

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

Abstract

5 The faces and hands of caregivers and other social partners offer a rich source of social and
6 causal information that may be critical for infants' cognitive and linguistic development.

7 Previous work using manual annotation strategies and cross-sectional data has found
8 systematic changes in the proportion of faces and hands in the egocentric perspective of
9 young infants. Here, we examine the prevalence of faces and hands in a longitudinal
10 collection of nearly 1700 headcam videos collected from three children along a span of 6 to
11 32 months of age—the SAYCam dataset (Sullivan, Mei, Perfors, Wojcik, & Frank, 2021).

12 To analyze these naturalistic infant egocentric videos, we first validated the use of a
13 modern convolutional neural network of pose detection (OpenPose) for the detection of
14 faces and hands. We then applied this model to the entire dataset, and found a higher
15 proportion of hands in view than previous reported and a moderate decrease the
16 proportion of faces in children's view across age. In addition, we found variability in the
17 proportion of faces/hands viewed by different children in different locations (e.g., living
18 room vs. kitchen), suggesting that individual activity contexts may shape the social
19 information that infants experience.

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

21 Word count: 3444

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

23 **Introduction**

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

32 Yet until recently most hypotheses about infants' visual experience have gone
33 untested. Though parents and scientists alike have strong intuitions about what infants
34 see, even the viewpoint of a walking child is hard to intuit (Clerkin, Hart, Rehg, Yu, &
35 Smith, 2017; Franchak, Kretch, Soska, & Adolph, 2011). By equipping infants and toddlers
36 with head-mounted cameras, researchers have begun to document the infant's egocentric
37 perspective on the world. Using these methods, a growing body of work now demonstrates
38 that the viewpoints of very young infants (less than 4 months of age) are indeed dominated
39 by frequent, persistent views of the faces of their caregivers (Jayaraman, Fausey, & Smith,
40 2015; Jayaraman & Smith, 2018; Sugden, Mohamed-Ali, & Moulson, 2014).

41 Beyond these early months, infants' motor and cognitive abilities mature, leading to
42 vastly different perspectives on the world. For example, crawlers see fewer faces and hands
43 than do walking children (Franchak, Kretch, & Adolph, 2017; Kretch, Franchak, & Adolph,
44 2014; Sanchez, Long, Kraus, & Frank, 2018) as well as different views of objects (Smith,
45 Yu, & Pereira, 2011). Further, as infants learn to use their own hands to act on the world,
46 they seem to focus on manual actions taken by their social partners, and their perspective
47 starts to capture views of hands manipulating objects (Fausey, Jayaraman, & Smith, 2016).

- 48 In turn, caregivers may also start to use their hands with more communicative intent,
49 directing infants' attention by pointing and gesturing to different events and objects during
50 play (Yu & Smith, 2013).

51 Here, we examine the social information present in the infant visual perspective—the
52 presence of faces and hands—by analyzing a longitudinal collection of more than 1700
53 headcam videos collected from three children along a span of 6 to 32 months of age—the
54 SAYCam dataset (Sullivan et al., 2021). In addition to its size and longitudinal nature,
55 this dataset is more naturalistic than those previously used in two key ways. First,
56 recordings were taken under a large variety of activity contexts (Bruner, 1985; Roy, Frank,
57 DeCamp, Miller, & Roy, 2015) encompassing infants' viewpoints during both activities
58 outside and inside the home. Even in other naturalistic datasets, the incredible variety in a
59 typical infant's experience has been largely underrepresented (see examples in Figure 1;
60 e.g., riding in the car, gardening, watching chickens during a walk, browsing magazines,
61 nursing, brushing teeth). Second, the head-mounted cameras used in the SAYCam dataset
62 captured a larger field of view than those typically used, allowing a more complete picture
63 of the infant perspective. While head-mounted cameras with a more restricted field of view
64 do represent where infants are foveating most of the time (Smith, Yu, Yoshida, & Fausey,
65 2015; Yoshida & Smith, 2008), they may fail to capture short saccades to either faces or
66 hands in the periphery, as the timescale of head movements is much longer.

67 With hundreds of hours of footage (>40M frames), however, this large dataset
68 necessitates a shift to an automated annotation strategy. Indeed, annotation of the frames
69 extracted from egocentric videos has been prohibitively time-consuming, meaning that
70 most frames are typically not inspected, even in the most comprehensive studies. For
71 example, Fausey et al. (2016) collected a total of 143 hours of head-mounted camera
72 footage (15.5 million frames), of which one frame every five seconds was hand-annotated
73 (by four coders), totalling 103,383 frames (per coder)—an impressive number of
74 annotations but nonetheless only 0.67% of the collected footage. To address this challenge,

75 we use a modern computer vision model of pose detection to automatically detect the
76 presence of hands and faces from the infant egocentric viewpoint. Specifically, we use
77 OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018), a model optimized for jointly
78 detecting human face, body, hand, and foot keypoints that operates well on scenes
79 including multiple people, even if they are partially-occluded (see Figure 1). In prior work
80 examining egocentric videos, OpenPose performed comparably to other modern face
81 detection models (Sanchez et al., 2018).

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

90

Method

91 **Dataset**

92 The dataset is described in detail in Sullivan et al. (2021); we summarize these
93 details here. Children wore Veho Muvi miniature cameras mounted on a custom camping
94 headlamp harness (“headcams”) at least twice weekly, for approximately one hour per
95 recording session. One weekly session was on the same day each week at a roughly constant
96 time of day, while the other(s) were chosen arbitrarily at the participating family’s
97 discretion. At the time of the recording, all three children were in single-child households.
98 Videos captured by the headcam were 640x480 pixels, and a fisheye lens was attached to
99 the camera to increase the field of view to approximately 109 degrees horizontal x 70

¹⁰⁰ degrees vertical. Videos¹ with technical errors or that were not taken from the egocentric
¹⁰¹ perspective were excluded from the dataset. We analyze 1745 videos, with a total duration
¹⁰² of 391.11 hours (>40 million frames).

¹⁰³ **Detection Method**

¹⁰⁴ To automatically annotate the millions of frames in SAYCam, we used a pose
¹⁰⁵ detector, OpenPose² (Cao et al., 2018; Simon, Joo, Matthews, & Sheikh, 2017), which
¹⁰⁶ provided the locations of 18 body parts (ears, nose, wrists, etc.). To do so, a convolutional
¹⁰⁷ neural network was used for initial anatomical detection, and part affinity fields were
¹⁰⁸ subsequently applied for part association to produce a series of body part candidates. Once
¹⁰⁹ these body part candidates were matched to a single individual in the frame, they were
¹¹⁰ finally assembled into a pose. Thus, while we only made use of the outputs of the face and
¹¹¹ hand detections, the entire set of pose information from an individual was used to
¹¹² determine the presence of a face/hand, making the process more robust to occlusion than
¹¹³ methods optimized to detect only faces or hands. Note, however, that these face/hand
¹¹⁴ detections are reliant on the detection of at least a partial pose, so some very up-close
¹¹⁵ views of faces/hands may go undetected.

¹¹⁶ **Detection Validation**

¹¹⁷ To test the validity of OpenPose’s hand and face detections, we compared the
¹¹⁸ accuracy of these detections relative to human annotations of 24,000 frames selected
¹¹⁹ uniformly at random from videos of two children (S and A); 24000 frames sampled from
¹²⁰ allocentric videos were excluded, and these videos were also excluded from the other
¹²¹ analyses. Frames were jointly annotated for the presence of faces and hands by one author.

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

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

122 A second set of coders recruited via AMT (Amazon Mechanical Turk) additionally
123 annotated 3150 frames; agreement with the primary coder was >95%.

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

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

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

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

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

161 Results and Discussion

162 Access to social information across age

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

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

169 linear mixed-models (see Tables 1 & 2).⁴

170 However, the most striking result from these analyses is a much overall greater
171 proportion of hands in view than have previously been reported (Fausey et al., 2016). We
172 found this to be true across all ages, in all three children, and regardless of whether we
173 analyzed human annotations (on the 24K random subset, see dotted lines in Appendix
174 Figure A1) or OpenPose annotations on the entire dataset (see solid lines in Figure 3).
175 This is notable especially given that OpenPose showed relatively low recall for hands,
176 indicating that this may be an underestimate of the proportion of hands in view. In fact,
177 analysis the human annotations underscores revealed a much higher proportion of hands
178 relative to faces than the automated annotations.

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

187 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 16%, but only
190 decreased face detections from 20% to 9.90%. This cropping likely removed both the
191 majority of detections of children's own hands but also some detections of adult hands (see
192 Figure 2), especially as OpenPose was biased to miss children's hands when they were in

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

193 view. Nonetheless, within this modified field of view, we still observed more hand
194 detections than face detections (see dashed lines in Figure 3). We also still found a higher
195 proportion of hands in view relative to faces when excluding any frames containing child
196 hand's from the human annotated gold sample (see Appendix FigureA1).

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
201 for each child and age range. As shown in Figure 5, we found that all three infants were
202 more likely to see hands independently – without the presence of a face – than they were
203 likely to see faces independently. That is, generally speaking when a face was present, a
204 hand also tended to be present.

205 **Variability in social information across learning contexts**

206 How does the child's context influence the social information in view? Bruner (1985)
207 discussed the role of children's activities in shaping the information present for learning.
208 Following this idea, we investigated whether there were differences in access to faces by the
209 activity that the child was engaged in. This hypothesis seems intuitively appealing.

210 Some activities seem likely to be characterized by a much higher proportion of faces
211 (e.g., diaper changes) than others (e.g., a car trip). Following this same idea, perhaps other
212 activities involve the presence of more hands in the field of view (e.g., playtime).

213 We did not have access to annotations of activity. Thus, following Roy et al. (2015),
214 we used spatial location as a proxy for activity context, taking advantage of the presence of
215 these annotations for a subset of the SAYCam videos. Of the 1745 videos in the dataset,
216 639 were annotated for the location or locations they were filmed in. These location
217 annotations were only available for two children, S and A. Annotated locations mostly

218 consisted of rooms of the house (e.g., “living room”) but also included some other locations
219 (e.g., “car,” “outside”). Of this set, 296 videos were filmed in only a single location (e.g.,
220 the location label did not change within the video), representing 17 percent of the dataset
221 and over 5 million frames. In our viewing of the SAYCam videos and in other annotations
222 available with the dataset, activities varied somewhat predictably by location: for example,
223 eating tended to occur in the kitchen, whereas playtime was the dominant activity in the
224 living room.

225 Figure 6 shows the proportion of faces vs. hands across locations. We found
226 substantial variation across locations and, to some extent, across children. Separate
227 chi-squared tests for each child and detection type revealed significant variability in
228 detections by location in each case, with all $ps < .001$. For example, while both A and S
229 saw a relatively similar proportion of faces and hands in the bedroom, the two children saw
230 quite different amounts of faces and hands from one another in the kitchen. This difference
231 is likely explained by differences in arrangement of the kitchen in the two children’s
232 households (Sullivan, personal communication), such that mealtimes in one kitchen
233 resulted in a face-to-face orientation while it did not in the other). This example illustrates
234 how specifics of the geometry of a particular context can play an outsize role in the child’s
235 access to social information during that context.

236 Fine-grained changes in the social information in view

237 In a third set of analyses, we explored fine-grained changes in the SAYCam infants’
238 access to social information across development. In these analyses, we capitalize on the fact
239 that OpenPose provides not only face and hand detections but also positional keypoints. In
240 particular, we explored this keypoint dataset with the idea that greater mobility allows
241 older children to be further from their caregivers on average. Thus, younger, less mobile
242 children may tend to see larger faces towards the center of their visual field while older,
243 more mobile children may experience more smaller, more variable views of faces. The same

²⁴⁴ dynamic would be predicted hold for hands as well, as it would be driven by overall
²⁴⁵ differences in distance.

²⁴⁶ Supporting this idea, we found that the averages sizes of the people, faces, and hands
²⁴⁷ in the infant view became smaller over development (Figure 7). This effect was relatively
²⁴⁸ consistent across the three children in the dataset, despite the fact that the three children
²⁴⁹ showed sometimes disparate overall proportions of faces/hands in view. Thus, children may
²⁵⁰ see closer, larger views of people, hands, and faces earlier in development.

²⁵¹ In keeping with this hypothesis, we also found evidence that faces tended to be
²⁵² farther away from older children. We restricted our analysis here to faces where both eyes
²⁵³ were detected and computed interpupillary distance as a rough metric of distance, since
²⁵⁴ eyes should be closer together on average when a face is further from the camera. Figure
²⁵⁵ 8A shows the average interpupillary distance on faces as a function of each child's age at
²⁵⁶ the time of recording. There is a trend from larger, closer faces (with a larger interpupillary
²⁵⁷ distance) to smaller faces that were farther away (with a smaller interpupillary distance).

²⁵⁸ Finally, we also examined whether there were changes in where faces tended to
²⁵⁹ appear in the camera's (and hence, by proxy, the child's) field of view. As expected, faces
²⁶⁰ tended to be located towards the upper field of view, while views of hands were more
²⁶¹ centrally distributed (see Appendix, Figure XX for average density distributions).
²⁶² However, we also found evidence that older children tended to see more faces in more
²⁶³ variable positions than younger children. Specifically, we examined how variable the
²⁶⁴ horizontal and vertical coordinates were of the faces in the infant view. To do so, we
²⁶⁵ calculated the coefficient of variation of the horizontal (x) and vertical (y) positions of
²⁶⁶ centers of the faces detected by OpenPose (see Figure 8B), and examined changes across
²⁶⁷ age. Faces tended to be more variable in the vertical than their horizontal position (see
²⁶⁸ Figure 8B). We also found that as children got older, they tended to see faces that varied
²⁶⁹ more in their horizontal – but not their vertical position – suggesting that older children

270 might be more likely to see more smaller faces in their periphery (see Figure 8B).

271 **General Discussion**

272 Here, we analyzed the social information in view in a dense, longitudinal dataset,
273 applying a modern computer-vision model to quantify the hands and faces seen from each
274 of three children's egocentric perspective from 6 to 32 months of age. First, we found a
275 moderate decrease across age in the proportion of faces in view in the videos, in keeping
276 with previous work (Fausey et al., 2016). This finding is particularly notable given that, in
277 previous cross-sectional data, this effect seems to be most strongly driven by infants
278 younger than 4 months of age (e.g., Fausey et al., 2016; Jayaraman et al., 2015; Sugden et
279 al., 2014) who see both more frequent and more persistent faces (Jayaraman & Smith,
280 2018). We also found this to be true when restricting our analyses to full-field faces,
281 suggesting this effect is not driven by a concurrent shift from more full-view to
282 partial-views of faces.

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

296 in the social information in view across more vs. less structured activity contexts could
297 further test this idea.

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

309 We also used these detailed pose annotations to explore finer-grained changes in how
310 children experience the faces and hands of their caregivers over development. We found
311 that the faces, hands, and people in the infant view tended to become smaller and that
312 faces tended to be farther away and in more variable horizontal positions. Overall, these
313 data support the idea that the social information in view changes across development as
314 infants become increasingly mobile and independent. As children can navigate the world on
315 their own, they may experience fewer close-up interactions with the caregivers and more
316 bouts of play where they are exploring the objects and things in the environment around
317 them.

318 More broadly, however, these analyses underscore the importance of how, when, from
319 whom, and what data we sample; these choices become central when we attempt to draw
320 conclusions about the regularities of experience. Indeed, while unprecedented in size, this
321 dataset still has many limitations. These videos only represent a small portion of the

322 everyday experience of these three children, all of whom come from relatively privileged
323 households in western societies and thus are not representative in many ways of the global
324 population. Any idiosyncrasies in how and when these particular families chose to film
325 these videos also undoubtedly influences the variability seen here. And without
326 eye-tracking data, we do not know if children are attending to the social information in
327 their visual field.

328 Nonetheless, we believe that these advances in datasets and methodologies represent
329 a step in the right direction. The present paper demonstrates the feasibility of using a
330 modern computer vision model to annotate the entirety of a very large dataset (here,
331 >40M million frames) for the presence and size of people, hands, and faces, representing
332 orders of magnitude more data relative to prior work. We propose that the large-scale
333 analysis of dense datasets, collected with different fields of view, cameras and from many
334 different laboratories, will lead to generalizable conclusions about the regularities of infant
335 experience that scaffold learning.

336 **Acknowledgements**

337 Thanks to the creators of the SAYCam dataset who made this work possible and to
338 Alessandro Sanchez for his contributions to the codebase. This work was funded by a
339 Jacobs Foundation Fellowship to MCF, a John Mereck Scholars award to MCF, and NSF
340 #1714726 to BLL.

References

- 341
- 342 Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands
343 and recognizing activities in complex egocentric interactions. In *Proc. Of the IEEE*
344 *international conference on computer vision* (pp. 1949–1957).
- 345 Bruner, J. (1985). The role of interaction formats in language acquisition. In *Language and*
346 *social situations* (pp. 31–46). Springer.
- 347 Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime
348 multi-person 2D pose estimation using Part Affinity Fields. In *ArXiv preprint*
349 *arXiv:1812.08008*.
- 350 Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual
351 statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B,*
352 372(1711), 20160055.
- 353 Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in
354 humans from birth. *Proceedings of the National Academy of Sciences*, 99(14),
355 9602–9605.
- 356 Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing
357 visual input in the first two years. *Cognition*, 152, 101–107.
- 358 Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2017). See and be seen: Infant–caregiver
359 social looking during locomotor free play. *Developmental Science*.
- 360 Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye
361 tracking: A new method to describe infant looking. *Child Development*, 82(6),
362 1738–1750.
- 363 Frank, M. C., Simmons, K., Yurovsky, D., & Pusiol, G. (2013). Developmental and
364 postural changes in children's visual access to faces. In *Proceedings of the 35th*
365 *annual meeting of the cognitive science society* (pp. 454–459).

- 366 Gredeback, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of
367 infants' gaze as they view adult shifts in overt attention. *Infancy*, 13(5), 533–543.
- 368 James, W. (1890). *The principles of psychology* (Vol. 1). Macmillan London.
- 369 Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective
370 scenes change over the first year of life. *PLoS One*.
371 <https://doi.org/10.1371/journal.pone.0123780>
- 372 Jayaraman, S., & Smith, L. B. (2018). Faces in early visual environments are persistent not
373 just frequent. *Vision Research*.
- 374 Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see
375 the world differently. *Child Development*, 85(4), 1503–1518.
- 376 Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of
377 a spoken word. *Proc. Of the National Academy of Sciences*, 112(41), 12663–12668.
- 378 Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments
379 modulate children's visual access to social information. In *Proceedings of the 40th*
380 *annual conference of the cognitive science society*.
- 381 Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single
382 images using multiview bootstrapping. In *CVPR*.
- 383 Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of
384 toddler visual experience. *Developmental Science*, 14(1), 9–17.
- 385 Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted
386 cameras to studying the visual environments of infants and young children. *Journal*
387 *of Cognition and Development*, 16(3), 407–419.
- 388 Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye:
389 Typical, daily exposure to faces documented from a first-person infant perspective.
390 *Developmental Psychobiology*, 56(2), 249–261.

- 391 Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A large,
392 longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*.
- 393 Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to
394 study visual experience. *Infancy*, 13, 229–248.
- 395 Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and
396 their parents coordinate visual attention to objects through eye-hand coordination.
397 *PloS One*, 8(11).

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	0.098	0.010	1.956	9.386	0.012
Age	-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	-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.

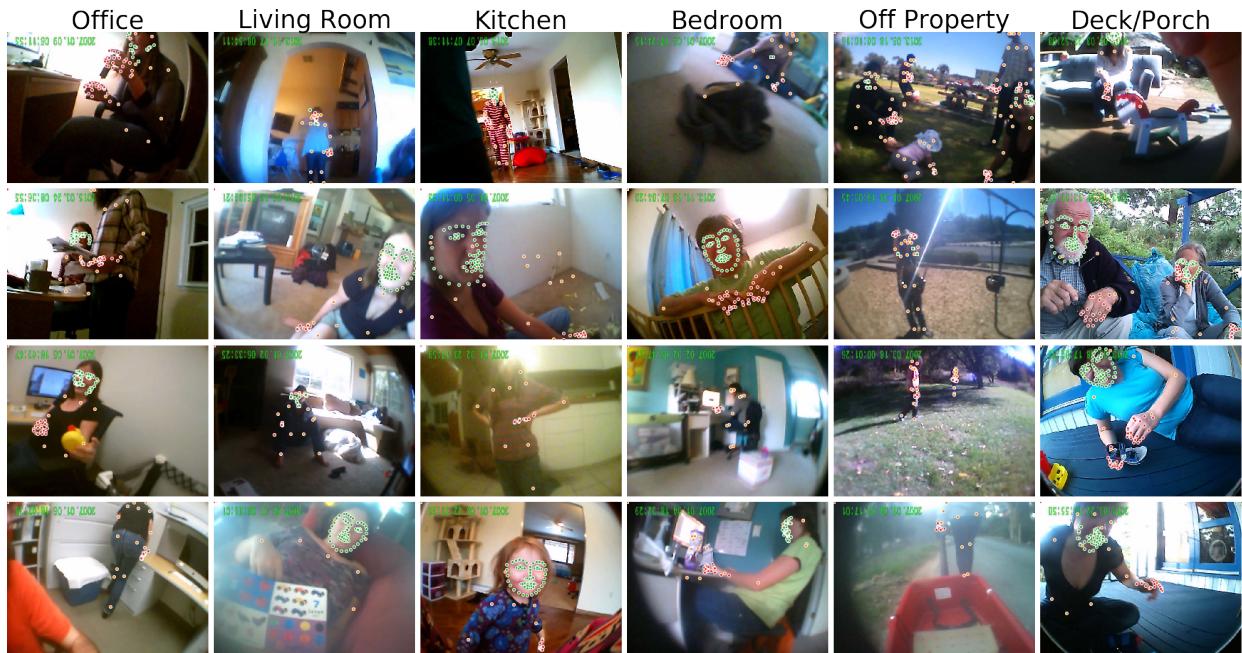


Figure 1. Example frames taken from the dataset, illustrating variability in the infant perspective across different locations. OpenPose detections are shown overlaid on these images (green dots = face, red dots = hands, orange dots = pose).

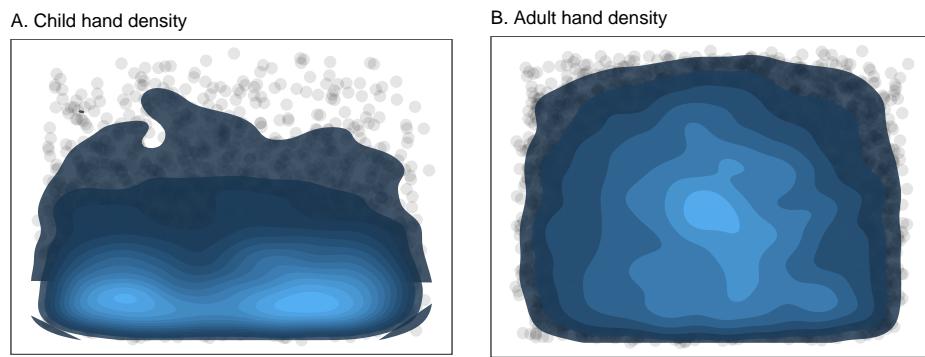


Figure 2. Density estimates for the child (left) and adult (right) hands that were detected in the 24K frame random gold set; each dot represents the center of a bounding box made by an adult participant. Brighter values indicate more detections.

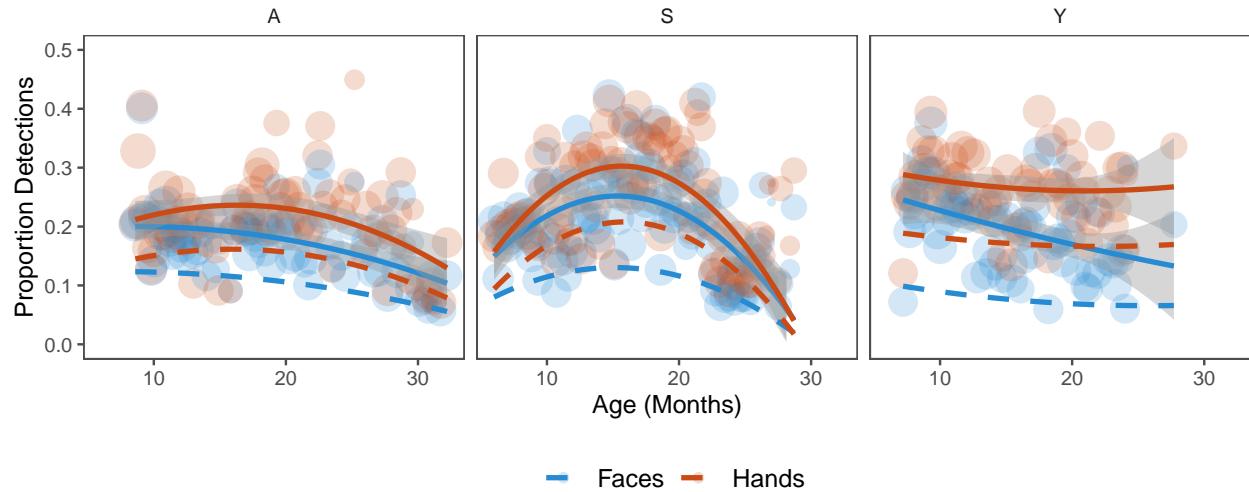


Figure 3. Proportion of faces and hands seen as a function of age for each child (A, S, and Y).

Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range. 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.

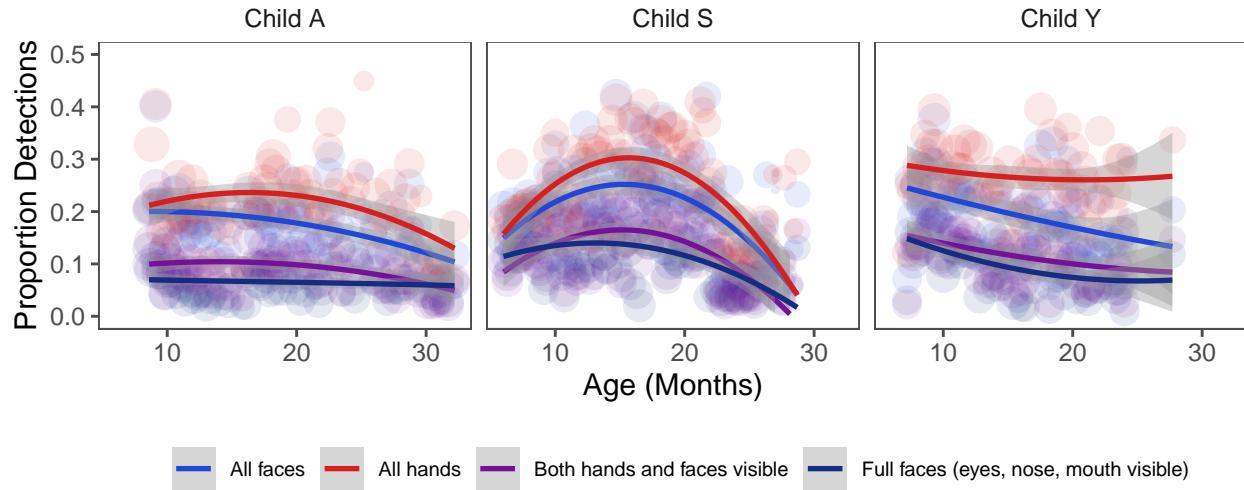


Figure 4. Proportion of face, hand, full face (i.e., nose, mouth, and nose visible), and joint face/hand detections as a function of age for each child in the dataset. Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range.

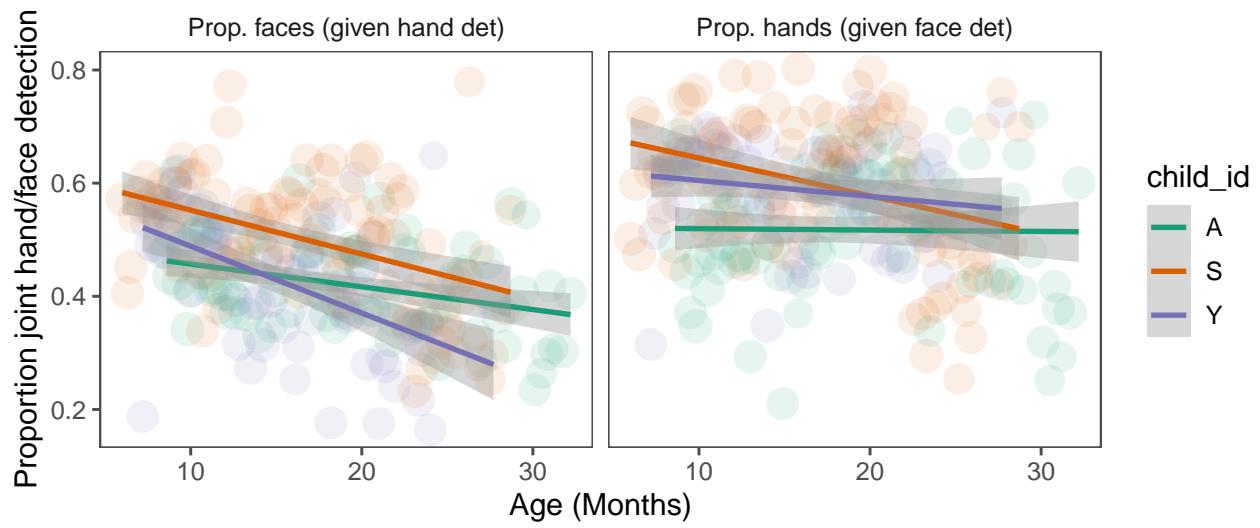


Figure 5. Proportion of joint face and hands detection within frames where hands (left) or faces (right) were detected.

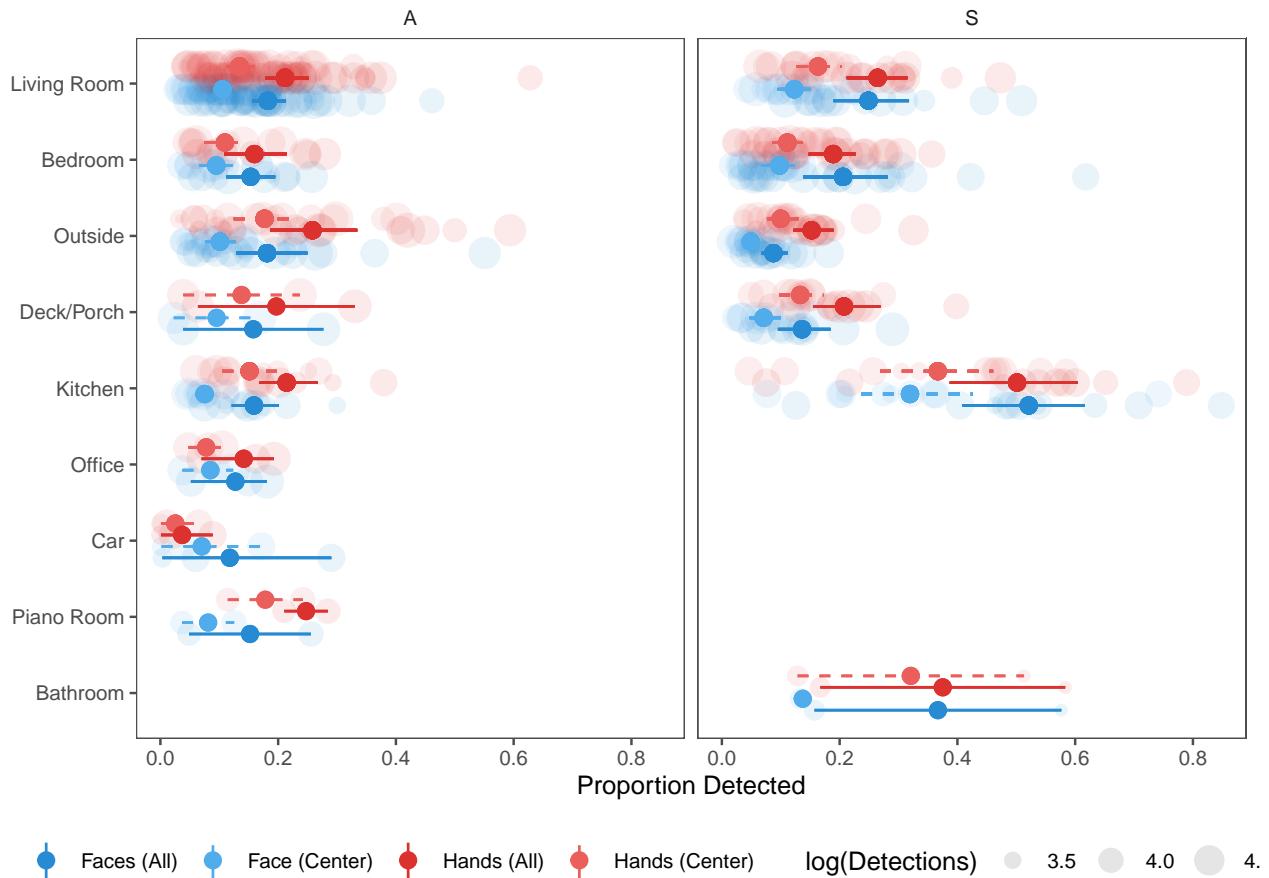


Figure 6. Proportion of faces and hands by location in which egocentric videos were filmed; each panel represents data from an individual child (location annotations were not yet available for Y). Each dot represents data from a week in which videos were filmed and are scaled by the number of frames.

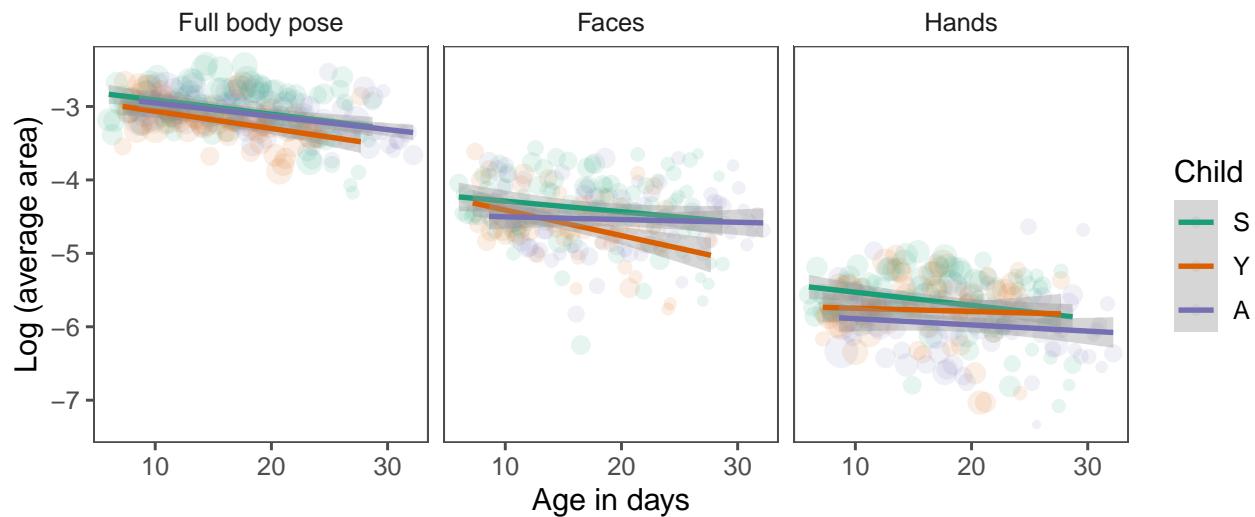


Figure 7. Average size of poses, faces, and hands detected in the dataset between eyes in faces detected as a function of age for each child in the dataset (each color = different child). Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range.

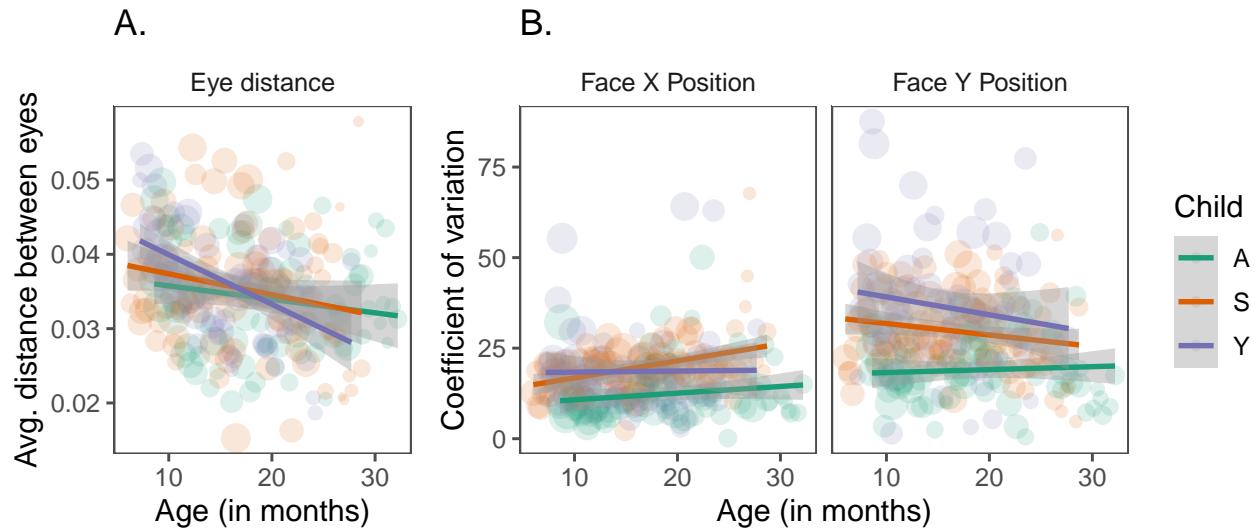


Figure 8. (A) Average distance between eyes and (B) average coefficient of variation for the x and y position of faces detected by OpenPose as a function of each child's age at the time of filming. Data in (A) are restricted to faces where both eyes were detected. Data are binned by each week that the videos were filmed and scaled by the number of face detections in that age range.

Appendix A

Face/hand detections relative to human annotations

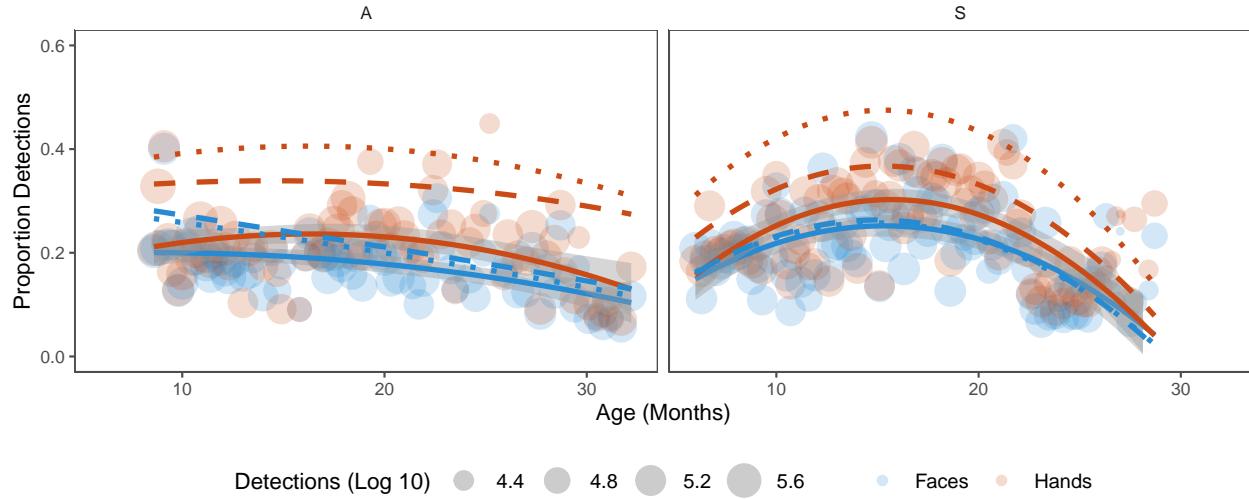


Figure A1. Proportion of faces and hands seen as a function of age for each child in the dataset. Data are binned by each week that the videos were filmed and scaled by the number of frames in that age range. Dashed lines show estimated trend lines from proportion of faces/hands in view when analyzing 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

Distribution of faces and hands in the visual field

³⁹⁸ We explored where in the visual field children tended to see faces and hands, suspecting
³⁹⁹ that these distributions might become wider as children grow older and learn to locomote
⁴⁰⁰ on their own, following preliminary analyses from Frank (2012).

⁴⁰¹ As expected, faces tended to appear in the upper visual field in contrast to hands,
⁴⁰² which tended to be more centrally located.

⁴⁰³ However, we found little evidence for any changes in the positions of faces and hands
⁴⁰⁴ across age, suggesting that this is a relatively stable property of infants' – and perhaps
⁴⁰⁵ adults' – visual environment from 6 months of age.

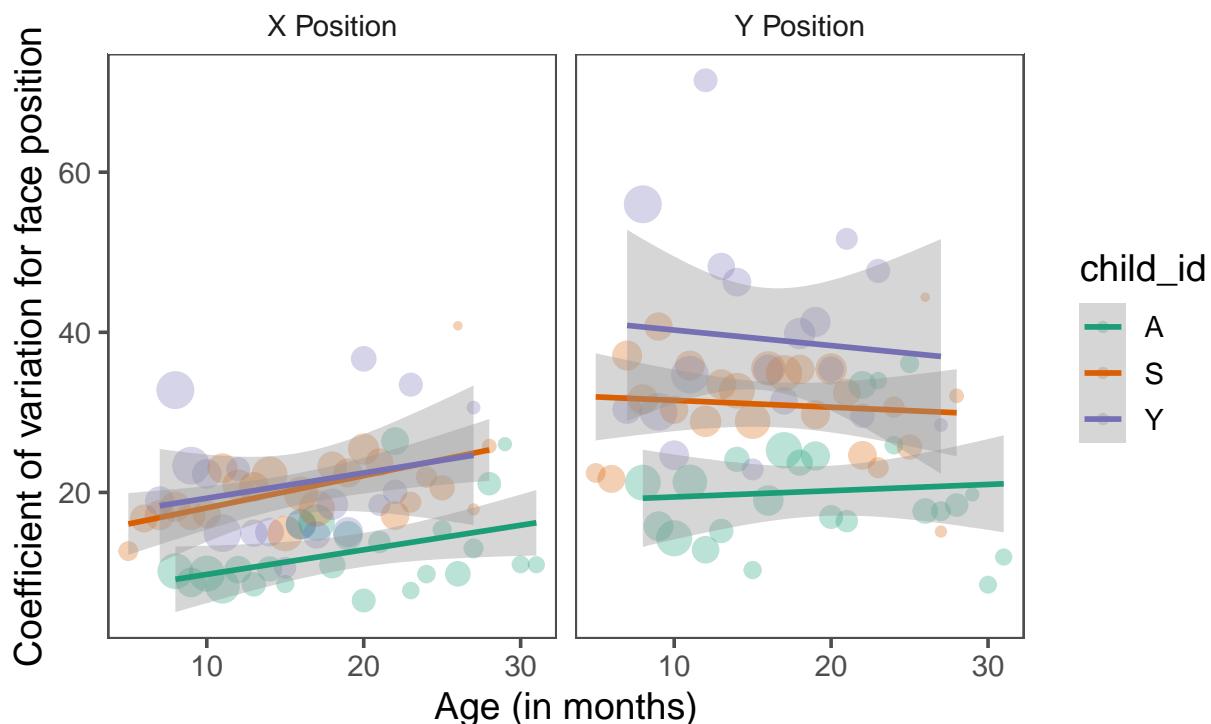


Figure B1. ToDo

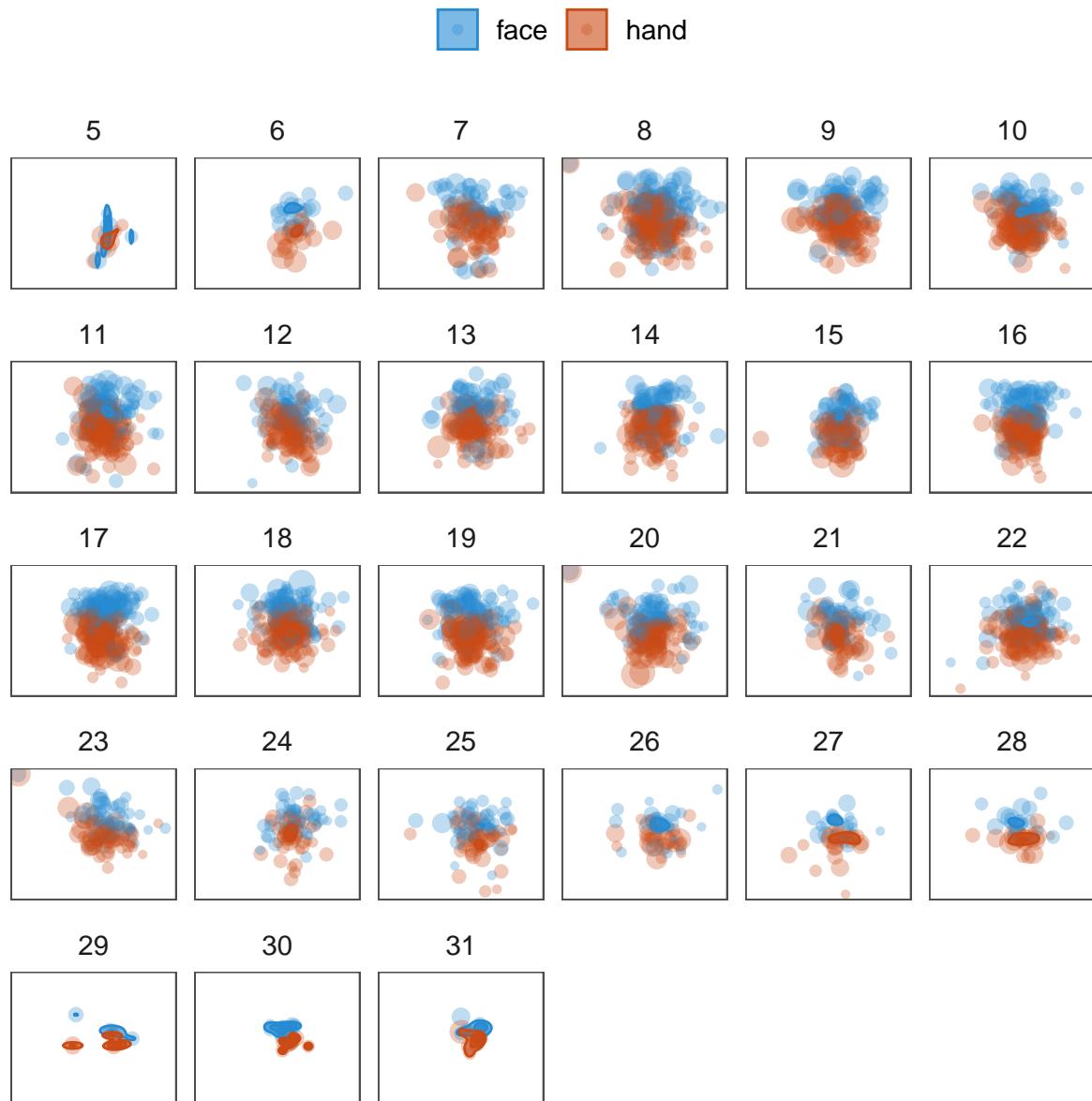


Figure B2. Each panel shows the average position of faces and hands in the visual field; each dot represents the average position from one video within a given age range.