Object Recognition, Computer Vision, and the Caltech 101: A Response to Pinto et al.

Original Article

Why is Real-World Visual Object Recognition Hard?

Object Recognition, Computer Vision, and the Caltech 101: A Response to Pinto et al.

Posted by **jmutch** on **23 Mar 2008** at **20:43 GMT**

Object Recognition, Computer Vision, and the Caltech 101: A Response to Pinto et al.

Yann LeCun, David G. Lowe, Jitendra Malik, Jim Mutch, Pietro Perona, and Tomaso Poggio

Readers of the recent paper "Why is Real-World Visual Object Recognition Hard?" [8] who are unfamiliar with the literature on computer vision are likely to come away with the impression that the problem of making visual recognition invariant with respect to position, scale, and pose has been overlooked. We would therefore like to clarify two main points.

(1) The paper criticizes the popular Caltech 101 benchmark dataset for not containing images of objects at a variety of positions, scales, and poses. It is true that Caltech 101 does not test these kinds of variability; however, this omission is intentional. Techniques for addressing these issues were the focus of much work in the 1980s [11]. For example, datasets like that of Murase and Nayar [6] focused on the problem of recognizing specific objects from a variety of 3d poses, but did not address the issue of object categories and the attendant intra-category variation in shape and texture. Pinto et al.'s synthetic dataset is in much the same spirit as Murase and Nayar's. Caltech 101 was created to test a system [4,3] that was already position, scale, and pose invariant, with the goal of focusing on the more difficult problem of categorization. Its lack of position, scale, and pose variation is stated explicitly on the Caltech 101 website [2], where the dataset is available for download, and is often explicitly restated in later papers that use the dataset (including three of the five cited in Fig. 1). This is not to say that Caltech 101 is without problems. For example, as the authors state, correlation of object classes and backgrounds is a concern, and the relative success of their "toy" model does seem to suggest that the baseline for what is considered good performance on this dataset should be raised.

(2) The paper mentions the existence of other standard datasets (LabelMe [10], Peekaboom [12], StreetScenes [1], NORB [5], PASCAL [7]), many of which contain other forms of variability

such as position, scale, and pose variation, occlusion, and multiple objects. But the authors do not mention that, unlike their "toy" model, most of the computer vision / bio-inspired algorithms they cite do address some of these issues as well, and have in fact been tested on more than one dataset. Thus, many of these algorithms should be capable of dealing fairly well with the "difficult" task of the paper's Fig. 2, on which the authors' algorithm – unsurprisingly – fails. Caltech 101 is one of the most popular datasets currently in use, but it is by no means the sole standard of success on the object recognition problem. See [9] for a recent review of current datasets and the types of variability contained in each.

In conclusion, researchers in computer vision are well aware of the need for invariance to position, scale, and pose, among other challenges in visual recognition. We wish to reassure PLoS readers that research on these topics is alive and well.

References

[1] Bileschi S (2006) StreetScenes: Towards scene understanding in still images. [Ph.D. Thesis]. Cambridge (Massachusetts): MIT EECS.

[2] Caltech 101 dataset (accessed 2008-02-17) Available: http://www.vision.caltech....

[3] Fei-Fei L, Fergus R, and Perona P (2004) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE CVPR 2004, Workshop on Generative-Model Based Vision.

[4] Fergus R, Perona P, Zisserman A (2003) Object Class Recognition by Unsupervised Scale-Invariant Learning. Proc. CVPR 1006: 264-271.

[5] LeCun Y, Huang FJ, and Bottou L (2004) Learning methods for generic object recognition with invariance to pose and lighting. IEEE CVPR 2004: 97–104.

[6] Murase H and Nayar SK (1995) Visual learning and recognition of 3-D objects from appearance. International Journal of Computer Vision, Vol. 14, pp. 5-24.

[7] PASCAL Object Recognition Database Collection, Visual Object Classes Challenge (accessed 2007-12-26) Available: http://www.pascal-network....

[8] Pinto N, Cox DD, and DiCarlo JJ (2008) Why is Real-World Visual Object Recognition Hard? PLoS Computational Biology, 4(1):e27.

[9] Ponce J, Berg TL, Everingham MR, Forsyth DA, Hebert M, Lazebnik S, Marszalek M, Schmid C,

Russell BC, Torralba A, Williams CKI, Zhang J, and Zisserman A (2006) Dataset Issues in Object Recognition. In Toward Category-Level Object Recognition, eds. Ponce J, Hebert M, Schmid C, and Zisserman A, LNCS 4170, Springer-Verlag, pp 29-48.

[10] Russell B, Torralba A, Murphy K, and Freeman WT (2005) LabelMe: a database and web-based tool for image annotation. Cambridge (Massachusetts): MIT Artificial Intelligence Lab Memo AIM-2005–025.

[11] Ullman S (1996) High-Level Vision: Object Recognition and Visual Cognition. MIT press.

[12] Von Ahn L, Liu R, and Blum M (2006) Peekaboom: a game for locating objects in images. ACM SIGCHI 2006: 55–64.

Report a Concern **Respond to this Posting**

See all ongoing discussions on this article