Active Vision: Learning Visual Objects through Egocentric Views of Children and Parents

Sven Bambach[1]
http://homes.soic.indiana.edu/sbambach/
David J. Crandall[1]
http://homes.soic.indiana.edu/djcran/
Linda B. Smith[2]
http://psych.indiana.edu/faculty/smith4.php
Chen Yu[2]
http://psych.indiana.edu/faculty/chenyu.php

[1] School of Informatics and Computing,
Indiana University
[2] Department of Psychological and Brain Sciences,
Indiana University

Abstract

Work in Cognitive Science has shown that infants are amazingly efficient at the complex task of learning to recognize objects in a world full of visual clutter. In fact, many computer vision researchers have drawn analogies between that process and the impressive recent performance of deep learning. This connection raises the exciting potential that better understanding human learning may give us insight into how to improve deep learning, while deep learning may give us new tools to model and understand how humans learn. We consider a first step towards the latter, in particular using deep learning to test the hypothesis that one reason toddlers are able to learn so efficiently is that they create high-quality visual training data for themselves by actively manipulating objects and thus self-selecting ideal object views for visual learning. We test this idea by collecting egocentric video data of free toy play between toddlers and parents, and then train separate Convolutional Neural Networks based on the toddlers' views and the parents' views. Our results show that the egocentric data collected by parents and toddlers have different properties, and that CNNs learn better models using the toddler than the parent views.

1 Introduction

Object recognition is of fundamental importance to humans, whose everyday lives rely on identifying a large variety of visual objects. A vexing question for cognitive scientists is how toddlers are able to learn to identify objects so quickly in a visually noisy and dynamic world where objects are often encountered under seemingly sub-optimal conditions. Many previous studies on early visual object recognition focus on exposing young visual learners to stimuli displayed on a computer screen. While these controlled experimental paradigms are powerful, we also know that these paradigms are very different from young children's everyday learning experiences: active toddlers do not just passively perceive visual information but instead actively manipulate objects, thereby self-selecting many views of the same objects [3].

Meanwhile, in computer vision, many researchers have noted the conceptual connection between this process of infant learning and recent deep learning-based algorithmic techniques that are able to learn surprisingly effective visual models from large, often noisy visual datasets with little prior information. Although the analogy between these two is probably largely conceptual (as opposed to actually occuring through the same mechanical processes), it nevertheless raises the interesting possibility that human learning may give us insight into how to improve deep learning, and that deep learning algorithms could give us new tools for understanding and modeling human learning.

This abstract is a summary of a recent paper [1] in which we use deep learning to test a cognitive science hypothesis: toddlers' active viewing of objects may create high-quality training data for visual object recognition. To do this, we used head-mounted cameras to collect video data in which parents and children were asked to jointly play with a set of toys. We then trained two separate Convolutional Neural Networks (CNNs) [2] with first-person data from the infant view and parent view, and tested them on a separate set of images taken in a well-controlled environment. The results show that the CNNs perform better (in multiple simulation conditions) when trained with the toddlers' data than with the parents', suggesting that toddlers' pattern of interaction with objects is especially well suited to generating better training data. To the best of our knowledge, this is the first study to collect and use egocentric video in everyday contexts and demonstrate a working learning system taking advantage of object view self-selection by active toddlers for visual object recognition.

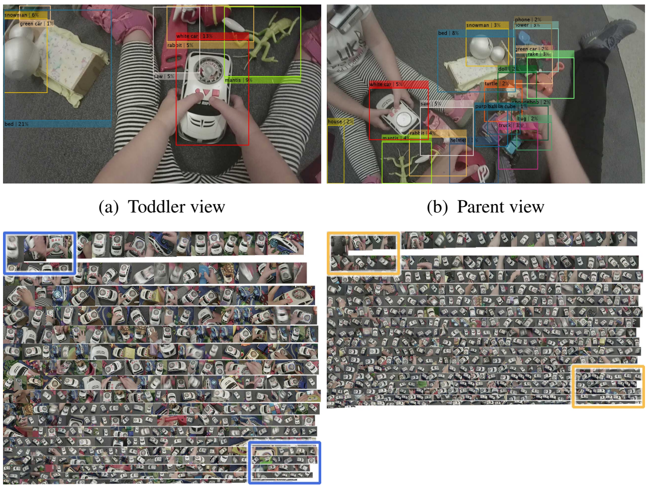
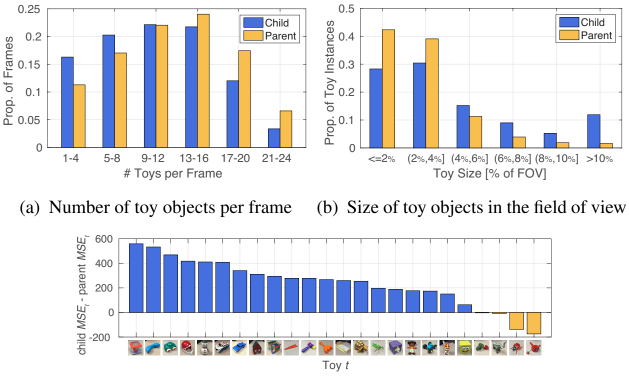

(a) Toddler view (b) Parent view
(c) All toy cars seen by toddlers (d) All toy cars seen by parents

Figure 1: **(a-b)** First-person video frames captured during joint child-parent play from **(a)** toddler and **(b)** parent views. **(c-d)** Views of a toy car as seen by **(c)** toddlers and **(d)** parents, showing the diversity of toddler views. (Objects shown to scale; colored boxes show field of view size.)



(a) Number of toy objects per frame (b) Size of toy objects in the field of view
(c) Differences in visual diversity accross toy instances

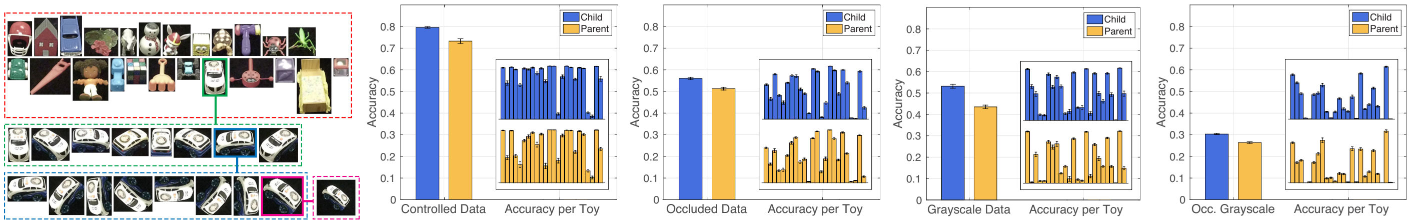
Figure 2: Comparison of objects in the fields of view of toddlers and parents, in terms of **(a)** number and **(b)** size of objects. **(c)** Difference between toddlers and parents in the visual diversity for each of the toys. Positive values indicate higher diversity for toddlers.

2 Data Collection

To test our hypotheses and models, we collected two types of image data, one for training and one for testing. For the training data, we used head-mounted cameras to capture first-person video of toddlers and parents as they jointly played with a set of 24 toys in a naturalistic, unconstrained setting (Figure 1(a-b)). For the test data, we collected a controlled dataset in which we photographed the same objects, but against a clean background and from a systematic set of canonical viewpoints (Figure 3(a)).

Training Data (First-Person)

We invited 10 child-parent pairs (4 boys, 6 girls, mean age 22.6 mos.) into our play room and equipped both with head-mounted cameras. On average, we captured about 8 minutes of video (720×1280px, 30Hz) per subject. After temporally synchronizing parent and child videos, we man-



(a) Controlled toy images (b) Controlled data (c) Occluded data (d) Grayscale data (e) Occluded grayscale data

Figure 3: **(a)** Samples from the test data. Each of 24 toys (red) was photographed from 8 viewpoints (green), and each image was further rotated 8 times (blue) and cropped at lower zoom levels (cyan) to add further variation. **(b-e)** CNN classification accuracies trained with first-person images from toddlers (blue) and parents (orange) tested on controlled image data of the same objects. Bars show standard errors across 10 trained networks.

ually annotated the location (bounding box) of each toy in view of each subject (for 1 frame every 5 sec.). Overall we annotated 9,646 toy instances from the child views, and 11,313 instances from the parent views. Figure 1(c-d) shows all annotated instances of a toy car as an example.

Testing Data (Third-Person)

We also created a separate test set of the same 24 toy objects. The goal of this test data was to have a large variety of clean, systematically-collected, unobstructed third-person views for each toy, to serve as a view-independent and therefore objective way to evaluate the performance of visual object recognition. We captured 8 photos from each toy, one from each 45° rotation around its vertical axis. Sample images from every toy are shown inside the red box in Figure 3(a). To add further variation, we rotated (8×) and rescaled (2×) each photo, resulting in a test set of $8 \times 8 \times 2 = 128$ images for each toy and 3,072 images total.

3 Study 1: Quantifying and Comparing Object Properties in Egocentric Views

During joint play, toddlers and parents generate many instances of visual objects within their self-selected fields of view. Our first study quantified and compared properties of object appearance across the two views.

Number and Sizes of Objects in View

Figure 2(a) presents histograms showing the number of objects that appear simultaneously in the field of view. Toddlers have a larger fraction of frames (16.3%) with only 1-4 objects compared to parents (11.3%). Conversely, parents are more likely to have most objects in view at once. We also investigated object size within the fields of view. Figure 2(b) shows that toddlers are much more likely to have toys prominently in view (object bounding box >10% of field of view) while parents are much more likely to see small objects (≤2% FOV).

Variation in Visual Object Appearance

Finally, we aim to quantify the visual diversity across toy instances among the two different views. We represent each instance into a fixed-length 300-d vector, and then compute the pixelwise mean squared error (MSE) distance between all pairs of instances for each object and subject. As shown in Figure 2(c), the mean MSE for each toy between toddler and parent views shows greater diversity for toddlers in 20 of the 24 toys.

4 Study 2: Object Recognition with Deep Models

We investigate how well a CNN trained with real-world toy instances (as captured during our joint play experiments) recognizes the same 24 objects in a separate, controlled testing environment. We do not claim that a CNN actually emulates visual object learning in toddlers, but are instead interested in CNNs as proxies for ideal learners. Given the differences between toddler and parent views summarized in Study 1, we hypothesize that the toddler data captures a richer representation of each object, leading to better classification performance on the controlled test data. All experiments use AlexNet [2], fine-tuned from pre-trained weights.

Simulation 1: CNNs Learn from the Training Data

Before we experiment with controlled test images, we are interested in evaluating the learnability of the first-person data. Training two networks (one with the toddler data and one with the parent data) with a 6-fold cross validation split yielded an average test accuracy of 89.9% for the toddler views and 93.1% for the parent views (4.2% random baseline).

Simulation 2: Using Testing Data from a Third-Person View

We investigate how well learned concepts from the first-person training data transfer to the clean test data. We trained 10 CNNs on the toddler training data and a separate set of 10 CNNs on the adult training data, and then tested both with the same controlled tes data described above (3,072 images). As shown in Figure 3(b), the toddler networks achieve higher accuracy by 6.3 percentage points. Figure 3(b) also compares the distribution of mean accuracies for each object, showing that the child networks outperform for 16 out of the 24 toys, indicating that the differences in overall accuracy are not caused by a minority of classes.

Simulation 3: Recognizing Occluded Objects

Another interesting question is how well the toddler and parent views allow the trained networks to deal with occlusion. To test this, we added occlusion to each test image by systematically blocking different image quadrants (resulting in 14 different occlusions per image). Figure 3(c) shows results from the 2×10 networks of Simulation 2 on the occluded data, showing that toddler networks retain better mean accuracy.

Simulation 4: The Effect of Color Information

The performance difference on the controlled images might be because one set of networks simply relies more on color information. To examine this idea, we repeated all experiments with grayscale images. First, we investigate if lack of color increases the difficulty to learn from the two datasets. The average test accuracy across splits decreased to 76.9% for toddler and to 83.1% for parent networks, a realitively small drop. Next, we repeated Simulations 2 and 3 and train two sets of 10 networks, one with grayscale toddler images and the other with grayscale parent images, and test on grayscale testing set images. Figure 3(d-e) shows that toddler again outperform parent networks, both for the non-occluded and the occluded testing data.

5 Summary and Future Work

We collected first-person video data of free toy play between toddler-parent dyads, and used it to train state-of-the-art machine learning models (CNNs). Our results showed that (1) CNNs were able to learn object models of the toys in this first-person data; (2) these models could generalize and recognize the same toys in a different context with different viewpoints; and (3) the visual data collected by toddlers is of particularly high quality as models trained with toddler data consistently outperformed those trained with parent data in multiple simulation conditions. In addition, we believe this to be a first step towards the exciting paradigm of using deep learning techniques to test hypotheses in cognitive science. Our future work will focus on further understanding the factors that may account for the observed performance differences.

Acknowledgements. This work was supported by NSF (CAREER IIS-1253549, CNS-0521433, BCS-15233982), NIH (R01 HD074601, R21 EY017843), the IU IUCRG and FRSP programs, a Paul Purdom Fellowship, and used compute facilities from NVidia, the Lilly Endowment through its support of the IU PTI, and the Indiana METACyt Initiative.

[1] S. Bambach, L. Smith, D. Crandall, and C. Yu. Active viewing in toddlers facilitates visual object learning: an egocentric vision approach. In *CogSci*, 2016.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[3] C. Yu, L. Smith, H. Shen, A. Pereira, and T. Smith. Active information selection: Visual attention through the hands. *IEEE Trans Auton Ment Dev*, 1(2):141–151, 2009.