Do we need more training data or better models for object detection?

Charless Fowlkes

with Xiangxin Zhu, Carl Vondrick, Deva Ramanan

Computer Science Department
University of California, Irvine
Object detection

Recognize and localize multiple object categories within an image
Current state of object detection

- PASCAL VOC *detection* challenge provides realistic benchmark of object detection performance.
  - 20 categories, cluttered consumer photographs
- Performance has steadily increased!

![Graph showing the average AP from 2006 to 2011](image-url)

*Graph illustrating the average AP (Avg. AP) from 2006 to 2011.*
Current state of object recognition

- PASCAL VOC detection challenge provides realistic benchmark of object detection performance.
  - 20 categories, cluttered consumer photographs
- Performance has steadily increased!
  .... but so has the amount of training data?
Scanning window classifier detection

Learn classifier $w$

$w^T x > 0$?

pos

neg

Face detection
Rowley, Baluja, & Kanade. CVPR 96
Viola & Jones IJCV 01

Pedestrian detection (and other objects)
Oren et al. CVPR 97
Dalal & Triggs CVPR 05
Felzenswalb et al. PAMI 09
The choice of feature space $X$ imposes fundamental limits on our classification performance.
Ideal Behaviour: Performance saturates at Bayes Risk
Model Bias

Bayes optimal decision boundary
Bayes optimal decision boundary

Linear decision boundary

Model Bias

Fixed model complexity may limit asymptotic performance
Ideal Behaviour: Complexity-Generalization tradeoff

Performance

Model Complexity
Ideal Behaviour: Complexity-Generalization tradeoff

Model Complexity

Under-fitting

Over-fitting
Experiment #1

- Face detection with single HOG template
- Train a linear classifier using SVM

$$\min_{w, \xi, b} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right\}$$

$$y_i (w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

- Vary the number of positive training examples
- Use “hard negative mining” for negative examples
Performance as training set grows

How will performance change with more training data?
Performance as training set grows

Worse performance with more training data ???
Performance as training set grows

![Graph showing average precision vs. number of training samples. The graph has two lines: one for Fixed C=0.002 and another for Crossval on C. The equation for the optimization problem is also shown: \( \min_{w, \xi, b} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right\} \).]
As amount of training data changes, one needs to tune the regularization parameter $C$. 

![Graph showing the relationship between $C$ and average precision for different values of $N$.]
Experiment #2

We want to detect faces at many different viewpoints. What positive training data should we use?

(a) include all viewpoints in training
(b) only train on a subset of views (e.g. frontal faces)
Single template trained with 200 frontal faces outperforms template trained with 800 images that include all views of faces.
Single template trained with 200 frontal faces outperforms template trained with 800 images that include all views of faces.

This holds true for both training and test performance!
SVM is sensitive to outliers

Hinge-loss charges a large cost for impossible to classify training examples

Interestingly, discriminative mixture models sometimes appear to solve this problem by assigning outliers to a “junk” cluster
Learned templates

All views

Frontal views only
Experiment #3

Increase model complexity by using mixture components to model different viewpoints.
Supervised clustering (e.g. using CMU MultiPIE viewpoint annotations)

Unsupervised clustering using k-means to perform recursive splitting

Hierarchical clustering allows varying number of clusters (K) while minimizing effects of sample variance (similar to stratified sampling)
Human-in-the-loop clustering can boost mixture model performance

![Graphs showing average precision for Face and Bus categories]

- **Face**
  - Human cluster, $K=5$
  - Kmeans cluster, $K=4$

- **Bus**
  - Human cluster, $K=5$
  - Kmeans cluster, $K=4$
Bus Category

![Graph showing AP vs. Number of training data for different K values.](image1)

![Graph showing AP vs. Number of mixtures for different N values.](image2)

**Legend:**
- Red: K=1
- Green: K=3
- Blue: K=5
- Cyan: K=11
- Magenta: K=21

**Legend:**
- Red: N=50
- Green: N=100
- Blue: N=500
- Cyan: N=1000
- Magenta: N=1898

**Data Performance:**
- Ideal
  - Model Complexity
    - Data
      - Actual

**Model Complexity:**
- Data
  - Performance
PASCAL 10x Dataset

Collect 10x as much positive training examples as original PASCAL dataset

- Flickr tag query + mechanical turk workers label images
- Dataset is available online for 10 categories

When to mark yes:
- If the image contains at least one horse.
- If the horses are clearly visible through glass.
- If the horses are in a mirror.
- If the horses are in poor lighting, but still visible.
- If the image has a picture of horses as long as they are realistic.

When to mark no:
- If all the horses are toys or photoshopped.
- If the image is about horses but does not contain a horse.
- If every horse is very tiny.
- If the image is taken inside a horse.
- If the image is poor quality or has bad motion blur.
- If the image is a collage or multiple images.
- If you don’t know what the image contains.

Draw a box around each individual horse in this image.
If there are more than 5 horses, then label the 5 largest.
You must read the instructions and examples as we hand review all work.
• Cross validation to choose optimal regularization and # of mixture components for each category

• Performance saturates with ~10 templates per category and ~100 positive training examples per template
Have we reached the Bayes Risk for the HOG feature space?
Deformable part model (DPM)

Represent local part appearance with discriminatively trained templates, connected by “springs” that encode relative locations.

In following experiments, we consider a fully supervised (non-latent) version

*Same input feature space but a richer representation*

[Felzenszwalb, McAllester, Ramaman. 2008]
Alternate view of DPM

Every possible placement of parts “synthesizes” a rigid template
Dynamic programming used in DPM is a fast way to index a very large collection of related rigid templates (each with its own bias)
Why does DPM outperform rigid mixtures?

Part appearance parameters are tied (shared across training instances with different poses)

Model can deform (extrapolate to unseen configurations)
**Experiment #5: Rigid Mixture of Parts**

Rigid Mixture of Parts (RMP):
- Learn part appearance templates from supervised annotations
- One mixture component (template) for each spatial configuration seen during training
- Fast evaluation at test time

Part appearance is shared across training examples but no extrapolation to new configurations
Performance of supervised DPM appears to be due to both shared part appearance and extrapolation to unseen spatial layouts.
State of the art face detection with only 100 training examples

[Zhu & Ramanan, 2012]
Lesson learned

More training data helps (but only if you are careful)

– Set regularization parameter using cross validation

– “Clean” training data and human supervision can help SVM which is sensitive to outliers

– Having the proper correspondence / alignment / clustering of training data can greatly improve model performance

Bigger gains from richer representations

– Mixture of linear classifiers + HOG feature space not necessarily saturated
Thanks!

Xiangxin Zhu  Carl Vondrick  Deva Ramanan  Charless Fowlkes

UCIrvine
University of California, Irvine