Active Viewing in Toddlers Facilitates Visual Object Learning: An Egocentric Vision Approach

Sven Bambach, David J. Crandall, Linda B. Smith†, Chen Yu†
{sbambach, djcran, smith4, chenyu}@indiana.edu
School of Informatics and Computing, Indiana University
†Department of Psychological and Brain Sciences, Indiana University
Bloomington, IN, 47405 USA

Abstract

Early visual object recognition in a world full of cluttered visual information is a complicated task at which toddlers are incredibly efficient. In their everyday lives, toddlers constantly create learning experiences by actively manipulating objects and thus self-selecting object views for visual learning. The work in this paper is based on the hypothesis that active viewing and exploration of toddlers actually creates high-quality training data for object recognition. We tested this idea by collecting egocentric video data of free toy play between toddler-parent dyads, and used it to train state-of-the-art machine learning models (Convolutional Neural Networks, or CNNs). Our results show that the data collected by parents and toddlers have different visual properties and that CNNs can take advantage of these differences to learn toddler-based object models that outperform their parent counterparts in a series of controlled simulations.

Keywords: vision, visual object learning, convolutional neural networks, head-mounted cameras

Introduction

Visual object recognition is of fundamental importance to humans and most animals, whose everyday lives rely on identifying a large variety of visual objects. Because of its importance, even human infants possess sophisticated perceptual and learning processes to form categorical representations of visual stimuli (Quinn & Eimas, 1996). Even as toddlers, they already seem to be able to easily recognize everyday objects. A vexing question for cognitive scientists is how young learners achieve this ability in a visually noisy and dynamic world where objects are often encountered under seemingly suboptimal conditions, including in unusual orientations, varying lighting conditions, or partial occlusions (Johnson & Aslin, 1995; Casasola, Cohen, & Chiarello, 2003). Despite recent progress in object recognition in the computer vision community (Krizhevsky, Sutskever, & Hinton, 2012), even the most powerful computational algorithms trained with large amounts of data are arguably not yet able to learn as efficiently as toddlers do.

Many previous studies on early visual object recognition focus on examining exactly what visual information is extracted from the retinal image to construct invariant descriptors of objects. For this purpose, many experimental paradigms have been invented that repeatedly expose young visual learners to stimuli displayed on a computer screen (familiarization phase), and then measure looking times towards familiar and novel stimuli (test phase). These paradigms are powerful, allowing us to examine, in a rigorously controlled way, which visual features are extracted, how they are stored in memory, and how they are activated to recognize new instances. However, we also know that these experimental paradigms are very different from young children's everyday learning experiences: active toddlers do not just passively perceive visual information but instead generate manual actions to objects, thereby creating self-selection of object views (Yu et al., 2009). Indeed, recent work shows that infants who have more experience in manual object exploration have more robust expectations about unseen views of novel objects (Soska, Adolph, & Johnson, 2010). Another study using head-mounted cameras to record toddlers fields of view found a preference towards planar views of objects: toddlers dwelled longer on these views while manually exploring held 3-d objects than would be expected if the objects were rotated randomly. This bias substantially increased between the ages of 12-36 months (Pereira et al., 2010).

Figure 1: All instances of a toy as seen by cameras mounted on heads of toddlers (left) and parents (right) during joint play between 10 toddler-parent dyads, showing greater diversity in toddler views. Instances are shown to scale and colored boxes depict the field of view size.

Visual object recognition depends on the specific views of objects experienced by the learner. In everyday contexts such as toy play, toddlers actively create many different views of the same object. In light of this, the overall hypothesis in the present study is that active viewing may create high-quality training data for visual object recognition. To test this hypothesis, we used head-mounted cameras to collect first-person video data from a naturalistic environment in which

parents and children were asked to jointly play with a set of toy objects. Figure 1 shows examples of different views of the same toy car, collected from toddlers' view (left) and parents' view (right) during the same play sessions. Clearly, toddlers created more diverse views in terms of relative object size, orientation, and occlusion compared to their parents. In the present study, we first quantify the differences in visual properties of objects between toddler and parent views, finding a higher variation among visual instances for the child. A learning system could take advantage of such variation by building more generalizable representations for recognizing unseen instances, thus better facilitating visual object recognition.

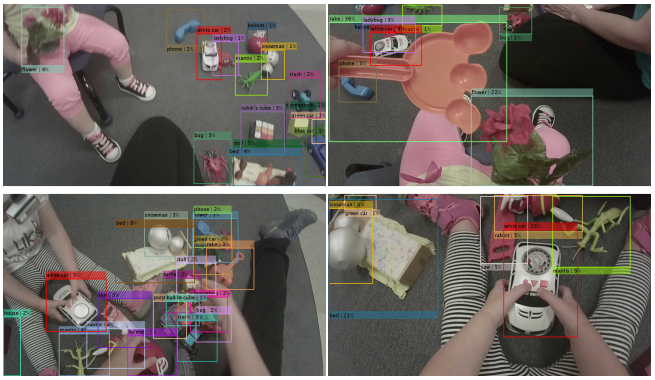
To test this idea, the main focus of the study was to train machine learning models based on Convolutional Neural Networks (CNNs), which are currently considered the most powerful visual learning models in the computer vision community (Krizhevsky et al., 2012), with data from the two different views, and to examine the extent to which these models take advantage of visual information created and perceived by toddlers. The results show that the CNNs perform better on object recognition in multiple simulation conditions when trained with the toddlers' data than with the parents' 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.

Data Collection

To test our hypotheses and models, we collected two types of image data, one for training our CNN models and one for testing them. 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 toys in a naturalistic, unconstrained setting. For the test data, we collected a controlled dataset in which we photographed the same set of objects, but against a clean background and from a systematic set of canonical viewpoints. We now describe each dataset in detail.

Training Data

The training data was collected in a small (~$15m^2$) room with a soft carpet to facilitate sitting on the floor. This "toy room" had an adult-sized chair and a toddler-sized chair, but otherwise no large objects or other distractions. Figure 2 gives an impression of the setting. Our data was collected from 10 child-parent dyads (9 mothers and 1 father; 6 girls and 4 boys, mean child age 22.6 months and SD = 2.1 months). Before entering the toy room, both the parent and toddler were equipped with a head-mounted camera. Both cameras were small (4.8cm×4.8cm×1.5cm), lightweight (22g) *Looxcie 3* cameras with a 100° diagonal field of view. Video data was recorded directly onto a microSD card. Cameras were attached (with velcro) to an adjustable headband to ensure a tight but comfortable fit on the center of each participant's forehead. Next, we randomly arranged 24 toys (Figure 3)



(a) Parent view (b) Child view

Figure 2: First-person video examples that were captured during joint child-parent play in our toy room, contrasting parent view (a) and child view (b). Each row shows one dyad. Also shown are bounding boxes and toy sizes (as % of FOV).

Figure 3: The 24 toys that were used in all of our experiments.

in the center of the floor and encouraged the dyad to play together as they pleased. Once they were engaged with the toys, we left the room and did not give further instructions. Most parents sat on the floor, while toddlers switched between sitting on the floor and walking or crawling around to pick up new toys. Two toddlers briefly sat in the small chair.

For each child-parent dyad, we extracted the greater of 10 minutes of video or the longest period of continuous toy play (uninterrupted by the child taking off the camera or losing interest), yielding at least 3 minutes 35 seconds and an average of 7 minutes 58 seconds of video per dyad. All videos were captured with a resolution of 720×1280 pixels at 30 frames per second, and each video pair between toddler and parent was synchronized.

The location of each of the 24 toy objects within the captured first-person video data was manually annotated. To do this, we subsampled the video stream at one frame every five seconds, and then manually drew bounding boxes around each toy in each frame. Figure 2 shows four annotated example frames. Since toys were often occluded by other objects or truncated at the frame boundaries, we used the following guideline: if only part of a toy was visible, we drew a box around the part if it was visually identifiable as the right toy; if multiple parts of an identifiable toy were visible, we drew a box that included all visible parts of the toy.

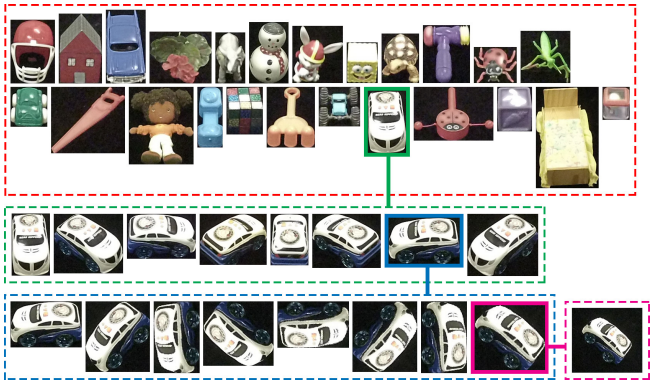
Overall, we captured 9,646 toy instances from the toddler views and 11,313 from the parent views, for an average of 401

instances per class across all toddlers and 471 across parents. There were no large outliers for any of the toys in terms of appearance frequency; the least frequent toy appeared 307 and 341 times for toddlers and parents, respectively, while the most frequent appeared 559 and 600 times.

Testing Data

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 again used the *Looxcie 3* camera but this time captured static photos (at the same resolution as the video). The toy room floor was covered in a black cloth to obscure background clutter, and the camera was mounted onto a tripod, pointing towards the floor at a 45° angle. Each toy was put on the floor at a distance of 50cm from the tripod center. The height of the camera was 45cm, creating a distance from lens to toy center of around 67cm, which approximately centered each toy in the camera frame.

We captured 8 photos from each toy, one from each 45° angle rotation around its vertical axis. Sample images from every toy are shown inside the red box in Figure 4, while the green box shows one of the toys from all 8 viewpoints. To create even more diversity, we additionally rotated each image around the optical center of the camera in 45° increments (blue box in Figure 4). Images were then cropped to a bounding box around the object. To add scale variation, we padded and rescaled images to simulate zooming out by a factor of two (cyan box in Figure 4). In total, our test data consisted of $8 \times 8 \times 2 = 128$ images for each toy and 3,072 images total.



(a) Controlled toy images

(b) Toy image with added occlusion

Figure 4: (a) Samples from the controlled test data. Each of the 24 toys (red) was photographed from 8 viewpoints (green), and each resulting image was further rotated 8 times (blue). To add scale variation, each image was also cropped at a lower zoom level (cyan). (b) Sample test images with synthetic occlusion.

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.

Object Appearance in the Field of View

We begin by studying how many toys are present within the field of view, as well as the perceptual size of those toys.

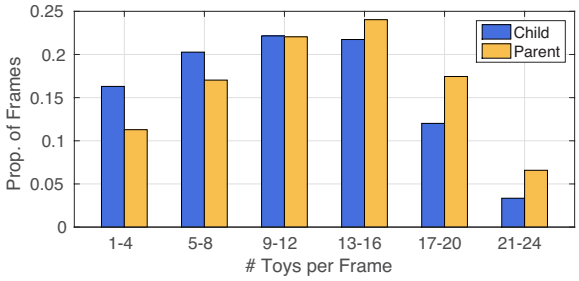
Number of Visual Objects Figure 5(a) presents histograms showing the number of visual objects that appear simultaneously in the field of view. Toddlers have a larger fraction of frames (16.3%) with only 1-4 objects compared to parents (11.3%). Conversely, parents are more likely to have most objects in view at once, with 24.0% of parent frames containing more than 17 objects versus only 15.4% of toddler frames.

Visual Object Sizes Next, we investigate the size of visual objects within the fields of view. We approximate the actual size of an object with the area of its bounding box, and measure the fraction of the field of view that is occupied by this box. Figure 5(b) shows that 42.3% of object instances occupy $\leq 2\%$ of parents' field of view, while only 3.5% of all objects appear dominantly in view ($> 8\%$ of FOV). On the other hand, toddlers exhibit a more spread-out distribution: only 28.3% of objects appear small ($\leq 2\%$ FOV) while 11.9% of objects occupy more than 10% of the view. For perspective, the white car (red bounding box) in the bottom row of Figure 2 occupies 5% of the parent view (a) and 13% in the child view (b). These results are consistent with findings from previous head-camera studies (Yu et al., 2009).

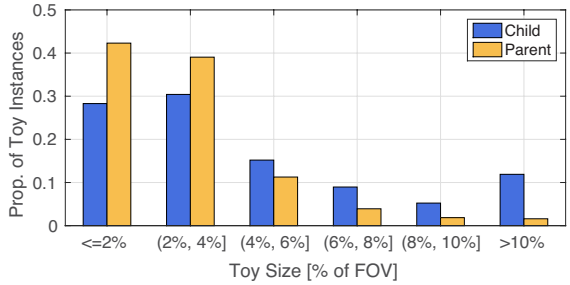
Variation in Visual Object Appearance

Finally, we aim to quantify the visual diversity across the views. We resize each toy image to a canonical size (10×10 pixels), and, for each subject, compute the pixelwise mean squared error (*MSE*) between all instances of the same object. In other words, we take each $10 \times 10 \times 3$ color image, represent it as a 300-dimensional vector, and then compute the mean *MSE* distance between all possible pairs of instances for each object and subject. This score should be low for a subject who sees many visually similar instances of an object, and high for one that sees much variation. We compute scores on a per subject basis to control for inter-subject differences in object appearance.

Figure 6 compares the visual diversity between views generated by toddlers and parents. For each object, we subtract the average MSE score across all toddlers from the corresponding score across parents, so that positive values indicate higher diversity in toddler views while negative values indicate higher diversity in parent views. Using this metric, we



(a) Number of toy objects visible per frame



(b) Size of toy objects within the field of view

Figure 5: Comparison of how objects appear in the fields of view of toddlers and parents, in terms of (a) number of objects appearing simultaneously, and (b) size of objects in view.

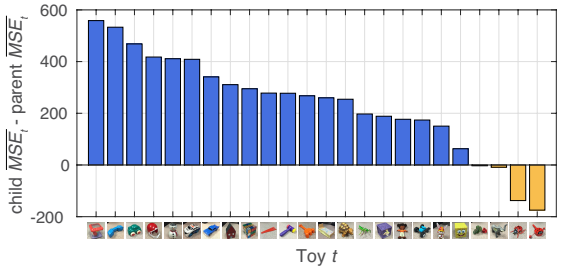


Figure 6: Difference between toddlers and parents in the visual diversity for each of the toys. Positive values indicate higher diversity for toddlers. See text for details.

find that toddlers on average generated more diverse views for 20 out of the 24 toys. We also experimented with other image representations such as grayscale image vectors and GIST features (Oliva & Torralba, 2001) (which capture shape and texture information), and found similar tendencies, with 21 and 15 toys being considered as more diverse respectively.

Discussion

Our study showed large differences in the views of objects that toddlers and adults interact with, even when jointly interacting with them at the same time. While parents are more likely to have "overview" views including many objects, toddlers are more likely to pick out and inspect single objects up close, resulting in fewer, larger objects in view. Additionally, this object selection process seems to create more diverse viewpoints for the children than for the parents.

Study 2: Visual Object Recognition Based on Deep Learning Models

Deep learning using Convolutional Neural Networks (CNNs) has recently shown impressive success in computer vision, improving the state-of-the-art for visual recognition by a large margin (Krizhevsky et al., 2012). We investigate how well a CNN trained with real-world toy instances (as captured during our joint play experiments) recognizes the same 24 visual objects in a separate, controlled testing environment. We do not claim that a CNN constitutes the perfect model to emulate visual object learning in toddlers (or humans in general). Instead, we are interested in CNNs as ideal learners. We assume that the network will learn to use whichever visual features are sufficient to distinguish the 24 objects. Given the differences in captured visual object views of parents and toddlers, two separate networks, one trained with toddler data and the other trained with parent data, might learn different (better) strategies. More directly, we hypothesize that the toddler data captures a richer representation of each object, leading to better classification performance on the controlled test data.

We first describe CNN implementation details and verify that the networks can learn visual object appearance based on first-person data. We then test the networks in a series of experiments based on our controlled views of each object.

Convolutional Neural Networks

CNNs are a special type of multi-layer, feed-forward neural networks, consisting of multiple convolutional layers followed by multiple fully-connected layers. Neurons between the convolutional layers are connected sparsely and with shared weights, effectively implementing a set of filters. Filter responses are passed to a non-linear activation function as well as a local pooling function before serving as input to the next layer. Intuitively, the convolutional layers learn filters (from low-level in early layers to high-level in deep layers) that extract image features, while the fully-connected layers act as a classifier. Please see (Krizhevsky et al., 2012) for more details on CNNs.

Implementation For all our experiments, we used the well established AlexNet CNN architecture (Krizhevsky et al., 2012), consisting of five convolutional layers and three fully-connected layers. The input layer of the network has a fixed size of 224×224×3 neurons, which means the network expects input images to be resized to 224×224 pixels. Instead of training the network from scratch, we follow the common protocol of beginning with a network pre-trained on the ImageNet dataset (Deng et al., 2009), which consists of millions of images. We adjust the final layer to have 24 neurons to accommodate our 24-way object classification task, and then use the parameters learned from ImageNet as initialization for training on our data. Each network is trained via back propagation with a softmax loss function, using batch-wise stochastic gradient descent with a learning rate of 0.001, momentum of 0.9, and batch size of 256 images.

Simulation 1: CNNs Learn from the Training Data

Before we experiment with controlled test images, we need to ensure that CNNs are indeed able to learn visual object models from our first-person data. As detailed in Data Collection, our training data includes 24 different visual objects with 11,313 parent views and 9,646 toddler views. Figure 1 illustrates how some of these images look. We now consider these images as two different datasets (toddler data and parent data), and perform a 6-fold cross validation split on both. This means that for each dataset we train six different CNN networks that each use a randomly-selected one-sixth of the data for testing and the remaining five-sixths for training.

For both parent and toddler data, we found that the back-propagation converged after about 10 epochs, i.e. after observing the training data around 10 times. The average test accuracy across splits was 89.9% for the toddler views and 93.1% for the parent views. To put this into perspective, random guessing achieves $1/24 \approx 4.2\%$, while guessing the majority class achieves 5.8% for toddlers and 5.3% for parents.

We investigated failure cases by inspecting the ten images that each network was least confident about (i.e., having the lowest predicted probability of the true class). Both parent and toddler networks showed similar patterns, where most mistakes were caused by either two or more objects overlapping each other, strong motion blur, or a combination of both. From this we conclude that the failures are reasonable and that CNNs are indeed able to learn from the first person data.

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

Now we investigate how well learned concepts from the first-person training data transfer to the clean testing data. We trained a CNN on the first-person toddler training data, and a separate CNN on the first-person adult training data, and then tested both with the same controlled testing data described above (3,072 images). To avoid learning frequency biases, since some objects have more training instances than others, we uniformly sampled the training data from each class. Given that CNN training is non-deterministic due to random training subset sampling and parameter initialization, we repeated the full training and test procedure over 10 independent trials. We stopped training each network after convergence (around 12 epochs).

As shown in Figure 7(a), the networks trained on toddler data achieve higher recognition accuracy by 6.3 percentage points compared to the networks trained on parent data, reaching 79.6% as opposed to 73.3%. Figure 7(a) also compares the distribution of mean accuracies for each object. Overall, the child networks achieve the same or better results for 16 out of the 24 toys, indicating that the differences in overall accuracy are not caused by some 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.

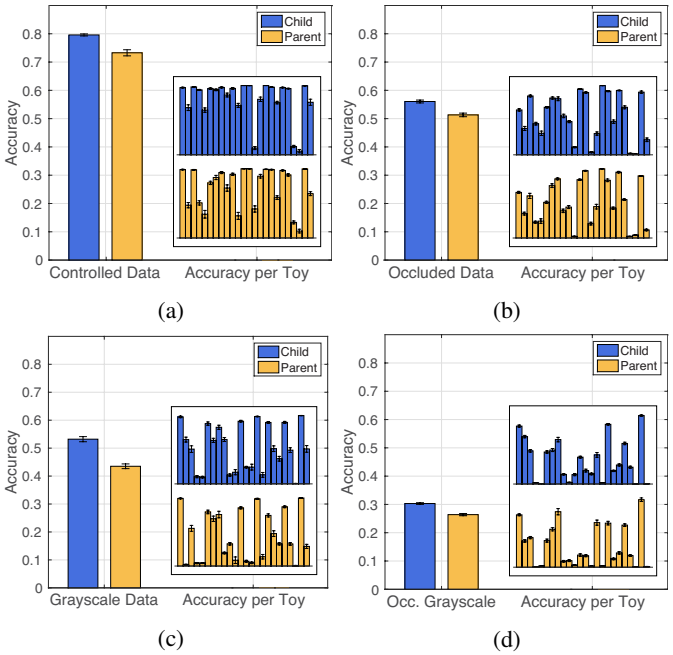


Figure 7: Classification accuracies of CNNs trained with first-person image data from toddlers (blue) and parents (orange), when tested on controlled image data of the same objects. Bars show standard errors across 10 trained networks.

To test this, we systematically added occlusion to each testing image, by splitting the image into quadrants and then occluding each possible combination of one to three quadrants with gray boxes. This resulted in 14 occlusions per image, as shown in Figure 4(b). The occluded testing data thus consisted of $14 \times 3{,}072 = 43{,}008$ images.

Figure 7(b) presents results of testing the same 2×10 networks from Simulation 2 on the occluded data. Toddler networks retain better overall mean accuracy compared to the parent networks (56.1% vs. 51.3%). The relative performances when compared to the non-occluded data drop by ~30% for both parent and toddler networks, which suggests that both are affected similarly by occlusion.

Simulation 4: The Effect of Color Information

The differences in performance on the controlled object images might be because one set of networks relies more on color information to learn object models based on the first-person training data. To examine this idea, we repeat all experiments with grayscale images.

Cross-validation First, we investigate if the absence of color information increases the difficulty to learn from the two first-person datasets. We repeat the same 6-fold cross validation experiments as described above in Simulation 1, but this time train the networks with grayscale images. The average test accuracy across splits decreased to 76.9% (from 89.9%) for the toddler networks and to 83.1% (from 93.1%) for the parent networks. Thus, the absence of color accounts for a rather small drop in learnability of the data.

Controlled Testing We repeat our Simulation 2 and 3 experiments and train two sets of 10 networks, one with the grayscale toddler images and the other with grayscale parent images, and test them on grayscale versions of the testing dataset images. Figure 7(c-d) summarizes the results. The toddler networks again significantly outperform parent networks in terms of overall mean accuracy, both for the non-occluded and the occluded testing data, with mean accuracies of 52.3% over 43.5%, and 30.3% over 26.4%, respectively. The relative performance drops compared to the color experiments are in favor of the toddler networks on the non-occluded data (-34% vs. -40%), but slightly in favor of the parent networks on the occluded data (-51% vs. -54%).

Discussion

Our results suggest that naturalistic first-person images of unconstrained toy play can be used to train CNN-based object models that generalize to recognizing the same objects in a different context. It appears that toddlers generate high quality object views that facilitate learning, as networks trained on toddler data consistently outperformed parent networks.

Summary and General Discussion

In the present paper, we collected egocentric 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 indeed able to learn object models of the toys in this first-person data and (2) that these models could generalize and recognize the same toys in a different context with different viewpoints. Finally, we showed that (3) the visual data collected by toddlers seems to be of particularly high quality as models trained with toddler data consistently outperformed those trained with parent data in multiple simulation conditions.

In the real world, toddlers spend hours every day playing with toys, actively manipulating objects and, as a result, create learning experiences by self-selecting object views for visual learning. It may not sound surprising that better data leads to better learning. Nonetheless, if active viewing by toddlers creates high-quality training data for object recognition, as evident in the present study, then the potential long-term impact of day-in and day-out object play that repeatedly and incrementally provides such data may be the key to why toddlers are incredibly efficient in visual object learning. If so, future research should focus not only on studying particular learning mechanisms in experimental tasks but also on how high-quality data is created by learners themselves. A better understanding of human learning systems would require a better understanding of both learning algorithms and the data fed into them. This paper represents a first step towards this direction by linking high-density video data collected in naturalistic contexts with state-of-the-art machine learning.

Our future work will focus on further understanding the factors that may account for the observed performance differences (e.g. different spatial scales and resolutions of object images versus innate differences in the viewpoints). Another future direction within our framework is to further examine how CNNs take advantage of visual instances captured from diverse viewpoints. Our results here seem to support view-based theories of visual object recognition, stating that human learners store viewpoint-dependent surface and/or shape information (Tarr & Vuong, 2002), and recognize objects by their similarity to stored views (Bülthoff & Edelman, 1992). By diagnosing the network directly and visualizing learned high-level filters and the image regions they are likely to fire on (Zeiler & Fergus, 2014), we may be able to provide more direct evidence on the mechanisms of object recognition.

Acknowledgments

This work was supported by the National Science Foundation (CAREER IIS-1253549, CNS-0521433, BCS-15233982), the National Institutes of Health (R01 HD074601, R21 EY017843), and the IU OVPR through the IUCRG and FRSP programs. It used compute facilities provided by NVidia, the Lilly Endowment through its support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. SB was supported by a Paul Purdom Fellowship. We thank Sam Dong, Steven Elmlinger, Seth Foster, and Charlene Tay for helping with the data collection.

References

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *PNAS*, *89*(1), 60–64.

Casasola, M., Cohen, L. B., & Chiarello, E. (2003). Six-month-old infants' categorization of containment spatial relations. *Child development*, *74*(3), 679–693.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*.

Johnson, S. P., & Aslin, R. N. (1995). Perception of object unity in 2-month-old infants. *Dev Psychol*, *31*(5), 739.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Proc. NIPS* (pp. 1097–1105).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, *42*(3), 145–175.

Pereira, A., James, K., Jones, S., & Smith, L. (2010). Early biases and developmental changes in self-generated object views. *Journal of vision*, *10*(11), 22.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *J Exp Child Psychology*, *63*(1), 189–211.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. *Dev Psychol*, *46*(1), 129.

Tarr, M. J., & Vuong, Q. C. (2002). Visual object recognition. In *Stevens' handbook of experimental psychology*. Wiley.

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

Zeiler, M., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proc. ECCV* (pp. 818–833).