Recap: Local Feature Matching Outline

1. Find a set of distinctive keypoints
2. Define a region around each keypoint
3. Extract and normalize the region content
4. Compute a local descriptor from the normalized region
5. Match local descriptors

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: Object Recognition 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 ≥ 3 correspondences
  - Fitting a homography from ≥ 4 correspondences
  - Affine: solve a system
    \[ A t = b \]
  - Homography: solve a system
    \[ A h = 0 \]
- Correspondences may be noisy and may contain outliers
  - Use RANSAC for robust fitting

Recap: Robust Estimation with RANSAC

RANSAC loop:
1. Randomly select a seed group of points on which to base transformation estimate (e.g., a group of matches)
2. Compute transformation from seed group
3. Find inliers to this transformation
4. If the number of inliers is sufficiently large, recompute least-squares estimate of transformation on all of the inliers
- Keep the transformation with the largest number of inliers
Recap: Generalized Hough Transform

- Suppose our features are scale- and rotation-invariant. Then a single feature match provides an alignment hypothesis (translation, scale, orientation).

Of course, a hypothesis from a single match is unreliable. Solution: let each match vote for its hypothesis in a Hough space with very coarse bins.

Topics of This Lecture

- Indexing with Local Features
  - Inverted file index
  - Visual Vocabularies
- 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
  - Multi-resolution models

Application: Mobile Visual Search

- Take photos of objects as queries for visual search

Large-Scale Image Matching Problem

- How can we perform this matching step efficiently?

Indexing Local Features

- Each patch / region has a descriptor, which is a point in some high-dimensional feature space (e.g., SIFT)

Database with thousands (millions) of images
Indexing Local Features

- When we see close points in feature space, we have similar descriptors, which indicates similar local content.

- This is of interest for many applications:
  - E.g. Image matching,
  - E.g. Retrieving images of similar objects,
  - E.g. Object recognition, categorization, 3d Reconstruction,

Text Retrieval vs. Image Search

- What makes the problems similar, different?

Visual Words: Main Idea

- Extract some local features from a number of images...
Visual Words: Main Idea

Each point is a local descriptor, e.g. SIFT vector.

Idea: quantize the feature space.

Indexing with Visual Words

Map high-dimensional descriptors to tokens/words by quantizing the feature space

• Quantize via clustering, let cluster centers be the prototype “words”

Indexing with Visual Words

Map high-dimensional descriptors to tokens/words by quantizing the feature space

• Determine which word to assign to each new image region by finding the closest cluster center.
Visual Words

- Example: each group of patches belongs to the same visual word.

Figure from Sivic & Zisserman, ICCV 2003

Slide credit: Kristen Grauman

Visual Words

- Often used for describing scenes and objects for the sake of indexing or classification.

Sivic & Zisserman 2003; Csurka, Bray, Dance, & Fan 2004; many others.

Slide credit: Kristen Grauman

Inverted File for Images of Visual Words

When will this give us a significant gain in efficiency?

Slide credit: Kristen Grauman

Sampling Strategies

- Sparse, at interest points
- Dense, uniformly
- Randomly
- Multiple interest operators
  - To find specific, textured objects, sparse sampling from interest points often more reliable.
  - Multiple complementary interest operators offer more image coverage.
  - For object categorization, dense sampling offers better coverage.
  
[See Nowak, Jurie & Triggs, ECCV 2004]

Slide credit: Kristen Grauman

Clustering / Quantization Methods

- k-means (typical choice), agglomerative clustering, mean-shift, ...
- Hierarchical clustering: allows faster insertion / word assignment while still allowing large vocabularies
  - Vocabulary tree [Nister & Stewenius, CVPR 2006]
Example: Recognition with Vocabulary Tree

- Tree construction:

Vocabulary Tree

- Training: Filling the tree

Vocabulary Tree

- Training: Filling the tree

Vocabulary Tree

- Training: Filling the tree
Vocabulary Tree
- Recognition
  - RANSAC verification

Quiz Questions
- What is the computational advantage of the hierarchical representation vs. a flat vocabulary?
- What dangers does such a representation carry?

Vocabulary Tree: Performance
- Evaluated on large databases
  - Indexing with up to 1M images
- Online recognition for database of 50,000 CD covers
  - Retrieval in ~1s
- Experimental finding that large vocabularies can be beneficial for recognition

Vocabulary Size
- Larger vocabularies can be advantageous...
- But what happens when the vocabulary gets too large?
  - Efficiency?
  - Robustness?

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)

Summary: Indexing features
- Detect or sample features
- Describe features
- Index each one into pool of descriptors from previously seen images
- Quantize to form “bag of words” vector for the image
Application for Content Based Img Retrieval

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

  Visually defined query

  “Groundhog Day” [Rammis, 1993]

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/video/index.html

Collecting Words Within a Query Region

- Example: Friends

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

Example Results

Query

More Results

Query

Retrieved shots

Applications: Specific Object Recognition

- Commercial services coming out: koooba

  Works well for mostly planar objects:
  - Movie posters,
  - Book covers,
  - CD/DVD covers,
  - Video games,

Source: http://www.koooba.com
Applications: Aachen Tourist Guide

Topics of This Lecture
- Indexing with Local Features
  - Inverted file index
  - Visual Vocabularies
- 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
  - Multi-resolution models

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
  - structure (configuration of parts)

Different Connectivity Structures
- Bag-of-visual-words
- Constellation
- Star shape
- Tree
- k-fan
- Hierarchy
- Sparse flexible model

Some Class-Specific Graphs
- Articulated motion
  - People
  - Animals
- Special parameterisations
  - Limb angles
Topics of This Lecture

• Indexing with Local Features
  - Inverted File Index
  - Visual Vocabularies
• 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
  - Multi-resolution models

Analogy to Documents

Off 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 are sent from our eyes. From this point the brain breaks down the pictures in the darkness into a series of so-called concept events. By reading along their position, Hubel and Wiesel have been able to demonstrate how the messages about the image falling on the retina undergo a step-wise analysis in a system of nerve cells stored in column 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 the US and Europe, which have been a target for a 10% rise to support the yen. The country also has to work hard to underwrite exports, only one of its biggest needs to be exported. The government has now 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.

Similarly, Bags-of-Textons for Texture Repr.

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.

\[ d_j \cdot q = \sum_{k=1}^{n} w_{j,k} \times w_{k} \]

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

Learning/Recognition with BoW Histograms

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

BoW for Object Categorization

- Works pretty well for image-level classification

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

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>98.6</td>
<td>97.1</td>
<td>96.2</td>
</tr>
<tr>
<td>cars (rear)</td>
<td>98.3</td>
<td>96.6</td>
<td>96.3</td>
</tr>
<tr>
<td>cars (side)</td>
<td>95.0</td>
<td>87.3</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?
**BoW Representation: Spatial Information**

- A bag of words is an orderless representation: throwing out spatial relationships between features
- Middle ground:
  - Visual “phrases” : frequently co-occurring words
  - Semi-local features : describe configuration, neighborhood
  - Let position be part of each feature
  - Count bags of words only within sub-grids of an image
  - After matching, verify spatial consistency (e.g., look at neighbors - are they the same too?)

**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**

- Indexing with Local Features
  - Inverted File Index
  - Visual Vocabularies
- 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
  - Multi-resolution models
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

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

---

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
  - Learn 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

---

Implicit Shape Model - Recognition

- Interest Points
- Matched Codebook Entries
- Probabilistic Voting

---

Implicit Shape Model: Basic Idea

- Test image

---

Implicit Shape Model: Basic Idea

- Training image

---

Implicit Shape Model: Basic Idea

- Visual codeword with displacement vectors

---

Implicit Shape Model - Representation

- Appearance codebook
- Spatial occurrence distributions

---

Implicit Shape Model - Recognition

- Backprojected Hypotheses
- Backprojection of Maxima

---

Implicit Shape Model - Recognition

- Image Feature
- Interpretation (Codebook match)
- Object Position

---

Implicit Shape Model - Recognition

- Probabilistic vote weighting

---

Implicit Shape Model - Recognition

- 3D Voting space (continuous)

---

Implicit Shape Model - Recognition


---

Implicit Shape Model - Recognition

- [B. Leibe, Leonardis, Schiele, SCV08a, LATEX]
Example: Results on Cows

Original image

Interest points

Matched patches

Prob. Votes

1st hypothesis

2nd hypothesis
Example: Results on Cows

3rd hypothesis

Scale Voting: Efficient Computation

- Continuous Generalized Hough Transform
  - Binned accumulator array similar to standard Gen. Hough Transf.
  - Quickly identify candidate maxima locations
  - Refine locations by Mean-Shift search only around those points
  - Avoid quantization effects by keeping exact vote locations.
  - Mean-shift interpretation as kernel prob. density estimation.

Detection Results

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

Scale Invariant Voting

- Scale-invariant feature selection
  - Scale-invariant interest points
  - Rescale extracted patches
  - Match to constant-size codebook

- Generate scale votes
  - Scale as 3rd dimension in voting space
    \[ x_{vote} = x_{max} - x_{min}(s_{max}/s_{min}) \]
    \[ y_{vote} = y_{max} - y_{min}(s_{max}/s_{min}) \]
    \[ s_{vote} = (s_{max}/s_{min}) \]
  - Search for maxima in 3D voting space

Scale Voting: Efficient Computation

- Scale-adaptive Mean-Shift search for refinement
  - Increase search window size with hypothesis scale
  - Scale-adaptive balloon density estimator
    \[ \hat{p}(o_n, x) = \frac{1}{V_b} \sum \lambda k \sum p(o_n, x) \delta_{k(x - x)} K \left( \frac{x - x}{b} \right) \]
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)

Implicit Shape Model - Segmentation

- Local Features
- Matched Codebook Entries
- Probabilistic Voting
- Backprojected Meta-Information
- 3D Voting Space (continuous)
- Segmentation
- Pixel Contributions
- Backprojected Hypotheses
- Backprojection of Maxima

Example Results: Motorbikes

You Can Try It At Home...

- Linux binaries available
  - Including datasets & several pre-trained detectors
  - http://www.vision.ee.ethz.ch/bleibe/code

Topics of This Lecture

- Indexing with Local Features
  - Inverted File Index
  - Visual Vocabularies

- 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
  - Multi-resolution models
Perceptual and Sensory Augmented Computing

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

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 a Hypothesis

score(p_0, ..., p_n) = \sum_{i=0} \phi(H, p_i)

score(z) = \beta \cdot \Psi(H, z)

concatenation of HOG features and part displacement features

Recognition Model

f_w(z) = w \cdot \Phi(x)

f_w(x) = \max_x 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

B. Leibe

Slide credit: Pedro Felzenszwalb

Results: Bicycles

B. Leibe

Slide adapted from Trevor Darrell

False Positives

- Bicycles

B. Leibe

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
  - 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.
  - Currently, state-of-the-art approach in object detection

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