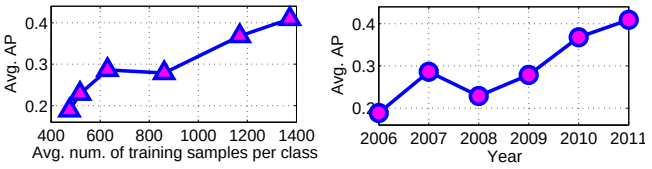
Do We Need More Training Data or Better Models for Object Detection?

Xiangxin Zhu[1]
xzhu@ics.uci.edu
Carl Vondrick[2]
vondrick@mit.edu
Deva Ramanan[1]
dramanan@ics.uci.edu
Charless C. Fowlkes[1]
fowlkes@ics.uci.edu

[1] Computer Science Department
University of California
Irvine, CA, USA

[2] CSAIL
Massachusetts Institute of Technology
Cambridge, MA, USA
(Work performed while at UC Irvine)



Much of the impressive progress in object detection is built on the methodologies of statistical machine learning, which makes use of large training datasets. Consider the benchmark results of the well-known PASCAL VOC object challenge over the past 5 years (above). We see a clear trend in increased performance over the years as methods have gotten better and training datasets have become larger. In this work, we ask a meta-level question about the field: will continued progress be driven faster by increasing amounts of training data or the development of better object detection models?

To answer this question, we collected a massive training set that is an order of magnitude larger than existing collections such as PASCAL [4]. We follow the dominant paradigm of scanning-window templates trained with linear SVMs on HOG features [1, 2, 5, 6], and evaluate detection performance as a function of the amount of positive training data (N) and the model complexity (K), where K is measured by the amount of mixture components capturing variations in object sub-categories, 3D viewpoint, etc.

We found there is a surprising amount of subtlety in scaling up training data sets in current systems. For a given model, one would expect performance to generally increase with the amount of data, but eventually saturate. Empirically, we found the bizarre result that off-the-shelf implementations often decrease in performance with additional data! One would also expect that to take advantage of additional training data N, it is necessary grow the model complexity K. However, we often found scenarios in which performance was relatively static even as model complexity and training data grew (Fig 2).

In this paper, we offer explanations and solutions for these phenomena. First, we found it crucial to set model regularization as a function of the amount of training data N using cross-validation, a standard technique not typically deployed in current object detection systems. Second, existing strategies for discovering subcategory structure, such as clustering aspect ratios [5] and appearance features [3] may not suffice. We found this was related to the inability of classifiers to deal with "polluted" data when mixture labels were improperly assigned (Fig. 3). Increasing model complexity K is thus only useful when mixture components capture the "right" sub-category structure (Fig. 4). Finally, we found that is was easier to capture the "right" structure with compositional representations; we show that one can implicitly encode an exponentially-large K by composing parts together, yielding substantial performance gains over explicit mixture models (Fig.5). We conclude that there is currently little benefit to simply increasing training dataset sizes. But there may be significant room to improve current representations and learning algorithms, even when restricted to existing feature descriptors.

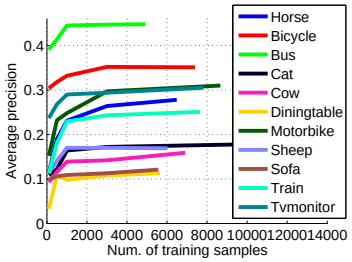


Figure 2: We plot the best performing mixture-models at varying amount of training data for 11 PASCAL categories. All the curves saturate with a relatively small amount of training data. In this work, we analyze how these apparant limits on performance can be broken.

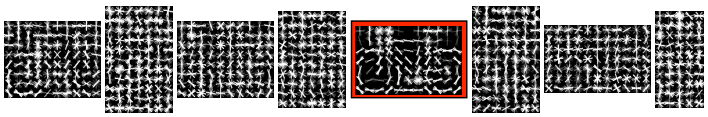


Figure 3: A single clean `bicycle` template (marked with red) alone achieves ap=29.4%, which is almost equivalent to the performance of using all 8 mixtures (ap=29.7%). This suggests that scaling up model complexity by simply adding additional mixture components may not suffice. Both models strongly outperform a single-mixture model trained on the full training set. This suggests that SVMs are sensitive to noisy examples, and one should train with "clean" data that does not pollute a template.

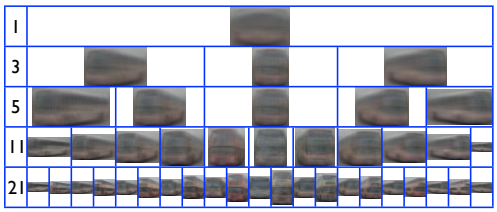


Figure 4: We describe supervised methods for hierarchically structuring data. In this case, we learn separate mixture components corresponding to `bus` viewpoints and object type (single vs double-decker).

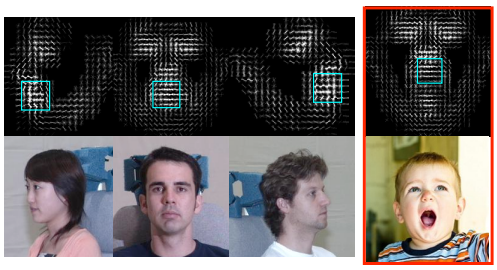


Figure 5: We describe two methods for increasing the performance of mixture models. First, we share spatially-localized regions (the blue "part") between mixture components, shown on the [left]. Second, we allow parts to be composed in novel spatial arrangements not seen in the training data [right]. These modifications define a spectrum of representations between classic mixture models and deformable part models.

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*.

[3] S. Divvala, A. Efros, and M. Hebert. How important are deformable parts in the deformable parts model? *CoRR*, abs/1206.3714, 2012.

[4] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010.

[6] T. Malisiewicz, A. Gupta, and A.A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.