Active Vision Dataset Benchmark

Phil Ammirato
UNC-Chapel Hill
ammirato@cs.unc.edu

Alexander C. Berg
UNC-Chapel Hill
aberg@cs.unc.edu

Jana Košecká
George Mason University
kosecka@gmu.edu

Abstract

Several recent efforts in computer vision indicate a trend toward studying and understanding problems in larger scale environments, beyond single images, and focus on connections to tasks in navigation, mobile manipulation, and visual question answering. A common goal of these tasks is the capability of moving in the environment, acquiring novel views during perception and while performing a task. This capability comes easily in synthetic environments, however achieving the same effect with real images is much more laborious. We propose using the existing Active Vision Dataset to form a benchmark for such problems in a real-world settings with real images. The dataset is well suited for evaluating tasks of multiview active recognition, target driven navigation, and target search, and also can be effective for studying the transfer of strategies learned in simulation to real settings.

1. Introduction

Vision is a key capability for robots to operate successfully in the everyday world around us, but necessary capabilities are not stressed in previous computer vision datasets and benchmarks. For instance, despite how central motion is to robotics, current state-of-the-art object detection approaches in computer vision [13, 12, 10], operate independently on each input image, at least in part because this is how datasets and benchmarks (*e.g.*, ImageNet[14] and COCO[11]) are designed.

There is a new crop of datasets of environments and APIs for training and testing visual perception approaches that can be helpful for robotics [22, 9, 15, 19, 20, 5]. These are designed to allow studying navigation, mobile manipulation, and visual question answering. The common structure of these APIs allows agents to move in, and sometimes interact with, an environment and then returns a new view of the environment. While this capability comes easily for synthetic environments, achieving the same functionality with real images is more challenging.

We build a new benchmark based on the Active Vision Dataset [1] to stress visual navigation and few-shot object detection in real world scenes. The main advantage of the Active Vision Dataset is that it is made up of real captured images (and depth maps) of real scenes, thus stressing aspects of approaches that computer graphics based evaluation may not.

The Active Vision Dataset consists of very dense collections of RGB-D imagery of real-world scenes in order to allow simulating a robot moving through an environment by simply sampling the appropriate captured view. The dataset includes everyday rooms: kitchens, living rooms, dining rooms, offices, bathrooms, etc. A common set of manipulable objects encountered in research on manipulation (BigBird[16]) are placed in each scanned scene, and these are labeled with bounding boxes. The data for a typical room will include thousands of RGB-D images, and thousands of object bounding boxes for a few dozen objects. The views are organized by camera position, so that an API can allow an agent to specify a relative motion in the scene and properly select the next view that is seen (e.g. what would I see if I turned right? went backwards?). More details can be found in [1].

The Active Vision Dataset **attempts to address several difficulties** in studying object detection and recognition relevant to robotics applications. Generally work on this area in the past suffered from the lack of realistic benchmarks which would enable comparisons of different approaches. Approaches were tested on robotic platforms with wide variations in the environments and objects encountered, making direct comparisons difficult. AVD allows approaches to be compared on exactly the same environment. More recently shared CG-based environments that do allow comparison, do not provide the variability of real images of scenes (with some exceptions for panoramic datasets, *e.g.* [8]). Very recent datasets including those based on Matterport3D[4] do use real imagery, but sample camera positions somewhat sparsely requiring interpolation (image-base rendering) between recorded views, again introducing artifacts. AVD provides a denser set of views that does not require interpolation.

Taking a step back to compare the state of work on manipulation, where repeatability is addressed by considering standard datasets of objects to be manipulated and distributing the physical objects to different groups. The underlying vision tasks include segmentation, object instance recognition, and pose recovery in clutter and these are exhibited in some recent challenges [21, 16, 3, 7]. These however do not exhibit the large variations in scale, viewpoint, and background clutter encountered in mobile manipulation problems, where the camera position relative to objects can vary more widely as the robot platform moves, and which require detection of, and navigation to, objects of interest.

2. Related work

There has been a recent influx of datasets focusing on active vision tasks, where visual observations are considered jointly with some control/action authority. The existing datasets vary in the level of visual realism they provide, their scale, the type of modalities they can simulate and ability of agents to interact in the world. They can be broadly partitioned into CG synthetic worlds derived from the original SUNCG [17] dataset or datasets derived from scans of the real world [4].

The MINOS [15] environment is a synthetic environment which contains both synthetic scenes from SUNCG (with over 45,000 scenes) and meshes of reconstructed scenes from the Matterport3D dataset [4] (with 90 multi-floor houses). While the scale of the dataset is appealing, the quality of the visual observations is limited to either synthetic renderings of SUNCG scenes or renderings of reconstructed meshes, which suffer from many reconstruction artifacts affecting the visual observations. Additional sensing modalities of depth, plane normals and semantic segmentation and capability of arbitrary viewpoints and continuous motion are enabled by the dataset. House3D [19] is also a fully simulated large scale environment derived from SUNCG and enables visual observations as rendered views along with depth and semantic segmentation.

Matterport3D [4] is a large-scale RGB-D dataset containing 10,800 panoramic views about 1-1.5m apart with surface reconstructions, camera poses, and 2D and 3D semantic segmentation annotations. The scale and visual realism of the data is impressive, but the poses where high resolution panoramas as available are quite sparse. Views generated outside of the panorama grid are obtained by rendering mesh reconstructions and have notable artifacts.

Efforts to eliminate some of the reconstruction artifacts have been tackled recently by [20] which used a novel image-based rendering approach to eliminate some of the visual artifacts. The resulting rendered views while free of some artifacts are still quite blurry.

The AI2-THOR [9] environment also falls into the category of CG synthetic environments allowing continuous and discrete motion and near photo-realistic 3D indoor scenes. The API for this environment does not provide access to other modalities and the initial scale of the environments is smaller than other synthetic datasets (2-3 bedroom houses compared to multi-story buildings provided by Matterport3D and SUNCG). Another effort at a simulated world is [18]

The tasks studied in the context of these datasets and environments include navigation [6], target driven navigation [22], visual question answering [5], and planning [2]. While it can make sense to study each of these in both completely artificial environments as well as real scenes, using real imagery allows probing aspects of visual perception that might overfit or otherwise yield unrealistic performance on CG data.

3. Active Vision Dataset Benchmark (AVDB)

The goal of AVDB is to help develop and compare approaches to vision problems that are relevant for robotics on a repeatable real-world environment. For each task, a training set, validation set, and test set from AVD is specified, and the testing methodology fixed. In some cases additional training data from outside sources will be expected. While the dataset can be used for benchmarking straightforward vision tasks like detection, we focus on exploring tasks involving active vision, transfer learning, and class-incremental learning.

For the benchmark we have augmented the original AVD collection with 7 new scans of scenes to be used in testing. Collectively these new scans account for over 9,000 images and 18,000 bounding boxes. We will release the images and movement pointers for these scenes, but will keep the labels private for the benchmark evaluation.

3.1. Task: Active Object Search

The goal of the first task in AVDB, Active Object Search, is to localize and navigate towards pre-specified objects of interest. Each training and validation episode starts at a random starting position in a random scene with a random target object. The episode is declared as success when the navigation strategy stops at a location within <1m of the object and the object is visible in the image at that location. For each scene/object pair in the test set, we have randomly chosen a fixed set of starting positions (location $+$ direction). This provides a large test sample while keeping the starting positions consistent for every evaluation, allowing fair comparisons. Systems will be evaluated based on the average number of steps needed to achieve success for each scene/object pair, with an upper bound. We provide an **OpenAI Gym** style environment that can be used with the AVD data for training and testing on this task.

Within this Active Object Search task, we define three specific train/test scenarios.

3.1.1 Known Environment

As described in [1], some of the scenes in AVD are scanned twice. In the second scan some objects are moved around, some are removed completely, and some new objects are introduced. In this task, the system is given the first scan of a scene during training, and is tested only on the second scan of that scene.

The training set consists of the first scan of the scene of interest, and optionally any other scans from AVD or other datasets. The test set is the second scan of the scene of interest, with a set of starting positions for each object as described above. We have three ”second” scans of scenes in our new test data, the first scans have already been publicly released. It is expected that a different model will be used for each test scene as this task is environment specific, and the final evaluation metric will be averaged over these three scenes.

For validation style data, there exist five scenes in AVD currently that have two scans. We recommend a system be evaluated using each of these scenes as the scene of interest, one at a time, for validation before testing.

3.1.2 Unknown Environment

In this task, the test scenes are not seen at all during training. We provide four new scenes for the test set in this task. The training set consists of the currently available AVD scenes, and we recommend holding out at least three scenes during training for validation. In both the known environment task and this task, the set of objects being searched for are the same in training and testing.

3.1.3 Transfer Learning

Learning approaches to both visual question answering tasks and navigation strategies typically require large amounts of training data, which is laborious to attain on real robotic platforms and does not support the repeatability of the results. Many existing approaches mentioned in the related work section opt to use simulated environments for these tasks. AVDB can be used for evaluation of the transfer of these tasks to real-world environments.

3.2. Task: Class-incremental Learning

A robot in the real world will likely encounter objects outside those in its initial training set. It would be useful if the robot could recognize these new objects with as few examples and in as little time as possible.

For this task we consider 27 object instances that are present in AVD. The idea is to train on scenes considering 17 of the instances, and test on scenes considering all 27 instances. In addition to training scenes with 17 objects, 1-10 “target images” of each of the 27 objects in isolation will be provided at test time. The aim is to compare object instance detection with training in scenes to object instance detection where only a small number of isolated example images are provided for training.

The training set will consist of all 30,000+ images currently available in AVD, but only seventeen instances will be considered foreground. The other ten instances must be treated as background, or blacked out, etc. In addition, 1-10 “target images” of the 17 instances are available at training time. The seven new scenes collected for AVDB compromise the test set, which contains all 27 instances. “Target images” for the remaining ten test instances are provided at test time. Use of any additional training data is permitted, as long as the ten test instances are not present as foreground in any training data. Systems will be evaluated using the mean average precision metric commonly used in object detection, over all 27 instances.

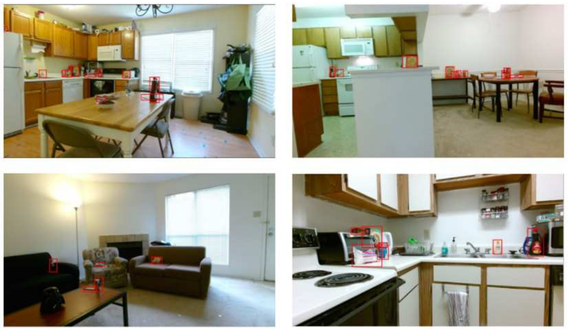

Figure 1: Some examples of labeled images.

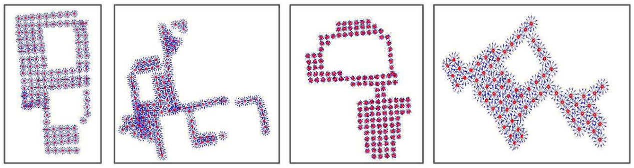

Figure 2: Camera locations (red) and directions (blue) from AVD. The dense sampling of images allows benchmarking active navigation tasks using visual input.

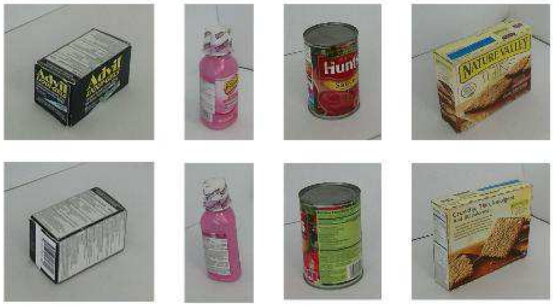

Figure 3: Example “target images”.

References

[1] P. Ammirato, P. Poirson, E. Park, J. Kosecka, and A. C. Berg. A dataset for developing and benchmarking active vision. In *ICRA*, 2017.

[2] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. *arXiv preprint arXiv:1711.07280*, 2017.

[3] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 2017.

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[5] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied Question Answering. *arXiv preprint arXiv:1711.11543*, 2017.

[6] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. *ACCV*, 2011.

[8] D. Jayaraman and K. Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In *ECCV*, 2016.

[9] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta1, and A. Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr. Focal loss for dense object detection (best student paper award). In *International Conference on Computer Vision (ICCV), Venice, Italy*, 2017. Oral.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zrich, 2014. Oral.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[15] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[16] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *ICRA*, 2014.

[17] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] Y. Z. S. Q. Z. X. T. S. K. Y. W. A. Y. Weichao Qiu, Fangwei Zhong. Unrealcv: Virtual worlds for computer vision. *ACM Multimedia Open Source Software Competition*, 2017.

[19] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian. House3d: A rich and realistic 3d environment. 2017.

[20] A. Zamir, F. Xia, and Z.-Y. H. et al. Gibson environment for embodied real-world active perception.

[21] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018.

[22] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017.