A longitudinal investigation of the social information in natural infant 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. The present research aims to test the generality of these findings using the SAYcam dataset (Sullivan, Mei, Perfors, Wojcik, & Frank, n.d.), a longitudinal collection of over 1700 headcam videos collected from three children along a span of 6 to 32 months of age. To do so, we validate the use of a modern convolutional neural network for pose detection (OpenPose) for the detection of people, faces, and hands to analyze these naturalistic infant egocentric videos. We then apply this model to the entire dataset, analyzing the prevalence of faces across age, individuals, and activity contexts. Overall, we find a higher prevalence of hands seen by infants than previously reported, considerably variability in the proportion of faces/hands seen across different locations (e.g., living room vs. kitchen), yet surprising consistency across both individual children.

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

Introduction

Infants are confronted by a blooming, buzzing onslaught of stimuli (James, 1891) which they must learn to parse to make sense of the world around them. Yet infants do not embark on this learning process alone: infants are engaged in learning from their caregivers from early infancy. From as early as 3 months of age, young infants follow overt gaze shifts (Gredeback, Fikke, & Melinder, 2010), and even newborns prefer to look at faces with direct vs. averted gaze (Farroni, Csibra, Simion, & Johnson, 2002), despite their limited acuity [CITE].

Faces are thus likely to be an important conduit of social information that scaffolds infants cognitive development. Given this importance, developmentalists have long hypothesized [CITE] that faces are prevalent in the visual experience of young infants. However, as even the viewpoint of a walking child is not easily predicted by our own adult intuitions (Clerkin, Hart, Rehg, Yu, & Smith, 2017; Franchak, Kretch, Soska, & Adolph (2011); Yoshida & Smith (2008)), researchers have begun to record the egocentric views collected from infants and toddlers wearing head-mounted cameras to test theories about the infant perspective.

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, Fausey, & Smith, 2015; Jayaraman & Smith, 2018; Sugden, Mohamed-Ali, & Moulson, n.d.). However, as infants mature, their perspective starts to capture views of hands paired with the objects they are acting on (Fausey, Jayaraman, & Smith, 2016). As infants learn to use their own hands to act on the world, they may focus on manual actions taken by their social partners. Furthermore, caregivers may start to use their hands more with communicative intent, directing infants attention with pointing and gestures to particular events or objects during play (Yu & Smith, 2013).

The present research aims to test the generality of these findings using the SAYcam dataset (Sullivan et al., n.d.), a longitudinal collection of over 1700 headcam videos collected from three children along a span of 6 to 32 months of age. In addition to its size and longitudinal nature, this dataset builds on those used in previous research in two key ways. First, recordings were from a variety of many different naturalistic contexts, encompassing infants' viewpoints during both activities outside and inside the home. Second, the cameras used in this longitudinal study encompassed a much wider field of view than those typically used, allowing a more complete picture of the infant perspective.

However, with hundreds of hours of footage (>30M frames), this large dataset truly necessitates a shift to an automated annotation strategy. Indeed, annotation of the frames extracted from egocentric videos has been to be prohibitively time-consuming, meaning that many of the frames are not inspected. For example, 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 first validate the use of a modern computer vision model (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018) to automatically detect the presence of hands and faces from the infant egocentric viewpoint. In particular, we focus on OpenPose (Cao et al., 2018), a model optimized for jointly detecting human face, body, hand, and foot keypoints (135 in total) that operates well on scenes including multiple people even if they are partially-occluded (see Figure 1).

We then apply these methods at scale to the larger dataset, allowing us to analyze the proportion of faces and hands observed by each child across age and activity context.

In the following paper, we first describe the dataset and the pose model used in the following analyses and validate the use of this model for extracting our key descriptive variables by comparing to a human-annotated gold set of 24,000 frames. Next, we analyze key descriptive variables, including those that have been previously reported to vary across age, including the relative proportions of faces vs. hands and the sizes of the faces, over the entire dataset (30M frames), finding a greater prevalence of hands than has been previously reported.

We then investigate sources of variability in our more naturalistic dataset that may explain these differences, including a diversity of activity contexts as well as a larger field of view captured by our cameras.

Method

Dataset

Videos captured by the headcam were 640x480 pixels, and a fisheye lens was attached to the camera to increase 109 degrees horizontal x 70 degrees vertical. Children wore 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[1] with technical errors or that were not taken from the egocentric perspective were excluded from the dataset. We analyze 1,636 videos, with a total duration of XX hours (XX million frames).

Part 1: How well can we capture social information using computer vision?

Computer vision model To automatically annotate the millions of frames in SAYcam, we use OpenPose (Cao et al., 2018; Simon, Joo, Matthews, & Sheikh, 2017), a computer vision model optimized for jointly detecting human face, body, hand, and foot keypoints (135 in total) that operates well on scenes including multiple people even if they are partially-occluded. This convolutional neural network (CNN)-based pose detector[2] provided the locations of 18 body parts (ears, nose, wrists, etc.). 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 outputs of the face and hand detection modules.

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 the 1,636 videos of two children (S and A). Frames were jointly annotated for the presence of faces and hands. These randomly sampled frames covered XX of the videos present in the dataset.

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

Detection accuracy for faces and hands Overall, we found that face detections were slightly more accurate than hand detections,

Precision and recall (F-score) variation across child/age for faces

Describe possible sources of variation that decrease scores for: -Faces: weird viewpoints, occluded/side viewpoint, faces in books -Hands: children's own hands, hands in books, side viewpoints

Describe additional child vs. hand annotation; P/R/F variation across child vs. adult hands (better for adult hands, still OK for child hands)

Part 2: Access to social information across age

Next, 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 XX). First, we saw that the proportion of faces in view showed a moderate decrease across this age range, both when analyzing the random 24K frames as well as the entire dataset. While these trends appear somewhat different than those observed in Fausey et al. (2016), note that Fausey et al. (2016) included data from very young infants (starting at 2 months while), while here the youngest videos coming from S and A around 6 and 9 months of age, respectively. Similarly, our age range extends 8 months later than those infants in Fausey et al. (2016), throughout a portion of third year of life.

However, the most striking result from the dataset is a much greater proportion of hands in view than have previously been reported. We found this to be true across all ages, in both children, and regardless of whether we analyzed human annotations (on the 24K random subset) or the entire dataset.

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 fisheye 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 used in Sullivan et al. (n.d.) was much wider (109 degrees horizontal x 70 degrees vertical) than the FOV of the lens used in Fausey et al. (2016) (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 more frequent

[1] All videos are available at https://nyu.databrary.org/volume/564

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

hand detections.

To assess this possibility, we obtained annotations for a subsample of XX frames in which a hand was detected in the random gold set; participants (recruited via AMT) were asked to draw bounding boxes around children's hands and adult hands. Overall, we found that 34% of the hands detected by OpenPose in the random 24K sample were of children's own hands (compared to 8% reported in Fausey et al. (2016)) suggesting that this difference in field of view did contribute substantially to the higher rate of detections in the frames. Heatmaps of the bounding boxes obtained from these annotations can be seen in Figure XX, showing that children's hands tended to appear in the lower half of the frames.

We thus re-analyzed the entire dataset while restricting our analysis to a smaller, middle portion of the frame comparable to the field of view used in Fausey et al. (2016) (XX vs XX degrees). To do so, we excluded hand detections that occurred in the bottom 40% of the frame, while retaining all detections that occurred in the top of the frame.
Overall, we found X, suggesting Y.

Intriguingly, we observed an unexpected trend in which the relative proportion of faces vs. hands increased during the third year of life; in particular, this seemed…(save for discussion?)

Variability by Location Next we examine variation in the presence of hands and faces across different locations. Of the XX videos, the content of 1,829 have been manually annotated for filming location, activites taking place, and visible objects (see Sullivan et al. (n.d.)). XX of the videos were filmed in a single location, representing XX% of the dataset and roughly XX million frames (see Sullivan et al. (n.d.)). To give a sense of the contexts the children experienced, the most frequent filming locations were the living room (339 videos), bedroom (182), kitchen (150), outside on property (129), child's bedroom (81), deck/porch (73), hallway (70), and off property (57). Filming only took place twice in the dining room (see Figure XX). Activities varied predictability by these locations: for example, eating and XX occured in the kitchen, whereas playtime was the dominant activity in the living room (see Figure XX). Bria: analyze activity

Not this, but point out that eating was not frequent: The most frequent activities were sitting (410), playing (375), being held (352), and standing (297). Eating was the 11th most-frequent activity (117 videos).

(goldset, full dataset)

Discussion

We demonstrate the feasibility of using modern computer vision models to vastly increase the efficiency of processing egocentric headcam footage, allowing us to annotate 100% of even very large datasets for the presence and size of people, hands, and faces. We validate the use of this model by comparing to a human-annotated gold set, and find X.

Analyzing this dataset has yielded a better understanding of infants' evolving access and attention to social information.

Need to think about the child's viewpoint relative to actual FOV as well as attention

Fausey 2016: 103,383 images; Here: 30,000,000 frames; 300 fold increase in data

Acknowledgements

We would like to thank X and Y for helpful comments, and…

References

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*.

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

Gredeback, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, *13*(6), 839–848.

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`

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

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

Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (n.d.). 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. (n.d.). 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*(3), 229–248.