The Need for Dynamic & Active Datasets

Jürgen Leitner, Donald G. Dansereau, Sareh Shirazi, Peter Corke
ARC Centre of Excellence for Robotic Vision (ACRV)
Queensland University of Technology, Australia
j.leitner@roboticvision.org

Abstract

Datasets provide open and accessible benchmarks for the refinement of existing algorithms and the testing of new techniques. Notably, recent advances in computer vision, particularly in convolutional and deep neural networks, have been enabled by the availability of appropriate datasets. Though datasets have evolved to contain ever more data, the focus has remained on static and passive information. We propose that in emerging application areas such as augmented reality and robotics, a more active and dynamic approach to datasets is required. We review key limitations of existing datasets, and propose directions for discussion and creation of deeper dataset capabilities, including accommodating shifts in camera pose, scene behaviour and affordances, and variations in camera configuration.

1. Introduction

Over the last decade datasets have played a vital role in a range of research fields, including computer vision, robotics, biology and medicine. Common benchmarks have provided the foundation for the empirical analysis and refinement of techniques. In machine learning, in particular, datasets have enabled not only the basis for learning, but also for the direct comparison of algorithmic performance, allowing the steady refinement and improvement of machine learning techniques.

A wide variety of computer vision datasets are presently available, and several websites exist that curate comprehensive lists of datasets tailored to specific tasks, including image labelling, object classification, and stereo processing [2, 10, 17]. Due to the widespread availability of an ever-increasing amount of data, a recent move to larger datasets is evident. This trend to 'big data' is also visible in industry, where machine learning is increasingly used to provide business intelligence.

While the number and size of datasets is ever increasing, they are still limited in terms of the domains to which they apply, and are prone to bias [20]. In this work we argue that rather than indefinitely increasing the size of datasets, yielding incremental improvements to the functionality they offer, there is a call for increasing the capabilities offered by datasets. Particularly in fields involving active or dynamic vision and interaction with the environment, e.g. robotics and augmented reality, there is a call for datasets to provide richer information.

We argue that it would be of great benefit to these fields to pursue datasets that capture deeper information, for example by allowing changes in camera pose or interaction with the scene. We further argue that, using well-established methods from image-based rendering, the dataset could abstract itself away from the camera that captured it, allowing simulated changes to key camera specifications such as field of view or depth of field, point of view and multiple-camera configurations.

2. Current Datasets and Their Limitations

Current computer vision datasets are enabling a more objective means for comparing a wide variety of algorithms, focusing on specific tasks, such as scene segmentation, object detection and classification. Figure 1 shows a selection of images taken from a few of the commonly known computer vision datasets. By providing a set of pre-recorded images these sets can help overcome commonly occurring problems, when trying to reproduce, verify and compare re-

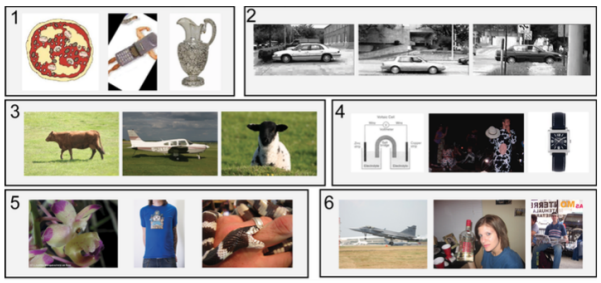

Figure 1. Pictures from commonly known vision datasets.

sults of different algorithms. In a way these datasets abstract the camera-related parameters, lighting and other variables when taking images.

In computer vision, the classification of objects in images is one of the main tasks where datasets have provided benchmarks. Certain images are in fact so commonly used that the researchers themselves can identify the datasets by just looking at a few samples [20]. It has been argued though that such datasets contain bias, and efforts have been made to measure this bias. Furthermore, recent publications have shown how easily current state of the art systems can be "fooled" into misclassification through slight image modification, for example by adding noise [14].

To overcome these biases and to test scalability to larger sizes, artificially created synthetic datasets have been proposed. The ability to provide objective ground truth is another advantage with generated data. Synthetic light field datasets have been used to validate techniques in challenging conditions, including transparencies, occlusions and reflections [21]. Computer games and game engines have been used to create vision datasets. For robotic applications, for example, a synthetic dataset based on 3D models and renderers was created to provided information to find grasp points. This techniques enabled to learn grasping for real objects that are similar in shape to the ones rendered [18]. A limitation of these datasets is generally the lack of texture information, seen in Figure 2.

Over the last years more datasets have been released that contain video data. The difference is that the temporal information can be exploited by the classification algorithm, to detect human actions [5]. Vision datasets have also extended the use of depth information, in e.g. action recognition [15], as the availability of RGB-D cameras has increased. In other areas of research the use of cameras has increased as well, leading to a number of applications in which the camera is in motion during sensing. Augmented reality applications, for example, depend on feature extraction and 3D reconstruction.In robotics, visual SLAM datasets are available [7], providing a joint dataset which contains the sensory input to the agent as well as the action taken – represented in the current state. Recent robotic datasets focus on 3D point clouds and the integration of a wide range of sensors [7, 16]. The recordings are usually taken from pre-defined traverses of mobile robots (or mobile sensors suites). While these provide a way of testing how to perform certain tasks, they are currently not providing a benchmark for how the choice of a specific action can yield change in the sensory input. Especially in robotic vision, where the agent is able to change its environment (or at least the position of the camera), a different kind of comparison needs to be done. Previously, robotics algorithms were evaluated by loading them onto a standard platform running in the same environment, providing realism and some ability to compare their performance [1]. The process was not fully repeatable due to variation in initial robot position and changes in the environment and lighting conditions. At Berkeley the BigBIRD dataset is collected using multiple RGB and depth cameras. The object is placed on a turntable, which is moved in small increments to record different viewpoints (Figure 4). The large dataset contains not just textures but also 3D point clouds.

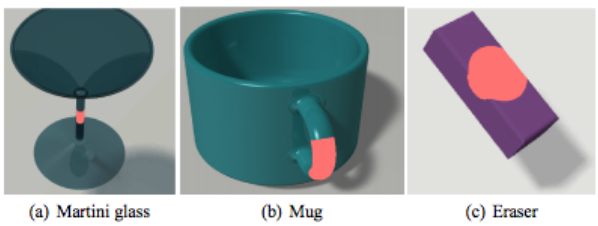

Figure 2. Synthetic pictures from a dataset for grasp learning [18].

3. Propositions

We propose that datasets can be made more useful in a range of applications by combining techniques from image synthesis, augmented reality, and image-based rendering. An appropriate approach has the potential to bolster existing datasets, exploiting recent improvements in sensing and rendering to yield very realistic synthetic data.

3.1. Augmented Reality for Virtual Visualization

For robotic vision datasets, it is desirable that each dataset include spatial and temporal representations along with raw imagery. Generally speaking, datasets should attempt to capture the behaviour of real objects under a wide variety of conditions. The combination of synthetic data and the physical surrounding environment would add valuable insight into the targeted task as well as a uniform validation scenario.

We propose that methods from augmented reality lend themselves well to this need. This technology provides an

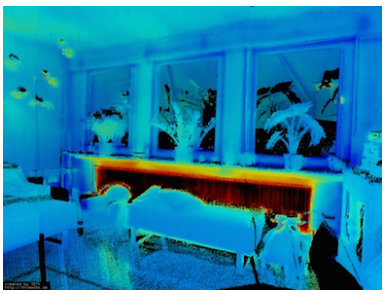

Figure 3. A snapshot from the Robotic 3D Scan Repository providing a point-cloud, which can arbitrarily rotated.

interactive interface for various applications including computer vision and robotics and etc. Besides, this framework provides an opportunity to use the same set up and formulate hypotheses for different robotic vision tasks, interact with the synthetic environment based on the desired task (e.g. adding challenges such as occlusion, illumination variation and etc.) and perform the tests with a clearly repeatable experimental set up. AR seems to decrease the gap of what we can generate or collect and what we can properly and fairly evaluate and analyse. In addition, to cover a broader range of tasks and simulate their challenges, there is a need for a robust multimodal system including multiple sensory channels and a proper sensor fusion algorithm.

3.2. Image-Based Datasets

An evident means by which a dataset might better serve applications employing mobile cameras is in supporting the synthesis of novel views. This might be carried out by estimating 3D textured representations of scenes and objects [11, 12, 13], such that each example in a dataset becomes an interactive object from which arbitrary views can be rendered. A key limitation of this approach, though, is that it stores derived quantities rather than directly measured data. This means that limitations in the 3D modelling algorithm, e.g. in dealing with specular reflection or transparency, become a permanent limitation of the dataset.

It is desirable, then, to store the raw source imagery along with the 3D model, to allow improvements in modelling algorithms to benefit the dataset. This approach was taken by [4], for example. The further possibility arises of foregoing model formation altogether by employing image-based rendering techniques [19, 9]. In such an approach, a scene or object is represented only as a set of images, with no explicit geometric model, and rendering is posed as the inference of the values of rays captured by the virtual camera. Image-based models can be unstructured, allowing cameras to exist at arbitrary locations in the scene [6, 3], or they can be structured as in the case of the light field, which captures a regular grid of camera poses [8]. The advantage of the light field approach is that rendering becomes computationally efficient, requiring only a 4D linear interpolation per pixel, but with the drawback of increased redundancy, ultimately requiring more storage than unstructured light fields.

In addition to rendering from novel poses, both 3D modelling and image-based rendering approaches present the possibility of rendering from novel camera configurations. This includes the ability to simulate cameras with different depths of field, fields of view and resolutions, as well as allowing for stereo, multi-camera rigs, and plenoptic cameras. This powerful capability frees the dataset from the parameters of the camera used to measure it, and allows the evaluation of hardware as well as algorithms.

3.3. Synthetic Datasets for Robotic Vision

Controlling where to place the sensors in the environment to better classify and detect objects is one of the main advantages that robotic vision has over standard computer vision. Building on BigBIRD style datasets we propose to include a standardized viewpoint control capability. A motion input can be used to move the camera virtually in the scene. The new view of the scene can be rendered using state-of-the-art computer game 3D engines. By interpolating from the existing data in the dataset this can be used to create new realistic images for classification.

Furthermore, one approach is to synthesise more complex scenes with multiple objects rendered into the environment. In these cases occlusions occur based on the known objects' geometries and the viewpoint of the camera. By modelling sufficiently accurate representations of the objects offline and allowing the camera position to be externally specified, a more applicable dataset for robotic vision can be created.

4. Discussion

Current datasets provide a standardized testing of computer vision techniques. Such benchmarks though are not yet available for scenarios where the choice of a specific action can yield change in the sensory input. Especially in robotic vision, where the agent is able to change its environment (or at least the position of the camera) a different kind of dataset is needed to allow for an objective comparison.

We propose more dynamic and active datasets, where based on data collected, additional views can be synthetically generated. Datasets like these allow for better reproducibility of results gained with, for example, robotic platforms, in which the camera motion can be controlled. They also provide a more systematic way of comparing techniques, by abstracting specifics of the hardware.

Figure 4. The BigBIRD dataset collection setup.

References

[1] J. Albus, R. Bostelman, T. Hong, T. Chang, W. Shackleford, and M. Shneier. The lagr project integrating learning into the 4d/rcs control hierarchy. In *International Conference in Control, Automation and Robotics (ICINCO)*, 2006.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2015. `http://archive.ics.uci.edu/ml`.

[3] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *28Th Annual Conference on Computer Graphics and Interactive Techniques*, pages 425–432. ACM, 2001.

[4] B. Çalli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols. *CoRR*, abs/1502.03143, 2015.

[5] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[6] A. Davis, M. Levoy, and F. Durand. Unstructured light fields. In *Computer Graphics Forum*, volume 31, pages 305–314. Wiley Online Library, 2012.

[7] A. S. Huang, M. Antone, E. Olson, L. Fletcher, D. Moore, S. Teller, and J. Leonard. A high-rate, heterogeneous data set from the darpa urban challenge. *The International Journal of Robotics Research*, 29(13):1595–1601, 2010.

[8] M. Levoy and P. Hanrahan. Light field rendering. In *23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 31–42, New York, NY, USA, 1996. ACM.

[9] A. Lumsdaine, G. Chunev, and T. Georgiev. Plenoptic rendering with interactive performance using GPUs. In *IS&T/SPIE Electronic Imaging*, pages 829513–829513. International Society for Optics and Photonics, 2012.

[10] G. Mazars. CV Datasets on the web, 2015. `http://www.cvpapers.com/datasets.html`.

[11] R. Newcombe and A. Davison. Live dense reconstruction with a single moving camera. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1498–1505, 2010.

[12] R. Newcombe, A. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136. IEEE, 2011.

[13] R. A. Newcombe, S. Lovegrove, and A. J. Davison. DTAM: dense tracking and mapping in real-time. In *International Conference on Computer Vision (ICCV)*, pages 2320–2327, 2011.

[14] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[15] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[16] T. Peynot, S. Scheding, and S. Terho. The Marulan Data Sets: Multi-Sensor Perception in Natural Environment with Challenging Conditions. *International Journal of Robotics Research*, 29(13), 2010.

[17] H. Riemenschneider. Yet Another Computer Vision Index To Datasets (YACVID), 2014. `http://riemenschneider.hayko.at/vision/dataset/`.

[18] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

[19] H.-Y. Shum, S.-C. Chan, and S. B. Kang. *Image-based rendering*. Springer Science & Business Media, 2008.

[20] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011.

[21] S. Wanner, S. Meister, and B. Goldlücke. Datasets and benchmarks for densely sampled 4d light fields. In *Annual Workshop on Vision, Modeling and Visualization: VMV*, pages 225–226, 2013.