Postural developments modulate children’s visual access to social information

Alessandro Sanchez*  Bria Long*  Allison M. Kraus  Michael C. Frank
sanchez78@stanford.edu  bria@stanford.edu  allison.m.kraus@gmail.com  mcfrank@stanford.edu
Department of Psychology  Department of Psychology  Department of Psychology
Stanford University  Stanford University  Stanford University

Abstract
The ability to process social information is a critical component of children’s early language and cognitive development. However, as children reach their first birthday, they begin to locomote themselves, dramatically affecting their visual access to this information. How do these postural and locomotor 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 36 infants’ (8-16 months of age) egocentric visual perspective and use computer vision algorithms to estimate the proportion of faces and hands in infants’ environments. We find that infants’ posture and orientation to their caregiver modulates their access to social information, confirming previous work that suggests 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.

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

Introduction
From early infancy, children are deeply engaged in learning from others (Csibra & Gergely, 2009; Meltzoff, 2007). 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 (Gredebäck, Fikke, & Melinder, 2010). Further, when infants attend to video stimuli, they tend to look mostly at faces at the expense of other visual information — though older infants start to look towards people’s hands and the actions they are performing (Frank, Amso, & Johnson, 2014; Frank, Vul, & Saxe, 2012).

Then, however, their view of the world radically changes (Adolph & Berger, 2006). Infants’ motor abilities improve dramatically near the end of the first year of life, allowing them to locomote independently. These motor changes have significant consequences for what children see; crawling and walking infants simply have 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 (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 (Kretch, Franchak, & Adolph, 2014).

These motor improvements lead to developmental cascades that impact children’s emerging social, cognitive, and linguistic abilities in various ways. For example, postural changes 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. Thus, children’s ability to independently locomote may change their ability to access social information (e.g., faces, gaze) and in turn accelerate their learning.

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 record the visual experiences of infants and children — which even for walking children are extremely different from the adult perspective (and not easily predicted by our own adult intuitions) (Clerkin, Hart, Rehg, Yu, & Smith, 2017; Franchak et al., 2011; Yoshida & Smith, 2008). Children’s views tend to be more restricted and dominated by objects and hands (Yoshida & Smith, 2008), and computational and empirical work suggests that this restricted viewpoint may be more effective for learning objects and their labels than the comparable adult perspective (Bambach, Crandall, Smith, & Yu, 2017; Yurovsky, Smith, & Yu, 2013). This perspective changes over the first two years of life, as visual input transitions from primarily containing close up views of faces to capturing views of hands paired with the objects they are acting on (Fausey, Jayaraman, & Smith, 2016). It is uncertain how much these changes are driven by the concurrent motoric changes described above versus by other developmental factors.

Here, we directly examine whether postural and locomotor developments change the availability of social information — the presence of faces and hands. To do so, we recorded the visual experience of a group of infants in three age ranges (8, 12, and 16 months) 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 then 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 (using wrists as a proxy for the latter) in the child’s visual environment, both overall and relative to naming events by their caregivers. We hypothesized that there would be differential access to social information based on children’s postural developments: crawling infants would see fewer faces because they would primarily be looking at the ground, while walking toddlers would have access to a richer visual landscape, thus providing children with greater access to the social information in their environment.

*These authors contributed equally to this work.
**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% English at home. Demographics and exclusion rates are given in the table below.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>% incl.</th>
<th>Avg age</th>
<th>Avg video length (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 mo.</td>
<td>12</td>
<td>0.46</td>
<td>8.71</td>
<td>14.41</td>
</tr>
<tr>
<td>12 mo.</td>
<td>12</td>
<td>0.40</td>
<td>12.62</td>
<td>13.48</td>
</tr>
<tr>
<td>16 mo.</td>
<td>12</td>
<td>0.31</td>
<td>16.29</td>
<td>15.00</td>
</tr>
</tbody>
</table>

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. Some of these data were also analyzed in Frank, Simmons, Yurovsky, & Pusiol (2013).

**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 516 minutes (almost a million frames) of video, with an average video length of 14.33 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

---

"Detailed instructions for creating this headcam can be found at [http://babieslearninglanguage.blogspot.com/2013/10/how-to-make-babycam.html](http://babieslearninglanguage.blogspot.com/2013/10/how-to-make-babycam.html)"
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. 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). Although MTCNN and OpenPose performed comparably with the random sample, MTCNN performed better on the high density sample (specifically looking at precision), suggesting that OpenPose generated more false positives than MTCNN. ViolaJones performed quite poorly relative to the other detectors, especially with respect to the random sample. We thus use MTCNN detections in the following analyses. 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sample Type</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTCNN-Faces</td>
<td>High density</td>
<td>0.89</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>MTCNN-Faces</td>
<td>Random</td>
<td>0.94</td>
<td>0.62</td>
<td>0.75</td>
</tr>
<tr>
<td>OpenPose-Faces</td>
<td>High density</td>
<td>0.78</td>
<td>0.93</td>
<td>0.84</td>
</tr>
<tr>
<td>OpenPose-Faces</td>
<td>Random</td>
<td>0.72</td>
<td>0.80</td>
<td>0.76</td>
</tr>
<tr>
<td>ViolaJones-Faces</td>
<td>High density</td>
<td>0.96</td>
<td>0.44</td>
<td>0.60</td>
</tr>
<tr>
<td>ViolaJones-Faces</td>
<td>Random</td>
<td>0.44</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>OpenPose-Wrists</td>
<td>High density</td>
<td>0.66</td>
<td>1.00</td>
<td>0.79</td>
</tr>
<tr>
<td>OpenPose-Wrists</td>
<td>Random</td>
<td>0.88</td>
<td>0.29</td>
<td>0.43</td>
</tr>
</tbody>
</table>

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. Scores in bold are the highest F-scores for each sample type.

Results

Changes in Posture and Orientation

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 lying/crawling, 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 (see Figure 3).

Changes in Access to Faces and Hands

We examined the proportion of face and hand detections across age (see Figure 4). We observed a slight U-shaped function in face detections, such that 12-month-olds appeared to have visual access to slightly fewer faces than 8 or 16-month-olds; conversely, hand detections perhaps appeared to generally increase with age.

Age related effects were much smaller than postural and locomotor changes on children’s visual access to faces and hands. Children’s posture was a major factor in how many faces and hands they saw during the play session. Infants who were sitting saw more faces than infants who were...
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 that infants in each age group spent in each posture/orientation relative to their caregiver.

Figure 4: Proportion of faces detected by the MTCNN model (left) and wrists detected by the OpenPose model (right) as a function of child’s age. Larger dots indicate children who had longer play sessions and thus for whom there was more data.

A summary of the coefficients of the models can be found in Table 2. While infants’ posture and orientation significantly impacted the proportion of faces and hands that infants saw, age was not a significant predictor in either model. Thus, these results suggest that infants’ visual access to social information is modulated by their posture and orientation, which is in turn a function of their general locomotor development.

Access to Faces and Hands During Labeling Events

Finally, we explored how face and hand detections changed during object labeling events as a function of infants’ posture and orientation. We analyzed a four-second window around each labeling event (e.g., “Look at the [zem]!”); these labeling events were hand-annotated and synchronized with the frame-by-frame face/hand detections. We found that infants’ posture and orientations modulated their visual access to faces and hands during labeling events; infants who were sitting or standing were more likely to have visual access to this social information (see Figure 6). However, we did not find that infants saw particularly more faces or hands during naming events relative to baseline (avg. difference in propor-
Estimate | Std. Error | z value | Pr(>|z|)  
--- | --- | --- | ---  
**Faces**  
Intercept | -5.78 | 0.33 | -17.55 | 0.000  
Close | 2.09 | 0.31 | 6.70 | 0.000  
Far | 2.58 | 0.34 | 7.54 | 0.000  
Prone | -0.18 | 0.08 | -2.39 | 0.017  
Sit | 1.00 | 0.07 | 13.54 | 0.000  
Stand | 1.50 | 0.08 | 19.98 | 0.000  
Age | 0.17 | 0.17 | 0.97 | 0.334  
**Hands**  
Intercept | -4.84 | 0.19 | -25.40 | 0.000  
Close | 0.88 | 0.17 | 5.25 | 0.000  
Far | 1.77 | 0.27 | 6.45 | 0.000  
Prone | 0.58 | 0.10 | 5.73 | 0.000  
Sit | 1.18 | 0.10 | 12.04 | 0.000  
Stand | 1.01 | 0.10 | 10.22 | 0.000  
Age | 0.14 | 0.10 | 1.40 | 0.161  

Table 2: Model coefficients from generalized linear models predicting the proportion of faces (upper panel) and hands (lower panel) seen by children.

Figure 5: Proportion of face and wrist detections in each age group as a function of children’s posture (top panel) and orientation (bottom panel).

Figure 6: Proportion of face and hand detections around a naming instance ('Look, a Zem'; +/- 2 seconds around each utterance) as a function of infants’ posture. Error bars represent non-parametric bootstrapped 95% confidence intervals.

General Discussion

We used a head-mounted camera to explore how children’s postural and locomotor development affects their visual access to social information, here operationalized as the presence of the faces and hands of their caregiver. Children’s posture and orientation towards their caregiver changed systematically across age, and both of these factors influenced the proportion of faces and hands that were available in the child’s visual field. This work suggests that motor development modulates how infants experience their visual world and the social information in it: Infants that are sitting and standing have a different view of their world, the people in it, and the actions that are being performed.

We created a situation in the lab in which the context of interaction was tightly controlled, but as children grow and change, the activities in which they engage with their caregivers also vary, leading to differences in the distribution of contexts they experience. Thus, more work is needed to understand how our results relate to children’s home experiences (Fausey et al., 2016). The ability to walk is part of a cascade of changes in children’s experience, and our study...
captures only a cross-sectional slice of this broader, multifaceted trajectory.

Understanding these changes 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 systems. We hope that these new tools can be leveraged to understand the changing infant perspective on the visual world and the implications of these changes 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 meeting of the cognitive science society in Frank et al. (2013).

All data and code for these analyses are available at https://github.com/amsan7/xs-face

References


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).


