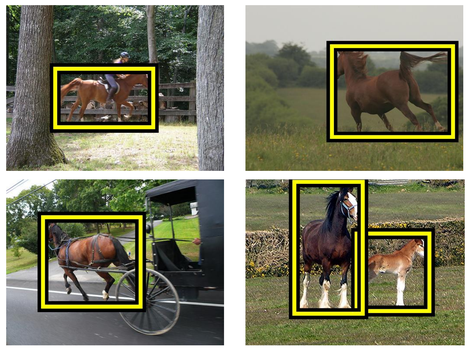
Training Object Class Detectors from Eye Tracking Data

Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller and Vittorio Ferrari

Introduction

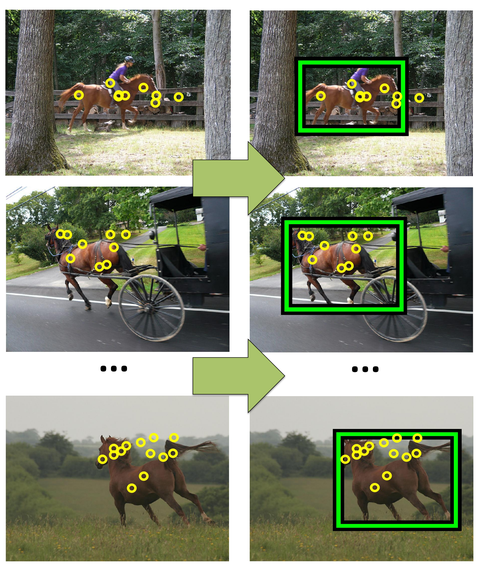
Training object detectors

Current way: draw bounding-boxes



- time consuming (26s-42s per box) [Su AAAI 2012]
- need detailed annotation guidelines

New way: bounding-boxes from eye-tracking data



+ annotation time (1s per image)
+ reduce annotation time by 6.8x
+ simple annotation guidelines
+ correct localizations in half images

Eye tracking dataset

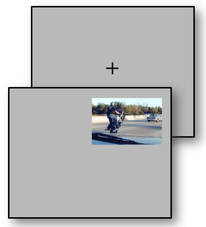
Data

- Large scale (**6270 images**)
- Pascal VOC 12 : *train+val* images of 10 classes
- 5 distinct viewers (28 in total) for each image

- 178,000 fixations (5.7 per viewer per image)
- **Mean response time** = 889 ms/image
- **Fixations on target objects** = 75.2%

Experiment

- **Visual search paradigm**
  - more fixations on target objects
  - faster than free-viewing
- **Pairs of classes**
  - two-alternative forced choice object discrimination
  - pair classes with similar background (e.g. cat, dog)



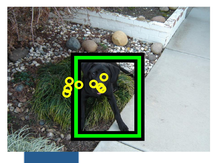
- Add random offset (central bias)
- Random image order

From fixations to bounding-boxes

Bounding-box estimation as figure-ground superpixel labeling

$\mathcal{R}_{bb+fix}$

- small subset (7%)
- learn to predict bounding-boxes from fixations



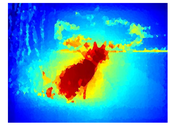
Feature extraction

$\mathcal{R}_{fix}$

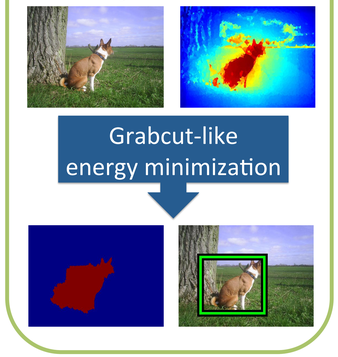
- derive new bounding-boxes from fixations

Initial object segmentation

- Train a superpixel classifier in $\mathcal{R}_{bb+fix}$ (linear SVM + Platt scaling)
- Apply model in $\mathcal{R}_{fix}$ set



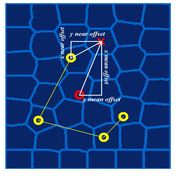
Segmentation refinement



Features

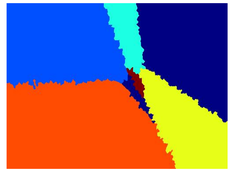
Fixation position

Visual search increases fixations on (or near) the target object



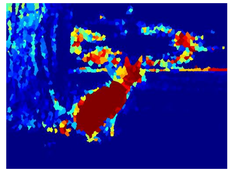
Fixation timing

Timing matters: the longer or the later, the more significant



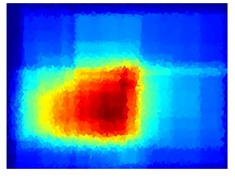
Fixation appearance

Learn color distribution of fg and bg superpixels from fixations

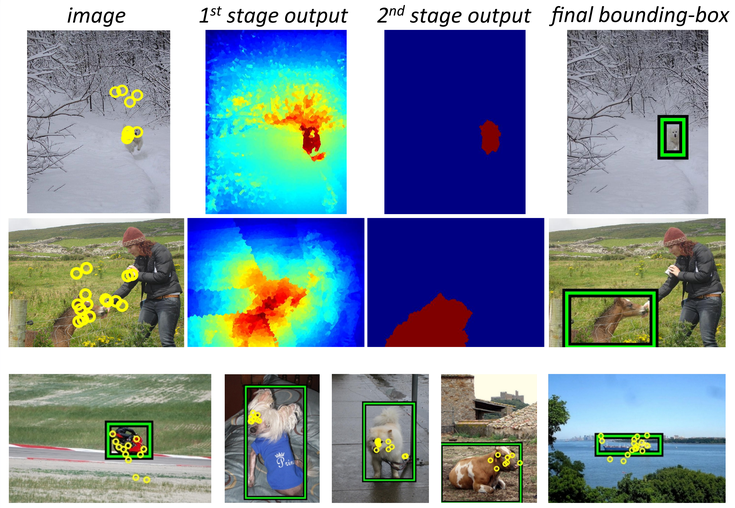


Objectness

Probability that a window contains object of ***any*** class

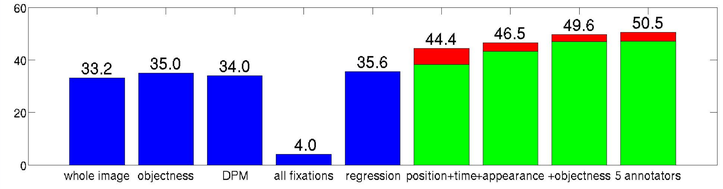


Results



Quantitative results

- 10 Pascal VOC 12 classes, 6270 images in trainval
- Evaluate predicted bounding-boxes in $\mathcal{R}_{fix}$
- Performance: percentage of images with correct predictions



+ all feature types contribute
+ full model outperforms all baselines
+ segmentation refinement always helps (+ 3-5%)

Train DPM detector from fixations [Felzenszwalb PAMI10]

- Pascal VOC 12: train on trainval, test on test set (10991 images)

All ground-truth bounding boxes

mAP = 25.5%

All predicted bounding-boxes

mAP = 12.5%

6.8x fewer ground-truth

mAP = 13.7%

Given same annotation time, get comparable mAP performance