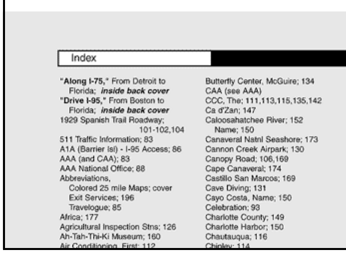
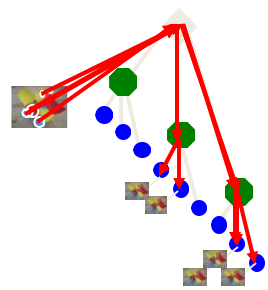
Instance recognition

Thurs Oct 29



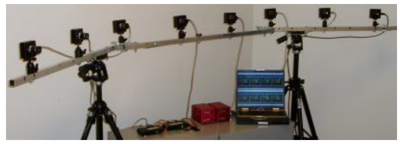
Index

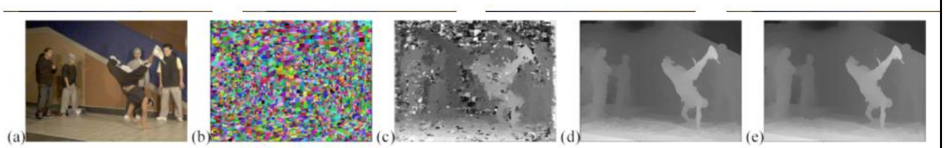


Last time

- Depth from stereo: main idea is to triangulate from corresponding image points.
- Epipolar geometry defined by two cameras
  - We've assumed known extrinsic parameters relating their poses
- Epipolar constraint limits where points from one view will be imaged in the other
  - Makes search for correspondences quicker
- To estimate depth
  - Limit search by epipolar constraint
  - Compute correspondences, incorporate matching preferences

Virtual viewpoint video





(a
Figure 6: *Sample results from stereo reconstruction stage: (a) input color image; (b) color-based segmentation; (c) initial disparity estimates* $\hat{d}_{ij}$*; (d) refined disparity estimates; (e) smoothed disparity estimates* $d_i(x)$.
Fig
d) A depth-matted object from earlier in the sequence is inserted into the video.

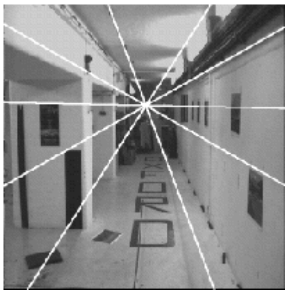
C. Zitnick et al, High-quality video view interpolation using a layered representation, SIGGRAPH 2004.

Virtual viewpoint video

C. Larry Zitnick et al, High-quality video view interpolation using a layered representation, SIGGRAPH 2004.

http://research.microsoft.com/IVM/VVV/

Review questions:
What stereo rig yielded these epipolar lines?



Epipole has same coordinates in both images.
Points move along lines radiating from e: “Focus of expansion”

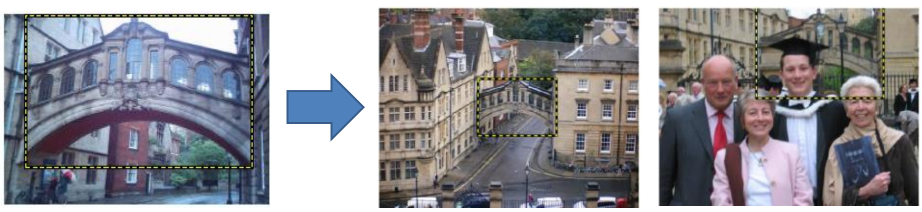
Figure from Hartley & Zisserman

Review questions

- When solving for stereo, when is it necessary to break the soft disparity gradient constraint?
- What can cause a disparity value to be undefined?
- What parameters relating the two cameras in the stereo rig must be known (or inferred) to compute depth?

Today

- Instance recognition
  - Indexing local features efficiently
  - Spatial verification models



Recognizing or retrieving specific objects

Example I: Visual search in feature films

Visually defined query

"Groundhog Day" [Rammis, 1993]

"Find this clock"

"Find this place"

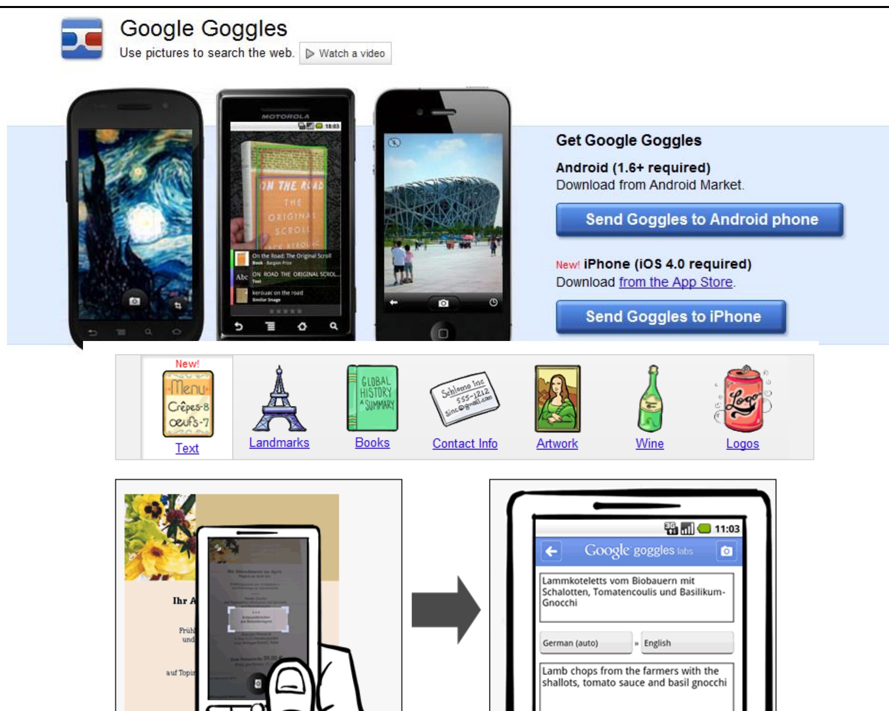
Slide credit: J. Sivic

Recognizing or retrieving specific objects

Example II: Search photos on the web for particular places

Find these landmarks ...in these images and 1M more

Slide credit: J. Sivic



Recall: matching local features

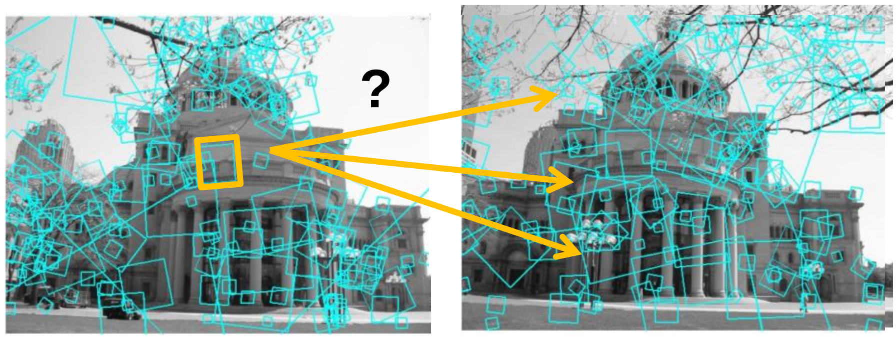
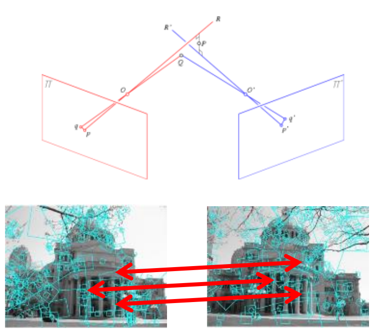


Image 1 Image 2

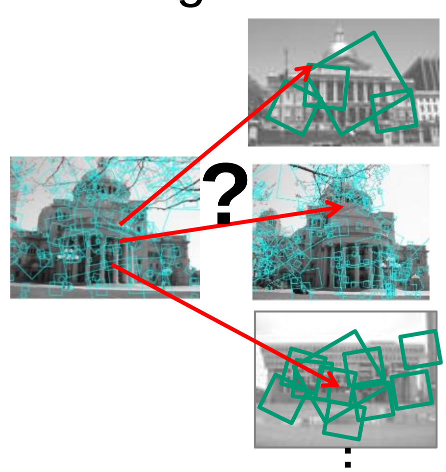
To generate **candidate matches**, find patches that have the most similar appearance (e.g., lowest SSD)

Simplest approach: compare them all, take the closest (or closest k, or within a thresholded distance)

Multi-view matching



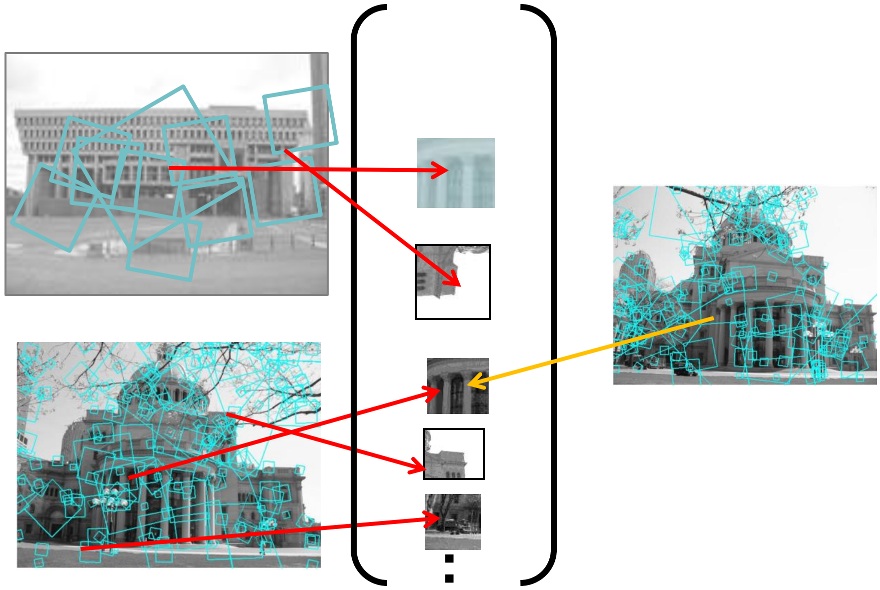
vs



Matching two given views for depth

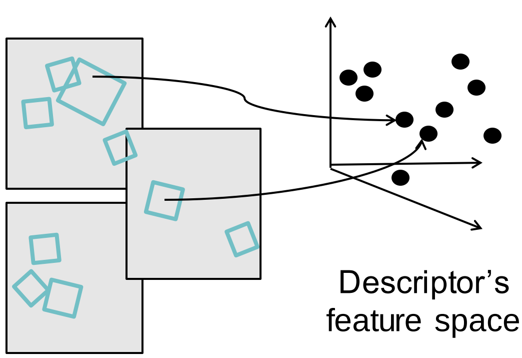
Search for a matching view for recognition

Indexing local features



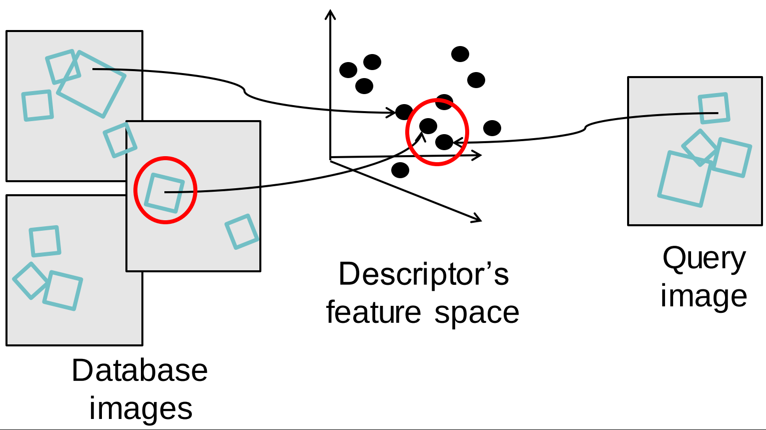
Indexing local features

- 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.

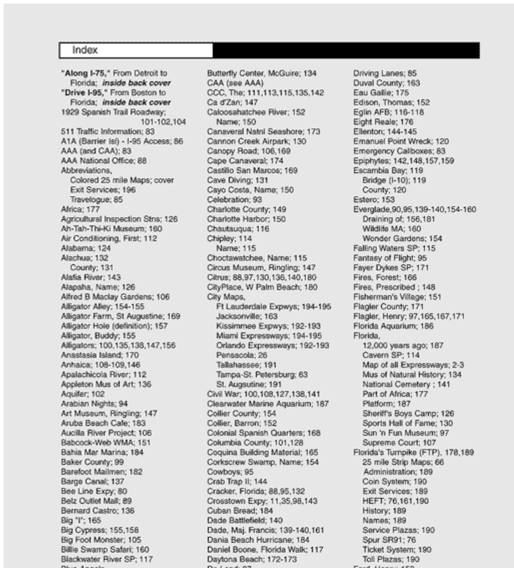


Indexing local features

- 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?

- Possible solutions:
  - Inverted file
  - Nearest neighbor data structures
    - Kd-trees
    - Hashing

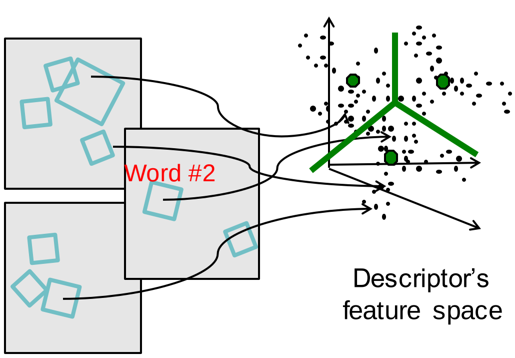
Indexing local features: inverted file index


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".

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.

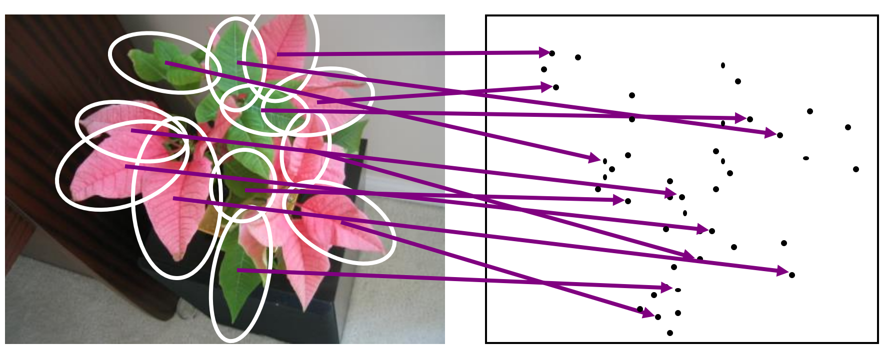
Visual words: main idea

- Extract some local features from a number of images …

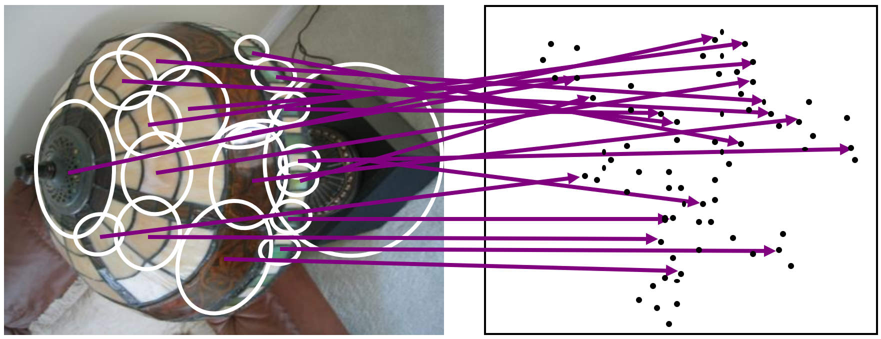
e.g., SIFT descriptor space: each point is 128-dimensional

Slide credit: D. Nister, CVPR 2006

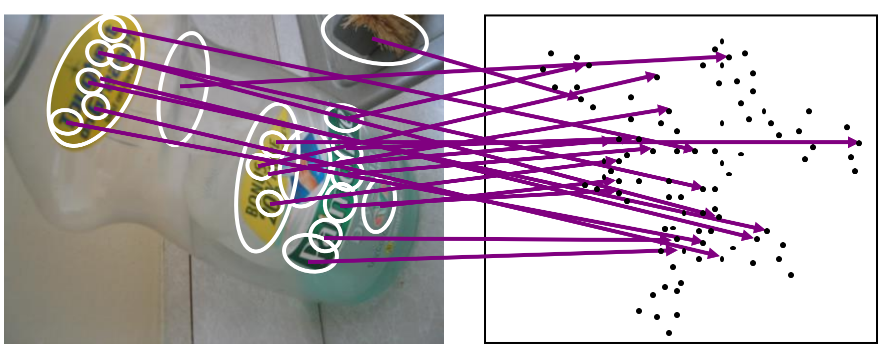
Visual words: main idea

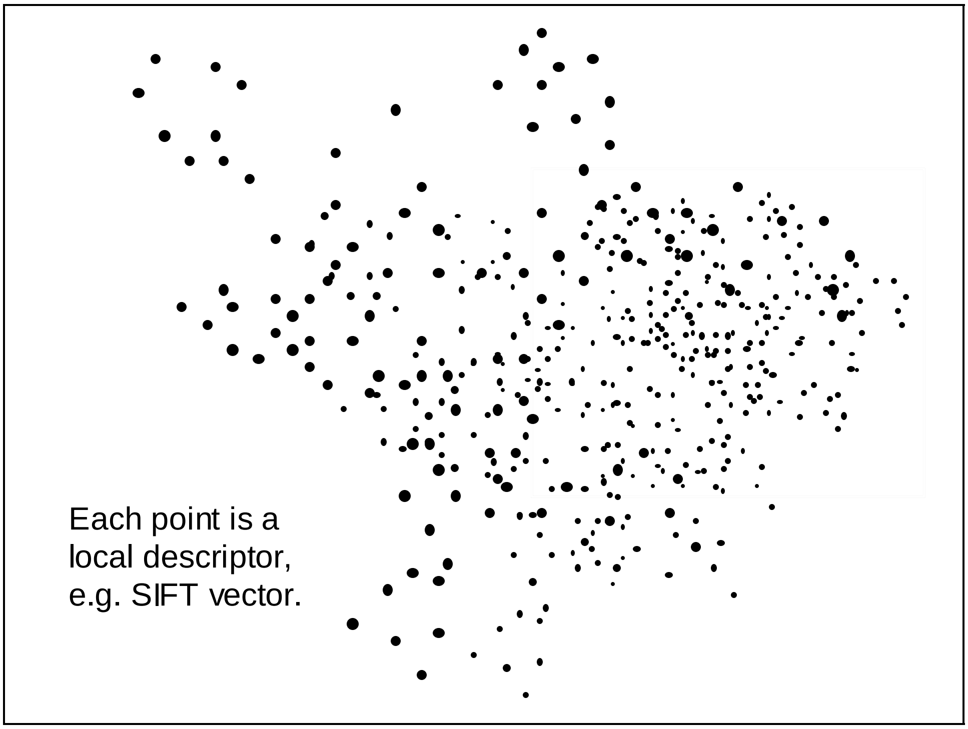


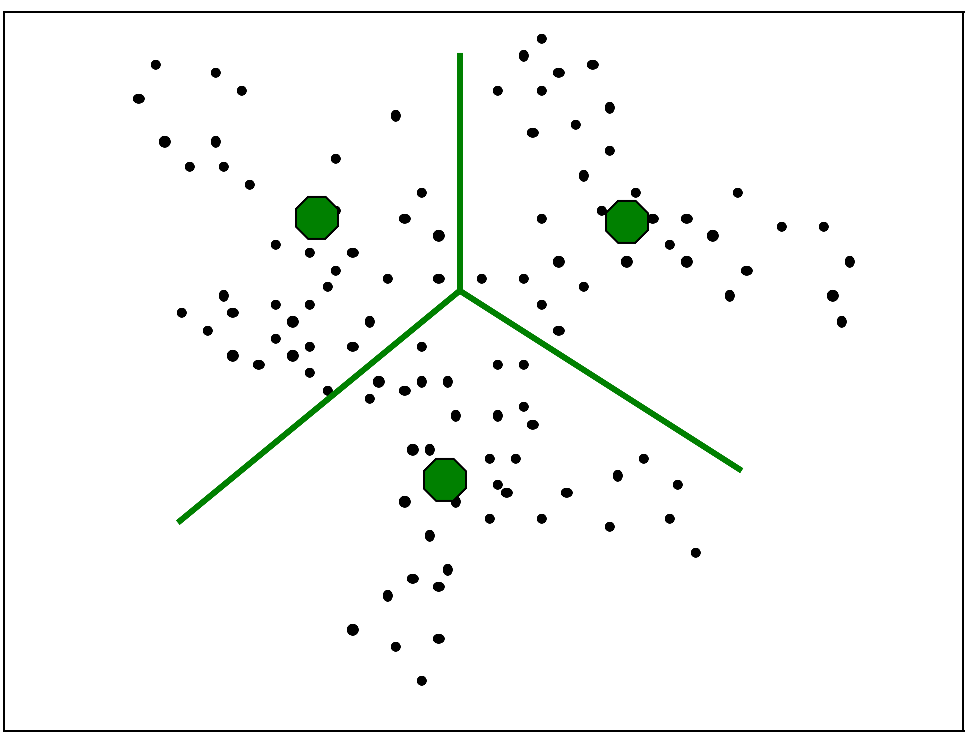
Visual words: main idea



Visual words: main idea

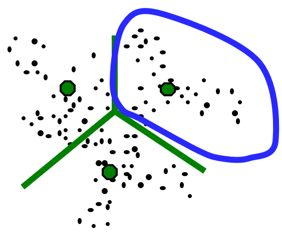



Each point is a
local descriptor,
e.g. SIFT vector.



Visual words

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



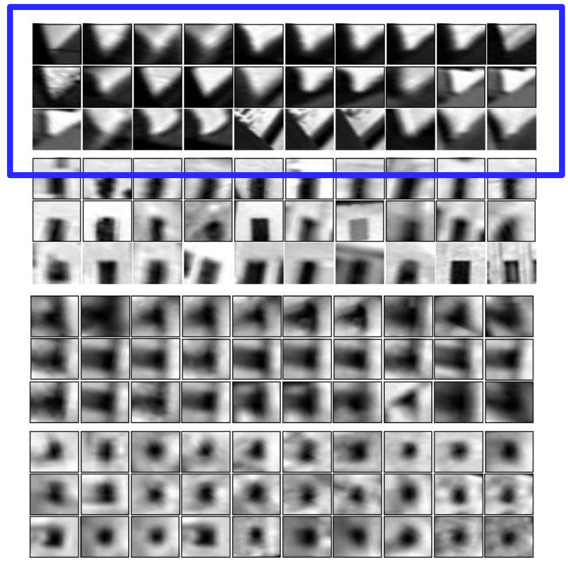
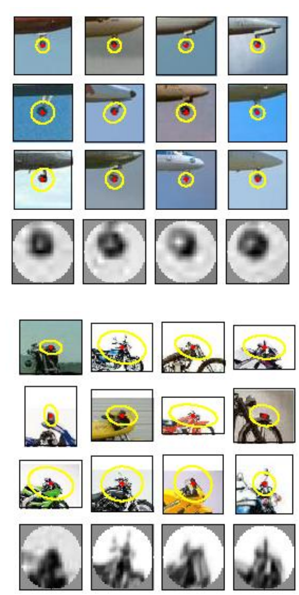


Figure from Sivic & Zisserman, ICCV 2003

Visual words

- Also used for describing scenes and object **categories** for the sake of indexing or classification.

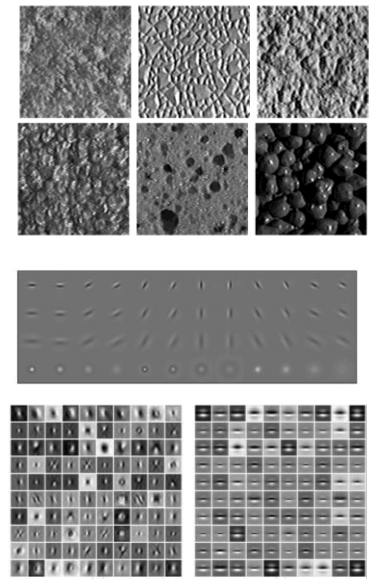


Sivic & Zisserman 2003; Csurka, Bray, Dance, & Fan 2004; many others.

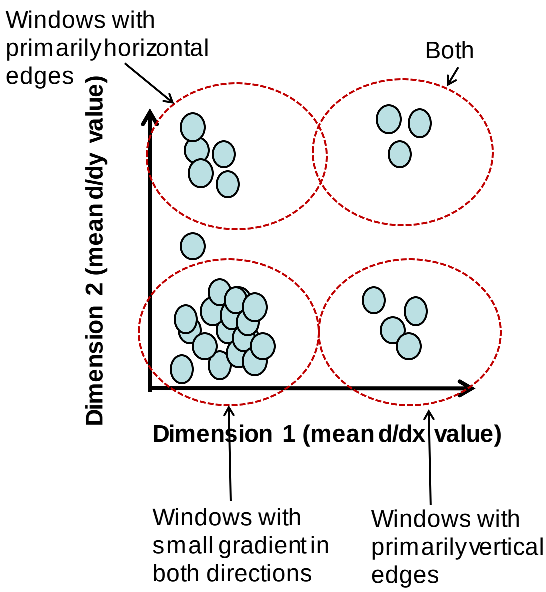
Visual words and textons

- First explored for texture and material representations
- *Texton* = cluster center of filter responses over collection of images
- Describe textures and materials based on distribution of prototypical texture elements.

Leung & Malik 1999; Varma & Zisserman, 2002



Recall: Texture representation example



	mean d/dx value	mean d/dy value
Win. #1	4	10
Win.#2	18	7
⋮		
Win.#9	20	20

⋮

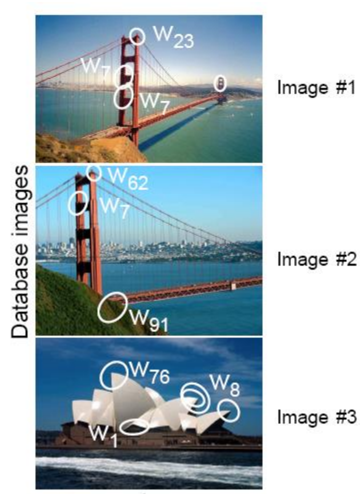
statistics to summarize patterns in small windows

Visual vocabulary formation

Issues:

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

Inverted file index



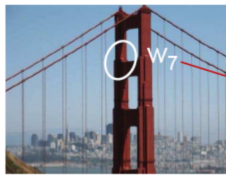
Word #	Image #
1	3
2 …	
7	1, 2
8	3
9	
10 …	
91	2

…

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

Inverted file index

When will this give us a significant gain in efficiency?



New query image

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

…

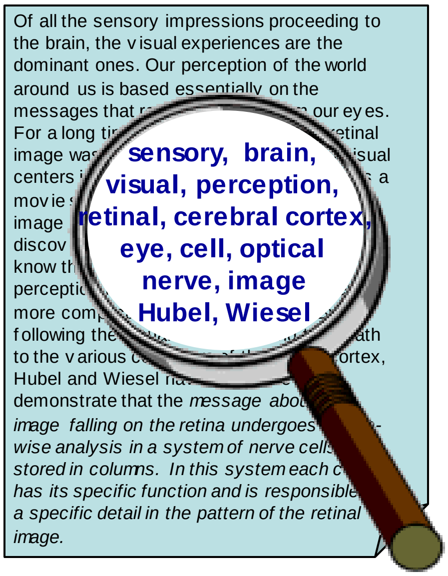
- New query image is mapped to indices of database images that share a word.

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

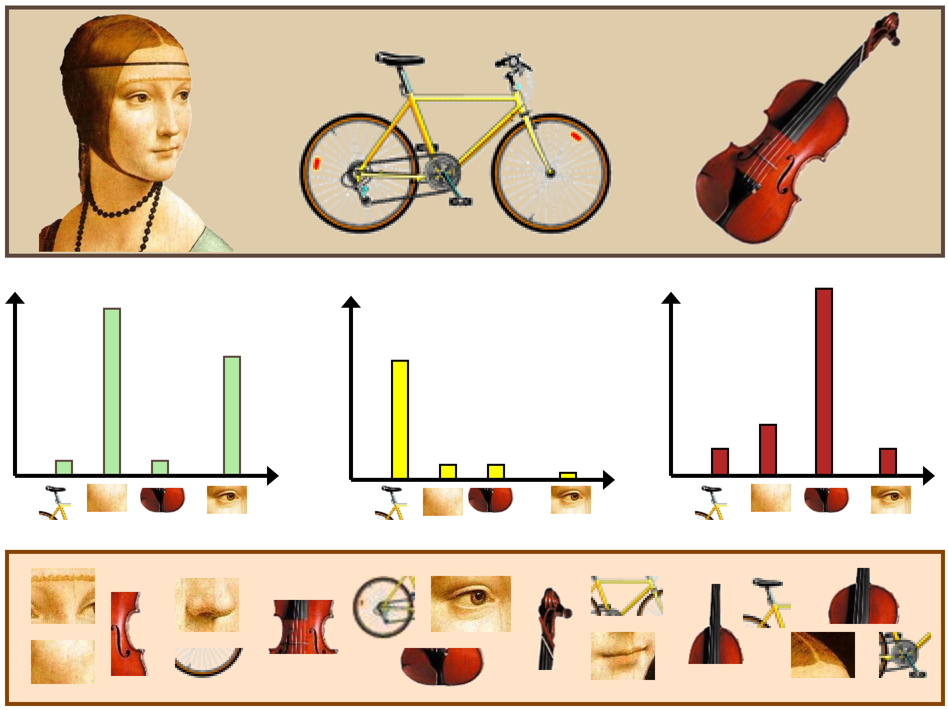
Analogy to documents



ICCV 2005 short course, L. Fei-Fei

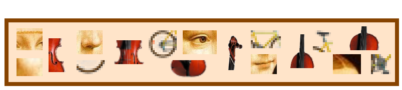
Object → **Bag of 'words'**

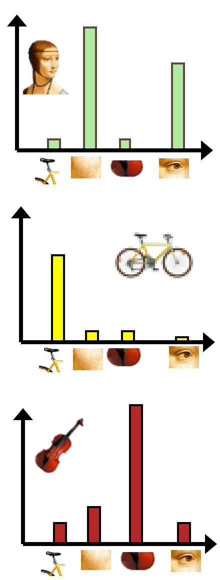
ICCV 2005 short course, L. Fei-Fei



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.

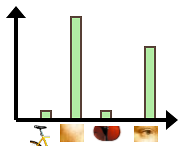




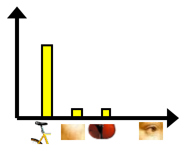
Comparing bags of words

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

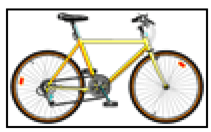
[1 8 1 4]



[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

tf-idf weighting

- **T**erm **f**requency – **i**nverse **d**ocument **f**requency
- Describe frame by frequency of each word within it, downweight words that appear often in the database
- (Standard weighting for text retrieval)

Number of occurrences of word i in document d

Number of words in document d

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

Total number of documents in database

Number of documents word i occurs in, in whole database

Inverted file index and bags of words similarity

New query image

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

1. Extract words in query
2. Inverted file index to find relevant frames
3. Compare word counts

Kristen Grauman

Bags of words for content-based image retrieval

Visually defined query

"Groundhog Day" [Rammis, 1993]

"Find this clock"

"Find this place"

Slide from Andrew Zisserman
Sivic & Zisserman, ICCV 2003

Example

retrieved shots

Slide from Andrew Zisserman
Sivic & Zisserman, ICCV 2003

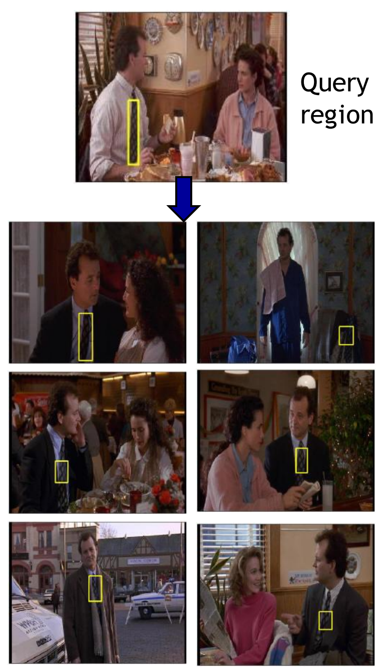
Visual Object Recognition Tutorial

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



Query region

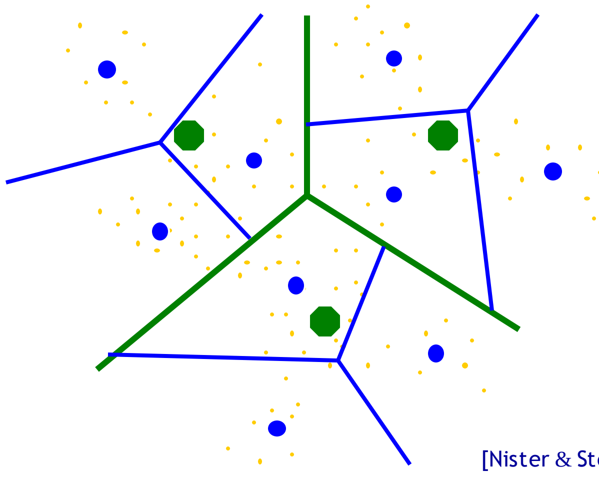
Retrieved frames

K. Grauman, B. Leibe

46

Vocabulary Trees: hierarchical clustering for large vocabularies

- Tree construction:



[Nister & Stewenius, CVPR'06]

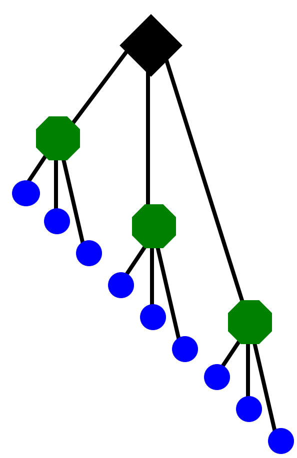
K. Grauman, B. Leibe

Slide credit: David Nister

Visual Object Recognition Tutorial

Vocabulary Tree

- Training: Filling the tree



[Nister & Stewenius, CVPR'06]

K. Grauman, B. Leibe

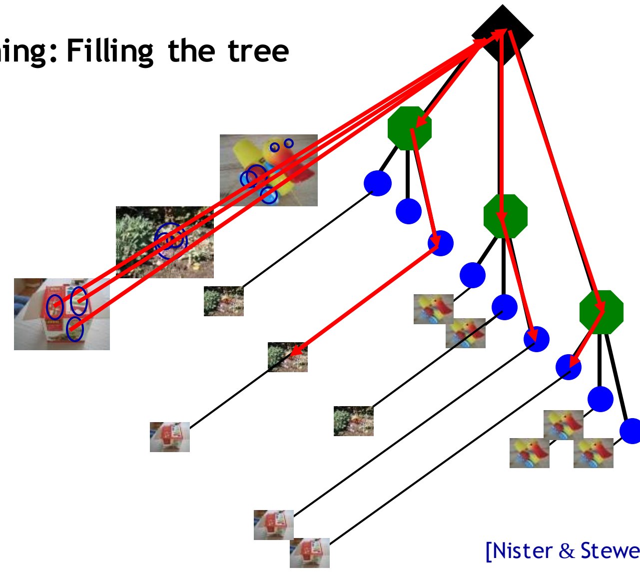
Slide credit: David Nister

Visual Object Recognition Tutorial

Visual Object Recognition Tutorial

Vocabulary Tree

- **Training: Filling the tree**



[Nister & Stewenius, CVPR'06]

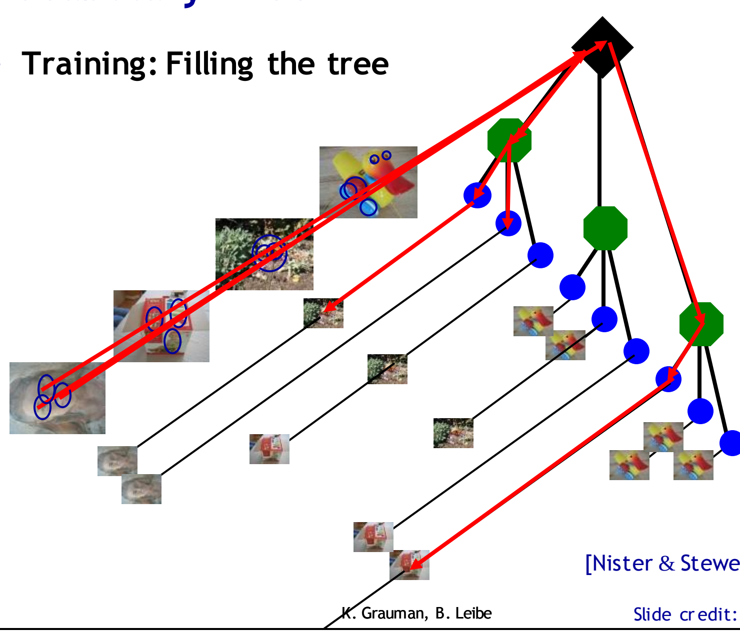
K. Grauman, B. Leibe

Slide credit: David Nister

Visual Object Recognition Tutorial

Vocabulary Tree

- **Training: Filling the tree**



[Nister & Stewenius, CVPR'06]

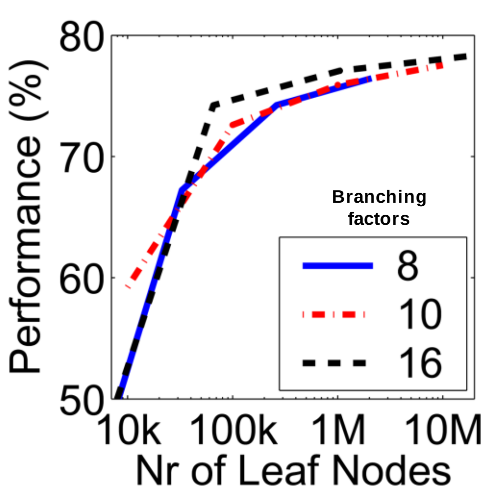
K. Grauman, B. Leibe

Slide credit: David Nister

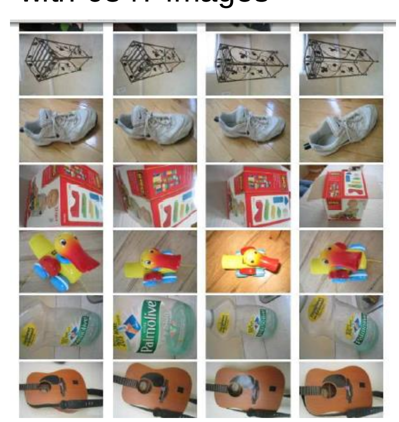
50

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

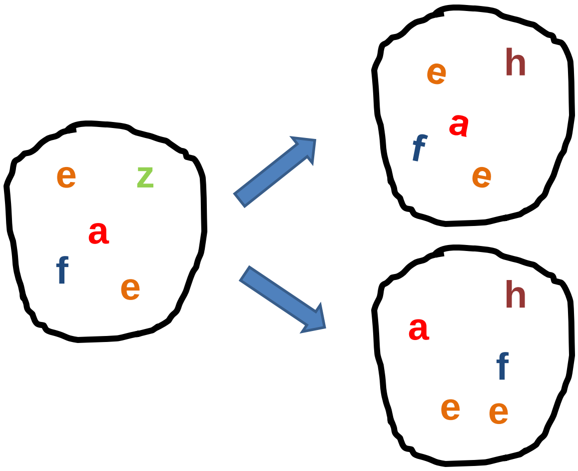
- basic model ignores geometry – must verify afterwards, or encode via features
- background and foreground mixed when bag covers whole image
- optimal vocabulary formation remains unclear

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

Which matches better?



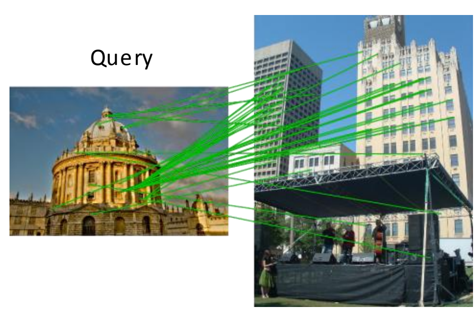
Derek Hoiem

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.

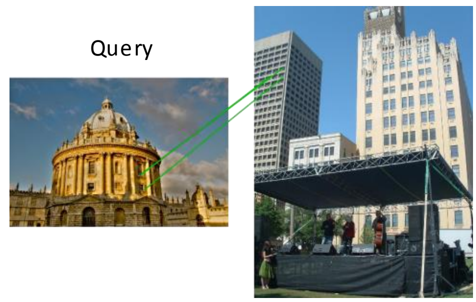
Slide credit: Ondrej Chum

Spatial Verification

Query

DB image with high BoW similarity

Query



DB image with high BoW similarity

Only some of the matches are mutually consistent

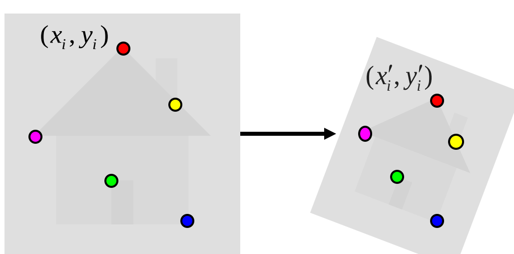
Slide credit: Ondrej Chum

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

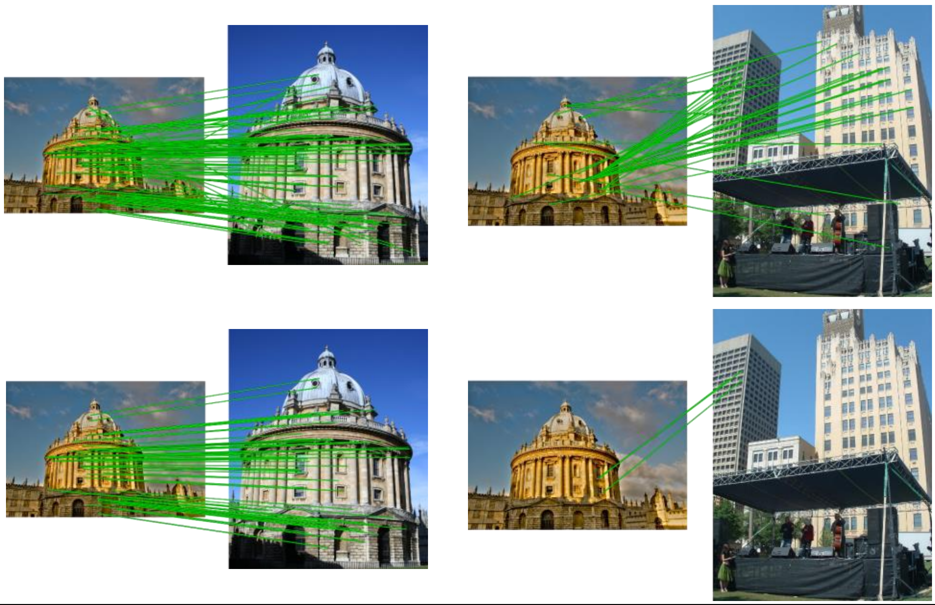


Approximates viewpoint changes for roughly planar objects and roughly orthographic cameras.

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

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

RANSAC verification

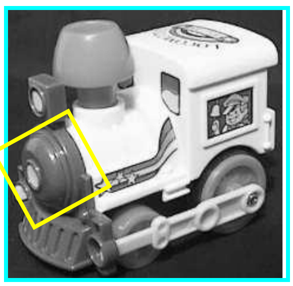


Spatial Verification: two basic strategies

- 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).



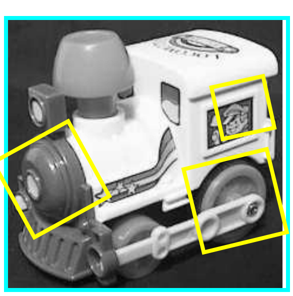
Model

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

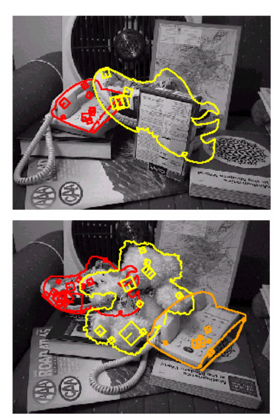
David G. Lowe. **"Distinctive image features from scale-invariant keypoints."** *IJCV* 60 (2), pp. 91-110, 2004.

Slide credit: Lana Lazebnik

Example result

Background subtract for model boundaries

Objects recognized,



Recognition in spite of occlusion

[Lowe]

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.

Visual Object Recognition Tutorial

Example Applications

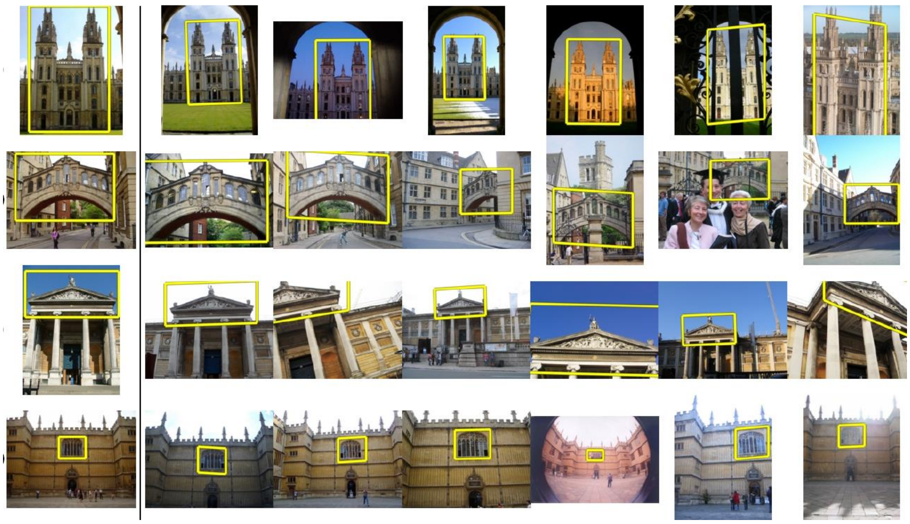
Mobile tourist guide
- Self-localization
- Object/building recognition
- Photo/video augmentation

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)

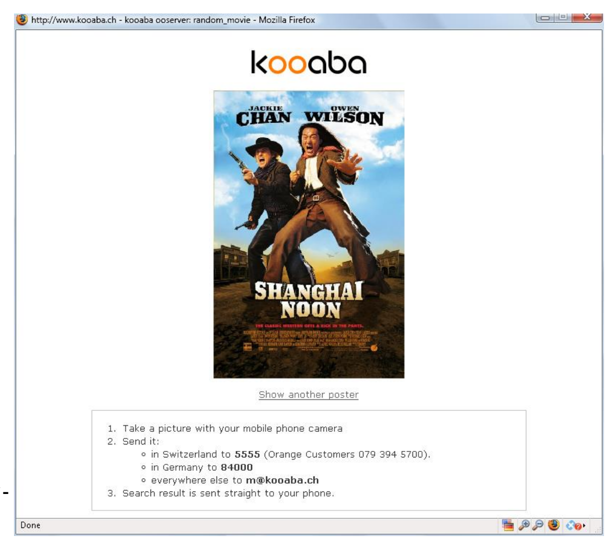
[Philbin CVPR'07]

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#

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?

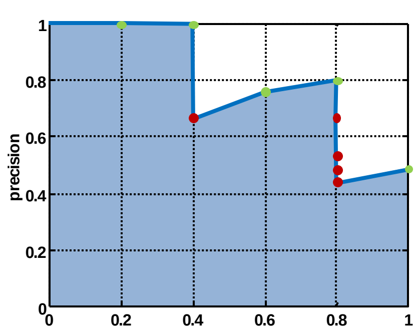
Kristen Grauman

Scoring retrieval quality

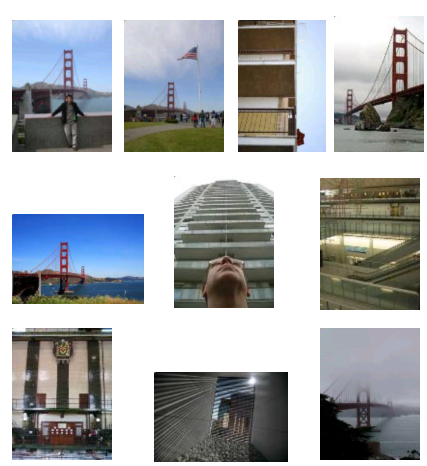
Query

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

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



Results (ordered):



Slide credit: Ondrej Chum

Recognition via alignment

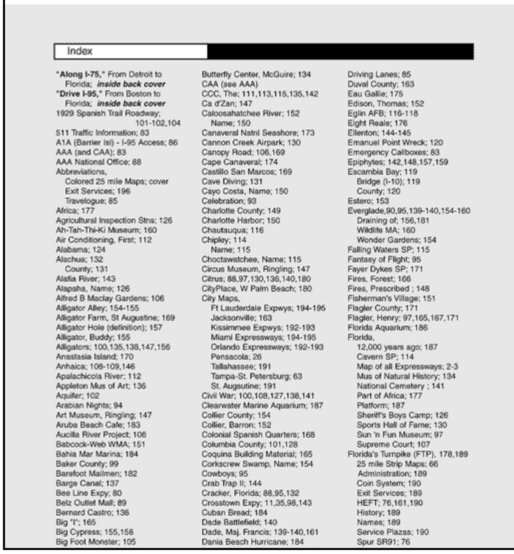
Pros:

- 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.

What else can we borrow from text retrieval?



Index

"**Along I-75,**" From Detroit to Florida; ***inside back cover***
"**Drive I-95,**" From Boston to Florida; ***inside back cover***
1929 Spanish Trail Roadway; 101-102,104
511 Traffic Information; 83
A1A (Barrier Isl) - I-95 Access; 86
AAA (and CAA); 83
AAA National Office; 88
Abbreviations,
  Colored 25 mile Maps; cover
  Exit Services; 196
  Travelogue; 85
Africa; 177
Agricultural Inspection Stns; 126
Ah-Tah-Thi-Ki Museum; 160
Air Conditioning, First; 112
Alabama; 124
Alachua; 132
  County; 131
Alafia River; 143
Alapaha, Name; 126
Alfred B Maclay Gardens; 106
Alligator Alley; 154-155
Alligator Farm, St Augustine; 169
Alligator Hole (definition); 157
Alligator, Buddy; 155
Alligators; 100,135,138,147,156
Anastasia Island; 170
Anhaica; 108-109,146
Apalachicola River; 112
Appleton Mus of Art; 136
Aquifer; 102
Arabian Nights; 94
Art Museum, Ringling; 147
Aruba Beach Cafe; 183
Aucilla River Project; 106
Babcock-Web WMA; 151
Bahia Mar Marina; 184
Baker County; 99
Barefoot Mailmen; 182
Barge Canal; 137
Bee Line Expy; 80
Belz Outlet Mall; 89
Bernard Castro; 136
Big "I"; 165
Big Cypress; 155,158
Big Foot Monster; 105
Butterfly Center, McGuire; 134
CAA (see AAA)
CCC, The; 111,113,115,135,142
Ca d'Zan; 147
Caloosahatchee River; 152
  Name; 150
Canaveral Natnl Seashore; 173
Cannon Creek Airpark; 130
Canopy Road; 106,169
Cape Canaveral; 174
Castillo San Marcos; 169
Cave Diving; 131
Cayo Costa, Name; 150
Celebration; 93
Charlotte County; 149
Charlotte Harbor; 150
Chautauqua; 116
Chipley; 114
  Name; 115
Choctawatchee, Name; 115
Circus Museum, Ringling; 147
Citrus; 88,97,130,136,140,180
CityPlace, W Palm Beach; 180
City Maps,
  Ft Lauderdale Expwys; 194-195
  Jacksonville; 163
  Kissimmee Expwys; 192-193
  Miami Expressways; 194-195
  Orlando Expressways; 192-193
  Pensacola; 26
  Tallahassee; 191
  Tampa-St. Petersburg; 63
  St. Augsutine; 191
Civil War; 100,108,127,138,141
Clearwater Marine Aquarium; 187
Collier County; 154
Collier, Barron; 152
Colonial Spanish Quarters; 168
Columbia County; 101,128
Coquina Building Material; 165
Corkscrew Swamp, Name; 154
Cowboys; 95
Crab Trap II; 144
Cracker, Florida; 88,95,132
Crosstown Expy; 11,35,98,143
Cuban Bread; 184
Dade Battlefield; 140
Dade, Maj. Francis; 139-140,161
Dania Beach Hurricane; 184
Driving Lanes; 85
Duval County; 163
Eau Gallie; 175
Edison, Thomas; 152
Eglin AFB; 116-118
Eight Reale; 176
Ellenton; 144-145
Emanuel Point Wreck; 120
Emergency Callboxes; 83
Epiphytes; 142,148,157,159
Escambia Bay; 119
  Bridge (I-10); 119
  County; 120
Estero; 153
Everglade,90,95,139-140,154-160
  Draining of; 156,181
  Wildlife MA; 160
  Wonder Gardens; 154
Falling Waters SP; 115
Fantasy of Flight; 95
Fayer Dykes SP; 171
Fires, Forest; 166
Fires, Prescribed ; 148
Fisherman's Village; 151
Flagler County; 171
Flagler, Henry; 97,165,167,171
Florida Aquarium; 186
Florida,
  12,000 years ago; 187
  Cavern SP; 114
  Map of all Expressways; 2-3
  Mus of Natural History; 134
  National Cemetery ; 141
  Part of Africa; 177
  Platform; 187
  Sheriff's Boys Camp; 126
  Sports Hall of Fame; 130
  Sun 'n Fun Museum; 97
  Supreme Court; 107
Florida's Turnpike (FTP), 178,189
  25 mile Strip Maps; 66
  Administration; 189
  Coin System; 190
  Exit Services; 189
  HEFT; 76,161,190
  History; 189
  Names; 189
  Service Plazas; 190
  Spur SR91; 76

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 ***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

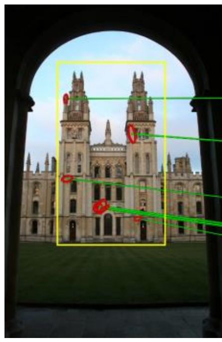
Slide credit: Ondrej Chum

Query Expansion

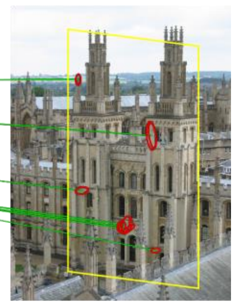
Chum, Philbin, Sivic, Isard, Zisserman: Total Recall..., ICCV 2007

Slide credit: Ondrej Chum

Query Expansion Step by Step



Query Image

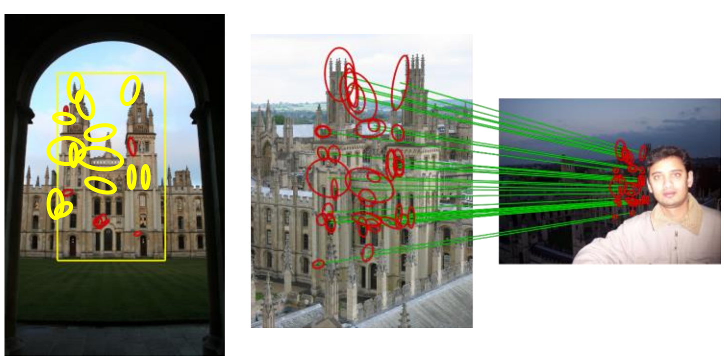


Retrieved image

Originally not retrieved

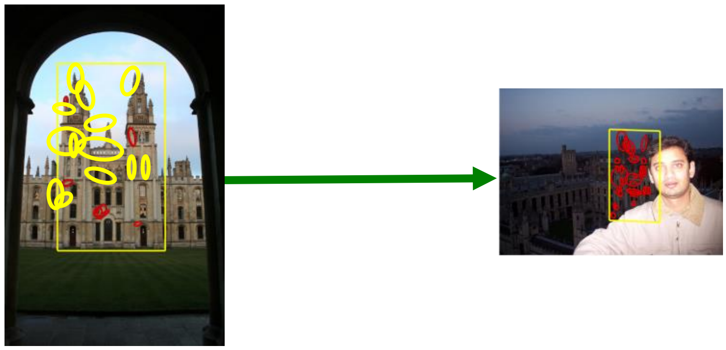
Slide credit: Ondrej Chum

Query Expansion Step by Step



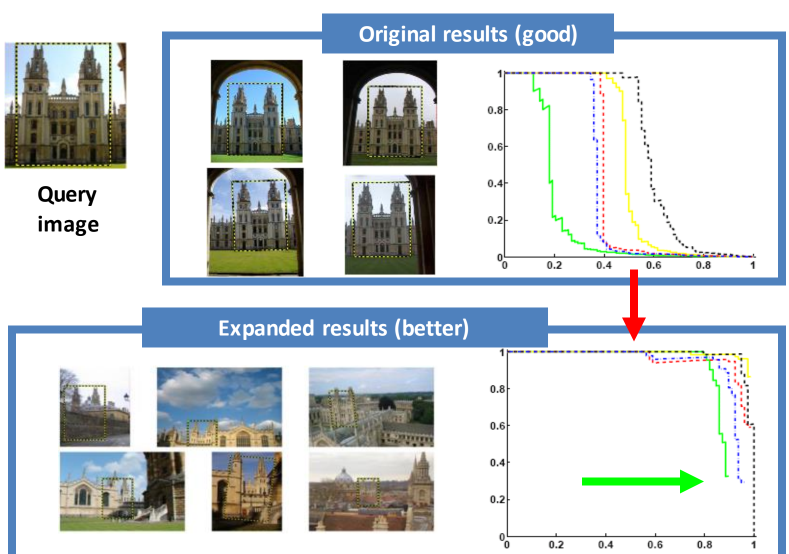
Slide credit: Ondrej Chum

Query Expansion Step by Step



Slide credit: Ondrej Chum

Query Expansion Results



Slide credit: Ondrej Chum

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