An Object Is More Than a Single Image: The Toybox Dataset of Visual Object Transformations

Xiaohan Wang* Tengyu Ma* Azhar Molla Seunghwan Cha
James Ainooson Xiaotian Wang Maithilee Kunda
Department of Electrical Engineering and Computer Science, Vanderbilt University
PMB 351679, 2301 Vanderbilt Place, Nashville, TN 37235-1679, USA
{xiaohan.wang, tengyu.ma, a.molla, seunghwan.cha,
james.ainooson, xiaotian.wang, mkunda}@vanderbilt.edu

Abstract

Machine learning-based approaches have made great achievements in computer vision thanks to the availability of high-quality data and the increase of computational power. However, most of the state-of-the-art techniques are mainly trying to solve the "what is where" problem and miss some other significant dimensions of an image, for example, what's the physics behind the object in the image and how human can manipulate the object. To fill in the gap from only recognizing the object to reasoning about the object, we introduce a new video dataset, Toybox. Videos in Toybox come from first-person, wearable camera recordings of common household objects and toys being manually manipulated to undergo structured transformations like rotations and translations. We also present results from initial experiments using deep convolutional neural networks that begin to examine how different perspectives in a 3D scene of training data can affect visual object recognition performance, and how our dataset can be used to learn hand-object-scene interaction.

1. Introduction

Recent breakthroughs in computer vision, particularly for the problem of visual object recognition have been largely driven by the creation and use of large-scale labeled image datasets, with ImageNet being the canonical example. Thanks to synergistic progress in datasets, CNN architectures, and computing hardware, we can now train a deep CNN that performs on par with or even better than humans in many vision tasks. However, deep CNNs that are trained from general purpose labeled dataset such as ImageNet missed the ability of reasoning about images and the scene of human-object interaction. **Therefore, we present a new dataset called Toybox that has been designed to enable an improved understanding of small sample learning and hand-object-scene interaction, though we expect this dataset will be valuable for many other areas of computer vision research as well.** Toybox contains videos of structured visual transformations over individual objects, as illustrated in Figure 1, which will enable many innovative scientific experiments with CNNs that are not possible with ImageNet or similar datasets.

The design of the Toybox dataset was motivated by the following research questions related to visual object recognition: 1) to what extent is a diversity of individual objects necessary and/or sufficient to train (or retrain) a CNN? 2) to what extent does having various perspectives of an object available during training affect recognition performance? 3) how can human hands stably manipulate various objects?

While by no means are we attempting to address all of these questions in this paper, we present details of the Toybox dataset, including comparisons with other similar datasets as listed in Table 1, and results from initial CNN-based experiments that highlight some of the unique contributions of the Toybox dataset.

1.1. Related work

Many common object recognition datasets (e.g., ImageNet, Microsoft COCO, etc.) contain only one image per real-world object. 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 to understand the training process of deep CNNs. Several existing datasets are already beginning to fill this gap, as listed in Table 1.

The Toybox 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

*These authors contributed equally to this work.

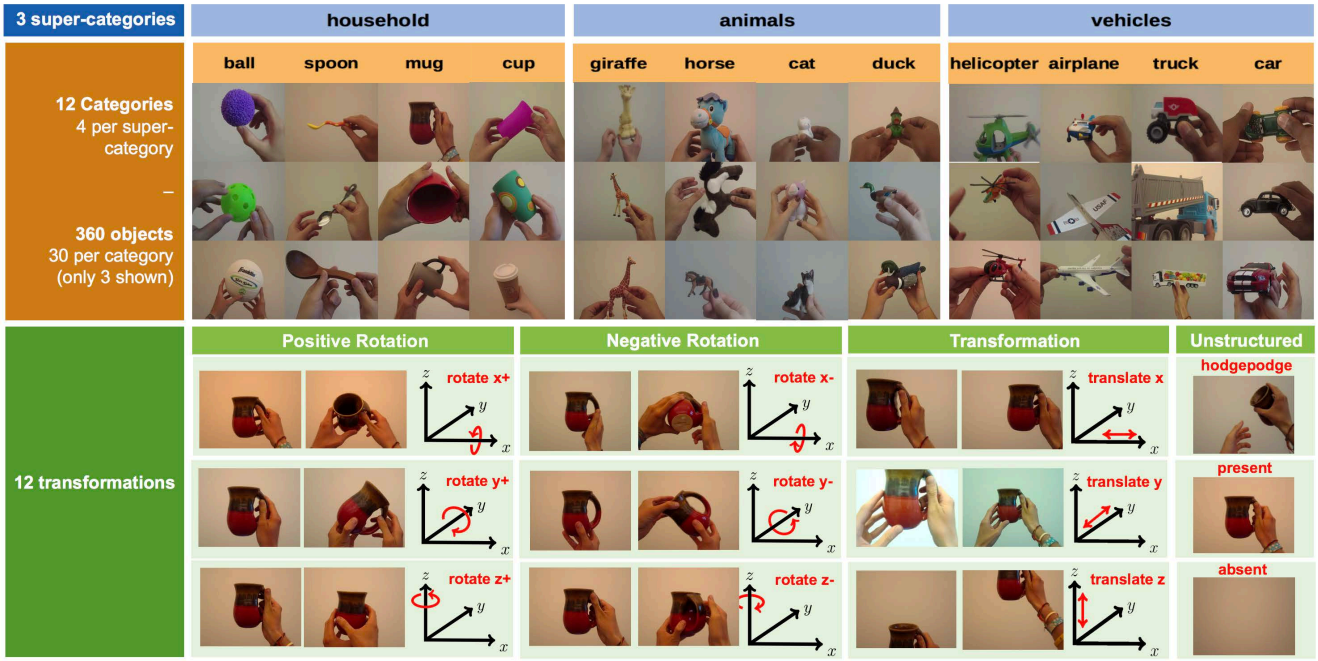

Figure 1. An overview of the Toybox dataset. There are 12 video clips for each object. Except for absent and present, all videos are 20 seconds long and contain a defined transformation of the object. Rotations and translations contain two revolutions and three translations along a defined axis, respectively; hodgepodge contains unstructured object motion. For animals and vehicles, we included both cartoony toys (e.g., top row) and scaled-down, realistic models (e.g., bottom row).

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 by every 3°. Toybox 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.

2. The Toybox dataset

Selection of categories and objects. Toybox 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). 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). Each category contains 30 different objects. For both animals and vehicles, we cannot include real objects, so these objects are either realistic, scaled-down model objects or "cartoony" toy objects (see Figure 1).

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-wearer's eyes.

Recording devices and Object videos. 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*. For each object, a set of 12 videos was recorded, as shown in Figure 1. Except for absent and present, all videos are about 20-second long. For rotations, each video contains two full revolutions of the object; for translations, each video contains three back-and-forth translations starting from the minus end of each axis. 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.**

3. Experiments

For initials, proof-of-concept experiments with Toybox, we used the transfer learning methodology appearing in many recent studies, e.g., [2, 11], which involves re-training the last layer of a pre-trained, deep convolutional neural network.

For Section 3.1, we used the ImageNet ILSVRC 2012 pre-trained Inception v3 network as a fixed feature extractor, and

Table 1. Review of image datasets that contain multiple real images of the same physical object.

Dataset	Categories (labels)	Objs/cat	Rotated views/obj	Other variants	Imgs/obj	Total imgs
COIL-100 [10]	100 (household: mug, cup, can, etc.)	∼1	72	n/a	72	7,200
SOIL-47 [4]	47 (household: lightbulb, mug, etc.)	∼1	21	lighting	42	1,974
NORB[1] [7]	5 (human figure, car, truck, etc)	10	324	lighting	1,944	97,200
ALOI [5]	1000 (household: duck, tissues, etc.)	∼1	75	lighting direction, lighting color	111	110,250
3D Object [13]	8 (household: bike, shoe, car, etc)	10	24	zooming	72	∼7,000
Intel Egocentric[5,6] [12]	42 (household: scissors, bowl, cup, wallet, etc.)	1	various	background, manual activity	1,600	70,000
RGB-D[2] [6]	51 (household: bowl, stapler, etc.)	3-14	750	camera resolution	>750	250,000
BigBIRD[2] [14]	100 (household: crayon, cereal, etc.)	∼1-8	600	n/a	600	60,000
iCubWorld-Trfms.[3,5] [11]	20 (household: lotion, book, phone, etc.)	10	∼1200	lighting, background, zooming	∼2,000	∼200,000
iLab-20M [3]	15 (vehicles: boat, bus, car, tank, train, etc.)	25-160	88	lighting, background, focus	>18,480	21,798,480
CORe50[2,4,5,6] [8]	10 (household: plug, phone, scissors, etc.)	5	∼1	indoor/outdoor, slight handheld movement	∼300	164,866
Toybox[5,6] [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.*

then re-trained the last layer using Toybox dataset and tested using the ImageNet dataset **Note that the choice of using ImageNet images (instead of hold-out Toybox 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). We used the Tensorflow software library for all experiments [1].

For Section 3.2, based on the modern methods proposed in the paper [9], we used Matlab to automatically draw the hand-skeleton and then manually adjust the skeleton to be more accurate. Currently, only a small part of our objects are labeled with the hand-skeleton, and we aim to label more in the future.

We first looked at the effect of *object diversity* on transfer learning, by varying the number of objects per category in the training dataset, 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). A training set with images of only a single Toybox object per category (i.e., 1100 images of a single object) yields an average error rate of 60.63%, 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 51.98%. Adding more objects per category (with total training images per category fixed at 1100) continues to improve performance significantly, with our final experiment using 30 objects per category yielding an average error rate of 21.43% (Figure 2A).

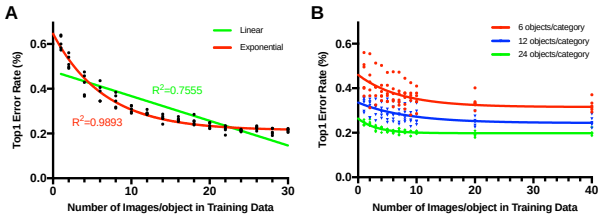

Figure 2. Effects of object diversity and view diversity on training performance. **A**: top-1 error rate on ImageNet test set as a function of object number per category (object diversity) in the Toybox training set, ranging from 1 to 30 distinct physical objects per category, with the total number of training images per category held constant at 1100. **B**: top-1 error rate on ImageNet test set as a function of image number per object (object view diversity) in the Toybox training set, ranging from 2 to 40 images per object. The effect of view diversity was also tested with different object numbers (i.e., 6, 12, 24) per category. For instance, the total number of training images per category varies from 24 to 480 for the 12 objects per category group. All data points were from 5-6 independent experiments with different objects selected randomly.

3.1. Effects of object and view diversity

We then looked at *object view diversity*, by varying the number of images per object used for transfer learning, with the total number of objects per category fixed at 6, 12, and 24, respectively. For instance, for the group of 12 objects per category, we gradually increased the total number of images per object from 2 to 100 (drawn uniformly across all the 12 objects). Without loss of generality, concentrating on the blue points and the exponential fitted curve (12 objects per category), with a single image per object, the average top1

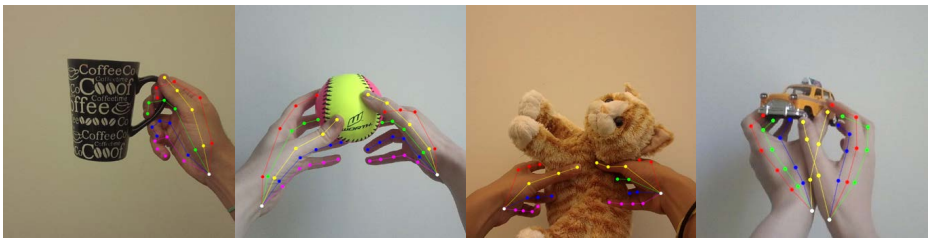

Figure 3. The Toybox dataset could serve as training data to learn how to naturally manipulate objects.

error rate is 33.0%. This error rate is subsequently reduced to 27.5% if we have 10 images per object, and is further reduced to 25.6% and 24.8% for 20 and 40 images per object, respectively. Increasing the number of views per object can apparently improve the performance of the classifier at the very beginning, i.e., 40 images per object achieve more than 8% lower error rate than the 1 image per object. On the contrary, if we keep increasing the number of images, for example, 100 images per object with average error rate 23.9%, the improvement becomes limited with only a 1.1% error rate decrease compared to the result obtained with 40 images per object.

3.2. Hand-object-scene interaction

In Human-object-scene interaction, human hand will be the primary organ to interact with objects. Our dataset could be potentially useful for hand-object-scene applications since the rich hand gestures as shows in Figure 3 in our Toybox dataset could serve a training dataset to learn how human handle objects stably and efficiently.

4. Discussion

We showed in this paper that our new Toybox dataset could complement existing dataset in studying the effect of small sample learning and hand-object-scene interaction. By providing structured videos showing a range of transformations using human hand, we can systematically analyze more dimensions of an image beyond "what is where" in a way that is not possible with the canonical datasets.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] V. Lomonaco and D. Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In S. Levine, V. Vanhoucke, and K. Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR, 13–15 Nov 2017.

[9] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.