Postural changes mediate children's visual access to social information

Alessandro Sanchez
sanchez7@stanford.edu
Department of Psychology
Stanford University

Bria Long
bria@stanford.edu
Department of Psychology
Stanford University

Ally Kraus
XX@XX
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

The ability to process social cues–including eye gaze–is a critical component of children's early language and cognitive development. However, as children reach their first birthday, they begin to locomote themselves, walking and exploring their visual environment in an entirely new way. How do these postural and locomotive changes affect children's access to the social information relevant for word-learning? Here, we explore this question by using head-mounted cameras to record infants' (8-16 months of age) egocentric visual perspective and use state-of-the-art computer vision algorithms to detect the proportion of faces in infants' environments. We find that infants' posture and orientation to their caregiver largely mediate infants' access to faces, suggesting that these postural and locomotive developments facilitate infants' emerging linguistic and social capacities. Broadly, we suggest that the combined use of head-mounted cameras and the application of novel deep learning algorithms is a promising avenue for understanding the statistics of infants' visual and linguistic experience.

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

Introduction

Children are remarkably skilled language learners, connecting arbitrary labels ("cup") with specific visual concepts at a rapid pace through the first two years of life. However, children do not learn words in a vacuum but in a rich social environment, where social cues provided by speakers (e.g., eye-gaze) provide strong scaffolding for this learning process. Indeed, children's ability to effectively process these social cues may be a key factor in their early language development. For example, in a longitudinal study, children's level of joint engagement with their mother was found to predict both their receptive and productive vocabularies (Carpenter, Nagell, & Tomasello, 1998). More recently, 10 month-olds who follow an adult's gaze in an experimental context have larger vocabularies at 18 months (R. Brooks & Meltzoff, 2005) and throughout the second year of life (Rechele Brooks & Meltzoff, 2008).

However, as children are learning their first words, their view of the world is also radically changing (K. Adolph & Berger, 2007). Infants' motor abilities improve dramatically near the end of the first year of life, allowing them to locomote independently. These motor changes have drastic consequences for what children see; crawling and walking infants have radically different views of the world. For example, during spontaneous play in a laboratory playroom, toddlers are more likely to look at the floor while crawling than while walking (J. Franchak, Kretch, Soska, & Adolph, 2011); in general, walking infants tend to have full visual access to their environment and the people in it, while crawling infants do not (K. S. Kretch, Franchak, & Adolph, 2014).

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 also impact how children interact with their mothers; walking (vs. crawling) infants make different kinds of object-related bids for attention from their mothers and tend to hear more action directed statements (e.g., "open it") (Karasik, Tamis-LeMonda, & Adolph, 2014). Further, in an observational study, Walle & Campos (2014) 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 social information (e.g., faces, gaze) relative to children who are still crawling and sitting. In other words, the ability to walk and move around independently and subsequently access more detailed social information may allow infants to learn words quicker and more efficiently, facilitating language growth.

Recent technological developments allow for testing of this hypothesis by documenting the experiences of infants and children from their own perspective. By using head-mounted cameras, researchers have begun to document visual experiences of infants and children — which even for walking children are radically different from the adult perspective (and not easily predicted by our own adult intuitions) (Clerkin, Hart, Rehg, Yu, & Smith, 2017; J. Franchak et al., 2011; Yoshida & Smith, 2008). Children's views tend to be restricted and to be dominated by objects and hands much more than that of adults (Yoshida & Smith, 2008), and both computational and empirical work suggest that this restricted viewpoint may be more effective for learning objects and their labels than the comparable adult perspective (Bambach, Crandall, Smith, & Yu, 2017; D. Yurovsky, Smith, & Yu, in press). Further, recent work also suggests dramatic changes in the child's perspective over the first two years of life, as views transition from primarily containing a close up view of faces to capturing views of hands paired with the objects they are acting on (Fausey, Jayaraman, & Smith, 2016).

Here, we directly ask whether postural and locomotive developments change the availability of the social information relevant for word learning. To do so, we recorded the visual experience of a group of infants and children using head-mounted cameras during a brief laboratory free-play session; children's posture and orientation relative to their caregiver were also recorded from a third-person perspective and hand-annotated. We capitalize on recent improvements in face detection algorithms (Cao, Simon, Wei, & Sheikh, 2017; K. Zhang, Zhang, Li, & Qiao, 2016) to analyze the frequencies

of faces in the child's visual environment, both overall and relative to naming events by their caregivers.

Methods

Participants

Our final sample consisted of 36 infants and children, with 8 participants in three age groups: 8 months (6 females), 12 months (7 females), and 16 months (6 females). Participants were recruited from the surrounding community via state birth records, had no documented disabilities, and were reported to hear at least 80% English at home. Demographics and exclusion rates are given in Table 1.

Group	N	% incl.	Mean age	Videos length (min)
8 mo.	12	0.46	8.71	14.41
12 mo.	12	0.40	12.62	13.48
16 mo.	12	0.31	16.29	15.00

Table 1: Demographics by age group.

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 CDI data, 2 for missing scene camera footage, 1 for fussiness, and one excluded for sample symmetry. All inclusion decisions were made independent of the results of subsequent analyses.

Head-mounted camera

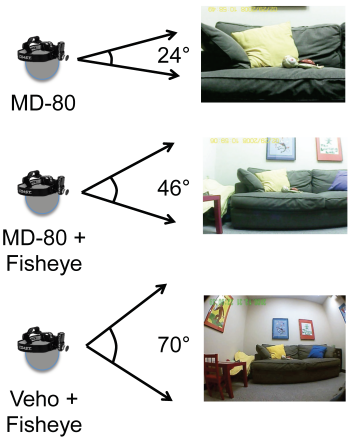

Figure 1: Field of view for three different headcam configurations, with the device we used in the middle. The lowest camera is pictured for comparison, but was not available until after our study was already in progress.

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.Detailed instructions for creating this headcam can be found at `http://babieslearninglanguage.blogspot.com/2013/10/how-to-make-babycam.html`. 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, left).

Even with the fish-eye lens, the vertical field of view of the camera is still considerably reduced compared to the child's approximate vertical field of view, which spans around 100–120° in the vertical 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 the faces of adults around them. This allowed us to maximize our chances of capturing faces that the child would have seen during the play session.

Procedure

First, all parents signed consent documents in a waiting room where children were fitted with the headcam. After the child was comfortable in the waiting room and with the experimenter, the experimenter placed the headcam on the child's head. 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 (Yoshida & Smith, 2008).

After the child was comfortable wearing the headcam, the child and caregiver were shown to a playroom for the free-play session–the focus of the current study. Parents were shown a box containing three pairs of novel and familiar objects (e.g., a ball and a feather 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 novel 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 videos of 516 minutes (almost a milion frames), with an average video length of 8.6 minutes (min = 4.53, max = 19.35).

Posture and Orientation Annotation We created a set of custom annotations that described the child's physical posture (e.g. standing) and the orientation of the caregiver relative to the child (e.g. far away). The child's posture was categorized

Figure 2: Example face detections made by MTCNN for the headcam videos from a child in each group (green squares).

as being held/carried, prone (crawling or lying), sitting, or standing. The caregiver's orientation was characterized as being close to the child, far from the child, and a global category of caregiver behind the child. For the first two annotations (close/far from the child), the caregiver could either be to the the front or to the side of the child. All annotations were made by a trained coder using the OpenSHAPA/Datavyu software (K. Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012), and times when the child was out of view of the tripod camera were marked as uncodable and were excluded from these annotations.

Face Detection

We used a total of three face detection systems to explore infants' changing access to social information and to avoid the cost of hand-annotating every frame. We first measured the performance of the most commonly-used and widely available face detection algorithms (Viola-Jones). We used this as a benchmark for performance, and 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 capitalized on recent improvements in computer vision, testing the performance of two state-of-the-art face detectors that both made use of 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 agent's bodyparts.

Face Detection Algorithms The first face detection system made use of a series of Haar feature-based cascade classifiers (Viola-Jones, (Viola & Jones, 2004)) applied to each individual frame. This detector provided information about the presence of a face as well as its size and position.

The second algorithm was based on the work by K. Zhang et al. (2016) using multi-task cascaded convolutional neural networks (MTCNNs). The system was built using a novel cascaded CNN-based framework for joint 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 provided by (`https://github.com/davidsandberg/facenet`). Like Viola-Jones, this detector provided information about the presence of a face as well as its size and position.

The third algorithm was a CNN-based pose detector that provided the locations of 18 body parts (ears, nose, wrists, etc.) called OpenPose (Cao et al., 2017; Simon, Joo, Matthews, & Sheikh, 2017; Wei, Ramakrishna, Kanade, & Sheikh, 2016) avaliable 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. For the purposes of our investigation we only made use of the body parts relevant to the face (ears, eyes, nose). We operationalized face detections as any frames containing a nose, as any half of a face necessarily contains a nose.

Detector evaluation In order to evaluate the performance of these detectors, we constructed a "gold set" of frames by hand labeling both a sample of frames with a high density of faces (as reported by the detectors) and a random sample from the remaining frames. This was done so as to not bias our evaluation by the relatively rare appearance of faces in the dataset. A face was present in a frame if at least half of the face was showing. Precision (hits / hits + false alarms), recall (hits / hits + misses), and F-score (harmonic mean of previous measures) were calculated and are reported in Table 2.

Both OpenPose and MTCNN detectors performed relatively well on the gold set, with MTCNN outperforming OpenPose on the random sample but trailing behind in the high density sample. Due to the high performance of these two detectors we have included the results of both in the analyses to follow. The ViolaJones detector did not perform well on either gold set, so the results from this detector were not included.

	Algorithm	Sample Type	P	R	F
1	MTCNN	High density	0.841	1.000	0.914
2	MTCNN	Random	0.947	0.750	0.837
3	OpenPose	High density	0.995	0.863	0.924
4	OpenPose	Random	0.710	0.917	0.800
5	ViolaJones	High Density	0.991	0.495	0.660
6	ViolaJones	Random	0.389	0.292	0.333

Table 2: Detector performance.

Results

First, we report developmental shifts in infants' posture and their orientation relative to their caregiver. Then, we explore how these changes influence children's visual access to faces across this developmental time range. Finally, we explore how these changes impact the accessibility of faces during labeling events.

Changes in Posture and Orientation

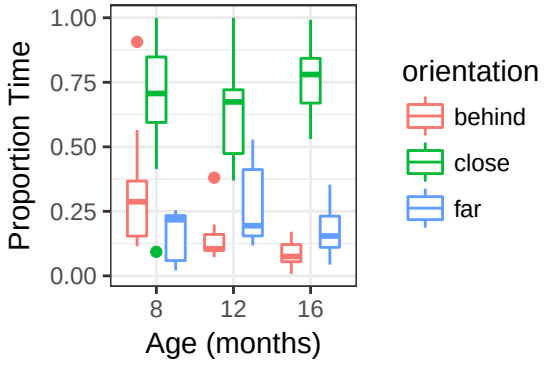

Figure 3: Proportion time that infants in each age group spent in each orientation relative to their caregiver.

We noted characteristic changes in infants' posture and orientation across this developmental time range. The proportion of time infants spent sitting decreased with age, and the proportion of time infants spent standing increased with age. Both 8-month-olds and 12-month-olds spent equivalent amounts of time either lying/crawling, which was markedly decreased in the 16-month-olds, who spent most of their time either sitting or standing (see Figure 4). We also observed characteristic 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 (see Figure 3).

Changes in Access to Faces

We first examined the proportion of face detections across age; a full summary can be seen in Figure 5. We observed a slight U-shaped function both when analyzing the output of the MTCNN and OpenPose detectors, such that 12-month-olds appeared to experience slighly fewer faces faces than 8 or 16-month-olds.

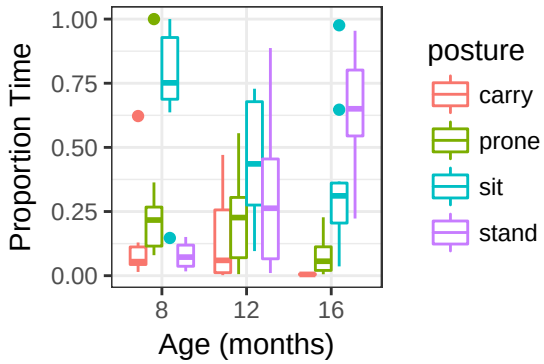

Figure 4: Proportion time that infants in each age group spent in each posture.

However, we found that any age related effects were much smaller compared to the impact of postural and locomotive changes on children's visual access to faces. Children's posture was a major factor both in how many faces they saw during the play session. Infants who were sitting saw more faces than infants who were lying down or being carried, while infants who were standing saw the most faces. We also examined how the child's orientation relative to their caregiver impacted their visual access to faces: children who were far away from their caregiver were more likely to see faces than children who were close to their caregiver; this was true within all age groups and for both face detectors.

To formalize these observations, we fit a generalized logistic mixed-effect model to all with the presence/absence of a face on every frame as the dependent variable, and participant's age, orientation, and posture as predictors. Interactions between predictors were not included as this maximal model failed to converge. A summary of the coefficients of this model can be found in Table 3 for the MTCNN detections, though we found the same pattern of results for the OpenPose detections. While age remained a significant predictor even when accounting for the effects of infants' posture and orientation, it accounted for the least amount of variance. Overall, these results suggest that, instead, infants' access to faces is heavily infleunced by their own posture and theirorientation towards their caregiver.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.847	0.064	-75.303	0.000
scale(age.at.test)	0.170	0.075	2.273	0.023
postureprone	-0.513	0.042	-12.326	0.000
posturesit	0.648	0.039	16.422	0.000
posturestand	1.005	0.041	24.721	0.000
orientationclose	1.549	0.022	71.940	0.000
orientationfar	2.258	0.022	100.357	0.000

Table 3: Results from GLMM model prediting the presence/absence of a face (MTCNN-Faces) across the entire dataset.

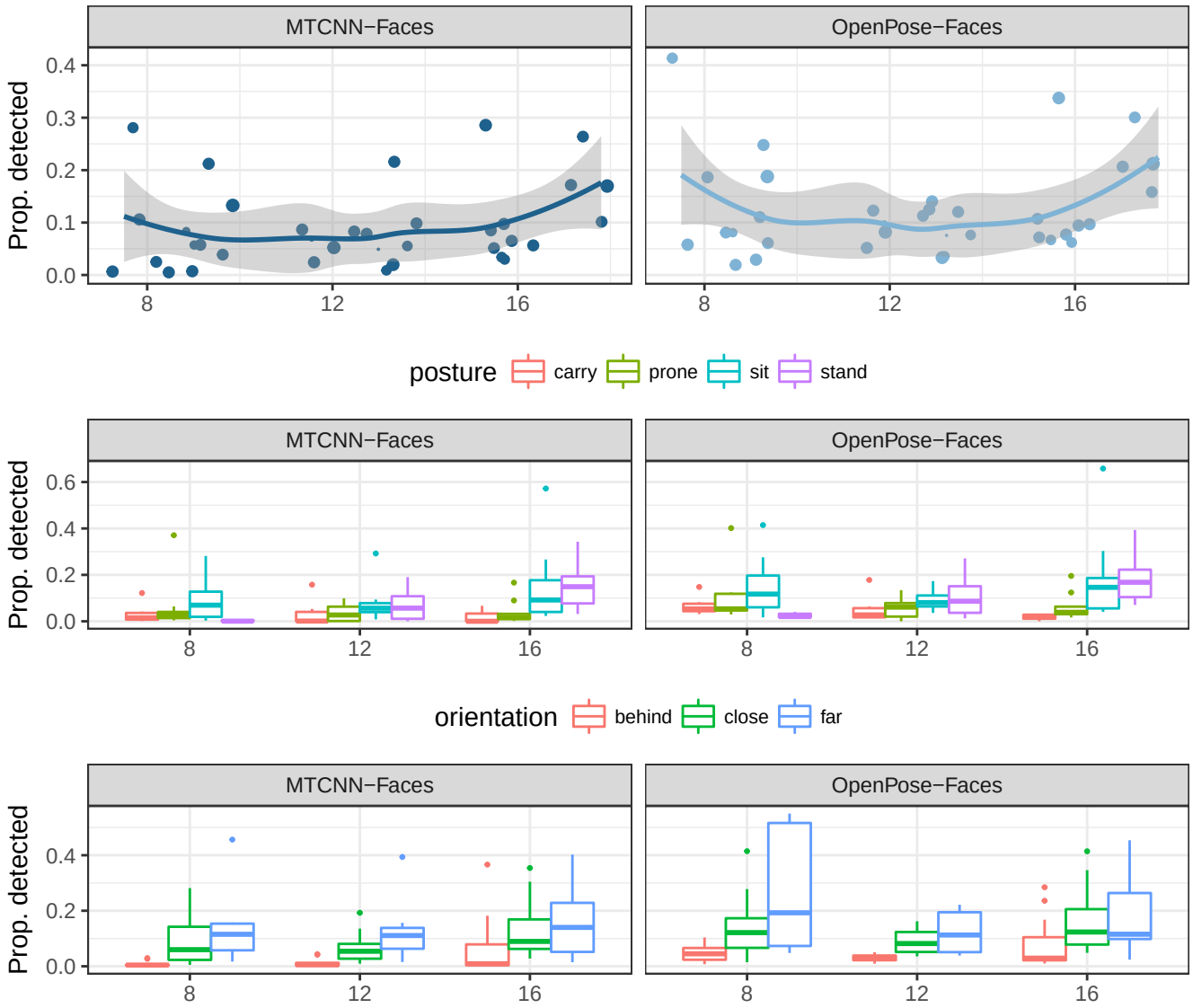

Figure 5: Proportion face (left) and hand (right) detections as a function of participant's age.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.847	0.064	-75.303	0.000
scale(age.at.test)	0.170	0.075	2.273	0.023
postureprone	-0.513	0.042	-12.326	0.000
posturesit	0.648	0.039	16.422	0.000
posturestand	1.005	0.041	24.721	0.000
orientationclose	1.549	0.022	71.940	0.000
orientationfar	2.258	0.022	100.357	0.000

Table 4: Results from GLMM model prediting the presence/absence of faces (OpenPose-Faces) across the entire dataset.

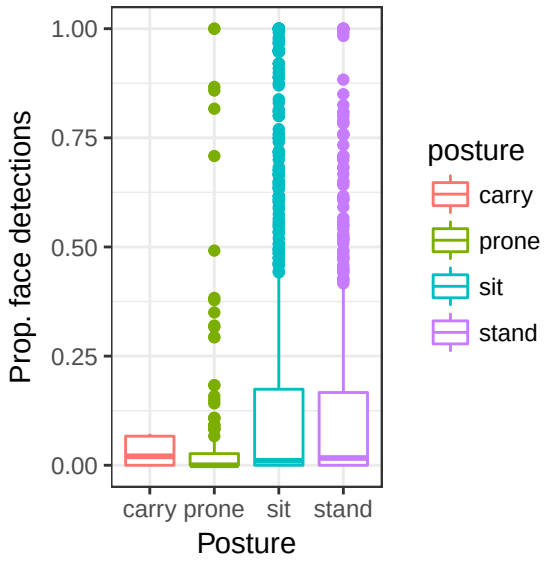

Figure 6: Proportion face detections around a naming instance ('Look, a Zem'; +/- 2 seconds around each utternace) as a function of infants's posture. Each point represents a naming instance for each child in the dataset.

Access to Faces During Labeling Events

Finally, we analyzed how face detections changed during object labeling events as a function of infant's posture and orientation. Specfically, we analyzed a four-second window around each labeling event (e.g., "Look at the [zem]!"); these labeling events were hand-annotated and automatically synchronized with the frame-by-frame face detections. We again found that infant's posture impacted the degree to which they saw their caregiver's face during a labeling event; infants who were sitting or standing were more likely to have access to their caregiver's face.

General Discussion

We use a head-mounted camera to explore how children's postural and locomotive development directly impacts their access to social information relevant for word-learning, here operationalized as the presence of the faces of their caregiver. We found that children's posture and orientation towards their caregiver changed systematically across age, and that all three of these factors dramatically impacted the proportion of faces that were avaliable in the child's visual field. Thus, infants' postural development is a mediating factor that explains age-related changes in the proportion of faces avaliable in infants' visual field.

This work also deploys novel advancements in computer vision to the study of developmental psychology. The field of object detection and recognition has advanced dramatically in the past five years since the re-birth of deep learning algorithms Krizhevsky, Sutskever, & Hinton (2012), creating a new generation of algorithmic tools. These tools are substantially better equipped to deal with noisier, more complicated datasets and can extract richer and more detailed information. Videos from the infant perspective provided substantial challenges (e.g., partially occluded faces) for the classic models of face detection (e.g., ViolaJones, Viola & Jones (2004)). Further, as the headcam technologies employed here were inexpensive (˜$60 a camera) and the computer vision algorithms freely avaliable, this method is a promising avenue for quanitfying the visual and social information avaliable to infant learners.

Thus, we suggest that the combined use of these new tools can be leveraged to understand the changing infant perspecive on the visual world and the implications of these changes for both linguistic, cognitive, and social development.

Acknowledgements

Thanks to Kaia Simmons, Kathy Woo, Aditi Maliwal, and other members of the Language and Cognition Lab for help in recruitment, data collection, and annotation. This research was supported by a John Merck Scholars grant to MCF. An earlier version of this work was presented to the Cognitive Science Society in Frank, Simmons, Yurovsky, & Pusiol (2013). Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall (Jordan Hall), Stanford, CA, 94305, tel: (650) 724-4003, email: mcfrank@stanford.edu.

References

Adolph, K., & Berger, S. (2007). Motor development. In *Handbook of child psychology*. Wiley Online Library.

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

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

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

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

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

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

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

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.

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., Kretch, K., Soska, K., & Adolph, K. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development*.

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).

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.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently.

Child Development, *85*(4), 1503–1518.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

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.

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, 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.

Yurovsky, D., Smith, L., & Yu, C. (in press). 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.