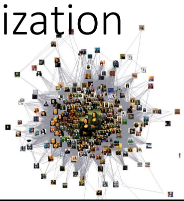
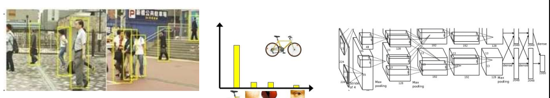
Mining, and Intro to Categorization

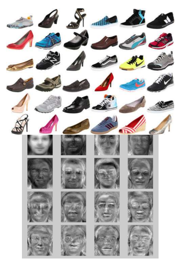


Tues April 11

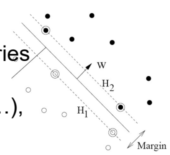
Kristen Grauman
UT Austin

Recognition and learning





Recognizing categories (objects, scenes, activities, attributes…), learning techniques



Last time

- Instance recognition wrap up:
  - Spatial verification
  - Sky mapping example
  - Query expansion

Review questions

- Does an inverted file index sacrifice accuracy in bag-of-words image retrieval? Why or why not?
- Name a pro and con of query expansion.
- Why does a single SIFT match cast a 4D vote for the Generalized Hough spatial verification model?
- What does a perfect precision recall curve look like?

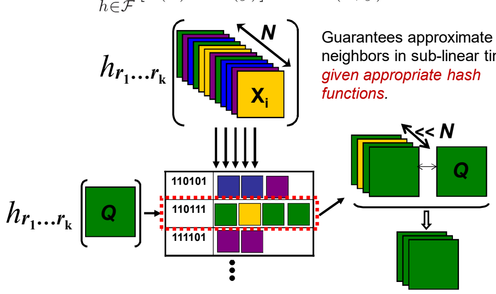
Today

- Discovering visual patterns
  - Randomized hashing algorithms
  - Mining large-scale image collections
- Introduction to visual categorization

Locality Sensitive Hashing (LSH)

[Indyk and Motwani '98, Gionis et al.'99, Charikar '02, Andoni et al. '04]

$$\Pr_{h \in \mathcal{F}} [h(x) = h(y)] = sim(x, y)$$



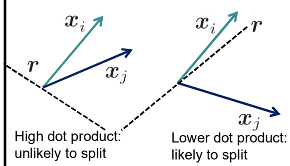
Guarantees approximate near neighbors in sub-linear time, *given appropriate hash functions*.

Kristen Grauman

LSH function example: inner product similarity

The probability that a *random hyperplane* separates two unit vectors depends on the angle between them:

$$\Pr[\mathrm{sign}(\boldsymbol{x}_i^T \boldsymbol{r}) = \mathrm{sign}(\boldsymbol{x}_j^T \boldsymbol{r})] = 1 - \frac{1}{\pi}\cos^{-1}(\boldsymbol{x}_i^T \boldsymbol{x}_j)$$



Corresponding hash function:

$$h_{\boldsymbol{r}}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{r}^T\boldsymbol{x} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

for $\vec{r}_i \sim N(\mu = 0, \sigma^2 = 1)$

[Goemans and Williamson 1995, Charikar 2004]

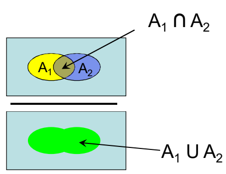
Kristen Grauman

LSH function example: Min-hash for set overlap similarity

[Broder, 1999]

$$\Pr_{h \in \mathcal{F}}[h(x) = h(y)] = sim(x, y)$$

$$\mathrm{sim}(\mathcal{A}_1, \mathcal{A}_2) = \frac{|\mathcal{A}_1 \cap \mathcal{A}_2|}{|\mathcal{A}_1 \cup \mathcal{A}_2|} \in \langle 0, 1 \rangle$$



Kristen Grauman

LSH function example: Min-hash for set overlap similarity

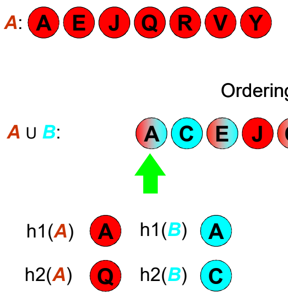
Vocabulary	Set ***A***	Set ***B***	Set ***C***
A B C D E F	A B C	B C D	A E F

Random orderings								min-Hash		
f_1:	0.41 3	0.90 6	0.22 2	0.59 5	0.75 4	0.07 1	~ Un (0,1)	C	C	F
f_2:	0.19 1	0.31 2	0.94 6	0.55 3	0.88 5	0.63 4	~ Un (0,1)	A	B	A
f_3:	3	2	1	6	4	5		C	C	A
f_4:	4	3	5	6	1	2		B	B	E

overlap (***A***,***B***) = 3/4 (1/2) overlap (***A***,***C***) = 1/4 (1/5) overlap (***B***,***C***) = 0 (0)

Slide credit: Ondrej Chum *[Broder, 1999]*

LSH function example: Min-hash for set overlap similarity



$$P(h(A) = h(B)) = \frac{|A \cap B|}{|A \cup B|}$$

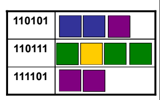
Slide credit: Ondrej Chum

[Broder, 1999]

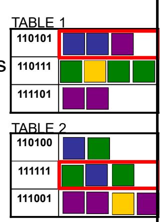
Multiple hash functions and tables

- Generate *k* such hash functions, concatenate outputs into hash key:

$$P\left(h_{1,\ldots,k}(x) = h_{1,\ldots,k}(y)\right) =$$



- To increase recall, search multiple independently generated hash tables
  - Search/rank the union of collisions in each table, or
  - Require that two examples in at least *T* of the tables to consider them similar.

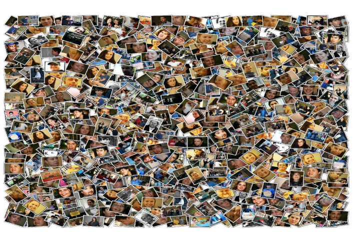


Kristen Grauman

Mining for common visual patterns

In addition to visual search, want to be able to **summarize, mine, and rank** the large collection as a whole.

- What is common?
- What is unusual?
- What co-occurs?
- Which exemplars are most representative?



Kristen Grauman

Mining for common visual patterns

In addition to visual search, want to be able to **summarize, mine, and rank** the large collection as a whole.

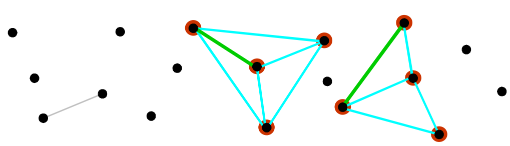
We'll look at a few examples:

- **Connected component clustering** via hashing
  - [Geometric Min-hash, Chum et al. 2009]
- **Visual Rank** to choose "image authorities"
  - [Jing and Baluja, 2008]
- **Frequent item-set mining** with spatial patterns
  - [Quack et al., 2007]

Kristen Grauman

Connected component clustering with hashing

1. Detect seed pairs via hash collisions
2. Hash to related images
3. Compute connected components of the graph



Contrast with frequently used quadratic-time clustering algorithms

Slide credit: Ondrej Chum

Geometric Min-hash

[Chum, Perdoch, Matas, CVPR 2009]

- Main idea: build spatial relationships into the hash key construction:
  - Select first hash output according to min hash ("central word")
  - Then append subsequent hash outputs from within its neighborhood

Figure from Ondrej Chum

Results: Geometric Min-hash clustering

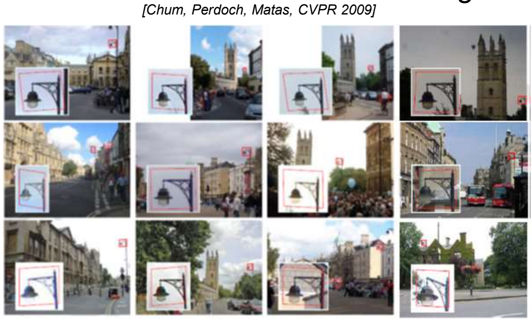
[Chum, Perdoch, Matas, CVPR 2009]

100 000 Images downloaded from FLICKR
Includes 11 Oxford Landmarks with manually labeled ground truth

Slide credit: Ondrej Chum

Results: Geometric Min-hash clustering

[Chum, Perdoch, Matas, CVPR 2009]



Discovering small objects

Slide credit: Ondrej Chum

Results: Geometric Min-hash clustering

[Chum, Perdoch, Matas, CVPR 2009]

Discovering small objects

Slide credit: Ondrej Chum

Mining for common visual patterns

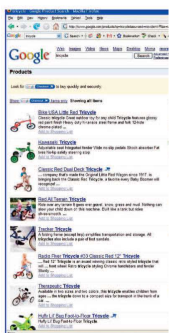
In addition to visual search, want to be able to **summarize, mine, and rank** the large collection as a whole.

We'll look briefly at a few recent examples:

- **Connected component clustering** via hashing [Geometric Min-hash, Chum et al. 2009]
- **Visual Rank** to choose "image authorities" [Jing and Baluja, 2008]
- **Frequent item-set mining** with spatial patterns [Quack et al., 2007]

Visual Rank: motivation

Product search



Mixed-type search

- **Goal:** select small set of "best" images to display among millions of candidates

Kristen Grauman

Visual Rank

[Jing and Baluja, PAMI 2008]

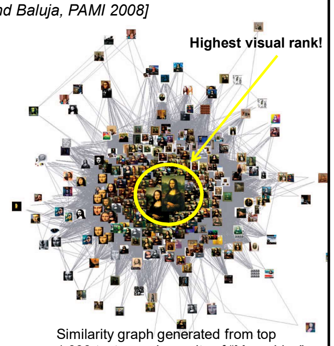
- Compute relative "authority" of an image based on random walk principle.
  - Application of PageRank to visual data
- **Main ideas**:
  - Graph weights = number of matched local features between two images
  - Exploit text search to narrow scope of each graph
  - Use LSH to make similarity computations efficient

Kristen Grauman

Results: Visual Rank

[Jing and Baluja, PAMI 2008]

Original has more matches to rest

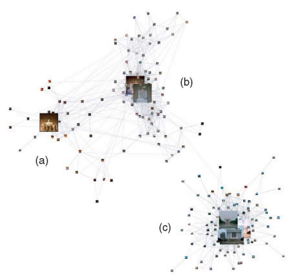


Similarity graph generated from top 1,000 text search results of "Mona-Lisa"

Kristen Grauman

Results: Visual Rank

[Jing and Baluja, PAMI 2008]



Similarity graph generated from top 1,000 text search results of "Lincoln Memorial".
Note the diversity of the high-ranked images.

Kristen Grauman

Mining for common visual patterns

In addition to visual search, want to be able to **summarize, mine, and rank** the large collection as a whole.

We'll look briefly at a few recent examples:

- **Connected component clustering** via hashing [Geometric Min-hash, Chum et al. 2009]
- **Visual Rank** to choose "image authorities" [Jing and Baluja, 2008]
- **Frequent item-set mining** with spatial patterns [Quack et al., 2007]

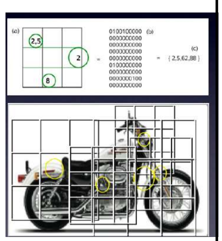
Frequent item-sets

Kristen Grauman

Frequent item-set mining for spatial visual patterns

[Quack, Ferrari, Leibe, Van Gool, CIVR 2006, ICCV 2007]

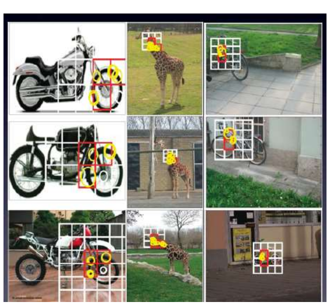
- What configurations of local features frequently occur in large collection?
- **Main idea**: Identify *item-sets* (visual word layouts) that often occur in *transactions* (images)
- Efficient algorithms from data mining (e.g., Apriori algorithm, Agrawal 1993)



Kristen Grauman

Frequent item-set mining for spatial visual patterns

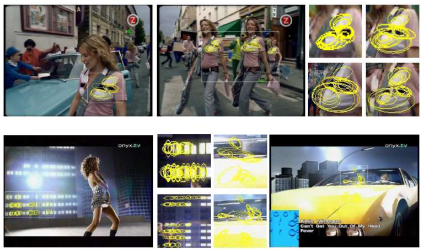
[Quack, Ferrari, Leibe, Van Gool, CIVR 2006, ICCV 2007]



Kristen Grauman

Frequent item-set mining for spatial visual patterns

[Quack, Ferrari, Leibe, Van Gool, CIVR 2006, ICCV 2007]

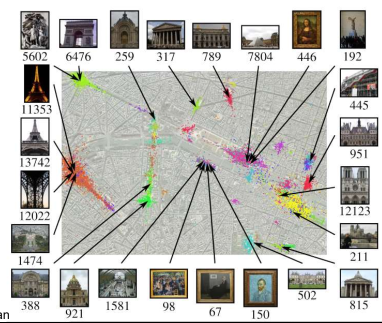


Two example itemset clusters

Kristen Grauman

Discovering favorite views

Discovering Favorite Views of Popular Places with Iconoid Shift. T. Weyand and B. Leibe. ICCV 2011.

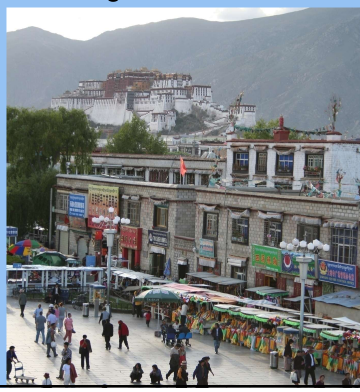


Kristen Grauman

Today

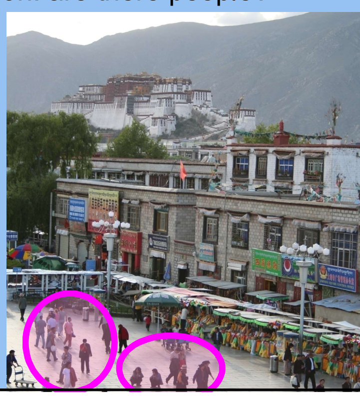
- Discovering visual patterns
  - Randomized hashing algorithms
  - Mining large-scale image collections
- **Introduction to visual categorization**

What does recognition involve?

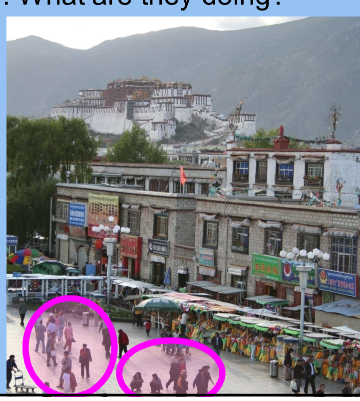


Fei-Fei Li

Detection: are there people?



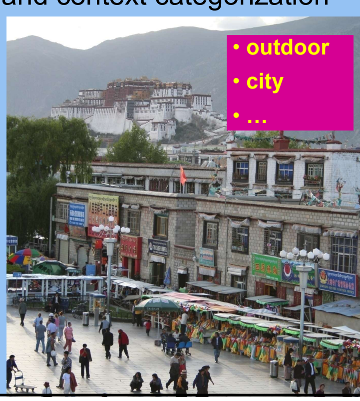
Activity: What are they doing?



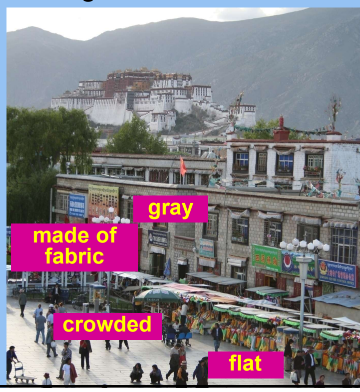
Object categorization

Instance recognition

Scene and context categorization



Attribute recognition



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

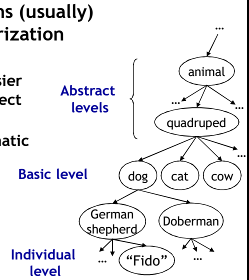
Visual Object Categories

- 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 Recognition Tutorial

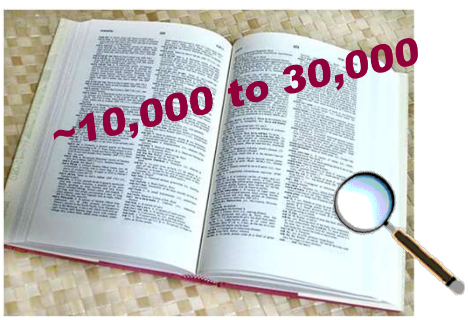
Visual Object Categories

- 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.
  - ⇒ Basic-level categorization is easier and faster for humans than object identification!
  - ⇒ How does this transfer to automatic classification algorithms?



K. Grauman, B. Leibe

How many object categories are there?



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

Biederman 1987

Other Types of Categories

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

Visual Object Recognition Tutorial

K. Grauman, B. Leibe

Why recognition?

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

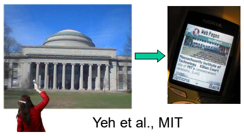
– Organize and give access to visual content
  - Connect to information
  - Detect trends and themes

Autonomous agents able to detect objects

Slide credit: Kristen Grauman

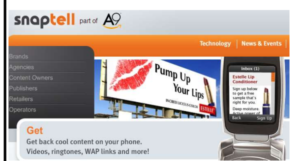
http://www.darpa.mil/grandchallenge/gallery.asp

Posing visual queries



Yeh et al., MIT

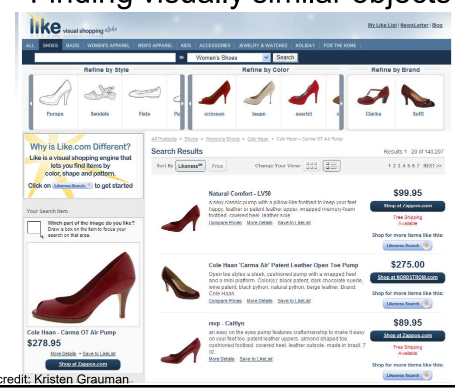
Belhumeur et al.



Kooaba, Bay & Quack et al.

Slide credit: Kristen Grauman

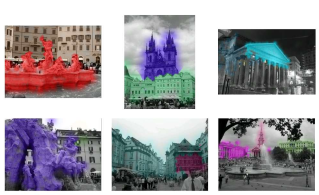
Finding visually similar objects



Slide credit: Kristen Grauman

Exploring community photo collections

Snavely et al.



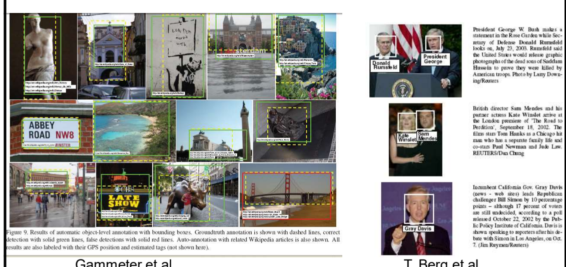
Slide credit: Kristen Grauman

Simon & Seitz

Discovering visual patterns

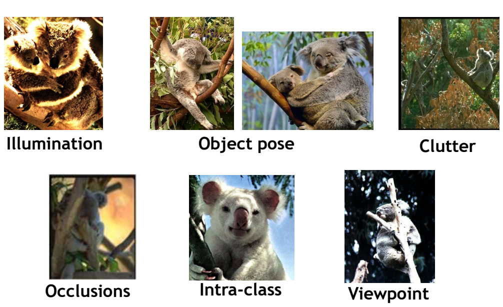
Slide credit: Kristen Grauman

Auto-annotation



Slide credit: Kristen Grauman

Challenges: robustness



Slide credit: Kristen Grauman

Challenges: context and human experience

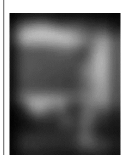
Context cues

Slide credit: Kristen Grauman

Challenges: context and human experience

Context cues

Function



Dynamics

Slide credit: Kristen Grauman

Video credit: J. Davis

Challenges: complexity

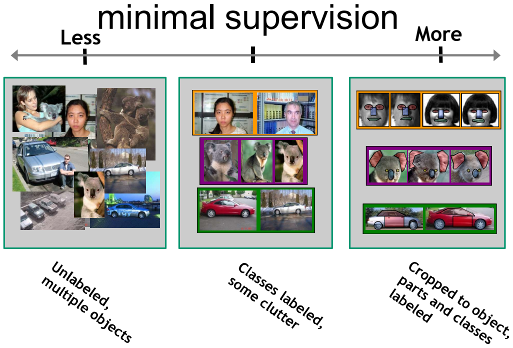
- Millions of pixels in an image
- 30,000 human recognizable object categories
- 30+ degrees of freedom in the pose of articulated objects (humans)
- Billions of images online
- 82 years to watch all videos uploaded to YouTube per day!

…

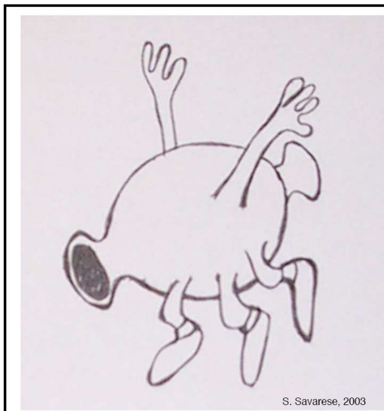
- About half of the cerebral cortex in primates is devoted to processing visual information [Felleman and van Essen 1991]

Slide credit: Kristen Grauman

Challenges: learning with minimal supervision



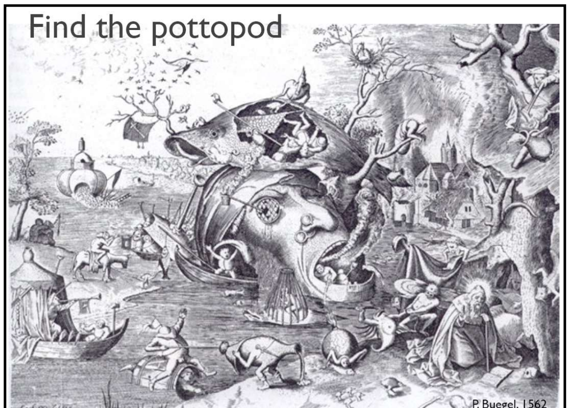
Slide credit: Kristen Grauman



This is a pottopod

Slide from Pietro Perona, 2004 Object Recognition workshop

Find the pottopod



Slide from Pietro Perona, 2004 Object Recognition workshop

What kinds of things work best today?

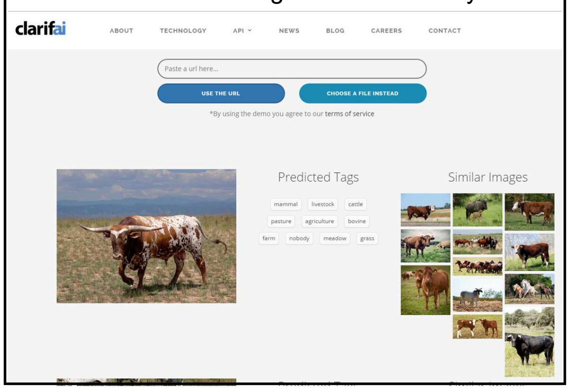
Reading license plates, zip codes, checks

Frontal face detection

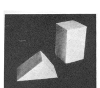
Recognizing flat, textured objects (like books, CD covers, posters)

Fingerprint recognition

What kinds of things work best today?

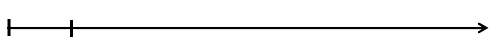


Progress charted by datasets



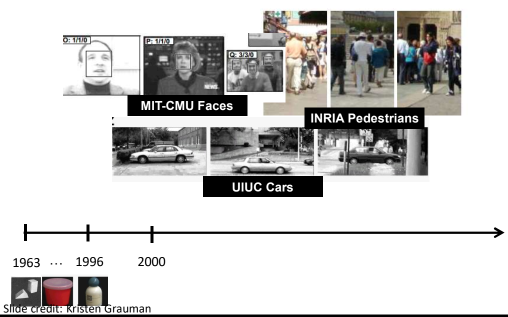
Roberts 1963

COIL

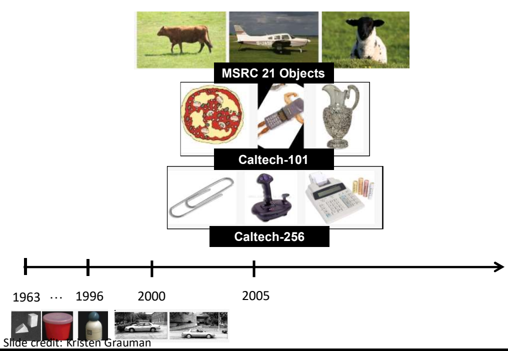


Slide credit: Kristen Grauman

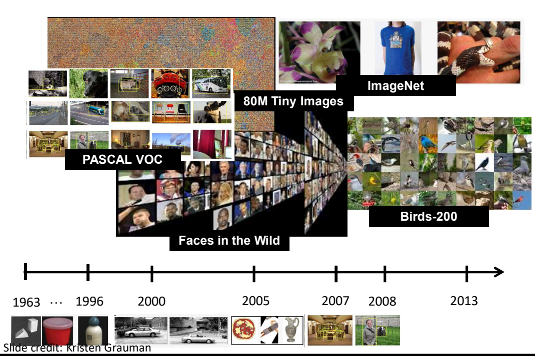
Progress charted by datasets



Progress charted by datasets



Progress charted by datasets



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*,**

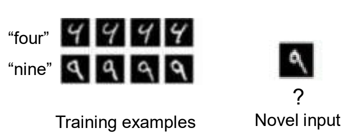
* Labeled data availability
** Architecture design decisions, parameters.

Next

- Sliding window object detection (Faces!)

Supervised classification

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



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

Supervised classification

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

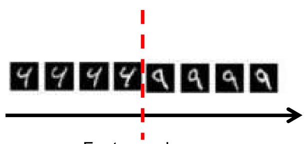
- Consider the two-class (binary) decision problem
  - $L(4 \rightarrow 9)$: Loss of classifying a 4 as a 9
  - $L(9 \rightarrow 4)$: Loss of classifying a 9 as a 4

- **Risk** of a classifier *s* is expected loss:

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

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

Supervised classification



Feature value **x**

Optimal classifier will minimize total risk.

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

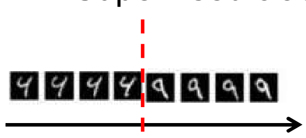
If we choose class “four” at boundary, expected loss is:

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

If we choose class “nine” at boundary, expected loss is:

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

Supervised classification



Feature value **x**

Optimal classifier will minimize total risk.

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

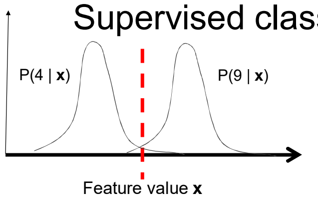
So, best decision boundary is at point **x** where

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

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

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

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

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

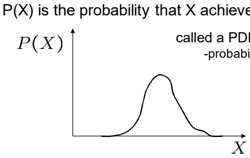
To classify a new point, choose class with lowest expected loss; i.e., choose "four" if

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

Probability

Basic probability

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



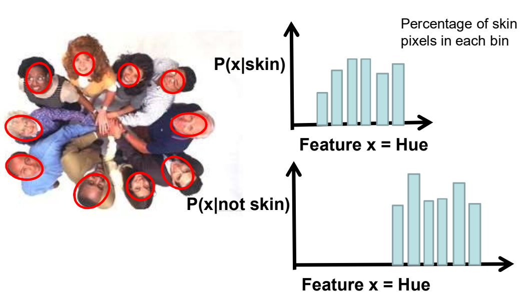
called a PDF
-probability distribution/density function

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

Source: Steve Seitz

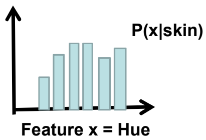
Example: learning skin colors

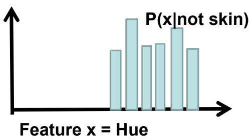
- We can represent a class-conditional density using a histogram (a "non-parametric" distribution)



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?

Bayes rule

$$\underbrace{P(skin \mid x)}_{\text{posterior}} = \frac{\overbrace{P(x \mid skin)}^{\text{likelihood}}\overbrace{P(skin)}^{\text{prior}}}{P(x)}$$

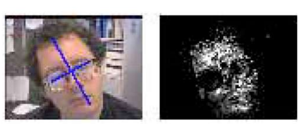
$$P(skin \mid x) \;\alpha\; P(x \mid skin)P(skin)$$

Where does the prior come from?

Why use a prior?

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

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

Example: classifying skin pixels

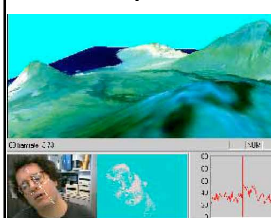


Figure 13: CAMSHIFT-based face tracker used to over a 3D graphic's model of Hawaii

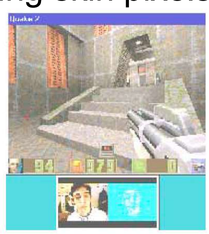


Figure 12: CAMSHIFT-based face tracker used to play Quake 2 hands free by inserting control variables into the mouse queue

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

Gary Bradski, 1998

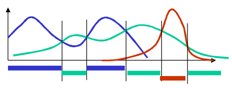
Supervised classification

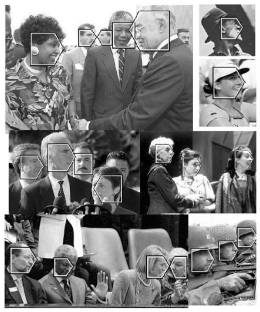
- Want to minimize the expected misclassification
- Two general strategies
  - Use the training data to build representative probability model; separately model class-conditional densities and priors (*generative*)
  - Directly construct a good decision boundary, model the posterior (*discriminative*)

General classification

This same procedure applies in more general circumstances

- More than two classes
- More than one dimension





H. Schneiderman and T.Kanade

Example: face detection

- Here, X is an image region
  - dimension = # pixels
  - each face can be thought of as a point in a high dimensional space

H. Schneiderman, T. Kanade. "A Statistical Method for 3D Object Detection Applied to Faces and Cars". IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)
http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/hws/www/CVPR00.pdf

Source: Steve Seitz

Next

- Sliding window object detection (Faces!)