

Lecture 14: Indexing with local features

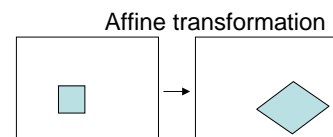
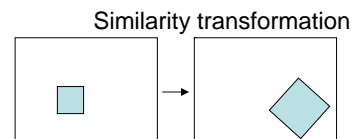
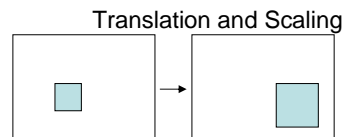
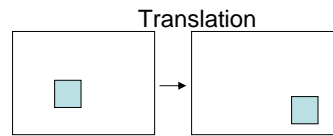
Thursday, Nov 1
Prof. Kristen Grauman

Outline

- Last time: local invariant features, scale invariant detection
- Applications, including stereo
- Indexing with invariant features
- Bag-of-words representation for images

Classes of transformations

- **Euclidean/rigid:**
Translation + rotation
 - Lengths and angles preserved
- **Similarity:** Translation + rotation + uniform scale
- **Affine:** Similarity + shear
 - Valid for orthographic camera, locally planar object
 - Lengths and angles **not** preserved

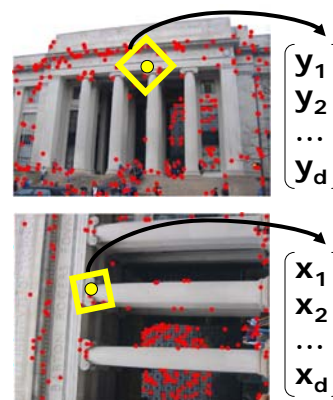


Invariant local features

Subset of local feature types designed to be *invariant* to

- Scale
- Translation
- Rotation
- Affine transformations
- Illumination

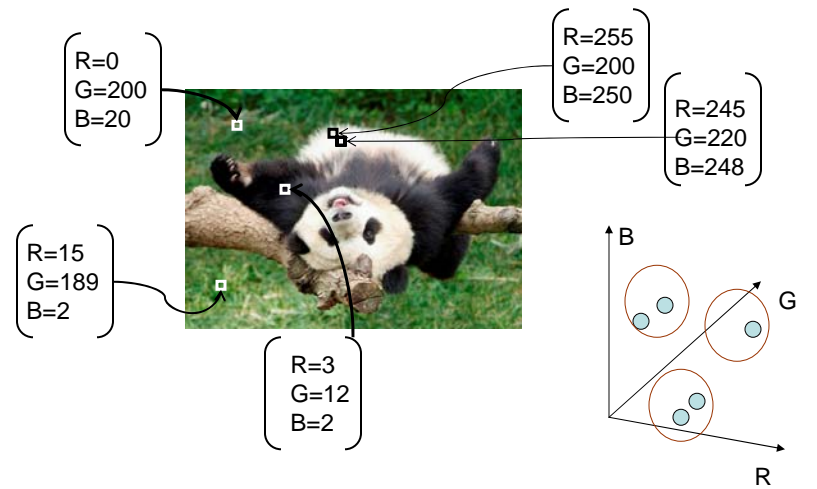
- 1) Detect distinctive interest points
- 2) Extract invariant descriptors



[Mikolajczyk & Schmid, Matas et al., Tuytelaars & Van Gool, Lowe, Kadir et al.,...]]

Recall: segmentation as clustering

- Previously we represented *pixels* with features, mapping each one to a d -dimensional vector



Recall: segmentation as clustering

- Previously we represented *pixels* with features, mapping each one to a d -dimensional vector

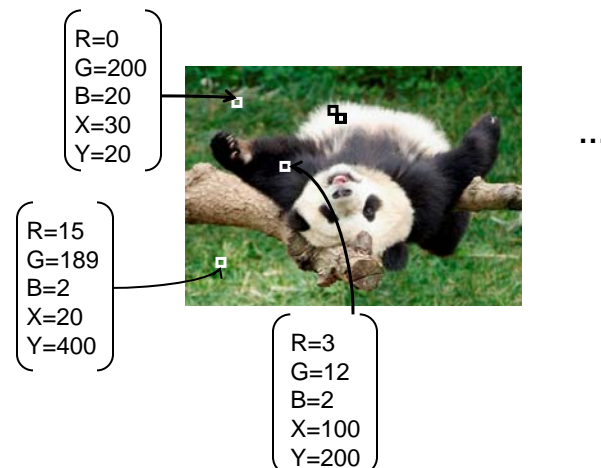


Image patches as vectors

Left Right

“Unwrap” image to form vector, using raster scan order

Each window is a vector in an m^2 dimensional vector space. Normalization makes them unit length.

Slide by Trevor Darrell, MIT

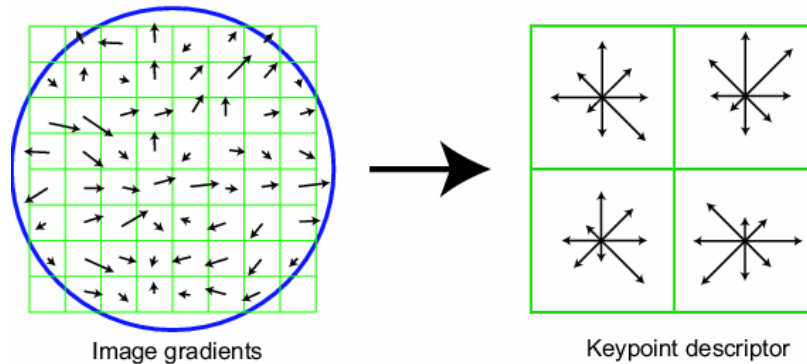
Image metrics

Can compare those vector descriptions

- SSD
- Dot product
- ...

SIFT descriptors: vector formation

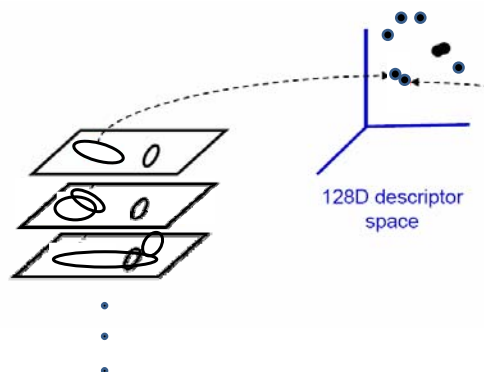
- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create **array of orientation histograms**
- 8 orientations x 4x4 histogram array = 128 dimensions



David Lowe, UBC

Indexing with local features

- Now we have patches or regions, still mapping each one to a d -dimensional vector (e.g., $d=128$ for SIFT)



Indexing with local features

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

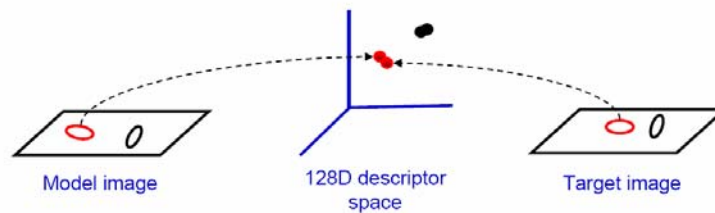


Figure from Andrew Zisserman, University of Oxford

What are the limitations of describing image patches with a stack of pixel intensities?

Why should something like a SIFT descriptor be more robust?

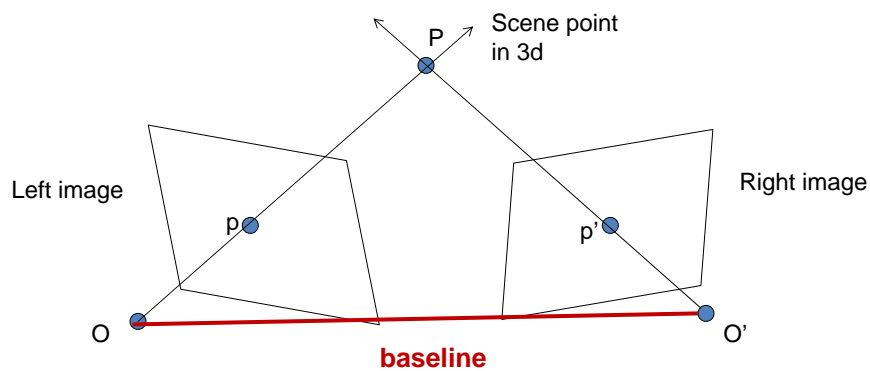
What role does the interest point detection play?



Many applications of local features

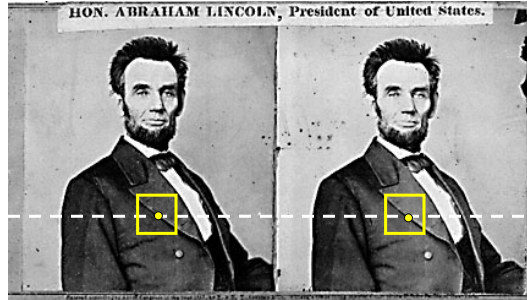
- Wide baseline stereo
- Motion tracking
- Panoramas
- Mobile robot navigation
- 3D reconstruction
- Recognition
 - Specific objects
 - Textures
 - Categories
- ...

Recall: Triangulation



Estimate scene point based on camera relationships and correspondence.

Dense correspondence search



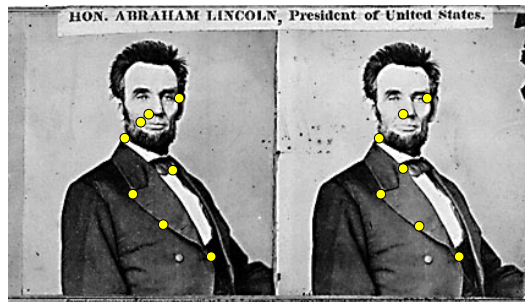
For each epipolar line

For each pixel / window in the left image

- compare with every pixel / window on same epipolar line in right image
- pick position with minimum match cost (e.g., SSD, correlation)

Adapted from Li Zhang

Sparse correspondence search



- Restrict search to sparse set of detected features
- Rather than pixel values (or lists of pixel values) use *feature descriptor* and an associated *feature distance*
- Still narrow search further by epipolar geometry

Wide baseline stereo

- 3d reconstruction depends on finding good correspondences
- Especially with wide-baseline views, local image deformations not well-approximated with rigid transformations
- Cannot simply compare regions of fixed shape (circles, rectangles) – shape is not preserved under affine transformations

Wide baseline stereo

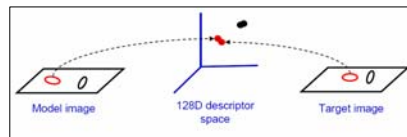


Figure 1: BOOKSHELF: Estimated epipolar geometry on indoor scene with significant scale change. In the cutouts the change in the resolution of detected DRs is clearly visible.

Wide baseline stereo



Figure 2: VALBONNE: Estimated epipolar geometry and points associated to the matched regions are shown in the first row. Cutouts in the second row show matched bricks.

J. Matas, O. Chum, M. Urban, T. Pajdla. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions, BMVC 2002.

Wide baseline stereo

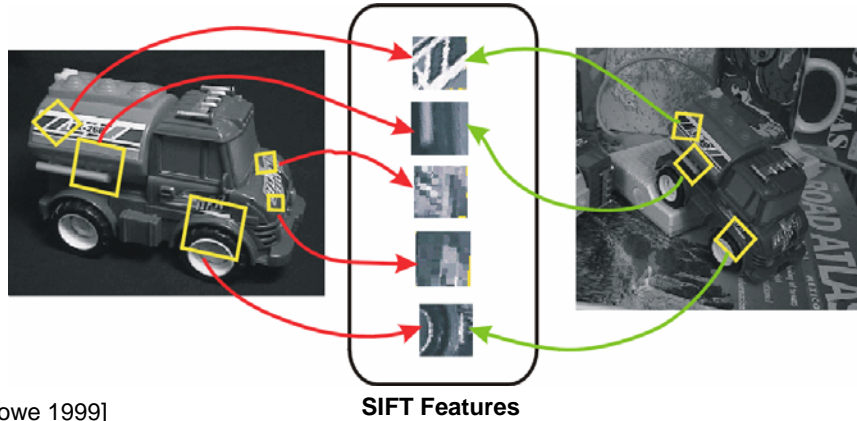


Figure 3: WASH: Epipolar geometry and dense matched regions with fully affine distortion.

J. Matas, O. Chum, M. Urban, T. Pajdla. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions, BMVC 2002.

SIFT matching and recognition

- Index descriptors
- Generalized Hough transform: vote for object poses
- Refine with geometric verification: affine fit, check for agreement between image features and model



Recognition of specific objects, scenes



Schmid and Mohr 1997



Sivic and Zisserman, 2003



Rothganger et al. 2003



Lowe 2002

Panorama stitching



(a) Matier data set (7 images)



(b) Matier final stitch

Brown, Szeliski, and Winder, 2005

Value of local (invariant) features

- Complexity reduction via selection of distinctive points
- Describe images, objects, parts without requiring segmentation
- Local character means robustness to clutter, occlusion
- Robustness: similar descriptors in spite of noise, blur, etc.

Comparative evaluations

Testing various detector and descriptor options for relative *repeatability* and *distinctiveness*



Planar objects / flat scenes:
Mikolajczyk & Schmid (2004)



3D objects:
Moreels & Perona (2005)

[Images from Lazechnik, Sicily 2006]

Affine Covariant Features

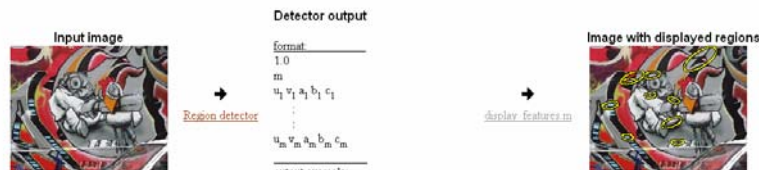


KATHOLIEKE UNIVERSITEIT
LEUVEN



Collaborative work between the Visual Geometry Group, Katholieke Universiteit Leuven, INRIA Sophia-Antipolis and the Center for Machine Perception.

Affine Covariant Region Detectors



Parameters defining an affine region

u, v, a, b, c in $a(x-u) + b(y-v) + c = 1$ with $(0,0)$ at image top left corner

Code

- provided by the authors, see [publications](#) for details and links to authors web sites.

Linux binaries

[Harris-Affine](#) & [Hessian-Affine](#)

[MSER](#) - Maximally stable extremal regions (also Windows)

[IBR](#) - Intensity extrema based detector

[EBR](#) - Edge based detector

[Salient](#) region detector

Example of use

```
prompt> ./h_affine.in -haraaff -1 img1.ppm -o img1.haraaff -thres 1000 matlab> ;
prompt> ./h_affine.in -hesaaff -1 img1.ppm -o img1.hesaaff -thres 500 matlab> ;
prompt> ./msr.in -t 2 -es 2 -1 img1.ppm -o img1.msrr matlab> ;
prompt> ./ibr.in img1.ppm img1.ibr -scalefactor 1.0 matlab> ;
prompt> ./ebr.in img1.ppm img1.ebr matlab> ;
prompt> ./saient.in img1.ppm img1.sai matlab> ;
```

Displaying

<http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html#binaries>

Outline

- Last time: local invariant features, scale invariant detection
- Applications, including stereo
- Indexing with invariant features
- Bag-of-words representation for images

Success of text retrieval



- efficient
- scalable
- high precision

Can we use retrieval mechanisms from text retrieval?

Need a visual analogy of a textual word.

Visual problem

- Retrieve key frames containing the same **object**



Slide from Andrew Zisserman

Problem specification: particular object retrieval

Example: visual search in feature films

Visually defined query

"Find this clock"



"Find this place"



"Groundhog Day" [Rammis, 1993]



Slide from Andrew Zisserman

Example



Slide from Andrew Zisserman

retrieved shots



Text retrieval vs. image search

- What makes the problems similar, different?

Object

Bag of 'words'



ICCV 2005 short course, L. Fei-Fei

Analogy to documents

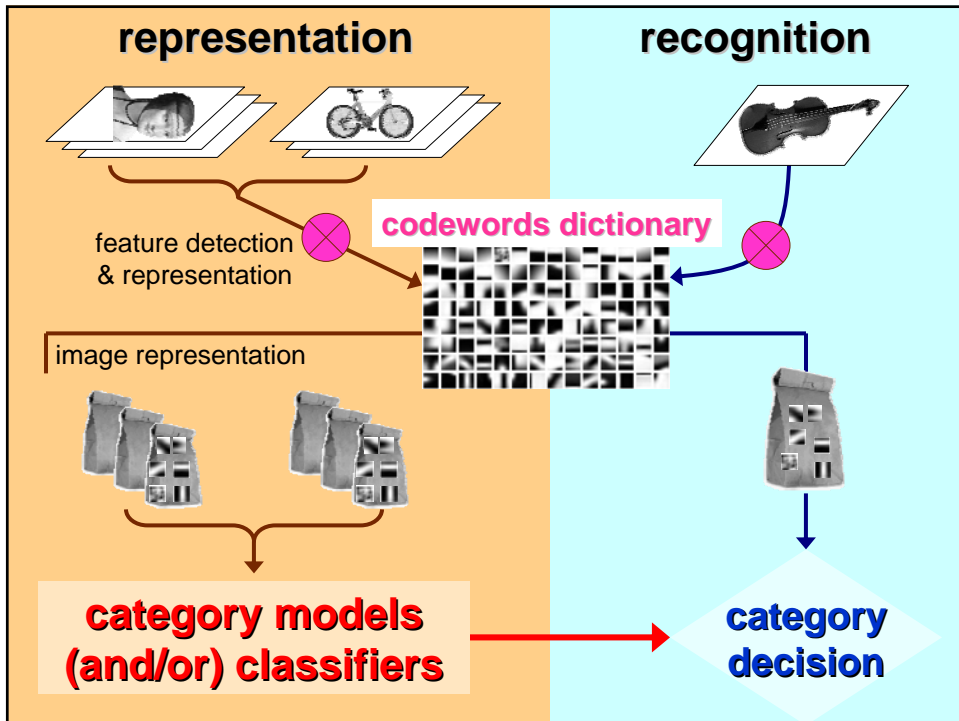
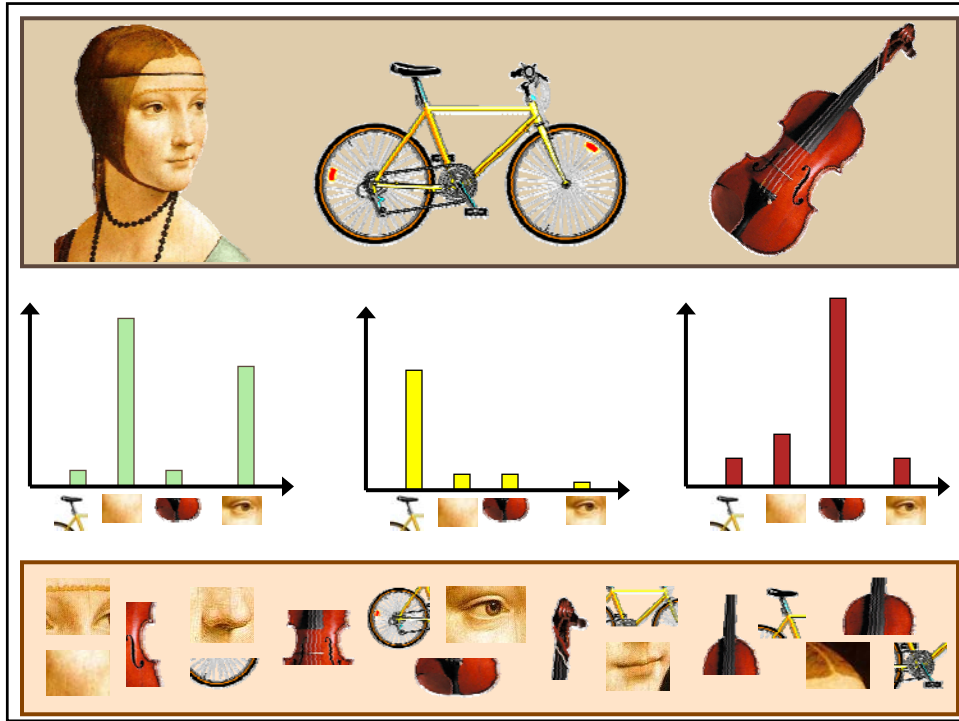
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach our eyes. For a long time, the visual image was considered as a simple picture. However, the discovery of the visual centers in the brain has shown that the image is processed in a more complex way. Hubel and Wiesel have discovered that the visual cortex is organized in columns, each with its specific function and is responsible for a specific detail in the pattern of the retinal image.

**sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel**

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% increase in exports to \$750bn, compared with \$575bn in 2004. The surplus of \$660bn. The government also needs to reduce the demand for foreign currency. China's government has also announced that it will permit it to trade within a narrow band, but the US wants the yuan to be allowed to rise 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.

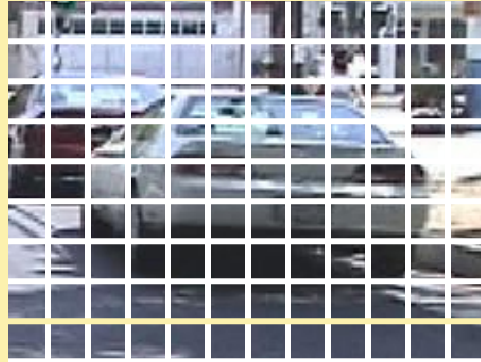
**China, trade,
surplus, commerce,
exports, imports, US,
yuan, bank, domestic,
foreign, increase,
trade, value**

ICCV 2005 short course, L. Fei-Fei



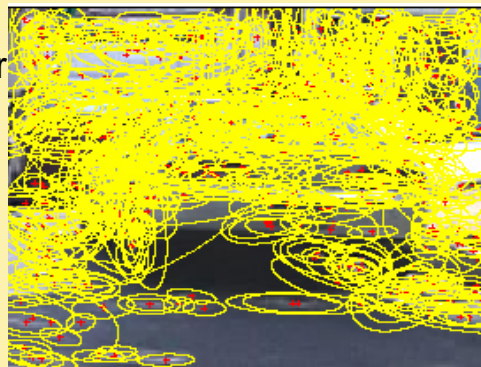
1.Feature detection and representation

- Regular grid



1.Feature detection and representation

- Regular grid
- Interest point detector



1.Feature detection and representation

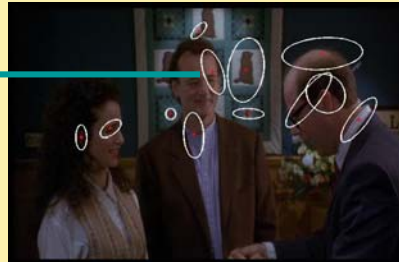
- Regular grid
- Interest point detector
- Other methods
 - Random sampling
 - Segmentation based patches

1.Feature detection and representation


**Compute
SIFT
descriptor**
[Lowe'99]



**Normalize
patch**



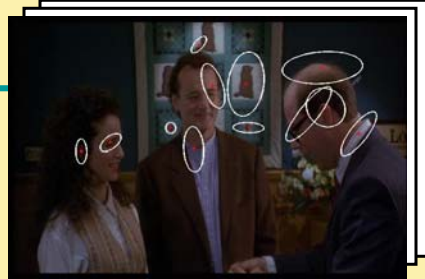
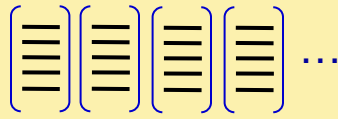
Detect patches

[Mikojaczyk and Schmid '02]

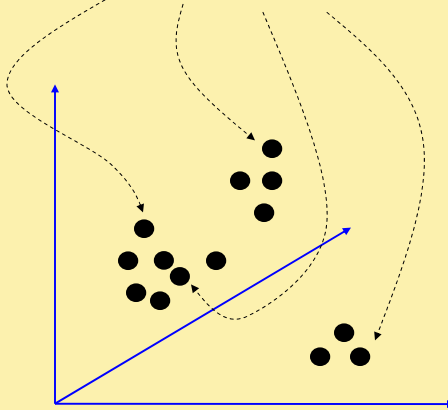
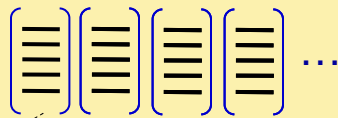
[Matas et al. '02]

[Sivic et al. '03]

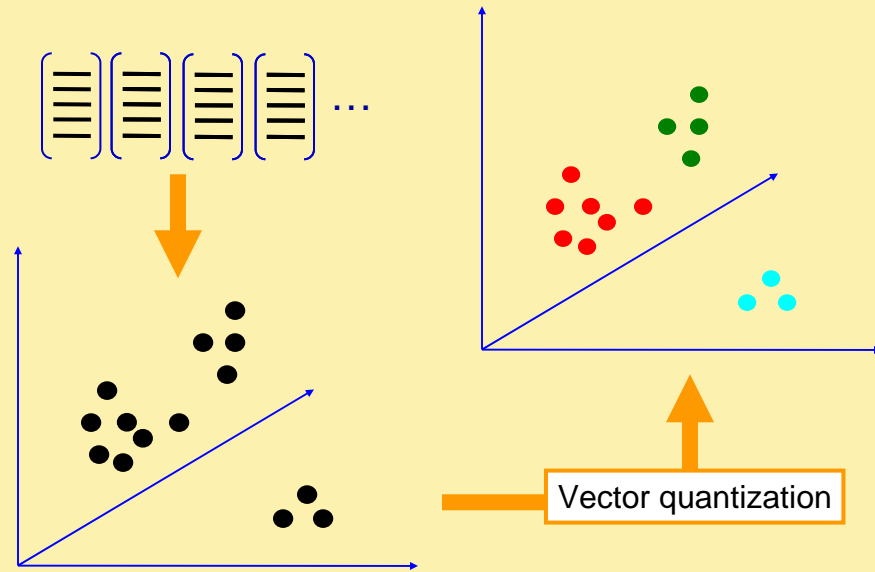
1. Feature detection and representation



2. Codewords dictionary formation

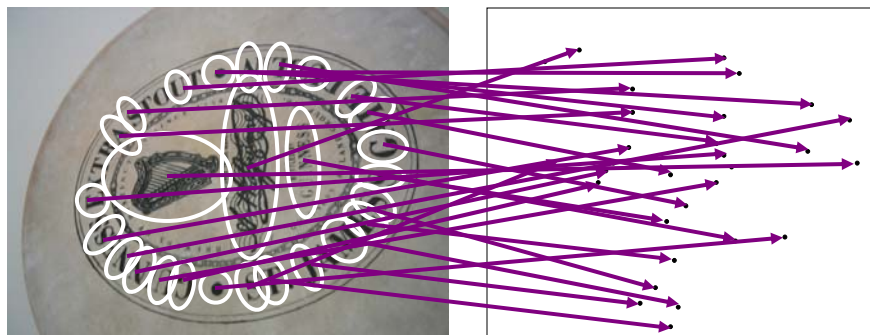


2. Codewords dictionary formation



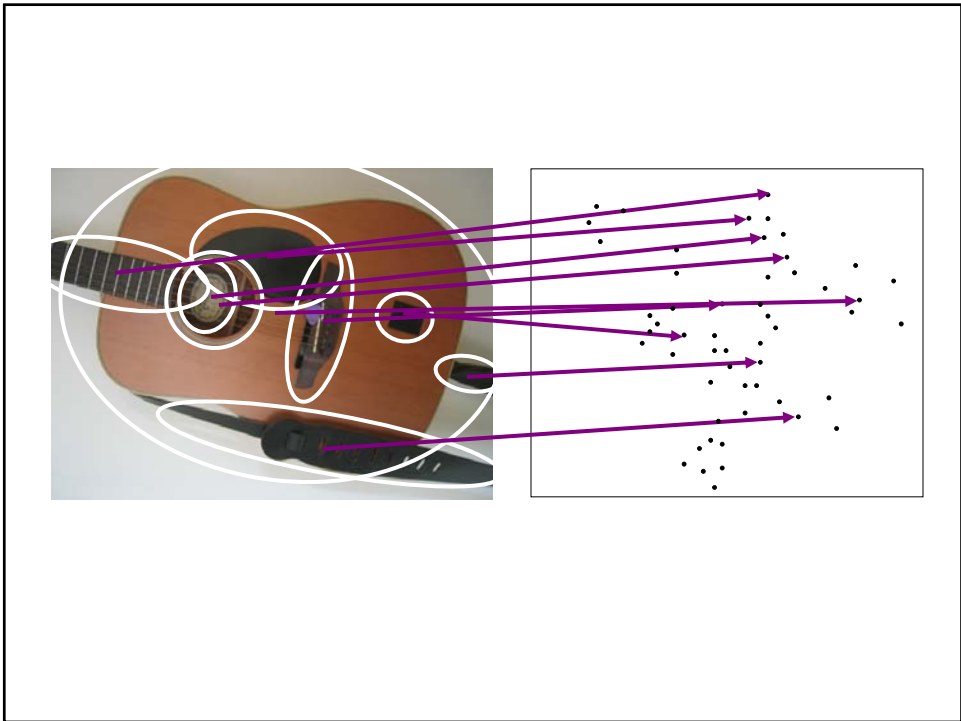
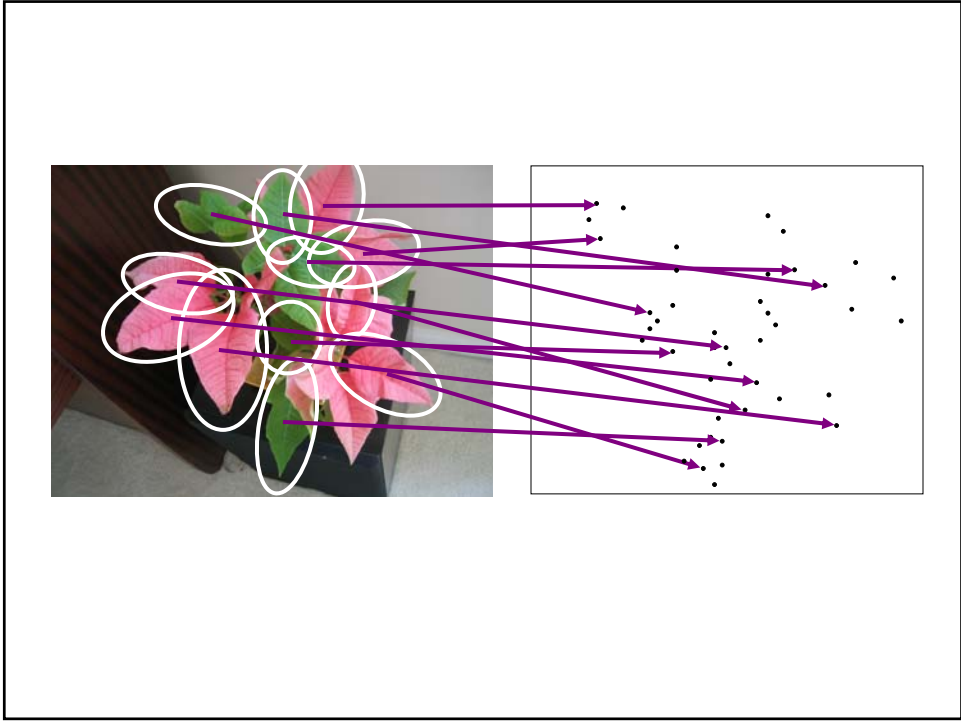
Slide credit: Josef Sivic

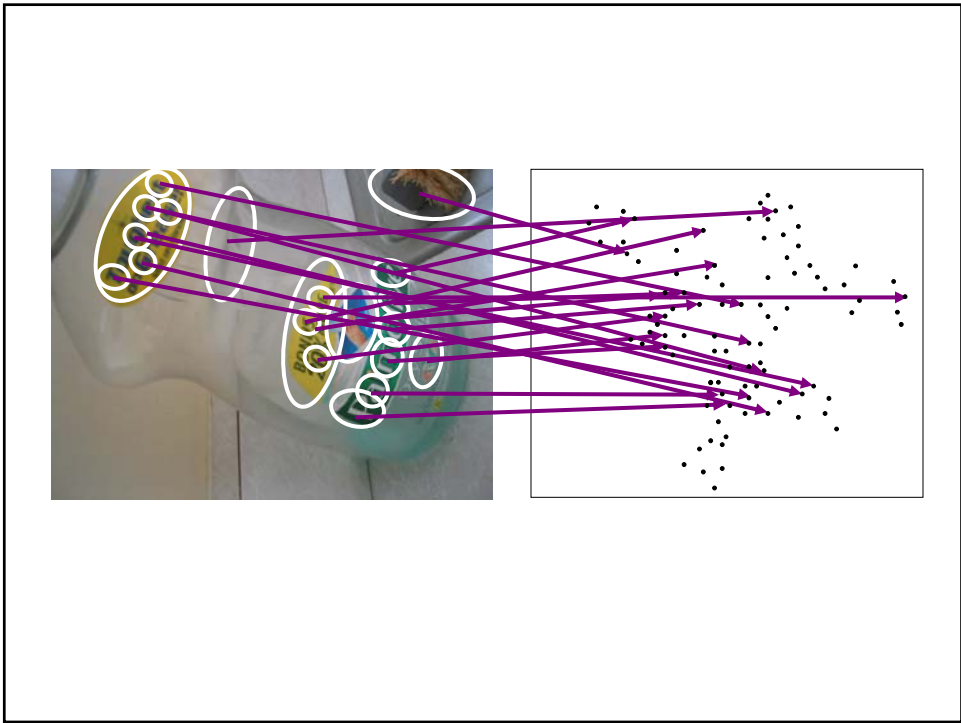
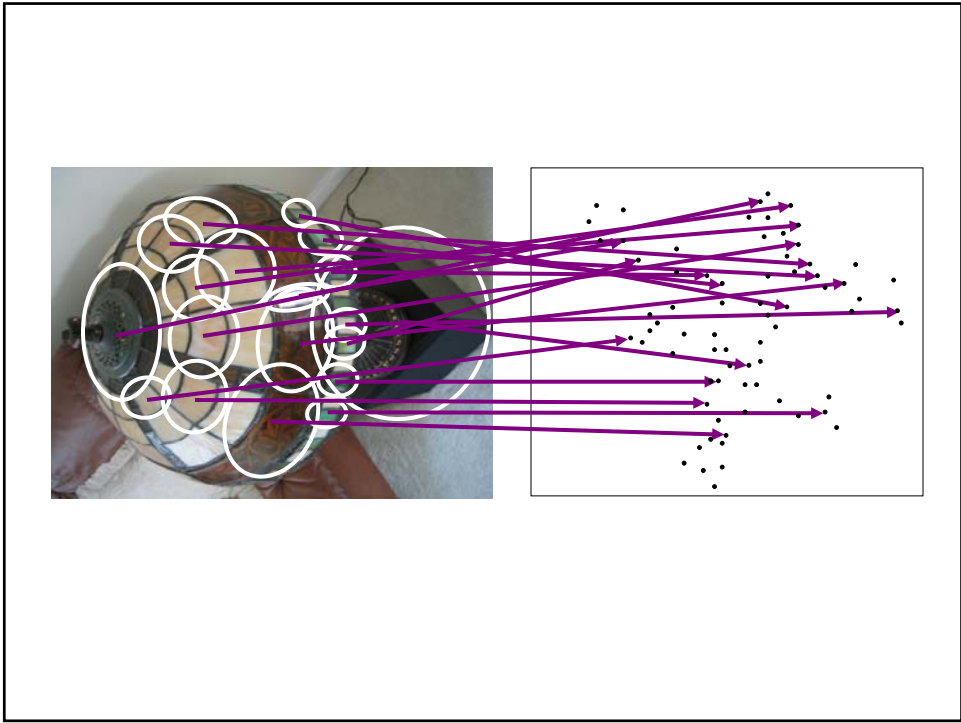
Extract some local features from a number of images ...



SIFT descriptor space: each point is 128-dimensional

Slides from D. Nister





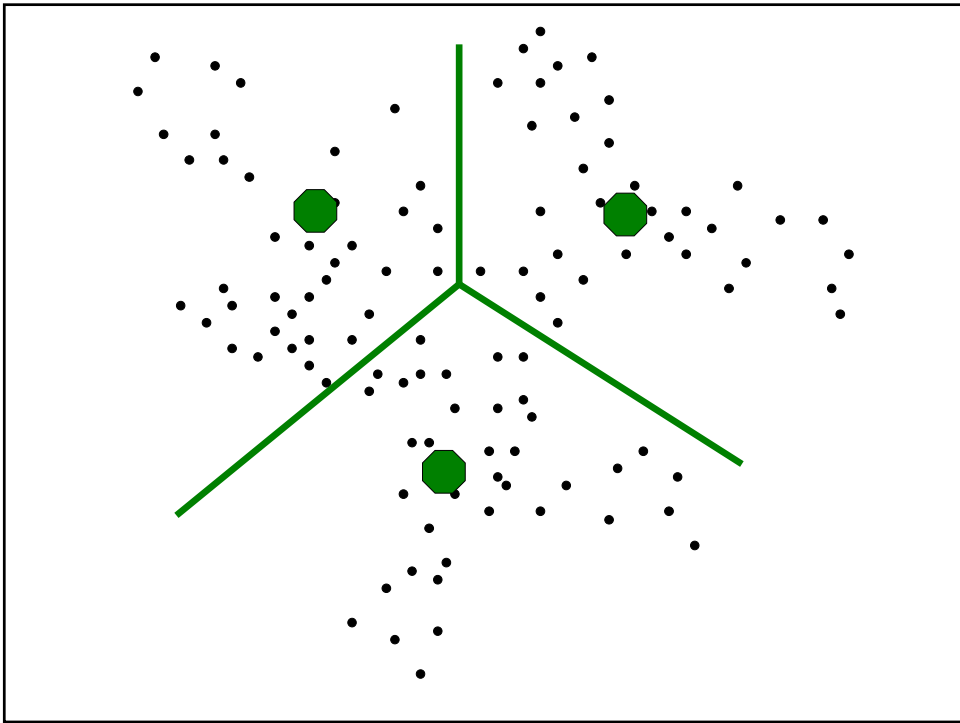
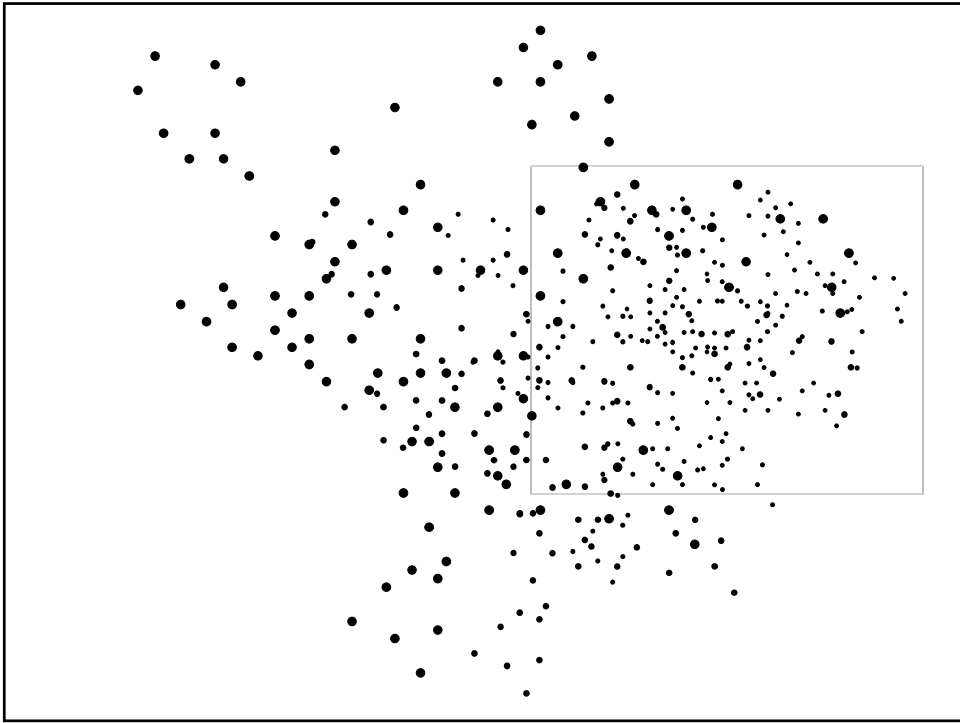
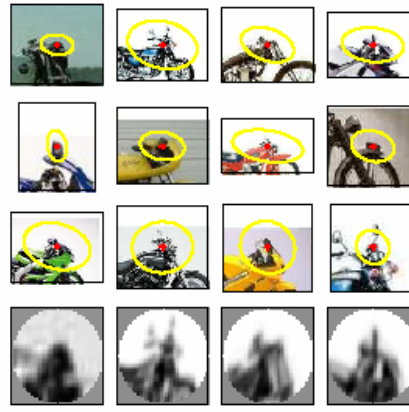
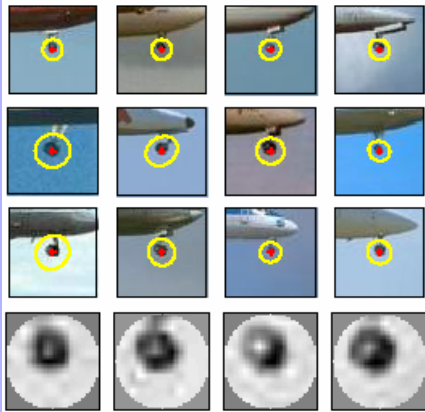
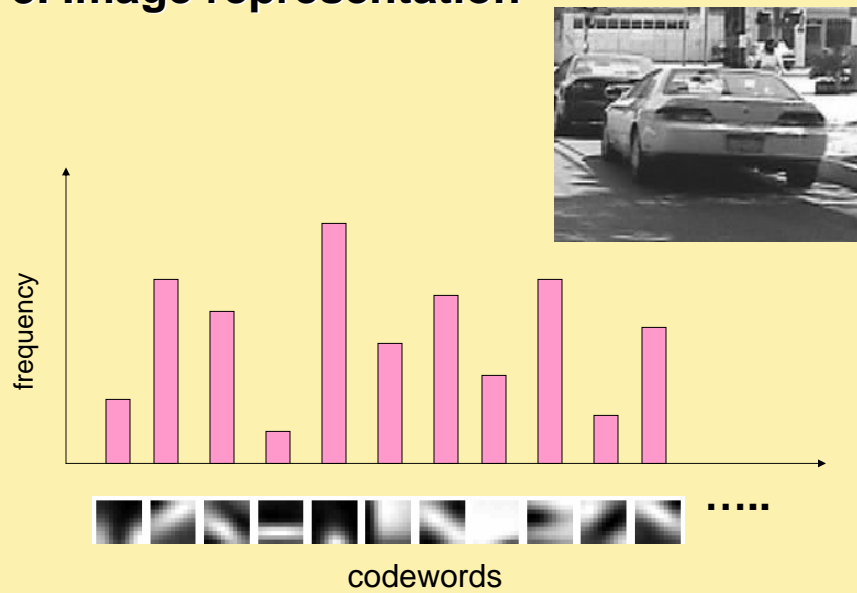


Image patch examples of codewords



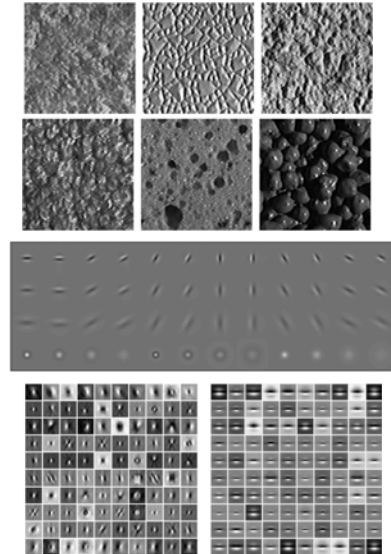
Sivic et al. 2005

3. Image representation



Visual words = textons

- *Texton* = cluster center of filter responses over collection of images [Leung and Malik, 1999]
- Represent texture or material with histogram of texton occurrences (or prototypes of whatever feature type employed)



Visual words and bags of words

- Have a way to represent
 - Individual local image regions as “tokens” / discrete set of visual words
 - Entire image in terms of its distribution of words
- How to use this for indexing task?
- Again, can look to text retrieval for inspiration

Inverted file index

- For each word, store list of documents (pages) where that word occurs

Index		
*Along 475," From Detroit to Florida; inside back cover	Butterfly Center, McGuire; 134	Driving Lanes; 85
"Drive I-95," From Boston to Florida; inside back cover	CAA (see AAA)	Duval County; 163
1929 Spanish Trail Roadway, 101-102,104	CCC, Tho; 111,113,115,135,142	Eau Gallie; 175
511 Traffic Information; 83	Ca d'Zan; 147	Edison, Thomas; 162
A1A (Barrier Is) - I-95 Access; 86	Caloosahatchee River; 152	Eight AF8; 118,119
AAA (and CAA); 83	Name; 150	Eight Reale; 178
AAA National Office; 88	Carriacoual Natal Seashore; 173	Elberton; 144-145
Abbreviations,	Cannon Creek Airpark; 130	Emanuel Point Wreck; 120
Coked 25 mile Maps; cover	Canopy Road; 106,109	Emergency Callboxes; 83
Exit Services; 196	Cape Canaveral; 114	Epiphany; 142,148,157,159
Travelogue; 85	Castillo San Marcos; 169	Escambia Bay; 119
Africa; 177	Cave Diving; 131	Bridge (I-10); 119
Agricultural Inspection Strip; 126	Cayo Costa, Name; 150	County; 120
Ar-Tan-Tha-Ki Museum; 180	Celebration; 99	Estero; 163
Air Conditioning, Fleet; 112	Charlotte County; 149	Everglades; 80,95,139-140,154-160
Alabama; 124	Charlotte Harbor; 150	Draining of; 156,181
Alachua; 132	Chautauque; 116	Wildlife MA; 160
County; 131	Chipley; 114	Wonder Gardens; 164
Atalla River; 143	Name; 115	Falling Waters SP; 115
Atropa; Name; 126	Chocowichee, Name; 115	Fantasy of Flight; 86
Alfred B Munday Gardens; 106	Circus Museum, Ringling; 147	Fayer Dykes SP; 171
Alligator Alley; 164-165	Citra; 88,97,100,130,140,180	Fires, Forest; 168
Alligator Farm, St Augustine; 169	CityPlace, W Palm Beach; 180	Fires, Prescribed; 148
Alligator Hole (definition); 157	City Maps,	Fisherman's Village; 151
Alligator, Buddy; 155	FL Lashbarkle Express; 194-195	Flagler County; 171
Alligators; 100,135,138,147,155	Jacksonville; 163	Flagler, Henry; 97,105,107,171
Amistada Island; 170	Kissimmee Express; 192-193	Florida Aquarium; 186
Anhaika; 108-109,146	Miami Expressways; 194-195	Florida,
Apalachicola River; 112	Orlando Expressways; 192-193	12,000 years ago; 187
Applenton Mas of Art; 126	Pensacola; 26	Cavern SP; 114
Aquifer; 102	Tallahassee; 191	Map of all Expressways; 2-3
Arabian Nights; 94	Tampa-St. Petersburg; 63	Mus of Natural History; 134
Art Museum, Ringling; 147	St. Augustine; 191	National Cemetery; 141
Aruba Beach Cade; 183	Civil War; 100,108,127,138,141	Part of Africa; 177
Aucilla River Project; 108	Clematis Marine Aquarium; 187	Platform; 187
Babcock-Web WMA; 151	Collier County; 154	Sherriff's Boys Camp; 126
Bahia Mir Marina; 184	Collier, Barron; 152	Sports Hall of Fame; 130
Baker County; 82	Colonial Spanish Quarters; 168	Sun 'n Fun Museum; 97
	Columbia County; 101,128	Supreme Court; 107
	Coquina Building Material; 165	Florida's Turnpike (FTP); 178,189
	Cookinze, Swamp, Name; 154	25-mile State Maps; 66

Inverted file index for images





When would using an inverted file reduce the amount of images we need to search/compare?

Video Google [Sivic & Zisserman, 2003]

In each frame independently
determine elliptical regions (segmentation covariant with camera viewpoint)
compute SIFT descriptor for each region [Lowe '99]



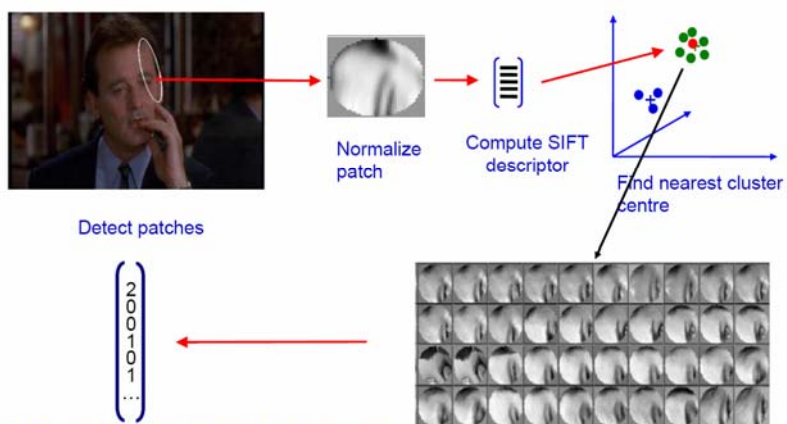
1000+ descriptors per frame

-  Harris-affine
-  Maximally stable regions

Slide from Andrew Zisserman, University of Oxford

Video Google [Sivic & Zisserman, 2003]

Assign visual words and compute histograms for each
key frame in the video



Slide from Andrew Zisserman

Video Google [Sivic & Zisserman, 2003]

- Stage 1: generate a short list of possible frames using bag of visual word representation:
 1. Accumulate all visual words within the query region
 2. Use “book index” to find other frames with these words
 3. Compute similarity for frames which share at least one word



- Generates a tf-idf ranked list of all the frames in dataset

Slide from Andrew Zisserman, University of Oxford

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 → n_{id}

Number of words in document d → n_d

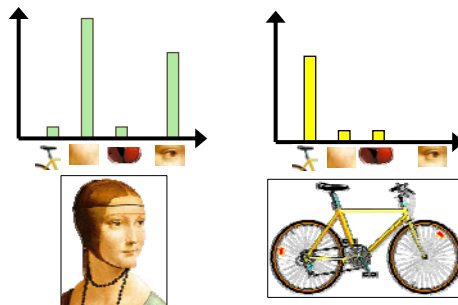
Total number of documents in database → N

Number of occurrences of word i in whole database → n_i

Comparing bags of words

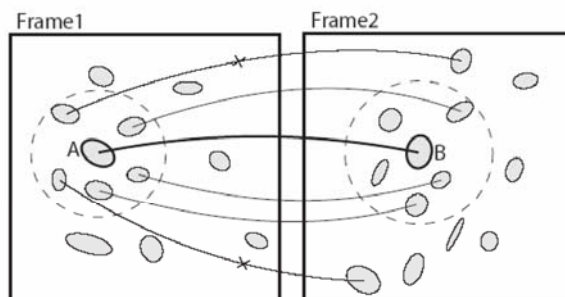
- Rank frames by dot product between their (tf-idf weighted) occurrence counts

$$[1 \ 8 \ 1 \ 4]' \cdot [5 \ 1 \ 1 \ 0]$$



Video Google [Sivic & Zisserman, 2003]

Stage 2: re-rank short list on spatial consistency



NB weak measure
of spatial
consistency

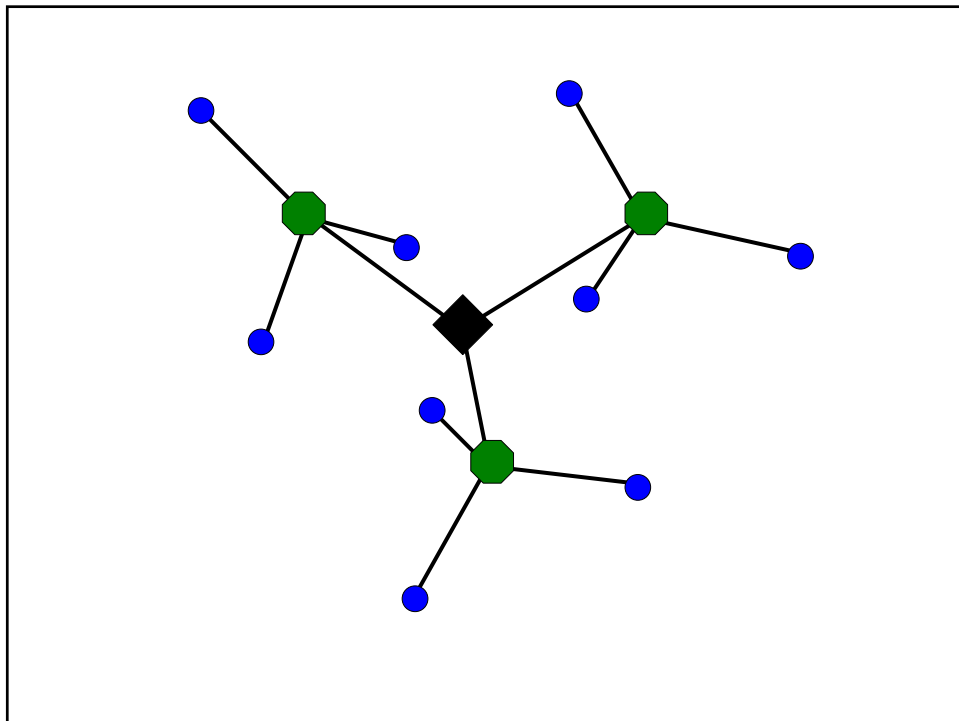
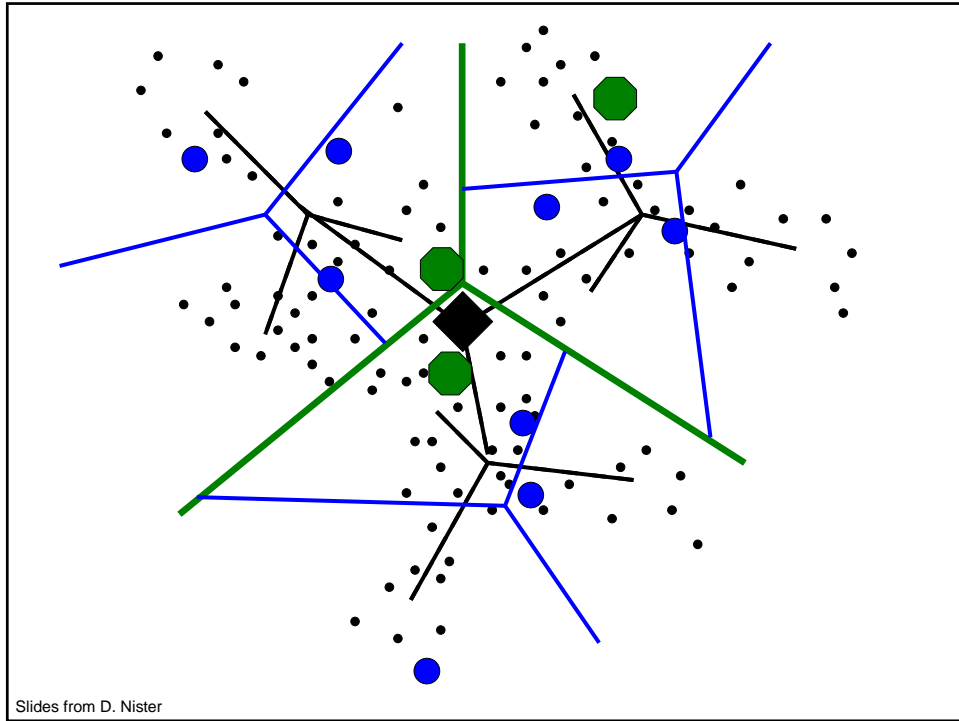
- Discard mismatches
 - require spatial agreement with the neighbouring matches
- Compute matching score
 - score each match with the number of agreement matches
 - accumulate the score from all matches

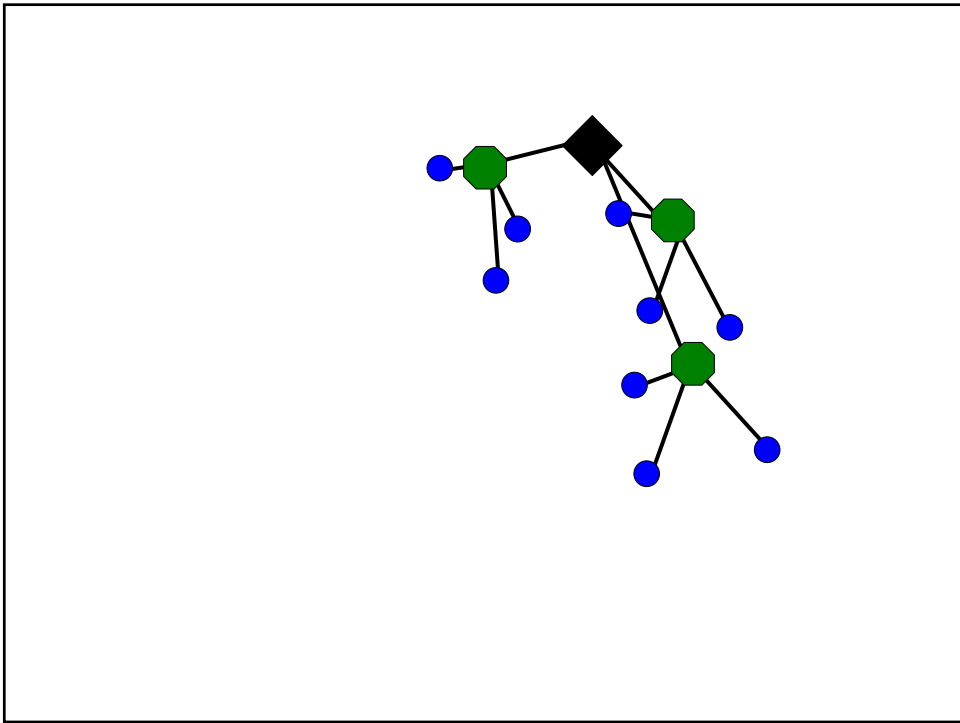
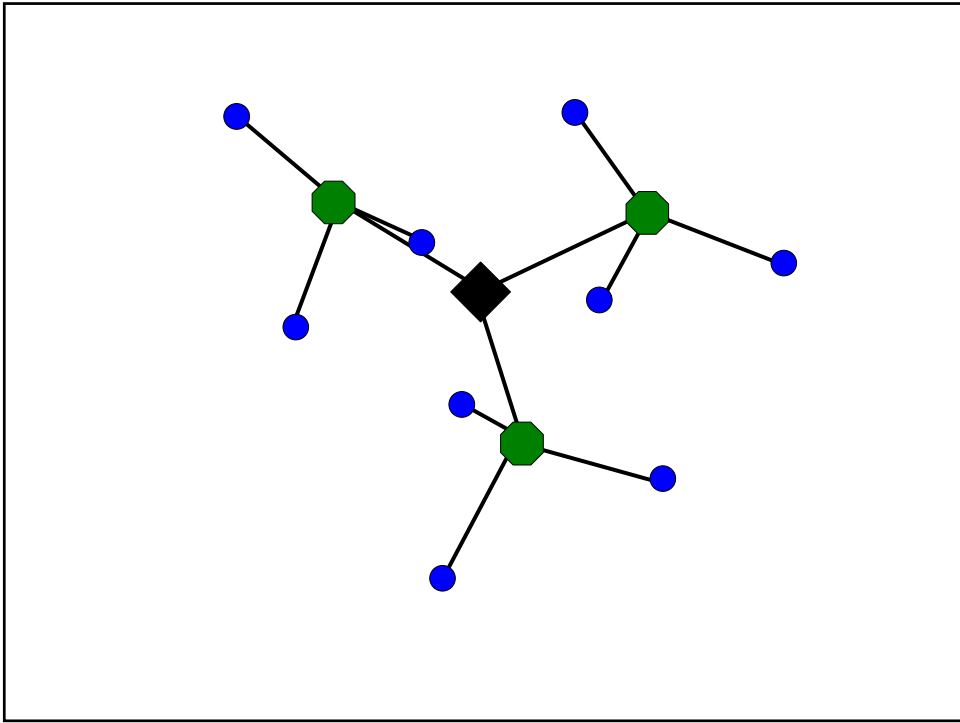
Video Google demo

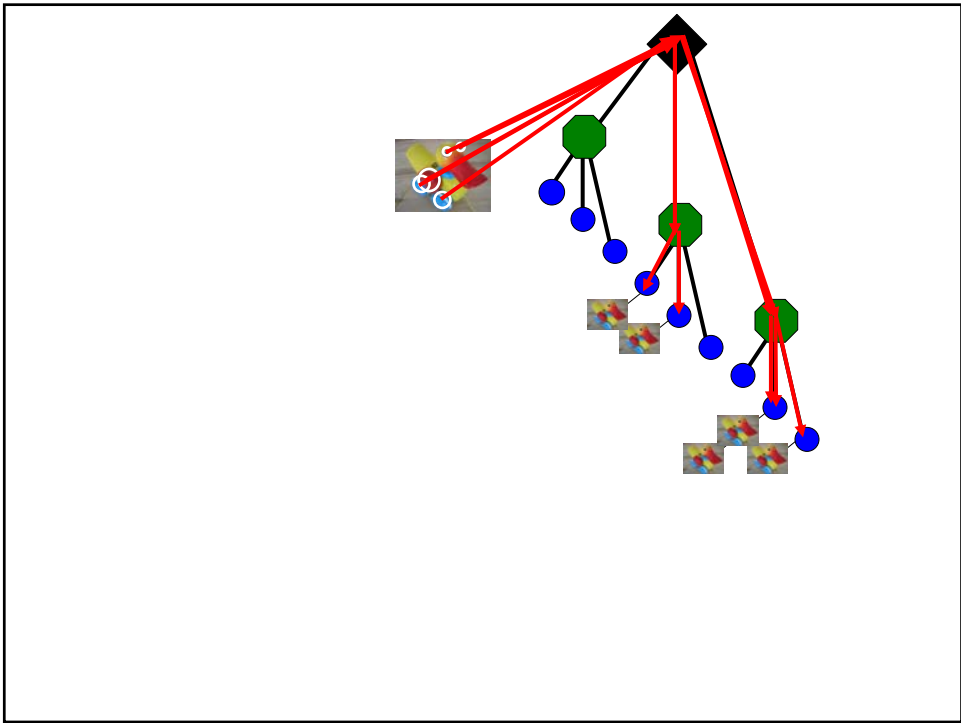
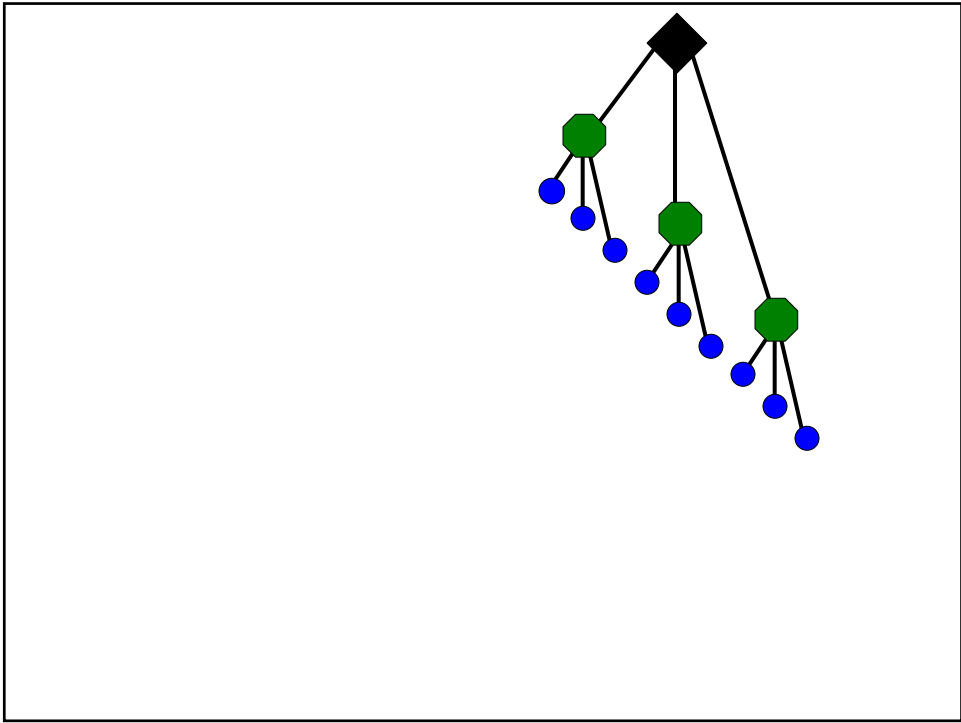
<http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html>

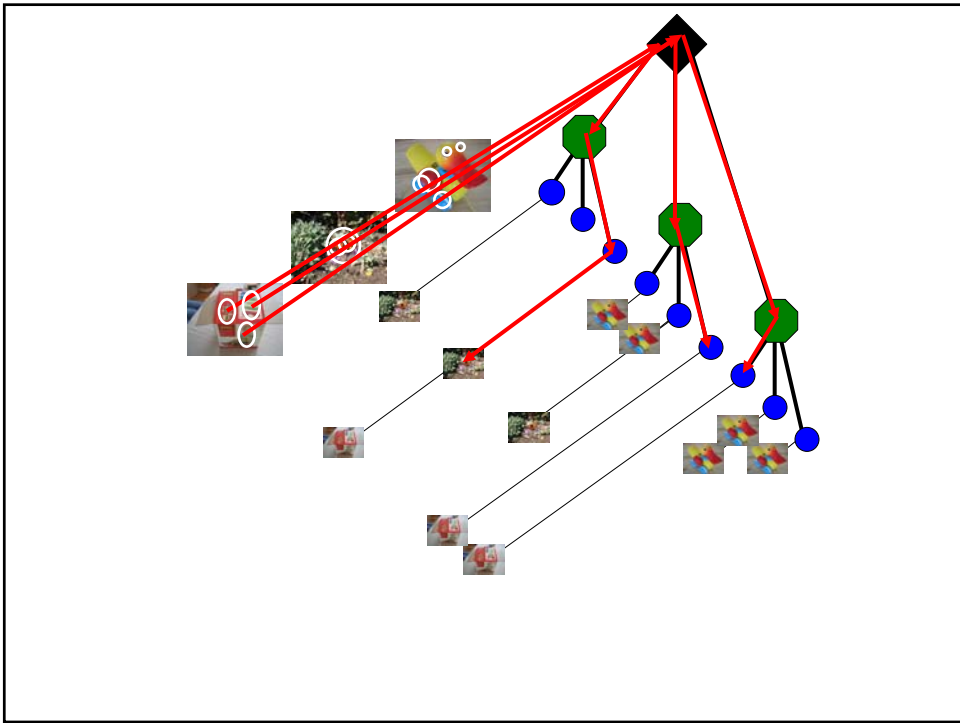
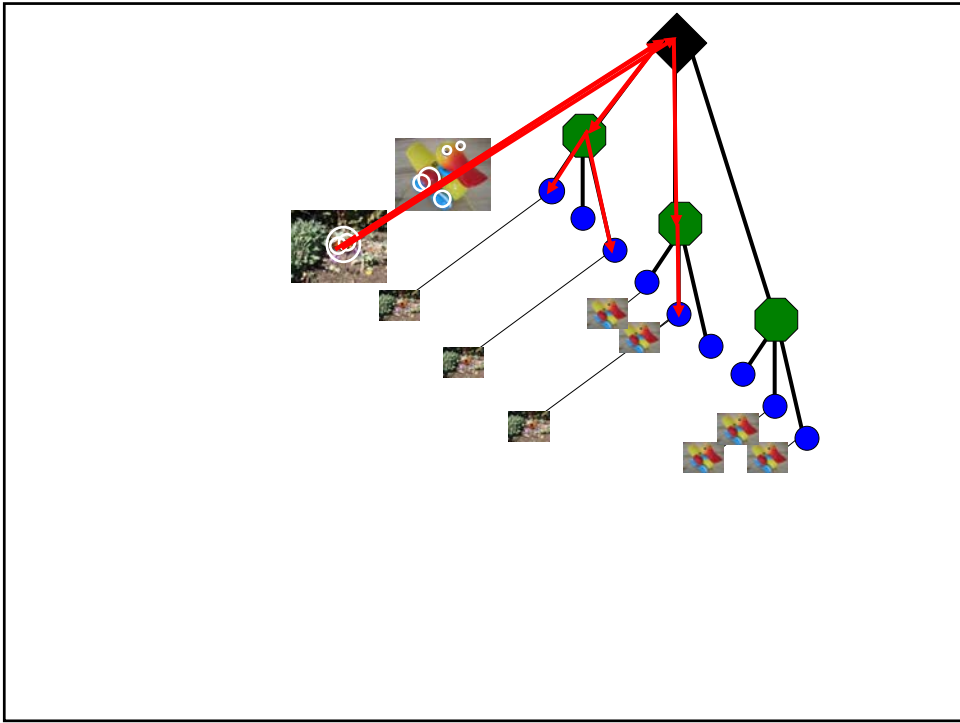
Hierarchical vocabulary

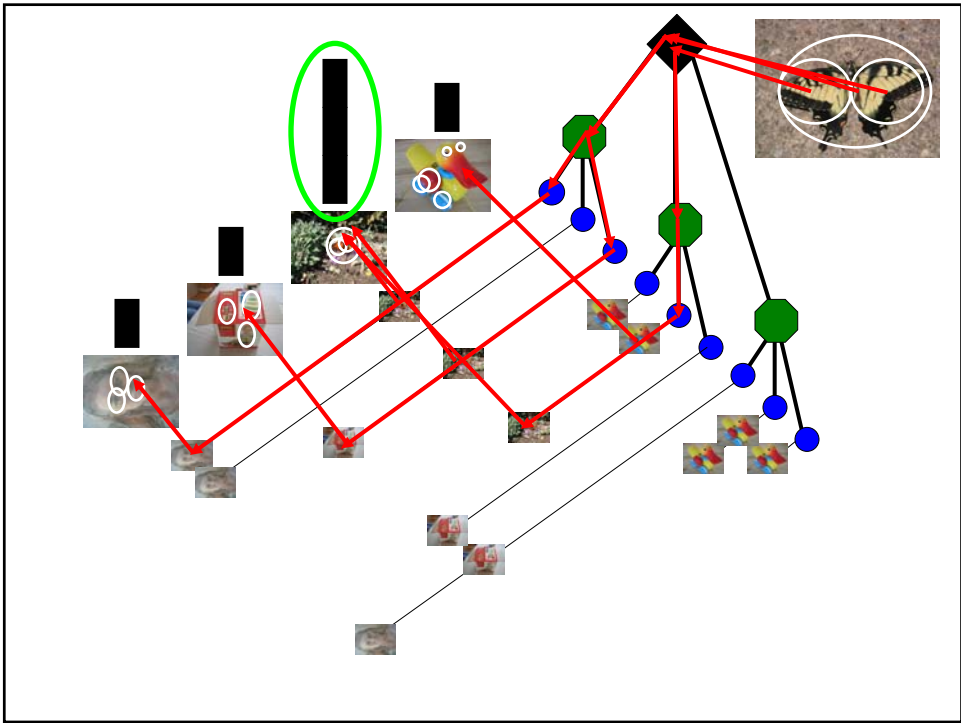
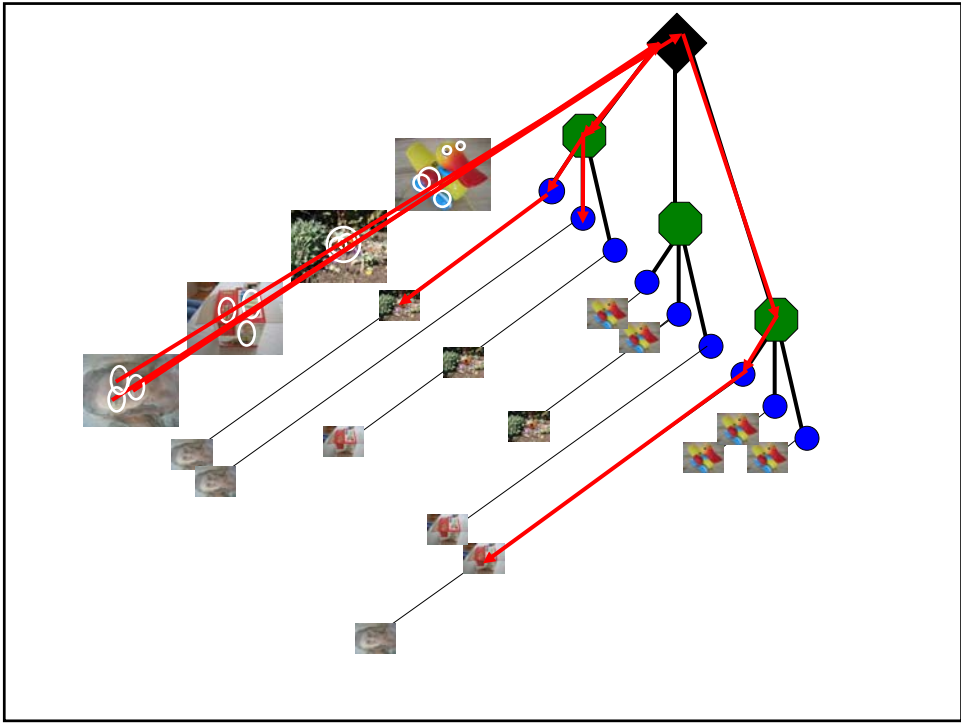
- To manage a large vocabulary efficiently, we can form the quantization of feature space in a hierarchical way
- David Nister & Henrik Stewenius, Scalable Recognition with a Vocabulary Tree, CVPR 2006



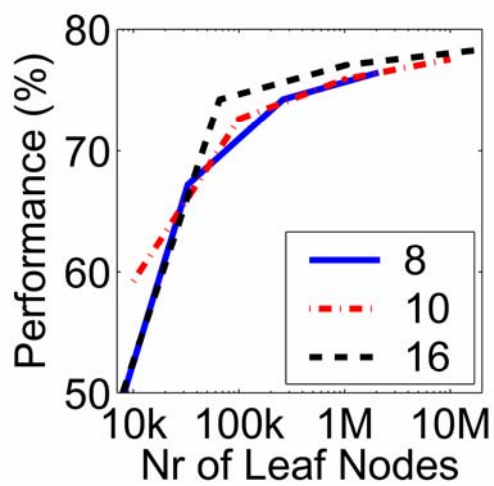








What is the computational advantage of the hierarchical representation bag of words, vs. a flat vocabulary?



Larger vocabularies can be advantageous...

But what happens if it is too large?

Bag of words representation: advantages

- Flexibility comes with ignoring geometry (?)
- Compact description, yet rich
- Local features → vector
 - Usable representation
 - Relatively efficient learning
- Yields good results in practice

Bag of words representation: Issues

- Flexibility comes with ignoring geometry (!)
- Background/foreground treated at once
- Vocabulary formation
 - Number of words/clusters?
 - Universal, or dataset specific?
 - May be expensive
- How to localize/segment object?

Making the Sky Searchable: Fast Geometric Hashing for Automated Astrometry

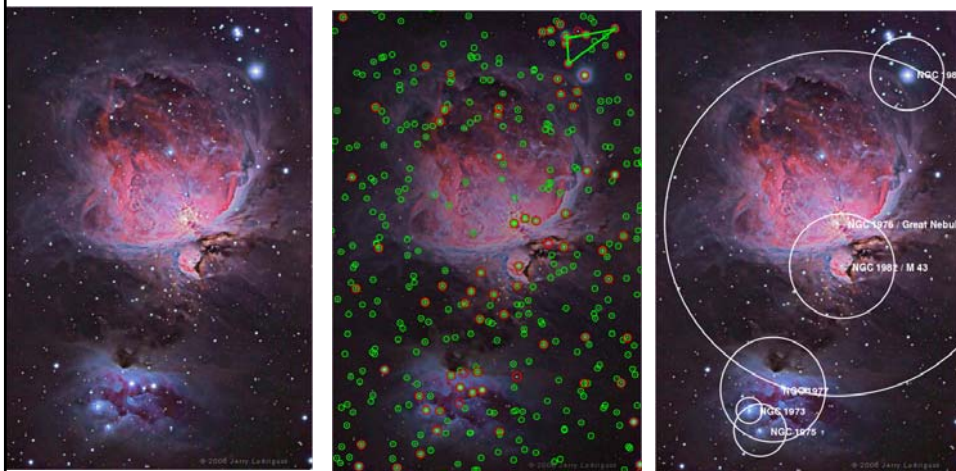
Sam Roweis, Dustin Lang & Keir Mierle
University of Toronto

David Hogg & Michael Blanton
New York University

Check out the slides at:
cosmo.nyu.edu/hogg/research/2006/09/28/astrometry_google.ppt

Example

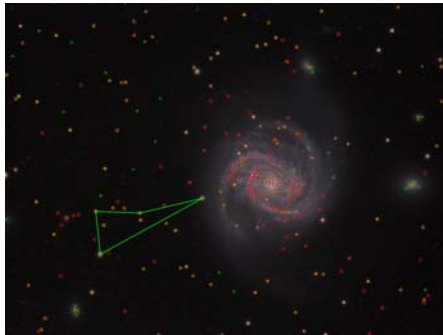
Roweis, Lang, Mierle, Hogg & Blanton



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

Example

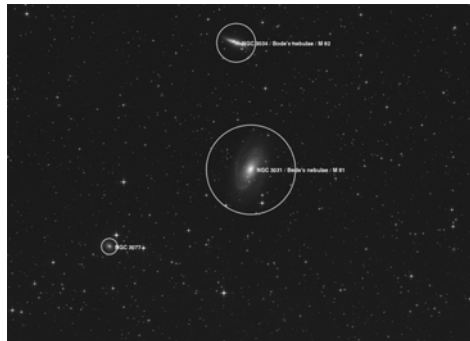
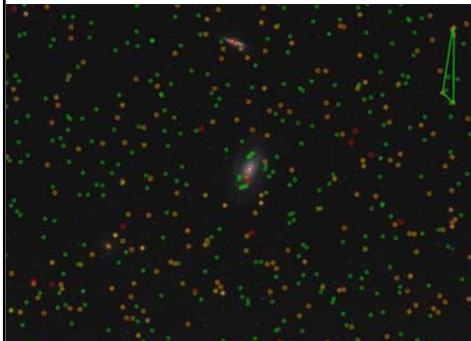
Roweis, Lang, Mierle, Hogg & Blanton



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

Example

Roweis, Lang, Mierle, Hogg & Blanton



A beautiful image of Bode's nebula (c.2007) by Peter Bresseler, from starlightfriend.de
<http://astrometry.net/gallery.html>

Today: key ideas

- Invariant features: distinctive matches possible in spite of significant view change, useful for wide baseline stereo
- 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

Coming up

- Next week:
 - Model-based object recognition
 - Face recognition, detection
- Read FP 18.1-18.5, FP 22.1-22.3