Cognition 152 (2016) 101–107

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

From faces to hands: Changing visual input in the first two years

Caitlin M. Fausey [a,*], Swapnaa Jayaraman [b], Linda B. Smith [b]

[a] Department of Psychology, University of Oregon, Eugene, OR 97403, United States
[b] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, United States

ARTICLE INFO

Article history:
Received 12 March 2015
Revised 7 March 2016
Accepted 8 March 2016

Keywords:
Scene statistics
Head camera
Faces
Hands
Infancy
Egocentric vision

ABSTRACT

Human development takes place in a social context. Two pervasive sources of social information are faces and hands. Here, we provide the first report of the visual frequency of faces and hands in the everyday scenes available to infants. These scenes were collected by having infants wear head cameras during unconstrained everyday activities. Our corpus of 143 hours of infant-perspective scenes, collected from 34 infants aged 1 month to 2 years, was sampled for analysis at 1/5 Hz. The major finding from this corpus is that the faces and hands of social partners are not equally available throughout the first two years of life. Instead, there is an earlier period of dense face input and a later period of dense hand input. At all ages, hands in these scenes were primarily in contact with objects and the spatio-temporal co-occurrence of hands and faces was greater than expected by chance. The orderliness of the shift from faces to hands suggests a principled transition in the contents of visual experiences and is discussed in terms of the role of developmental gates on the timing and statistics of visual experiences.

1. Introduction

The world is characterized by many regularities and human learners are sensitive to these, as evident in extensive research on vision, language, causal reasoning, and social intelligence (e.g., Anderson & Schooler, 1991; Griffiths, Steyvers, & Tenenbaum, 2007; Kahneman, 2011; Simoncelli, 2003). A core theoretical problem concerns how the learner discovers which regularities are relevant for learning and how those regularities segregate into different domains of knowledge (e.g., Aslin & Newport, 2012; Frost, Armstrong, Siegelman, & Christiansen, 2015; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). The relevant data for different domains and tasks could be determined by the regularities in the data themselves (e.g., Colunga & Smith, 2005; Rogers & McClelland, 2004; Tenenbaum et al., 2011) or from internal biases that define distinct domains (e.g., Frost et al., 2015; Spelke, 2000). Here, we present evidence for another way in which data for learning may be bundled into segregated sets, by development itself: visual experiences present different regularities at different developmental points and in so doing development may effectively define distinct datasets of visual information.

Our example case concerns two powerful sources of information for developing infants: human faces and human hands. Faces convey information about the emotional and attentional states of social partners. Hands act on the world; they make things happen. Experimental evidence indicates that infants develop specialized knowledge about the visual properties of faces, enabling the rapid recognition of faces and the meaningful interpretation of facial gestures (see Johnson, 2011). Infants also develop specialized knowledge about seen hand movements, knowledge that supports causal inferences about instrumental actions on objects (e.g., Cannon & Woodward, 2012; Woodward, 2009) and that links gestures and points to reference and word learning (e.g., Carpenter, Nagell, & Tomasello, 1998; Namy & Waxman, 1998; Rader & Zukow-Goldring, 2012). Overall, the evidence suggests a protracted course of development of both kinds of knowledge (see De Heering, Rossion, & Maurer, 2012; Goldin-Meadow & Alibali, 2013) and mature cortical visual representations for faces and hands that are distinct (e.g., Bracci, Ietswaart, Peelen, & Cavina-Pratesi, 2010; Peelen & Downing, 2007).

Although human beings, with their faces and hands, are plentiful in the larger dataset that is human experience, we hypothesize that early visual samples of people are dense with faces (regularities relevant to face processing) and that later samples are dense with hands (regularities relevant for instrumental acts on objects). This hypothesis is suggested by recent discoveries using a new technology, head cameras worn by infants. Although conducted for a variety of purposes by different investigators, all of these studies aimed to capture the visual world of infants and in aggregate they have provided a set of new insights pertinent to the present hypothesis: First, the scenes directly in front of infants are highly selective with respect to the visual information in the larger environment (e.g., Smith, Yu, Yoshida, & Fausey, 2015;

* Corresponding author.
E-mail address: fausey@uoregon.edu (C.M. Fausey).

http://dx.doi.org/10.1016/j.cognition.2016.03.005
0010-0277/

Yu & Smith, 2012). Second, properties of these scenes differ systematically from adult-perspective scenes (e.g., Smith, Yu, & Pereira, 2011), from third-person perspective scenes (e.g., Aslin, 2009; Yoshida & Smith, 2008; Yurovsky, Smith, & Yu, 2013), and are not easily predicted by adult intuitions (e.g., Franchak, Kretch, Soska, & Adolph, 2011; Yurovsky et al., 2013). Third, and most critically, the properties of these scenes are different for children of different ages and developmental abilities (e.g. Frank, Simmons, Yurovsky, & Pusiol, 2013; Kretch, Franchak, & Adolph, 2014; Pereira, James, Jones, & Smith, 2010; Raudies & Gilmore, 2014). Infant-perspective scenes change systematically with development because they depend on the perceiver's body morphology, typical postures and motor skills, abilities, interests, motivations, and caretaking needs. These all change dramatically over the first two years of life, and thus collectively serve as developmental gates to different kinds of visual datasets. In brief, the overarching hypothesis is that development bundles visual experiences into separate datasets for infant learners (see also Adolph & Robinson, 2015; Bertenthal & Campos, 1990; Campos et al., 2000).

One result that has now been reported from studies using head cameras to record everyday at-home experiences is that faces were very frequent in infant-perspective scenes for infants younger than 4 months of age (e.g., Jayaraman, Fausey, & Smith, 2015; Sugden, Mohamed-Ali, & Moulson, 2014). In contrast, laboratory studies of toddler-perspective views found that the faces of social partners were rarely in the toddlers' views but the hands of the partners were frequently in view (e.g., Deák, Krasno, Triesch, Lewis, & Sepeta, 2014; Franchak et al., 2011; Yu & Smith, 2013). Because the contexts of these studies with younger and older infants were different, this developmental pattern – from visual experiences dense with faces to those that were dense with hands – could be the product of the home versus laboratory contexts of the social interactions. Alternatively, the developmental pattern could be broadly characteristic of age-related changes in infant experiences and could indicate a more pervasive temporal segregation of visual datasets about social agents. Here, we provide evidence by using head cameras to collect a large corpus of infant-perspective scenes during unconstrained at-home activities for infants as young as 1 month and as old as 24 months.

Our use of head cameras builds on the prior developmental research using this method (see Smith et al., 2015, for review) as well as growing multi-disciplinary efforts directed toward understanding egocentric vision (e.g., Fathi, Ren, & Rehg, 2011; Pirsiavash & Ramanan, 2012). Considerable progress in understanding adult vision has been made by studying "natural scenes" (e.g., Geisler, 2008; Simoncelli, 2003). However, these scenes are photographs taken by adults and differ systematically in content and visual properties from the scenes sampled by perceivers as they move about in the world (e.g., Pinto, Cox, & DiCarlo, 2008; see also Foulsham, Walker, & Kingstone, 2011). As noted by Braddick and Atkinson (2011), body-worn cameras are especially important for building a developmentally-indexed corpus of scenes that captures how the visual data change as infants' bodies, postures, interests, and activities change with development. Here, we provide evidence for the general importance of a developmentally-indexed description of egocentric scenes by showing that the content of those scenes changes systematically with age for two important classes of social information.

2. Method

2.1. Participants

The participating infants ($n = 34$, 17 male) varied in age from 1 to 24 months (see Fausey, Jayaraman, & Smith, 2015, for additional participant information). Prior work suggests that a shift from scenes dense with the faces of social partners to those dense with their hands could occur with increasing engagement in instrumental acts (e.g., in the period around 5 to 11 months; Rochat, 1992; Soska & Adolph, 2014; Woodward, 1998) or perhaps around one year when infants show increased interest in and imitation of others' instrumental acts (e.g., Fagard & Lockman, 2010; Karasik, Tamis-LeMonda, & Adolph, 2011). Because there is no strong prior basis for making fine-grained predictions about the ages across which a transition from many faces to many hands might occur, we sampled infants continuously within the expected broad age range of this transition – from 1 to 16 months. Because some of the laboratory studies indicating a toddler focus on hands have included older infants (near their second birthday, e.g., Smith et al., 2011; Yu & Smith, 2013), we also included more advanced 24-month-olds to measure the distribution of hands and faces in experiences at the end of infancy. The sample of infants was recruited from a database of families maintained for research purposes that is broadly representative of Monroe County, Indiana: 84% European American, 5% African American, 5% Asian American, 2% Latino, 4% Other) and consisted of predominantly working- and middle-class families.

2.2. Capturing the scenes

Recording the availability of faces and hands in infants' everyday environments requires a method that does not distort the statistics of those daily environments. Accordingly, we used a commercial wearable camera that was easy for parents to use (Looxcie). The diagonal field of view (FOV) was 75 degrees, vertical FOV was 41 degrees, and horizontal FOV was 69 degrees, with a 2″ to infinity depth of focus. The camera recorded at 30 Hz. The battery life of each camera was approximately two continuous hours; parents were given multiple cameras to use and could alternate and charge the cameras to full battery capacity as they needed. Video was stored on the camera until parents had completed their recording and then was transferred to laboratory computers for storage and processing.

The camera was secured to a hat that was custom fit to the infant so that when the hat was securely placed on the infant the lens was above the nose and did not move. Because the central interest of this project was the faces and hands of others (not the infant's own hands), the camera was situated and adjusted to capture the broad view in front of the infant; as a result, the camera could miss the infant's own in-view hands if those hands were below the infant's chin and close (within 2 in.) to the infant's body (see Smith et al., 2015, for a discussion of these issues). Parents were not told that we were interested in faces or hands but were told that we were interested in their infant's everyday activities and to try to record six hours of video when their child was awake. Hours of recording did not always reach the six hour goal and varied across participants ($M = 4.22$, $SD = 1.76$), but did not vary with age ($r(32) = -.12$, *n.s.*). The total number of scenes collected across all infants was 15,507,450; the analyzed scenes were sampled from this larger dataset as described below. Activities and contexts were primarily captured at home (over 80% of all scenes) but also included some out-of-home settings such as stores and group activities. A time-sampling study of the larger population from which these families were selected (Jayaraman, Fausey, & Smith, submitted for publication) indicated similar proportions of (awake) time in the home that changed little over this age range.

2.3. Coding for the presence of faces and hands

To estimate the rate of faces and hands in the collected scenes, scenes were sampled at 1/5 Hz (Fig. 1; see also Fausey et al., 2015,