CS294-43: Visual Object and Activity Recognition

Prof. Trevor Darrell
Spring 2009

March 3rd, 2009
Last Week– Voting, Hashing, and Random Forest techniques


Today – (More) Discriminative approaches (SVM, HCRF)

- Classic SVM on “bags of features”:

- ISM + SVM + Local Kernels:
  M. Fritz; B. Leibe; B. Caputo; B. Schiele: Integrating Representative and Discriminant Models for Object Category Detection, ICCV'05, Beijing, China, 2005 [M. Fritz]

- Local SVM:

- “Latent” SVM with deformable parts:

- Hidden Conditional Random Fields:
But first...

Some hints from Subransu Maji’s latest work on discriminative voting....
Generative vs. Discriminative

“Model the world”

“Model the decision”
Today – Discriminative approaches


• M. Fritz; B. Leibe; B. Caputo; B. Schiele: Integrating Representative and Discriminant Models for Object Category Detection, ICCV'05, Beijing, China, 2005


Basic recognition flow

- Detect or sample features
  - List of positions, scales, orientations

- Describe features
  - Associated list of d-dimensional descriptors

Index each one into pool of descriptors from previously seen images

Quantize to form bag of words vector for the image

Compute match with another image

K. Grauman, B. Leibe
Detect or sample features
List of positions, scales, orientations

Describe features
Associated list of d-dimensional descriptors

Quantize to form bag of words vector for the image

SVM
SVM Review...
Separable by a hyperplane in 2-d:
Which one?
Linear SVM Classifier

Data: \( \{ x_i, y_i \} \quad i = 1, 2, 3 \ldots N \quad y_i = \{-1, +1\} \)

Discriminant: \( f(x) = (w \cdot x + b) > 0 \)

minimize \( \| w \| \)

subject to \( y_i (w \cdot x_i + b) > 1 \quad \text{for all } i \)

Solution: QP gives \( \{ \alpha_i \} \)

\( w_{opt} = \sum \alpha_i y_i x_i \)

\( f(x) = \sum \alpha_i y_i (x_i \cdot x) + b \)
Non-separable by a hyperplane in 2-d
Non-separable by a hyperplane in 2-d
Separable by a hyperplane in 3-d
Embedding

Figure 1.6  The idea of SVMs: map the training data into a higher-dimensional feature space via $\Phi$, and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in input space. By the use of a kernel function (1.2), it is possible to compute the separating hyperplane without explicitly carrying out the map into the feature space.
Kernels

• linear classifier:

\[ f(x) = \text{sign}(w^T x + b) \]

• Kernel classifier:

\[ K(u, v) = \Phi(u) \cdot \Phi(v) \]

\[ f(x) = \text{sign}(\sum_i y_i \alpha_i K(x, x_i) + b). \]

[Dance et al.]
Figure 1.7 Example of an SV classifier found using a radial basis function kernel $k(x, x') = \exp(-||x - x'||^2)$ (here, the input space is $\mathcal{X} = [-1, 1]^2$). Circles and disks are two classes of training examples; the middle line is the decision surface; the outer lines precisely meet the constraint (1.25). Note that the SVs found by the algorithm (marked by extra circles) are not centers of clusters, but examples which are critical for the given classification task. Gray values code $|\sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b|$, the modulus of the argument of the decision function (1.35). The top and the bottom lines indicate places where it takes the value 1 (from [471]).
Example Kernel functions

• Polynomials
• Gaussians
• Sigmoids
• Radial basis functions
• Local feature Kernels (c.f. Fritz et al. and correspondence Kernels in next lecture)
Tried linear, quadratic, cubic; linear had best performance....

$$K(\text{[Image 1]}, \text{[Image 2]}) = K(\text{[Image 3]}, \text{[Image 4]}) = \langle \text{[Bar Chart 1]}, \text{[Bar Chart 2]} \rangle$$

[Dance et al.]
Fig. 5. Images correctly classified containing multiple objects of the same category.

Table 2. Confusion matrix and mean rank for SVM ($k=1000$, linear kernel).

<table>
<thead>
<tr>
<th>True classes</th>
<th>faces</th>
<th>buildings</th>
<th>trees</th>
<th>cars</th>
<th>phones</th>
<th>bikes</th>
<th>books</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>faces</strong></td>
<td>98</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>34</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td><strong>buildings</strong></td>
<td>1</td>
<td><strong>63</strong></td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td><strong>trees</strong></td>
<td>1</td>
<td>10</td>
<td><strong>81</strong></td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td><strong>cars</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td><strong>85</strong></td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><strong>phones</strong></td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td><strong>55</strong></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>bikes</strong></td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td><strong>91</strong></td>
<td>9</td>
</tr>
<tr>
<td><strong>books</strong></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td><strong>73</strong></td>
</tr>
</tbody>
</table>

Mean ranks: 1.04, 1.77, 1.28, 1.30, 1.83, 1.09, 1.39

Table 1. Confusion matrix and the mean rank for the best vocabulary ($k=1000$).

<table>
<thead>
<tr>
<th>True classes</th>
<th>faces</th>
<th>buildings</th>
<th>trees</th>
<th>cars</th>
<th>phones</th>
<th>bikes</th>
<th>books</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>faces</strong></td>
<td>76</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td><strong>buildings</strong></td>
<td>2</td>
<td><strong>44</strong></td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><strong>trees</strong></td>
<td>3</td>
<td>2</td>
<td><strong>80</strong></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><strong>cars</strong></td>
<td>4</td>
<td>1</td>
<td>0</td>
<td><strong>75</strong></td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>phones</strong></td>
<td>9</td>
<td>15</td>
<td>1</td>
<td>16</td>
<td><strong>70</strong></td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td><strong>bikes</strong></td>
<td>2</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>8</td>
<td><strong>73</strong></td>
<td>0</td>
</tr>
<tr>
<td><strong>books</strong></td>
<td>4</td>
<td>19</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td><strong>69</strong></td>
</tr>
</tbody>
</table>

Mean ranks: 1.49, 1.88, 1.33, 1.33, 1.63, 1.57, 1.57

[Dance et al.]
Today – Discriminative approaches


• M. Fritz; B. Leibe; B. Caputo; B. Schiele: Integrating Representative and Discriminant Models for Object Category Detection, ICCV'05, Beijing, China, 2005


Today – Discriminative approaches


- M. Fritz; B. Leibe; B. Caputo; B. Schiele: Integrating Representative and Discriminant Models for Object Category Detection, ICCV’05, Beijing, China, 2005


Today – Discriminative approaches


• M. Fritz; B. Leibe; B. Caputo; B. Schiele: Integrating Representative and Discriminant Models for Object Category Detection, ICCV'05, Beijing, China, 2005


Discriminatively Trained Mixtures of Deformable Part Models

Pedro Felzenszwalb and Ross Girshick
University of Chicago

David McAllester
Toyota Technological Institute at Chicago

Deva Ramanan
UC Irvine

http://www.cs.uchicago.edu/~pff/latent
PASCAL Challenge

• ~10,000 images, with ~25,000 target objects.
  - Objects from 20 categories (person, car, bicycle, cow, table...).
  - Objects are annotated with labeled bounding boxes.
Model Overview

- Mixture of deformable part models (pictorial structures)
- Each component has global template + deformable parts
- Fully trained from bounding boxes alone
2 component bicycle model

- Root filters
- Part filters
- Coarse resolution
- Finer resolution
- Deformation models
Object Hypothesis

Score of filter is dot product of filter with HOG features underneath it.

Score of object hypothesis is sum of filter scores minus deformation costs.

Multiscale model captures features at two resolutions.
Model

\[ f_w(x) = w \cdot \Phi(x) \]

\[ f_w(x) = \max_z w \cdot \Phi(x, z) \]

\[ Z = \text{vector of part offsets} \]

\[ \Phi(x, z) = \text{vector of HOG features (from root filter & appropriate part sub-windows) and part offsets} \]
Latent SVM

\[ f_w(x) = \max_z w \cdot \Phi(x, z) \]

Linear in \( w \) if \( z \) is fixed

Training data: \((x_1, y_1), \ldots, (x_n, y_n)\) with \( y_i \in \{-1, 1\} \)

Learning: find \( w \) such that \( y_i f_w(x_i) > 0 \)

\[ w^* = \arg\min_w \lambda \|w\|^2 + \sum_{i=1}^{n} \max(0, 1 - y_i f_w(x_i)) \]

Regularization \hspace{2cm} \text{Hinge loss}
Latent SVM training

\[ w^* = \arg\min_w \lambda \| w \|^2 + \sum_{i=1}^{n} \max(0, 1 - y_i f_w(x_i)) \]

- Non-convex optimization
- Huge number of negative examples
- Convex if we fix \( z \) for positive examples
- Optimization:
  - Initialize \( w \) and iterate:
    - Pick best \( z \) for each positive example
    - Optimize \( w \) via gradient descent with data mining
Initializing $w$

- For $k$ component mixture model:
  - Split examples into $k$ sets based on bounding box aspect ratio
  - Learn $k$ root filters using standard SVM
    - Training data: warped positive examples and random windows from negative images (Dalal & Triggs)
  - Initialize parts by selecting patches from root filters
    - Subwindows with strong coefficients
    - Interpolate to get higher resolution filters
  - Initialize spatial model using fixed spring constants
Car model

root filters
coarse resolution

part filters
finer resolution

defformation models
Person model

- Root filters
- Coarse resolution
- Part filters
- Finer resolution
- Deformation models
Bottle model

root filters
coarse resolution

part filters
finer resolution

deformation models
Histogram of Gradient (HOG) features

- Dalal & Triggs:
  - Histogram gradient orientations in 8x8 pixel blocks (9 bins)
  - Normalize with respect to 4 different neighborhoods and truncate
  - 9 orientations * 4 normalizations = 36 features per block
- PCA gives ~10 features that capture all information
  - Fewer parameters, speeds up convolution, but costly projection at runtime
- Analytic projection: spans PCA subspace and easy to compute
  - 9 orientations + 4 normalizations = 13 features
- We also use 2*9 contrast sensitive features for 31 features total
Bounding box prediction

- predict \((x_1, y_1)\) and \((x_2, y_2)\) from part locations
- linear function trained using least-squares regression
Context rescoring

- Rescore a detection using “context” defined by all detections
- Let $v_i$ be the max score of detector for class $i$ in the image
- Let $s$ be the score of a particular detection
- Let $(x_1, y_1), (x_2, y_2)$ be normalized bounding box coordinates
- $f = (s, x_1, y_1, x_2, y_2, v_1, v_2, \ldots, v_{20})$
- Train class specific classifier
  - $f$ is positive example if true positive detection
  - $f$ is negative example if false positive detection
Bicycle detection
More bicycles

False positives
Code

Source code for the system and models trained on PASCAL 2006, 2007 and 2008 data are available here:

http://www.cs.uchicago.edu/~pff/latent
Today – Discriminative approaches


- M. Fritz; B. Leibe; B. Caputo; B. Schiele: Integrating Representative and Discriminant Models for Object Category Detection, ICCV'05, Beijing, China, 2005


Previous Work

Large-scale feature
[e.g. Efros et al., ICCV03]

Local patches
[e.g. Laptev & Perez, ICCV07]

Wang and Mori NIPS 2008
Large vs. Small Scale Features

Wang and Mori NIPS 2008: Explore Hidden-state conditional random field model integrating local features and global template
CRF Part Based Models

Given $n$ pairs $(x_i, y_i)$, learn a model that maps images to object categories (where $x_i$ is an image, $y_i$ is an object category).

$\phi(x_i) \in \mathbb{R}^d$

$x_i \rightarrow \{\phi(x_1), \ldots, \phi(x_m)\}$

Quattoni et al. 2004, 2007 develop Hidden-state CRF model for category recognition: capture inter-part dependencies with a hidden (or ‘latent’) part label…
Quattoni et al. 2004:

Graph Structure

$G(V, E)$
Quattoni et al. 2004:

**CRFs with hidden variables for Object Recognition**

We introduce a hidden variable: $h_i = \{h_1 \ldots h_m\}$, $h_j \in H$ and define the conditional model:

$$P(y, h \mid x; \theta) = \frac{e^{\psi(y, h, x; \theta)}}{\sum_{y', h} e^{\psi(y', h, x; \theta)}}$$

$$P(y \mid x; \theta) = \sum_h P(y, h \mid x; \theta) = \frac{\sum_h e^{\psi(y, h, x; \theta)}}{\sum_{y', h} e^{\psi(y', h, x; \theta)}}$$

$$\psi(y, h, x; \theta) \rightarrow \text{Maps a configuration to} \quad \mathbb{R}$$
Potentials

\[ G(V, E) \text{ is a minimum spanning tree.} \]
Weight \((i, j) = \text{distance between patches } x_i \text{ and } x_j \)

**Compatibility functions**

- **Compatibility between a pair of part labels and a category**
  \[ f^2(j, k, y, h_j, h_k, x; \theta) = \theta(y, h_j, h_k) \]

- **Compatibility between a part label and a category**
  \[ f^1(j, y, h_j, x; \theta) = \phi(x_j) \bullet \theta(h_j) + \theta(y, h_j) \]

- **Compatibility between a patch and a part label**

**Potential function**

\[
\psi(y, h, x; \theta) = \sum_{j \in V} f^1(j, y, h_j, x; \theta) + \sum_{(j, k) \in E} f^2(j, k, y, h_j, h_k, x; \theta)
\]

**Obtained** with Lowe’s detector
SIFT features + relative location and scale
Wang and Mori, NIPS ‘08

• extend Quattoni et al. to include a global descriptor
• develop an efficient initialization scheme similar to Felzenswab et al.
• apply to activities using local and global spatio-temporal features...
Hidden Conditional Random Field with a global feature:

\[ \ell = \sum_t \log p(y^t | x^t) = \sum_t \log \left( \sum_h p(y^t, h | x^t) \right) \]

\( p(y, h | x) \propto \exp(\Psi(y, h, x)) \)

Wang and Mori, NIPS ‘08
Learning a HCRF Model

Wang and Mori, NIPS ‘08
Visualization of Learned Model

Wang and Mori, NIPS ‘08
Results: Weizmann dataset

Wang and Mori, NIPS ‘08
Max-Margin Hidden Conditional Random Fields for Human Action Recognition

Yang Wang and Greg Mori
TR 2008-21
School of Computing Science
Simon Fraser University
{ywang12, mori}@cs.sfu.ca

Abstract

We present a new method for classification with structured latent variables. Our model is formulated using the max-margin formalism in the discriminative learning literature. We propose an efficient learning algorithm based on the cutting plane method and decomposed dual optimization. We apply our model to the problem of recognizing human actions from video sequences, where we model a human action as a global root template and a constellation of several “parts”. We show that our model outperforms another similar method that uses hidden conditional random fields, and is comparable to other state-of-the-art approaches. More importantly, our proposed work is quite general and can potentially be applied in a wide variety of vision problems that involve various complex, interdependent latent structures.
Wang and Mori, CVPR’09:

• Max-margin version extension of NIPS’08
• Similar to LSVMs:
  – semi-convex
  – hinge-loss
• But:
  – inherently multi-class
  – inter-part constraints
  – does not explicitly solve for latent part position
Wang and Mori, CVPR’09:

Figure 2. Confusion matrices of classification results on Weizmann and KTH dataset. The confusion matrix of per-video classification on the Weizmann dataset is not shown, since it is simply a perfect diagonal matrix.
Wang and Mori, CVPR’09:

Table 2. Comparison of classification accuracy with previous work on the Weizmann dataset.

<table>
<thead>
<tr>
<th>method</th>
<th>per-frame</th>
<th>per-video</th>
<th>per-cube</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>0.9311</td>
<td>1.0000</td>
<td>N/A</td>
</tr>
<tr>
<td>Wang &amp; Mori [31]</td>
<td>0.9029</td>
<td>0.9722</td>
<td>N/A</td>
</tr>
<tr>
<td>Jhuang et al. [12]</td>
<td>N/A</td>
<td>0.9880</td>
<td>N/A</td>
</tr>
<tr>
<td>Niebles &amp; Fei-Fei [20]</td>
<td>0.5500</td>
<td>0.7280</td>
<td>N/A</td>
</tr>
<tr>
<td>Blank et al. [3]</td>
<td>N/A</td>
<td>N/A</td>
<td>0.9964</td>
</tr>
</tbody>
</table>

Table 3. Comparison of our approach with the HCRF model on the KTH dataset. The first number in each cell is the accuracy of per-frame classification. The second number is the accuracy of per-video classification.

| method              | $|\mathcal{H}|=6$ | $|\mathcal{H}|=10$ | $|\mathcal{H}|=20$ |
|---------------------|---------|---------|---------|
| HCRF                | 0.6633  | 0.6698  | 0.6444  |
|                     | 0.7855  | 0.8760  | 0.7512  |
| Our approach        | 0.7064  | 0.7853  | 0.7486  |
|                     | 0.8475  | 0.9251  | 0.8966  |

Table 4. Comparison of per-video classification accuracy with previous approaches on the KTH dataset.

<table>
<thead>
<tr>
<th>method</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>0.9251</td>
</tr>
<tr>
<td>Liu &amp; Shah [18]</td>
<td>0.9416</td>
</tr>
<tr>
<td>Jhuang et al. [12]</td>
<td>0.9170</td>
</tr>
<tr>
<td>Wang &amp; Mori [31]</td>
<td>0.8760</td>
</tr>
<tr>
<td>Niebles et al. [21]</td>
<td>0.8150</td>
</tr>
<tr>
<td>Dollár et al. [8]</td>
<td>0.8117</td>
</tr>
<tr>
<td>Schuldt et al. [25]</td>
<td>0.7172</td>
</tr>
</tbody>
</table>
Mar 17th – Correspondence and Pyramid-based techniques