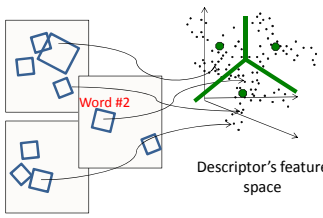
Generic object recognition

May 16th, 2019

Yong Jae Lee
UC Davis

Visual words

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

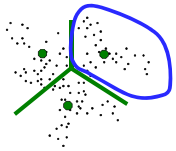


- Quantize via clustering, let cluster centers be the prototype "words"
- Determine which word to assign to each new image region by finding the closest cluster center.

Kristen Grauman

Visual words

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



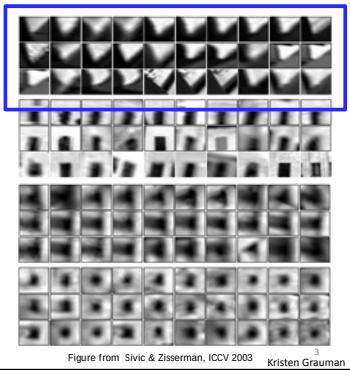


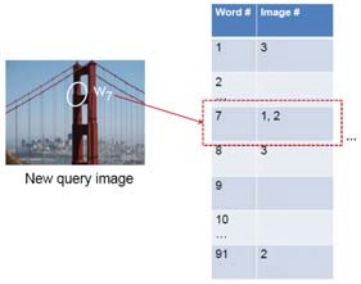
Figure from Sivic & Zisserman, ICCV 2003

Kristen Grauman

Inverted file index

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

Inverted file index



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

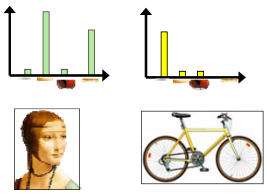
Bags of visual words

- 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

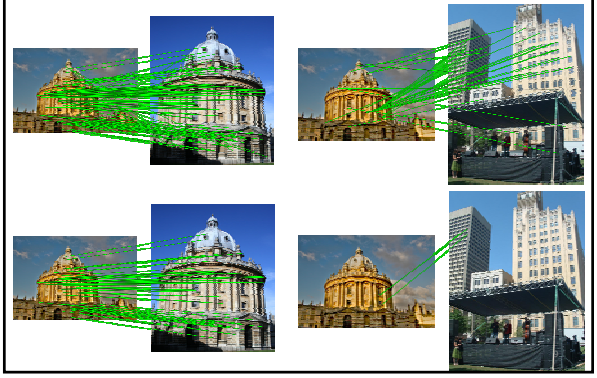
Kristen Grauman

Application: Large-Scale Retrieval

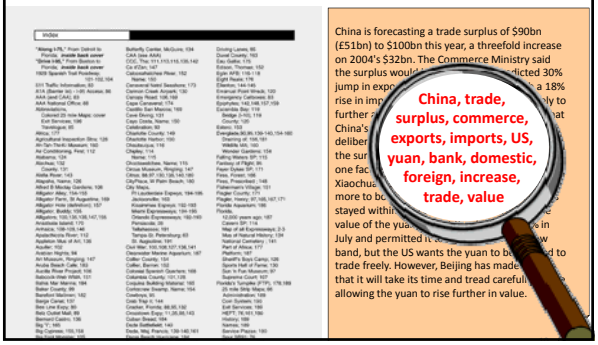
Query Results from 5k Flickr images (demo available for 100k set)

[Philbin CVPR'07]

Spatial Verification: RANSAC



What else can we borrow from text retrieval?



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)

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

Number of occurrences of word i in document d

Number of words in document d

Total number of documents in database

Number of documents word i occurs in, in whole database

11

Kristen Grauman

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

12

Slide credit: Ondrej Chum

Query expansion

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

13

Slide credit: Ondrej Chum

Recognition via local-feature based alignment

Pros:

- Effective when we are able to find reliable features within clutter
- Great results for matching specific instances

Cons:

- Spatial verification as post-processing – not seamless, expensive for large-scale problems
- Not suited for generic category recognition

14

Kristen Grauman

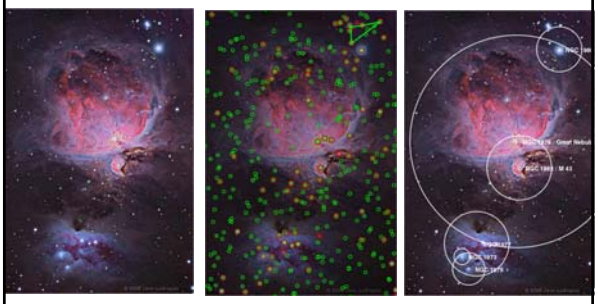
Making the Sky Searchable:
Fast Geometric Hashing for Automated Astrometry

Sam Roweis, Dustin Lang & Keir Mierle
University of Toronto

David Hogg & Michael Blanton
New York University

15

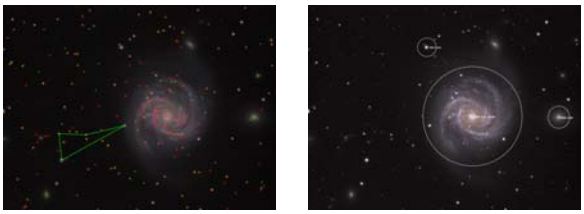
Example



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

16

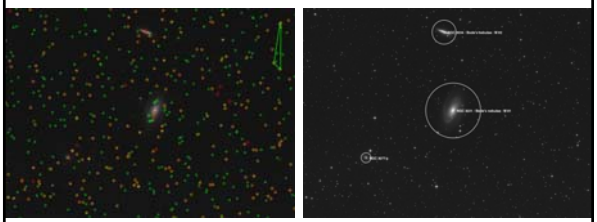
Example



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

17

Example



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

18

Today

- Generic object recognition

19

What does recognition involve?

20

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Verification: is that a lamp?

21

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Detection: are there people?

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

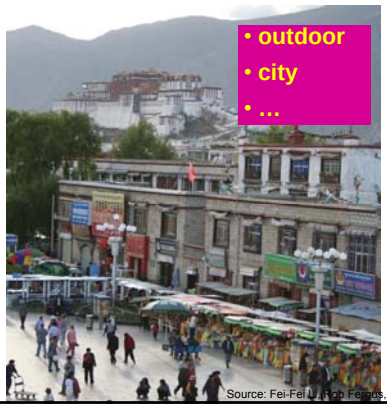
Identification: is that Potala Palace?

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Object categorization

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Scene and context categorization



Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

25

Instance-level recognition problem

26

Generic categorization problem

27

Object Categorization

- Task Description
  - "Given a small number of training images of a category, recognize a-priori unknown instances of that category and assign the correct category label."
- Which categories are feasible visually?

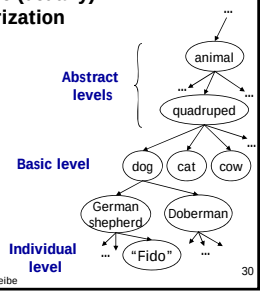
"Fido" | German shepherd | dog | animal | living being

Visual Object Categories

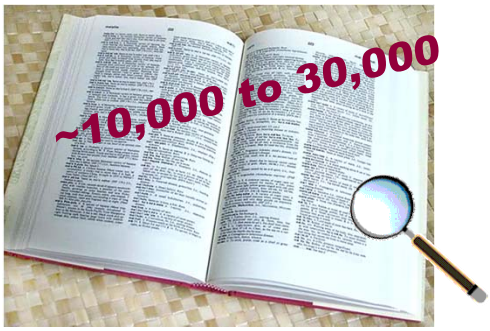
- Basic Level Categories in human categorization [Rosch 76, Lakoff 87]
  - The highest level at which category members have similar perceived shape
  - The highest level at which a single mental image reflects the entire category
  - The level at which human subjects are usually fastest at identifying category members
  - The first level named and understood by children
  - The highest level at which a person uses similar motor actions for interaction with category members

Visual Object Categories

- Basic-level categories in humans seem to be defined predominantly visually.
- There is evidence that humans (usually) start with basic-level categorization *before* doing identification.



How many object categories are there?



Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Biederman 1987

31

32

Other Types of Categories

- **Functional Categories**
  - **e.g. chairs =** *"something you can sit on"*

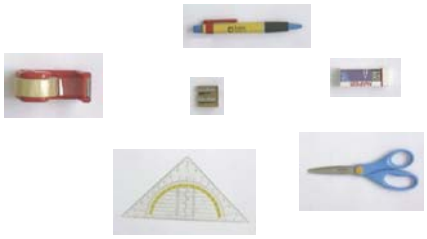
K. Grauman, B. Leibe

Visual Object Recognition Tutorial

33

Other Types of Categories

- **Ad-hoc categories**
  - **e.g.** *"something you can find in an office environment"*



Why recognition?

– Recognition a fundamental part of perception
  - e.g., robots, autonomous agents

– Organize and give access to visual content
  - Connect to information
  - Detect trends and themes

Posing visual queries

Yeh et al., MIT

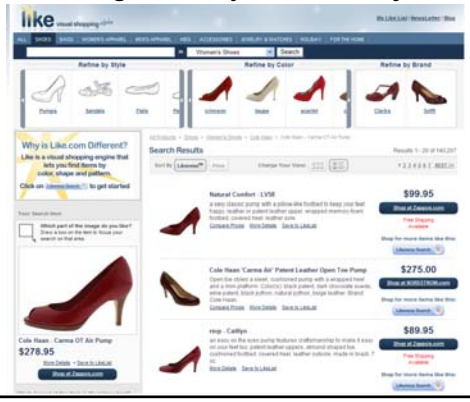
Belhumeur et al.

Kooaba, Bay & Quack et al.

Autonomous agents able to detect objects

37

Finding visually similar objects



38

Discovering visual patterns

Objects Sivic & Zisserman

Categories Lee & Grauman

Actions Wang et al.

39

Kristen Grauman

Auto-annotation

Gammeter et al.

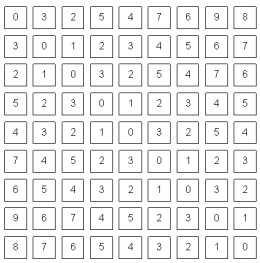
T. Berg et al.

40

Kristen Grauman

What are the challenges?

What we see



What a computer sees

Challenges: robustness


Illumination


Object pose


Clutter


Occlusions


Intra-class appearance


Viewpoint

43

Kristen Grauman

Challenges: robustness

Realistic scenes are crowded, cluttered, have overlapping objects.

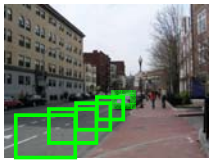
44

Challenges: importance of context

45

slide credit: Fei-Fei, Fergus & Torralba

Challenges: importance of context



46

Challenges: complexity

flickr
10 billion images

facebook
250 billion images

imgur
1 billion images served daily

YouTube
300 hours uploaded per minute

photobucket
10 billion images

cisco

Almost 90% of web traffic is visual!

47

Challenges: complexity

- Thousands to millions of pixels in an image
- 30+ degrees of freedom in the pose of articulated objects (humans)
- About half of the cerebral cortex in primates is devoted to processing visual information [Felleman and van Essen 1991]

48

Kristen Grauman

Challenges: learning with minimal supervision

49

Kristen Grauman

What works today

- Reading license plates, zip codes, checks

50

Source: Lana Lazebnik

What works today

- Reading license plates, zip codes, checks
- Fingerprint recognition

51

Source: Lana Lazebnik

What works today

- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection

52

Source: Lana Lazebnik

What works today

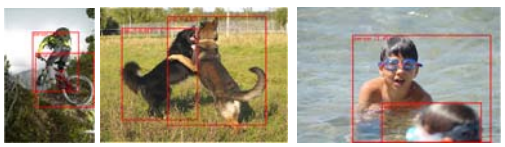
- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection
- Recognition of flat textured objects (CD covers, book covers, etc.)

53

Source: Lana Lazebnik

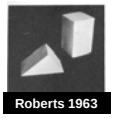
What works today

- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection
- Recognition of flat textured objects (CD covers, book covers, etc.)
- Recognition of generic categories(*)!



54

Progress charted by datasets



Roberts 1963

COIL

1963 … 1996

Slide credit: Kristen Grauman

Progress charted by datasets

MIT-CMU Faces

INRIA Pedestrians

UIUC Cars

1963 … 1996 2000

Slide credit: Kristen Grauman

Progress charted by datasets

MSRC 21 Objects

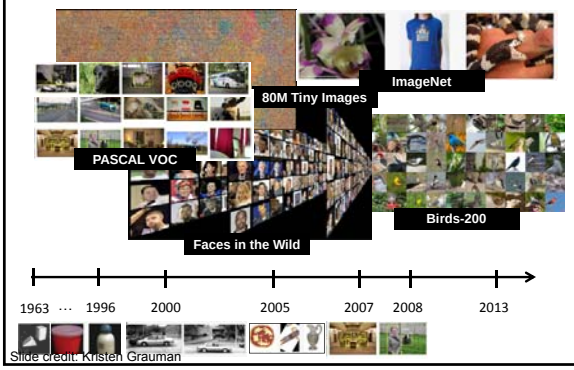
Caltech-101

Caltech-256

1963 … 1996 2000 2005

Slide credit: Kristen Grauman

Progress charted by datasets



Slide credit: Kristen Grauman

Evolution of methods

- Hand-crafted models
- 3D geometry
- Hypothesize and align

- Hand-crafted features
- Learned models
- Data-driven

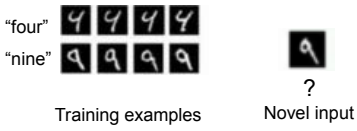
- "End-to-end" learning of features and models*,**

* Labeled data availability
** Architecture design decisions, parameters.

Kristen Grauman

Supervised classification

- Given a collection of *labeled* examples, come up with a function that will predict the labels of new examples.



Training examples Novel input

- How good is the function we come up with to do the classification?
- Depends on
  - Mistakes made
  - Cost associated with the mistakes

60

Kristen Grauman

Supervised classification

- Given a collection of *labeled* examples, come up with a function that will predict the labels of new examples.
- Consider the two-class (binary) decision problem
  - L(4→9): Loss of classifying a 4 as a 9
  - L(9→4): Loss of classifying a 9 as a 4
- **Risk** of a classifier *s* is expected loss:

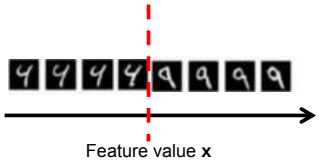
$$R(s) = \Pr(4 \to 9 \mid \text{using } s)L(4 \to 9) + \Pr(9 \to 4 \mid \text{using } s)L(9 \to 4)$$

- We want to choose a classifier so as to minimize this total risk

61

Kristen Grauman

Supervised classification



Feature value **x**

Optimal classifier will minimize total risk.

At decision boundary, either choice of label yields same expected loss.

If we choose class "four" for point x at boundary, expected loss is:

$$= P(\text{class is } 9 \mid \mathbf{x})\, L(9 \to 4) + P(\text{class is } 4 \mid \mathbf{x})L(4 \to 4)$$

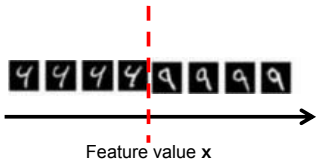
If we choose class "nine" for point x at boundary, expected loss is:

$$= P(\text{class is } 4 \mid \mathbf{x})\, L(4 \to 9)$$

62

Kristen Grauman

Supervised classification



Feature value **x**

Optimal classifier will minimize total risk.

At decision boundary, either choice of label yields same expected loss.

So, best decision boundary is at point **x** where

$$P(\text{class is } 9 \mid \mathbf{x})\, L(9 \to 4) = \mathrm{P}(\text{class is } 4 \mid \mathbf{x})L(4 \to 9)$$

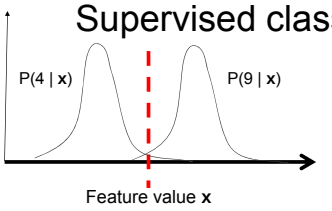
To classify a new point, choose class with lowest expected loss; i.e., choose "four" if

$$P(4 \mid \mathbf{x})L(4 \to 9) > P(9 \mid \mathbf{x})L(9 \to 4)$$

63

Kristen Grauman

Supervised classification



Optimal classifier will minimize total risk.

At decision boundary, either choice of label yields same expected loss.

So, best decision boundary is at point **x** where

$$P(\text{class is } 9 \mid \mathbf{x})\, L(9 \to 4) = P(\text{class is } 4 \mid \mathbf{x}) L(4 \to 9)$$

To classify a new point, choose class with lowest expected loss; i.e., choose "four" if

$$P(4 \mid \mathbf{x}) L(4 \to 9) > P(9 \mid \mathbf{x}) L(9 \to 4)$$

How to evaluate these probabilities?

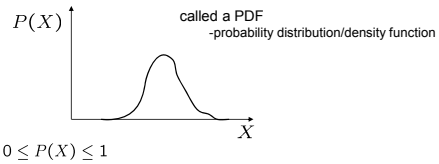
64

Kristen Grauman

Probability

Basic probability

- X is a random variable
- P(X) is the probability that X achieves a certain value



called a PDF
-probability distribution/density function

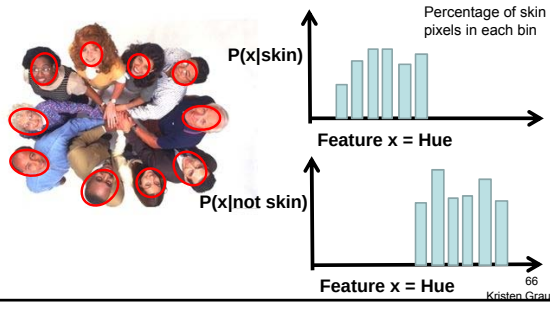
- $0 \le P(X) \le 1$
- $\int_{-\infty}^{\infty} P(X)dX = 1$ or $\sum P(X) = 1$
  continuous X discrete X
- Conditional probability: P(X | Y)
  - probability of X given that we already know Y

65

Source: Steve Seitz

Example: learning skin colors

- We can represent a class-conditional density using a histogram (a "non-parametric" distribution)



66

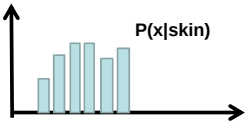
Kristen Grauman

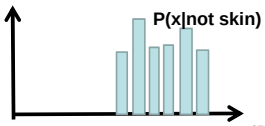
Example: learning skin colors

- We can represent a class-conditional density using a histogram (a "non-parametric" distribution)

Now we get a new image, and want to label each pixel as skin or non-skin.

What's the probability we care about to do skin detection?





67

Kristen Grauman

Bayes rule

$$P(skin \mid x) = \frac{P(x \mid skin)P(skin)}{P(x)}$$

posterior — likelihood — prior

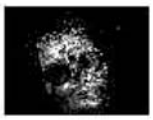
Where does the prior come from?

Why use a prior?

68

Example: classifying skin pixels

Now for every pixel in a new image, we can estimate probability that it is generated by skin.



Brighter pixels → higher probability of being skin

Classify pixels based on these probabilities

- if $p(\text{skin}|\boldsymbol{x}) > \theta$, classify as skin
- if $p(\text{skin}|\boldsymbol{x}) < \theta$, classify as not skin

69

Example: classifying skin pixels

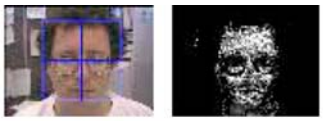


Figure 6: A video image and its flesh probability image

Figure 7: Orientation of the flesh probability distribution marked on the source video image

Gary Bradski, 1998

70

Example: classifying skin pixels

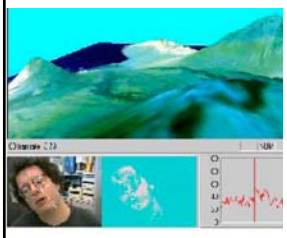


Figure 13: CAMSHIFT-based face tracker used to over a 3D graphic's model of Hawaii

Figure 12: CAMSHIFT-based face tracker used to play Quake 2 hands free by inserting control variables into the mouse queue

Using skin color-based face detection and pose estimation as a video-based interface

Gary Bradski, 1998

71

Supervised classification

- Want to minimize the expected misclassification
- Two general strategies
  - Use the training data to build representative probability model; separately model class-conditional densities and priors (*generative*)
  - Directly construct a good decision boundary, model the posterior (*discriminative*)

72

Coming up

- Face detection
- Categorization with local features and part-based models
- Deep convolutional neural networks

73

Questions?

See you Tuesday!

74