Why Does Data-Driven Beat Theory-Driven Computer Vision?

John K. Tsotsos, Iuliia Kotseruba
York University
tsotsos, yulia_k@eecs.yorku.ca

Alexander Andreopoulos
IBM Research
aandreo@us.ibm.com

Yulong Wu
Aion Foundation
yulong@aion.network

Abstract

This paper proposes that despite the success of deep learning methods in computer vision, the dominance we see would not have been possible by the methods of deep learning alone: the tacit change has been the evolution of empirical practice in computer vision. We demonstrate this by examining the distribution of sensor settings in vision datasets, only one potential dataset bias, and performance of both classic and deep learning algorithms under various camera settings. This reveals a strong mismatch between optimal performance ranges of theory-driven algorithms and sensor setting distributions in common vision datasets.

1. Introduction

There are many classic volumes that define the field of computer vision (e.g., [15]). There, the theoretical foundations of image and video processing, analysis, and perception are developed theoretically and practically, representing what we term theory-driven computer vision. A geometrical and physical understanding of the image formation process, from illuminant to camera optics to image creation, as well as the material properties of the surfaces that interact with incident light was mathematically modeled so that when those equations were simulated by a computer, they would result in the percepts of human vision. It is difficult to deny the theoretical validity of those approaches and from the earliest days of computer vision, the performance of these theory-based solutions had always appeared promising, with much supporting literature (see [14] for early reviews).

However, during most of the history of computer vision, the discipline suffered from two main problems [1]. Firstly, computational power and memory were too meagre to deal with the requirements of vision [16]. Secondly, the availability of large sets of test data that could be shared and could permit replication of results was limited. An empirical methodology and tradition to guide testing and replication was also missing.

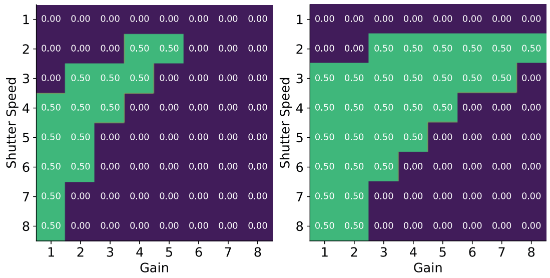
The first problem improved as Moore's Law played out. Especially important, was the advent of GPUs in the late 1990s, with their rapid general availability. Major progress was made on the second problem with the introduction of cheaper memory and the possibility of large collections of images. Whereas the early scarcity of data precluded extensive use of learning methods, the emergence of large image sets encouraged exploration of learning systems. Early papers pointed to the utility of images of handwritten digits for testing recognition and learning methods (e.g., [6]) so the creation of the MNIST set [7] was timely and impactful. The community witnessed the emergence of data-driven computer vision models created by extracting statistical regularities from a large number of image samples. The MNIST set was soon joined by others; PASCAL Visual Object Classes (VOC) Challenge [3], ImageNet [13], and more. The contribution of these data sets and challenges is undeniable towards the acceleration of developments in computer vision.

2. Effect of Sensor Settings for Interest Point and Saliency Algorithms

Previous work explored how performance of several interest point and saliency algorithms changes with varying camera parameters [1]. The experiments revealed a strong dependence on settings. Performance patterns seemed orderly as if determined by some physical law, exhibiting a strong and clear structure.

The authors created a dataset that reflected different cameras, camera settings, and illumination levels (experimental details in [1]). They tested several algorithms (including Harris-Affine and Hessian-Affine region detectors [10]) to reveal the effects of camera shutter speed and voltage gain, under simultaneous changes in illumination, and demonstrated significant differences in their sensitivities.

Figure 1 shows two examples; several others can be seen in [1]. The results show that such algorithms have very specific ranges where good performance can be obtained. Simply put, if one wished to use one of these specific algorithms for a particular application, then it is necessary to ensure that the images processed are acquired using the sensor setting ranges that yield good performance (Figure 1). Such



(a) Harris

(b) Hessian

Figure 1: a) Harris-Affine; b) Hessian-affine. Adapted from [1] showing precision-recall values for combinations of sensor settings (collapsed across illumination conditions). Shutter speed increases from top to bottom and gain increases from left to right. [1] thresholds precision and recall values at 0.5; here, those bins are set to 0.5.

considerations are rarely observed.

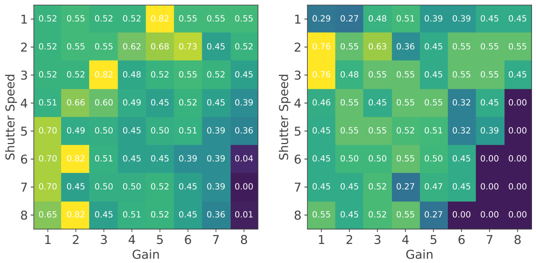
3. Effect of Sensor Settings on Object Detection Algorithms

The same test for more recent recognition algorithms, both classic and deep learning methods, was performed in [19] where experimental details can be found. Four popular object detection algorithms were evaluated including the Deformable Part Models (DPM) [4] and the Bag-of-Words model (BoW) [18], shown in Figure 2; the others can be seen in [19]. Mean average precision (mAP) values were not thresholded and are plotted intact. Although not shown here, [19] also showed that performance depends significantly on illumination level as well as sensor settings and does not easily generalize across these variables. As before, if one wished to use one of these specific algorithms for a particular application, then it is necessary to ensure that the images are acquired using the sensor setting ranges that yield good performance (Figure 2).

In general, it can be seen that there is less orderly structure when compared to the previous set of tests (thus making any characterization of 'good performing' sensor settings more difficult) and the authors wondered about the reason. Could the difference be due to an uneven distribution of training samples along those dimensions? Could overall performance be influenced by such bias?

4. Distributions of Sensor Parameters in Common Computer Vision Datasets

As mentioned, the two above studies caused us to be curious about the reasons behind the uneven and unexpected performance patterns across algorithms. After thorough verifications of the methods employed, we concluded that some imbalance in data distribution across sensing parameters might be the cause. Surprisingly, among works on various biases in vision datasets, few acknowledge the existence of sensor bias (or capture bias [15]) and none provide quantitative statistics.

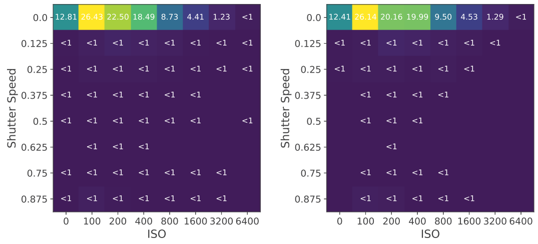


(a) DPM

(b) BoW

Figure 2: Results for 2 object detection algorithms (DPM and BoW) for different shutter speed and gain values for the high illumination condition (adapted from [19]). Shutter speed increases from top to bottom and gain increases from left to right. mAP values are shown for sensor setting combinations.

To explore this further, we selected two common datasets, Common Objects in Context (COCO) [9] and VOC 2007, the dataset used in the PASCAL Visual Object Classes Challenge in 2007 [2]. Since both datasets consist of images gathered from Flickr, we used Flickr API to recover EXIF data (tags for camera settings provided by the camera vendor) for each image. We examined both sets in detail, but present results from COCO here, while those from VOC can be found in [17]. In the COCO dataset 59% and 58% of train and validation data respectively had EXIF data available. We use the `trainval35K` split commonly used for training object detection algorithms.



(a) COCO train set

(b) COCO val set

Figure 3: Distribution of exposure times and ISO in a) training and b) validation sets in terms of % of the total images in the COCO set (with EXIF data). Some of the bins are empty since there are no images in the dataset obtained with those camera settings.

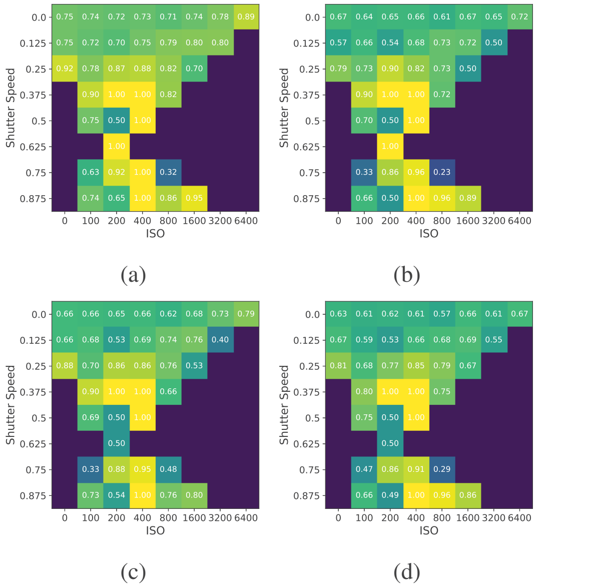
Using shutter speed, f-number and ISO we can compute exposure value (EV) using the formula in [20]. From EV we can derive the illumination level. We define low illumination between -4 and 7 EV (up to 320lx), mid-level illumination between 8 and 10 EV (640 to 2560lx) and high-level illumination above 11 EV (more than 5120lx) which approximately matches the setup in [19]. The distributions of exposure times (shutter speeds) shows that 'auto settings' on cameras dominate (see [17]). We also tabulated the image counts in each illumination level, not surprisingly, nearly 90% of all images are acquired under high to medium illumination conditions (see [17]).

5. Object Detection on Images With Different Sensor Parameters from COCO Dataset

We next investigated how different sensor parameters affect the performance of object detection algorithms, namely, Faster R-CNN [12], Mask R-CNN [5], YOLOv3 [11] and RetinaNet [8], state-of-the-art object detection algorithms trained on COCO `trainval35K` set. Figures 3a and 3b show the percentages of images for a range of the shutter speed and ISO settings in COCO train and validation sets. The bin edges of heatmaps approximately match the ranges reported in [1] and [19]. Since shutter speed in the previous studies was limited to 1s, in our setup all images with exposure time $>$ 1s fall into the last bin and exposure time values between 0 and 1s are split into 8 equal intervals. Both [1] and [19] report gain, which is not available on most consumer cameras, therefore we use ISO values as a proxy. The following ISO bin ranges [0, 100, 200, 400, 800, 1600, 3200, 6400, 10000] approximately correspond to the gain values used in [1] and [19].

Figure 4 shows evaluation results in terms of mean average precision (mAP) for object detection algorithms trained on COCO and evaluated on the portion of COCO 5K minival set with available EXIF data and presented in the same style as the previous tests. However, it is difficult to compare our results with the results of the previous works directly because of the differences in the evaluation datasets, algorithms (interest point vs object detection), camera parameters (gain vs ISO), inability to precisely establish illumination level in common vision datasets and possible inconsistencies in computing average precision in each case.

Note that nearly 90% of training and validation data in COCO is concentrated in the top row of the diagram (very short exposure times and ISO values of up to 800). Figure 4 reveals very similar results from all 4 algorithms that are trained on this dataset suggesting possible training bias. It is also apparent that the mAP values in the top row are consistent with the reported performance of the algorithms but fluctuate wildly in bins that contain less representative camera parameter ranges. It is hard to attribute this fluctuation entirely to sensor bias, as other factors may be at play (e.g. types and number of objects, small number of images in the underrepresented bins). This should be investigated further. It is never a useful property for an algorithm to display such significant sensitivity to small parameter changes. One might expect a small shift in shutter speed, for example, to lead to only small changes in subsequent detection performance; this experiment shows this is not the case.



(a)

(b)

(c)

(d)

Figure 4: Performance of a) YOLOv3, b) Faster R-CNN, c) Mask R-CNN and d) RetinaNet on `minival` subset of COCO for different sets of shutter speed and gain values. mAP for bounding boxes at 0.5 IoU ($mAP^{IoU=.50}$) is computed for each bin using COCO API.

6. Discussion

This topic deserves far more discussion but in this short paper, space will not permit (but see [17]). First, theory-based algorithms seem to have an orderly pattern of performance with respect to the sensor settings we examined. This may be due to their analytic definitions; they were not designed to be parameterized for the full range of sensor settings. If good performance is sought from any of these algorithms, they should be employed with cameras set to the algorithm's inherent optimal ranges. Second, the same test on modern algorithms reveals a haphazard performance pattern. It might be that some of the variations are due to biases in the data, maybe some due to the particular objects in question, others may be due to the properties of the network architectures. This needs more extensive analysis. Third, an examination of two popular image sets, VOC2007 and COCO, shows that image metadata (sensor settings, camera pose, illumination, etc.) is often not available. This means that for any given "in the wild" set of images, the performance of data-driven methods may be predicted by how well the distribution of images along dimensions of sensor setting and illumination parameters of a test set matches the distribution resulting from the training set. This requires further verification. Finally, as can be seen in Figure 3, the variability required to train is not even available in the large datasets we considered. The distributions of images across these parameters was uneven so training algorithms are impeded with respect to learning the variations. Any expectation of generalization may be misplaced. It might be good practice to require specification of image distributions across relevant parameters in order to ensure that not only training, but evaluations, are properly performed.

With all due respect to all the terrific advances made in computer vision, we propose here and provide some justification, that the empirical methodology that led to the turning point in the discipline was based on an oversight that none of us noticed at the time. Sensor settings matter and each algorithm, perhaps most especially the theory-driven ones, have ranges within which one might expect good performance and ranges where one should not expect it. Testing outside the ranges is unfair and inappropriate.

The evolution of our discipline's empirical methodology may need a corrective push. If sensor settings (maybe also illumination levels or other variables) had been properly accounted for in the large scale testing of theory-driven algorithms, perhaps they would have performed at higher levels. In comparing the data-driven with theory-driven algorithms, the distribution of camera settings favored the data-driven algorithms because they were trained on such a random distribution while the theory-driven algorithms were tested on data for which they were not designed to operate. But no one realized this at the time. Thus the empirical strategy favored data-driven models.

A sound empirical method involves the use of objective, quantitative observation in a systematically controlled, replicable situation, in order to test or refine a theory. At the very least, a discussion on how to firm up empiricism in computer vision needs to take place.

References

[1] A. Andreopoulos and J. K. Tsotsos. On sensor bias in experimental methods for comparing interest-point, saliency, and recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):110–126, 2011. 1, 2, 3

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2

[3] M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, et al. The 2005 PASCAL Visual Object Classes Challenge. In *Machine Learning Challenges Workshop*, pages 117–176, 2005. 1

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2009. 2

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 3

[6] G. E. Hinton, C. K. Williams, and M. D. Revow. Adaptive elastic models for hand-printed character recognition. In *Advances in Neural Information Processing Systems*, pages 512–519, 1992. 1

[7] Y. LeCun. The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/, 1998. 1

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 2

[10] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 1

[11] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 3

[13] 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. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[14] S. C. Shapiro. *Encyclopedia of Artificial Intelligence, Second Edition*. John Wiley and Sons, 1992. 1

[15] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, pages 37–55. Springer, 2017. 1, 2

[16] J. K. Tsotsos. The complexity of perceptual search tasks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 89, pages 1571–1577, 1989. 1

[17] J. K. Tsotsos, I. Kotseruba, A. Andreopoulos, and Y. Wu. A possible reason for why data-driven beats theory-driven computer vision. *arXiv preprint arXiv:1908.10933*, 2019. 2, 3

[18] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 2

[19] Y. Wu and J. Tsotsos. Active control of camera parameters for object detection algorithms. *arXiv preprint arXiv:1705.05685*, 2017. 2, 3

[20] D. Wueller and R. Fageth. Statistic analysis of millions of digital photos. In *Digital Photography IV*, volume 6817, page 68170L, 2008. 2