Int J Comput Vis (2016) 119:1–2
DOI 10.1007/s11263-016-0914-5

Guest Editorial: Big Data

Alyosha Efros[1] · Antonio Torralba[2]

Published online: 14 June 2016

Computer vision has a split personality. Within the same field, and largely guided by the same set of fundamental algorithms, it combines two problems that are utterly disparate in their aims and philosophy—here we will call them "Vision as Measurement" and "Vision as Understanding". Measurement problems deal with obtaining objective, quantifiable information about the physical world (e.g. scene depth in meters, visual angle in radians, light-source brightness in candelas-per-meter-squared, etc.). Measurement problems are akin to physics—they are well-posed and the validity of a solution can always be tested with an experiment. Employing careful physical or geometric modeling and rigorous mathematics, this area has been quite successful in solving a number of important problems, such as stereo and structure-from-motion.

Vision as Understanding, on the other hand, has much more to do with psychology and philosophy than physics and mathematics. The goals are defined not in terms of objective quantities, but as subjective, observer-centric tasks. Implicit in tasks such as "find a table in the image" are much deeper issues involving the notion of what is meant by "table", which could vary across cultures, contexts, and even individual observers. Because of this, approaches based on concise models and elegant mathematics, that proved so successful at Vision as Measurement, have largely been found unhelpful for Vision as Understanding.

It's worth noting that this difficulty in describing natural phenomena in terms of concise models is present in a number of other well-known problems, such as speech recognition and machine translation. Both have been considered extremely hard research problems only a decade ago, and yet today both have reached the level of maturity suitable for commercial use. In both cases, the catalyst has been the sudden availability of huge amounts of training data, that, more than anything else, allowed for dramatic improvements in performance.

Recently, the field of computer vision has also been experiencing an extreme makeover (one might say revolution) due to the rapid shift toward much more data-driven methods. Interestingly, this transformation did not happen merely due to the availability of big visual data (after all, datasets of millions of images have been used in computer vision since the mid-2000s). The second, equally-important ingredient has been the rise of high-capacity, computationally-efficient models, primarily convolutional neural networks, that could take advantage of all this big data without underfitting.

The compilation of this special issue has been caught in the middle of this revolution. While the papers presented here have been written before Bastille has been taken, so to speak, they certainly raise many of the issues that made the revolution, in some sense, inevitable. For example, "SUN Database: Exploring a Large Collection of Scene Categories" (doi:10.1007/s11263-014-0748-y) presents large-scale new dataset of visual scenes. Arguing that images need to be understood holistically as a scene, rather than just a collection of individual objects, this work has pushed the cause of big visual data, reinforced the agenda of scene understanding, and has already been influential for training deep features for scenes. But how does one move from a large, disjoint

✉ Antonio Torralba
torralba@mit.edu

Alyosha Efros
efros@eecs.berkeley.edu

[1] Computer Science Division, Electrical Engineering and Computer Science Department, UC Berkeley, Berkeley, USA

[2] Computer Science and Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA

dataset of images to a visually-connected representation? "Joint Inference in Weakly-Annotated Image Datasets via Dense Correspondence" (doi:10.1007/s11263-016-0894-5) explores this question by finding dense correspondences between instances of the same object category in an effort to create visual connections within the vast data. Another way of connecting visual data, this time labelled with unlabelled, is explored in "Large Scale Retrieval and Generation of Image Descriptions" (doi:10.1007/s11263-015-0840-y), which introduces a large-scale graphical model for label propagation. Starting from a set of fully segmented images and lots of unlabeled images, the graphical model is used to propagate the annotations. "Sparse Output Coding for Scalable Visual Recognition" (doi:10.1007/s11263-015-0839-4) deals with situations when the big data also has high cardinality in the number of classes, proposing an efficient sparse-coding scheme. Finally, the paper provocatively named "Do we need more training data?" (doi:10.1007/s11263-015-0812-2) is asking the question: do we really need all this data, when our learning model might not have enough capacity to take advantage of it all? The findings of this work have directly anticipated the advancement of newer, high-capacity models, such as convolutional neural networks.