Modern Object Detection

Most slides from Ali Farhadi
## Comparison of Classifiers

### Learning Objective

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Objective Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>$\max \sum \log P(x_{ij}</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>$\max \sum \log(P(y_i</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>$\min \lambda \sum \xi_i + \frac{1}{2} |\theta|$ such that $y_i \theta^T x \geq 1 - \xi_i \forall i, \xi_i \geq 0$</td>
</tr>
<tr>
<td>Kernelized SVM</td>
<td>complicated to write</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>most similar features $\rightarrow$ same label</td>
</tr>
</tbody>
</table>

### Training

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Training Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>$\theta_{ij} = \frac{\sum \delta(x_{ij} = 1 \land y_i = k) + r}{\sum \delta(y_i = k) + Kr}$</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Gradient ascent $\theta^T x &gt; t$</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>Quadratic programming or subgradient opt. $\theta^T x &gt; t$</td>
</tr>
<tr>
<td>Kernelized SVM</td>
<td>Quadratic programming $\sum_i y_i \alpha_i K(\hat{x}_i, x) &gt; 0$</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>Record data $y_i$ where $i = \arg\min_i K(\hat{x}_i, x)$</td>
</tr>
</tbody>
</table>

### Inference

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>$\theta_1^T x + \theta_0^T (1-x) &gt; 0$ where $\theta_1 = \log P(x_{ij} = 1</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>$\theta^T x &gt; t$</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>$\theta^T x &gt; t$</td>
</tr>
<tr>
<td>Kernelized SVM</td>
<td>$\sum_i y_i \alpha_i K(\hat{x}_i, x) &gt; 0$</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>$y_i$ where $i = \arg\min_i K(\hat{x}_i, x)$</td>
</tr>
</tbody>
</table>
Image Categorization

Training
- Training Images
- Training Labels
- Image Features
- Classifier Training
- Trained Classifier

Testing
- Test Image
- Image Features
- Trained Classifier
- Prediction
  - Outdoor
1. Extract fixed-sized (64x128 pixel) window at each position and scale
2. Compute HOG (histogram of gradient) features within each window
3. Score the window with a linear SVM classifier
4. Perform non-maxima suppression to remove overlapping detections with lower scores
• Tested with
  – RGB
  – LAB
  – Grayscale

Slightly better performance vs. grayscale
**Histogram of gradient orientations**

- Votes weighted by magnitude
- Bilinear interpolation between cells

Histograms in 8x8 pixel cells

Orientation: 9 bins (for unsigned angles)
Normalize with respect to surrounding cells

\[ L_2 - \text{norm} : v \rightarrow \frac{v}{\sqrt{\|v\|_2^2 + \epsilon^2}} \]
Histograms of Oriented Gradients for Human Detection

# features = 15 × 7 × 9 × 4 = 3780

# orientations

# cells

# normalizations by neighboring cells
Training set
0.16 = w^T x - b

\text{sign}(0.16) = 1

\Rightarrow \text{pedestrian}
Detection examples
Each window is separately classified
What about this one?

Can the model we trained for pedestrians detect the person in this image?
Specifying an object model

Statistical Template in Bounding Box

- Object is some (x,y,w,h) in image
- Features defined wrt bounding box coordinates
When do statistical templates make sense?

Caltech 101 Average Object Images
Deformable objects

Images from Caltech-256

Slide Credit: Duan Tran
Deformable objects

Images from D. Ramanan’s dataset

Slide Credit: Duan Tran
Parts-based Models

Define objects by collection of parts modeled by

1. Appearance
2. Spatial configuration

Slide credit: Rob Fergus
Explicit Models

Hybrid template/parts model

Detections

Template Visualization

root filters
coarse resolution

part filters
finer resolution

deformation models

Felzenszwalb et al. 2008
How to model spatial relations?

- Many others...

- **O(N^6)**
  - Constellation
  - Fergus et al. '03
  - Fei-Fei et al. '03

- **O(N^2)**
  - Star shape
  - Leibe et al. '04, '08
  - Crandall et al. '05
  - Fergus et al. '05

- **O(N^3)**
  - k-fan ($k = 2$)
  - Crandall et al. '05

- **O(N^2)**
  - Tree
  - Felzenszwalb & Huttenlocher '05

- **O(N^2)**
  - Bag of features
  - Csurka '04
  - Vasconcelos '00

- **O(N^3)**
  - Hierarchy
  - Bouchard & Triggs '05

- **O(N^2)**
  - Sparse flexible model
  - Carneiro & Lowe '06

from [Carneiro & Lowe, ECCV’06]
Tree-shaped model
Pictorial Structures Model

Part = oriented rectangle

Spatial model = relative size/orientation

Felzenszwalb and Huttenlocher 2005
Pictorial Structures Model

\[ P(L|I, \theta) \propto \left( \prod_{i=1}^{n} p(I|l_i, u_i) \prod_{(v_i, v_j) \in E} p(l_i, l_j|c_{ij}) \right) \]

Appearance likelihood

Geometry likelihood
Part representation

- Background subtraction
Pictorial Structures

- Model is represented by a graph $G = (V, E)$.
  - $V = \{v_1, \ldots, v_n\}$ are the parts.
  - $(v_i, v_j) \in E$ indicates a connection between parts.
- $m_i(l_i)$ is the cost of placing part $i$ at location $l_i$.
- $d_{ij}(l_i, l_j)$ is a deformation cost.
- Optimal location for object is given by $L^* = (l_1^*, \ldots, l_n^*)$,

$$L^* = \arg\min_L \left( \sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right)$$
Results for person matching
Results for person matching
Enhanced pictorial structures
Deformable Latent Parts Model

Useful parts discovered during training

Detections

Template Visualization

root filters
coarse resolution

part filters
finer resolution

deforation models

Felzenszwalb et al. 2008
Score = $F_0 \cdot \Phi(p_0, H) + \sum F_i \cdot \Phi(p_i, H) - \sum d_i \cdot \Phi_d(x, y)$

\[
\left( \sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right)
\]
State-of-the-art Detector: Deformable Parts Model (DPM)

1. Strong low-level features based on HOG
2. Efficient matching algorithms for deformable part-based models (pictorial structures)
3. Discriminative learning with latent variables (latent SVM)

Person model

root filters  
coarse resolution

part filters  
finer resolution

deformation models
Person detections

high scoring true positives

high scoring false positives (not enough overlap)
Car

- root filters (coarse resolution)
- part filters (finer resolution)
- deformation models
Car detections

high scoring true positives

high scoring false positives
Cat

root filters
coarse resolution

part filters
finer resolution

deformation
models
Cat detections

high scoring true positives

high scoring false positives (not enough overlap)
Person riding horse
Person riding bicycle
Structure