Measuring children's visual access to social information using face detection

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

Abstract

Other people are the most important source of information in a child's life, and one important channel for social information is faces. Faces can convey affective, linguistic, and referential information through expressions, speech, and eye-gaze. But in order for children to apprehend this information, it must be accessible. How much of the time can children actually see the faces of the people around them? We use data from a head-mounted camera, in combination with face-detection methods from computer vision, to address this question in a scalable, automatic fashion. We develop a detection system using off-the-shelf methods and show that it produces robust results. Data from a single child's visual experience suggest the possibility of systematic changes in the visibility of faces across the first year, possibly due to postural shifts.

Keywords: Social development; face processing; head-camera.

Introduction

Faces are perhaps the most important source of social information for young children. Infants show a preference for faces and face-like configurations from birth (Johnson, Dziurawiec, Ellis, & Morton, 1991; Farroni et al., 2005), and they will fixate faces to the exclusion of nearly everything else when attending to complex naturalistic stimuli (Frank, Vul, & Johnson, 2009; Frank, Vul, & Saxe, 2011). By their first birthday, they are sensitive to facial information about emotion (Cohn & Tronick, 1983) and social group (Kelly et al., 2005), and they will readily follow gaze to an attended target (Scaife & Bruner, 1975). As they begin to speak and understand language, joint attention becomes a powerful cue for learning the meanings of words (Baldwin, 1991).

To extract all of this important information in the natural environment, infants and children must attend to people's faces. Nearly all of what we know about children's attention to—and understanding of—faces comes from tightly-controlled lab experiments. In such experiments, the stimuli are typically presented in a very accessible format: at eye-level, large enough so that all details can be appreciated. How often do children actually see the faces of the people around them, though? And how often are the faces large enough to discern details from?

Head-mounted cameras provide a new technique for measuring access to faces during development. While the method of placing a miniature camera on the head of an infant or young child is still relatively new, a number of investigators have begun using it to record children's first-person perspective (Yoshida & Smith, 2008; Aslin, 2009; Smith, Yu, & Pereira, in press). Some studies have even used head-mounted eye-trackers to measure what part of the visual scene the child is fixating, a good proxy for what parts of the world the child is attending to (Franchak, Kretch, Soska, Babcock, & Adolph, 2010).

Of particular interest is the result, reported by Franchak et al. (2010), that 14-month-olds rarely fixated their mother's face, even when she spoke to them directly. They looked instead at her hands or other parts of her body. The authors speculated that this result might have been due to the mother's location, usually high above the child. When mothers were sitting down, their faces were much more visible to their children. In our current investigation we follow up on this suggestion, investigating the possibility that the posture of caregivers and the infant's own posture work together to cause developmental changes in the accessibility of social information.

The introduction of these new methods mean that for the first time, we can see what babies are looking at as they interact with—and learn from—the people around them. This development opens up many new questions for investigation. Yet work of this type is hindered by the tremendously slow and resource-intensive task of manually annotating videos, frame by frame. Up until now, only a few research groups have grappled with the task of how to analyze the massive datasets captured using these methods.

The current study thus serves two purposes. First, it is designed to measure the accessibility of social information—in the form of faces—to infants. To investigate this question across development, we make use of a previously-described dataset (Aslin, 2009), in which a head-mounted camera recorded 2 – 3 hours of the visual experience of a single child at ages 3, 8, and 12 months (sample frames shown in Figure 1). Second, we investigate the possibility of using automated face detection to measure social information. It might in principle be possible to hand-annotate the presence of faces in each of the million-odd frames in our dataset (such annotation can be done around 4–8 times slower than real-time, yielding around 25–50 hours of total annotation time). For any larger study with more participants, annotation costs would quickly become prohibitive. Our study thus was designed to serve as proof-of-concept for the automated strategy.

Detection of upright faces in static images is widely considered to be a solved problem in computer vision, with the work of Viola and Jones (2004) providing a computationally-efficient solution that is now used in a wide variety of systems and consumer electronics. Nevertheless, the dataset we used presents a distinct set of challenges for such methods. In what follows, we describe our method for handling these challenges using a collection of out-of-the-box techniques from computer vision and machine learning. We end by describing

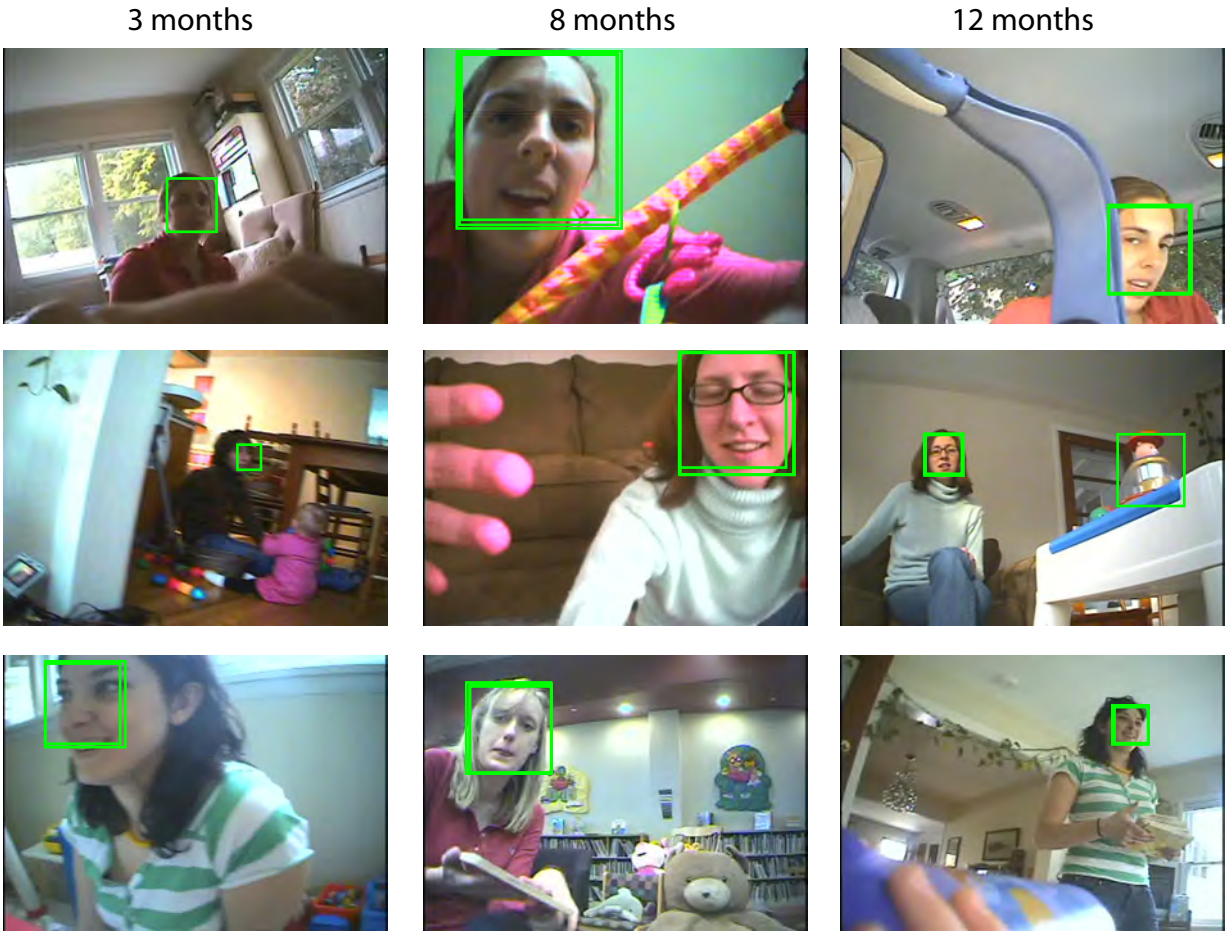

Figure 1: Sample frames showing head-camera data plotted along with face detector data. A separate rectangle is plotted for each active detector. Frames were selected in which annotations and model predictions matched.

developmental changes in the prevalence and size of faces in the field of view of the infant we studied. These changes suggest that there may be a number of important factors influencing the accessibility of social information during early development.

Methods

Although in principle a single joint detection and tracking system could be constructed to detect faces in head-camera video, in practice such as system would be complex and computationally-intensive. Thus, we pursued a two-step approach to face-detection (Figure 2). We first preprocessed each frame of our data separately using simple but noisy detectors, which find faces in static images. We then tested a number of supervised post-processing models on their performance in picking frames with successful rather than spurious detections.

Because of this two-step scheme, conventional annotations of a gold standard training sample (e.g. face/no face) were not maximally effective. If the detectors did not find a face in a frame, training the post-processing model that the frame contained a face would be counterproductive. Instead, our strategy was to create two annotated sets. The first was a training set that indicated whether, for each frame, the detectors had correctly identified a face. The second was a generalization dataset that indicated whether a face was in fact present in the frame, allowing us to test what proportion of faces our models identified on a completely independent dataset (different clips from the same corpus). In addition, we annotated the child's posture in each video of the corpus. These annotations (along with the details of the dataset) are described below.

Data and annotation

Aslin (2009) head-camera dataset Data for the study consisted of videos collected on three days during the infancy of a single child, at ages 3, 8, and 12 months. This dataset was originally collected by Aslin (2009); the data are described in detail in Cicchino, Aslin, and Rakison (2010). The method of collection was a small wireless camera mounted

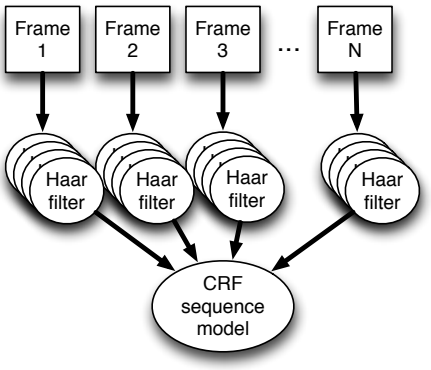

Figure 2: Schematic of our face-processing approach. Step 1: process frames with noisy Haar-style detectors. Step 2: filter detections with conditional random field (CRF) model.

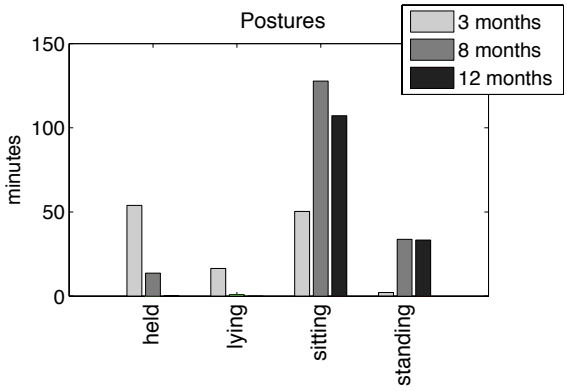

Figure 3: Time spent in each coded posture at each age.

on the infant's head, allowing recording of a large portion of the infant's visual field. The camera was a Sony 480TVL CCD "bullet" camera, embedded in a headband and wirelessly transmitted to a digital video recorder. Videos were approximately 126, 190, and 140 minutes long for the segments collected at 3, 8, and 12 months, respectively. Recordings were made while the infant was in a number of different locations, including in the home, on a shopping trip, on a walk, and at a playgroup. Due to the variation in activities across ages, the natural statistics of these three samples were unmatched (likely due to both sampling issues and true differences in the distribution of activities across ages); thus we will not attempt to compare across activity types.

Annotation of detectors (training set) We annotated a sample of videos to provide training data for our models. For this annotation effort our goal was to select frames in which the raw face detectors had correctly selected a face (and reject those for which the detections were incorrect). We classed a frame as containing a correct detection if there was at least one detector around the face of a person (thus a frame could still contain some spurious detections, though in practice this was relatively rare). We annotated nine clips of one minute each (16k frames). Three minute-long clips were selected for each age group randomly, with the caveat that they included some correct face detections in each.

Annotation of face presence (generalization set) We additionally performed frame-by-frame annotations of whether a face was present in the video frame. We selected 3–4 one-minute clips at each of the three ages for a total of 11 minutes of video at 30 frames per second (20k frames). One-minute clips were selected randomly, again with the caveat that they needed to contain at least some instances of faces. We counted a frame as containing a face when a face was fully visible with no occlusions at three-quarter view or greater (both eyes visible). This stringent annotation criterion was used because occluded or profile-view faces are much less likely to be useful for inferring eye-gaze direction, emotional state, or other social information.

Posture annotation We additionally annotated the posture of the child during the videos, in order to use this factor in our analysis of position and size of detected faces. We attempted to estimate the child's posture wherever possible, categorizing it as lying, sitting, standing, crawling, or being held. Figure 3 shows descriptive data for this measure. Annotation of this measure was somewhat subjective, but inter-rater agreement was relatively high with $\kappa = .72$ for five categories.

Models

Although face detection is generally considered to be a solved problem (Viola & Jones, 2004), face detection in developmental, first-person data presents a number of challenges that do not usually occur in static photographs or standard videos. First, faces are often occluded and at odd orientations for children. Second, in our case, the video was transmitted wirelessly and contained some artifacts due to the transmission method. Third, the head-mounted camera was subject to quick movements as the child moved his head, meaning that many methods applicable for scene segmentation or motion tracking in static-camera applications could not be used here. Our modeling goal in this project was to combine computationally-inexpensive techniques to address these challenges.

Our preprocessing step made use of off-the-shelf Haar-style detectors from the OpenCV package (Bradski & Kaehler, 2008). Each frame was processed with four separate detectors: three full-face and one profile detector. These detectors were noisy, capturing many faces but also spuriously identifying many background elements as faces as well (e.g. doorknobs, high contrast windowpanes, see Figure 4). This processing step ran at approx. 10% of real time on a quad-core machine, taking around 4 days to process all detectors.

Next, we trained post-processing models to discriminate valid detections from invalid detections, using our detector-annotated training set. Our primary model of interest was a

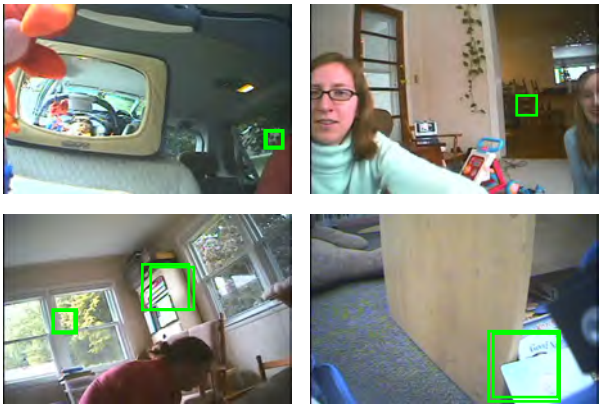

Figure 4: Frames in which CRF model incorrectly predicted that the detectors had correctly identified a face.

conditional random field (CRF) model (Lafferty, McCallum, & Pereira, 2001). CRFs are discriminative sequence models: they take input data of sequences of observations (with some feature set describing each observation) and return a classification of each observation in the sequence. Their key difference from feature-based classifiers (e.g. Naive Bayes or MaxEnt) is their ability to use sequential information; likewise, their key difference from sequence models (e.g. hidden Markov models) is their ability to incorporate rich featural information about each observation. They have been applied successfully to a number of tasks including natural language processing and computer vision. For this application, we used the Matlab CRF toolbox (Schmidt & Swersky, 2008).

We included two other simpler models for comparison: a Naive Bayes (NB) classifier and a hidden Markov model (HMM). The classifier made use of exactly the same feature set but considered each frame in isolation (neglecting sequential dependencies). The HMM considered only the sequence of decisions and the number of detectors that were active. Thus, the difference in performance between the CRF and the classifier provides a rough measure of the contribution of sequential information (provided by the video), while the difference between CRF and HMM provides an estimate of the gain due to adding featural information.

We created a set of binary features to describe the detections in each frame. These included a separate feature for whether each detector was active, a feature for each detector pair to indicate whether the detector centers fell within a certain threshold (5 pixels) of one another, and features for each detector indicating whether it changed in size or disappeared in either the preceding or following frame. We used this feature set to train the models to classify the training data as containing correct or incorrect detections.

Table 1: Model performance on detector-annotated training dataset ("Tr," 9 minutes, only frames with successful detections) and generalization dataset ("Gen," 11 minutes, all frames with human-visible faces). P = precision, R = recall, F = F-score (harmonic mean of precision and recall).

	Tr P	Tr R	Tr F	Gen P	Gen R	Gen F
NB	.64	.82	.72	.76	.55	.64
HMM	.72	**.85**	.78	.81	**.57**	**.67**
CRF	**.85**	.77	**.81**	**.85**	.53	.65

Table 2: CRF model performance on generalization training set by age. Prop. faces refers to the proportion of total faces in the gold-standard dataset for that age.

	3 months	8 months	12 months
Precision	.92	.89	.33
Recall	.60	.48	.35
F-score	.73	.62	.34
Prop. faces	.46	.45	.07

Results

Table 1 shows evaluation results for each of the models on the two datasets we annotated, the detector-annotated training dataset and the gold-standard generalization dataset. (Rather than using a technique like cross-validation to test generalization performance, we report results on both the training data and an independent generalization set that was never used for training). The CRF model performed best on the training data, capturing a slightly better tradeoff between precision and recall. The gain in performance was relatively slight from the HMM to the CRF, indicating that the majority of the value of the CRF was due to the sequential dependencies enforced by the model. Knowing that a previous frame contained a successful detection was helpful in deciding whether the current one did as well.

When we applied the three models to the generalization dataset, F-scores were within a small range of one another, with the HMM outperforming the CRF, perhaps indicating some overfitting of feature weights to the training data. Nevertheless, the CRF produced the highest precision on the generalization dataset. Because our aim was to measure the quantity and spatial distribution of faces at each age, we judged precision more valuable than recall and chose the CRF model for our analysis (though we note that results do not change meaningfully if the other models are chosen).

Performance on the generalization set was highly asymmetric across the three ages, with high precision and recall for the 3-month data, mid-level performance on the 8-month data, and very low performance on the 12-month data (Table 2). A number of experiments attempting age-specific training failed to find major gains in performance by training only on e.g. 12-month data. There were few faces in the 12-month

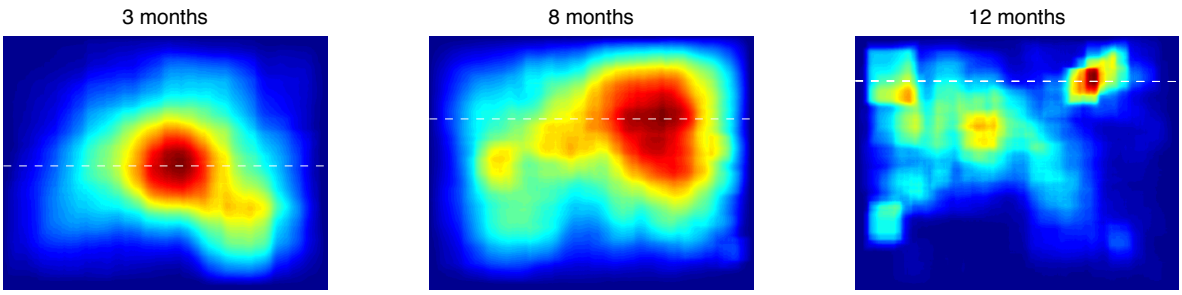

Figure 5: Heat maps showing probability of finding a face in each location of the camera field for 3, 8, and 12-month-old data. Dotted lines show the vertical locations of maxima.

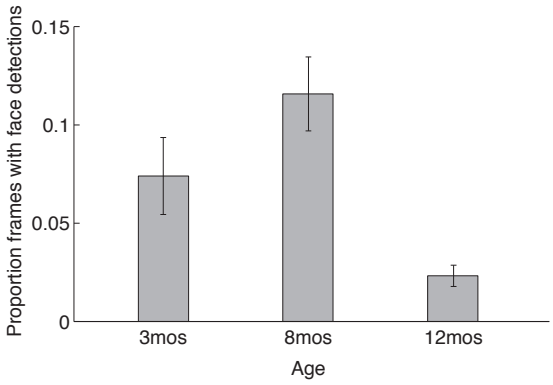

Figure 6: Proportion of faces detected by CRF model at each age. Error bars show standard error across video clips.

data (7% of frames, compared with 46% of frames in the 3-month data), and those that were present were very hard to detect correctly, perhaps because of their small size. Figure 4 shows frames in which the CRF model incorrectly reported a face; these typically showed consistent spurious detections for some superficially face-like configuration of objects.

We evaluated the CRF model on the entire dataset, using the settings established in training. Congruent with the generalization data, we found very few faces in the 12-month data relative to the other two ages. Figure 6 shows the estimated proportion of face-containing frames across clips at each age. Nevertheless, we should be cautious in interpreting these results, due to the relatively small amount of data available in this dataset. It may be the case that these results are skewed due to, e.g., participating in a play-group at 8 months with many children present.

Figure 5 shows a heat map of the probability of finding a face at each location in the camera's field for each of the three samples.[1] Faces were higher in the image plane at 8 and 12 months than at 3 months. This shift could potentially be due to postural differences: the child was more likely to be held or lying down at 3 months and more likely to be sitting or standing at 8 and 12 months. In a sitting or standing position, faces tend to be higher in the visual field than when lying down and looking up over the edge of the crib.

Faces were also different sizes in the older videos. The 3-month videos had a qualitatively different distribution of detected face sizes (Figure 7). We cannot completely rule out the possibility that some of the smaller faces in the 8- and 12-month videos were spurious detections. Nevertheless, the relatively similar distribution for each of these (compared with the drop in precision from 8 to 12 months) suggests that decreasing precision of detections was not the only factor here. Though speculative, a postural explanation for the shift in size might also be proposed: at older ages, the child was less likely to be lying or being held close to the face of a caregiver. Instead, in a seated or standing position, the faces of others would be further away.

Our final analysis directly measured size and vertical position of faces by posture (due to the limited overlap in postures between ages, regression analysis was not possible). The lying posture, seen only at 3 months, had a much larger face size than the other postures (almost 9% of camera field, as opposed to 2.5% for holding, 3% for sitting, and 4% for standing). Both lying and being held also had lower average vertical positions (.50 and .52 respectively, where 1 was the top of the screen) than sitting and standing (.60 and .57, respectively).

General Discussion

We investigated the possibility of using automated face detection techniques to measure the accessibility of social information to infants. With head-mounted videos from a single infant at 3, 8, and 12 months, we constructed a discriminative model of face-detection that made use of inexpensive but noisy detectors and a secondary filtering step using a conditional random field model. This approach was relatively suc-

[1]For this and the remaining analyses, we averaged across all detections for each frame, potentially including some noise due to spurious detectors in correct frames. Future work should look estimate the location of correct detections as well as frames in which they occur.

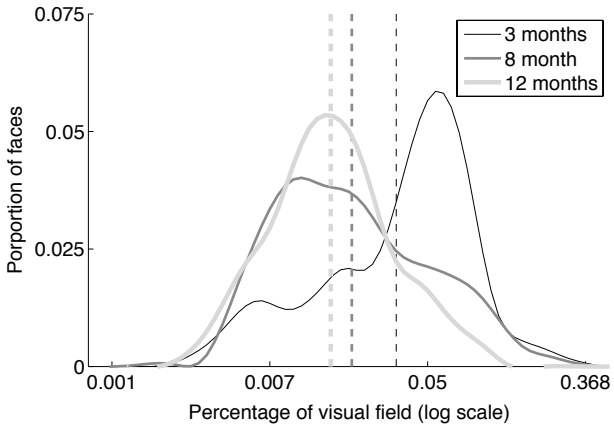

Figure 7: Smoothed histogram of detected face sizes at 3, 8, and 12 months. Height shows proportion of detections at each size; horizontal axis is scaled in log proportion of camera field. Dashed lines give means.

cessful in picking out correct detections for the dataset as a whole.

The face-detector data revealed a surprising pattern. Faces were far less frequent in the 12-month data (and harder to detect, providing a potential caveat to our descriptive results). In addition, those faces that were detected in the older part of the dataset were both smaller and higher in the visual field of the infant. These differences seemed related to the distribution of postures across different ages, and indeed size and horizontal position did vary with posture. Nevertheless, further research (and considerably more data) will be necessary to check these conclusions.

The speculative picture that emerges is nevertheless congruent with previous work (Franchak et al., 2010). As children grow and become more adept at locomotion, they create a situation where the faces of others in their environment are further away from them and less visible. While the young infant is constantly having the faces of others pressed into his, the toddler lives in a world populated by knees.

More broadly, the methodological upshot of this work is that head-camera footage may be an extremely valuable tool for studying social attention and access to social information "in the wild." Nevertheless, this work cannot proceed if hand-annotation is the only solution. Computer vision methods that are appropriate for data of this type must be developed, and this study took a first step in that direction, revealing suggestive developmental differences.

Acknowledgments

Special thanks to Richard Aslin for generously sharing the primary dataset used in this project. Additional thanks to Ally Kraus for help with data organization and annotation, as well as to Adam Vogel and Evan Rosen for work on an earlier version of this project. This work was supported by a John Merck Scholars Fellowship.

References

Aslin, R. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry & Vision Science*, *86*, 561.

Baldwin, D. (1991). Infants'contribution to the achievement of joint reference. *Child Development*, 875–890.

Bradski, G., & Kaehler, A. (2008). *Learning opencv: Computer vision with the opencv library*. O'Reilly Media.

Cicchino, J., Aslin, R., & Rakison, D. (2010). Correspondences between what infants see and know about causal and self-propelled motion. *Cognition*.

Cohn, J. F., & Tronick, E. Z. (1983). Three-month-old infants' reaction to simulated maternal depression. *Child Development*, *54*, 185–193.

Farroni, T., Johnson, M., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 17245.

Franchak, J., Kretch, K., Soska, K., Babcock, J., & Adolph, K. (2010). Head-mounted eye-tracking of infants natural interactions: A new method.

Frank, M., Vul, E., & Johnson, S. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*, 160–170.

Frank, M., Vul, E., & Saxe, R. (2011). Measuring the development of social attention using free-viewing. *Infancy*, 1–21.

Johnson, M., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*, 1–19.

Kelly, D., Quinn, P., Slater, A., Lee, K., Gibson, A., Smith, M., et al. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*, F31.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. int'l conf. on machine learning (icml)* (pp. 282–289).

Scaife, M., & Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature*.

Schmidt, M., & Swersky, K. (2008). *Conditional Random Field Toolbox for Matlab.* (http://www.di.ens.fr/ mschmidt/Software/crfChain.html)

Smith, L., Yu, C., & Pereira, A. (in press). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*.

Viola, P., & Jones, D. H. (2004). Robust real-time face detection. *International Journal of Computer Vision*.

Yoshida, H., & Smith, L. (2008). What's in view for toddlers? using a head camera to study visual experience. *Infancy*, *13*, 229–248.