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 \( \geq 3 \) correspondences
  - Fitting a homography from \( \geq 4 \) correspondences
  - Affine: solve a system
  - Homography: solve a system
  \[
  A t = b \\
  A h = 0
  \]
- 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
Strategy 2: 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.

Pose Clustering and Verification with SIFT

- To detect instances of objects from a model base:
  1. Index descriptors
     - Distinctive features narrow down possible matches
  2. Generalized Hough transform to vote for poses
     - Keypoints have record of parameters relative to model coordinate system
  3. Affine fit to check for agreement between model and image features
     - Fit and verify using features from Hough bins with 3+ votes

Object Recognition Results

- Objects recognized
- Recognition in spite of occlusion
- Background subtract for model boundaries

Location Recognition

[Lowe, IJCV'04]
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)

- 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,...

Indexing Local Features

- With potentially thousands of features per image, and hundreds to millions of images to search, how to efficiently find those that are relevant to a new image?

  - Low-dimensional descriptors (e.g. through PCA):
    - Can use standard efficient data structures for nearest neighbor search
  
  - High-dimensional descriptors
    - Approximate nearest neighbor search methods more practical

  - Inverted file indexing schemes
Indexing Local Features: Inverted File Index

- For text documents, an efficient way to find all pages on which a word occurs is to use an index...
- We want to find all images in which a feature occurs.
- To use this idea, we’ll need to map our features to "visual words".

Text Retrieval vs. Image Search

- What makes the problems similar, different?
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"

Descriptor space

Visual Words

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

Figure from Sivic & Zisserman, ICCV 2003

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

Sivic & Zisserman 2003; Csurka, Bray, Dance, & Fan 2004; many others.
Inverted File for Images of Visual Words

Word number List of image numbers
1 5, 10, ...
2 10, ...

When will this give us a significant gain in efficiency?

Visual Vocabulary Formation

Design choices:
- Sampling strategy: where to extract features?
- Clustering / quantization algorithm
- Unsupervised vs. supervised
- What corpus provides features (universal vocabulary?)
- Vocabulary size, number of words

Sampling Strategies

- Sparse, at interest points
- Dense, uniformly
- Randomly

- 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 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

Quiz Questions

• What is the computational advantage of the hierarchical representation vs. a flat vocabulary?

• What dangers does such a representation carry?

Vocabulary Tree

• Recognition

Quiz Questions

• Evaluated on large databases
  • Indexing with up to 1M images

• Online recognition for database of 50,000 CD covers
  • Retrieval in ~1s (in 2006)

• Experimental finding that large vocabularies can be beneficial for recognition

[Nister & Stewenius, CVPR'06]
Vocabulary Size

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

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

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

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
- List of positions, scales, orientations
- Associated list of d-dimensional descriptors
- 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?

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/vggoogle/index.html

Collecting Words Within a Query Region

- Example: Friends
Example Results

Query

More Results

Query

Retrieved shots

Applications: Specific Object Recognition

• Commercial services coming out:
  - Movie posters,
  - Book covers,
  - CD/DVD covers,
  - Video games,
  - ... [Source: http://www.kooaba.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 (2D image fragments)
  - structure (configuration of parts)

Different Connectivity Structures

- Bag of visual words
  - Caio et al. '04
  - Vasconcelos et al. '00
- Constellation
  - Ferras et al. '03
  - Fei-Fei et al. '03
- Star shape
  - Leibe et al. '04
  - Crandall et al. '05
  - Feng et al. '05
- k-fan (k = 2)
  - Crandall et al. '05
- Hierarchy
  - Bouchard et al. '05
- Sparse flexible model
  - Carneiro & Lowe '06

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

- sensory, brain, visual, perception, eye, cell, optical nerve, image
- Hubel, Wiesel

Object

Bag of ‘words’

Source: ECCV 2005 short course, Li Fei-Fei
**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.

**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.

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

**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).
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)

Slide credit: Svetlana Lazebnik

BoW for Object Categorization

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

Slide credit: Svetlana Lazebnik

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?

Slide adapted from Bill Freeman

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?)

Slide credit: Kristen Grauman

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

Slide credit: Svetlana Lazebnik
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

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