Detection II: Deformable Part Models

Prof. Adriana Kovashka
University of Pittsburgh
November 12, 2015
Today: Object category detection

• Window-based approaches:
  – Review: Viola-Jones detector
  – Dalal-Triggs pedestrian detector

• Part-based approaches:
  – Implicit shape model
  – Deformable parts model
Object Category Detection

• Focus on object search: “Where is it?”
• Build templates that quickly differentiate object patch from background patch
Viola-Jones detector: features

“Rectangular” filters

Feature output is difference between adjacent regions

Value = Σ (pixels in white area) – Σ (pixels in black area)

Efficiently computable with integral image: any sum can be computed in constant time

Integral image

Adapted from Kristen Grauman and Lana Lazebnik
Considering all possible filter parameters: position, scale, and type:

180,000+ possible features associated with each 24 x 24 window

Which subset of these features should we use to determine if a window has a face?

Use AdaBoost both to select the informative features and to form the classifier.
Viola-Jones detector: AdaBoost

- Want to select the single rectangle feature and threshold that best separates positive (faces) and negative (non-faces) training examples, in terms of weighted error.

Outputs of a possible rectangle feature on faces and non-faces.

Resulting weak classifier:

\[ h_t(x) = \begin{cases} 
+1 & \text{if } f_t(x) > \theta_t \\
-1 & \text{otherwise} 
\end{cases} \]

For next round, reweight the examples according to errors, choose another filter/threshold combo.
Cascading classifiers for detection

- Form a *cascade* with low false negative rates early on
- Apply less accurate but faster classifiers first to immediately discard windows that clearly appear to be negative
Dalal-Triggs pedestrian detector

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

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05
Resolving detection scores

Non-max suppression

Score = 0.1

Score = 0.8

Score = 0.8

Adapted from Derek Hoiem
Map each grid cell in the input window to a histogram counting the gradients per orientation.

Train a linear SVM using training set of pedestrian vs. non-pedestrian windows.

Code available:
http://pascal.inrialpes.fr/soft/olt/
• Histogram of gradient orientations

Orientation: 9 bins (for unsigned angles)  Histograms in 8x8 pixel cells

- Votes weighted by magnitude
- Bilinear interpolation between cells
Histograms of oriented gradients (HOG)

N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005

Image credit: N. Snavely
Historograms of oriented gradients (HOG)

N. Dalal and B. Triggs, [Histograms of Oriented Gradients for Human Detection](http://www.cs.ubc.ca/~nando/), CVPR 2005
Histograms of oriented gradients (HOG)

N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005
Histograms of oriented gradients (HOG)

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05
\[ 0.16 = w^T x \]

\[ \text{sign}(0.16) = 1 \]

\[ \implies \text{pedestrian} \]
Detection examples
Are we done?

• Single rigid template usually not enough to represent a category
  – Many objects (e.g. humans) are articulated, or have parts that can vary in configuration
  – Many object categories look very different from different viewpoints, or from instance to instance
Deformable objects

Images from Caltech-256

Slide Credit: Duan Tran
Deformable objects

Images from D. Ramanan’s dataset
Parts-based Models

Define object by collection of parts modeled by

1. Appearance
2. Spatial configuration

Slide credit: Rob Fergus
How to model spatial relations?

• One extreme: fixed template
Part-based template

- Object model = sum of scores of features at fixed positions

\[ +3 +2 -2 -1 -2.5 = -0.5 \]

? > 7.5

Non-object

\[ +4 +1 +0.5 +3 +0.5 = 10.5 \]

? > 7.5

Object
How to model spatial relations?

- Another extreme: bag of words
How to model spatial relations?

- Star-shaped model
How to model spatial relations?

• Star-shaped model
Parts-based Models

- Articulated parts model
  - Object is configuration of parts
  - Each part is detectable and can move around

Adapted from Derek Hoiem, images from Felzenszwalb
Implicit shape models

- Visual vocabulary is used to index votes for object position [a visual word = “part”]

B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV Workshop on Statistical Learning in Computer Vision 2004
Implicit shape models

- Visual vocabulary is used to index votes for object position [a visual word = “part”]

B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV Workshop on Statistical Learning in Computer Vision 2004
Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering
Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest word
Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest word
3. For each word, store all positions it was found, relative to object center
Implicit shape models: Testing

1. Given new test image, extract patches, match to vocabulary words
2. Cast votes for possible positions of object center
3. Search for maxima in voting space

Original Image  Interest Points
Example: Results on Cows

Original image

K. Grauman, B. Leibe
Example: Results on Cows

Interest points

K. Grauman, B. Leibe
Example: Results on Cows

Matched patches

K. Grauman, B. Leibe
Example: Results on Cows

1st hypothesis

K. Grauman, B. Leibe
Example: Results on Cows

2nd hypothesis

K. Grauman, B. Leibe
Example: Results on Cows

3rd hypothesis
K. Grauman, B. Leibe
Detection Results

- Qualitative Performance
  - Recognizes different kinds of objects
  - Robust to clutter, occlusion, noise, low contrast
Discriminative part-based models

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, PAMI 32(9), 2010

Lana Lazebnik
Discriminative part-based models

Multiple components

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, PAMI 32(9), 2010

Lana Lazebnik
Discriminative part-based models

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, PAMI 32(9), 2010
Scoring an object hypothesis

- The score of a hypothesis is the sum of appearance scores minus the sum of deformation costs.

\[ z = (p_0, \ldots, p_n) \]
\[ p_0 : \text{location of root} \]
\[ p_1, \ldots, p_n : \text{location of parts} \]

\[
\text{score}(p_0, \ldots, p_n) = \sum_{i=0}^{n} F_i \cdot H(p_i) - \sum_{i=1}^{n} D_i \cdot (dx_i, dy_i, dx_i^2, dy_i^2)
\]

Adapted from Lana Lazebnik
Detection

- Feature map
- Feature map at twice the resolution
- Model

- Response of root filter
- Color encoding of filter response values
- Combined score of root locations
- Response of part filters
- Transformed responses
Training

- Training data consists of images with labeled bounding boxes
- Need to learn the filters and deformation parameters
Training

- Our classifier has the form

\[ f(x) = w \cdot H(x) \]

- \( w \) are model parameters, \( z \) are *latent* hypotheses

- **Latent SVM** training:
  - Initialize \( w \) and iterate:
    - Fix \( w \) and find the best \( z \) for each training example
    - Fix \( z \) and solve for \( w \) (standard SVM training)

\( z = (p_0, ..., p_n) \)

\( p_0 \): location of root

\( p_1, ..., p_n \): location of parts
Car model

Component 1

Component 2
Car detections

high scoring true positives

high scoring false positives
Person model
Person detections

high scoring true positives

high scoring false positives
(not enough overlap)
Bottle model
Cat model
Cat detections

high scoring true positives

high scoring false positives (not enough overlap)
Detection state of the art

**R-CNN: Regions with CNN features**

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. **R-CNN achieves a mean average precision (mAP) of 53.7% on PASCAL VOC 2010.** For comparison, Uijlings et al. (2013) report 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%.

Summary

• Window-based approaches
  – Assume object appears in roughly the same configuration in different images
  – Look for alignment with a global template

• Part-based methods
  – Allow parts to move somewhat from their usual locations
  – Look for good fits in appearance, for both the global template and the individual part templates

• Next time: Analyzing and debugging detection methods