Computer Vision - Lecture 14

Part-based Models for Object Categorization


Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de
leibe@vision.rwth-aachen.de
Course Outline

- Image Processing Basics
- Segmentation & Grouping
- Object Recognition
- Object Categorization I
  - Sliding Window based Object Detection
- Local Features & Matching
  - Local Features - Detection and Description
  - Recognition with Local Features
  - Indexing & Visual Vocabularies
- Object Categorization II
  - Bag-of-Words Approaches & Part-based Approaches
  - Deep Learning Methods
- 3D Reconstruction
Topics of This Lecture

• Recap: Specific Object Recognition with Local Features
  - Matching & Indexing
  - Geometric Verification

• Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures

• Bag-of-Words Model
  - Use for image classification

• Implicit Shape Model
  - Generalized Hough Transform for object category detection

• Deformable Part-based Model
  - Discriminative part-based detection
Recap: Recognition with Local Features

• Image content is transformed into local features that are invariant to translation, rotation, and scale
• Goal: Verify if they belong to a consistent configuration

Local Features, e.g. SIFT

Slide credit: David Lowe
Recap: Indexing features

Detect or sample features

List of positions, scales, orientations

Describe features

Associated list of d-dimensional descriptors

⇒ Shortlist of possibly matching images + feature correspondences

Index each one into pool of descriptors from previously seen images

or

Match to quantized descriptors (visual words)
Extension: *tf-idf* Weighting

- **Term frequency - inverse document frequency**
  - Describe frame by frequency of each word within it, downweight words that appear often in the database
  - (Standard weighting for text retrieval)

\[
t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}
\]

- **Number of occurrences of word** \(i\) **in document** \(d\)
- **Number of words in document** \(d\)
- **Total number of documents in database**
- **Number of occurrences of word** \(i\) **in whole database**

Slide credit: Kristen Grauman
Recap: Fast Indexing with Vocabulary Trees

- Recognition

Geometric verification

[Nister & Stewenius, CVPR’06]
Recap: Geometric Verification by Alignment

• Assumption
  - Known object, rigid transformation compared to model image
  \[ \Rightarrow \text{If we can find evidence for such a transformation, we have recognized the object.} \]

• You learned methods for
  - Fitting an *affine transformation* from \( \geq 3 \) correspondences
  - Fitting a *homography* from \( \geq 4 \) correspondences

\[
\text{Affine: solve a system} \quad \begin{align*}
    A_t &= b \\
    \text{Homography: solve a system} \quad A_h &= 0
\end{align*}
\]

• Correspondences may be noisy and may contain outliers
  \[ \Rightarrow \text{Need to use robust methods that can filter out outliers} \]
  \[ \Rightarrow \text{Use RANSAC or the Generalized Hough Transform} \]
Topics of This Lecture

• Recap: Specific Object Recognition with Local Features

• Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures

• Bag-of-Words Model
  - Use for image classification

• Implicit Shape Model
  - Generalized Hough Transform for object category detection

• Deformable Part-based Model
  - Discriminative part-based detection
Recognition of Object Categories

• We no longer have exact correspondences...

• On a local level, we can still detect similar parts.

• Represent objects by their parts ⇒ Bag-of-features

• How can we improve on this?
  ➢ Encode structure

Slide credit: Rob Fergus
Part-Based Models

- Fischler & Elschlager 1973

- Model has two components
  - parts
    - (2D image fragments)
  - structure
    - (configuration of parts)
Different Connectivity Structures

\[ O(N) \]

- a) Bag of visual words
  - Csurka et al. '04
  - Vasconcelos et al. '00

\[ O(N^k) \]

- b) Constellation
  - Fergus et al. '03
  - Fei-Fei et al. '03

\[ O(N^2) \]

- c) Star shape
  - Leibe et al. '04, '08
  - Crandall et al. '05
  - Fergus et al. '05

\[ O(N^2) \]

- d) Tree
  - Felzenszwalb & Huttenlocher '05

\[ O(N^3) \]

- e) k-fan \((k = 2)\)
  - Crandall et al. '05

- f) Hierarchy
  - Bouchard & Triggs '05

- g) Sparse flexible model
  - Carneiro & Lowe '06

Slide adapted from Rob Fergus

Image from [Carneiro & Lowe, ECCV'06]
Topics of This Lecture

• Recap: Specific Object Recognition with Local Features

• Part-Based Models for Object Categorization
  ➢ Structure representations
  ➢ Different connectivity structures

• **Bag-of-Words Model**
  ➢ Use for image classification

• Implicit Shape Model
  ➢ Generalized Hough Transform for object category detection

• Deformable Part-based Model
  ➢ Discriminative part-based detection
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex served as a movie screen, so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China's exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
Recap: Visual Words

- Quantize the feature space into “visual words”
- Perform matching only to those visual words.

Exact feature matching $\rightarrow$ Match to same visual word

Figure from Sivic & Zisserman, ICCV 2003

Slide adapted from Kristen Grauman
Recap: Bag-of-Word Representations (BoW)

Object → Bag of “words”

Source: ICCV 2005 short course, Li Fei-Fei
Recap: Categorization with Bags-of-Words

- Compute the word activation histogram for each image.
- Let each such BoW histogram be a feature vector.
- Use images from each class to train a classifier (e.g., an SVM).

Slide adapted from Kristen Grauman
Recap: Advantage of BoW Histograms

- Bag of words representations make it possible to describe the unordered point set with a single vector (of fixed dimension across image examples).

- Provides easy way to use distribution of feature types with various learning algorithms requiring vector input.
Limitations of BoW Representations

• The bag of words removes spatial layout.

• This is both a strength and a weakness.

• Why a strength?

• Why a weakness?
Spatial Pyramid Representation

• Representation in-between orderless BoW and global appearance
Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

Slide credit: Svetlana Lazebnik

B. Leibe

[Lazebnik, Schmid & Ponce, CVPR’06]
Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance
Summary: Bag-of-Words

- **Pros:**
  - Flexible to geometry / deformations / viewpoint
  - Compact summary of image content
  - Provides vector representation for sets
  - Empirically good recognition results in practice

- **Cons:**
  - Basic model ignores geometry - must verify afterwards, or encode via features.
  - Background and foreground mixed when bag covers whole image
  - When using interest points or sampling: no guarantee to capture object-level parts \(\Rightarrow\) Dense sampling is often better.
Topics of This Lecture

• Recap: Specific Object Recognition with Local Features

• Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures

• Bag-of-Words Model
  - Use for image classification

• Implicit Shape Model
  - Generalized Hough Transform for object category detection

• Deformable Part-based Model
  - Discriminative part-based detection
Implicit Shape Model (ISM)

- Basic ideas
  - Learn an appearance codebook
  - Learn a star-topology structural model
    - Features are considered independent given obj. center

- Algorithm: probabilistic Gen. Hough Transform
  - Exact correspondences $\rightarrow$ Prob. match to object part
  - NN matching $\rightarrow$ Soft matching
  - Feature location on obj. $\rightarrow$ Part location distribution
  - Uniform votes $\rightarrow$ Probabilistic vote weighting
  - Quantized Hough array $\rightarrow$ Continuous Hough space
Implicit Shape Model: Basic Idea

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

Implicit Shape Model: Basic Idea

- Objects are detected as consistent configurations of the observed parts (visual words).

Test image

Implicit Shape Model - Representation

- Learn appearance codebook
  - Extract local features at interest points
  - Agglomerative clustering $\Rightarrow$ codebook

- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
Implicit Shape Model - Recognition

Interest Points → Matched Codebook Entries → Probabilistic Voting

Image Feature → Interpretation (Codebook match) → Object Position

\[ p(C_i | f) \quad p(o_n, x | C_i, \ell) \]

Probabilistic vote weighting

[Leibe, Leonardis, Schiele, SLCV’04; IJCV’08]
Implicit Shape Model - Recognition

Interest Points → Matched Codebook Entries → Probabilistic Voting

Backprojected Hypotheses

3D Voting Space (continuous)

Backprojection of Maxima

[Leibe, Leonardis, Schiele, SLCV’04; IJCV’08]
Example: Results on Cows

Original image
Example: Results on Cows

Interest points

B. Leibe
Example: Results on Cows

Matched patches

B. Leibe
Example: Results on Cows

Prob. Votes
Example: Results on Cows

1st hypothesis

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

2nd hypothesis

B. Leibe
Example: Results on Cows

3rd hypothesis

B. Leibe
Scale Invariant Voting

- Scale-invariant feature selection
  - Scale-invariant interest regions
  - Extract scale-invariant descriptors
  - Match to appearance codebook

- Generate scale votes
  - Scale as 3rd dimension in voting space
    \[
    \begin{align*}
    x_{vote} &= x_{img} - x_{occ}(s_{img}/s_{occ}) \\
    y_{vote} &= y_{img} - y_{occ}(s_{img}/s_{occ}) \\
    s_{vote} &= (s_{img}/s_{occ}).
    \end{align*}
    \]
  - Search for maxima in 3D voting space
Detection Results

- Qualitative Performance
  - Recognizes different kinds of objects
  - Robust to clutter, occlusion, noise, low contrast
Detections Using Ground Plane Constraints

Battery of 5 ISM detectors for different car views

[Leibe, Cornelis, Cornelis, Van Gool, CVPR’07]
Extension: Rotation-Invariant Detection

• Polar instead of Cartesian voting scheme

  ![Diagram showing polar instead of Cartesian voting scheme](image)

• Benefits:
  - Recognize objects under image-plane rotations
  - Possibility to share parts between articulations.

• Caveats:
  - Rotation invariance should only be used when it’s really needed. (Also increases false positive detections)

[Mikolajczyk, Leibe, Schiele, CVPR’06]
Sometimes, Rotation Invariance Is Needed…

Figure from [Mikolajczyk et al., CVPR’06]
Implicit Shape Model - Segmentation

Local Features → Matched Codebook Entries → Probabilistic Voting

Segmentation

Backprojected Hypotheses

Pixel Contributions

Backprojection of Maxima

3D Voting Space (continuous)

Leibe, Leonardis, Schiele, DAGM’04; IJCV’08
Example Results: Motorbikes

[Leibe, Leonardis, Schiele, SLCV’04; IJCV’08]
You Can Try It At Home...

- Linux source code & binaries available
  - Including datasets & several pre-trained detectors
  - [http://www.vision.rwth-aachen.de/software](http://www.vision.rwth-aachen.de/software)
Topics of This Lecture

• Recap: Specific Object Recognition with Local Features

• Part-Based Models for Object Categorization
  ➢ Structure representations
  ➢ Different connectivity structures

• Bag-of-Words Model
  ➢ Use for image classification

• Implicit Shape Model
  ➢ Generalized Hough Transform for object category detection

• Deformable Part-based Model
  ➢ Discriminative part-based detection
Starting Point: HOG Sliding-Window Detector

- Array of weights for features in window of HOG pyramid
- Score is dot product of filter and vector
Deformable Part-based Models

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

Slide credit: Pedro Felzenszwalb
2-Component Bicycle Model

Root filters  
coarse resolution

Part filters  
finer resolution

Deformation models

Slide credit: Pedro Felzenszwalb  
B. Leibe
Object Hypothesis

- Multiscale model captures features at two resolutions

Score of filter: dot product of filter with HOG features underneath it

Score of object hypothesis is sum of filter scores minus deformation costs
Score of a Hypothesis

\[
\text{score}(p_0, \ldots, p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot (d_x^2, d_y^2) \]

“data term”

filters

displacements

deformation parameters

“spatial prior”

\[
\text{score}(z) = \beta \cdot \Psi(H, z) 
\]

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features
Recognition 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

Slide credit: Pedro Felzenszwalb
Results: Persons

- Results (after non-maximum suppression)
  - ~1s to search all scales

Slide credit: Pedro Felzenszwalb
Results: Bicycles

B. Leibe

Slide adapted from Trevor Darrell
False Positives

- Bicycles
Results: Cats

High-scoring true positives

High-scoring false positives (not enough overlap)

Slide credit: Pedro Felzenszwalb
You Can Try It At Home...

• Deformable part-based models have been very successful at several recent evaluations.

⇒ State-of-the-art approach in object detection for several years

• Source code and models trained on PASCAL 2006, 2007, and 2008 data are available here:
  
  http://www.cs.uchicago.edu/~pff/latent
References and Further Reading

• Details about the ISM approach can be found in

• Details about the DPMs can be found in

• Try the ISM Linux binaries
  - http://www.vision.ee.ethz.ch/bleibe/code

• Try the Deformable Part-based Models
  - http://www.cs.uchicago.edu/~pff/latent