-hand coordination.
375 *PloS One*, 8(11).

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	0.098	0.010	1.956	9.386	0.012
age_scale	-0.001	0.000	430.976	-3.150	0.002

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)	0.161	0.009	1.881	18.046	0.004
age_scale	-0.001	0.001	428.651	-1.680	0.094

Table 2

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

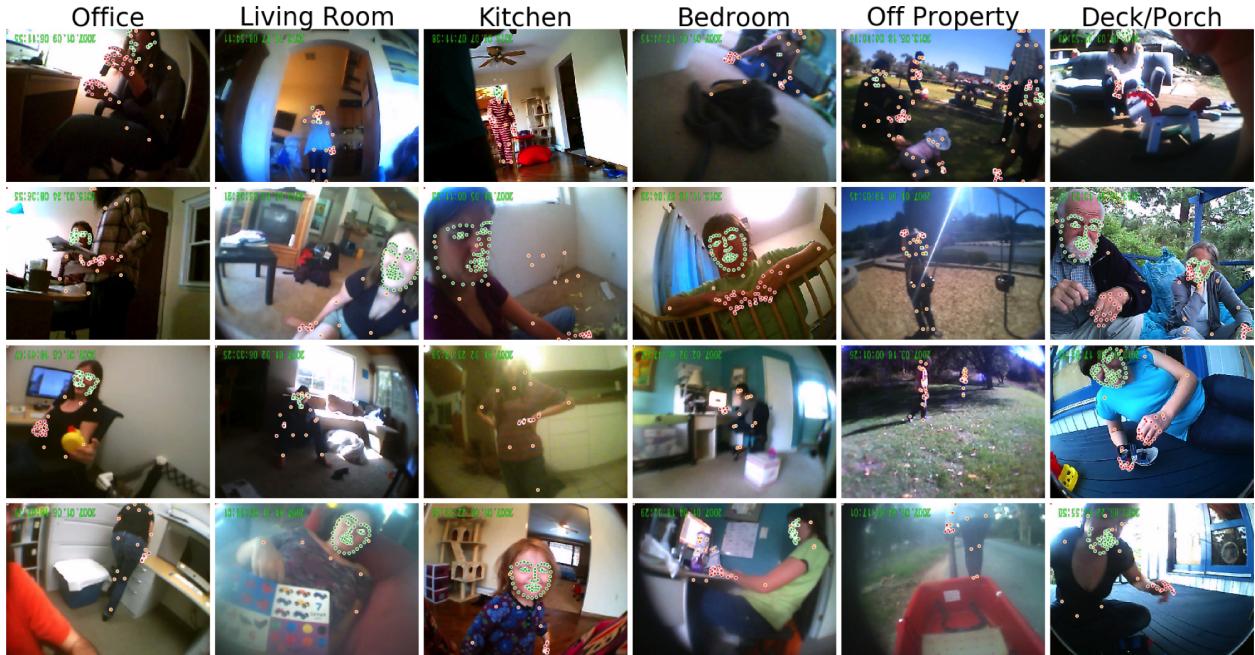


Figure 1. Example frames taken from the dataset, illustrating variability in the infant perspective across different locations. OpenPose detections are shown overlaid on these images (green dots = face, red dots = hands, orange dots = pose).

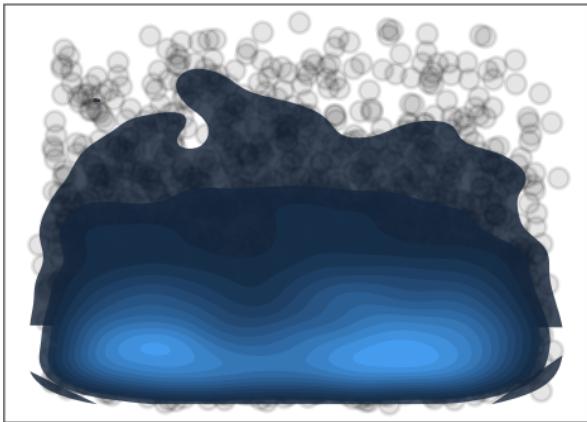
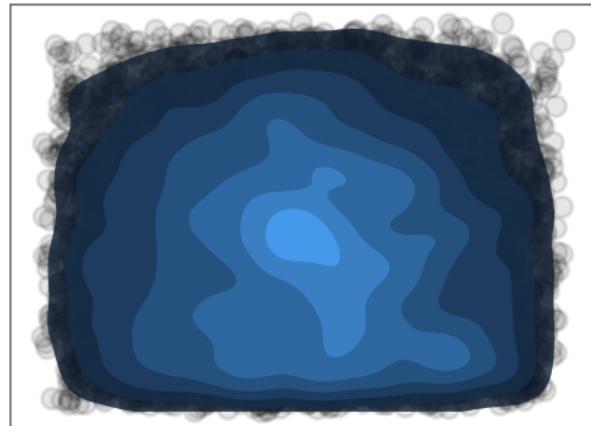
A. Child hand density**B. Adult hand density**

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.

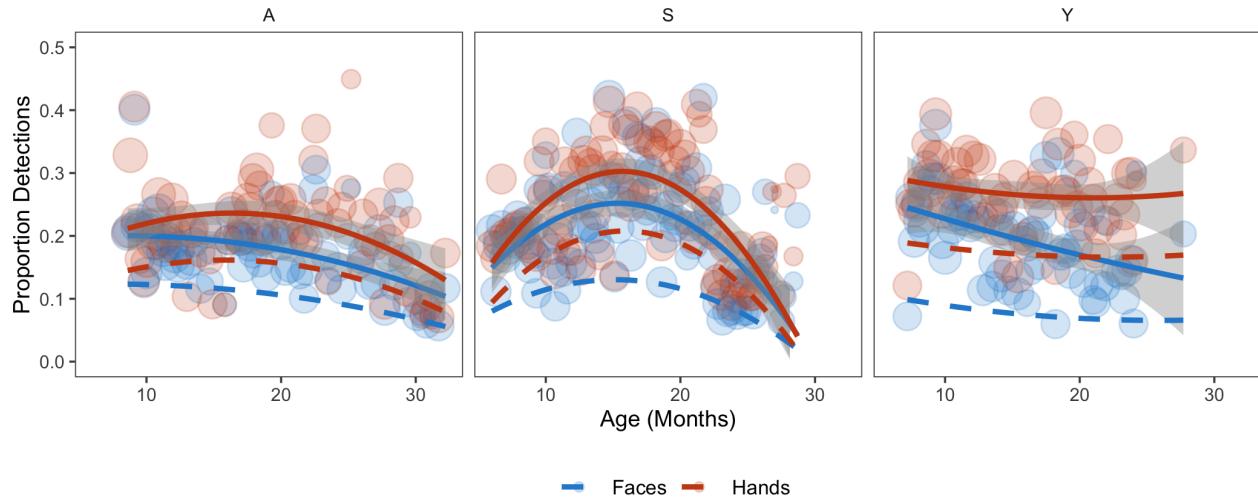


Figure 3. Proportion of faces and hands seen as a function of age for each child in the dataset.

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.

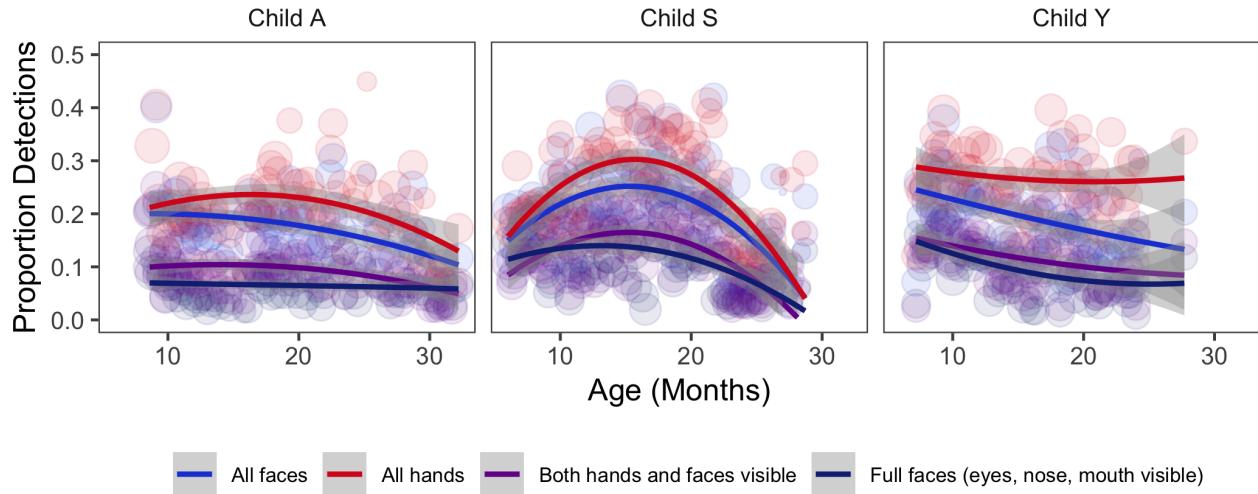


Figure 4. Proportion of face, hand, 'full' face (i.e. nose, mouth, and nose visible), and joint face/hand detections as a function of age for each child in the dataset. Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range.

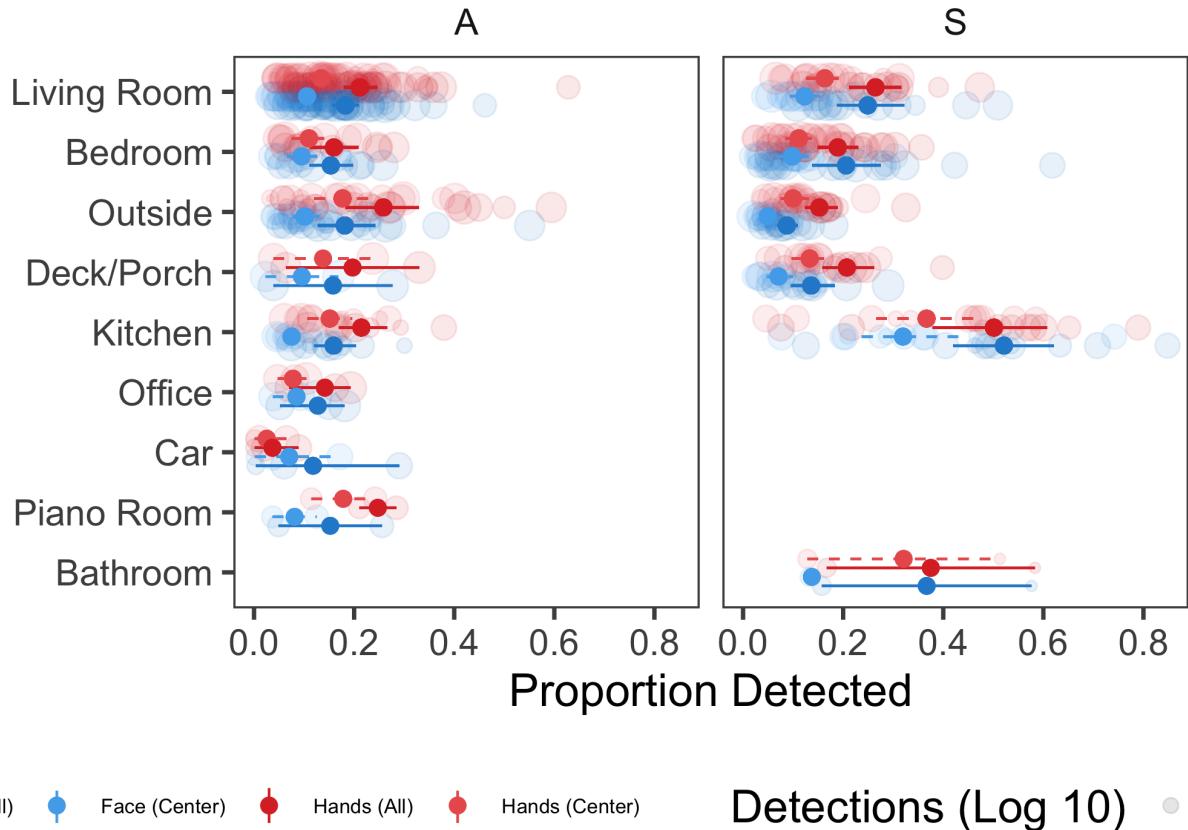


Figure 5. Proportion of faces and hands by location in which egocentric videos were filmed; each panel represents data from an individual child (location annotations were not yet available for Y). Each dot represents data from a week in which videos were filmed and are scaled by the number of frames.

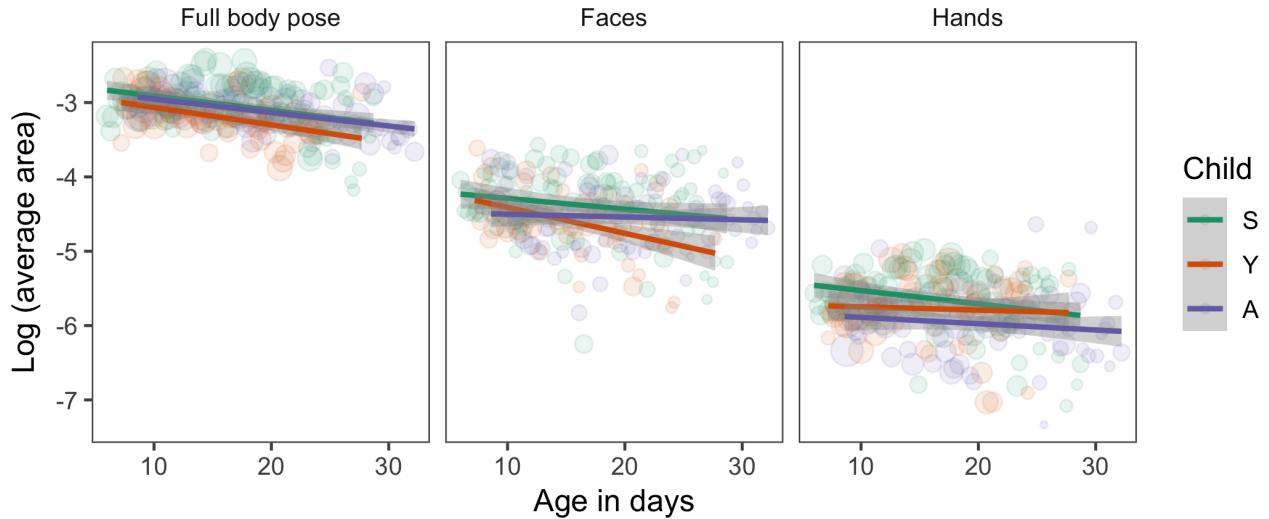


Figure 6. Average size of poses, faces, and hands detected in the dataset between eyes in faces detected as a function of age for each child in the dataset (each color = different child). Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range.

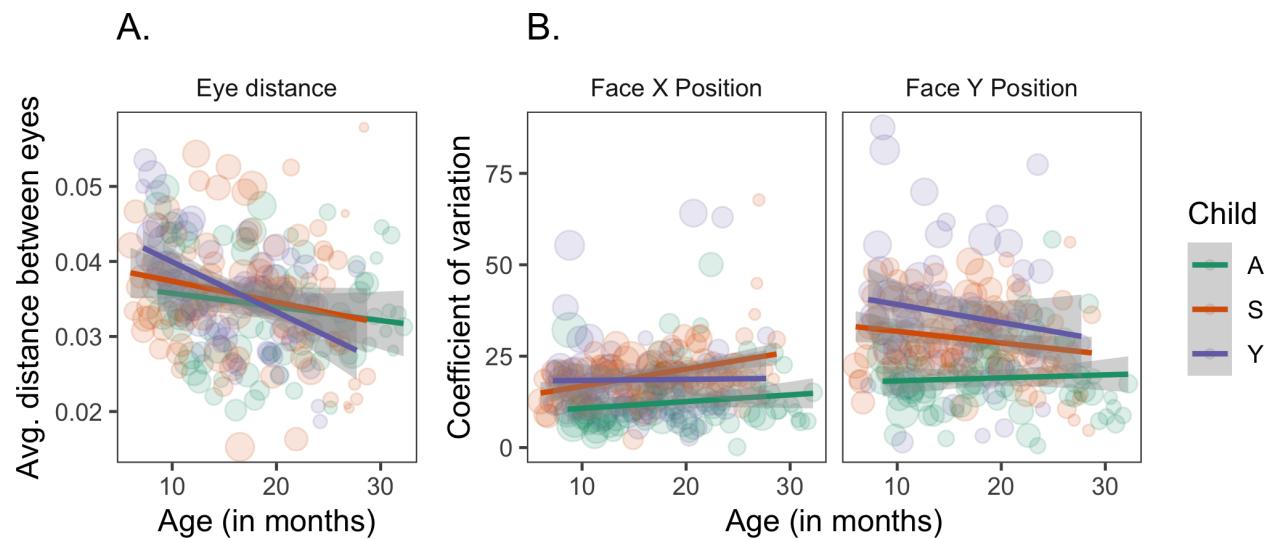


Figure 7. (A).Average distance between eyes and (B) average coefficient of variation for the x and y position of faces detected by OpenPose as a function of each child's age at the time of filming. Data in (A) are restricted to faces where both eyes were detected. Data are binned by each week that the videos were filmed and scaled by the number of face detections in that age range.