Announcements

- PS3 out; due 6/3, 11:59 pm
- Sign attendance sheet (3rd one)
Visual words

- Map high-dimensional descriptors to tokens/words by quantizing the feature space
  
- Quantize via clustering, let cluster centers be the prototype “words”
  
- 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

Inverted file index

- Database images are loaded into the index mapping words to image numbers
Inverted file index

When will this give us a significant gain in efficiency?

- New query image is mapped to indices of database images that share a word.

---

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

- Rank frames by normalized scalar product between their (possibly weighted) occurrence counts—nearest neighbor search for similar images.

\[
\text{sim}(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\|\|q\|}
\]

\[
= \frac{\sum_{i=1}^{V} d_j(i) \cdot q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} \cdot \sqrt{\sum_{i=1}^{V} q(i)^2}}
\]

for vocabulary of \(V\) words
Application: Large-Scale Retrieval

Spatial Verification: two basic strategies

- RANSAC
  - Typically sort by BoW similarity as initial filter
  - Verify by checking support (inliers) for possible transformations
    - e.g., “success” if find a transformation with > N inlier correspondences

- Generalized Hough Transform
  - Let each matched feature cast a vote on location, scale, orientation of the model object
  - Verify parameters with enough votes

RANSAC verification
Voting: Generalized Hough Transform

• If we use scale, rotation, and translation invariant local features, then each feature match gives an alignment hypothesis (for scale, translation, and orientation of model in image).

Model

Novel image

Voting: Generalized Hough Transform

• A hypothesis generated by a single match may be unreliable,
• So let each match vote for a hypothesis in Hough space

Model

Novel image

What else can we borrow from text retrieval?

China is forecasting a trade surplus of $840bn ($1.2 trillion) for this year, a threefold increase on 2004’s $324bn. The Commerce Ministry said the surplus would arise from a 30% jump in exports and a 20% rise in imports, and the trade deficit would be reduced to $4 billion. The US wants the yuan to be allowed to trade freely. However, Beijing has made it clear it will take its time and tread carefully, allowing the yuan to rise further in value.
**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)

\[
I_i = \frac{n_{id}}{n_d} \log \left( \frac{N}{n_i} \right)
\]

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

---

**Query expansion**

**Query:** golf green

**Results:**

- How can the grass on the greens at a golf course be so perfect?
- For example, a skilled golfer expects to reach the green on a par-four hole in ...
- Manufactures and sells synthetic golf putting greens and mats.

Irrelevant result can cause a ‘topic drift’:

- Volkswagen Golf, 1999, Green, 2000cc, petrol, manual, , hatchback, 94000miles, 2.0 GTi, 2 Registered Keepers, HPI Checked, Air Conditioning, Front and Rear Parking Sensors, ABS, Alarm, Alloy

---

**Query expansion**

**Results:**

- Spatial verification
- New results

Chum, Philbin, Sivic, Isard, Zisserman: Total Recall, ..., ICCV 2007
Recognition via alignment

**Pros:**
- Effective when we are able to find reliable features within clutter
- Great results for matching specific instances

**Cons:**
- Scaling with number of models
- Spatial verification as post-processing – not seamless, expensive for large-scale problems
- Not suited for generic category recognition

Summary

- Matching local invariant features
  - Useful to find objects and scenes
- Bag of words representation: quantize feature space to make discrete set of visual words
  - Summarize image by distribution of words
  - Index individual words
- Inverted index: pre-compute index to enable faster search at query time
- Recognition of instances via alignment: matching local features followed by spatial verification
  - Robust fitting: RANSAC, GHT

Making the Sky Searchable: Fast Geometric Hashing for Automated Astrometry

Sam Roweis, Dustin Lang & Keir Mierle
University of Toronto

David Hogg & Michael Blanton
New York University
Example

A shot of the Great Nebula, by Jerry Lodriguss (c.2006), from astropix.com
http://astrometry.net/gallery.html

Example

An amateur shot of M100, by Filippo Ciferri (c.2007) from flickr.com
http://astrometry.net/gallery.html

Example

A new full image of Bode’s Nebula (c.2007) by Peter Bresseler, from starlightfriend.de
http://astrometry.net/gallery.html
Today

• Generic object recognition

What does recognition involve?

Verification: is that a lamp?
Detection: are there people?

Identification: is that Potala Palace?

Object categorization

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.
Scene and context categorization

- outdoor
- city
- ...

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Instance-level recognition problem

John’s car

Generic categorization problem
### Object Categorization

- **Task Description**
  - "Given a small number of training images of a category, recognize a-priori unknown instances of that category and assign the correct category label."

- **Which categories are feasible visually?**

![Diagram of "Fido", German shepherd, dog, animal, living being]

---

### Visual Object Categories

- **Basic Level Categories in human categorization**
  - [Rosch 76, Lakoff 87]
    - The highest level at which category members have similar perceived shape
    - The highest level at which a single mental image reflects the entire category
    - The level at which human subjects are usually fastest at identifying category members
    - The first level named and understood by children
    - The highest level at which a person uses similar motor actions for interaction with category members

---

### Visual Object Categories

- Basic-level categories in humans seem to be defined predominantly visually.
- There is evidence that humans (usually) start with basic-level categorization before doing identification.
How many object categories are there?

~10,000 to 30,000

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Biederman 1987

Other Types of Categories

- Functional Categories
  - e.g. chairs = "something you can sit on"
Other Types of Categories

- Ad-hoc categories
  - e.g. “something you can find in an office environment”

Why recognition?

- Recognition a fundamental part of perception
  - e.g., robots, autonomous agents

- Organize and give access to visual content
  - Connect to information
  - Detect trends and themes

Posing visual queries

Kooaba, Bay & Quack et al.
Autonomous agents able to detect objects

Finding visually similar objects

Discovering visual patterns
Auto-annotation

Challenges: robustness
- Illumination
- Object pose
- Clutter
- Occlusions
- Intra-class appearance
- Viewpoint

Realistic scenes are crowded, cluttered, have overlapping objects.
Challenges: importance of context

Challenges: complexity

- **flickr**: 6 billion images
- **facebook**: 70 billion images
- **Imgur**: 1 billion images served daily
- **YouTube**: 10 billion images
- **photobucket**: (icon)

*Almost 90% of web traffic is visual!*

5/19/2015
Challenges: complexity

- Thousands to millions of pixels in an image
- 30+ degrees of freedom in the pose of articulated objects (humans)
- About half of the cerebral cortex in primates is devoted to processing visual information [Felleman and van Essen 1991]

Challenges: learning with minimal supervision

What works most reliably today

- Reading license plates, zip codes, checks

Source: Lana Lazebnik
What works most reliably today

- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection
- Recognition of flat textured objects (CD covers, book covers, etc.)

Source: Lana Lazebnik
What works most reliably today

- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection
- Recognition of flat textured objects (CD covers, book covers, etc.)
- Recognition of generic categories beginning to work!

Generic category recognition: basic framework

- Build/train object model
  - Choose a representation
  - Learn or fit parameters of model / classifier
- Generate candidates in new image
- Score the candidates

Generic category recognition: representation choice

Window-based	Part-based
Supervised classification

- Given a collection of labeled examples, come up with a function that will predict the labels of new examples.

  "four"  "nine"
  ?

  Training examples  Novel input

- How good is some function we come up with to do the classification?
- Depends on
  - Mistakes made
  - Cost associated with the mistakes

"four"  "nine"

Training examples  Novel input

Supervised classification

- Given a collection of labeled examples, come up with a function that will predict the labels of new examples.

  Feature value \( x \)

- Consider the two-class (binary) decision problem
  - \( L(4 \rightarrow 9) \): Loss of classifying a 4 as a 9
  - \( L(9 \rightarrow 4) \): Loss of classifying a 9 as a 4

- Risk of a classifier \( s \) is expected loss:

\[
R(s) = \Pr(4 \rightarrow 9 \mid x)L(4 \rightarrow 9) + \Pr(9 \rightarrow 4 \mid x)L(9 \rightarrow 4)
\]

- We want to choose a classifier so as to minimize this total risk

Supervised classification

- Optimal classifier will minimize total risk.

- At decision boundary, either choice of label yields same expected loss.

  If we choose class "four" at boundary, expected loss is:

\[
P(\text{class is } 9 \mid x)L(9 \rightarrow 4) + P(\text{class is } 4 \mid x)L(4 \rightarrow 4)
\]

  If we choose class "nine" at boundary, expected loss is:

\[
P(\text{class is } 4 \mid x)L(4 \rightarrow 9)
\]
Supervised classification

Optimal classifier will minimize total risk.

At decision boundary, either choice of label yields same expected loss.

So, best decision boundary is at point \( x \) where

\[
P(\text{class is } 9 \mid x) L(9 \to 4) = P(\text{class is } 4 \mid x) L(4 \to 9)
\]

To classify a new point, choose class with lowest expected loss; i.e., choose “four” if

\[
P(4 \mid x) L(4 \to 9) > P(9 \mid x) L(9 \to 4)
\]

How to evaluate these probabilities?

---

Probability

Basic probability

- \( X \) is a random variable
- \( P(X) \) is the probability that \( X \) achieves a certain value

\[
P(X)\quad \text{called a PDF – probability distribution/density function}
\]

- \( 0 \leq P(X) \leq 1 \)
- \( \int_{-\infty}^{\infty} P(X) dX = 1 \) or \( \sum P(X) = 1 \) for continuous \( X \) or discrete \( X \)

- Conditional probability: \( P(X \mid Y) \) – probability of \( X \) given that we already know \( Y \)

---
Example: learning skin colors

- We can represent a class-conditional density using a histogram (a “non-parametric” distribution)

Feature \( x \) = Hue

\[
P(x|\text{skin})
\]

\[
P(x|\text{not skin})
\]

Now we get a new image, and want to label each pixel as skin or non-skin. What’s the probability we care about to do skin detection?

Bayes rule

\[
P(\text{skin} \mid x) = \frac{P(x \mid \text{skin})P(\text{skin})}{P(x)}
\]

\[P(\text{skin} \mid x) \propto P(x \mid \text{skin})P(\text{skin})\]

Where does the prior come from?

Why use a prior?
Example: classifying skin pixels

Now for every pixel in a new image, we can estimate probability that it is generated by skin.

Classify pixels based on these probabilities

- if \( p(\text{skin} | x) > \theta \), classify as skin
- if \( p(\text{skin} | x) < \theta \), classify as not skin

Brighter pixels \( \rightarrow \) higher probability of being skin

---

Example: classifying skin pixels

Using skin color-based face detection and pose estimation as a video-based interface

Gary Bradski, 1998

---

Example: classifying skin pixels

Gary Bradski, 1998
Supervised classification

- Want to minimize the expected misclassification
- Two general strategies
  - Use the training data to build representative probability model; separately model class-conditional densities and priors (generative)
  - Directly construct a good decision boundary, model the posterior (discriminative)

Coming up

Face detection
Categorization with local features and part-based models
Deep convolutional neural networks

Questions?

See you Thursday!