

REVIEW

How Infants View Natural Scenes Gathered From a Head-Mounted Camera

Richard N. Aslin*

ABSTRACT

The role of early visual experience in human infant development has been inferred primarily by studies of visual deprivation (e.g., cataracts). Another approach, described here, is to provide a detailed description of the visual input gathered by normal infants in their natural environment. Recently, several labs have begun the laborious process of obtaining video images from a head-mounted camera to provide an infant's eye-view of their visual world. Preliminary findings from one such study are reviewed and discussed in the context of the power and limitations of this approach for revealing important insights about the role of early visual experience, as well as the broader implications for studies of cognitive, language, and social development. (Optom Vis Sci 2009;86:561–565)

Key Words: infant, head-camera, eye tracking, perceptual development, visual learning

Studies of infant perception, cognition, and language have for four decades relied on a variety of behavioral measures, the most ubiquitous of which assesses how infants' gaze is directed to visual or auditory-visual stimuli.¹ Most of these measures of infant gaze involve a global definition; i.e., whether the infant is looking at a stimulus or looking away from that stimulus. A small number of studies use more detailed measures of gaze, some aimed at assessing eye-movement control to a small target,² and others at gathering sequential shifts in gaze (scan paths) as the infant views more complex patterns or scenes.^{3–6} These studies of detailed scanning build on classic research^{7,8} by using commercial eye-trackers to collect and analyze infant gaze under more user-friendly circumstances.⁹

In all these studies, the experimenter seeks tight control over what the infant sees in an attempt to test a specific hypothesis that requires a particular stimulus contrast. Missing from these studies is the ability of the infant to select more broadly from potentially available scenes, for example by moving its head to open the field of view to regions of interest that are beyond the default head position. Even when researchers present dynamic scenes on a video display,⁴ which some argue are more natural than static images, the implicit assumption is that these dynamic video displays are a

reasonable proxy for natural scenes selected by the infant rather than by the experimenter.

In the past few years, several labs have begun to ask whether studies of infants' vision, learning, social interaction, and language development have been missing a key perspective on what the infant is attending to in the natural environment. The goal of these studies is not to determine what infants attend to when a stimulus is presented by an experimenter within a restricted field-of-view (i.e., a TV monitor), but rather what infants seek out in the natural environment as momentary targets of attention when there are few constraints on where their gaze might be directed. This goal is related to a substantial literature in the study of the adult visual system that seeks to describe the statistics of natural scenes.^{10,11} A core assumption in this literature is that many of the visual feature analyzers in the brain have evolved to be optimally tuned to the stable features of the visual environment. But rather than asking what information is present in the visual world, and presuming that adults sample that world extensively, here, we ask what subset of that visual world comprises the input to the infant's visual system.

The ideal situation for assessing where infants direct their gaze would be a head-mounted eye-tracker because it would allow the infant to turn its head to sample a much larger field-of-view than is available on a TV monitor. However, it has proven extremely difficult to gather infant gaze with such a head-mounted system (but see ref. 12 for a recent exception). Other researchers, most recently Yoshida and Smith,¹³ have relied on a head-mounted camera to obtain an approximate view of what infants attend to in natural visual environments. However, as they note, infants can

*PhD

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.optvissci.com).

direct their gaze anywhere within ± 30 deg of where their head is pointed, and so the more subtle details of where gaze is directed are lost from these head-camera videos. Moreover, the context used by Yoshida and Smith was restricted to a small table on which three objects were placed, with the mother facing the infant on the other side of the table. Although this context is more dynamic than those used in most eye-tracking studies, it, nevertheless, does not approach the complexity of scenes available to infants in their everyday environment.

Despite the poor precision of a head-mounted camera for documenting dynamic changes in gaze (focal attention), several labs are pushing this approach as a low-resolution proxy for the deployment of visual attention in the natural dynamic environment (P. Sinha, personal communication, C. von Hofsten, personal communication). It will be illuminating to see what these studies can tell us about the early input to the visual, motor, and language systems that could not be captured by so-called third-person views in which an external camera mounted near the infant provides a recording of both the infant and the surrounding visual world.

An alternative approach to a head-mounted eye-tracker was described by Fiser et al.¹⁴ They adopted a hybrid strategy that takes advantage of the best features of a head-mounted video camera and a commercial (head-fixed) eye-tracking system. The first component, as in Yoshida and Smith,¹³ was to use a head-mounted camera to obtain the central $\pm 45^\circ$ of the infant's field of attention under a variety of natural viewing conditions. The second component, as in recent eye-tracking studies,^{3–6,15} was to obtain a detailed record of infants' shifts in gaze as they viewed the dynamic videos obtained from the head-mounted camera on a large-screen display. Although this hybrid strategy is clearly not ideal, because the dynamic videos are not provided by the same infant and in real-time synchrony with their active control of gaze, it does provide a first detailed glimpse at how infants deploy their gaze to natural scenes that are defined from a "dynamic first-person view" (i.e., from the infant's perspective in their visual world).

METHODS

In Fiser et al.,¹⁴ head-mounted videos were gathered from one male infant aged 15 weeks at the time of the first collection date and 38 weeks at the second collection date. The infant wore a headband onto which was attached a small SONY "bullet" camera whose color images were relayed to a digital video recorder via a wireless transmitter. The angle of the bullet camera was adjusted so that the field-of-view (90°) captured the central, straight-ahead gaze position of the infant. Several hours of video recordings were collected at the two ages as the mother and infant went about their everyday activities (play, feeding, sitting in a car-seat, or stroller). From these videos, a set of 3-min vignettes was extracted to illustrate a variety of behavioral contexts that included people, salient objects, and dynamic events (e.g., moving in a grocery cart through the supermarket—see Appendix, Supplemental Digital Content 1, which describes each video vignette, <http://links.lww.com/A1111>), which contained no artifacts from the wireless transmission to the digital video recorder.

These 3-min video vignettes (without audio) were then presented to different infants (4- and 8-month-old infants) and to adults on a 42-inch SONY plasma display screen. The viewing

TABLE 1.

Mean percentage of each video scenario that elicited a reliable measure of gaze from the eye-tracker (adult data in parentheses), and the percentage of the three video scenarios at the two ages that contained images of each of the three coding categories

	Eye tracks	Faces	Hands	Objects
4-month-old infants				
Wegmans	72.7 (85.0)	83.8	51.9	97.7
Home	49.5 (83.7)	77.9	59.1	96.3
Wal-Mart	62.2 (79.4)	81.3	21.3	99.8
8-month-old infants				
Play	49.2 (85.7)	89.2	36.9	87.0
Blocks	31.3 (77.1)	44.5	89.7	95.6
School	25.3 (77.9)	96.8	5.9	70.1

distance was 85 cm, so the images in the videos were approximately $\frac{1}{2}$ life size. An ASL model 504 automated corneal-reflection eye-tracking system (with head-movement compensation) was used to obtain detailed records of gaze position as the infants and adults watched these dynamic video images. The eye-tracking data were coded by human observers (from cross-hairs on each video frame) and sorted into three primary categories: fixations on faces, hands, and salient objects (small and back-grounded objects were excluded, see Appendix, Supplemental Digital Content 1, <http://links.lww.com/A1111>). The 4-month-old infants and adult controls viewed three video scenarios gathered from the head-mounted camera worn by the male infant at 15 weeks of age. The 8-month-old infants and adult controls viewed three video scenarios gathered from the head-mounted camera at 38 weeks of age. Thus, infants who were being eye-tracked viewed age-matched video scenarios. Twenty percent of all video scenarios were coded by a second observer to obtain inter-rater reliabilities, which were 96.2% (range, 94.6 to 97.7) for infants and 97.5% (range, 96.2 to 99.3) for adults.

RESULTS

Table 1 shows that the percentage of natural video scenarios that contain images of faces and salient objects (i.e., large and foregrounded) is very high ($>70\%$), whereas the percentage of videos that contain images of hands (either the infant's own hands or the hands of others) is considerably less (6 to 59%). The sole exception to these trends was the "Blocks" video shown to the 8-month-old infants, which specifically involved a series of interactions with nearby objects. These summary data extend the report of Yoshida and Smith,¹³ who focused their analyses of head-camera videos on a behavioral context more similar to the blocks video. Thus, it is not surprising that the majority of the gaze estimates of the 15-month-old infants from the study of Yoshida and Smith were directed to hands and objects rather than to faces.

Fig. 1 shows the mean percentages of gaze that were directed to faces, hands, and objects in the three video scenarios presented to



FIGURE 1. Percentage of gaze in the 4-month-old infants (and adult controls) directed to the three categories of interest (Faces, Hands, Objects). Error bars = SEM.

the 4-month-old infants and their adult controls. Two findings are readily apparent from these summary data. First, 4-month-old infants rarely look at images of hands; rather, they spend most of their time looking at salient objects and faces. Second, there is remarkable consistency between the distributions of looking times for infants and adults. Three separate analyses of variance (ANOVA) were run on each dependent measure (fixations of people, hands, and objects). None of these three ANOVAs revealed significant effects of age (4-month-old infants vs. adult), video (grocery store, home, Wal-Mart), or their interaction, except for a significant age effect for the salient-object fixations ($F(1,5) = 7.87, p = 0.038$).

Fig. 2 shows the mean percentages of gaze that were directed to faces, hands, and objects in the three video scenarios presented to the 8-month-old infants and their adult controls. Similar to the younger infants, the category that was least fixated was images of hands, except in the blocks video where hands were present in 89% of the video frames. Thus, in general, the 8-month-old infants also predominantly gazed at faces and objects. In contrast to the younger infants, however, there was less consistency in the distribution of gaze across the three categories between 8-month-old infants and adults. As with the 4-month-old infants, three separate ANOVAs were run on each dependent measure (fixations of people, hands, and objects). For fixations to people, there were significant effects of age ($F(1,7) = 25.72, p = 0.002$), video scenario ($F(2,14) = 5.76, p = 0.015$), and their interaction ($F(2,14) = 8.63, p = 0.004$). These results reflected the greater percentage of looking to images of people by adults than by 8-month-old infants in the school and blocks video scenarios. For fixations to hands and objects, there were only significant main effects of video ($F(2,14) = 38.2, p < 0.0001$; $F(2,14) = 5.89, p = 0.014$), reflecting the difference between the blocks video scenario compared to the play and school scenarios.

DISCUSSION

The Fiser et al.¹⁴ study represents a first attempt at gathering detailed eye-tracking data from infants as they view dynamic natural scenes obtained from a first-person perspective. Similar to Yoshida and Smith,¹³ we collected video images from an infant's point-of-view using a head-mounted camera. However, unlike Yoshida and Smith, we relied on detailed measures of gaze directed to these natural-scene video scenarios using a commercial eye-tracker, rather than off-line video coding of fixations from raw camera images. Yoshida and Smith limited their behavioral context to three locations on a small table, in part to study how infants look at objects in near space, and in part because their video coding of gaze did not have the spatial resolution of an eye-tracker. Our head-mounted videos were obtained in a much more diverse set of behavioral contexts, but use of an eye-tracker required us to display these videos to other infants (and adult controls) as they viewed them on a large-screen television display. Thus, our infants were looking at age-matched, infant-selected natural scenes in a laboratory context on a large-screen display, rather than in the natural context where these videos were collected. This hybrid paradigm introduced a mismatch between what the viewing infant saw on the TV display and what that infant could see if they moved their head. This mismatch was not present for the infant who served as

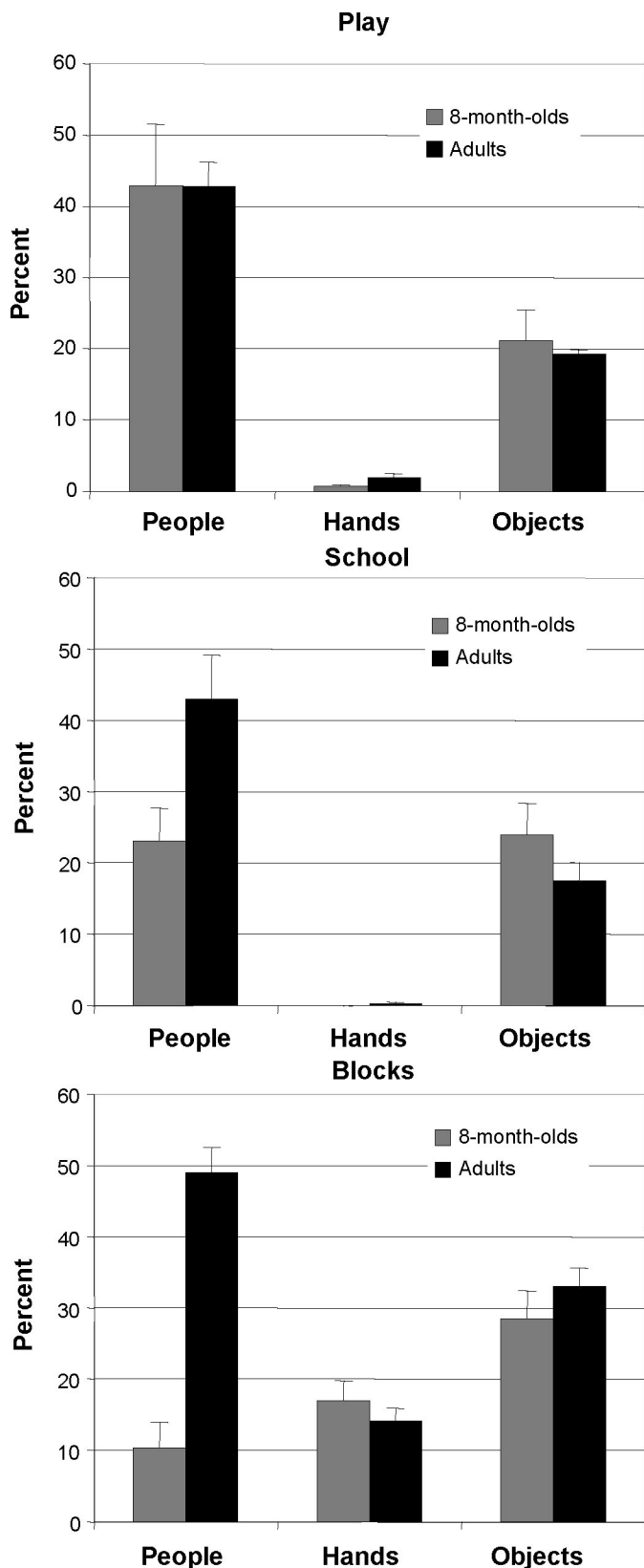


FIGURE 2. Percentage of gaze in the 8-month-old infants (and adult controls) directed to the three categories of interest (Faces, Hands, Objects). Error bars = SEM.

the gatherer of the head-mounted videos (but we did not have access to his changes in gaze).

Despite the foregoing limitations, the Fiser et al.¹⁴ findings revealed two important insights about how infants look at images in dynamic natural scenes. First, both 4- and 8-month-old infants spend very little time looking at images of hands (either their own or those of other people). Although these results appear to conflict with Yoshida and Smith,¹³ it is important to note that there was one exception to this conclusion: the 8-month-old infants (who are experienced reachers) viewing the blocks video, where there were many images of hands grasping objects in near space, spent nearly 20% of their time looking at hands (and less time looking at people). These results strongly suggest that the behavioral context in which infants are confronted with dynamic natural scenes influences how they deploy their gaze (and presumably their attention) to fixate different scene elements.

The second major finding is the close correspondence between scene elements fixated by infants at both 4 and 8 months of age and scene elements fixated by adults under identical circumstances. Although the specific visual patterns available in the video vignettes were not determined by the infant or adult whose fixations were being assessed with the eye-tracker, given identical dynamic video displays the infants and adults deployed their gaze in a similar manner. An intriguing speculation that arises from this finding is that, in the absence of a specific task, images in natural contexts elicit the same kinds of spontaneous attention not because of the “meaning” of these image components, but rather because of their lower-level perceptual properties such as movement, size, or auditory-visual characteristics. The age and task contexts that lead higher-level “meaning” to play a role over and above these physical properties is a topic for future research.

In summary, the results of Fiser et al.¹⁴ and related studies by Yoshida and Smith,¹³ as well as on-going work from several laboratories, provide a preliminary window on how infants deploy their gaze to access information in dynamic natural scenes gathered from the infant’s point-of-view. Future work will be directed to two key questions in infant development. First, what are the statistical properties of natural scenes to which infants deploy their attention? These statistics are undoubtedly a subset of the potential visual information available in the ambient optical array. Because low-level visual features, even if provided innately to human infants by their evolutionary history, must be combined to encode higher-order objects, a powerful learning mechanism is required for infants to represent objects in memory. Fiser and Aslin¹⁶ have provided empirical evidence for how this statistical learning might take place in young infants, and Johnson et al.¹⁷ have confirmed that the manner in which infants fixate the features of dynamic displays enables them to bind together image fragments under conditions of partial object occlusion.

Second, given the utility of multimodal information in natural scenes for adults’ acquisition of the words of a non-native language and for matching those words to meaningful components of dynamic images,¹⁸ the results of Fiser et al.¹⁴ and Yoshida and Smith¹³ suggest that further exploration of how infants perform similar tasks of language acquisition and cognitive development could provide useful insights about how infants learn in natural contexts.

ACKNOWLEDGMENTS

I thank Rachel White for coordinating the data collection and analysis.

This research was supported by grants from NIH (HD-37,082) and the McDonnell Foundation (220020096).

Received October 15, 2008; accepted February 6, 2009.

REFERENCES

1. Aslin RN. What's in a look? *Dev Sci* 2007;10:48–53.
2. von Hofsten C, Rosander K. Development of smooth pursuit tracking in young infants. *Vision Res* 1997;37:1799–810.
3. Gredebäck G, von Hofsten C. Infants' evolving representations of object motion during occlusion: a longitudinal study of 6- to 12-month-old infants. *Infancy* 2004;6:165–84.
4. Hunnius S, Geuze RH. Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: a longitudinal study. *Infancy* 2004;6:231–55.
5. Johnson SP, Slemmer JA, Amso D. Where infants look determines how they see: eye movements and object perception performance in 3-month-olds. *Infancy* 2004;6:185–201.
6. McMurray B, Aslin RN. Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy* 2004;6:203–29.
7. Bronson GW. *The Scanning Patterns of Human Infants: Implications for Visual Learning*. Norwood, NJ: ABLEX Pub. Corp.; 1982.
8. Haith MM. *Rules that Babies Look By: The Organization of Newborn Visual Activity*. Hillsdale, NJ: L. Erlbaum; 1980.
9. Aslin RN, McMurray B. Automated corneal-reflection eye-tracking in infancy: methodological developments and applications to cognition. *Infancy* 2004;6:155–63.
10. Geisler WS. Visual perception and the statistical properties of natural scenes. *Annu Rev Psychol* 2008;59:167–92.
11. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. *Annu Rev Neurosci* 2001;24:1193–216.
12. Corbetta D, Williams J, Snapp-Childs W. Object scanning and its impact on reaching in 6- to 10-month-old infants. In: Paper presented at the annual meeting of the Society for Research in Child Development, Boston, Massachusetts, March 31, 2007.
13. Yoshida H, Smith LB. What's in view for toddlers? Using a head-camera to study visual experience. *Infancy* 2008;13:229–48.
14. Fiser J, Aslin RN, Lathrop A, Rothkopf C, Markant J. An infants' eye view of the world: implications for learning in natural contexts. In: Paper presented at the International Conference on Infant Studies, Kyoto, Japan, June 19–23, 2006.
15. von Hofsten C, Dahlstrom E, Fredriksson Y. 12-month-old infants' perception of attention direction in static video images. *Infancy* 2005;8:217–31.
16. Fiser J, Aslin RN. Statistical learning of new visual feature combinations by infants. *Proc Natl Acad Sci U S A* 2002;99:15822–6.
17. Johnson SP, Davidow J, Hall-Haro C, Frank MC. Development of perceptual completion originates in information acquisition. *Dev Psychol* 2008;44:1214–24.
18. Yu C, Ballard DH, Aslin RN. The role of embodied intention in early lexical acquisition. *Cogn Sci* 2005;29:961–1005.

Richard N. Aslin

*Department of Brain and Cognitive Sciences
Meliora Hall, River Campus
University of Rochester
Rochester, New York 14627
e-mail: aslin@cvs.rochester.edu*