1   Postural developments modulate children's visual access to social information

2   Bria L. Long[1], Alessandro Sanchez[1], Allison M. Kraus[1], Ketan Agrawal[1], & Michael C.

3   Frank[1]

4   [1] Department of Psychology, Stanford University

5   Author Note

6   Correspondence concerning this article should be addressed to Bria L. Long, 450 Serra

7   Mall, Stanford CA 94305. E-mail: bria@stanford.edu

8   Abstract

9   The ability to process social information is a critical component of children's early language

10   and cognitive development. However, as children reach their first birthday, they begin to

11   locomote themselves, dramatically affecting their visual access to this information. How do

12   these postural and locomotor changes affect children's access to the social information

13   relevant for word-learning? Here, we explore this question by using head-mounted cameras

14   to record 36 infants' (8-16 months of age) egocentric visual perspective and use computer

15   vision algorithms to estimate the proportion of faces and hands in infants' environments. We

16   find that infants' posture and orientation to their caregiver modulates their access to social

17   information, confirming previous work that suggests motoric developments play a significant

18   role in the emergence of children's linguistic and social capacities. In addition, we applied

19   our automated analysis to the headcam data from a recent paper showing similar effects

20   postre on visual attention (Franchak et al., 2017), finding convergence across methodologies,

21   laboratories, and datasets. We suggest that the combined use of head-mounted cameras and

22   the application of new computer vision techniques is a promising avenue for understanding

23   the statistics of infants' visual and linguistic experience as they change over development.

24   *Keywords:* Postural developments modulate children's visual access to social

25   information

26   Word count: X

²⁷   Postural developments modulate children's visual access to social information

²⁸   From their earliest months, infants are deeply engaged in learning from others. Even
²⁹ newborns tend to prefer to look at faces with direct vs. averted gaze (Farroni, Csibra, Simion,
³⁰ & Johnson, 2002) and young infants follow overt gaze shifts. And as infants learn to parse
³¹ and segment their native language(s), the social cues provided by speakers (e.g., eye-gaze)
³² may provide strong scaffolding for early word learning. Indeed, empirical work suggests that
³³ children's ability to process social cues is a key factor in early language development: in a
³⁴ longitudinal study, children's level of joint engagement with their mother at 9-12 months was
³⁵ found to predict both their receptive and productive vocabularies (Carpenter, Nagell, &
³⁶ Tomasello, 1998). More specifically, 10 month-olds who follow an adult's gaze in an
³⁷ experimental context have larger vocabularies at 18 months and throughout the second year
³⁸ of life (R. Brooks & Meltzoff, 2005; Rechele Brooks & Meltzoff, 2008).

³⁹   Around their first birthday, however, children's ability to interact with their world
⁴⁰ radically changes (Adolph & Berger, 2006) as they are no longer constrained to the same
⁴¹ spot that their caregivers last placed them in. Before the first birthday, children often begin
⁴² crawling; soon after, they begin to walk independently. While not all children crawl
⁴³ (**???**)—or crawl in the same way—children tend to find creative ways of moving (e.g.,
⁴⁴ scooting) or getting others to help them move (e.g cruising) to begin to explore their world
⁴⁵ on their own (**???**). And as independent agents in the world, children are more active in the
⁴⁶ construction of their own learning environments. Yet while toddlers do determine their own
⁴⁷ input to a much greater degree, they spend much of their time in a world primarily
⁴⁸ populated by knees

⁴⁹   One possibility is that these motor improvements have strong developmental cascades,
⁵⁰ impacting children's emerging social, cognitive, and linguistic abilities (Iverson, 2010).
⁵¹ Indeed, these postural changes change how children interact with their mothers; walking

52 infants make different kinds of object-related bids for attention from their mothers than

53 crawling infants, and tend to hear more action directed statements (e.g., "open it") (Karasik,

54 Tamis-LeMonda, & Adolph, 2014). In an observational study, Walle and Campos (2014) also

55 found that children who were able to walk had both higher receptive and productive

56 vocabularies. On their account, children's ability to stand and independently locomote may

57 fundamentally change their ability to access the social information (e.g., faces, gaze) relative

58 to children who are still crawling and sitting. In other words, the ability of walking infants

59 to access more detailed social information may allow infants to learn words quicker and more

60 efficiently, facilitating language growth.

61   Over the past decade, researchers have started to use egocentric, head-mounted

62 cameras to document the social information that infants and children have access to across

63 early development, and to understand the degree to which there are substantial shifts in

64 their viewpoints that may have downstream developmental consequences (**???**). Indeed, the

65 infant visual perspective has proved to be both remarkably different than the adult

66 perspective and not easily predicted by our own intuitions (Clerkin, Hart, Rehg, Yu, &

67 Smith, 2017; Franchak, Kretch, Soska, & Adolph, 2011, Yoshida and Smith (2008)). Over

68 the first two years of life, the infant viewpoint transitions from primarily containing close up

69 view of faces to capturing relatively restricted views of hands paired with the objects they

70 are acting on (Fausey, Jayaraman, & Smith, 2016); both computational and empirical work

71 suggest that this restricted viewpoint is more effective for learning about objects and their

72 labels than the comparable adult perspective (D. Yurovsky, Smith, & Yu, in press, Bambach,

73 Crandall, Smith, and Yu (2017)).

74   Here, we directly examine the role that postural developments that infants' experience

75 as they reach their first birthday change the social information in the infant view. We

76 hypothesize that children's postural developments may be partially responsible for some of

77 these changes in the infant perspective. During spontaneous play, toddlers are more likely to

78 look at the floor while crawling than while walking (Franchak et al., 2011), when they have

79 full visual access to their environment and the people in it (Kretch, Franchak, and Adolph,

80 2014). More recently, when 12-month-olds (**???**) participated in a free-play session with

81 their caregivers, their own posture and as well as their caregiver's posture influenced the

82 proportion of time they spent looking at the faces and bodies of their caregivers (**???**).

83   Here, we expand on these previous findings in three key ways. First, using

84 head-mounted cameras during a brief laboratory free-play session, we record the visual

85 experience of three groups of children at 8, 12, and 16 months-of-age, covering a broad range

86 of ages and locomotor abilities around the first birthday; Children's posture and orientation

87 relative to their caregiver were also recorded from a third-person perspective and

88 hand-annotated. This cross-sectional design thus allows us to directly examine the relative

89 contributions of age vs. postural developments on children's visual access to social

90 information. Second, we use a modern computer vision algorithm for the automated

91 detection of faces and hands in the egocentric viewpoints of infants. In particular, we

92 capitalize on recent improvements in face and pose detection algorithms (Cao, Simon, Wei,

93 & Sheikh, 2017; K. Zhang, Zhang, Li, & Qiao, 2016) to analyze the frequencies of faces and

94 hands in the child's visual environment, both overall and relative to naming events by their

95 caregivers. Use of this emerging technology allows us to annotate the entirety of the dataset,

96 allowing a more complete picture of the changing infant perspective. In addition, we extend

97 this technique to the same dataset used in (**???**), replicating their main pattern of results

98 using this novel methodology and finding convergence between our datasets with different

99 play sessions, populations, and head-mounted cameras. Finally, as parents were presented

100 with pairs of novel and familiar objects with their children during the play sessions, we

101 directly examine the availability of social signals during naming events.

102   Thus, our current dataset allows us to analyze changes in the visual access to faces and

103 hands according to children's age, posture, and linguistic input. We predicted that there

104 would be differential access to social information based on children's postural developments:

105 crawling infants would see fewer faces/hands because they would primarily be looking at the

106 ground, while walking toddlers would have access to a richer visual landscape with greater

107 access to the social information in their environment. [language predictions?]

Methods

108

109 **Participants.**   Our final sample consisted of 36 infants and children, with 12

110 participants in three age groups: 8 months (6 F), 12 months (7 F), and 16 months (6 F).

111 Participants were recruited from the surrounding community via state birth records, had no

112 documented disabilities, and were reported to hear at least 80 percent English at home.

113 Demographics and exclusion rates are given in the table below.

Group	N	% incl.	Avg age	Avg video length (min)
8 mo.	12	0.46	8.71	14.41
12 mo.	12	0.40	12.62	12.71
16 mo.	12	0.31	16.29	15.10

114 To obtain this final sample, we tested 95 children, excluding 59 children for the

115 following reasons: 20 for technical issues related to the headcam, 15 for failing to wear the

116 headcam, 10 for fewer than 4 minutes of headcam footage, 5 for having multiple adults

117 present, 5 for missing Communicative Development Inventory (CDI) data, 2 for missing

118 scene camera footage, 1 for fussiness, and one for sample symmetry. All inclusion decisions

119 were made independent of the results of subsequent analyses. Some of these data were also

120 analyzed in Frank, Simmons, Yurovsky, and Pusiol (2013).

121 **Procedure.**   All parents signed consent documents while children were fitted with

122 the headcam. If the child was uninterested in wearing the headcam or tried to take it off, the

123 experimenter presented engaging toys to try to draw the child's focus away from the

124 headcam. When the child was comfortable wearing the headcam, the child and caregiver

125 were shown to a playroom for the free-play session. Parents were shown a box containing

126 three pairs of novel and familiar objects (e.g., a ball and a microfiber duster, named a

127 "zem"), and were instructed to play with the object pairs with their child one at a time, "as

128 they typically would." All parents confirmed that their child had not previously seen the

129 novel toys and were instructed to use the novel labels to refer to the toys. The experimenter

130 then left the playroom for approximately 15 minutes, during which a tripod-mounted camera

131 in the corner of the room recorded the session and the headcam captured video from the

132 child's perspective.

133 **Data Processing and Annotation.** Headcam videos were trimmed such that they

134 excluded the instruction phase when the experimenter was in the room and were

135 automatically synchronized with the tripod-mounted videos using FinalCut Pro Software.

136 These sessions yielded 507 minutes (almost a million frames) of video, with an average video

137 length of 14.07 minutes (min = 4.53, max = 19.35).

138 **Posture and Orientation Annotation.** We created custom annotations to

139 describe the child's physical posture (e.g. standing) and the orientation of the child relative

140 to the caregiver (e.g. far away). The child's posture was categorized as being held/carried,

141 prone (crawling or lying), sitting, or standing. The caregiver's orientation was characterized

142 as being close, far, or behind the child (independent of distance). For the first two

143 annotations (close/far from the child), the caregiver could either be to the front or side of

144 the child. All annotations were made by a trained coder using the OpenSHAPA/Datavyu

145 software (Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012). Times when the child

146 was out of view of the tripod camera were marked as uncodable and were excluded from

147 these annotations.

Face and Hand Detection

We used three face detection systems to measure infants' access to faces. The first of these is the most commonly-used and widely available face detection algorithm: Viola-Jones. We used this algorithm as a benchmark for performance, as while it can achieve impressive accuracy in some situations, it is notoriously bad at dealing with occluded faces (Scheirer, Anthony, Nakayama, & Cox, 2014). We next tested the performance of two face detectors that both made use of recently developed Convolutional Neural Networks (CNNs) to extract face information. The first algorithm was specifically optimized for face detection, and the second algorithm was optimized to extract information about the position of 18 different body parts. For the second algorithm (OpenPose; Cao et al., 2017), we used the agent's nose (one of the body parts detected) to operationalize the presence of faces, as any half of a face necessarily contains a nose.

The OpenPose detector also provided us with the location of an agent's wrists, which we used as a proxy for hands for two reasons. First, as we did not capture children's entire visual field, the presence of a wrist is likely often indicative of the presence of a hand within the field of view. Second, hands are often occluded by objects when caregivers are interacting with children, yet still visually accessible by the child and part of their joint interaction.

Algorithms. The first face detection system made use of a series of Haar feature-based cascade classifiers (Viola & Jones, 2004) applied to each individual frame. The second algorithm (based on work by K. Zhang et al. (2016)) uses multi-task cascaded convolutional neural networks (MTCNNs) for joint face detection and alignment, built to perform well in real-world environments where varying illuminations and occlusions are present. We used a Tensorflow implementation of this algorithm avaliable at https://github.com/davidsandberg/facenet.

The CNN-based pose detector (OpenPose; Cao et al., 2017; Simon, Joo, Matthews, &

173 Sheikh, 2017; Wei, Ramakrishna, Kanade, & Sheikh, 2016) provided the locations of 18 body

174 parts (ears, nose, wrists, etc.) and is available at

175 https://github.com/CMU-Perceptual-Computing-Lab/openpose. The system uses a CNN

176 for initial anatomical detection and subsequently applies part affinity fields (PAFs) for part

177 association, producing a series of body part candidates. The candidates are then matched to

178 a single individual and finally assembled into a pose; here, we only made use of the body

179 parts relevant to the face and hands (nose and wrists).

180   **Detector evaluation.**   To evaluate face detector performance, we hand-labeled a

181 "gold set" of labeled frames. To account for the relatively rare appearance of faces in the

182 dataset, we hand-labeled two types of samples: a sample containing a high density of faces

183 (half reported by MTCNN, half by OpenPose) and a random sample from the remaining

184 frames. Each sample was comprised of an equal number of frames taken from each child's

185 video. For wrist detections, the "gold set" was constructed in the same manner, except

186 frames with a high density of wrists came only from detections made by OpenPose. Faces

187 were classified as present if at least half of the face was showing; wrists were classified as

188 present if any part of the wrist was showing. Two authors labelled the frames independently

189 and resolved disagreements on a case-by-case basis. Precision (hits / hits + false alarms),

190 recall (hits / hits + misses), and F-score (harmonic mean of precision and recall) were

191 calculated for all detectors and are reported in Table 1.

192   For face detection, MTCNN outperformed OpenPose when taking into account only

193 the composite F-score (0.89 MTCNN vs. 0.83 OpenPose). Although MTCNN and OpenPose

194 performed comparably with the random sample, MTCNN performed better on the high

195 density sample (specifically looking at precision), suggesting that OpenPose generated more

196 false positives than MTCNN. ViolaJones performed quite poorly relative to the other

197 detectors, especially with respect to the random sample. We thus use MTCNN detections in

198 the following analyses. For wrist detection, OpenPose performed moderately well (F = 0.74)

¹⁹⁹ with relatively high precision but low recall on the randomly sampled frames (see Table 1).

²⁰⁰ We thus analyze wrist detections, with the caveat that we are likely underestimating the

²⁰¹ proportion of hands in the dataset.

Algorithm	Sample Type	P	R	F
MTCNN-Faces	High density	0.89	0.92	**0.90**
MTCNN-Faces	Random	0.94	0.62	0.75
OpenPose-Faces	High density	0.78	0.93	0.84
OpenPose-Faces	Random	0.72	0.80	**0.76**
ViolaJones-Faces	High density	0.96	0.44	0.60
ViolaJones-Faces	Random	0.44	0.38	0.41
OpenPose-Wrists	High density	0.66	1.00	0.79
OpenPose-Wrists	Random	0.88	0.29	0.44

Table 1

Detector performance on both high density samples (where proportion of targets detected was high) and random samples (where frames were randomly selected). P, R, and F denote precision, recall, and F-score, respectively. Scores in bold are the highest F-scores for each sample type.

Results

²⁰³ First, we report developmental shifts in infants' posture and their orientation relative

²⁰⁴ to their caregiver, consistent with previous literature (**???**, **???**). Then, we examine how

²⁰⁵ these changes influence children's visual access to faces and wrists/hands across this

²⁰⁶ developmental time range, examining the relative contributions of age vs. postural

²⁰⁷ developments. We also explore how these changes impact the accessibility of faces and

²⁰⁸ wrists/hands during labeling events, both when parents were labeling familiar objects (e.g.,

²⁰⁹ "cat") as well as novel objects (e.g., "zem"). Finally, we apply the same automated detection

²¹⁰ method to a egocentric video dataset collected by a different lab (**???**) during which

²¹¹ children's in-the-moment posture was annotated, replicating their existing results and finding

²¹² converging support across methods, labs, and paradigms for the impact of postural

²¹³ developments on children's visual access to social information.

Changes in Posture and Orientation

²¹⁵ How do childen's in-the-moment posture and orientation change across age? In our

²¹⁶ dataset, the proportion of time infants spent sitting decreased with age, and the proportion

²¹⁷ of time infants spent standing increased with infants' age. As children got older, their

²¹⁸ locomotive abilities allowed them to become more independent. Both 8-month-olds and

²¹⁹ 12-month-olds spent relatively equivalent amounts of time lying/crawling (i.e., "prone")

²²⁰ which was markedly decreased in the 16-month-olds, who spent most of their time sitting or

²²¹ standing (see Figure 2). We also observed changes in children's orientation relative to their

²²² caregivers: the 8-month-olds spent more time with their caregiver behind them supporting

²²³ their sitting positions than did children at other ages (see Figure 2). However, we also saw

²²⁴ considerable variability across children: some infants spent almost their entire time sitting at

²²⁵ a close distance from their caregiver, whereas others showed more considerable variability

²²⁶ (see Figure 3).

Changes in Access to Faces and Hands

²²⁸ We first examined the proportion of face and hand detections as a function of

²²⁹ children's age without considering their posture (see Figure 4). While faces tended to be in

²³⁰ the field-of-view overall more often than hands, children's head-mounted cameras were

²³¹ angled slightly upward to capture the presence of faces, and hand detections suffered from

²³² somewhat lower recall than face detections. Going forward, we thus analyze differences in

233 the relative proportion of faces or hands in view as a function of age, posture, and

234 orientation, rather than comparing them directly. Overall, we observed that 12-month-olds

235 appeared to have visual access to slightly fewer faces than 8 or 16-month-olds, creating a

236 slight U-shaped function in face detections; conversely, hand detections were showed a slight

237 increase across this age range, as reported in prior literature (**???**).

238   However, these age-related trends were much smaller than the effect of infant's

239 postural developments on children's visual access to faces and hands. Children's

240 in-the-moment posture was a major factor both in how many faces and hands were in view

241 during the play session, as was their orientation relative to their caregiver. Infants who were

242 sitting or standing had more faces in view than infants who were lying down/crawling

243 (i.e. prone), which was most frequent among 12-month-olds relative to the other age groups

244 (Figure 5). When caregivers were behind their children, supporting their children's sitting or

245 standing positions, children saw fewer faces and wrists. In particular, children who were

246 sitting or in front of their caregiver had a high proportion of faces and hands in their field of

247 view. (Figure 5, lower panel).

248   These trends were quantified using two generalized linear mixed-effect models

249 estimating the proportion of faces and hands that were in view, with orientation, posture,

250 their interaction, and scaled participant's age as fixed effects, and with random slopes for

251 infants' orientation and posture. A summary of the coefficients of the models can be found

252 in Table 2, confirming that infants who were sitting/standing and in front of their caregivers

253 saw the most faces and hands. When modeling the proportion of faces seen, age was not a

254 significant predictor; however, age remained a significant predictor when modeling the

255 proportion of hands seen by infants. Overall, these results suggest that infants' visual access

256 to social information is largely modulated by their posture and orientation to their caregiver,

257 which is in turn a function of their general locomotor development. Nonetheless, these

258 results also suggest that infants may still increase their attention to hands throughout this

259 age range as they continue to learn about objects, their functions, and their names.

Access to Faces and Hands During Labeling Events

261 Our play session was designed to provide parents with opporunities to label
262 objects–both familiar and novel–such that we could examine whether children seek out
263 different kinds of social information around naming events. We thus explored how face and
264 hand detections changed during object labeling events, analyzing a four-second window
265 around each labeling event (e.g., "Look at the [zem]!"). Every utterance of one of the
266 labelled objects (e.g., "ball") was counted as a "labeling event"; timestamps of the beginning
267 of each word were hand-annotated from transcripts ad synchronized with the frame-by-frame
268 detections. Overall, we did not find major differences in face/hand detections during naming
269 events relative to baseline, either for novel objects or for familiar objects (see Figure 6). This
270 was true both when we analyzed naming rates with or without taking into account children's
271 posture and orientation relative to their caregiver. These results suggest that either children
272 did not actively seek out social information during this particular play session around naming
273 events, perhaps because were a limited number of possible referents at any given time.

Extension to Franchak et al., 2017

275 In the present work, we found that infant's in-the-moment posture changed with their
276 age, as did infants' orientation relative their caregiver. In a related study with
277 12-month-olds, (**???**) also that children's in-the-moment posture changed the amount of
278 time infants spent looking at faces. Here, we sought to replicate the findings from (**???**)
279 using our automated methodology (OpenPose detections), using the footage from their
280 head-mounted cameras (shared on Databrary, (**???**)).

281 This dataset differed from the present in two key ways that could present a challenge

for our automated methodology. First, the environment that infants were immersed in with their caregivers (and experimenters) was much larger and more varied than the play room used in the present dataset, containing multiple structures and toys in different parts of the room for infants to explore, unlike the present room dataset which was relatively small (approximately 10' x 10'). Thus, our automated methodology could fail to generalize to scenes from these more complex environments, where detecting faces and hands could arguably be a much harder task. Second, (**???**) used a head-mounted eye-tracker, finding that children's posture affected where children were looking within their visual field. Nonetheless, if children are often orienting their heads towards where they are allocating their attention (**???**) then we should expect to find the same pattern of results only when analyzing the information in view.

Overall, we found convergence between our two methodologies: we replicated (**???**)"s main results, finding that the proportion of faces in view was greater when infants were sitting or standing vs. prone. We also found the same patterns of results with hands (i.e., wrist detectinos), which were not originally annotated in (**???**) (see Figure 7); results were validated with generalized mixed effect models as in our previous analysis. Interestingly, there was a relatively higher proportion of hands in view in (**???**) relative to the present dataset; this is likely mostly due to the fact that there were often multiple people in view: experimenters stayed in the room to film children and assist with technical difficulties, while in our present dataset the experiments left the room for the majority of the play session. Nonetheless, we still found that posture modulated the hands in view as well, suggesting that this is a major factor that structures infants" access to visual information.

Discussion

References

Adolph, K. E., & Berger, S. E. (2006). Motor development. *Handbook of Child Psychology.*

Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry, 23*(3), 244–247.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2017). An egocentric perspective on active vision and visual object learning in toddlers. In *Proceedings of the seventh joint ieee conference on development and learning and on epigenetic robotics.*

Brooks, R., & Meltzoff, A. (2005). The development of gaze following and its relation to language. *Developmental Science, 8*(6), 535–543.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language, 35*(1), 207–220.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR.*

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63*(4).

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

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

328   9602–9605.

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

331   Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye
332   tracking: A new method to describe infant looking. *Child Development*, *82*(6),
333   1738–1750.

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

337   Iverson, J. M. (2010). Developing language in a developing body: The relationship between
338   motor development and language development. *Journal of Child Language*, *37*(2),
339   229–261.

340   Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants
341   elicit different verbal responses from mothers. *Developmental Science*, *17*(3), 388–395.

342   Scheirer, W. J., Anthony, S. E., Nakayama, K., & Cox, D. D. (2014). Perceptual annotation:
343   Measuring human vision to improve computer vision. *IEEE Transactions on Pattern*
344   *Analysis and Machine Intelligence*, *36*(8), 1679–1686.

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

347   Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of*
348   *Computer Vision*, *57*(2), 137–154.

349   Walle, E. A., & Campos, J. J. (2014). Infant language development is related to the

₃₅₀   acquisition of walking. *Developmental Psychology, 50*(2), 336.

₃₅₁ Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines.

₃₅₂   In *CVPR*.

₃₅₃ Yoshida, H., & Smith, L. (2008). What's in view for toddlers? Using a head camera to study

₃₅₄   visual experience. *Infancy, 13*, 229–248.

₃₅₅ Yurovsky, D., Smith, L., & Yu, C. (in press). Statistical word learning at scale: The baby's

₃₅₆   view is better. *Developmental Science.*

₃₅₇ Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using

₃₅₈   multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10),

₃₅₉   1499–1503.
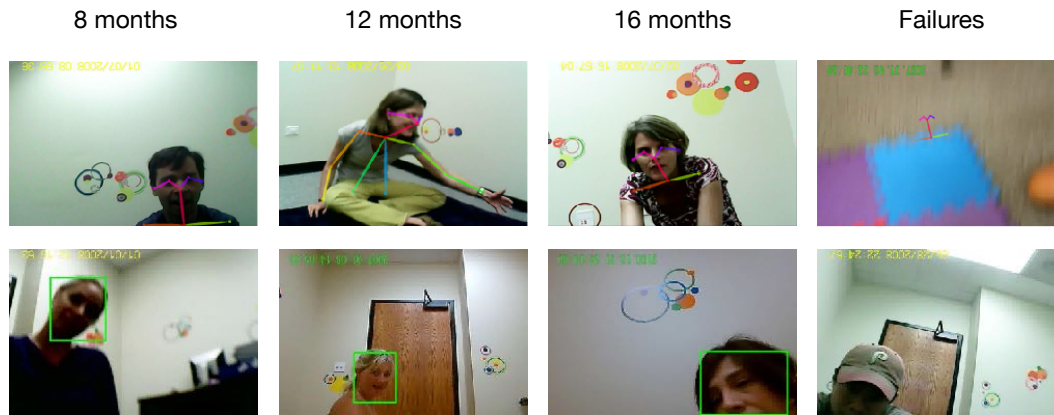
Figure 1. Example face and pose detections made by OpenPose (top row) and MTCNN (bottom row) from a child in each age group. The last column features a false positive from OpenPose and a false negative from MTCNN.
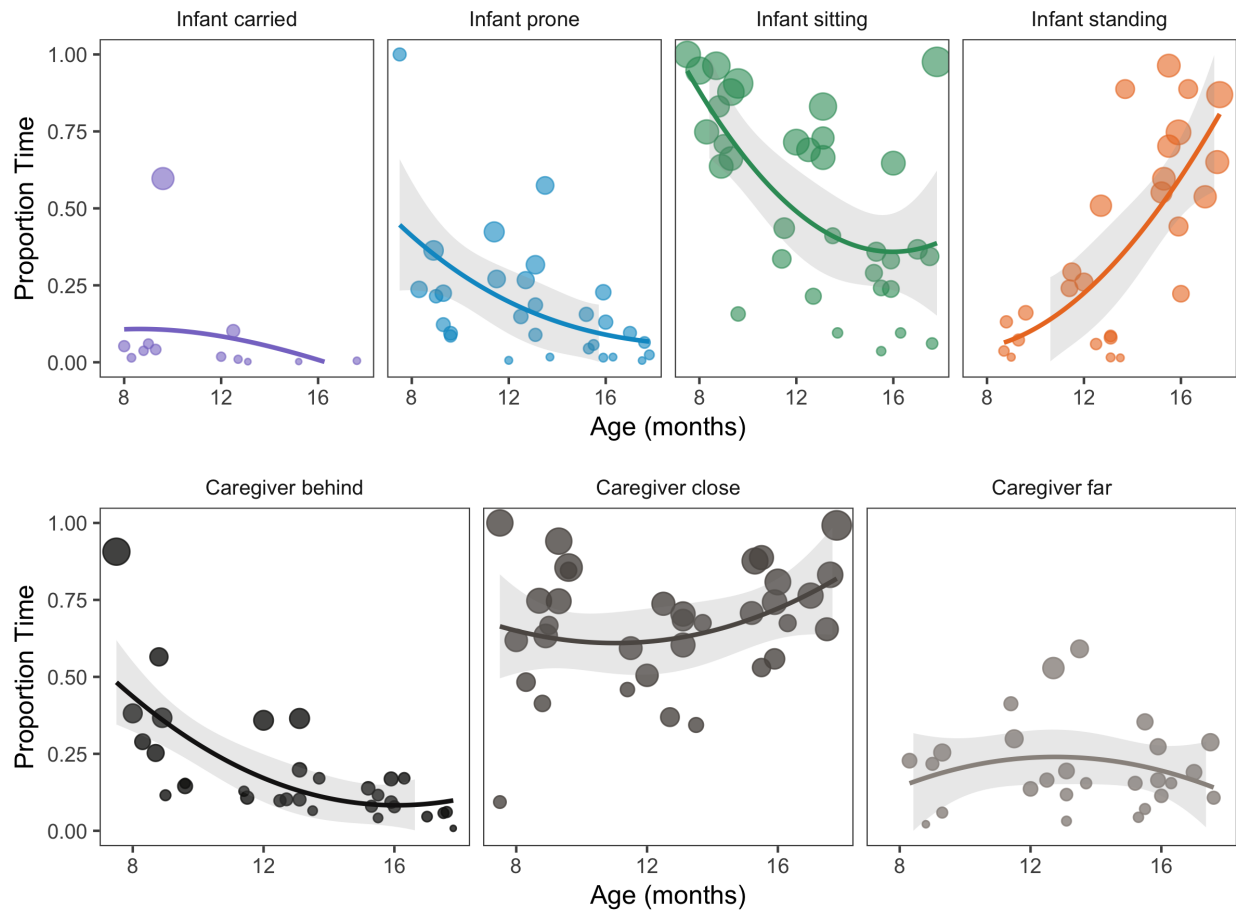
Figure 2. Proportion of time spent by each infant in different postures and orientations relative to their caregivers (CG); times where posture was not codable are ommitted for visualization purposes
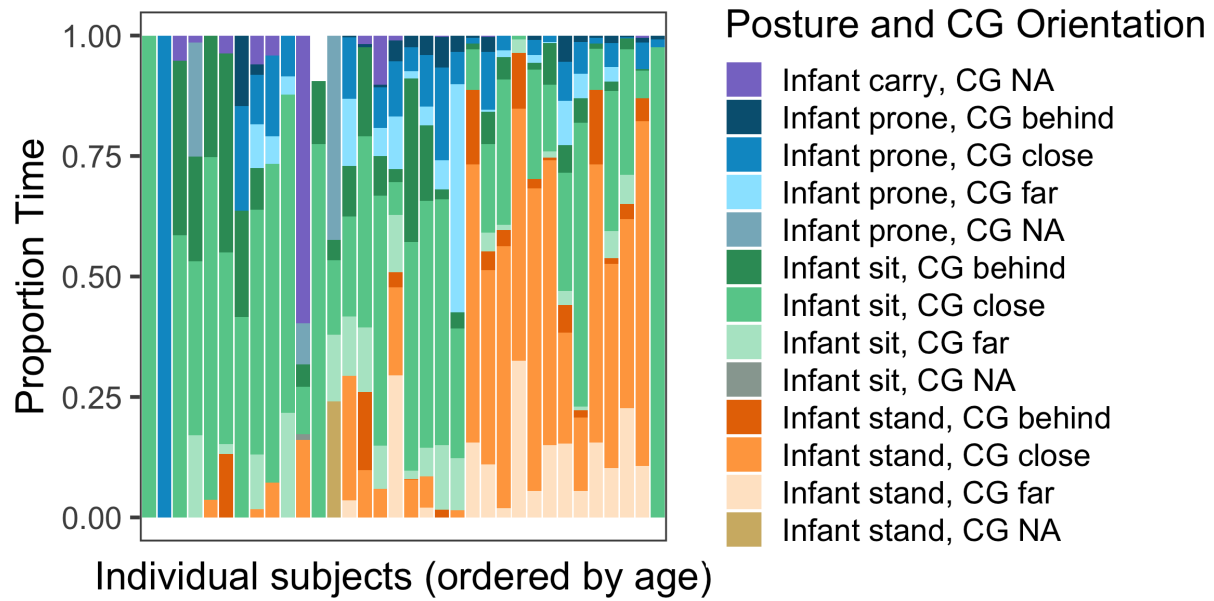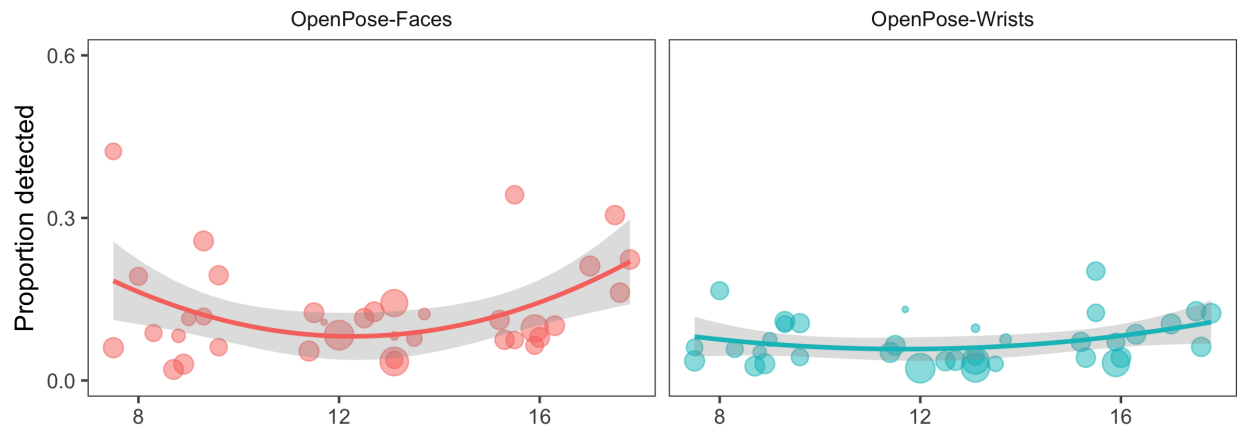
Figure 3. Proportion of time spent by each infant in different postures and orientations relative to their caregivers (CG); times where posture was not codable are ommitted for visualization purposes

Figure 4. Proportion of faces (left) and wrists (right) detected by the OpenPose model as a function of child's age. Larger dots indicate children who had longer play sessions and thus for whom there was more data.
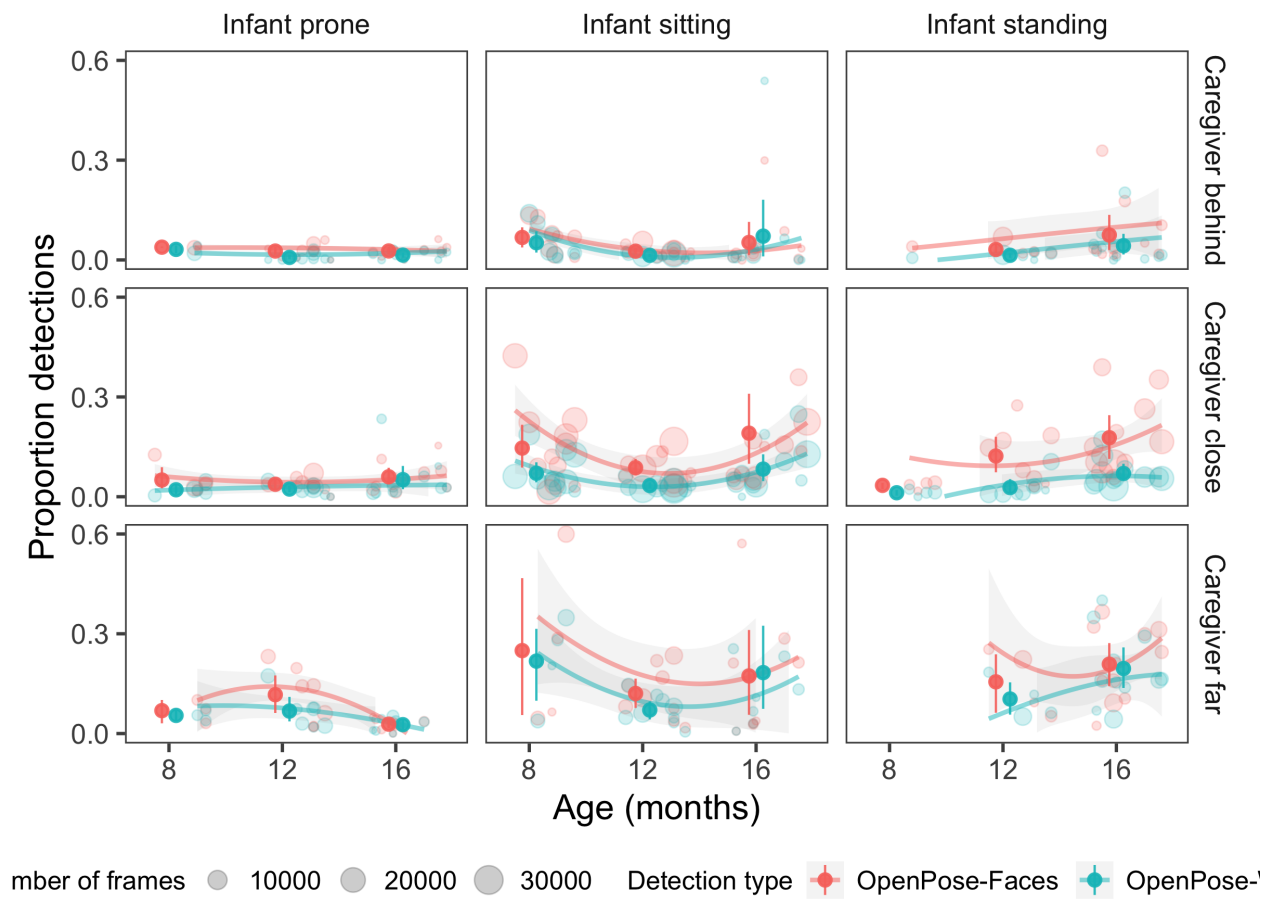
Figure 5. Proportion of face / wrist detections by children's age, their posture, and their caregivers orientation. Data points are scaled by the amount of time spent in each orientation/posture combination; times when posture/orientation annotatinos were unavaliable or the infant was carried are not plotted. Error bars represent 95% bootstrapped confidence intervals
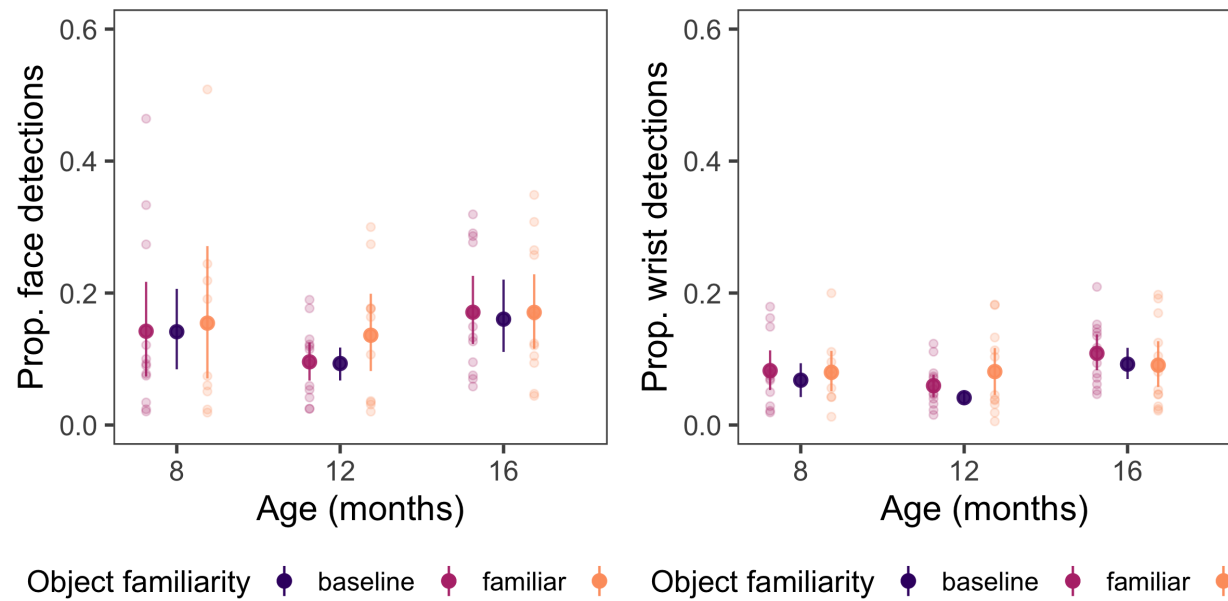
Figure 6. Proportion of face / wrist detections during naming events (+/- 2 seconds around label)for familiar and novel objects; these rates are put into context relative to baseline Error bars represent 95% bootstrapped confidence intervals.
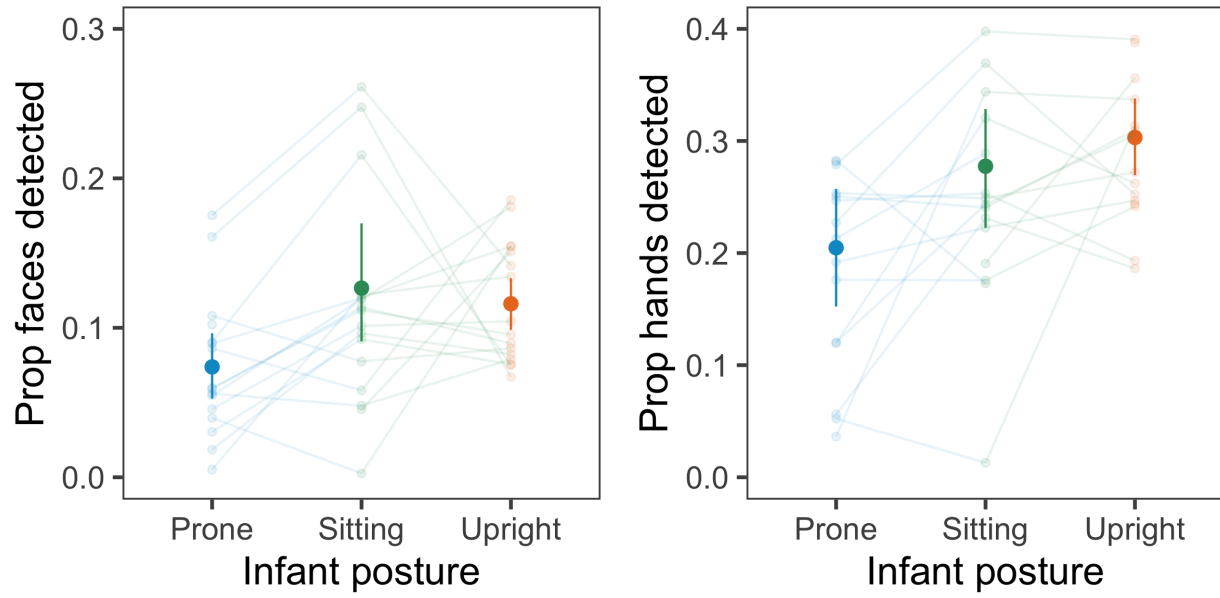
Figure 7. Proportion of face / wrist detections for 12-month-olds in Franchak et al., 2017 as a function of children's in-the-moment posture. Error bars represent 95 percent bootstrapped confidence intervals.