Part-based models


Code available
Discriminative deformable models

mixtures of deformable templates

somewhat simple and somewhat successful approach

False positives due to overlap criteria
Pascal VOC Object models

Training + testing code will be online this week (in time for VOC2009!)
Image features - histograms of gradients (HOG)

- Bin gradients into 9 orientations over 8x8 pixel neighborhoods & normalize
- Dalal & Triggs CVPR05
Scanning-window templates

Dalal and Triggs CVPR05 (HOG)

Papageorgiou and Poggio ICIP99 (wavelets)

\[ f(x) = w^T x \quad f(x) > 0 \]

Train with SVM
How to interpret negative weights?

\[ w^T x > 0 \]

\[ (w_+ - w_-)^T x > 0 \]

\[ w_+^T x > w_-^T x \]

Pedestrian template > Pedestrian background template

\[ w_+, w_- = \text{sum of positive, negative support vectors} \]

Right approach is to compete pedestrian, pillar, doorway... models

Background class is hard to model - easier to penalize particular vertical edges

Discriminative models come “equipped” with their own background
Part-based model (Pictorial Structure)

Root filter ($w_{\text{head}}$) 8x8 resolution
Part filter ($w_{\text{part}}$) 4x4 resolution
Deformation model ($w_{ij}$)

$$E(z, x) = \sum_i w_i \cdot \phi(x, z_i) + \sum_{i,j} w_{ij} \cdot d(z_i, z_j)$$

When $E$ is a tree graph, $\min_l E(l, x)$ computable with dynamic programming
For us, $E$ is a star graph

Rich related work: Fischler & Eschlager, Felzenswalb & Huttenlocher, Fergus et al,
“Cleaner” multiscale model

Part filters are not 4×4, but 8×8 at larger scale in image pyramid
Mixture models

Add a discrete component label to $z$ \( f_w(x) = \min_z E(z, x) \)

Allow for unbounded deformation using distance transform techniques
Example images (run on-the-fly at MSR)
...Consumer photo dataset...
Some statistics
PASCAL VOC 2008

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>pers</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASIA_Det</td>
<td>25.2</td>
<td>14.6</td>
<td>9.8</td>
<td>10.5</td>
<td>6.3</td>
<td>23.2</td>
<td>17.6</td>
<td>9.0</td>
<td>9.6</td>
<td>10.0</td>
<td>13.0</td>
<td>5.5</td>
<td>14.0</td>
<td>24.1</td>
<td>11.2</td>
<td>3.0</td>
<td>2.8</td>
<td>3.0</td>
<td>28.2</td>
<td>14.6</td>
</tr>
<tr>
<td>Jena</td>
<td>4.8</td>
<td>1.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>1.0</td>
<td>1.3</td>
<td>-</td>
<td>0.1</td>
<td>4.7</td>
<td>0.4</td>
<td>1.9</td>
<td>0.3</td>
<td>3.1</td>
<td>2.0</td>
<td>0.3</td>
<td>0.4</td>
<td>2.2</td>
<td>6.4</td>
<td>13.7</td>
</tr>
<tr>
<td>LEAR_PlusClass</td>
<td>36.5</td>
<td>34.3</td>
<td>10.7</td>
<td>11.4</td>
<td>22.1</td>
<td>23.8</td>
<td>36.6</td>
<td>16.6</td>
<td>11.1</td>
<td>17.7</td>
<td>15.1</td>
<td>9.0</td>
<td>36.1</td>
<td>40.3</td>
<td>19.7</td>
<td>11.5</td>
<td>19.4</td>
<td>17.3</td>
<td>29.6</td>
<td>34.0</td>
</tr>
<tr>
<td>MPI_struct</td>
<td>25.9</td>
<td>8.0</td>
<td>10.1</td>
<td>5.6</td>
<td>0.1</td>
<td>11.3</td>
<td>10.6</td>
<td>21.3</td>
<td>0.3</td>
<td>4.5</td>
<td>10.1</td>
<td>14.9</td>
<td>16.6</td>
<td>20.0</td>
<td>2.5</td>
<td>0.2</td>
<td>9.3</td>
<td>12.3</td>
<td>23.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Oxford</td>
<td>33.3</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>29.1</td>
<td>-</td>
<td>12.5</td>
<td>-</td>
<td>-</td>
<td>32.5</td>
<td>34.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UoCTTIUCI</td>
<td>32.6</td>
<td>42.0</td>
<td>11.3</td>
<td>11.0</td>
<td>28.2</td>
<td>23.2</td>
<td>32.0</td>
<td>17.9</td>
<td>14.6</td>
<td>11.1</td>
<td>6.6</td>
<td>10.2</td>
<td>32.7</td>
<td>38.6</td>
<td>42.0</td>
<td>12.6</td>
<td>16.1</td>
<td>13.6</td>
<td>24.4</td>
<td>37.1</td>
</tr>
<tr>
<td>KXRC_Det</td>
<td>26.4</td>
<td>10.5</td>
<td>1.4</td>
<td>4.5</td>
<td>0.0</td>
<td>10.8</td>
<td>4.0</td>
<td>7.6</td>
<td>2.0</td>
<td>1.8</td>
<td>4.5</td>
<td>10.5</td>
<td>11.8</td>
<td>13.6</td>
<td>9.0</td>
<td>1.5</td>
<td>6.1</td>
<td>1.8</td>
<td>7.3</td>
<td>6.8</td>
</tr>
</tbody>
</table>

UoCTTIUCI 1st on 7 classes, 2nd on 8

Test: ~ 1 second / image
Train: ~ 4 hours

All training + testing code online
Model Overview

Model consists of root filter + deformable parts

Training consists of bounding boxes
(part structure learned automatically)
Formal model

\[ E(z, x) = \sum_i w_i \cdot \phi(x, z_i) + \sum_{i,j} w_{ij} \cdot d(z_i, z_j) \]

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

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

Detection: \( f_w(x) > 0 \)
Linear vs. convex models

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

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

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

Given positive and negative training windows \( \{ x_i \} \)

\[ L(w) = \lambda \| w \|^2 + \sum_{i \in \text{pos}} \max(0, 1 - f_w(x_i)) + \sum_{i \in \text{neg}} \max(0, 1 + f_w(x_i)) \]

If \( f_w \) is linear in \( w \), \( L(w) \) is convex (standard SVM)
If \( f_w \) is arbitrary, \( L(w) \) is not convex (in general)
If \( f_w \) is convex in \( w \), \( L(w) \) is “semi-convex”

“semi-convex” - \( L(w) \) is convex if we fix latent values for positives
Coordinate descent

Define $L(w,z_p) = \text{loss when latent } \text{positive} \text{ values fixed to } z_p$

$L(w) = \min_{z_p} L(w,z_p)$

1) $\min_w L(w,z_p)$
Run detector on positives to estimate part locations

2) $\min_{z_p} L(w,z_p)$
Given positive part locations, learn detector with convex program
Part filter initialization

- Learn root filter with SVM
- Initialize part filters to regions in root filter with lots of energy, doubled in spatial resolution

(a)

(b)

(c)
Fig. 9. Some of the models learned on the PASCAL 2007 dataset.
Extensions: add mirror flipping to latent “z”

- Root filters
  - Coarse resolution
- Part filters
  - Finer resolution
- Deformation models
Car detections
Bicycle detections
Fig. 2. Detections obtained with a 2 component bicycle model. These examples illustrate the importance of deformations mixture models. In this model the first component captures sideways views of bicycles while the second component captures frontal and near frontal views. The sideways component can deform to match a "wheelie". The background to find a relatively small number of potential false positives.

A methodology of data mining for hard negative examples was adopted by Dalal and Triggs [76] but goes back at least to the bootstrapping methods used by [9] and [9]. Here we analyze data mining algorithms for SVM and LSVM training. We prove that data mining methods can be made to converge to the optimal model defined in terms of the entire training set.

Our object models are defined using filters that score subwindows of a feature pyramid. We have investigated feature sets similar to HOG [76] and found lower dimensional features which perform as well as the original ones. By doing principal component analysis on HOG features the dimensionality of the feature vector can be significantly reduced with no noticeable loss of information. Moreover, by examining the principal eigenvectors we discover structure that leads to "analytic" versions of low dimensional features which are easily interpretable and can be computed efficiently.

We have also considered some specific problems that arise in the PASCAL object detection challenge and similar datasets. We show how the locations of parts in an object hypothesis can be used to predict a bounding box for the object. This is done by training a model specific predictor using least squares regression. We also demonstrate a simple method for aggregating the output of several object detectors. The basic idea is that objects of some categories provide evidence for or against objects of other categories in the same image. We exploit this idea by training a category specific classifier that rescores every detection of that category using its original score and the highest scoring detection from each of the other categories.

There is a significant body of work on deformable models of various types for object detection, including several kinds of deformable template models, and a variety of part-based models. In the constellation models from parts are constrained to be in a sparse set of locations determined by an interest point operator, and their geometric arrangement is captured by a Gaussian distribution. In contrast, pictorial structure models define a matching problem where parts have an individual match cost in a dense set of locations, and their geometric arrangement is constrained by a set of "springs" connecting pairs of parts. The patchwork of parts model is similar, but it explicitly considers how the appearance model of overlapping parts interact to define a dense appearance model for images.

Our models are largely based on the pictorial structures framework. We use a dense set of possible positions and scales in an image, and define a...
Fig. 4. The matching process at one scale. Responses from the root and part filters are computed at different resolutions in the feature pyramid. The transformed responses are combined to yield a final score for each root location. We show the responses and transformed responses for the “head” and “right shoulder” parts. Note how the “head” filter is more discriminative. The combined scores clearly show two good hypotheses for the object at this scale.