CS 1674: Intro to Computer Vision

Deformable Part Models

Prof. Adriana Kovashka
University of Pittsburgh
November 9, 2016
Today: Object category detection

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

• Part-based approaches:
  – Implicit shape model
  – Deformable parts model (DPM)

• Improvements:
  – Speeding up DPM
  – Analyzing the failures of DPM
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
• **Histogram of gradient orientations**

Orientation: 9 bins (for unsigned angles)

Histograms in 8x8 pixel cells

— Votes weighted by magnitude

Adapted from Pete Barnum

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05
Histograms of oriented gradients (HOG)


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


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

N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005
Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05
0.16 = w^T x

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

\[ \Rightarrow \text{pedestrian} \]
Resolving detection scores

Non-max suppression

Score = 0.1

Score = 0.8

Score = 0.8

Adapted from Derek Hoiem
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
Deformable objects

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

Define object by collection of parts modeled by

1. Appearance
2. Spatial configuration
How to model spatial relations?

• One extreme: fixed template
Part-based template

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

\[
\begin{align*}
+3 & \quad +2 & \quad -2 & \quad -1 & \quad -2.5 & \quad = \quad -0.5 & \quad > \quad 7.5 & \quad \text{Non-object} \\
+4 & \quad +1 & \quad +0.5 & \quad +3 & \quad +0.5 & \quad = \quad 10.5 & \quad > \quad 7.5 & \quad \text{Object}
\end{align*}
\]
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: 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
Recall: Generalized Hough transform

- Template representation: for each type of landmark point, store all possible displacement vectors towards the center.

Template

Model
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
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
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
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

\[ z = (p_0, \ldots, p_n) \]

- \( p_0 \): location of root
- \( p_1, \ldots, p_n \): location of parts

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

- Appearance weights
- Displacements
- i.e. how much the part \( p_i \) moved from its expected location in the x, y directions
- i.e. how much we'll penalize the part \( p_i \) for moving from its expected location

Adapted from Lana Lazebnik
Detection

feature map

feature map at twice the resolution

model

response of root filter

response of part filters

transformed responses

color encoding of filter response values

combined score of root locations

Lana Lazebnik
Training

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

- Our classifier has the form

\[ f(x) = \mathbf{w} \cdot \mathbf{H}(x) \]

- \( \mathbf{w} \) are model parameters, \( \mathbf{z} \) are latent hypotheses

- **Latent SVM** training:
  - Initialize \( \mathbf{w} \) and iterate:
    - Fix \( \mathbf{w} \) and find the best \( \mathbf{z} \) for each training example
    - Fix \( \mathbf{z} \) and solve for \( \mathbf{w} \) (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)
Cat model
Cat detections

high scoring true positives

high scoring false positives (not enough overlap)
Today: Object category detection

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

• Part-based approaches:
  – Implicit shape model
  – Deformable parts model (DPM)

• Improvements:
  – Speeding up DPM
  – Analyzing the failures of DPM
Speeding up detection: Restrict set of windows we pass through SVM to those w/ high “objectness”

**Fig. 1:** Desired behavior of an objectness measure. The desired objectness measure should score the blue windows, partially covering the objects, lower than the ground truth windows (green), and score even lower the red windows containing only stuff or small parts of objects.
Objectness cue #1: Where people look

Fig. 2: MS success and failure.
Objectness cue #2: color contrast at boundary

Fig. 3: CC success and failure. Success: the windows containing the objects (cyan) have high color contrast with their surrounding ring (yellow) in images (a) and (b). Failure: the color contrast for windows in cyan in image (c) is much lower.

Alexe et al., CVPR 2010
Objectness cue #3:
no segments “straddling” the object box

Fig. 5: The SS cue. Given the segmentation (b) of image (a), for a window \( w \) we compute \( SS(w, \theta_{SS}) \) (eq. 4). In (c), most of the surface of \( w_1 \) is covered by superpixels contained almost entirely inside it. Instead, all superpixels passing by \( w_2 \) continue largely outside it. Therefore, \( w_1 \) has a higher SS score than \( w_2 \). The window \( w_3 \) has an even higher score as it fits the object tightly.

Alexe et al., CVPR 2010
Boxes found to have high “objectness”

Cyan = ground truth bounding boxes, yellow = correct and red = incorrect predictions for “objectness”

Only run the sheep / horse / chair etc. classifier on the yellow/red boxes.

Alexe et al., CVPR 2010
Today: Object category detection

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

• Part-based approaches:
  – Implicit shape model
  – Deformable parts model (DPM)

• Improvements:
  – Speeding up DPM
  – Analyzing the failures of DPM
How do detectors fail?

- Most errors that detectors make are reasonable
  - Localization error and confusion with similar objects
  - Misdetection of occluded or small objects

- Detectors have different sensitivity to different factors
  - E.g. less sensitive to truncation than to size differences

- Failure analysis code and annotations available online
  - [http://web.engr.illinois.edu/~dhoiem/projects/detectionAnalysis/](http://web.engr.illinois.edu/~dhoiem/projects/detectionAnalysis/)

Adapted from Hoiem et al., ECCV 2012
Top false positives: Airplane (DPM)

AP = 0.36

- Background: 27%
- Localization: 29%
- Similar Objects: 33%
  - Bird, Boat, Car
- Other Objects: 11%

Hoiem et al., ECCV 2012
Top false positives: Dog (DPM)

AP = 0.03

Similar Objects 50%
Person, Cat, Horse

Localization 17%

Background 23%

Other Objects 10%

Hoiem et al., ECCV 2012
Analysis of object characteristics

Additional **annotations** for seven categories: occlusion level, parts visible, sides visible

Level of occlusion: 2 (moderate)
Parts visible: bike body, handlebars, wheel
Parts not visible: seat
View: side visible (front, top, etc., not visible)
Area: 3233 pixels
Aspect Ratio (w/h): 1.24
Object characteristics: Aeroplane

Occlusion: poor robustness to occlusion, but little impact on overall performance

Hoiem et al., ECCV 2012
**Object characteristics: Aeroplane**

**Size**: strong preference for average to above average sized airplanes
Object characteristics: Aeroplane

Aspect Ratio: 2-3x better at detecting wide (side) views than tall views

Easier (Wide) — Harder (Tall)

Hoiem et al., ECCV 2012
Object characteristics: Aeroplane

**Sides/Parts:** best performance = direct side view with all parts visible

---

Easier (Side)

Harder (Non-Side)

Hoiem et al., ECCV 2012
Detection in 2014

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
  – Speed up by only scoring boxes that look like *any* object
  – Models prefer that objects appear in certain views