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

Anonymous CogSci submission

Abstract

The faces and hands of infants' caregivers and other social partners offer a rich source of social and causal information that may be critical for infants' cognitive and linguistic development. Previous work using manual annotation strategies and cross-sectional data has found systematic changes in the proportion of faces and hands in the egocentric perspective of young infants. Here, we examine the prevalence of faces and hands in a longitudinal collection of over 1400 headcam videos collected from three children along a span of 6 to 32 months of age—the SAYcam dataset (Sullivan, Mei, Perfors, Wojcik, & Frank, under review). To analyze these naturalistic infant egocentric videos, we first validate the use of a modern convolutional neural network for pose detection (OpenPose) for the detection of faces and hands. We then apply this model to the entire dataset, finding a decrease in the proportion of faces vs. hands across age up to the end of the second year of life (C. M. Fausey, Jayaraman, & Smith, 2016). In addition, we find considerable variability in the proportion of faces/hands viewed by children in different locations (e.g., living room vs. kitchen), demonstrating that variability across activity contexts may be a driving factor in the amount of social information that infants' experience.

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

Introduction

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

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

Beyond these early months,infants motor and cognitive abilities mature, leading to a vastly different perspectives on the world. For example, crawlers see fewer faces and hands than do walking children (Franchak, Kretch, & Adolph, 2017; Kretch, Franchak, & Adolph, 2014; A. Sanchez, Long, Kraus, & Frank, 2018) as well as different view of objects (Smith, Yu, & Pereira, 2011). Further, as infants learn to use their own hands to act on the world, they seem to focus on manual actions taken by their social partners, and their perspective starts to capture views of hands paired with the objects they are acting on (C. M. Fausey et al., 2016). In turn, caregivers may also start to use their hands with more communicative intent, directing infants attention by pointing and gesturing to different events and objects during play (Yu & Smith, 2013).

Here, we examine the social information present in the infant visual perspective–the presence of faces and hands– by analyzing a longitudinal collection of over 1400 headcam videos collected from three children along a span of 6 to 32 months of age—the SAYcam dataset (Sullivan et al., under review). In addition to its size and longitudinal nature, this dataset is more naturalistic than those previously used in two key ways. First, recordings were taken under a large variety of activity contexts (Bruner, 1985; B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015) encompassing infants' viewpoints during both activities outside and inside the home. Even in other naturalistic datasets, the incredible variety in a typical infant's experience has been largely underrepresented (see examples in Figure 1; e.g., riding in the car, gardening, watching chickens during a walk, browsing magazines, nursing, brushing teeth). Second, the head-mounted cameras used in the SAYcam dataset captured a larger field of view than those typically used, allowing a more complete picture of the infant perspective. While head-mounted cameras with a more restricted field of view do represent where infants are foveating most of the time (H. Yoshida & Smith, 2008, Smith, Yu, Yoshida, & Fausey (2015)), they may fail to capture short saccades to either faces or hands in the periphery, as the timescale of head movements is much longer.

With hundreds of hours of footage (>29M frames), how-

ever, this large dataset necessitates a shift to an automated annotation strategy. Indeed, annotation of the frames extracted from egocentric videos has been prohibitively time-consuming, meaning that many frames are typically not inspected, even in the most comprehensive studies. For example, C. M. Fausey et al. (2016) collected a total of 143 hours of head-mounted camera footage (15.5 million frames), of which one frame every five seconds was hand annotated (by four coders), totalling 103,383 frames (per coder)—an impressive number of annotations but nonetheless only 0.67% of the collected footage. To address this challenge, we use a modern computer vision model of pose detection to automatically detect the presence of hands and faces from the infant egocentric viewpoint. Specifically, we use OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018), a model optimized for jointly detecting human face, body, hand, and foot keypoints that operates well on scenes including multiple people, even if they are partially-occluded (see Figure 1).

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

Method

Dataset

The dataset is described in detail in Sullivan et al. (under review); we summarize these details here. Children wore Veho Muvi miniature cameras mounted on a custom camping headlamp harness ("headcams") at least twice weekly, for approximately one hour per recording session. One weekly session was on the same day each week at a roughly constant time of day, while the other(s) were chosen arbitrarily at the participating family's discretion. At the time of the recording, all three children were in single-child households. Videos captured by the headcam were 640x480 pixels, and a fisheye lens was attached to the camera to increase the field of view to approximately 109 degrees horizontal x 70 degrees vertical. Videos[^1] with technical errors or that were not taken from the egocentric perspective were excluded from the dataset. While data collection for the third child (Y) is still ongoing, here we analyze 1622 videos, with a total duration of 347.81225 hours (>29 million frames). [^1]: All videos are available at `https://nyu.databrary.org/volume/564`

Assessing automated social annotations

Computer vision model To automatically annotate the millions of frames in SAYcam, we use OpenPose (Cao et al., 2018; Simon, Joo, Matthews, & Sheikh, 2017a). This pose detector[1] provided the locations of 18 body parts (ears, nose, wrists, etc.). The system uses a convolutional neural network for initial anatomical detection and subsequently applies part affinity fields 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. Thus, while we only make use of the outputs of the face and hand detections, the entire set of pose information from an individual is used to determine the presence of a face/hand, making the process much more robust to occlusion than methods optimized to detect only faces or hands. Specifically, we tagged frames with detected nose keypoints as having a face, and frames with a wrist keypoint were tagged as having a hand.

Manual annotation strategy To test the validity of OpenPose's hand and face detections, we compared the accuracy of these detections relative to human annotations of 24,000 frames selected uniformly at random from videos of two children (S and A). Frames were jointly annotated for the presence of faces and hands. A second set of coders recruited via AMT (Amazon Mechanical Turk) additionally annotated 3150 frames; agreement with the primary coder was >95%.

Detection accuracy for faces and hands As has been observed in other studies on automated annotation of headcam data (e.g. Frank, Simmons, Yurovsky, & Pusiol, 2013, Alessandro Sanchez, Long, Kraus, & Frank (2018), Bambach, Lee, Crandall, & Yu (2015)), detection tasks that are easy in third-person video can be quite challenging in egocentric videos, due to difficult angles and sizes as well as substantial occlusion. For example, the infant perspective often contains non-canonical viewpoints of faces (e.g., looking up at a caregiver's chin) as well as partially-occluded or oblique viewpoints of both faces and hands. Furthermore, hand detection tends to be a somewhat harder problem than face detection (Bambach et al., 2015; Simon, Joo, Matthews, & Sheikh, 2017b), and certainly one that has received less attention than face detection. We thus expected overall performance to be lower in these naturalistic videos than on either photos taken from the adult perspective or in egocentric videos in controlled, laboratory settings (@.g., Alessandro Sanchez et al., 2018).

To evaluate OpenPose's performance, we compared its detections to the manually-annotated gold set of frames, calculating precision (hits / hits + false alarms), recall (hits / hits + misses), and f-score (the harmonic mean of precision and recall). In our data, for faces, the F-score was 0.64, with a precision of 0.7 and recall of 0.58. For hands, the F-score was 0.51, with a precision of 0.73 and recall of 0.4. While face and hand detections showed moderately good precision, face detections were overall slightly more accurate than hand detections. In general, hand detections suffered from fairly low recall, indicating that OpenPose likely underestimated the proportion of hands in the dataset.

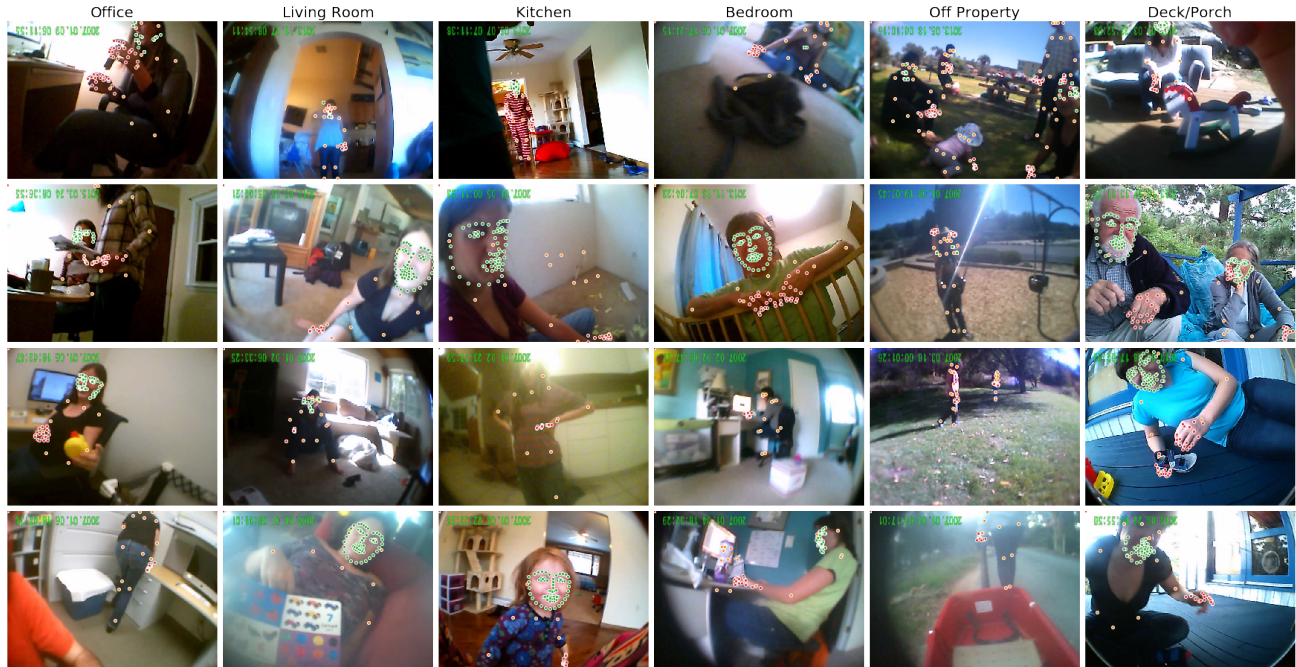[1] `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

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.
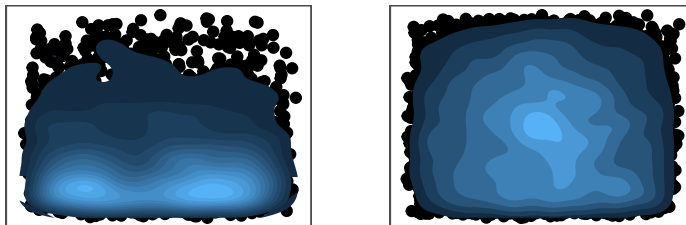

Figure 2: Density estimates for the child (left) and adult (right) hands that were detected in teh 24K random gold set; each dot represents the center of a bounding box made by an adult participant.

We suspected that this is in part because children's own hands were often in view of the camera and unconnected to a pose—a notoriously challenging detection problem (Bambach et al., 2015). To assess this possibility, we obtained human annotations for a subsample of 9051 frames in which a hand was detected; participants (recruited via AMT) were asked to draw bounding boxes around children's hands and adult's hands. Overall, we found that almost half of missed hand detections were of child hands. When frames with children's hands were removed from the gold set, recall did improve to 0.5656277 somewhat. We also observed that children's hands tended to appear in the lower half of the frames; heatmaps of the bounding boxes obtained from these annotations can be seen in Figure 3.

Finally, we examined whether the precision, recall, and F-score for hands and faces varied with age or child, and did not find substantial variation. Thus, while OpenPose was trained on photographs from the adult perspective, this model still generalized relatively well to the egocentric infant viewpoint with no fine-tuning or post-processing of the detections.

Access to social information

Changes across age We analyzed the social information in view across the entire dataset, looking specifically at the proportions of faces and hands that were in view for each child. Data from videos were binned according to the age of the child (in weeks) (see Figure 2). First, when examining the younger age ranges (< 24 months), as in C. M. Fausey et al. (2016), we saw that the proportion of faces vs. hands in view showed a moderate decrease across this age range, both when analyzing the random 24K frames as well as the entire dataset (see Figure 2A).

However, the most striking result is a much greater proportion of hands in view than have previously been reported (C. M. Fausey et al., 2016). We found this to be true across all ages, in all three children, and regardless of whether we analyzed human annotations (on the 24K random subset) or OpenPose annotations on the entire dataset. This is notable especially given that OpenPose showed relatively low recall for hands, indicating that this is still an underestimate of the proportion of hands in view. Nonetheless, one reason this could be the case is the much larger field of view that was captured by the cameras used in this study: unlike previous studies, our cameras were outfitted with a fish-eye lens in an attempt to capture as much of the children's field of view as possible. Thus, the field of view (FOV) of the fisheye lens
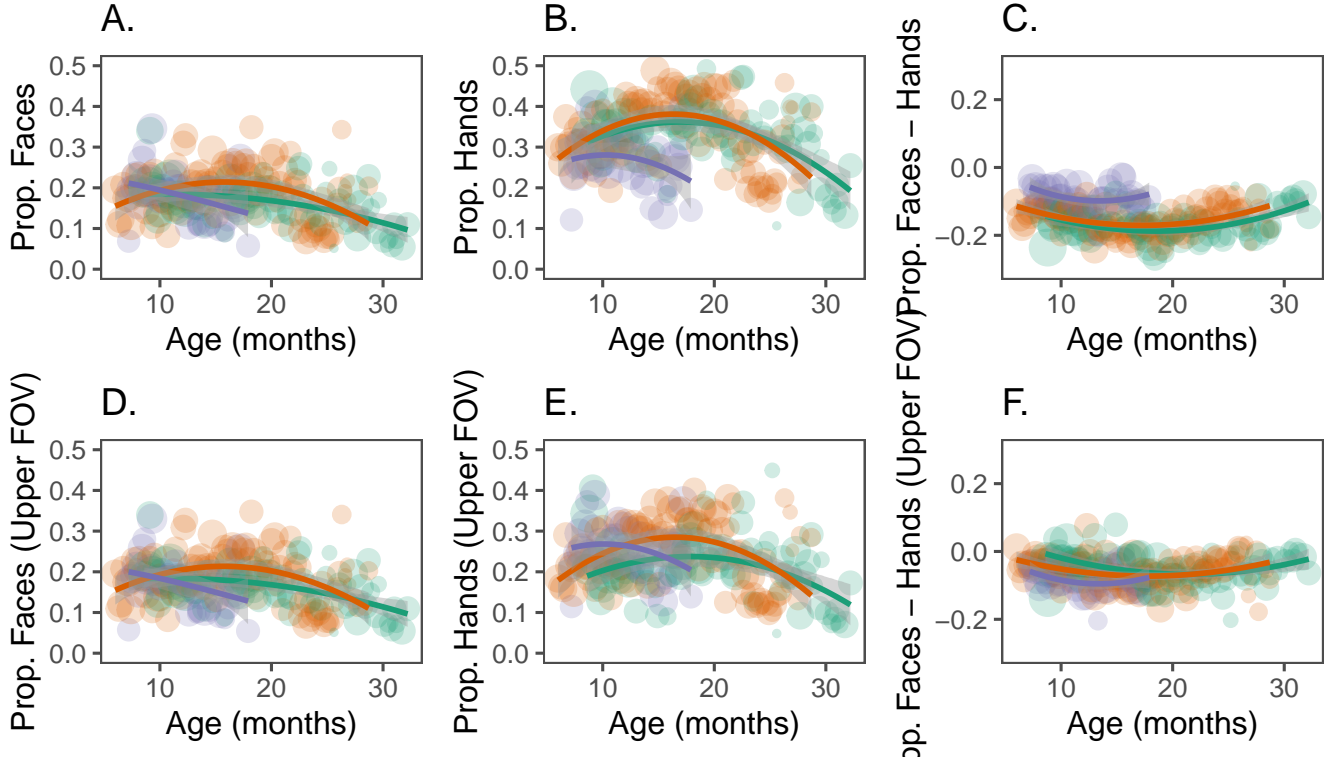
Figure 3: Proportion of faces (A), hands (B), and faces vs. hands (C) in the entire dataset and (D). when hand detections are restricted to the upper 60 percent of the field of view (B). Data from each child are plotted separately and binned by each week that the videos were filmed; data pointsscaled by the number of frames in that age range.

used was larger (109 degrees horizontal x 70 degrees vertical) than the FOV of the lens used in many previous studies; for example, in C. M. Fausey et al. (2016) the FOV was 69 x 41 degrees. This larger field of view may have allowed the SAYcam cameras to capture not only the presence of a social partner's hands interacting with objects, but also the children's own hands, leading to these more frequent hand detections.

We thus re-analyzed the entire dataset while restricting our analysis to a smaller, upper portion of the frame comparable to the field of view used in C. M. Fausey et al. (2016). To do so, we excluded hand detections that occurred in the bottom 40% of the frame, while retaining all hand detections that occurred in the top 60% of the frame. This coarse cropping of decreased the proportion of hand detections from 0.33percent to NA percent, but only decreased face detections from 0.18 to NA. Within this modified field of view, we still observed, on average, more hand detections than face detections.

However, we also still observed a moderate decrease in the proportion of faces vs. hands in view across age. To quantify this trend, we fit a generalized linear mixed model to the binned detection rates from this smaller field of field to estimate changes in the proportion of faces vs. hands over age. To better approximate C. M. Fausey et al. (2016), we restricted our analysis to the age range seen in C. M. Fausey et al. (2016), excluding videos when children were over 24

months of age, and indeed found that the proportion of faces vs. hands declined modestly with age (see Table 1). Note that C. M. Fausey et al. (2016) also included data from very young infants (from 1-4 months of age), while here the youngest videos coming from S and A around 6 and 9 months of age, respectively.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.55	0.03	-52.47	0.000
scale(age_days)	-0.13	0.00	-261.57	0.000

Table 1: Model coefficients from a generalized linear mixed model predicting the proportion of faces seen by infants.

However, when we considered the entire age range, we also observed an unexpected trend in which the relative proportion of faces vs. hands increased again after 24 months of life; this trend was stronger in the full dataset. While we did not have any a priori hypotheses about this age range, we suspect that this could have to do with an increase in children's independence from their caregiver as they master walking around and interacting with objects in the world on their own.

Variability by Location How does variability across different contexts influence the social information in the infant view? Intuitively, some activities in different contexts may be characterized by a much higher proportion of faces (e.g., di-

aper changes in bedrooms) than others (e.g., playtime in the living room). We thus next examined variation in presence of hands and faces across different locations. Of the 1622 videos, 639 were annotated (Sullivan et al., under review) for the location o videos were filmed in. 296 were filmed in single location, representing 17 percent of the dataset and over 5 million frames (see Sullivan et al. (under review)). Activities varied somewhat predictability by these contexts: for example, eating tended to occur in the kitchen, whereas playtime was the dominant activity in the living room. Overall, we found that the proportion of faces vs. hands varied across filming locations, and, to some extent, across children. For example, while both A and S saw a relatively similar proportion of faces vs. hands in the bedroom, they saw quite different amounts of faces vs. hands in kitchen.
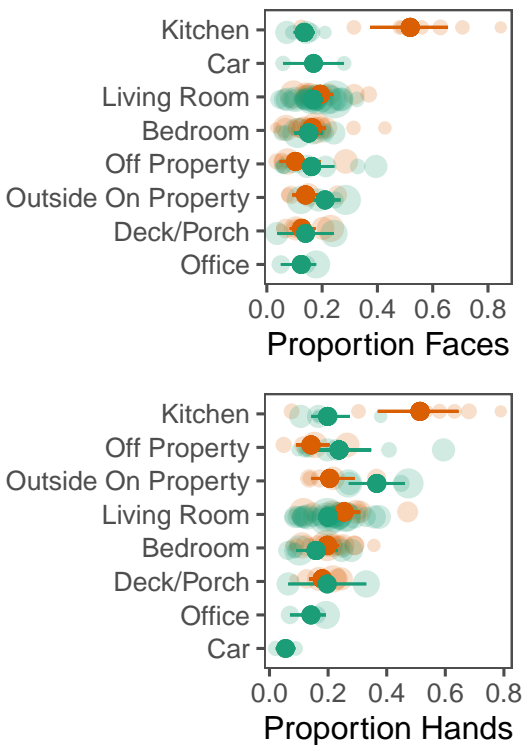

Figure 4: Proportion of faces (A), hands (B), and faces vs. hands (C) by location in which egocentric videos were filmed; data from individual children are plotted separately (location annotations were not available for Y). Data are only taken from the top 60 percent of the frames so as to minimize the contribution of childrens own hands. Each dot represents data from a week in which videos were filmed and are scaled by the number of frames (i.e., amount of video) in that datapoint.

General Discussion

Broadly, the present analysis of this dense, longitudinal dataset has yielded a better understanding of infants' evolving access to social information. First, we found a relatively high proportion of faces in view in the videos from the youngest age ranges in each child, confirming previous findings (C.

M. Fausey et al., 2016). This is particularly notable given that, in cross-sectional data, this effect seems to be most strongly driven by infants younger than 4 months of age (e.g., Fausey 2016; Jayaraman, Fausey, & Smith, 2015; Sugden, Mohamed-Ali, & Moulson, 2014) who see both more frequent and more persistent faces (Jayaraman & Smith, 2018).

Second, we found considerable variability in the social information in view depending on the activity context in which children were present. Even within a given, well-defined context—mealtime—we found variability across individual children. Intuitively, there are at least three ways to feed a young child in a kitchen (i.e., forward facing the child in a high chair, holding the child outward, or sitting side by side with them) and these regularities in a child's life may hold a larger influence over the amount of social information in view than previously considered.

Finally, we demonstrate the feasibility of using a modern, off-the-shelf computer vision model to vastly increase the efficiency of processing egocentric headcam footage, allowing us to annotate the entirety of very large datasets (here, >30 million frames) for the presence and size of people, hands, and faces, representing a 300-fold increase in data relative to prior work (C. M. Fausey et al., 2016). As these detections were imperfect compared to human annotators, fine-tuning these models to better optimize for the infant viewpoint remains an open avenue for future work. Standard computer vision models are rarely exposed on these naturalistic, egocentric viewpoints, and we suspect that training these models on more naturalistic data may lead to more robust, generalizable detectors.

These analyses have highlighted the importance of considering the infants field-of-view when analyzing egocentric videos. In many frames children's own hands were likely visible, due to the wide field of view—capturing valuable information about what these children were interacting with in different contexts across the first few years of their life. Furthermore, though this choice more closely approximates children's actual perspective, these viewpoints are still a proxy as we do not know where in the frames children are allocating their attention.

Indeed, the questions of how, when, from whom, and what data we sample become critically important when we attempt to construct naturalistic datasets, and we would be remiss if we did not acknowledge the limitations of the present dataset. While these data are longitudinal, they are only sampled from representative of few children, all growing up in relatively privileged households in western societies. Any idiosyncrasies in how and when these particular families chose to film these videos undoubtedly influences the variability seen here. Nonetheless, we believe that these advances in methodologies and datasets are a step in the right direction. The large-scale analysis of dense datasets—collected with different fields of view, cameras and from many different laboratories—has the potential to create generalizable conclusions about the regularities of infant experience that scaf-

fold learning.

Acknowledgements

References

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the ieee international conference on computer vision* (pp. 1949–1957).

Bruner, J. (1985). The role of interaction formats in language acquisition. In *Language and social situations* (pp. 31–46). Springer.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. In *ArXiv preprint arXiv:1812.08008.*

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), 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 childrens visual access to faces. In *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 454–459).

Gredebäck, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of infants' gaze as they view adult shifts in overt attention. *Infancy*, *13*(5), 533–543.

James, W., Burkhardt, F., Bowers, F., & Skrupskelis, I. K. (1890). *The principles of psychology* (Vol. 1). Macmillan London.

Jayaraman, S., & Smith, L. B. (2018). Faces in early visual environments are persistent not just frequent. *Vision Research*.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PLoS One*. `http://doi.org/10.1371/journal.pone.0123780`

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, *85*(4), 1503–1518.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments modulate children's visual access to social information. In *Proceedings of the 40th annual conference of the cognitive science society*.

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

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

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

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother?s view: The dynamics of toddler visual experience. *Developmental Science*, *14*(1), 9–17.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407–419.

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

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

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).