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
- 3D Reconstruction
- Optical Flow

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
- Describe features
- Associated list of d-dimensional descriptors
- Index each one into pool of descriptors from previously seen images
- 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 \( d \)
- Number of words in document \( d \)
- Total number of documents in database
- Number of occurrences of word \( i \) in whole database
Recap: Fast Indexing with Vocabulary Trees

- Recognition
- Geometric verification

[Nister & Stewenius, CVPR’06]

Application for Content Based Img Retrieval

- What if query of interest is a portion of a frame?

[“Groundhog Day” [Rammis, 1993]]

Slide credit: David Nister

Perceptual and Sensory Augmented Computing

Video Google System

1. Collect all words within query region
2. Inverted file index to find relevant frames
3. Compare word counts
4. Spatial verification

Sivic & Zisserman, ICCV 2003

• Demo online at:
  http://www.robots.ox.ac.uk/~vgg/research/egoogle/index.html

Collecting Words Within a Query Region

- Example: Friends

Query region: pull out only the SIFT descriptors whose positions are within the polygon

Slide credit: Kristen Grauman

Example Results

Query

Retrieved frames

More Results

Query

Retrieved shots

Slide credit: Kristen Grauman

Perceptual and Sensory Augmented Computing

Computer Vision WS 14/15

B. Leibe
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
    - Affine: solve a system
    - Homography: solve a system
    \[
    A \cdot t = b \\
    A \cdot h = 0
    \]

- **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

Applications: Aachen Tourist Guide

Applications: Fast Image Registration

Applications: Mobile Augmented Reality

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

- Bag of visual words
  - Caesura et al. '04
  - Vasconcelos et al. '00
- Constellation
  - Fergus et al. '03
  - Fei-Fei et al. '03
- Star shape
  - Crandall et al. '05
- Sparse flexible model
  - Carnetino & Lowe '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

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 transmitted to us by our eyes. For many years, it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex was 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 the brain. 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, this has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
Bags of Visual Words

- Summarize entire image based on its distribution (histogram) of word occurrences.
- Analogous to bag of words representation commonly used for documents.
- Main difference to text: visual words are not given a priori, but obtained through clustering (e.g., using k-means)

Comparing Bags of Words

- We build up histograms of word activations, so any histogram comparison measure can be used here.
- E.g. we can rank frames by normalized scalar product between their (possibly weighted) occurrence counts.
  
  Nearest neighbor search for similar images.

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).

Similarly, Bags-of-Textons for Texture Repr.

- Bag of words representation makes 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.

BoW for Object Categorization

- Works pretty well for image-level classification
**BoW for Object Categorization**

**Caltech6 dataset**

<table>
<thead>
<tr>
<th>class</th>
<th>bag of features</th>
<th>bag of features</th>
<th>Parts-and-shape model</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplanes</td>
<td>98.8</td>
<td>97.1</td>
<td>90.3</td>
</tr>
<tr>
<td>cars (rear)</td>
<td>98.3</td>
<td>98.6</td>
<td>90.3</td>
</tr>
<tr>
<td>cars (side)</td>
<td>95.0</td>
<td>98.6</td>
<td>88.5</td>
</tr>
<tr>
<td>faces</td>
<td>100</td>
<td>99.3</td>
<td>96.4</td>
</tr>
<tr>
<td>motorbikes</td>
<td>98.5</td>
<td>98.0</td>
<td>92.5</td>
</tr>
<tr>
<td>spotted cats</td>
<td>97.0</td>
<td></td>
<td>90.0</td>
</tr>
</tbody>
</table>

- Good performance for pure classification (object present/absent)
  - Better than more elaborate part-based models with spatial constraints...
  - What could be possible reasons why?

**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

**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
  - Interest points or sampling: no guarantee to capture object-level parts.
  - Optimal vocabulary formation remains unclear.
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 object 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).

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

Training image

Visual codeword with displacement vectors

Test image


B. Leibe

B. Leibe

B. Leibe

B. Leibe
Implicit Shape Model - Recognition

- Interest Points
- Matched Codebook Entries
- Probabilistic Voting
- Backprojected Hypotheses
- Backprojection of Maxima

3D Voting Space (continuous)

Example: Results on Cows

- Original image

Example: Results on Cows

- Interest points

Example: Results on Cows

- Matched patches

Example: Results on Cows

- Prob. Votes

Example: Results on Cows

- 1st hypothesis
Example: Results on Cows

2nd hypothesis

3rd hypothesis

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
    \[ x_{scale} = \frac{x_{img} - x_{loc}(s_{img}/s_{loc})}{s_{loc}} \]
    \[ y_{scale} = \frac{y_{img} - y_{loc}(s_{img}/s_{loc})}{s_{loc}} \]
    \[ s_{scale} = \frac{s_{img}}{s_{loc}}. \]
  - 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 DA 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...

Figure from [Wiskott et al., CVPR’06]

Implicit Shape Model - Segmentation

- Local Features
- Matched Codebook Entries
- Probabilistic Voting

- Backprojected Meta-Information
- 3D Voting Space (continuous)

Pixel Contributions
Backprojected Hypotheses
Backprojection of Maxima

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

Example Results: Motorbikes

You Can Try It At Home...

- Linux source code & binaries available
  - Including datasets & several pre-trained detectors
  - http://www.vision.rwth-aachen.de/software

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

**Object Hypothesis**

- Multiscale model captures features at two resolutions

**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

**Score of a Hypothesis**

\[ \text{score}(p_0, \ldots, p_n) = \sum_{f=0}^{n_f} F_f \cdot \phi(H, p_f) - \sum_{d=1}^{n_d} d_i \cdot (dx_i^2, dy_i^2) \]

- “data term”
- “spatial prior”

- concatenation filters and deformation parameters
- concatenation of HOG features and part displacement features

**Results: Persons**

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

Slide adapted from Trevor Darrell

False Positives

- Bicycles

Slide credit: Pedro Felzenszwalb

Results: Cats

High-scoring true positives

High-scoring false positives
(not enough overlap)

You Can Try It At Home...

- Deformable part-based models have been very successful at several recent evaluations.
  \( \Rightarrow \) Currently, state-of-the-art approach in object detection

- Source code and models trained on PASCAL 2006, 2007, and 2008 data are available here:
  \[ \text{http://www.cs.uchicago.edu/~pff/latent} \]

References and Further Reading

- Details about the ISM approach can be found in
  - B. Leibe, A. Leonardis, and B. Schiele,

- Details about the DPMs can be found in
  - P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan,

- Try the ISM Linux binaries
  \[ \text{http://www.vision.ee.ethz.ch/bleibe/code} \]

- Try the Deformable Part-based Models
  \[ \text{http://www.cs.uchicago.edu/~pff/latent} \]