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

Xiangxin Zhu
Carl Vondrick
Deva Ramanan
Charless Fowlkes

University of California, Irvine

Appeared in BMVC 2012
Slides adapted from Charless Fowlkes
Motivations
Current state of object recognition

- PASCAL VOC detection challenge provides realistic benchmark of object detection performance.
- Performance has steadily increased!
Current state of object recognition

- PASCAL VOC detection challenge provides realistic benchmark of object detection performance.
- Performance has steadily increased!
  .... but so has the amount of training data??
Bayes Risk

Feature space may limit our ultimate classification performance
Performance saturation

![Graph showing performance saturation](image-url)
Class of models may not be flexible enough
Ideal

Performance

Model Complexity

Actual
Experiments
Experiment #1

• Single Face Template
• HOG Feature vector
• Train a linear classifier using SVM

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

\[
y_i (\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0
\]

• Positive examples + “hard negative” mining
Performance vs #training examples

Single template face model

Average precision

Num. of training samples

0

0.3

0.4

0.5

0.6
Performance vs #training examples

Worse performance with more training data?!?!?
Performance vs #training examples

Single template face model

Average precision

Num. of training samples

Fixed C=0.002

Crossval on C
Need to make cross validation easy for everyday users!

\[
\min_{w,\xi,b} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right\}
\]
<table>
<thead>
<tr>
<th></th>
<th>train Images</th>
<th>train Objects</th>
<th>val Images</th>
<th>val Objects</th>
<th>trainval Images</th>
<th>trainval Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aeroplane</td>
<td>327</td>
<td>432</td>
<td>343</td>
<td>433</td>
<td>670</td>
<td>865</td>
</tr>
<tr>
<td>Bicycle</td>
<td>268</td>
<td>353</td>
<td>284</td>
<td>358</td>
<td>552</td>
<td>711</td>
</tr>
<tr>
<td>Bird</td>
<td>395</td>
<td>560</td>
<td>370</td>
<td>559</td>
<td>765</td>
<td>1119</td>
</tr>
<tr>
<td>Boat</td>
<td>260</td>
<td>426</td>
<td>248</td>
<td>424</td>
<td>508</td>
<td>850</td>
</tr>
<tr>
<td>Bottle</td>
<td>365</td>
<td>629</td>
<td>341</td>
<td>630</td>
<td>706</td>
<td>1259</td>
</tr>
<tr>
<td>Bus</td>
<td>213</td>
<td>292</td>
<td>208</td>
<td>301</td>
<td>421</td>
<td>593</td>
</tr>
<tr>
<td>Car</td>
<td>590</td>
<td>1013</td>
<td>571</td>
<td>1004</td>
<td>1161</td>
<td>2017</td>
</tr>
<tr>
<td>Cat</td>
<td>539</td>
<td>605</td>
<td>541</td>
<td>612</td>
<td>1080</td>
<td>1217</td>
</tr>
<tr>
<td>Chair</td>
<td>566</td>
<td>1178</td>
<td>553</td>
<td>1176</td>
<td>1119</td>
<td>2354</td>
</tr>
<tr>
<td>Cow</td>
<td>151</td>
<td>290</td>
<td>152</td>
<td>298</td>
<td>303</td>
<td>588</td>
</tr>
<tr>
<td>Diningtable</td>
<td>269</td>
<td>304</td>
<td>269</td>
<td>305</td>
<td>538</td>
<td>609</td>
</tr>
<tr>
<td>Dog</td>
<td>632</td>
<td>756</td>
<td>654</td>
<td>759</td>
<td>1286</td>
<td>1515</td>
</tr>
<tr>
<td>Horse</td>
<td>237</td>
<td>350</td>
<td>245</td>
<td>360</td>
<td>482</td>
<td>710</td>
</tr>
<tr>
<td>Motorbike</td>
<td>265</td>
<td>357</td>
<td>261</td>
<td>356</td>
<td>526</td>
<td>713</td>
</tr>
<tr>
<td>Person</td>
<td>1994</td>
<td>4194</td>
<td>2093</td>
<td>4372</td>
<td>4087</td>
<td>8566</td>
</tr>
<tr>
<td>Pottedplant</td>
<td>269</td>
<td>484</td>
<td>258</td>
<td>489</td>
<td>527</td>
<td>973</td>
</tr>
<tr>
<td>Sheep</td>
<td>171</td>
<td>400</td>
<td>154</td>
<td>413</td>
<td>325</td>
<td>813</td>
</tr>
<tr>
<td>Sofa</td>
<td>257</td>
<td>281</td>
<td>250</td>
<td>285</td>
<td>507</td>
<td>566</td>
</tr>
<tr>
<td>Train</td>
<td>273</td>
<td>313</td>
<td>271</td>
<td>315</td>
<td>544</td>
<td>628</td>
</tr>
<tr>
<td>Tvmonitor</td>
<td>290</td>
<td>392</td>
<td>285</td>
<td>392</td>
<td>575</td>
<td>784</td>
</tr>
</tbody>
</table>
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)
Worse performance with more training data?!?!?
Single template trained with 200 “clean” frontal faces outperforms template trained with 800 images that include all views of faces.

This holds true for both training and test performance.
Learned templates

All views

Frontal views only
SVM is sensitive to outliers

\[
\begin{align*}
\min_{w,\xi,b} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right\} \\
y_i(w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0
\end{align*}
\]

All has lower training objective, but higher 0-1 loss!
Experiment #3

Increase model complexity by using mixture components to model different viewpoints.
Model wider range of variability by using a “mixture” of rigid templates
Discriminative clustering uses mixture components to take care of outliers?
Human supervised clustering

<table>
<thead>
<tr>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>13</td>
</tr>
</tbody>
</table>
Human-in-the-loop clustering can boost mixture model performance

![Face clustering results and graph](image-url)
Human-in-the-loop clustering can boost mixture model performance

![Diagram showing average precision vs. number of training samples for human and Kmeans clusters, K=5 and K=4, respectively.](image)
Bus Category

(c) Bus: N vs. AP

(d) Bus: K vs. AP

Number of training data

Number of mixtures

K=1
K=3
K=5
K=11
K=21

N=50
N=100
N=500
N=1000
N=1898

Tuesday, September 11, 12
PASCAL 10x Dataset

• Collected 10 times as much positive training data as original PASCAL dataset

• Collect images from Flickr, MTurk users label images
• 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
Experiment #4: Have we reached Bayes Risk for linear classifiers with HOG features?
Represent local part appearance with templates, connected by “springs” that encode relative locations. Trained using SVM.

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

Every placement of parts “synthesizes” a rigid template

Dynamic programming used in DPM is a fast way to index a very large collection of rigid templates
Why does DPM do better than rigid mixtures?

- Part appearances are shared during training
- Can extrapolate to new unseen configurations
Rigid Part Model (RPM):

- part appearance is learned from training
- only score spatial configurations of parts seen during training
- very fast to test
State of the art face detection with only 100 training examples

DPM with shared parameters

[Zhu & Ramanan, 2012]
Do we need more data?
Do we need more data?

• More training data helps, but only if you are careful
  – “Clean” training data can help SVM which is sensitive to outliers
  – Having the proper correspondence / alignment / clustering can greatly improve model performance
  – Better models might provide “more bang for the buck”
## Dataset Bias: Distributions Match

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Us</th>
<th>2010</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncated</td>
<td>30.8</td>
<td>31.5</td>
<td>15.8</td>
</tr>
<tr>
<td>Occluded</td>
<td>5.9</td>
<td>8.6</td>
<td>7.1</td>
</tr>
<tr>
<td>Jumping</td>
<td>4.0</td>
<td>4.3</td>
<td>15.8</td>
</tr>
<tr>
<td>Standing</td>
<td>69.9</td>
<td>68.8</td>
<td>54.6</td>
</tr>
<tr>
<td>Trotting</td>
<td>23.5</td>
<td>24.9</td>
<td>26.6</td>
</tr>
<tr>
<td>Sitting</td>
<td>2.0</td>
<td>1.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Other</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Person Top</td>
<td>24.8</td>
<td>29.1</td>
<td>57.5</td>
</tr>
<tr>
<td>Person Besides</td>
<td>8.8</td>
<td>10.0</td>
<td>8.6</td>
</tr>
<tr>
<td>No Person</td>
<td>66.0</td>
<td>59.8</td>
<td>33.8</td>
</tr>
</tbody>
</table>