History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features
- Present trends: combination of local and global methods, data-driven methods, context
Global scene descriptors

• The “gist” of a scene: Oliva & Torralba (2001)

http://people.csail.mit.edu/torralba/code/spatialenvelope/
Data-driven methods

J. Hays and A. Efros, Scene Completion using Millions of Photographs, SIGGRAPH 2007
Data-driven methods

(a) Query Image

(b) Retrieval Set

(c) Superpixels

(d) Per-class Likelihoods

(e) Final Labeling

J. Tighe and S. Lazebnik, ECCV 2010
Geometric context

What Matters in Recognition?

• Learning Techniques
  – E.g. choice of classifier or inference method

• Representation
  – Low level: SIFT, HoG, gist, edges
  – Mid level: Bag of words, sliding window, deformable model
  – High level: Contextual dependence

• Data
  – More is always better
  – Annotation is the hard part
What Matters in Scene Recognition?

• Learning Techniques
  – ?

• Representation
  – ?

• Data
  – ?
Large-scale Instance Retrieval

Computer Vision
CS 143, Brown

James Hays

Many slides from Derek Hoiem and Kristen Grauman
Multi-view matching

Matching two given views for depth vs Search for a matching view for recognition

Kristen Grauman
How to quickly find images in a large database that match a given image region?
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/vgoogle/index.html
Example Applications

Mobile tourist guide
- Self-localization
- Object/building recognition
- Photo/video augmentation

[Quack, Leibe, Van Gool, CIVR’08]
Application: Large-Scale Retrieval

Query

Results from 5k Flickr images (demo available for 100k set)

[Philbin CVPR’07]
Application: Image Auto-Annotation

Left: Wikipedia image
Right: closest match from Flickr

Moulin Rouge
Old Town Square (Prague)
Tour Montparnasse
Colosseum
Viktualienmarkt Maypole

K. Grauman, B. Leibe
Google Goggles
Use pictures to search the web.

Get Google Goggles
Android (1.6+ required)
Download from Android Market.

Get Google Goggles to Android phone

New! iPhone (iOS 4.0 required)
Download from the App Store.

Send Goggles to iPhone

New! [Text] [Landmarks] [Books] [Contact Info] [Artwork] [Wine] [Logos]
Simple idea

See how many keypoints are close to keypoints in each other image

Lots of Matches

Few or No Matches

But this will be really, really slow!
Indexing local features

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

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

Easily can have millions of features to search!
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”.

Kristen Grauman
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
Visual vocabulary formation

Issues:
• Vocabulary size, number of words
• Sampling strategy: where to extract features?
• Clustering / quantization algorithm
• Unsupervised vs. supervised
• What corpus provides features (universal vocabulary?)
Inverted file index

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

<table>
<thead>
<tr>
<th>Word #</th>
<th>Image #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1, 2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>2</td>
</tr>
</tbody>
</table>
New query image is mapped to indices of database images that share a word.
Instance recognition: remaining issues

• How to summarize the content of an entire image? And gauge overall similarity?

• How large should the vocabulary be? How to perform quantization efficiently?

• Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?

• How to score the retrieval results?
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 through our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain just as a movie strip is projected onto a screen. Through the discoveries of Hubel and Wiesel we now know that the mechanism of perception in the brain is considerably more complex. The visual impulses following the 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 step-wise analysis in a system of nerve cells stored in columns. 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 a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to 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 needs to ramp up domestic demand so more goods stay 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, Beijing 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.
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

Kristen Grauman
Inverted file index and bags of words similarity

1. Extract words in query
2. Inverted file index to find relevant frames
3. Compare word counts
Instance recognition: remaining issues

• How to summarize the content of an entire image? And gauge overall similarity?

• How large should the vocabulary be? How to perform quantization efficiently?

• Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?

• How to score the retrieval results?
Vocabulary size

Results for recognition task with 6347 images

Influence on performance, sparsity
Recognition with K-tree

Following slides by David Nister (CVPR 2006)
Vocabulary trees: complexity

Number of words given tree parameters: branching factor and number of levels

Word assignment cost vs. flat vocabulary
110,000,000 Images in 5.8 Seconds

Slide Credit: Nister
Performance

Database Size

Performance (%)

Full set
Test set

ImageSearch at the VizCentre

New query:

Browse... Send File

File is 560x320

Top n results of your query.

bourne/m1000043322.pgm  bourne/m1000043323.pgm  bourne/m1000043326.pgm  bourne/m1000043327.pgm

UK Center for Visualization & Virtual Environments
Higher branch factor works better (but slower)
Sampling strategies

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]
Instance recognition: remaining issues

• How to summarize the content of an entire image? And gauge overall similarity?

• How large should the vocabulary be? How to perform quantization efficiently?

• Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?

• How to score the retrieval results?
Can we be more accurate?

So far, we treat each image as containing a “bag of words”, with no spatial information.
Can we be more accurate?

So far, we treat each image as containing a “bag of words”, with no spatial information.

Real objects have consistent geometry.
Both image pairs have many visual words in common.
Spatial Verification

Only some of the matches are mutually consistent

Slide credit: Ondrej Chum
Instance recognition: remaining issues

- How to summarize the content of an entire image? And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?
Scoring retrieval quality

Database size: 10 images
Relevant (total): 5 images

Results (ordered):

precision = #relevant / #returned
recall = #relevant / #total relevant

Slide credit: Ondrej Chum
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. This will 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, Beijing has made it clear that it will take its time and tread carefully before 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)

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

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

Query: \textit{golf green}

Results:

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

Irrelevant result can cause a `topic drift':

Query Expansion

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

Slide credit: Ondrej Chum
Things to remember

• Object instance recognition
  – Find keypoints, compute descriptors
  – Match descriptors
  – Vote for / fit affine parameters
  – Return object if # inliers > T

• Keys to efficiency
  – Visual words
    • Used for many applications
  – Inverse document file
    • Used for web-scale search