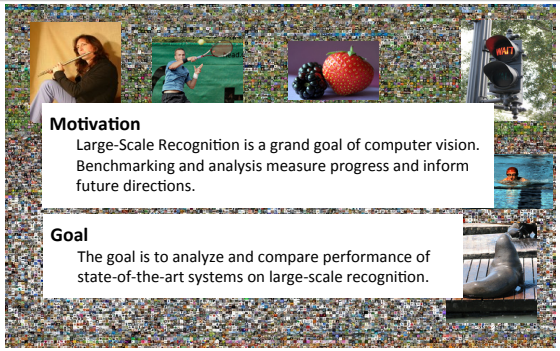
Detecting avocados to zucchinis: what have we done, and where are we going?

Olga Russakovsky[1] Jia Deng[1] Zhiheng Huang[1] Alexander C. Berg[2] Li Fei-Fei[1]

[1] Stanford University
[2] UNC Chapel Hill

Introduction



Motivation

Large-Scale Recognition is a grand goal of computer vision. Benchmarking and analysis measure progress and inform future directions.

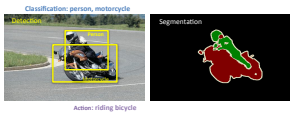
Goal

The goal is to analyze and compare performance of state-of-the-art systems on large-scale recognition.

Dataset

The ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 is much larger and more diverse than previous datasets.

PASCAL VOC 2005-2012



20 object classes **22,591 images**

IMAGENET

http://image-net.org/challenges/LSVRC/{2010,2011,2012,2013}

1000 object classes **1,431,167 images**

Clasification+localization challenge (ILSVRC2012)

Task: To determine the presence and location of an object class.

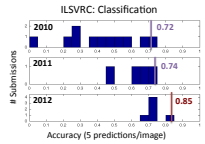
$$\text{Accuracy} = \frac{1}{100{,}000}\sum_{100{,}000 \text{ images}} 1[\text{correct on image } i]$$

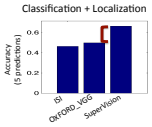
Only one object class is annotated per image (due to high cost of annotation at this scale) so an algorithm is allowed to produce multiple (up to 5) guesses without penalty.

Analysis setup

Why run analysis?

Reason #1: Surprisingly strong performance of the winning entry.





Reason #2: The scale of 1000 object categories allows for an unprecedented look at how object properties affect accuracy of leading algorithms.

State-of-the-art large-scale object localization algorithms

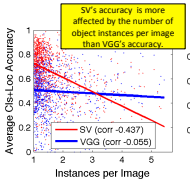
SuperVision (SV) by A. Krizhevsky, I. Sutskever, G. Hinton [1]
- Classification: Deep convolutional neural network; 7 hidden layers, rectified linear units, max pooling, dropout trick, trained with SGD on two GPUs for a week
- Localization: regression on (x, y, w, h)

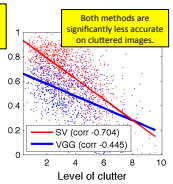
OxfordVGG (VGG) by K. Simonyan, Y. Aytar, A. Vedaldi, A. Zisserman [2]
- Classification: Root-SIFT, color statistics, Fisher vector (1024 Gaussians), product quantization, linear SVM, one-v-rest SVM trained with Pegasos SGD
- Localization: Deformable parts model, root-only

What images are difficult?

Protocol

For every one of the 1000 object categories
- Compute average measure of difficulty on validation images (x)
- Compute accuracy of algorithms on test images (y)





Level of clutter

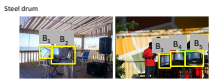
For every image, generate generic object location hypotheses using method of [3] until target object is localized

$$\text{Clutter} = \log_2(\text{average number of windows required})$$

Low clutter => target object is most salient in image
High clutter => object is in a complex image (hard)

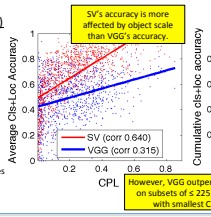
Chance Performance of Localization (CPL)

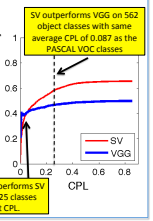
Take all instances of a class on all images: $B_1, B_2, \ldots B_N$



$$\text{CPL} = \frac{\sum_i \sum_{j \neq i} IOU(B_i, B_j) \geq 0.5}{N(N-1)}$$

High CPL => object at the same location/scale in all images
Low CPL => object at varied locations/scales (hard)

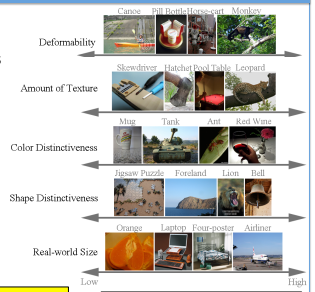




What objects are difficult?

Protocol

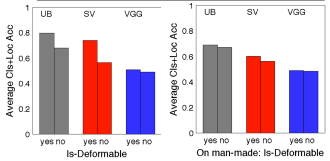
For every one of the 1000 classes
- Ask humans to annotate different properties, e.g., is this object deformable? (x)
- Compute accuracy of algorithms on test images (y)



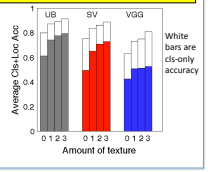
Upper bound (UB)

Optimally combines the output of SV and VGG (using an oracle) to demonstrate the current limit of object localization accuracy.

Deformable objects are much easier for current algorithms to localize, but when considering just man-made objects the effect disappears.



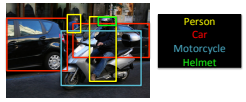
Highly textured objects are much easier for current algorithms to localize (especially for SV).



Where are we going?

- **Cluttered images remain very challenging for object localization**
  - Proposed measure of clutter can be used for creating and evaluating datasets.
- **Untextured and man-made objects are still challenging** even for the best algorithms.
- **Complementary advantages of SV and VGG** can be used to design the next generation of detectors:
  - SV algorithm is very strong at learning object texture, and
  - VGG algorithm is less sensitive to number of instances and object scale.
- **ILSVRC dataset is a promising benchmark for detection algorithms.**

ILSVRC 2013
200 object classes
fully annotated on 60K images
http://image-net.org/challenges/LSVRC/2013



Bibliography

[1] SV details at http://image-net.org/challenges/LSVRC/2012/supervision.pdf and in Krizhevsky et al. NIPS 2012
[2] VGG details at http://image-net.org/challenges/LSVRC/2012/oxford_vgg.pdf and in Sánchez CVPR 2011 and PRL2012, Arandjelović CVPR12, Felzenszwalb PAMI 2012
[3] Alexe, Deselaers, Ferrari. Measuring the objectness of image windows. PAMI 2012