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

Recap: Indexing features

- Detect or sample features
- List of positions, scales, orientations
  - Detect features
  - Associated list of d-dimensional descriptors
- Describe features
- Match to quantized descriptors (visual words)

⇒ Shortlist of possibly matching images + feature correspondences

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 \( d \)

Total number of documents in database

Number of occurrences of word \( i \) in whole database

Slide credits: Kristen Grauman; David Lowe
Recap: Fast Indexing with Vocabulary Trees

- Recognition

Recap: Geometric Verification by Alignment

- Assumption
  - Known object, rigid transformation compared to model image
  - 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

  \[ \begin{align*}
  \text{Affine: solve a system} \\
  A \mathbf{t} = \mathbf{b} \\
  \text{Homography: solve a system} \\
  \mathbf{Ah} = \mathbf{0}
  \end{align*} \]

- Correspondences may be noisy and may contain outliers
  - Need to use robust methods that can filter out outliers
  - 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

Part-Based Models

- Fischler & Elschlager 1973

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

Different Connectivity Structures

- $O(N)$
- $O(N^2)$
- $O(N^3)$

- Bag of visual words
  - Courville et al. ’04
  - Vasconcelos et al. ’00
- Constellation
  - Fergus et al. ’03
  - Fei-Fei et al. ’03
- Star shape
  - Leibe et al. ’04
  - Crandall et al. ’05
- Fergus et al. ’05
- Treet
  - Feiters et al. ’05
- Sparse flexible model
  - Carneiro & Lowe ’06

- $O(N^3)$
- $O(N^4)$

- k-fan ($k = 2$)
  - Crandall et al. ’05
- Hierarchy
  - Bouchard & Triggs ’05
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

Recap: Analogy to Documents

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 us from our eyes. Figure from  Sick & Zisserman, ICCV 2003

Recap: Visual Words

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

Recap: Bag-of-Word Representations (BoW)

Canada is forecasting a trade surplus of $40bn (€25bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by China's 20% jump in exports but US demand for Chinese goods is likely to only rise by 10%. The US has allowed it to trade within a narrow band, but the yuan is currently only 30% above its exchange rate in 1994. Zhou Xiaochuan said the country only needs to do more to boost domestic demand on its own. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a 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: 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).

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

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 ⇒ 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 → Prob. match to object part
  - NN matching → Soft matching
  - Feature location on obj. → Part location distribution
  - Uniform votes → Probabilistic vote weighting
  - Quantized Hough array → 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 - Representation**

- Learn appearance codebook
  - Extract local features at interest points
  - Agglomerative clustering of codebook
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object

**Spatial occurrence distributions**

**Implicit Shape Model - Recognition**

- Interest Points
- Matched Codebook Entries
- Probabilistic Voting

- Image Feature
- Interpretation (Codebook match)
- Object Position
- 3D Voting Space (continuous)

- Probabilistic vote weighting

**Implicit Shape Model: Basic Idea**

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

**Implicit Shape Model - Recognition**

- Interest Points
- Matched Codebook Entries
- Probabilistic Voting

- Backprojected Hypotheses
- Backprojection of Maxima
Example: Results on Cows

Original image

Interest points

Matched patches

Prob. Votes

1st hypothesis

2nd hypothesis
Example: Results on Cows

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
  - 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

Extension: Rotation-Invariant Detection
- Polar instead of Cartesian voting scheme
  - 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)

Sometimes, Rotation Invariance Is Needed...
**Implicit Shape Model - Segmentation**

Local Features → Matched Codebook Entries → Probabilistic Voting → Backprojected Hypotheses → Backprojection of Maxima → Segmentation

- Pixel Contributions
- 3D Voting Space (continuous)

---

**Example Results: Motorbikes**

- Backprojection Meta-information

---

**You Can Try It At Home...**

- Linux source code & binaries available
  - Including datasets & several pre-trained detectors
  - 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

\[ F \cdot \phi(p, H) \]

\( \phi(p, H) \) = concatenation of HOG features from window specified by \( p \).

---

**Deformable Part-based Model**

- 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 object hypothesis is sum of filter scores minus deformation costs

Score of a Hypothesis

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

\[
\text{Score of filter: dot product of filter with HOG features underneath it}
\]

Recognition Model

\[
f_w(x) = \max_{z} w \cdot \Phi(x, z)
\]

- \(z\): vector of part offsets
- \(\Phi(x, z)\): vector of HOG features (from root filter & appropriate part sub-windows) and part offsets

Results: Persons

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

Results: Bicycles

- Slide adapted from Trevor Darrell
False Positives

- Bicycles

Results: Cats

- High-scoring true positives
- High-scoring false positives (not enough overlap)

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

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