Announcements

- PS2 due Saturday 11:59 am

2

Approximating the Laplacian

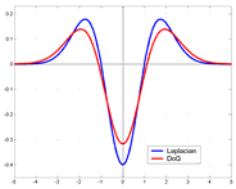
- We can approximate the Laplacian with a difference of Gaussians; more efficient to implement.

$$L = \sigma^2 \left(G_{xx}(x,y,\sigma) + G_{yy}(x,y,\sigma) \right)$$

(Laplacian)

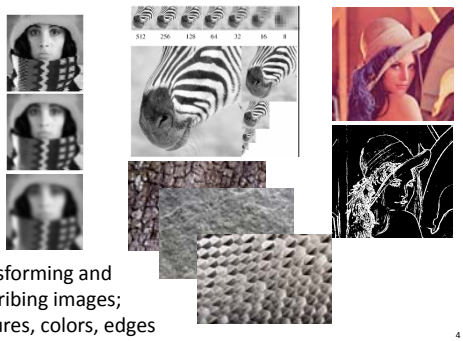
$$DoG = G(x,y,k\sigma) - G(x,y,\sigma)$$

(Difference of Gaussians)



3

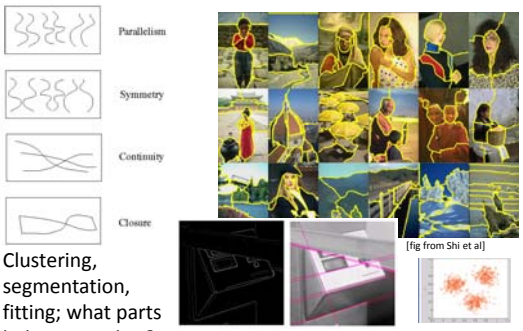
Recap: Features and filters



Transforming and describing images; textures, colors, edges

4
Kristen Grauman

Recap: Grouping & fitting

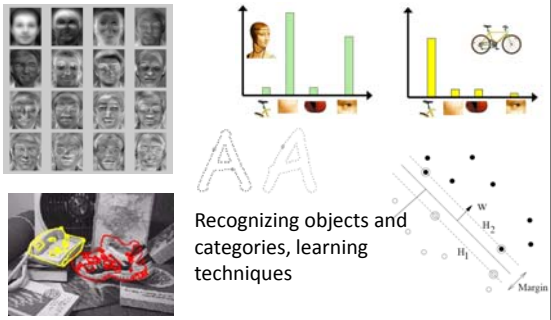


[fig from Shi et al]

Clustering, segmentation, fitting; what parts belong together?

5
Kristen Grauman

Recognition and learning



Recognizing objects and categories, learning techniques

6
Kristen Grauman

Matching local features

7
Kristen Grauman

Matching local features

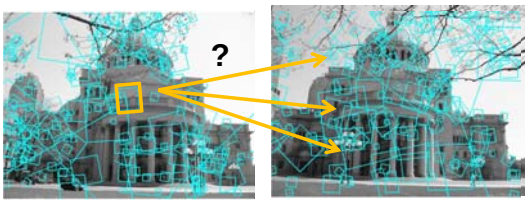


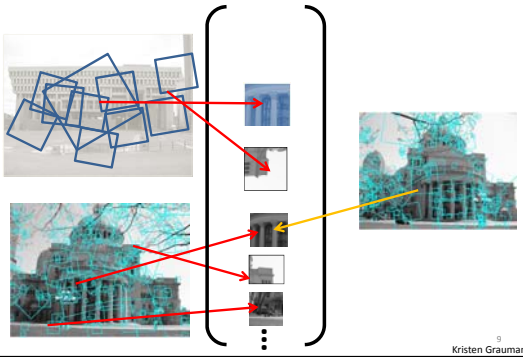
Image 1 Image 2

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)

8
Kristen Grauman

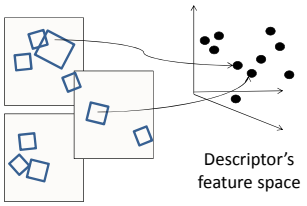
Indexing local features



9
Kristen Grauman

Indexing local features

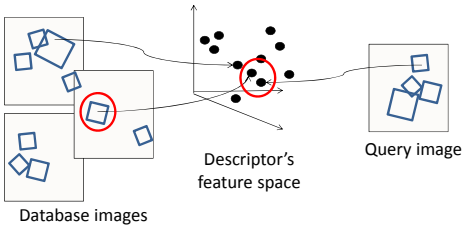
- Each patch / region has a descriptor, which is a point in some high-dimensional feature space (e.g., SIFT)



10
Kristen Grauman

Indexing local features

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



11
Kristen Grauman

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?

12
Kristen Grauman

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".

13
Kristen Grauman

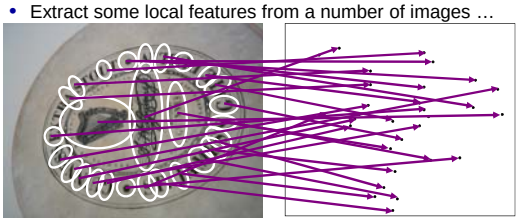
Text retrieval vs. image search

- What makes the problems similar, different?

14
Kristen Grauman

Visual words: main idea

- Extract some local features from a number of images …

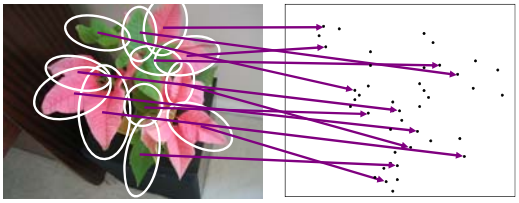


e.g., SIFT descriptor space: each point is 128-dimensional

Slide credit: D. Nister, CVPR 2006

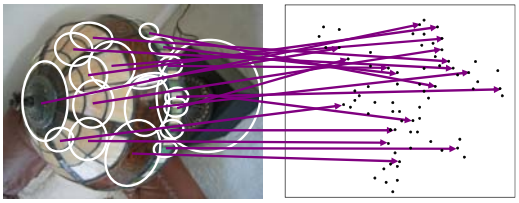
15

Visual words: main idea



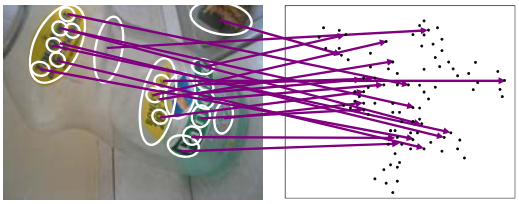
16

Visual words: main idea

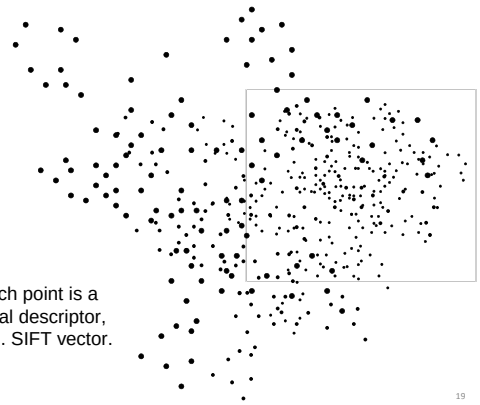


17

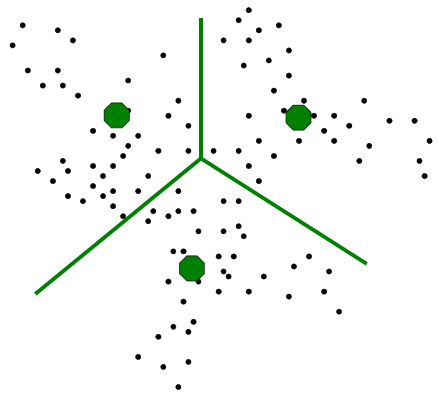
Visual words: main idea



18



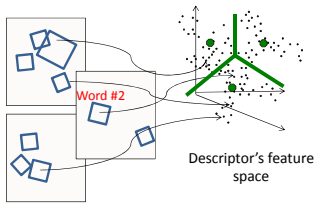
19



20

Visual words

- Map high-dimensional descriptors to tokens/words by quantizing the feature space



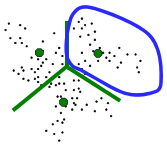
- Quantize via clustering, let cluster centers be the prototype "words"
- Determine which word to assign to each new image region by finding the closest cluster center.

21

Kristen Grauman

Visual words

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



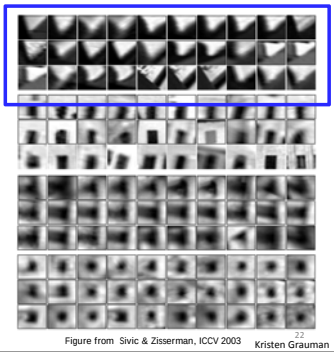
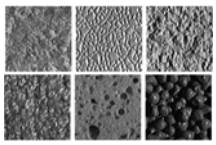


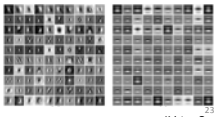
Figure from Sivic & Zisserman, ICCV 2003

22
Kristen Grauman

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.

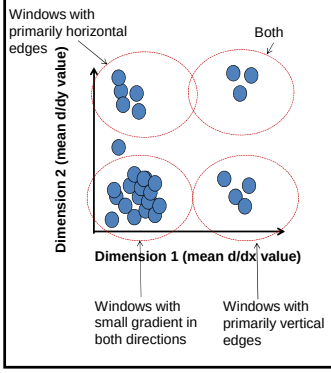




Leung & Malik 1999; Varma & Zisserman, 2002

23
Kristen Grauman

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

24
Kristen Grauman

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

25
Kristen Grauman

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

26
Kristen Grauman

Inverted file index

When will this give us a significant gain in efficiency?

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

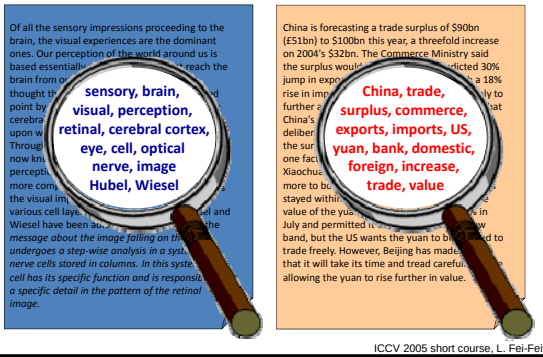
- New query image is mapped to indices of database images that share a word.

27
Kristen Grauman

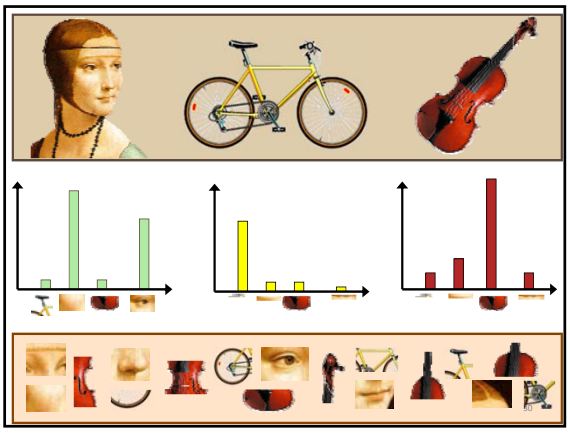
- If a local image region is a visual word, how can we summarize an image (the document)?

28

Analogy to documents

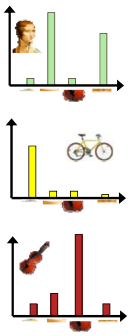


ICCV 2005 short course, L. Fei-Fei



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.

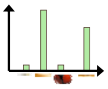


31

Comparing bags of words

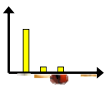
- Rank frames by normalized scalar product between their occurrence counts---*nearest neighbor* search for similar images.

[1 8 1 4]



$\vec{d}_j$

[5 1 1 0]



$\vec{q}$

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

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

for vocabulary of *V* words

32

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

33

Example

retrieved shots

Slide from Andrew Zisserman
Sivic & Zisserman, ICCV 2003

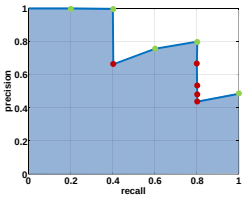
34

Scoring retrieval quality

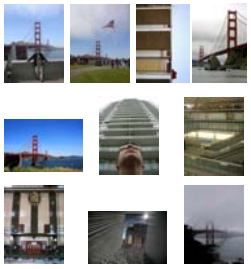
Query

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

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



Results (ordered):

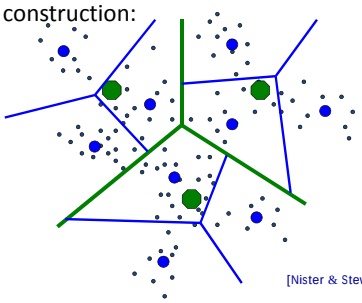


35

Slide credit: Ondrej Chum

Vocabulary Trees: hierarchical clustering for large vocabularies

- Tree construction:



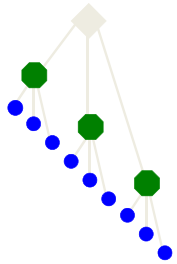
[Nister & Stewenius, CVPR'06]

36

Slide credit: David Nister

Vocabulary Tree

- Training: Filling the tree



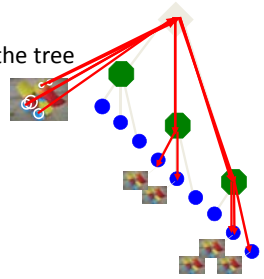
[Nister & Stewenius, CVPR'06]

37

Slide credit: David Nister

Vocabulary Tree

- Training: Filling the tree



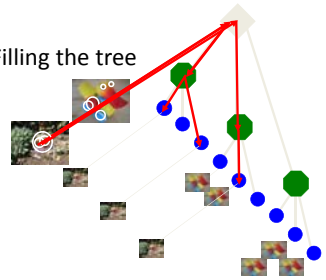
[Nister & Stewenius, CVPR'06]

38

Slide credit: David Nister

Vocabulary Tree

- Training: Filling the tree



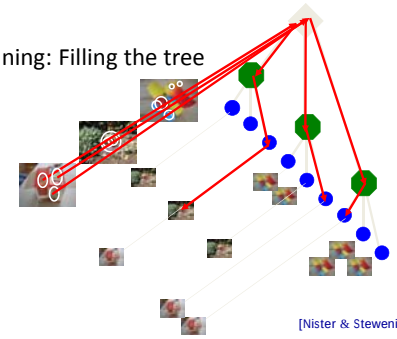
[Nister & Stewenius, CVPR'06]

39

Slide credit: David Nister

Vocabulary Tree

- Training: Filling the tree



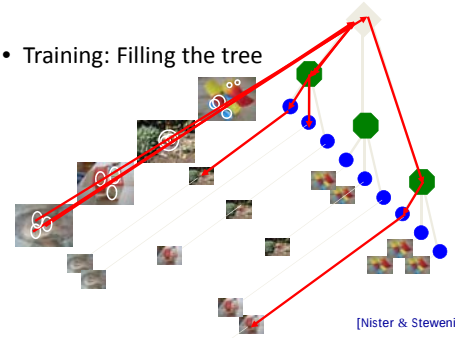
[Nister & Stewenius, CVPR'06]

40

Slide credit: David Nister

Vocabulary Tree

- Training: Filling the tree



[Nister & Stewenius, CVPR'06]

41

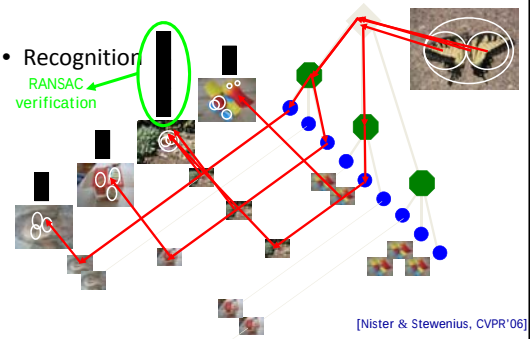
Slide credit: David Nister

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

42

Vocabulary Tree

- Recognition



[Nister & Stewenius, CVPR'06]

43

Slide credit: David Nister

Bags of words: pros and cons

+ flexible to geometry / deformations / viewpoint
+ compact summary of image content
+ provides vector representation for sets
+ 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

44

Summary So Far

- **Matching local invariant features**: useful to provide matches to find objects and scenes.
- **Bag of words** representation: quantize feature space to make discrete set of visual words
- **Inverted index**: pre-compute index to enable faster search at query time

45

Instance recognition

- Motivation – visual search
- Visual words
  - quantization, index, bags of words
- Spatial verification
  - affine; RANSAC, Hough
- Other text retrieval tools
  - tf-idf, query expansion
- Example applications

46

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?

47

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?

48

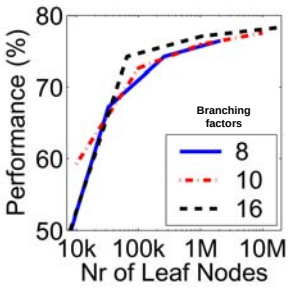
Kristen Grauman

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?

49
Kristen Grauman

Vocabulary size



Results for recognition task with 6347 images

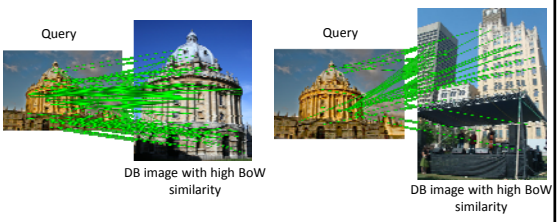
50
Nister & Stewenius, CVPR 2006
Kristen Grauman

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?

51
Kristen Grauman

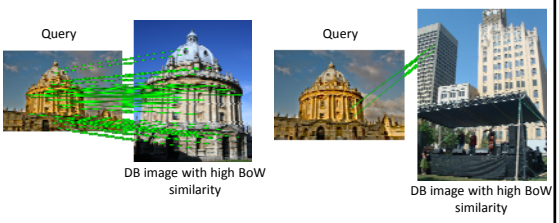
Spatial Verification



Both image pairs have many visual words in common.

52
Slide credit: Ondrej Chum

Spatial Verification



Only some of the matches are mutually consistent

53
Slide credit: Ondrej Chum

Spatial Verification: two basic strategies

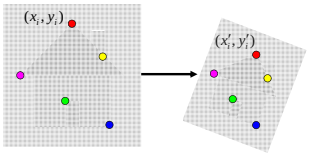
- RANSAC
  - Typically sort by BoW similarity as initial filter
  - Verify by checking support (inliers) for possible transformations
    - e.g., "success" if find a transformation with > N inlier correspondences
- Generalized Hough Transform
  - Let each matched feature cast a vote on location, scale, orientation of the model object
  - Verify parameters with enough votes

54
Kristen Grauman

RANSAC verification

55

Recall: Fitting an affine transformation



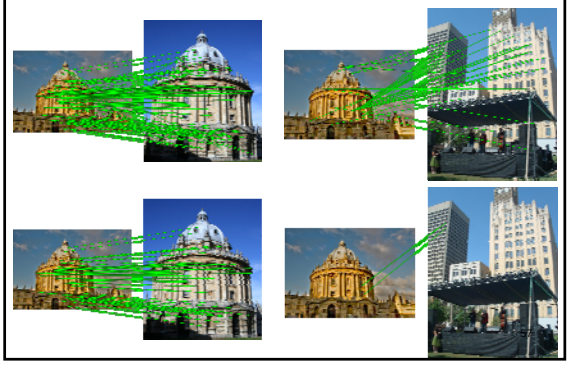
Approximates viewpoint changes for roughly planar objects and roughly orthographic cameras.

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

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

56

RANSAC verification



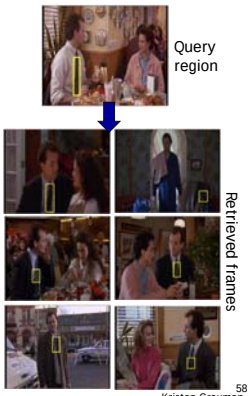
57

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



Visual Object Recognition Tutorial

Kristen Grauman 58

Example Applications

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

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

Visual Object Recognition Tutorial 59

Application: Large-Scale Retrieval

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

[Philbin CVPR'07]

Visual Object Recognition Tutorial 60

Visual Object Recognition Tutorial

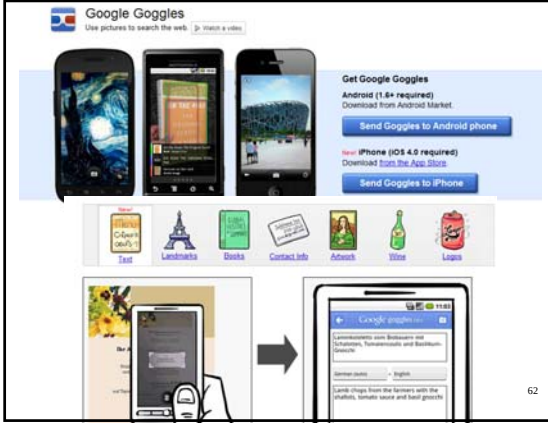
Web Demo: Movie Poster Recognition

50'000 movie posters indexed

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

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

61



62

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

63
Kristen Grauman

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

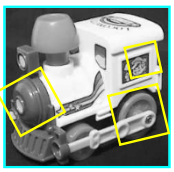
Novel image

64

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

65

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.

66

Slide credit: Lana Lazebnik

Example result

Background subtract for model boundaries

Objects recognized,

Recognition in spite of occlusion

67

[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.

68

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

69

Kristen Grauman

Questions?

See you Tuesday!

70