Training object class detectors with click supervision

Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller and Vittorio Ferrari

Train object detectors

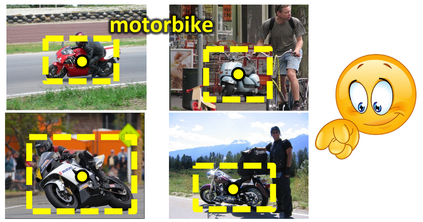
Full supervision: draw bounding boxes

time consuming (35s per box):
ImageNet protocol [Su AAAIW 12]

Weak supervision: image labels

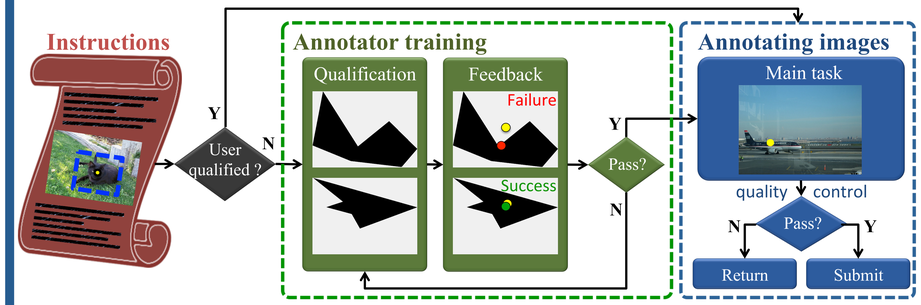
very cheap, but low quality detectors
[Bilen CVPR 16, Cinbis CVPR 14,
Deselaers ECCV 10, Siva ICCV 11]

Ours: center click supervision



- reduce annotation time by 9x-18x
- high quality detectors without ever drawing any bounding-boxes

Crowd-sourcing clicks

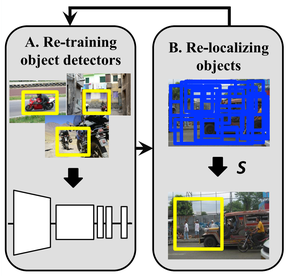


- **Amazon Mechanical Turk**
  - PASCAL VOC 2007 trainval set
  - 14612 clicks for 5011 images
- **Annotation time:** 1.9s per click
- **Average error:** 19.5 pixels
- Just **75$** in total

Clicks available online:

Incorporating clicks into WSOL

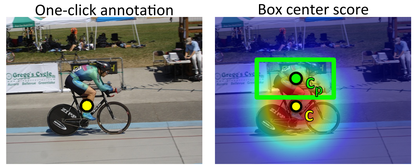
Multiple Instance Learning (MIL)



Initialization: full images [Cinbis CVPR 14, Pandey ICCV 11]

Re-localization: pick proposal with the highest appearance score $S = S_{ap}$

One-click supervision



$$S_{bc}(p; c, \sigma_{bc}) = \mathrm{e}^{-\frac{\|c_p - c\|^2}{2\sigma_{bc}^2}}$$

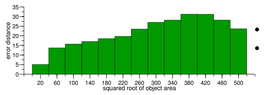
Initialization: *largest proposal centered on click*

Re-localization: *use both appearance and center click*

Two-click supervision

- Estimate object center even more accurately

- Estimate object area from distance between 2 clicks



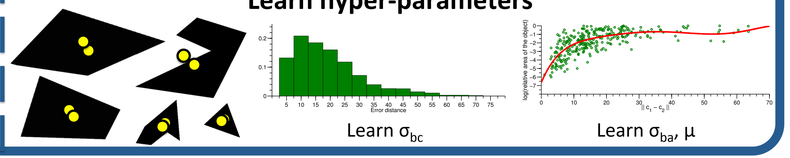
- error distance relates to area
- errors are independent

$$S_{ba}(p; c_1, c_2, \sigma_{ba}) = \mathrm{e}^{-\frac{(a_p - \mu(\|c_1 - c_2\|))^2}{2\sigma_{ba}^2}}$$

Initialization: *largest proposal centered on mean click*

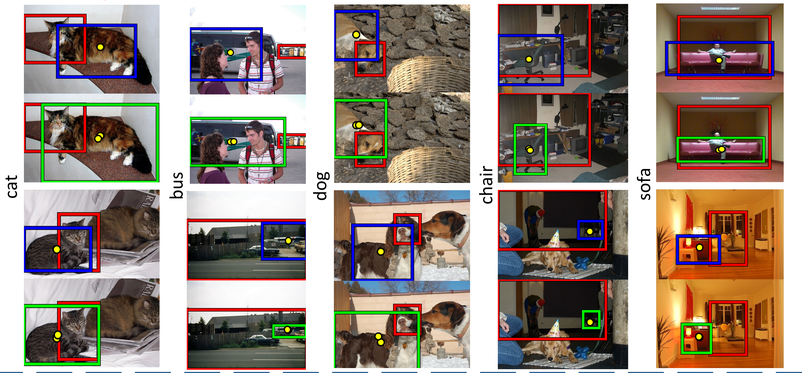
Re-localization: *use appearance, box center and area estimates*

Learn hyper-parameters



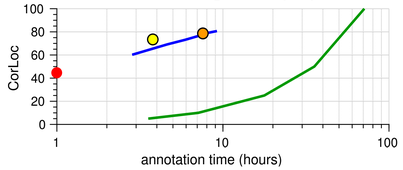
Results

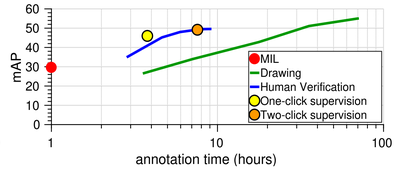
Weak supervision vs 1-click vs 2-click



Quantitative results

- PASCAL VOC 2007
- Fast R-CNN, AlexNet, EdgeBoxes proposals





- substantially better than WSOL at a modest cost
- two-click supervision performs even better
- reduces annotation time by 9x-18x, almost as good as fully supervised
- even better trade-off than human verification [Papadopoulos CVPR16]

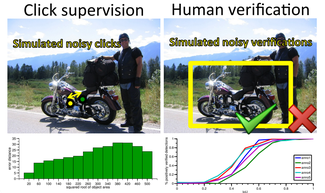
mAP	2-click	Full supervision
AlexNet	49.1	55.5
VGG16	57.5	65.9

90% mAP of full supervision, 9x less human annotation time

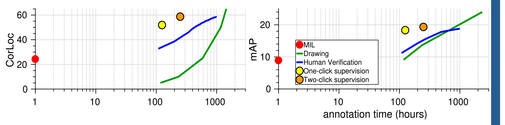
Simulated results on COCO

Create a realistic senario

Click supervision Human verification



- **Training:** training set (80 classes – 82,783 images)
- **Test:** validation set (40,137 images)



Center clicks clearly outperform human verification (3.5x cheaper)