Postural developments modulate children's visual access to social information

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

[1] Department of Psychology, Stanford University

Author Note

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

Abstract

The ability to process social information is a critical component of children's early linguistic and cognitive development. However, as children reach their first birthday, they begin to locomote themselves, dramatically changing the way they see the world around them. How do these postural and locomotor developments affect children's access to the social information relevant for word-learning? We explored this question by using head-mounted cameras to record the egocentric visual perspective of 36 infants' (at 8, 12, and 16 months of age) during a naturalistic play session with their caregiver. To estimate the proportion of faces and hands in the infant view, we used a computer vision algorithm to detect the presence of faces and hands in the entire dataset. We found that infants\'posture and orientation to their caregiver changed dramatically across this age range and modulated their access to this social information; infants who were sitting or standing with their caregiver at a close distance tended to have the most faces/hands in their visual field. We also applied our automated analysis to the egocentric video data from a recent paper documenting how posture changes one-year-olds visual access to faces (Franchak et al., 2017), finding convergence across both methodologies and dataset and confirming previous work that motoric developments play a significant role in the emergence of children's linguistic and social capacities. We suggest that the combined use of head-mounted cameras and the application of new computer vision techniques is a promising avenue for understanding the statistics of infants\' visual and linguistic experience as they change over development.

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

Word count: X

Postural developments modulate children's visual access to social information

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

Around their first birthday, however, children's ability to interact with their world radically changes (Adolph & Berger, 2006) as they are no longer constrained to the same spot that their caregivers last placed them in. Before the first birthday, children often begin crawling; soon after, they begin to walk independently. While not all children crawl (Adolph, Vereijken, & Denny, 1998)—or crawl in the same way—children tend to find creative ways of moving (e.g., scooting) or getting others to help them move (e.g cruising) to begin to explore their world on their own (Patrick, Noah, & Yang, 2012). As independent agents in the world, children are more active in the construction of their own learning environments–yet spend much of their time in a world primarily populated by knees

One possibility is that these motor improvements have strong developmental cascades, impacting children's emerging social, cognitive, and linguistic abilities (Iverson, 2010). Indeed, these postural changes change how children interact with their mothers; walking infants make different kinds of object-related bids for attention from their mothers than

crawling infants, and tend to hear more action directed statements (e.g., "open it") (Karasik, Tamis-LeMonda, & Adolph, 2014). In an observational study, Walle and Campos (2014) also found that children who were able to walk had both higher receptive and productive vocabularies. On their account, children's ability to stand and independently locomote may fundamentally change their ability to access the social information (e.g., facial expressions, gaze cues, pointing) relative to children who are still crawling and sitting. In other words, the ability of walking infants to access more detailed social information may allow infants to learn words quicker and more efficiently, facilitating language growth.

Over the past decade, researchers have started to use egocentric, head-mounted cameras to document the social information that infants and children have access to across early development, and to understand the degree to which there are substantial shifts in their viewpoints that may have downstream developmental consequences (Yoshida & Smith, 2008). Indeed, the infant visual perspective has proved to be both remarkably different than the adult perspective and not easily predicted by our own intuitions (Franchak, Kretch, Soska, & Adolph, 2011, Yoshida and Smith (2008), Clerkin, Hart, Rehg, Yu, and Smith (2017)). Over the first two years of life, the infant viewpoint seems to transition from primarily containing close up view of faces to capturing relatively restricted views of hands paired with the objects they are acting on (Fausey, Jayaraman, & Smith, 2016); both computational and empirical work suggest that this restricted viewpoint may be more effective for learning about objects and their labels than the comparable adult perspective (D. Yurovsky, Smith, & Yu, 2012, Bambach, Crandall, Smith, and Yu (2017)).

We hypothesized that children's postural developments may be partially responsible for some of these changes in the infant perspective. During spontaneous play, toddlers are more likely to look at the floor while crawling than while walking (Franchak et al., 2011), when they have full visual access to their environment and the people in it (Kretch, Franchak, and Adolph, 2014). More recently, when 12-month-olds (Franchak, Kretch, & Adolph, 2017)

participated in a free-play session with their caregivers, their own posture and as well as their caregiver's posture influenced the proportion of time they spent looking at the faces and bodies of their caregivers (Franchak et al., 2017).

Here, we directly examine the role that postural developments that infants' experience as they reach their first birthday change the social information in the infant view. We expand on previous findings in three key ways. First, using head-mounted cameras during a brief laboratory free-play session, we record the visual experience of three groups of children at 8, 12, and 16 months-of-age, covering a broad range of ages and locomotor abilities around the first birthday; Children's posture and orientation relative to their caregiver were also recorded from a third-person perspective and hand-annotated. This cross-sectional design thus allows us to directly examine the relative contributions of age vs. postural developments on children's visual access to social information. Second, we use a modern computer vision algorithm for the automated detection of faces and hands in the egocentric viewpoints of infants. In particular, we capitalize on recent improvements in face and pose detection algorithms (Cao, Simon, Wei, & Sheikh, 2017; K. Zhang, Zhang, Li, & Qiao, 2016) to analyze the frequencies of faces and hands in the child's visual environment, both overall and relative to naming events by their caregivers. Use of this emerging technology allows us to annotate the entirety of the dataset, allowing a more complete picture of the changing infant perspective. We apply this same automated method to the dataset used in Franchak et al. (2017), validating the generalizability of this technique. Third, as parents were presented with pairs of novel and familiar objects with their children during the play sessions, we were able to explore the availability of social signals during naming events.

Thus, the current study allows us to analyze changes in the visual access to faces and hands according to children's age, posture, and linguistic input. Broadly, we predicted that there would be differential access to social information based on children's postural developments: crawling infants would see fewer faces/hands because they would primarily be

looking at the ground, while walking toddlers would have access to a richer visual landscape with greater access to the social information in their environment.

Methods

Participants. Our final sample consisted of 36 infants and children, with 12 participants in three age groups: 8 months (6 F), 12 months (7 F), and 16 months (6 F). Participants were recruited from the surrounding community via state birth records, had no documented disabilities, and were reported to hear at least 80 percent English at home. Demographics and exclusion rates are given in the table below.

To obtain this final sample, we tested 95 children, excluding 59 children for the following reasons: 20 for technical issues related to the headcam, 15 for failing to wear the headcam, 10 for fewer than 4 minutes of headcam footage, 5 for having multiple adults present, 5 for missing Communicative Development Inventory (CDI) data, 2 for missing scene camera footage, 1 for fussiness, and one for sample symmetry. All inclusion decisions were made independent of the results of subsequent analyses. These data were also analyzed in Frank, Simmons, Yurovsky, and Pusiol (2013) and Sanchez, Long, Kraus, and Frank (2018).

Head-mounted camera. We used a small, head-mounted camera ("headcam") that was constructed from a MD80 model camera attached to a soft elastic headband. Videos captured by the headcam were 720x480 pixels with 25 frames per second.[1] A fisheye lens was attached to the camera to increase the view angle from 32° horizontal by 24° vertical to 64° horizontal by 46° vertical (see Figure 1, above).

Even with the fish-eye lens, the vertical field of view of the camera is still considerably reduced compared to the child's field of view, which spans around 100–120° in the vertical

[1]Detailed instructions for creating this headcam can be found at http://babieslearninglanguage.blogspot.com/2013/10/how-to-make-babycam.html.

dimension by 6-7 months of age (Cummings, Van Hof-Van Duin, Mayer, Hansen, & Fulton, 1988; Mayer, Fulton, & Cummings, 1988). As we were primarily interested in the presence of faces in the child's field of view, we chose to orient the camera upwards to capture the entirety of the child's upper visual field where the child is likely to see adult faces, understanding that this decision limited our ability to detect hands (especially those of the child, which are typically found at the bottom of the visual field).

Procedure. All parents signed consent documents while children were fitted with the headcam. If the child was uninterested in wearing the headcam or tried to take it off, the experimenter presented engaging toys to try to draw the child's focus away from the headcam. When the child was comfortable wearing the headcam, the child and caregiver were shown to a playroom for the free-play session. Parents were shown a box containing three pairs of novel and familiar objects (e.g., a ball and a microfiber duster, named a "zem"), and were instructed to play with the object pairs with their child one at a time, "as they typically would." All parents confirmed that their child had not previously seen the novel toys and were instructed to use the novel labels to refer to the toys. The experimenter then left the playroom for approximately 15 minutes, during which a tripod-mounted camera in the corner of the room recorded the session and the headcam captured video from the child's perspective.

Data Processing and Annotation. Headcam videos were trimmed such that they excluded the instruction phase when the experimenter was in the room and were automatically synchronized with the tripod-mounted videos using FinalCut Pro Software. These sessions yielded 507 minutes (almost a million frames) of video, with an average video length of 14.07 minutes (min = 4.53, max = 19.35).

Posture and Orientation Annotation. We created custom annotations to describe the child's physical posture (e.g. standing) and the orientation of the child relative to the caregiver (e.g. far away). The child's posture was categorized as being held/carried,

prone (crawling or lying), sitting, or standing. The caregiver's orientation was characterized as being close, far, or behind the child (independent of distance). For the first two annotations (close/far from the child), the caregiver could either be to the front or side of the child. All annotations were made by a trained coder using the OpenSHAPA/Datavyu software (Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012). Times when the child was out of view of the tripod camera were marked as uncodable and were excluded from 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 available at https://github.com/davidsandberg/facenet.

The CNN-based pose detector (OpenPose; Cao et al., 2017; Simon, Joo, Matthews, & Sheikh, 2017; Wei, Ramakrishna, Kanade, & Sheikh, 2016) provided the locations of 18 body parts (ears, nose, wrists, etc.) and is available at https://github.com/CMU-Perceptual-Computing-Lab/openpose. The system uses a CNN for initial anatomical detection and subsequently applies part affinity fields (PAFs) for part association, producing a series of body part candidates. The candidates are then matched to a single individual and finally assembled into a pose; here, we only made use of the body parts relevant to the face and hands (nose and wrists).

Detector evaluation. To evaluate face detector performance, we hand-labeled a "gold set" of labeled frames. To account for the relatively rare appearance of faces in the dataset, we hand-labeled two types of samples: a sample containing a high density of faces (half reported by MTCNN, half by OpenPose) and a random sample from the remaining frames. Each sample was comprised of an equal number of frames taken from each child's video. For wrist detections, the "gold set" was constructed in the same manner, except frames with a high density of wrists came only from detections made by OpenPose. Faces were classified as present if at least half of the face was showing; wrists were classified as present if any part of the wrist was showing. Two authors labelled the frames independently and resolved disagreements on a case-by-case basis. Precision (hits / hits + false alarms), recall (hits / hits + misses), and F-score (harmonic mean of precision and recall) were

calculated for all detectors and are reported in Table 1.

For face detection, MTCNN outperformed OpenPose when taking into account only the composite F-score (0.89 MTCNN vs. 0.83 OpenPose). MTCNN and OpenPose performed comparably with the random sample, with MTCNN having higher precision on the high density sample, suggesting that OpenPose generated slightly more false positives than MTCNN. ViolaJones performed quite poorly relative to the other detectors, especially with respect to the random sample. For wrist detection, OpenPose performed moderately well (F = 0.74) with relatively high precision but low recall on the randomly sampled frames (see Table 1). We thus analyze wrist detections, with the caveat that we are likely underestimating the proportion of hands in the dataset.

Results

First, we report developmental shifts in infants' posture and their orientation relative to their caregiver, consistent with previous literature (Adolph & Berger, 2006; Franchak et al., 2017). Then, we examine how these changes influence children's visual access to faces and wrists/hands across this developmental time range, examining the relative contributions of age vs. postural developments. We also explore how these changes impact the accessibility of faces and wrists/hands during labeling events, both when parents were labeling familiar objects (e.g., "cat") as well as novel objects (e.g., "zem"). Finally, we apply the same automated detection method to a egocentric video dataset collected by a different lab (Franchak et al., 2017) during which children's in-the-moment posture was annotated.

Changes in Posture and Orientation

How do childen's in-the-moment posture and orientation change across age? In our dataset, the proportion of time infants spent sitting decreased with age, and the proportion

of time infants spent standing increased with infants' age. As children got older, their locomotive abilities allowed them to become more independent. Both 8-month-olds and 12-month-olds spent relatively equivalent amounts of time lying/crawling (i.e., "prone") which was markedly decreased in the 16-month-olds, who spent most of their time sitting or standing (see Figure 3). We also observed changes in children's orientation relative to their caregivers: the 8-month-olds spent more time with their caregiver behind them supporting their sitting positions than did children at other ages (see Figure 3). However, we also saw considerable variability across children: some infants spent almost their entire time sitting at a close distance from their caregiver, whereas others showed more considerable variability (see Figure 4).

Changes in Access to Faces and Hands

We first examined the proportion of face and hand detections as a function of children's age without considering their posture (see Figure 5); here, we report face and wrist detections using OpenPose, although the same pattern of results for face detections was found using the outputs from MTCNN. While faces tended to be in the field-of-view overall more often than hands, children's head-mounted cameras were angled slightly upward to capture the presence of faces, and hand detections suffered from somewhat lower recall than face detections. Going forward, we thus analyze differences in the relative proportion of faces or hands in view as a function of age, posture, and orientation, rather than comparing them directly. Overall, we observed that 12-month-olds appeared to have visual access to slightly fewer faces than 8 or 16-month-olds, creating a slight U-shaped function in face detections; conversely, hand detections were showed a slight increase across this age range, as reported in prior literature (Fausey et al., 2016).

However, these age-related trends were much smaller than the effect of infant's postural developments on children's visual access to faces and hands. Children's

in-the-moment posture was a major factor both in how many faces and hands were in view during the play session, as was their orientation relative to their caregiver. Infants who were sitting or standing had more faces in view than infants who were lying down/crawling (i.e. prone), which was most frequent among 12-month-olds relative to the other age groups (Figure 6). When caregivers were behind their children, supporting their children's sitting or standing positions, children saw fewer faces and wrists. In particular, children who were sitting or in front of their caregiver had a high proportion of faces and hands in their field of view (Figure 6).

These trends were quantified using two generalized linear mixed-effect models estimating the proportion of faces and hands that were in view, with orientation, posture, their interaction, and scaled participant's age as fixed effects, and with random slopes for infants' orientation and posture. A summary of the coefficients of the models can be found in Table 2, confirming that infants who were sitting/standing and in front of their caregivers saw the most faces and hands. When modeling the proportion of faces seen, age was not a significant predictor; however, age remained a significant predictor when modeling the proportion of hands seen by infants. Overall, these results suggest that infants' visual access to social information is largely modulated by their posture and orientation to their caregiver, which is in turn a function of their general locomotor development. Nonetheless, these results also suggest that infants may still increase their attention to hands throughout this age range as they continue to learn about objects, their functions, and their names.

Access to Faces and Hands During Labeling Events

Our play session was designed to provide parents with opporunities to label objects–both familiar and novel–such that we could examine whether children sought out different kinds of social information around naming events. We thus explored how face and hand detections changed during object labeling events, analyzing a four-second window

around each labeling event (e.g., "Look at the [zem]!"). Every utterance of one of the labelled objects (e.g., "ball") was counted as a "labeling event"; timestamps of the beginning of each word were hand-annotated from transcripts ad synchronized with the frame-by-frame detections. Overall, we did not find major differences in face/hand detections during naming events relative to baseline, either for novel objects or for familiar objects (see Figure 7). This was true both when we analyzed naming rates with or without taking into account children's posture and orientation relative to their caregiver. These results suggest that either children did not actively seek out social information during this particular play session around naming events, perhaps because were a limited number of possible referents at any given time.

Extension to Franchak et al., 2017

In the present work, we found that infant's in-the-moment posture changed with their age, as did infants' orientation relative their caregiver. In a related study with 12-month-olds, Franchak et al. (2017) also that children's in-the-moment posture changed the amount of time infants spent looking at faces. Here, we sought to replicate the findings from Franchak et al. (2017) using our automated methodology (OpenPose detections), using the footage from their head-mounted cameras (D. A. Simon, Gordon, Steiger, & Gilmore, 2015).

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 room used for the present 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, Franchak et al. (2017) 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 (Yoshida & Smith, 2008) 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 the main results from Franchak et al. (2017), 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 detections), though these were not originally annotated in Franchak et al. (2017) (see Figure 8); 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 Franchak et al. (2017) 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 the caregiver, while in our present dataset the experimenters left the room for the majority of the play session. Nonetheless, we still found that posture modulated the proportion of hands in view, suggesting that this is a major factor that structures infants' access to visual information.

General Discussion

How do postural and locomotive developments influence the social information that infants see? We used a head-mounted camera to explore how these emerging motoric abilities affect children's visual access to social information, here operationalized as the presence of the faces and hands of their caregiver. First, we found systematic changes around the first birthday in children's in-the-moment posture and their orientation relative to the caregivers; older children spent more time standing and less time sitting; older children's caregivers spent less time supporting their standing or sitting postures. Children's changing posture and orientation to their caregiver jointly shaped the amount of social information that was in their view during one-on-one play sessions with their caregivers: Children saw

the most faces/hands when they were sitting or standing and close to their caregiver. Motor development appears to modulate how infants experience their visual world and the social information in it.

However, we did not see differences in visual access to social information during naming events, suggesting that children did not change where they were looking when hearing a label for either a novel or familiar object. We see several potential explanations for this: first, other work (Yoshida & Smith, 2008; Yu & Smith, 2013, 2017), including Franchak et al. (2017), has found that infants spend much more time looking at the toys vs. their caregiver's faces during these play sessions; infants may have thus been primarily interested in exploring these new toys rather than learning their names. Furthermore, moving their own neck upwards towards their caregiver still requires quite a physical effort at this age (especially when prone or sitting) and simply may not have been a priority for children in this context. A second factor is that this particular play session did not present many opportunities where children would need to use social cues to disambiguate referents. Indeed, there were only two possible referents in the room at a time—-and one of them was always a familiar category (i.e., car, kitty).

Broadly,these results show promise for the use of automated methods to detect the social information in the egocentric infant viewpoint during more naturalistic parent-child interactons. The use of this automated methodology allowed us to easily annotate the entirety of the dataset and analyze all of the frames without the need for hundreds of hours of human coding. Importantly, we also applied these same methods to an addiontal dataset that contained more variation in the type and structure of the play session, replicating and extending the effects of children's in-the-moment posture on visual access to social information (Franchak et al., 2017).

Yet more work is needed to understand how these results relate to children's home experiences (Fausey et al., 2016). Both this play session and that of Franchak et al. (2017)

were relatively controlled interactions where caregivers are highly aware that they are being montiored by experimenters. Furthermore, as children grow and change, the activities in which they engage with their caregivers are likely to also vary, leading to differences in the distribution of contexts they experience that may not be captured in these one-on-one play sessions. Finally, the ability to walk is of course only of a cascade of changes in children's abilities and experiences, and this study captures only a cross-sectional slice of this broader, multifaceted trajectory.

Understanding these changes and how they relate to one another has been a persistent challenge for developmental psychology, but the field of computer vision has advanced dramatically in recent years, creating a new generation of algorithmic tools. These tools deal better with noisier, more complicated datasets and extract richer information than previous system. We hope that these new tools can be leveraged to understand the changing infant perspective on the visual world, both in controlled experimental contexts and in children's home environments. By doing so, we believe that we can understand the implications of the changing infant perspective consequences for linguistic, cognitive, and social development.

Acknowledgements

Thanks to Kaia Simmons, Kathy Woo, and Aditi Maliwal for help in recruitment, data collection, and annotation. This work was funded by a Jacobs Foundation Fellowship to MCF, a John Merck Scholars award to MCF, and NSF #1714726 to BLL. An earlier version of this work was presented to the Cognitive Science Society in Frank et al. (2013) and Sanchez et al. (2018).

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.

Adolph, K. E., Vereijken, B., & Denny, M. (1998). Roles of variability and experience in development of crawling. *Child Development*, *69*(1299), 312.

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.

Cummings, M., Van Hof-Van Duin, J., Mayer, D., Hansen, R., & Fulton, A. (1988). Visual fields of young children. *Behavioural and Brain Research*, *29*(1), 7–16.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, *99*(14), 9602–9605.

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

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2017). See and be seen: Infant–caregiver social looking during locomotor free play. *Developmental Science*.

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

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

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

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

Mayer, D., Fulton, A., & Cummings, M. (1988). Visual fields of infants assessed with a new

perimetric technique. *Investigative Ophthalmology & Visual Science*, *29*(3), 452–459.

Patrick, S. K., Noah, J. A., & Yang, J. F. (2012). Developmental constraints of quadrupedal coordination across crawling styles in human infants. *Journal of Neurophysiology*, *107*(11), 3050–3061.

Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments modulate children's visual access to social information.

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

Simon, D. A., Gordon, A. S., Steiger, L., & Gilmore, R. O. (2015). Databrary: Enabling sharing and reuse of research video. In *Proceedings of the 15th acm/ieee-cs joint conference on digital libraries* (pp. 279–280).

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

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

Walle, E. A., & Campos, J. J. (2014). Infant language development is related to the acquisition of walking. *Developmental Psychology*, *50*(2), 336.

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

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

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

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS One*, *8*(11).

Yu, C., & Smith, L. B. (2017). Hand–eye coordination predicts joint attention. *Child Development*, *88*(6), 2060–2078.

Yurovsky, D., Smith, L., & Yu, C. (2012). Statistical word learning at scale: The baby's view is better. *Developmental Science*.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

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

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.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-3.35	0.17	-19.49	0.000
Sit	0.31	0.20	1.58	0.114
Stand	-0.02	0.20	-0.12	0.905
Close	0.03	0.19	0.15	0.882
Far	0.46	0.24	1.93	0.053
Age (Scaled)	0.14	0.11	1.22	0.223
Sit*Close	0.92	0.07	13.31	0.000
Stand*Close	1.30	0.08	16.02	0.000
Sit*Far	0.54	0.07	7.37	0.000
Stand*Far	1.24	0.09	14.18	0.000

Table 2

Model coefficients from a generalized linear mixed models predicting the proportion of faces seen by infants (as detected by OpenPose).

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-4.28	0.23	-18.67	0.000
Sit	0.54	0.19	2.90	0.004
Stand	0.61	0.23	2.63	0.009
Close	0.67	0.22	3.05	0.002
Far	0.33	0.24	1.35	0.176
Age (Scaled)	0.45	0.13	3.35	0.001
Sit*Close	0.34	0.09	3.93	0.000
Stand*Close	0.73	0.09	7.92	0.000
Sit*Far	-0.15	0.11	-1.39	0.165
Stand*Far	1.32	0.11	11.70	0.000

Table 3

Model coefficients from a generalized linear mixed models predicting the proportion of wrists seen by infants (as detected by OpenPose).

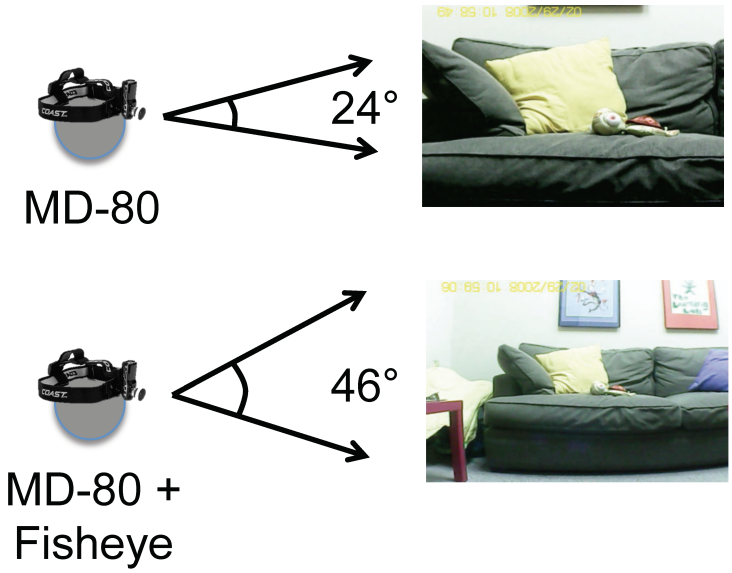

Figure 1. Vertical field of view for two different headcam configurations (we used the lower in our current study).

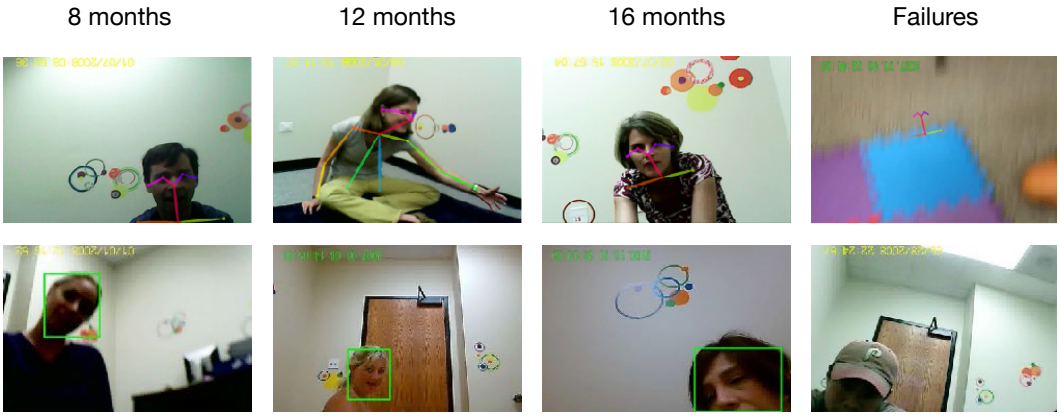

Figure 2. Example face and pose detections made by OpenPose (top row) and MTCNN (bottom row) from a child in each age group. The last column features a false positive from OpenPose and a false negative from MTCNN.

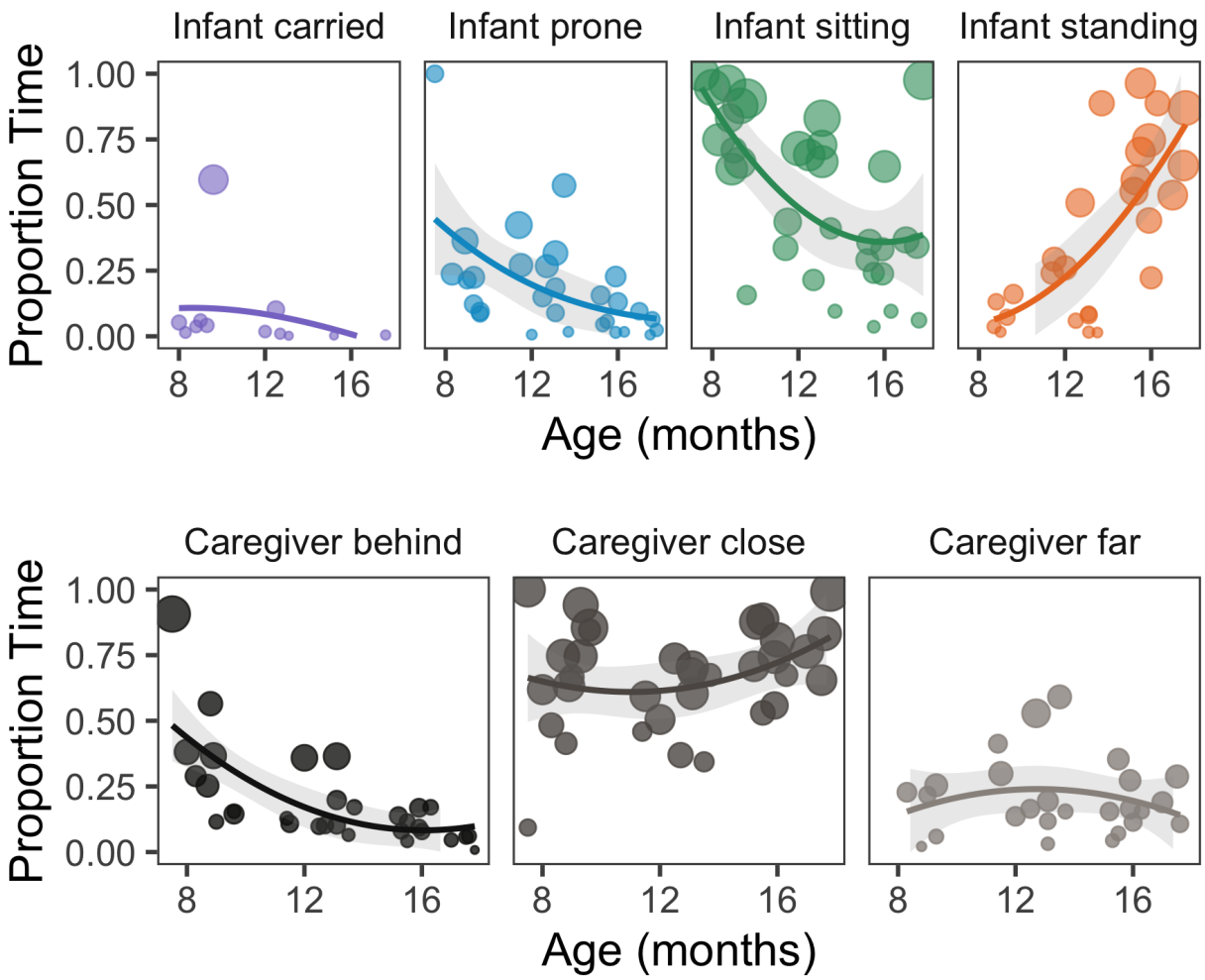

Figure 3. Proportion of time spent by each infant in different postures and orientations relative to their caregivers (CG); times where posture was not codable are ommitted for visualization purposes

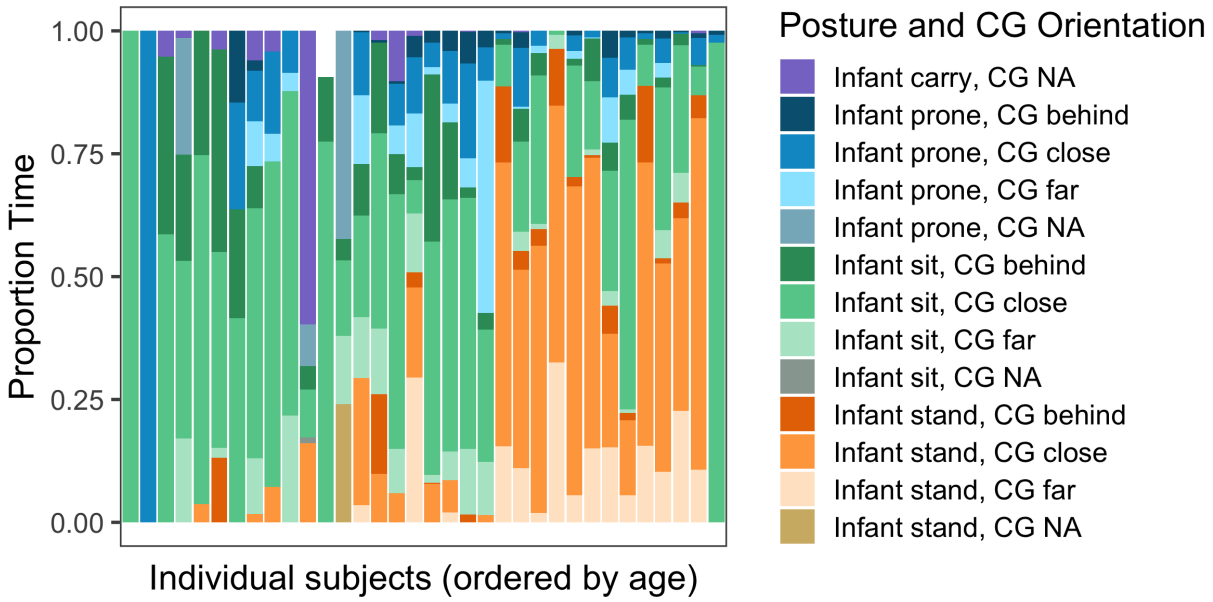

Figure 4. Proportion of time spent by each infant in different postures and orientations relative to their caregivers (CG); times where posture was not codable are ommitted for visualization purposes

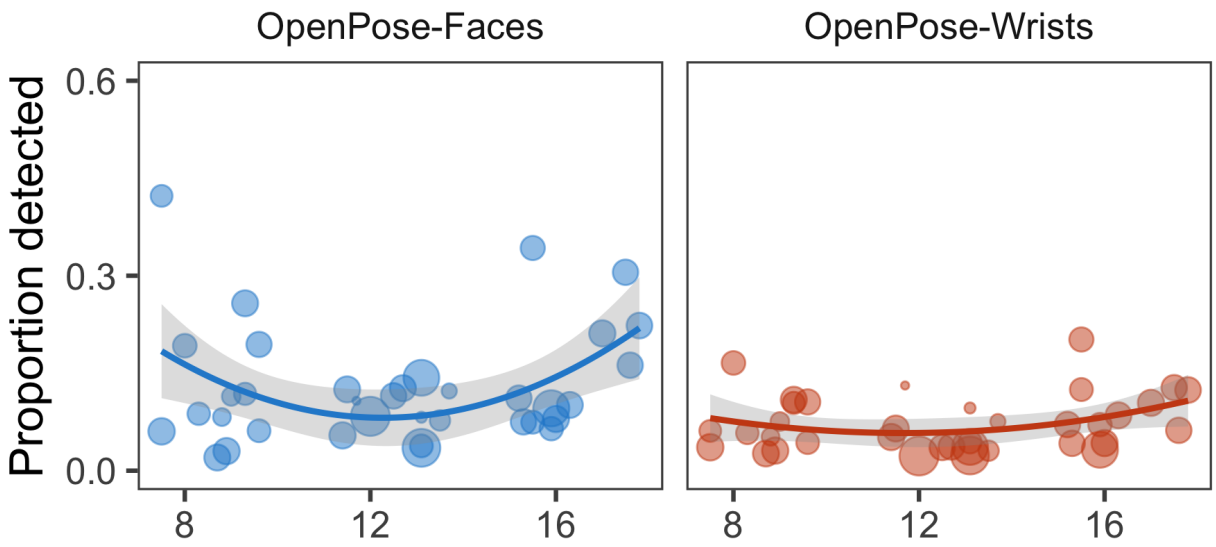

Figure 5. Proportion of faces (left) and wrists (right) detected by the OpenPose model as a function of child's age. Larger dots indicate children who had longer play sessions and thus for whom there was more data.

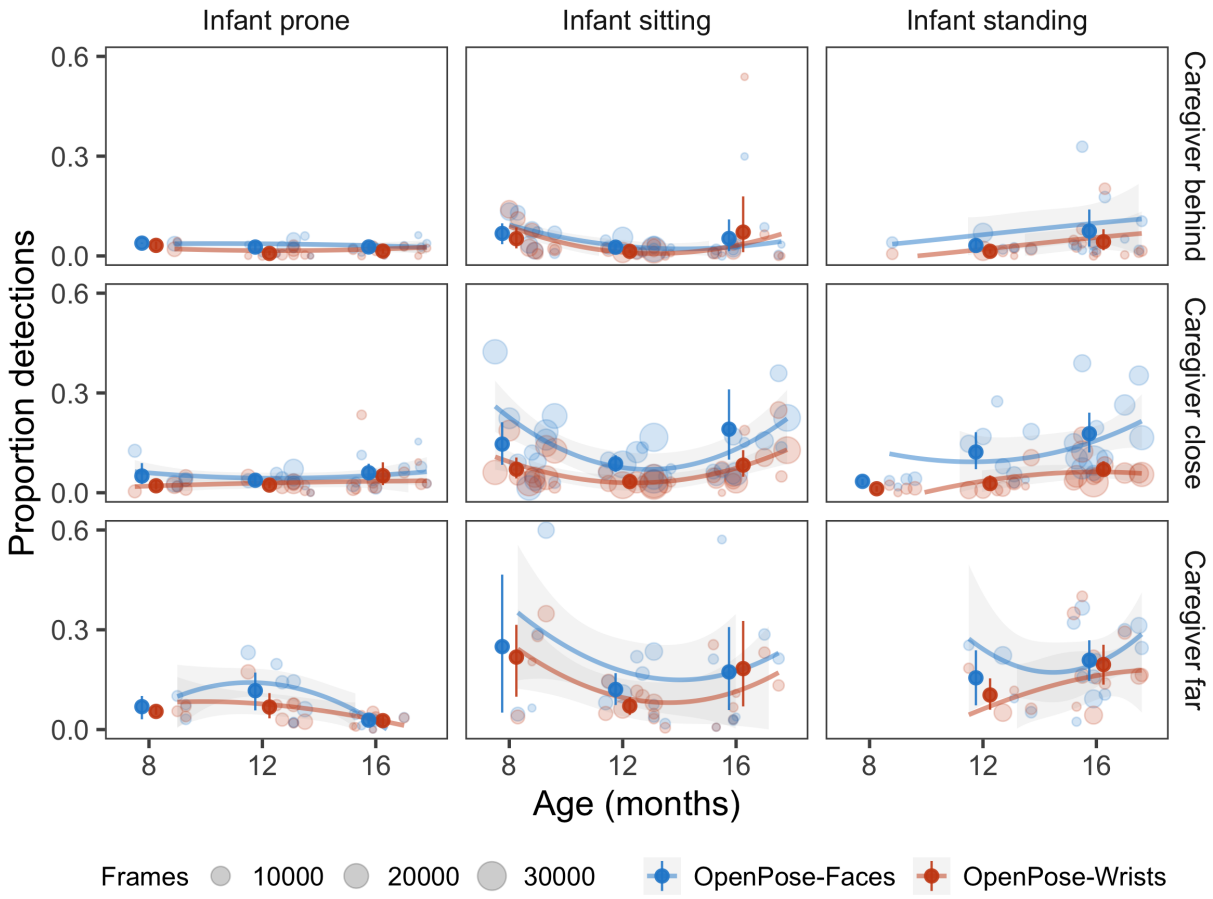

Figure 6. Proportion of face / wrist detections by children's age, their posture, and their caregivers orientation. Data points are scaled by the amount of time spent in each orientation/posture combination; times when posture/orientation annotatinos were unavaliable or the infant was carried are not plotted. Error bars represent 95% bootstrapped confidence intervals

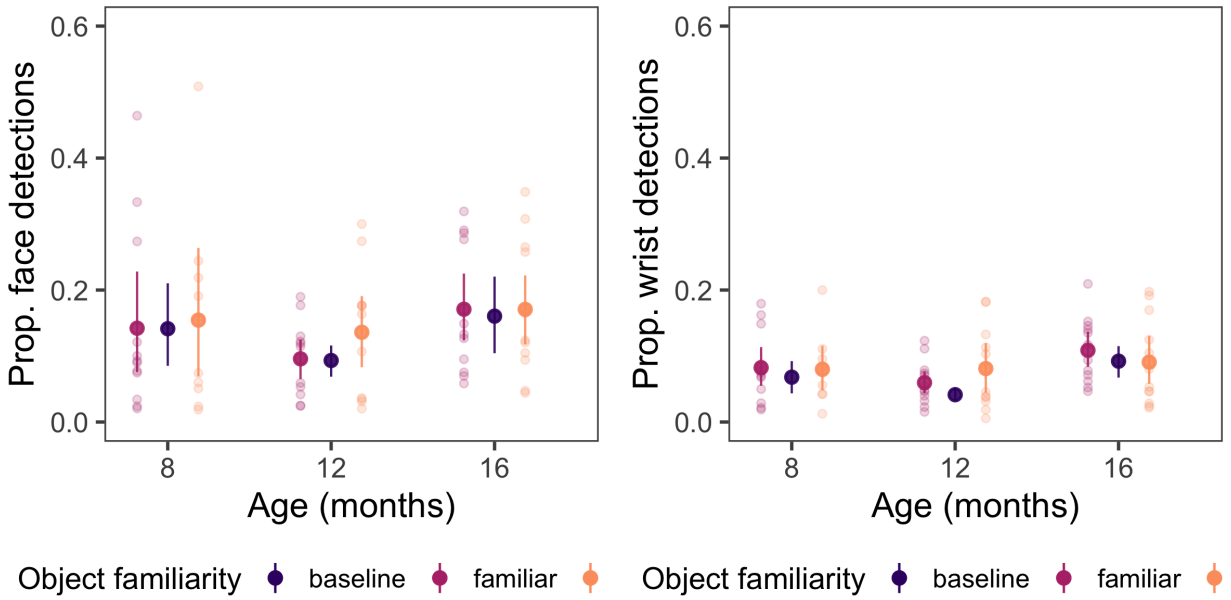

Figure 7. Proportion of face / wrist detections during naming events (+/- 2 seconds around label)for familiar and novel objects; these rates are put into context relative to baseline Error bars represent 95% bootstrapped confidence intervals.

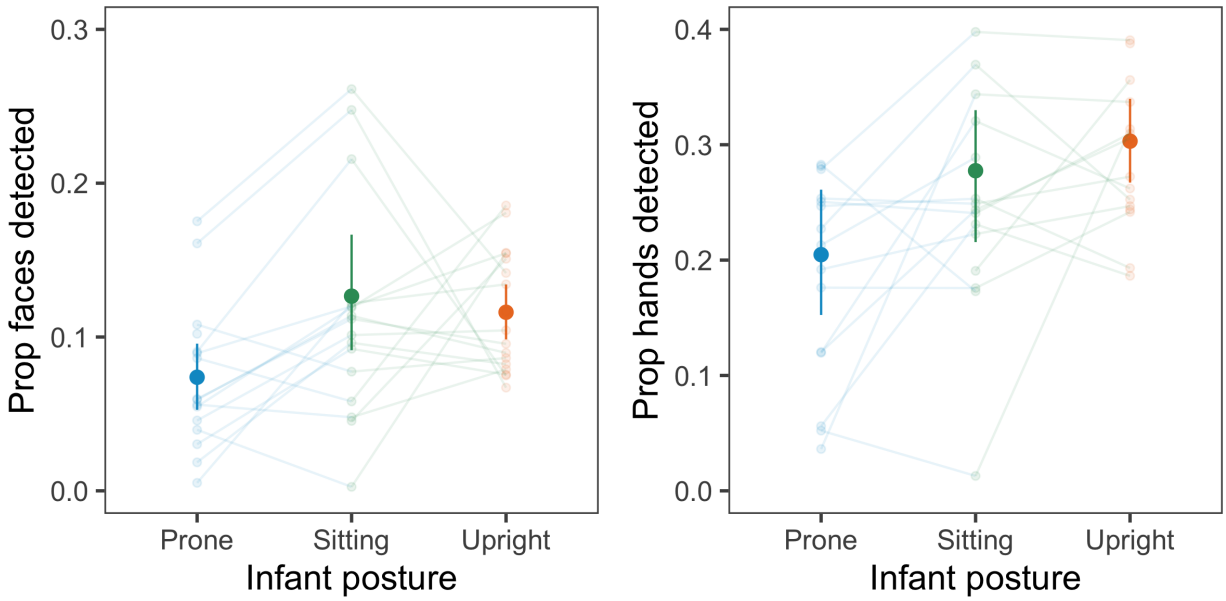

Figure 8. Proportion of face / wrist detections for 12-month-olds in Franchak et al., 2017 as a function of children's in-the-moment posture. Error bars represent 95 percent bootstrapped confidence intervals.