Quantifying 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.

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

Introduction

Infants are famously confronted by a blooming, buzzing onslaught of stimuli (James, 1891) which they must learn to parse and navigate as their cognitive and social skills develop. Fortunately, they can depend on regularities not only in their visual environments (Aslin, 2009) and linguistic environments, but also in the presence and actions of their caregivers and other social partners [@]. From a very young age, infants show great attention to faces CITE, and indeed faces are a critical part of infants' visual experience as an important conduit of social and linguistic information. As infants mature, they begin to also attend more to the manual actions taken by their social partners, engaging in joint attention to objects and events around them. Such episodes are prompted not only by a glance from a caregiver, but also through pointing and offering; thus, hands are also an important carrier of information, especially relevant for learning language and actions. It is unsurprising, then, that both hands and faces are prevalent in even young infants' visual experiences, as evidenced through analyses of egocentric views collected with head-mounted cameras (i.e., headcams; CITES).

Of course, not only is children's interest and attention to hands and faces

Previous work has suggested:

Previous work has found changes in the prevalence of faces vs. hands for infants of different ages. For example, Fausey et al. (2016) found that infants less than 12 months of age received face-dense input, relative to 1- to 2-year-olds who received more hand-dense input. However, it may be that this effect is driven by infants younger than 4 months of age (e.g., Jayaraman, Fausey, & Smith, 2015; Sugden, Mohamed-Ali, & Moulson, n.d.) who see both more frequent and more persistent faces (Jayaraman & Smith, 2018). Once infants begin to crawl (˜6 months) they may see far fewer hands and faces, overall.

Earlier work has also found surprising attention to hand movements and their interactions with objects (Yu & Smith, 2013), particularly in older infants (ages?).

Differences in the availability of social information depending on motor abilities (Franchak, Kretch, Soska, & Adolph, 2011; Sanchez, Long, Kraus, & Frank, 2018) (more Franchak papers?)

One limitation of past work is that it has relied on cross-sectional data, and thus cannot speak to whether these trajectories are present in individual children.

Here, we analyze the SAYcam dataset (Sullivan, Mei, Perfors, Wojcik, & Frank, n.d.), a longitudinal corpus of head-mounted camera data comprising more than 1700 videos from three children, for a total of over 300 hours of videos (>100 million frames). Over a span of 6 to 32 months of age, the three children (S, A, and Y) in the dataset 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.

This dataset differs in four key ways: 1) number of hours/frames 2) not just mealtimes–naturalistic sample of many contexts, 3) longitudinal, 4) much larger field-of-view.

To do so, we first test and validate novel computer vision methods for extracting social information from these egocentric viewpoints on a small subset of randomly selected frames from the dataset.

We then apply these methods at scale to the larger dataset, allowing us to extract key descriptive variables hypothesized to vary across development.

Method

Dataset

(briefly describe dataset; sampling strategy: location of two households, number of hours of video, variability in location, etc; reference published paper on what this dataset is, large field of view (fisheye lens)) (109 degrees horizontal x 70 degrees vertical)

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

Description of OpenPose (Figure 1) To automatically annotate the millions of frames in SAYcam, we use OpenPose

(Cao, Hidalgo, Simon, Wei, & Sheikh, 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.

Description of annotation strategy (24K by Ketan, 4K on Amazon Mechanical Turk, reliability) To test the validity of OpenPose's hand and face detections, we compared to human annotations of 24,000 frames selected uniformly at random from the videos of two children (S and A).

Describe main PRF statistics for 24K for faces and hands; interpret. Relatively higher precision vs. recall. - P/R/F 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

Prevalence of hands vs faces across age (in goldset, full dataset) (Figure 2)

Why so many hands?

More child hands as in gold set

Field-of-View Comparison

The field of view (FOV) of the fisheye lens used in Sullivan et al. (n.d.) is much wider (109 degrees horizontal x 70 degrees vertical) than the FOV of the lens used in Fausey et al. (2016) (69 deg. x 41 deg.). Looks like child hands make up about ˜34% of the hands detected in our gold set (in Fausey 2016, they are only 8% of the hands). Furthermore, a lot of the lower proportions of hands come from the infants <6 months of age.

Variability by Location

Next we examine variation in the presence of hands and faces across different locations. Of the 3,027 videos, the content of 1,829 have been manually manuallly annotated for filming location, activities taking place, and visible objects (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.

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)

List multiple references alphabetically and separate them by semicolons (Frank, 2012; Smith, Yu, & Pereira, 2011).

You might want to display a wide figure across both columns. To do this, you change the `fig.env` chunk option to `figure*`. To align the image in the center of the page, set `fig.align` option to `center`. To format the width of your caption text, you set the `num.cols.cap` option to `2`.

One-column images

Single column is the default option, but if you want set it explicitly, set `fig.env` to `figure`. Notice that the `num.cols` option for the caption width is set to `1`.

Figure 1: One column image.

R Plots

You can use R chunks directly to plot graphs. And you can use latex floats in the fig.pos chunk option to have more control over the location of your plot on the page. For more information on latex placement specifiers see **here**

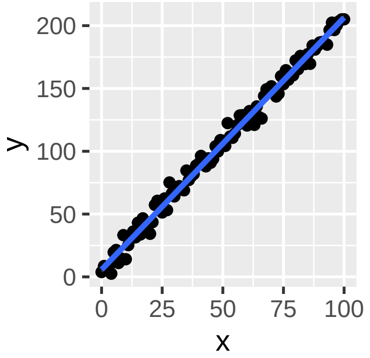

Figure 2: R plot

Tables

You can use the xtable function in the xtable package.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.09	0.10	-0.9	0.35
x	2.06	0.11	18.5	0.00

Table 1: This table prints across one column.

Discussion

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

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, *86*(6), 561–565.

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*.

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.

Frank, M. C. (2012). Measuring children's visual access to social information using face detection. In *Proceedings of the nth annual conference of the cognitive science society* (pp. XXX–XXX). Hillsdale, NJ: Cognitive Science Society.

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*.

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*.

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

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

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.