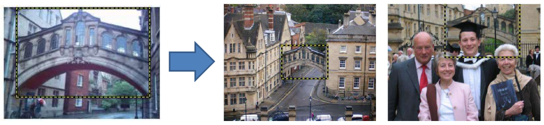
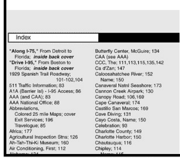
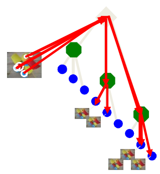


Instance recognition

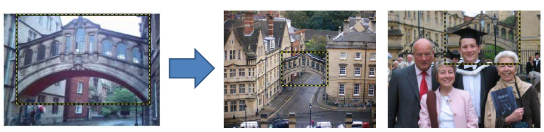


Thurs April 6
Kristen Grauman
UT Austin



Instance recognition

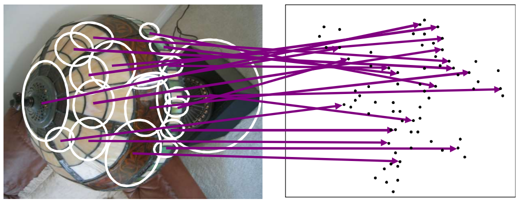
– Indexing local features efficiently (last time)
– Spatial verification models



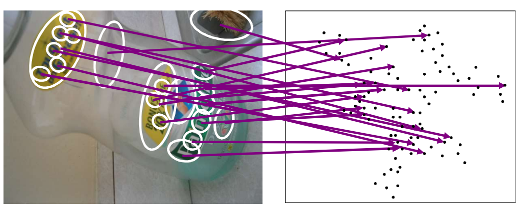
Picking up from last time

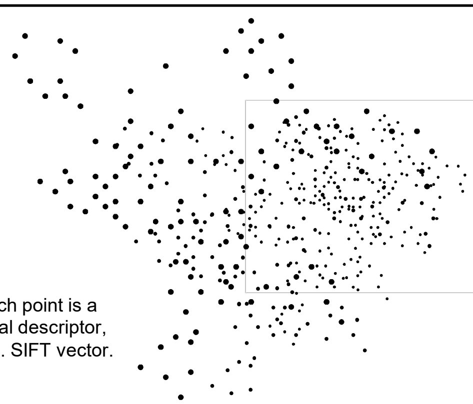
- Instance recognition wrap up:
  - Impact of vocabulary tree
  - Spatial verification
  - Sky mapping example
  - Query expansion

Visual words: main idea

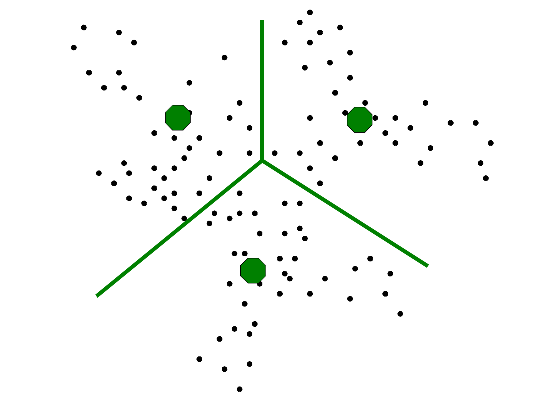


Visual words: main idea



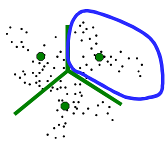


Each point is a local descriptor, e.g. SIFT vector.



Visual words

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



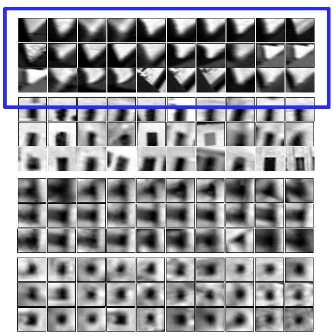
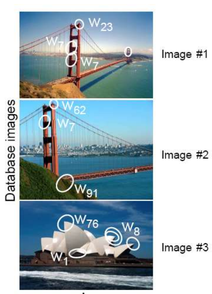


Figure from Sivic & Zisserman, ICCV 2003

Inverted file index



Word #	Image #
1	3
2	
…	
7	1, 2
8	3
9	
10	
…	
91	2

…

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

Slide credit: Kristen Grauman

Inverted file index

New query image

Word #	Image #
1	3
2	
7	1, 2
8	3
9	
10	
…	
91	2

…

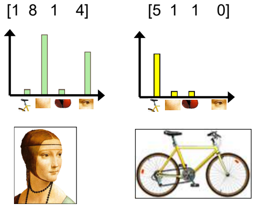
- New query image is mapped to indices of database images that share a word.

Slide credit: Kristen Grauman

Comparing bags of words

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

[1 8 1 4] [5 1 1 0]



$\vec{d}_j$ $\vec{q}$

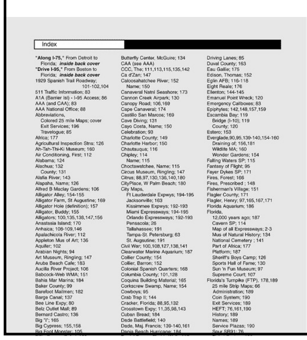
$$sim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|}$$

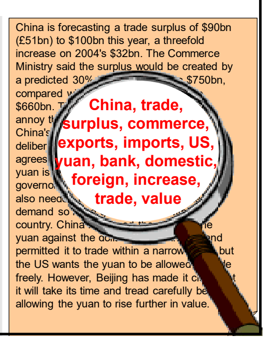
$$= \frac{\sum_{i=1}^{V} d_j(i) * q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} * \sqrt{\sum_{i=1}^{V} q(i)^2}}$$

for vocabulary of *V* words

Slide credit: Kristen Grauman

What else can we borrow from text retrieval?





Query expansion

Query: ***golf green***

Results:

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

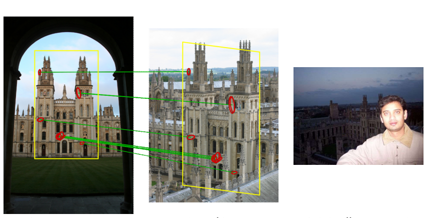
Slide credit: Ondrej Chum

Query Expansion

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

Slide credit: Ondrej Chum

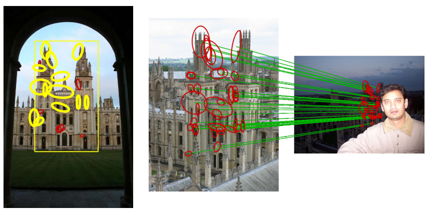
Query Expansion Step by Step



Query Image Retrieved image Originally not retrieved

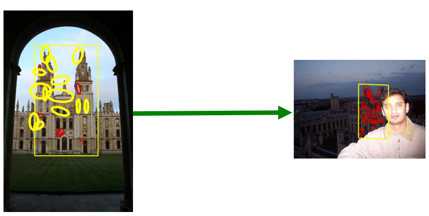
Slide credit: Ondrej Chum

Query Expansion Step by Step



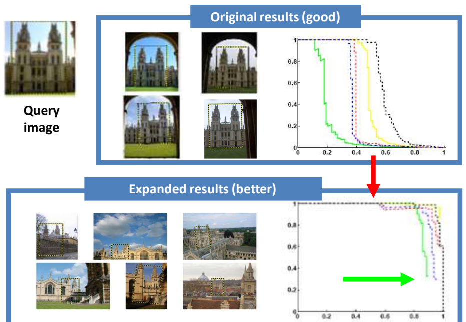
Slide credit: Ondrej Chum

Query Expansion Step by Step



Slide credit: Ondrej Chum

Query Expansion Results



Slide credit: Ondrej Chum

Instance recognition: remaining issues

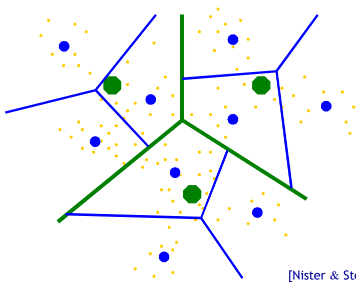
- How to summarize the content of an entire image? And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?

Slide credit: Kristen Grauman

Visual Object Recognition Tutorial

Vocabulary Trees: hierarchical clustering for large vocabularies

- **Tree construction:**



[Nister & Stewenius, CVPR'06]

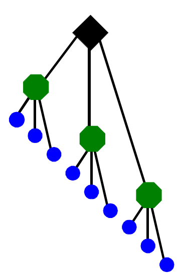
K. Grauman, B. Leibe

Slide credit: David Nister

Visual Object Recognition Tutorial

Vocabulary Tree

- **Training: Filling the tree**



[Nister & Stewenius, CVPR'06]

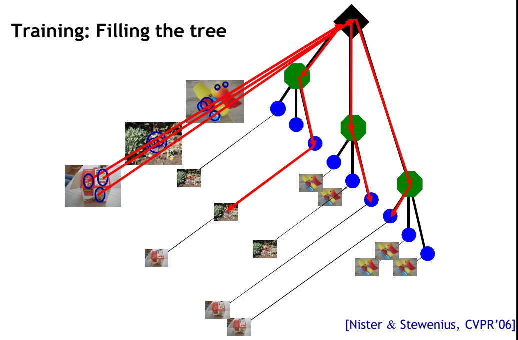
K. Grauman, B. Leibe

Slide credit: David Nister

Visual Object Recognition Tutorial

Vocabulary Tree

- Training: Filling the tree



[Nister & Stewenius, CVPR'06]

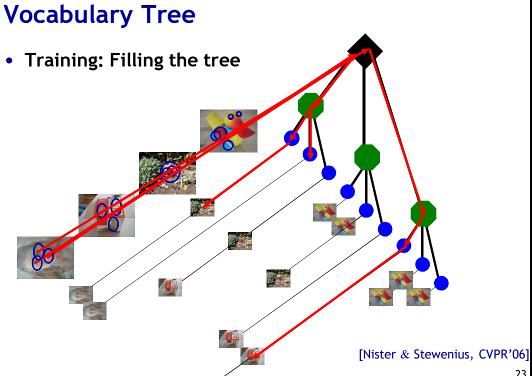
K. Grauman, B. Leibe

Slide credit: David Nister

Visual Object Recognition Tutorial

Vocabulary Tree

- Training: Filling the tree



[Nister & Stewenius, CVPR'06]

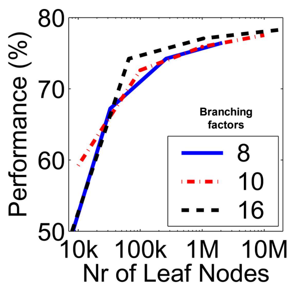
K. Grauman, B. Leibe

Slide credit: David Nister

23

What is the computational advantage of the hierarchical representation bag of words, vs. a flat vocabulary?

Vocabulary size



Results for recognition task with 6347 images

Influence on performance, sparsity?

Nister & Stewenius, CVPR 2006

Bags of words: pros and cons

+ flexible to geometry / deformations / viewpoint
+ compact summary of image content
+ provides vector representation for sets
+ very good results in practice

- basic model ignores geometry – must verify afterwards, or encode via features
- background and foreground mixed when bag covers whole image
- optimal vocabulary formation remains unclear

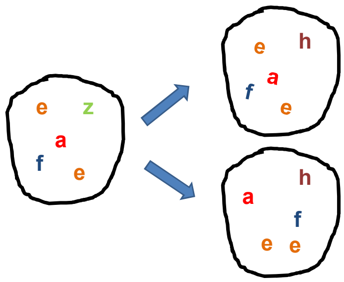
Slide credit: Kristen Grauman

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?

Slide credit: Kristen Grauman

Which matches better?



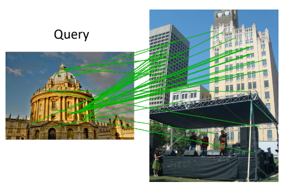
Derek Hoiem

Spatial Verification

Query

DB image with high BoW similarity

Query



DB image with high BoW similarity

Both image pairs have many visual words in common.

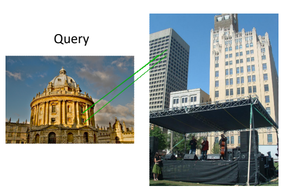
Slide credit: Ondrej Chum

Spatial Verification

Query

DB image with high BoW similarity

Query



DB image with high BoW similarity

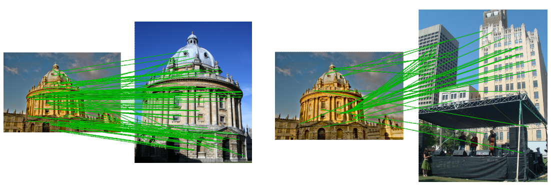
Only some of the matches are mutually consistent

Slide credit: Ondrej Chum

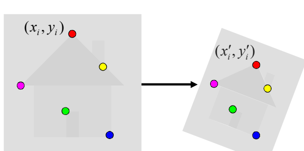
Spatial Verification: two basic strategies

- 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



Recall: Fitting an affine transformation

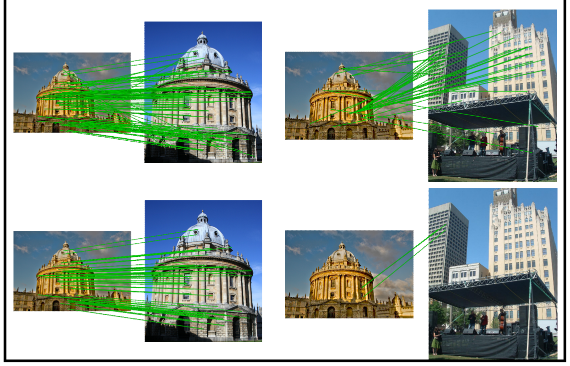


Approximates viewpoint changes for roughly planar objects and roughly orthographic cameras.

$$\begin{bmatrix} x_i' \\ y_i' \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\begin{bmatrix} & & & \cdots & & \\ x_i & y_i & 0 & 0 & 1 & 0 \\ 0 & 0 & x_i & y_i & 0 & 1 \\ & & & \cdots & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} \cdots \\ x_i' \\ y_i' \\ \cdots \end{bmatrix}$$

RANSAC verification

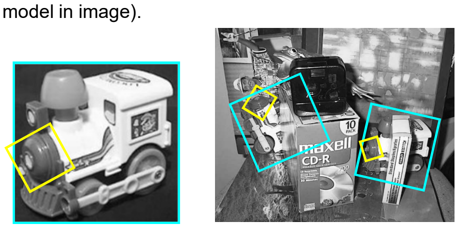


Spatial Verification: two basic strategies

- 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

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).



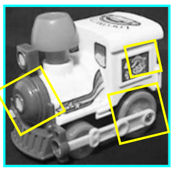
Model

Novel image

Adapted from Lana Lazebnik

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

Gen Hough Transform details (Lowe's system)

- **Training phase:** For each model feature, record 2D location, scale, and orientation of model (relative to normalized feature frame)
- **Test phase:** Let each match btwn a test SIFT feature and a model feature vote in a 4D Hough space
  - Use broad bin sizes of 30 degrees for orientation, a factor of 2 for scale, and 0.25 times image size for location
  - Vote for two closest bins in each dimension
- Find all bins with at least three votes and perform geometric verification
  - Estimate least squares *affine* transformation
  - Search for additional features that agree with the alignment

David G. Lowe. **"Distinctive image features from scale-invariant keypoints."** *IJCV* 60 (2), pp. 91-110, 2004.

Slide credit: Lana Lazebnik

Recall: difficulties of voting

- Noise/clutter can lead to as many votes as true target
- Bin size for the accumulator array must be chosen carefully
- In practice, good idea to make broad bins and spread votes to nearby bins, since verification stage can prune bad vote peaks.

Example result

Background subtract for model boundaries

Objects recognized,

Recognition in spite of occlusion

[Lowe]

Gen Hough vs RANSAC

GHT

- Single correspondence -> vote for all consistent parameters
- Represents uncertainty in the model parameter space
- Linear complexity in number of correspondences and number of voting cells; beyond 4D vote space impractical
- Can handle high outlier ratio

RANSAC

- Minimal subset of correspondences to estimate model -> count inliers
- Represents uncertainty in image space
- Must search all data points to check for inliers each iteration
- Scales better to high-d parameter spaces

Slide credit: Kristen Grauman

Instance recognition applications

- Snap, pick, pay

- https://www.usatoday.com/videos/tech/2014/10/31/18261641/

Slide credit: Kristen Grauman

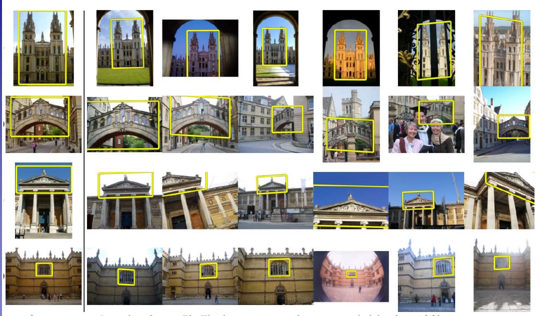
Example Applications

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

B. Leibe [Quack, Leibe, Van Gool, CIVR'08]

Visual Object Recognition Tutorial

Application: Large-Scale Retrieval



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

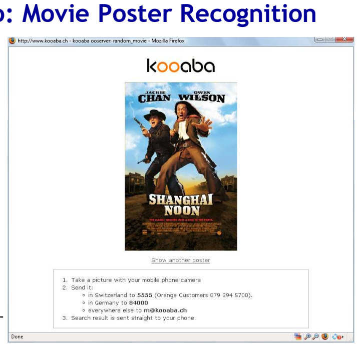
[Philbin CVPR'07]

Visual Object Recognition Tutorial

Web Demo: Movie Poster Recognition

50'000 movie posters indexed

Query-by-image from mobile phone available in Switzer-land



http://www.kooaba.com/en/products_engine.html#

Visual Object Recognition Tutorial

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?

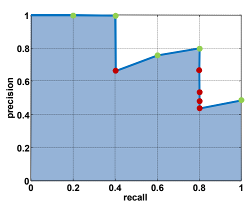
Kristen Grauman

Scoring retrieval quality

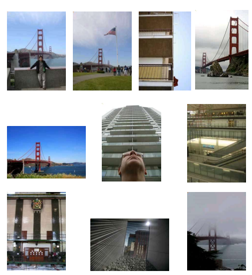
Query

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

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



Results (ordered):



Slide credit: Ondrej Chum

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 category recognition.

Summary

- **Matching local invariant features**
  - Useful not only to provide matches for multi-view geometry, but also 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

Kristen Grauman

Coming up

- Mining and visual pattern discovery
- Category recognition / supervised learning
- Sliding window object detection (Faces!)