Consistency and variability in two children's home visual environment across development

Anonymous CogSci submission

Abstract

What do children tend to see in their everyday lives? While an understanding of children's visual environment is central to both theories of language acquisition and visual development, relatively little work has examined the categories and objects that tend to be in the infant view during everyday experience. Here, we analyzed the prevalence of the superordinate categories (e.g., people, animals, food) in the infant view in a longitudinal dataset of egocentric infant visual experience (Sullivan et al., 2020). Overall, we found a surprising amount of consistency in the broad characteristics of children's visual environment across individuals and across developmental time, in contrast to prior work examining the changing nature of the social signals in the infant view. In addition, we analyzed the distribution and identity of the categories that children tended to be touching or interacting with in this data, confirming previous findings that these objects tended to be distributed in a Zipfian manner (Clerkin et al., 2017). Taken together, these findings take a first step towards characterizing infants' changing visual environment, and call for future work to examine the generalizability of these results and to link them to learning outcomes.

Keywords: Object categorization; infant visual experience; head-mounted cameras; longitudinal data.

Introduction

What do children tend to see in their everyday lives? While an understanding of children's visual environment is central to both theories of language acquisition and visual development, we know remarkably little about the categories and objects that tend to be in the infant view, or in what format they are experienced. For example, how often do infants tend to see animals in real-life vs. in storybooks or as toys? How consistent are the broad characteristics of children's visual environments across individuals and across developmental time?

Over the past decade, researchers have begun to answer these questions by documenting the infant egocentric perspective using head-mounted cameras (Franchak, Kretch, Soska, & Adolph, 2011; Yoshida & Smith, 2008), quantifying the degree to which there are substantial shifts in infants' viewpoints that may have downstream developmental consequences.

Indeed, as adults it is hard to intuit how strange this viewpoint can be, and how much it varies across development, transitioning over the first two years of life from close-up views of faces to restricted views of hands manipulating objects (Fausey, Jayaraman, & Smith, 2016; Long, Kachergis, Agrawal, & Frank, 2020), with children's postural developments to a large extent shaping what they see (Sanchez, Long,

Kraus, & Frank, 2018). Most work, however, has focused on documenting the social information that infants and children have access to across early development (Fausey et al., 2016; Sanchez et al., 2018; Yoshida & Smith, 2008).

More recent research has made progress towards understanding what objects tend to be the infant view, starting with annotating the basic-level categories (e.g., spoons, cups) in the view of 8-month-olds during mealtime. This research suggests that a small number of objects are both pervasively present during mealtime and among infants' first-learned words (Clerkin, Hart, Rehg, Yu, & Smith, 2017), suggesting a link between visual experience and early word learning. These findings suggest that a more complete understanding of the visual environment of infants and young children could yield insights about the inputs to both category learning and word learning.

Here, we take a step towards characterizing the visual environment of young children by analyzing the categories of objects (e.g., animals, vehicles, toys, people) present in the infant view in a longitudinal corpus of head-mounted camera data (Sullivan, Mei, Perfors, Wojcik, & Frank, 2020). To start, we choose to annotate these broad categories – rather than the basic-level identities of all objects that were present – as these less detailed annotations allow us to collect data about many more frames from the dataset, and later hone in on objects in categories of interest. We thus collected human annotations of a randomly sampled set of 24,000 frames from two children in the longitudinal dataset, allowing the analysis of the proportion of broad categories in the infant view from 6-32 months of age. To complement this broad sample, we hand-annotated the basic-level categories of the objects that children were interacting with in the subset of frames in which children's hands were visible (3000 frames), providing a closer look into the kinds of objects with which children have the most intensive visual and haptic experiences.

Using these annotations, we conducted four sets of analyses. First, using the category annotations, we examined whether the proportion of animals vs. inanimate objects are relatively equal in the infant view. A long literature has documented that even newborns have a tendency to attend to animate agents (Farroni et al., 2005), and visual cortex dedicates a remarkable amount of space to making fine-grained distinctions among different animals (Konkle & Oliva, 2012). Furthermore, animal words tend to be among children's first-

learned words (CITES Tardif 2008?). However, at present, it is unknown whether (non-human) animals (e.g., cats, dogs, other animals) are prevalent in the infant viewpoint, and whether they tend to be exemplars of real-life animals (e.g., ducks at a park, pet dogs) or mostly illustrations in storybooks or as toy stuffed animals. Second, we examined the co-occurrence between these broad object categories across visual scenes. While some activity contexts (e.g., storytime) are characterized by the conjunction of certain object categories (e.g., between books and people) that jibe with our intuitions, not all activities are intuitive or consistent. We conducted a set of data-driven, exploratory analyses of the co-occurrence statistics of these broad categories to identify other, reliable patterns in how infants experience their visual world. Third, prior work documenting the proportion of faces/hands in the infant view has suggested some developmental changes in how children experience their visual across this same age range–including in this same dataset (Long et al., 2020). Thus, one possibility is that as children learn to crawl and walk on their own (Franchak et al., 2011; Long et al., 2020; Sanchez et al., 2018), some categories that children are likely to interact with (i.e., toys, small objects) could become more prevalent in the child's view across age, whereas other, more stable categories (i.e., furniture) might show relative consistency. On the other hand, the broad characteristics children's visual environments may be relatively stable and determined mostly by the activities that they tend to engage in. We thus examined this hypothesis by exploring the prevalence of each of the broad categories that were annotated across developmental time and across both children.

Finally, we turned to the basic-level annotations of the categories that children tended to interact with, evaluating the degree to which these basic-level objects tend to be distributed in a Zipfian manner and thus the generality of Clerkin et al. (2017)'s results beyond mealtime, and across development.

Method

Dataset

The dataset is described in detail in Sullivan et al. (2020). Three children wore Veho Muvi miniature cameras mounted on a custom camping headlamp harness ("headcams") at least twice weekly, for approximately one hour per recording session. One weekly session was on the same day each week at a roughly constant time of day, while the other(s) were chosen arbitrarily at the participating family's discretion. At the time of the recording, all three children were in single-child households.

Videos captured by the headcam were 640x480 pixels, and a fisheye lens was attached to the camera to increase the field of view to approximately 109 degrees horizontal x 70 degrees vertical. We randomly sampled 24000 frames from videos of the children in the dataset (S, A) over the entire age range.[1]

---

[1] Data from the third child was until recently still being collected and processed, but will be sampled and annotated in the future.

Annotation procedures

Broad categories in view Annotations of the broad categories in the dataset were obtained using the AWS SageMaker Ground Truth service with credits from [blinded]. Participants were instructed to select all of the categories that could be applied to an image; two workers annotated each image, and each category that was annotated for an image was assigned a confidence score (possible range: 0-1, range in dataset: .5-1) provided by SageMaker on the basis of the cross-worker agreement on workers' past annotations. Participants selected whether the following 15 categories were present in the shown image: Animal (real), Animal (toy/drawing), Vehicle (real), Vehicle (toy/drawing), Plant, Clothing, Person, Furniture, Food, Utensil/Dish, Other Small Object, Other Big Object, Book, Other, or Nothing visible.

We included "other small objects" and "other big objects" as categories that participants could use to indicate objects that fell outside of these traditional superordinate categories (i.e. furniture, plant, toy). Additional instructions were provided to specify that 'other big object" refers to an object that is bigger than a chair, and that 'other small object' refers to an object that is small enough to be held with one or two hands (Konkle & Oliva, 2012). Annotators were required to select at least one category before proceeding, and could select as many categories as they chose. Workers selected an average of 2.11 categories per image (median=median(table(d$img))). Individual annotations that had confidence scores below the 25%th percentile were excluded from analyses (although all conclusions hold with or without these low-confidence annotations). Each child's age was calculated in days relative to the date that the videos were filmed and converted to months for plotting purposes.

Objects in the child's reach We further annotated the objects that children were interacting with in a subset of these frames. Specifically, we selected the frames in which a previous set of participants recruited via Amazon Mechanical Turk indicated that one or more hands were present in the frame, and that the visible hands belonged to an adult or a child, and drew bounding boxes around those hands. This resulted in a set of 3,050 images that were then annotated by the authors who established labeling conventions prior to annotation. Annotators noted what object the child was interacting with in frames containing children's hands, using basic level object categories, such as "bird" and "cracker." When children were interacting with drawing or toy versions, these annotations were marked a '-drawing' and '-toy' modifier. If a view was allocentric or there were no child hands in view, these were excluded from analysis (X%). Finally, if there was no object or the object was unclear, these frames were marked accordingly.
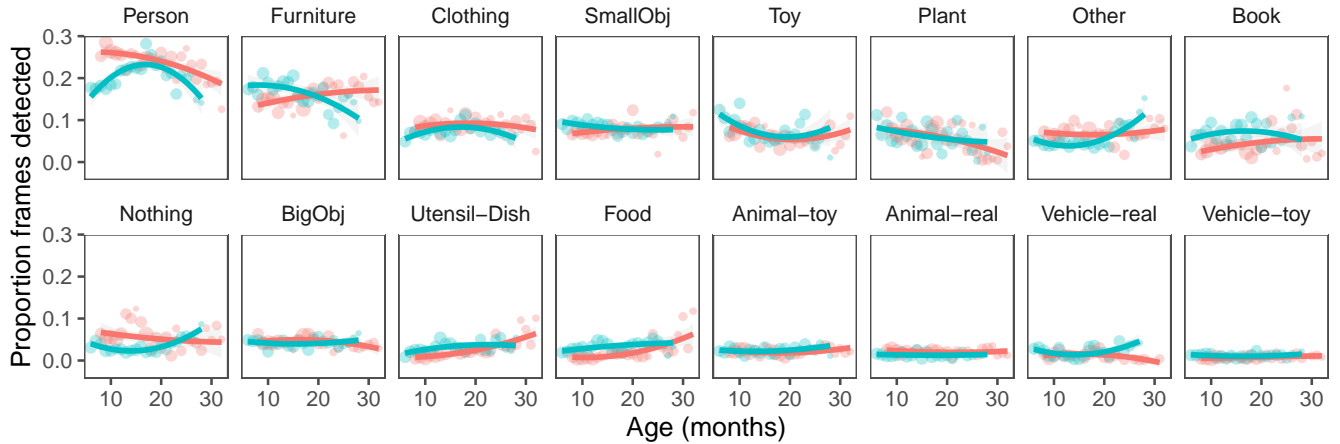
Figure 1: Frequency of categories annotated across the 24K random frames plotted as a function of each child's age (in months).

Results

Which categories are prevalent in the child's view?

First, we examined the overall prevalence of each category in the infant view. Somewhat surprisingly, we found that the prevalence of most of these broad categories was relatively stable both across the two children in the dataset as well as over developmental time. This stands in contrast to prior work on the prevalence of faces and hands in the infant view, suggesting that these broader characteristics of children's visual experience may be more consistent across children and more stable over time.

We next examined the details of these environments. We found that people were by far the most prevalent of these categories: over 20% of the frames that were annotated contained people, far more than any categories (including all kinds of toys combined). In contrast, there were relatively few instances of animals in the infant view–either as toys or their real-life counterparts. Less than 5% of the frames contained any kind of depicted or real animal, and those few frames that did contained depicted vs. real animals in equal proportion. Manual inspection of these frames containing animals revealed that the "real" animals had relatively little variety – they were overwhelmingly frames containing images of household pets (i.e., cats, dogs, and chickens, in the case of A), whereas the animals that were "toys/drawings" depicted a much larger variety of animals, as one might expect. Overall, these results suggest that – at least for these children – people are much more frequent that depictions or real-life versions of animals, indicating that toys and drawings may provide frequent input to their representations of these categories – despite the fact that animal names are often among children's first words (Frank, Braginsky, Yurovsky, & Marchman, n.d.) and often referenced in storybooks.

Far more prevalent than animals, instead, were objects. Views of furniture were the next most common category after "people". However, in older age ranges speaking, "big objects" – including furniture, vehicles, and other big objects – tended to be less frequently in the view of infants than "small"

objects – including toys (of all kinds), food, utensils, books, and other small objects (see Figure 2). This effect was amplified when we conducted this analysis on a subset of the frames in which children's hands were also in view – as a proxy for times when children were interacting with objects. In these frames, small objects tended to be much more prevalent in the frames that we annotated. These data are consistent children grow and become more adept at handling objects on their own, they may tend to be more often in view.

Which categories co-occur in children's visual environment?

Finally, we next examined the degree to which these broad categories appeared together in different frames. While some frequent activity context activities (i.e. mealtime) may have a fixed set of categories that are usually co-present (i.e. utensils/dishes and food), others may not be so easy to intuit. Figure 3 shows the co-occurrence of the broad categories that were annotated. First, it is apparent that people appear quite often, and in scenes with many of the other object categories (in particular, with clothing, books, utensils, and food). Furniture is another promiscuous category, although sensibly not often appearing with plants or real vehicles. Food and utensils are distinctively frequent neighbors, likely identifying mealtime. Books often appear with toy animals and people, but otherwise there is no distinctive profile for bookreading. Plants often appear with "big objects" and real vehicles, likely identifying outdoor contexts, which future research will have to further elaborate.

Which categories do children tend to interact with?

While many different categories may be in the child's view, not all of these objects may be equally important in the child's environment. In particular, it may be that children are more likely to form robust representations of objects that they physically interact with more often, and by extension they may also learn the labels of these objects earlier. In this analysis, we sought to analyze the basic-level identities of the objects
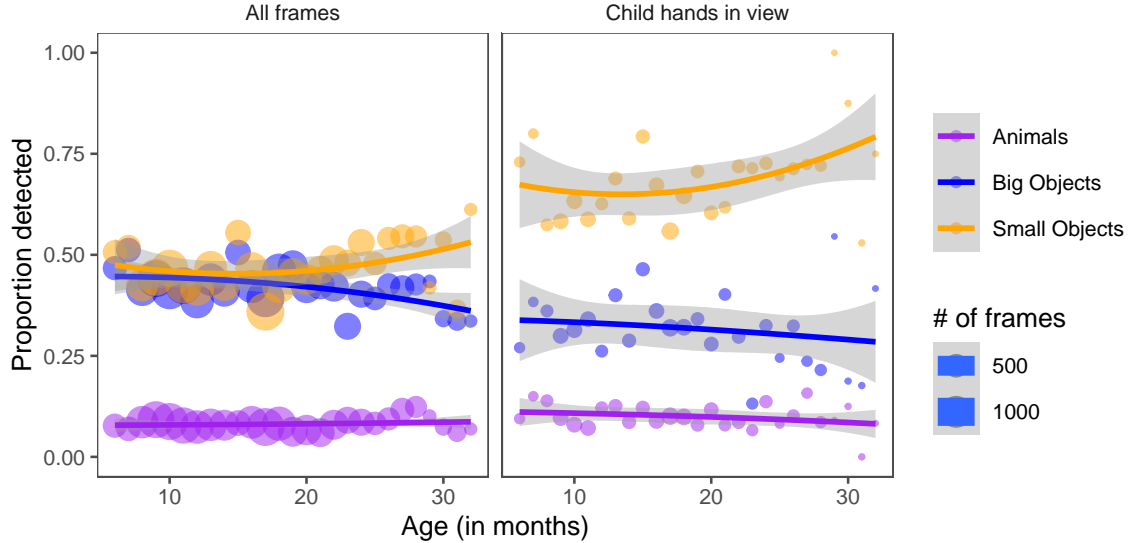
Figure 2: Frequency of animals (including toys) relative to big and small inanimate objects detected in the dataset, both when analyzing all frames that were annotated (left) and the subset of frames where a child's hand was visible in the frame (right).

that children tended to be interacting with in their home environments, and the distributions of those object categories. While some work has found that the objects in view during mealtime tend to have a Zipfian distribution (Clerkin et al., 2017), it is not yet known whether this finding will extend to objects that do not appear during mealtime and that children interact with during a wide range of activities. For example, there may be far fewer objects that are only interacted with a limited number of times vs. seen a limited number of times.

First, we found that the distribution of the objects in view roughly followed a power law distribution (with $\alpha = 1.9$), confirming that the distribution of the objects that children interact with in general (not only during mealtime as measured by Clerkin et al., 2017) is highly skewed – as is the distribution of categories that they tend to see. When we examined which categories were most frequent, we found that books were overwhelmingly the most present object in the views of these two children, comprising over 15% of the objects that children were seen to be manipulating. Generic baby toys (that were unidentifiable to the authors as specific toys) were the next most frequent object category, and children were often seen to be touching or holding on to their caregivers (see top 20 most frequent categories in Figure 4). Additionally, frames in which objects were occluded or unidentifiable had a high frequency of appearance, accounting for around 6% of frames, affirming previous assertions that infant egocentric views are nonintuitive. However, we also saw evidence of some coarse developmental differences, attesting to different play habits across childhood: for example, when these children were older than 18 months they were often seen holding crayons or paper, which was untrue in the views captured during their early infancy (i.e. 6-12 months of age).

General Discussion

Here, we analyzed the consitency and variablity of the categories and the objects in the infant view, examining a sample of random frames taken from a longitudinal dataset of two children (Sullivan et al., 2020). Overall, we found relative consistency in children's visual environment over development, in contrast to prior work on the prevalence of social signals over this same developmental time period. The relative proportions of broad categories of objects (i.e., furniture, toys, animals, people) was relatively consistent among the two individuals here, and across developmental time. People were most frequent, and a non-trivial proportion of frames didn't contain any discernible objects at all. Data-driven analysis of these category co-occurrence revealed stereotypical combinations (i.e. utensils and food, people and clothing).

While the proportion of small objects in view did seem to show something of a U-shaped curve around the first birthday – consistent with the hypothesis that children may see fewer toys as they are learning to crawl and walk (Long et al., 2020) – this remains to be confirmed with future research. Instead, postural developments and cognitive milestones may lead to finer-grained changes in how children view objects or the diversity of exemplars that tend to be view.

However, while people were incredibly frequent in the child's view, animals – either as toys or their real-life versions – were relatively infrequent and occurred in equal proportions. Instead, the child's view was most likely to be dominated by small objects that they were interacting with – such as food, books, or toys. We also found that the proportion of these small objects increased dramatically when we restricted our analysis to frames where the child's hands are in view, suggesting that the statistics of children's visual environment shift substantially when they are acting on the world themselves (i.e., while playing). Further, given that animal words
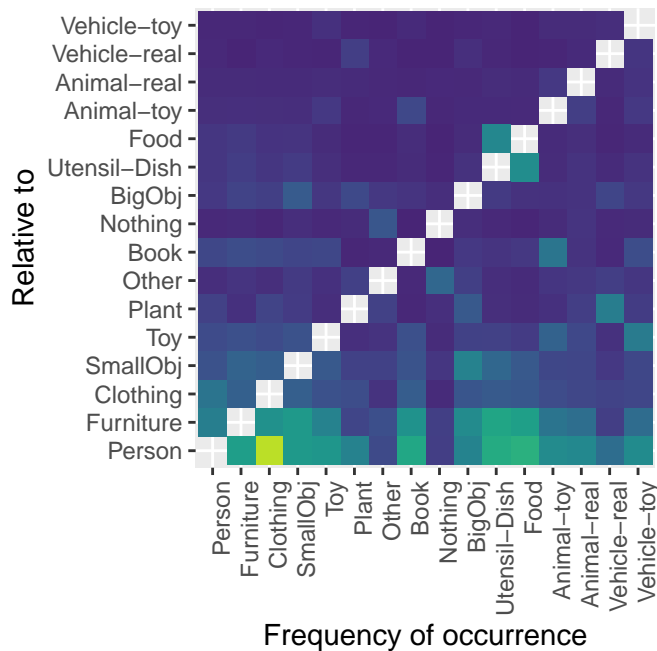
Figure 3: Co-ccurrence between different categories detected in the dataset across all frames. Lighter values indicate higher probabilities of co-occurrence (max=.8, min=0).

Figure 4: Top 20 most frequent categories that children's hands were interacting with in these egocentric videos.

are among children's first-learned words, these results underscore that children's heightened attention to animals (Farroni et al., 2005) likely interacts with frequency of occurrence in the visual field to drive early category learning.

Finally, we examined the categories that children tend to interact with in these egocentric videos, finding that these children were often manipulating books and, more generally, that the distribution of these objects seem to follow a Zipfian distribution, as does word usage in natural language. Indeed, as mealtime has previously been used to characterize the objects in the infant view (Clerkin et al., 2017) and frames with food or utensil and dishes accounted for less than 5% of views in the SAYcam dataset, we were unsure whether this would be the case. However, this analysis suggests that – at least in this sample – that infants may more generally interact with different object categories in a relatively Zipfian manner.

Overall, this work takes a first step in characterizing the categories in the visual environment of children over development, calling for future work to understand the generalizability of these findings beyond the present dataset. While we found relatively consistent results across both age and the two children in the dataset, both of these children are from relatively similar households and cultural contexts. Nonetheless, we predict broad generality of the findings that objects are more frequent than animals and that categories will be relatively stable in their prevalence across age.

More broadly, this work also highlights the need for systematic investigations in how the frequency of categories in the child's view interacts with different attentional biases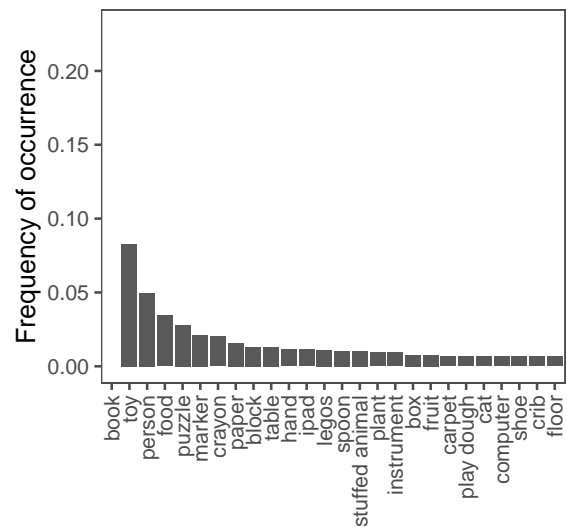, learning mechanisms, and social cues to produce robust representations that support category and early language learning. An understanding of what is – and what is not – learnable solely from frequent exposures will provide constraints on our accounts of the learning mechanisms that allow children to learn so much so quickly.

Acknowledgements

References

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.

Farroni, T., Johnson, M. H., 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*, *102*(47), 17245–17250.

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. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development*, *82*(6), 1738–1750.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (n.d.). *Variability and Consistency in Early Language Learning: The Wordbank Project*. Cambridge, MA: MIT Press. Retrieved from http://wordbank-book.stanford.edu

Konkle, T., & Oliva, A. (2012). A familiar-size stroop effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 561.

Long, B., Kachergis, G., Agrawal, K., & Frank, M. C. (2020). Detecting social information in a dense database of infants' natural visual experience. *Https://Psyarxiv.com/Z7tdg/*.

Sanchez, A., Long, B., Kraus, A. M., & Frank, M. C. (2018). Postural developments modulate children's visual access to social information. In *Proceedings of the 40th annual conference of the cognitive science society*.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2020). SAYCam: A large, longitudinal audiovisual dataset recorded from the infants perspective. *PsyArXiv*. Retrieved from `https://psyarxiv.com/fy8zx/`

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy*, *13*(3), 229–248.