Wang, X., Eliott, F., Ainooson, J., Palmer, J., and Kunda, M. (2017). An object is worth six thousand pictures: The egocentric, manual, multi-image (EMMI) dataset. To appear, International Conference on Computer Vision (ICCV) Workshop on Egocentric Perception, Interaction, and Computing (EPIC), Venice, Italy.

An Object is Worth Six Thousand Pictures: The Egocentric, Manual, Multi-Image (EMMI) Dataset

Xiaohan Wang, Fernanda M. Eliott, James Ainooson, Joshua H. Palmer, Maithilee Kunda
{xiaohan.wang, fernanda.m.eliott, james.ainooson, joshua.h.palmer, mkunda}@vanderbilt.edu
Department of Electrical Engineering and Computer Science, Vanderbilt University

Abstract

We describe a new image dataset, the Egocentric, Manual, Multi-Image (EMMI) dataset, collected to enable the study of how appearance-related and distributional properties of visual experience affect learning outcomes. Images in EMMI come from first-person, wearable camera recordings of common household objects and toys being manually manipulated to undergo structured transformations like rotation and translation. We also present results from initial experiments, using deep convolutional neural networks, that begin to examine how different distributions of training data can affect visual object recognition, and how the representation of properties like rotation invariance can be studied in novel ways using the unique properties of EMMI.

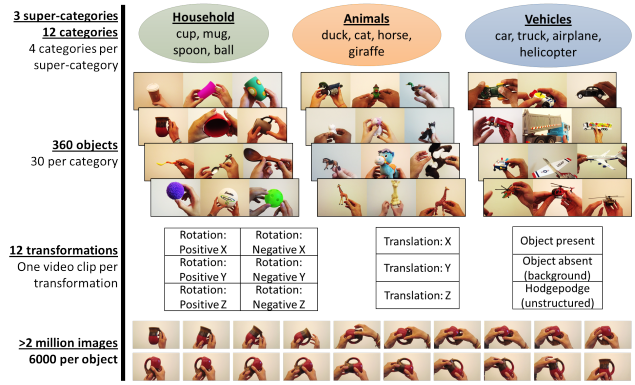

Figure 1. Structure of Egocentric, Manual, Multi-Image (EMMI) dataset, to support research on how appearance-related and distributional properties of first-person visual experience affect visual learning. (EMMI images brightened for PDF viewing.)

1. Introduction

Humans are fundamentally *egocentric* learners. Everything we learn comes in through our first-person sensorimotor systems. For example, the experiences that make up our "training data" to learn a task like object recognition are, for the most part, first-person views of our immediate surroundings. (The advent of television and the Internet has altered this somewhat, as we now have ready access to arbitrary images, but at least for many thousands of years of human existence, visual learning was necessarily driven by direct, first-person views of the world around us.)

First-person views can be characterized, and indeed are constrained, in two important ways. First, there are *appearance-related* properties inherent to first-person views, i.e., properties that are intrinsic and local to the views themselves. In contrast to distal views, for example, first-person views often contain close-ups of objects held by the viewer, with hands clearly visible [4, 19, 6, 17, 10].

Second, there are *distributional* properties of first-person views, i.e., non-local properties that involve relationships among all of the various first-person views experienced by a learner. For example, in contrast to distal views, which may have arbitrary distributions of content, first-person views will contain many views of the same objects or places, due to the fact that the viewer exists as a physical entity at a particular time and place in the world [11, 9].

Developmental psychology has begun to study how both types of properties play a role in human learning. For example, several studies now use wearable cameras to capture first-person views from young children in age ranges typically characterized by massive increases in object recognition ability [27]. These studies have found, for example, that toddlers obtain wonderfully clear, close-up, and unobstructed views of single objects—an appearance-related property—because they are old enough to reliably hold objects but young enough that their arms are quite short, which automatically brings the object close to the child's eyes [26]. Another study of infants in home settings found that views of objects are not uniformly distributed across object categories but instead strongly favor early-learned categories—a distributional property [9].

Of course, observing such properties in the first-person visual experiences of children, and even matching these

Figure 2. Top: Subset of EMMI objects. Bottom: Mug rotating around the $Y+$ axis. (EMMI images brightened for PDF viewing.)

properties to individual learning outcomes, does not by itself prove that these properties affect learning; observational studies can show correlation but not causation. Alternately, experimental studies with children (e.g., in lab settings) can manipulate children's experiences and then observe effects on learning in order to establish causal links, but are typically limited in ecological validity; lab studies can explain short-term learning scenarios but are harder to generalize to long-term childhood development.

A third approach is to use machine learning as an experimental platform to study relationships between properties of first-person visual experience and learning. While such computational experiments cannot directly tell us about processes of human learning, they *can* tell us something about the sufficiency of different combinations of representations, algorithms, and data for producing certain learning outcomes. Moreover, through systematic experimentation, we can begin to *quantify* the effects that these properties have on learning. Results will not only help us better understand visual learning in people but also will advance the state of the art in machine learning and computer vision.

Here, we describe a new image dataset collected to enable the study of how appearance-related and distributional properties of visual experience affect learning outcomes, called the Egocentric, Manual, Multi-Image (EMMI) dataset. Images in EMMI come from first-person, wearable camera recordings of common household objects and toys being manually manipulated to undergo structured transformations. What is unique about EMMI compared to many existing object recognition datasets is the number of distinct views generated per object ($\sim$6,000 image frames per object), combined with the coherence of views contained in these images (e.g., continuously varying, known rotations along different axes, etc.). Taken across a diversity of objects (30 objects per category) and categories (12 categories), this comes to over 2 million images in total.

We also present results from initial experiments that combine EMMI with convolutional neural networks to examine how different distributions of training data can affect visual object recognition, and how properties like rotation invariance can be studied in novel ways.

In particular, we investigate how the number of training images available per object, and the number of objects available per category, affect recognition performance, and how, at classification time, continuously varying views of the same object (e.g., a mug) smoothly alter the recognition behavior of the network (e.g., when the handle is in view vs. when it is hidden).

2. Related work

Many common object recognition datasets (e.g., ImageNet, Microsoft COCO, etc.) contain only one image per real-world object, a.k.a. the "Google image search" style of dataset. While these datasets have driven much exciting research in computer vision in recent years, they are, by their construction, limited in their applicability for supporting experiments on the effects of appearance-related or distributional properties of training inputs, such as the number of images or types of views available per object.

Several existing datasets do begin to fill this gap, as listed in Table 1. Each of these datasets contains more than one naturally captured image per object. We do *not* include synthetic image manipulations in this table, such as artificially skewing or scaling original images to create new ones.

The Egocentric, Manual, Multi-Image (EMMI) dataset presented in this paper continues and extends these prior efforts by providing a more structured and more dense sampling of viewpoints for objects in a variety of common categories. While other datasets have captured viewpoint variations (e.g., COIL, NORB, RGB-D, iLab-20M, etc.), many of these datasets have captured only a discrete collection of viewpoints, using, for example, a turntable turned to every 3°. EMMI contains images captured continuously at 30fps spanning full object rotations along all three rotational axes, as well as horizontal, vertical, and front-to-back (i.e., zooming) object translations.

Also, while at least one other dataset has captured manually performed, continuously varying rotations (e.g., iCubWorld-Transformations), these rotations are labeled only by broad type of rotation (e.g., in-plane or in-depth), and thus for a given image frame, specific pose information is not immediately available (i.e., would need be annotated).

Table 1. Review of image datasets that contain multiple real (i.e., not synthetically generated) images of the same physical object.

Dataset	Categories (labels)	Objs/cat	Rotated views/obj	Other variants	Imgs/obj	Total imgs
COIL-100 [20]	100 (household: mug, cup, can, bubblegum, block house, etc.)	∼1	72	n/a	72	7,200
SOIL-47 [8]	47 (household: cereal, crackers, mug, lightbulb, etc.)	∼1	21	lighting	42	1,974
NORB[1] [16]	5 (four-legged animal, human figure, airplane, truck, car)	10	324	lighting	1,944	97,200
ALOI [13]	1000 (household: tissues, duck, pineapple, ball, etc.)	∼1	75	lighting direction, lighting color	111	110,250
3D Object [23]	8 (bike, shoe, car, iron, mouse, cellphone, stapler, toaster)	10	24	zooming	72	∼7,000
Intel Egocentric[5,6] [22]	42 (household: bowl, cup, wallet, scissors, etc.)	1	various	background, manual activity	1,600	70,000
RGB-D[2] [14]	51 (household: mushroom, bowl, stapler, keyboard, etc.)	3-14	750	camera resolution	>750	250,000
BigBIRD[2] [25]	100 (household: crayon, toothpaste, cereal, etc.)	∼1-8	600	n/a	600	60,000
iCubWorld-Trfms.[3,5] [21]	20 (household: lotion, book, phone, flower, etc.)	10	∼1200	lighting, background, zooming	∼2,000	∼200,000
iLab-20M [7]	15 (vehicles: boat, bus, car, tank, train, etc.)	25-160	88	lighting, background, focus	>18,480	21,798,480
CORe50[2,4,5,6] [18]	10 (plug, phone, scissors, lightbulb, can, glasses, ball, marker, cup, remote)	5	∼1	indoor/outdoor, slight handheld movement	∼300	164,866
EMMI[5,6,7] [this paper]	**12 (cup, mug, spoon, ball, cat, duck, horse, giraffe, car, truck, airplane, helicopter)**	**30**	**∼4,200**	**translating, zooming**	**∼6,600**	**∼2,300,000**

[1] *Stereo pair images are not included in image counts.* [2] *Images collected as RGB-D video.* [3] *Updated counts taken from dataset website.*
[4] *From arXiv preprint.* [5] *Handheld objects.* [6] *Egocentric video.* [7] *Expected EMMI dataset size by date of publication.*

Manual rotations and other transformations in EMMI were timed to follow fixed patterns, so that estimates of object pose can be calculated for the majority of EMMI images.

As a final note on datasets, the datasets in Table 1 can also be divided by whether objects are on a table or other fixed surface (e.g., COIL, NORB, etc.) versus handheld (e.g., Intel, iCubWorld-Transformations, CORe50). Because objects in EMMI are handheld, images do contain some occlusions, which are unstructured; people collecting EMMI data were instructed to hold objects naturally while performing each object manipulation. Thus, EMMI may also be interesting as a testbed for studying the manual affordances of objects. For example, smaller objects in EMMI are often held with one hand, while larger objects in EMMI require holding with two hands.

How do datasets like the ones in Table 1 support studying how appearance-related and/or distributional properties affect learning? Obviously, many studies (practically all object recognition research) aims to *deal* with appearance-related and/or distributional properties of inputs, though often, the goal is to achieve invariance with respect to these properties—for example, a classifier that can recognize a mug in many different poses, or a learning algorithm whose performance does not depend on the distribution of training examples. Fewer studies explicitly examine the *effects* of such variations on learning.

Perhaps the largest bodies of relevant work are in active learning [24] and curriculum learning [5]. In active learning, the learner tries to choose, from a large pool of available training examples, which examples it would like to learn from next. In curriculum learning, an external agent tries to order training examples in such a way as to maximize learning outcomes for the learner, similar to the way that, in human education, a teacher can order material into a coherent "curriculum," to best scaffold the learning processes of a student. Both of these approaches highlight the ideas that both the content and ordering of training matters for learning. In one interesting demonstration, a study attempted to quantify the "learning value" contained in different training examples or subsets of examples [15].

Finally, we give two examples of studies that are very similar to the experimental work presented in the second half of this paper, and that exemplify the kinds of research that can be supported by the EMMI dataset. A study using the iCubWorld-Transformations dataset looked at effects of different distributions of training images on object recognition performance, based on the type of transformation represented in the images (e.g., rotation, zooming, etc.) or on the number of distinct images used per object [21]. A study using wearable camera data collected from adults and in-

Figure 3. Comparison of images from EMMI (top) and ImageNet (bottom). For EMMI images, household objects (left) are real, functional objects, though they do come in "adult" and "kiddie" versions. Animals (center) and vehicles (right) are replicas, either "realistic" scale models or "cartoony" toy objects. (EMMI images brightened for PDF viewing.)

fants playing with toys looked at how differing object sizes (i.e., handheld objects appear smaller to adults but larger to infants, due to differences in arm length) affect object recognition [3]. Both of these studies were conducted by fine-tuning a pre-trained deep network on different training sets, which is also the approach we use here (see Section 4).

3. EMMI Dataset Collection

Selection of categories. EMMI contains 12 categories, roughly grouped into three super-categories: household items (cup, mug, spoon, ball), animals (duck, cat, horse, giraffe), and vehicles (car, truck, airplane, helicopter). To maximize the usefulness of EMMI for comparisons with studies of human learning, all 12 of these categories are among the most common early-learned nouns for typically developing children in the U.S. [12]. Categories were selected both to provide ample shape variety in each super-category (e.g., spoon vs. ball, duck vs. cat, etc.) as well as shape similarity (e.g., cup vs. mug, car vs. truck, etc).

Selection of objects. Each category contains 30 different objects purchased from a combination of local household and toy stores, and Amazon.com. Individual objects were selected to provide variety within each category. In the super-category of "household items," objects are functional, real-world examples of that category, i.e., real cups, mugs, spoons, and balls (though we did include both adult and "kiddie" versions, see Figure 3). For both animals and vehicles, in contrast, we cannot include real ducks, cars, or helicopters, and so these objects are replicas. In these categories, we included both realistic, scaled-down model objects as well as "cartoony" toy objects (see Figure 3).

Recording devices and format. All videos were recorded using Pivothead Original Series wearable cameras, which are worn like a pair of sunglasses and have the camera located just above the bridge of the wearer's nose. Specific Pivothead settings included: video resolution set to *1920* x*1080*; frame rate set to *30 fps*; quality set to *SFine*; focus set to *auto*; and exposure set to *auto*.

Canonical views. For all objects, we defined a canonical view, which has the object held at a specified orientation, roughly centered in front of the camera-wearers eyes. For cups, the canonical view is defined as the object held upright. For mugs, the canonical view is defined as upright with the handle pointing to the right. For spoons, likewise, the canonical view has the handle pointed to the right and the bowl of the spoon turned up. For animals and vehicles, the canonical view is defined as the object facing towards the left (or standing with its head towards the left side, if its face is not aligned with its body).

Object videos. For each object, a set of 12 videos was recorded, as detailed in Table 2. For all rotations, each video contains two full revolutions of the object, over a fixed time course of about 20 seconds. For translations, each video contains three back-and-forth translations starting from the minus end of each axis, over a fixed time course of about 20 seconds. Rotations and translations were controlled to have an approximately constant velocity over the 20 second duration of the video. To do this, we developed a set of audio "temporal instruction templates" that camera-wearers would listen to while creating each video. **Thus, the pose of the object in every frame of a given video can be estimated according to the time of the frame.**

Table 2. Set of 12 videos collected per object in EMMI.

Label	Description	Time (s)
absent	background shot	2
present	steady hold at canonical view	2
rotation, X+	somersaulting towards viewer	20
rotation, X-	somersaulting away from viewer	20
rotation, Y+	in-plane clockwise rotation	20
rotation, Y-	in-plane counter-clockwise rotation	20
rotation, Z+	spinning (like a carousel) to the right	20
rotation, Z-	spinning (like a carousel) to the left	20
translation, X	horizontal, back-and-forth motion	20
translation, Y	in depth, back-and-forth (zooming) motion	20
translation, Z	vertical, back-and-forth motion	20
hodgepodge	unstructured object motion	20

Recording procedures. Objects were semi-randomly

assigned to several individual camera-wearers (all members of our research lab) for data collection. Efforts were made to ensure that no individual was over-represented in any category or object size class, to reduce any biases related to specific personal attributes or individual hand gestures. All videos were collected in an indoor setting against a white wall. No requirements were set as to time of day or specific lighting conditions, so there is variation in lighting across different objects (as can be seen in the example images in Figure 3).

Video-to-image conversion for experiments. The experiments described here used EMMI after the first half of data collection was complete, i.e. 15 objects per category across 12 categories, instead of the full 30 objects per category. Object videos were converted to images in jpeg format using FFmpeg. Around 1.2 million images were generated in this way. However, due to the limited field of view of the Pivothead wearable camera, we found that in some images, the object was almost or completely out of the image frame. To eliminate these "blank" images, the whole dataset was first used to re-train the Inception v3 neural network (as described in Section 4), with all 12 categories plus a 13th "blank" category that contained images from the "absent" videos, i.e., videos that recorded background only (see Table 2). The re-trained neural network was then applied back to the whole dataset to screen for "blank" images. About 10,000 "blank" images ($\sim$1%) were found using this classifier and, after manual confirmation, were subsequently removed for the experiments that are described next.

4. Methods

For initial, proof-of-concept experiments with EMMI, we used the transfer learning methodology appearing in many recent studies, e.g., [21, 3], which involves re-training the last layer of a pre-trained, deep convolutional neural network. Specifically, we used the ImageNet ILSVRC 2012 pre-trained Inception v3 network as a fixed feature extractor, and then re-trained the last layer using various subsets of our data to study the effects of different training regimes on object recognition performance. We used the Tensorflow software library for all experiments [2].

Inception is a representative convolutional neural network that has been shown to be highly successful in image recognition tasks [28]. The Inception v3 model we used here was pre-trained on the ImageNet ILSVRC 2012 dataset, which contains 1.2 million images from 1,000 categories. As a note on the pretrained network, most EMMI categories did appear in the original 1,000 categories used for pretraining—all except for helicopter, giraffe, horse, and duck. As the architecture of Inception v3 has been described extensively in other sources, we omit any detailed description here and refer readers to [28]. We conducted two experiments using the Inception v3 re-training methodology, using resources from the Tensorflow library [2].

Experiment 1, looking at *object diversity*, varied the number of objects per category used for re-training, with the total number of training images per category fixed at 1100 across conditions. For example, with one object per category, each of the 12 categories is represented by 1100 images of a single object from that category. With two objects per category, each category is represented by 1100 images uniformly drawn from two objects-550 images per object on average. As a comparison, we also used images from ImageNet for retraining, with 1100 images per category (which corresponds to having 1100 objects per category, since ImageNet essentially has one image per physical object).

Experiment 2, looking at *object view diversity*, varied the number of images per object used for re-training, with the total number of objects per category fixed at 12 across conditions, and the number of training images per category (drawn uniformly across the 12 objects) varied from 24—i.e., average of 2 images per object—to 1100—i.e., average of 92 images per object. Here, average number of images per object, drawn uniformly across all EMMI videos (except hodgepodge), is used to approximate views per object.

Both experiments were conducted using EMMI after data collection was halfway complete (15 objects per category out of the full 30 objects). For both experiments, training images were sampled uniformly across all videos in EMMI, excluding the hodgepodge videos. We generated different training datasets with these varying numbers and compositions of images from EMMI, and then retrained Inception v3 on each respective training dataset. For retraining, we modified a publicly available retraining script [1] published as part of the Tensorflow software package [2].

We used images from ImageNet, with 100 images sampled from each of the 12 EMMI categories, as the test set. **Note that the choice of using ImageNet images (instead of hold-out EMMI images) as the test set for our experiments was deliberate.** We aimed to explore how well training on a small number of handheld, often toy objects would be able to generalize to the very different objects represented in ImageNet (e.g., training on toy cats to recognize real cats). Certainly other testing approaches would also be interesting and will be pursued in future work.

5. Results

5.1. Experiment 1: Object diversity

In our first experiment, as described in Section 4, we tested how changing the number of distinct, physical objects represented in each category would affect recognition performance, with the total number of training images per category held constant. Training was done using images from the EMMI dataset, and testing was done using images from ImageNet. Results are shown in Figure 4.

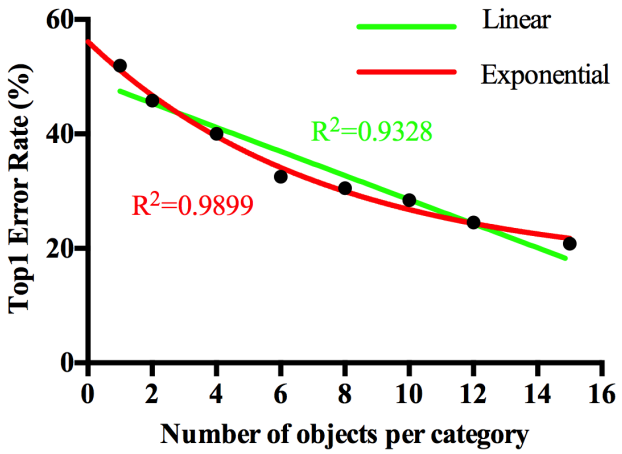

Figure 4. Experiment 1: Top-1 error rate on ImageNet test set as a function of object diversity in EMMI training set, ranging from 1 to 15 distinct physical objects per category, with total number of training images per category held constant at 1100. (Random baseline error rate for 12 categories is $\frac{11}{12} \approx 92\%$.)

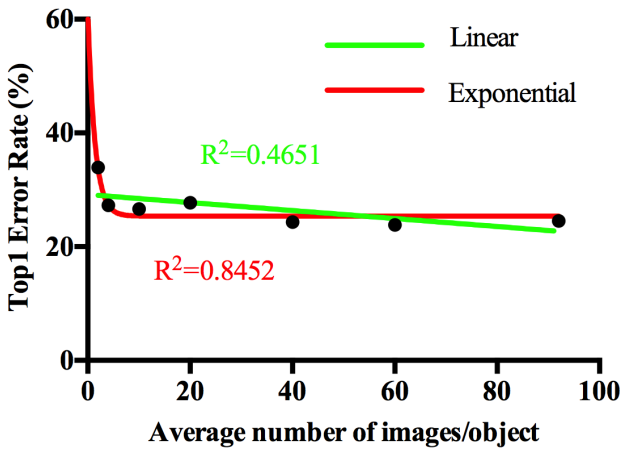

Figure 5. Experiment 2: Top-1 error rate on ImageNet test set as a function of object view diversity in EMMI training set, ranging from 2 to 92 images per object, with 12 objects per category. Total number of training images per category varies from 24 to 1100. (Random baseline error rate for 12 categories is $\frac{11}{12} \approx 92\%$.)

Using a training set with images of only a single EMMI object per category (i.e., 1100 images of a single object) yields an error rate of 51.92%, which while not excellent, is well below the random-guessing baseline error rate of 91.7%. Adding a second object (i.e., about 550 images of each of two objects) further reduces error to 45.83% Adding more objects per category (with total training images per category fixed at 1100) continues to improve performance significantly, with our final experiment using 15 objects per category yielding an error rate of just over 20%.

We also aimed to characterize the performance improvement by computing best-fit lines using both linear and exponential models. As shown in Figure 4, the exponential curve yields a better fit. Therefore, at least from the perspective of this model fitting, it appears that increasing object diversity will reduce the error rate in an exponential manner, with much greater improvements in performance for the first few added objects, and smaller increases thereafter.

Confusion matrices for results using 1 object per category and 15 objects per category are shown in Figure 6A and 6B, respectively. With fewer objects per category during training, there tend to be more false positives in certain categories. Increasing the number of objects per category during training greatly reduces the false positive rate.

For comparison, we performed the same retraining procedure using a training set drawn from ImageNet with 1100 images per category, corresponding to 1100 distinct objects per category. (This "data point" could be imagined to fall at the point $x = 1100$ on the plot shown in Figure 4.) The error rate obtained for this ImageNet retraining was 4.9%, and the confusion matrix is shown in Figure 6F.

Plugging in 1100 into the best-fit exponential curve shown in Figure 4 yields an expected error rate of 16.56%, and so the actual results using ImageNet are much better. This difference could reflect a fundamental limitation of training with toy objects from EMMI, or it could reflect the fact that the deep-level features in our network were originally trained on ImageNet categories to begin with. These are just a sampling of the interesting open questions raised by Experiment 1 results.

5.2. Experiment 2: Object view diversity

The second experiment, as described in Section 4, studied how varying the number of distinct views of each object would affect recognition performance. We approximated number of views as "number of images" uniformly sampled across all EMMI videos (excluding hodgepodge), with the total number of objects per category held constant at 12, and different average numbers of training images per object in each condition. As with Experiment 1, training used images from the EMMI dataset, and testing used images from ImageNet. Results are shown in Figure 5.

When there are only two images per object on average, the top-1 ImageNet error rate is 33.9%. This error is reduced to 27.3% after doubling the average number of images per object to 4. When increasing the average number of images per object to 92 (1100 images per category, as in Experiment 1), the error rate is around 24%, representing a 10% improvement relative to just 2 images per object. Again, the results better match an exponential fit ($R^2 = 0.85$) than a linear fit ($R^2 = 0.47$).

Confusion matrices for training using 2, 20, and 92 images per object are shown in Figure 6C, D, and E, respectively. As in Experiment 1, as the diversity of object views

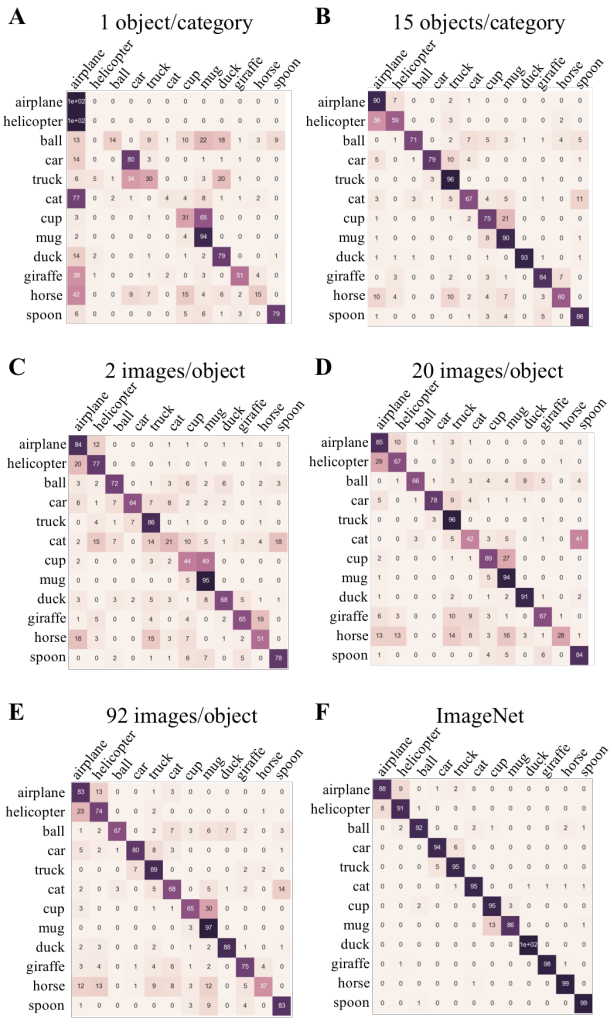

Figure 6. Confusion matrices showing the effects of increasing object diversity (A, B) or object view diversity (C, D, E), along with a comparison using ImageNet (F). Row labels are true categories and column labels are predictions.

increases, accuracy improvements, especially for certain categories, are clearly visible in these matrices.

While the results from Experiment 1 and Experiment 2 shown here—that increasing the diversity of objects and object views improves object recognition performance—are not likely to be surprising to anyone familiar with machine learning, what these experiments do offer is a unique opportunity to quantify and characterize these improvements.

6. Discussion

Large scale image datasets such as ImageNet have been the driving force for recent breakthroughs in image recognition. But the nature of these datasets (i.e., one image per physical, real-world object) makes it hard to study certain factors that are likely to be important for visual learning, especially factors like appearance-related and distributional properties that are central to the nature of egocentric vision. We created the EMMI dataset with the hope of complementing the roles of existing datasets like ImageNet.

A typical ImageNet category contains about 1200 images from 1200 objects. With EMMI, we showed that with as few as 15 objects per category, the transfer learning of the pretrained Inception v3 convolutional neural network can achieve 20% top 1 error rate, using ImageNet as the test set. It is worth pointing out that for categories such as car, truck and airplane, the images from the EMMI dataset are either small replicas or cartoony toys, while we tested recognition performance against the real-world images from ImageNet. It is surprising that the neural network, retrained only on small replicas and cartoony toys, is able to recognize real world objects from ImageNet with decent performance. In addition, all experiments were done by fine tuning the last layer. One interesting future direction would be to train the neural network from scratch using EMMI, and see how the early features are learned.

It is also worth noting that all images in EMMI contains the camera-wearer's hands holding, and partially occluding, the object. No image segmentation was performed as part of these experiments; the neural network received training inputs consisting of entire images from EMMI, hands and all. However, the network was still able to learn to classify non-egocentric, handless images from ImageNet.

We also showed that multiple views of the same object have the potential to enhance recognition performance, even when the number of objects per category remains the same. We did not include any data augmentation in our retraining process. One interesting future direction will be to systematically test how training with specific types of views of the same objects affect recognition performance, in comparison to traditional methods for synthetic data augmentation, and whether the two methods can work synergistically.

Another valuable research aspect of EMMI is having timed rotations and translations in the collected videos (although values are approximated). For example, because of the dense and controlled nature of the visual transformations represented in EMMI, we can reveal intriguing aspects of the neural network activations at classification time.

Figure 7A shows representative EMMI images from one mug and one cup undergoing rotation about the $Z+$ axis. The rest of the plots in this figure show activations of the neural network generated while classifying, in sequence, frames from this video. Note that the mug and cup images show one full rotation, but the plots show network changes over two full rotations.

First, we looked at neuron activations of the last hidden layer from Inception v3, as shown in Figure 7B. Inception v3 contains 2048 neurons in the last hidden layer, which are fully connected to the 12 output neurons.

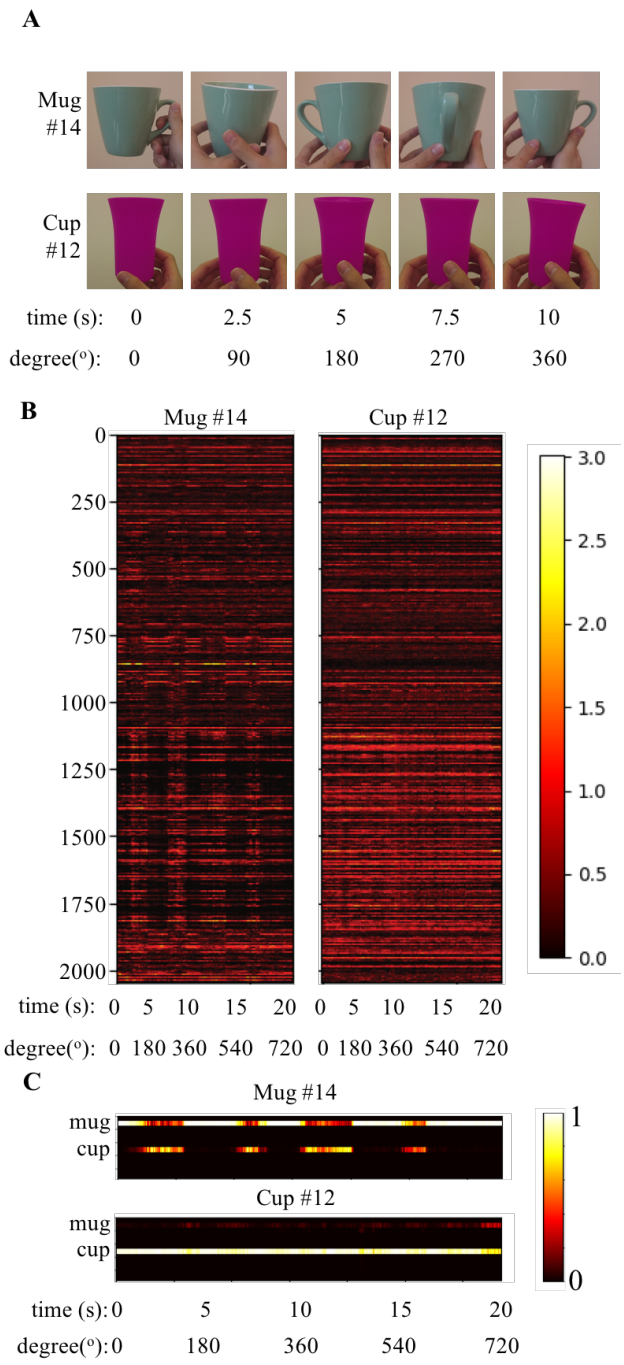

Figure 7. Representative object rotations (top) and corresponding neuron activations from the last hidden layer (middle) and the output layer (bottom) of the retrained neural network, at classification time, plotted as a function of time when receiving continuous inputs from EMMI video of object undergoing $Z+$ axis (i.e., carousel-like) rotation. Note the clear phases of activation demonstrated by the network while recognizing the mug, which appears "cup-like" when the handle is hidden, versus the relatively constant activation demonstrated by the network while recognizing the cup, which is rotationally symmetric about the $Z+$ axis.

Note that these neuron activations originate purely from the pre-trained Inception v3 network. They are not changed during our retraining process; only the links coming from this last hidden layer to the final output layer are subject to change during retraining. Each horizontal line represents the activity of a given neuron over the entire time span.

When "watching" the mug rotate, some neurons remain silent, some neurons get activated periodically as the mug rotates, and some neurons fire throughout the time course irrespective of the rotation. In contrast, the neuron firing pattern for a rotating cup is more constant. Because of the timed rotation, the periodical firing can be easily matched to certain rotating positions, and in fact we see that certain parts of this periodic signal correspond to moments when the mug looks "cup-like" because the handle is out of view.

We also plotted neuron activations of the output layer (the retrained classification layer) during these rotations. As shown in Figure 7C, the "mug neuron" fires at high levels overall when the network is watching the mug rotation, and likewise for the "cup neuron" while watching cup rotation. Interestingly, the mug neuron firing will dip and the cup neuron firing will go up when the mug handle is either behind or in front of the mug body. While this observation makes sense when the handle is hidden behind the mug, it shows limitations of the network's performance when the handle is in front; perhaps the handle is too low-contrast to be detected, or is interpreted as a flat pattern on the cup.

More generally, studying results in this way could give insights into which neurons in the network are responsible for representing certain category features (e.g. mug handles), and could also reveal how properties like rotation-invariance are encoded in a deep network.

7. Contributions

In this paper, we have presented a new dataset, called the Egocentric, Manual, Multi-Image (EMMI) dataset, to support research into visual learning. EMMI will enable systematic research into the effects of appearance-related and distributional properties of first-person visual experience on many kinds of visual learning, including (but not limited to) research on object recognition, as presented in proof-of-concept experiments here. Ultimately, we expect that EMMI will serve as a highly complementary resource to accompany the use of existing datasets like ImageNet.

Acknowledgments

Many thanks to Ellis Brown, Max de Groot, and Harsha Vankayalapati for help in data collection, and to Max de Groot, Fuxin Li, Jim Rehg, Linda Smith, and Chen Yu for valuable conversations. This work was supported in part by a Vanderbilt Discovery Grant, titled "New Explorations in Visual Object Recognition."

References

[1] TensorFlow retraining script. https://github.com/tensorflow/tensorflow/tree/master/tensorflow/examples/image_retraining.

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] S. Bambach, D. J. Crandall, L. B. Smith, and C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2016.

[4] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[6] A. Betancourt, M. M. López, C. S. Regazzoni, and M. Rauterberg. A sequential classifier for hand detection in the framework of egocentric vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 586–591, 2014.

[7] A. Borji, S. Izadi, and L. Itti. ilab-20m: A large-scale controlled object dataset to investigate deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2221–2230, 2016.

[8] J. Burianek, A. Ahmadyfard, and J. Kittler. Soil-47, the surrey object image library. *Centre for Vision, Speech and Signal processing, Univerisity of Surrey.[Online]. Available: http://www. ee. surrey. ac. uk/Research/VSSP/demos/colour/soil47*, 2000.

[9] E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith. Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, 372(1711):20160055, 2017.

[10] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 407–414. IEEE, 2011.

[11] C. M. Fausey, S. Jayaraman, and L. B. Smith. From faces to hands: Changing visual input in the first two years. *Cognition*, 152:101–107, 2016.

[12] L. Fenson, E. Bates, P. S. Dale, V. A. Marchman, J. S. Reznick, and D. J. Thal. *MacArthur-Bates communicative development inventories*. Paul H. Brookes Publishing Company, 2007.

[13] J.-M. Geusebroek, G. J. Burghouts, and A. W. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

[14] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.

[15] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba. Are all training examples equally valuable? *arXiv preprint arXiv:1311.6510*, 2013.

[16] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–104. IEEE, 2004.

[17] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2013.

[18] V. Lomonaco and D. Maltoni. Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*, 2017.

[19] W. W. Mayol and D. W. Murray. Wearable hand activity recognition for event summarization. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium On*, pages 122–129. IEEE, 2005.

[20] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100). 1996.

[21] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale. Object identification from few examples by improving the invariance of a deep convolutional neural network. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 4904–4911. IEEE, 2016.

[22] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2009.

[23] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[24] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[25] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 509–516. IEEE, 2014.

[26] L. B. Smith, C. Yu, and A. F. Pereira. Not your mothers view: The dynamics of toddler visual experience. *Developmental science*, 14(1):9–17, 2011.

[27] L. B. Smith, C. Yu, H. Yoshida, and C. M. Fausey. Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, 16(3):407–419, 2015.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.