Object Detection with Discriminatively Trained Part Based Models

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan

Presented by
Fabricio Santolin da Silva
Kaustav Basu

Some slides were taken from Prof. Trevor Darrell lecture
Introduction

Problem:
Recognition of generic objects with different shapes, colors or poses
• Cars, People, Bycicles, Animals
Introduction

• Detection System that represents highly variable objects using mixture of multiscale deformable part models.

• Performance gap among simpler models (rigid templates and bag-of-features) and richer models (pictorial structures)
Introduction

- Dalal-Triggs detector
  - HOG filter
  - Sliding box
Problem with bag-of-words

All have equal probability for bag-of-words methods

Location information is important
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  
coarse resolution

part filters  
finer resolution

deformation  
models
Object Hypothesis

Multiscale model captures features at two resolutions

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 plus the bias.
Model

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

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

Z = vector of part offsets

\( \Phi(x, z) \) = vector of HOG features (from root filter & appropriate part sub-windows) and part offsets
Mixture Models
SVM Review…

Separable by a hyperplane in 2-d:
Which one?
Maximum Margin:
Linear SVM Mathematically

- **Goal:** 1) Correctly classify all training data
  \[ wx_i + b \geq 1 \quad \text{if } y_i = +1 \]
  \[ wx_i + b \leq 1 \quad \text{if } y_i = -1 \]
  \[ y_i (wx_i + b) \geq 1 \quad \text{for all } i \]

  2) Maximize the Margin
  \[ M = \frac{2}{||w||} \]
  \[ \frac{1}{2} w^t w \]

- We can formulate a Quadratic Optimization Problem and solve for \( w \) and \( b \)

- **Minimize** \( \Phi(w) = \frac{1}{2} w^t w \)

  subject to \( y_i (wx_i + b) \geq 1 \quad \forall i \)
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 \( y_i f_w(x_i) > 0 \)

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

root filters
coarse resolution

part filters
finer resolution

deformation 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
• 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 highest score of a particular detection

• Let $(x_1, y_1), (x_2, y_2)$ be normalized bounding box coordinates

• $f = (s, x_1, y_1, x_2, y_2, v_1, v_2... , v_{20})$

• Train class specific classifier
  - $f$ is positive example if true positive detection
  - $f$ is negative example if false positive detection
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.
Bicycle detection
More bicycles

False positives
Car
For other information and the implementanition of their code
http://people.cs.uchicago.edu/~pff/latent/
Thank you!