343H: Honors AI

Lecture 25:

Neural networks

Applications, part 1

4/24/2014

Kristen Grauman

UT Austin

Today

- Neural networks
- Supervised learning in visual recognition

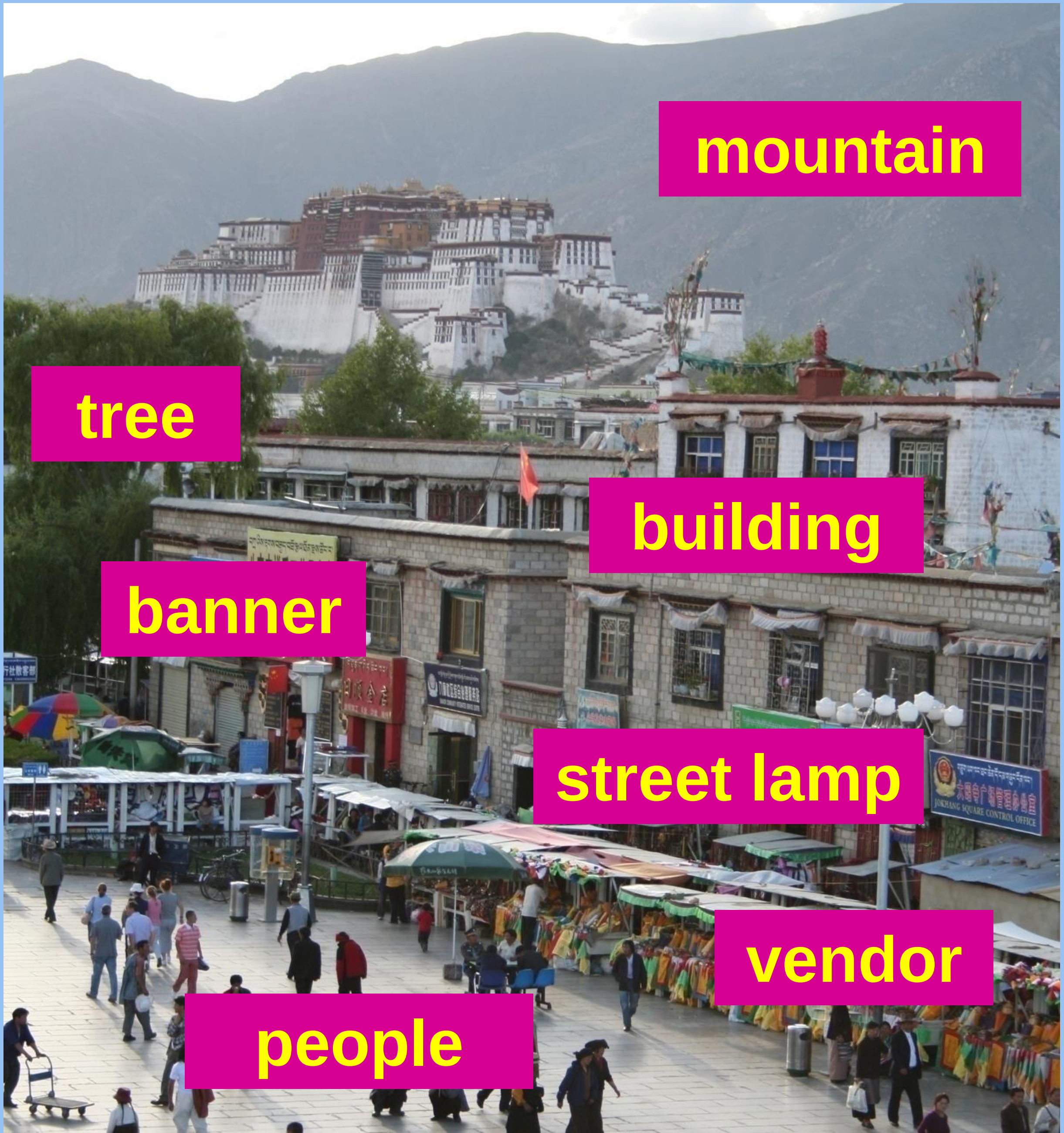
What does recognition involve?

Verification: is that a lamp?

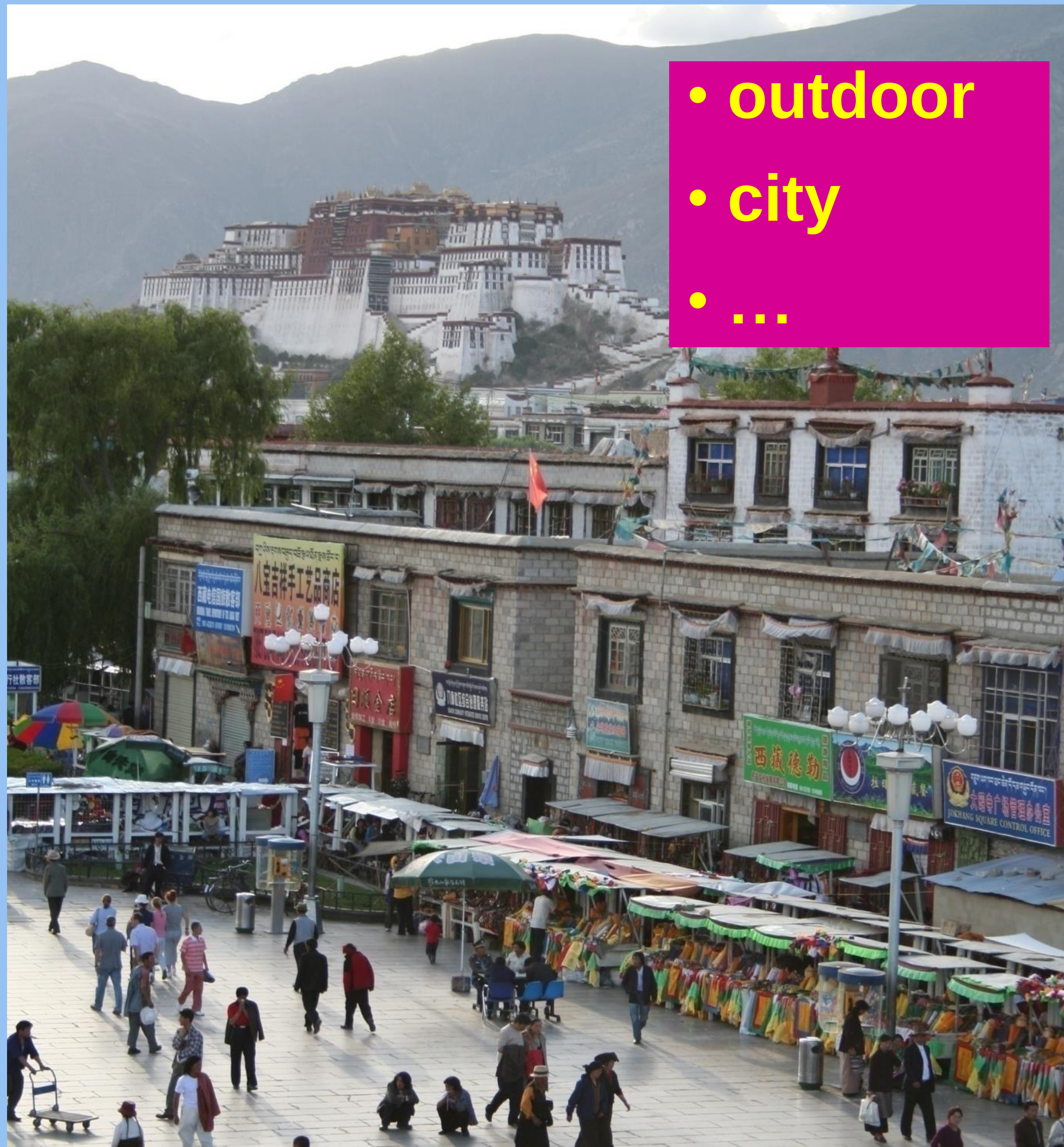
Detection: are there people?

Identification: is that Potala Palace?

Object categorization



Scene and context categorization



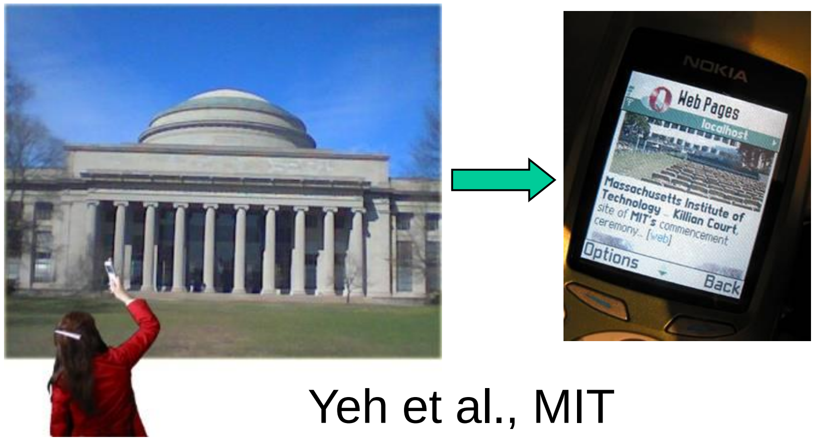
- **outdoor**
- **city**
- **…**

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



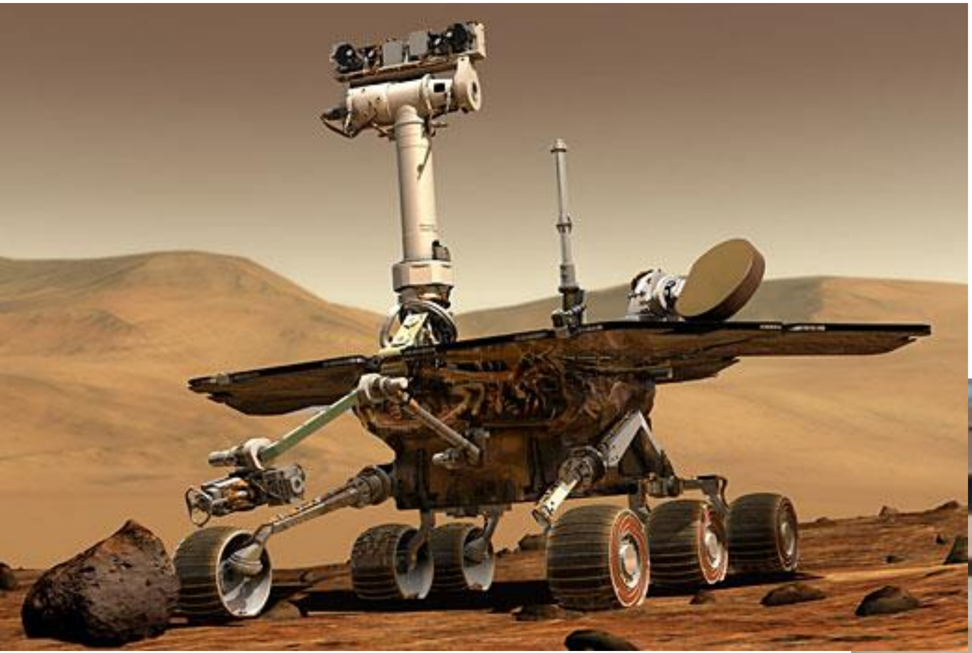
Yeh et al., MIT

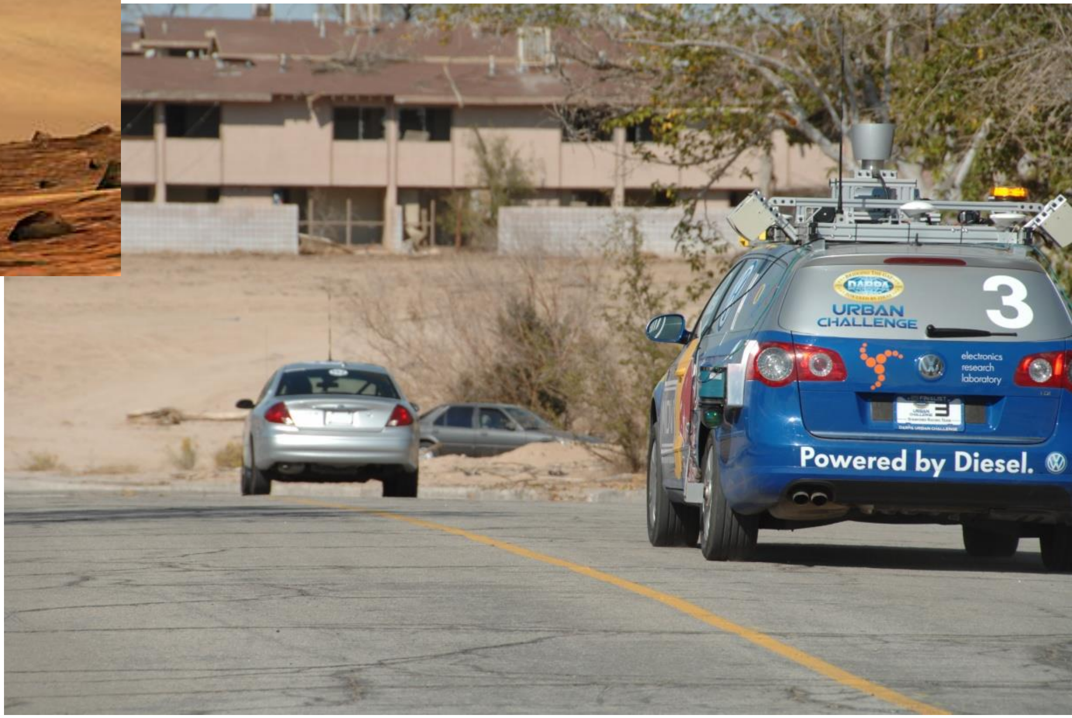
Belhumeur et al.

Kooaba, Bay & Quack et al.

Slide credit: Kristen Grauman

Autonomous agents able to detect objects

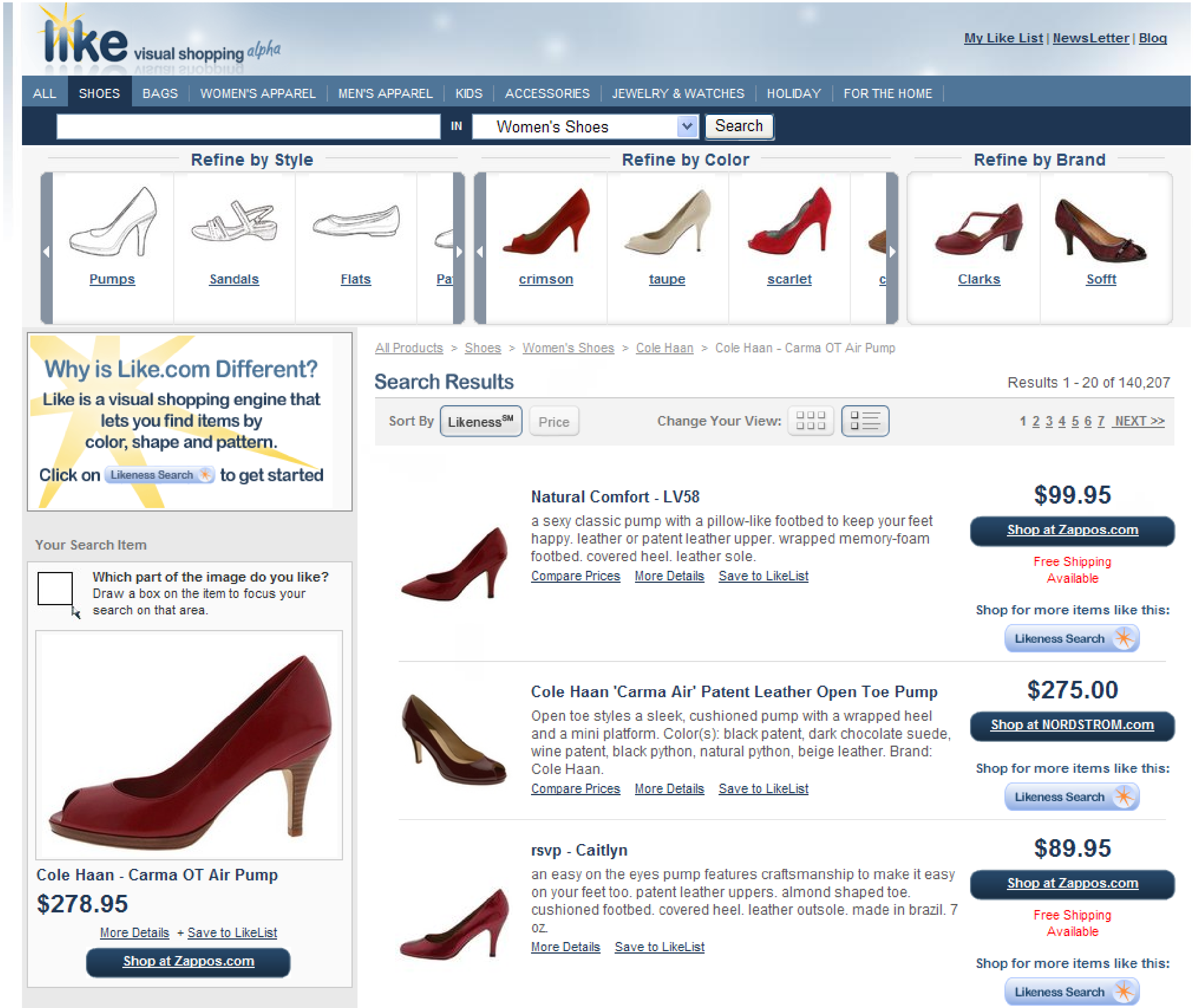




Slide credit: Kristen Grauman

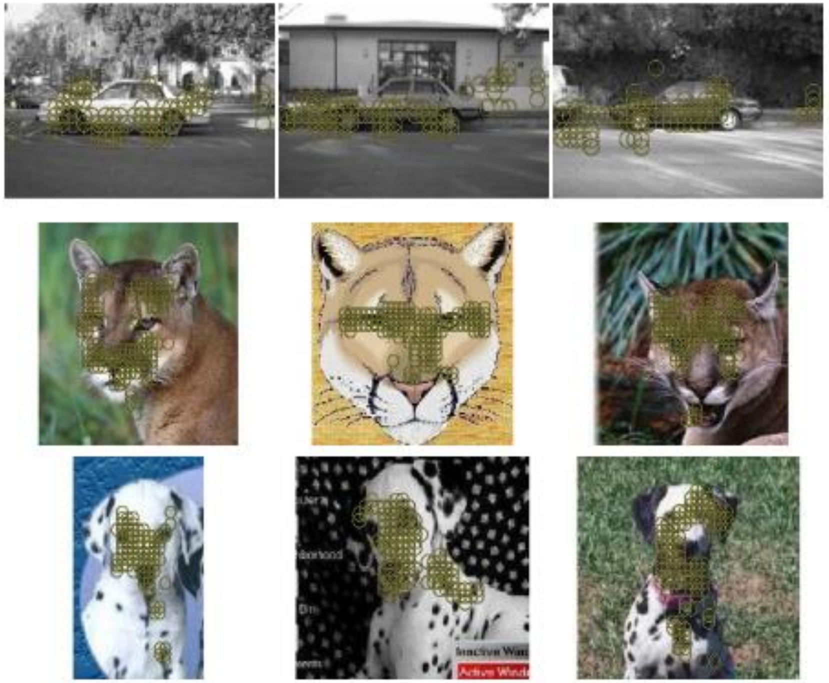
http://www.darpa.mil/grandchallenge/gallery.asp

Finding visually similar objects



Discovering visual patterns

Objects Sivic & Zisserman



Lee & Grauman

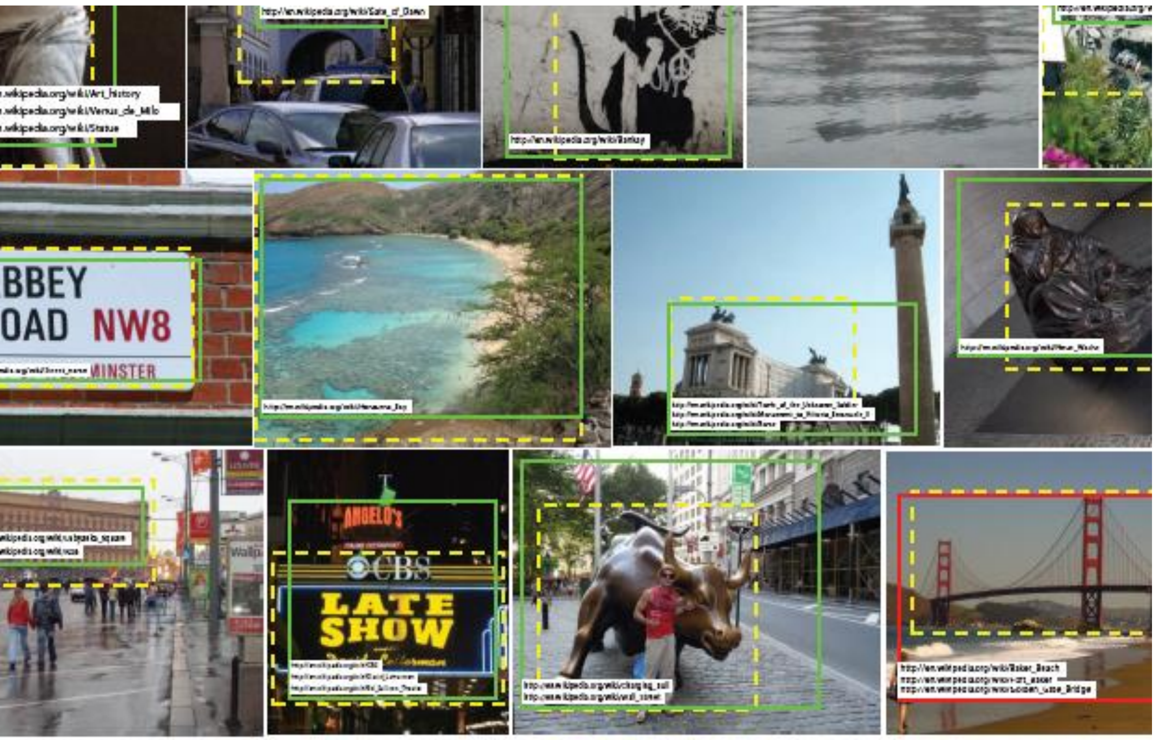
Categories

Wang et al.

Actions

Slide credit: Kristen Grauman

Auto-annotation



esults of automatic object-level annotation with bounding boxes. Groundtruth annotation is shown with dash
ith solid green lines, false detections with solid red lines. Auto-annotation with related Wikipedia articles is
lso labeled with their GPS position and estimated tags (not shown here).

Gammeter et al.

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

British director Sam Mendes and his partner actress Kate Winslet arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars Tom Hanks as a Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung

Incumbent California Gov. Gray Davis (news - web sites) leads Republican challenger Bill Simon by 10 percentage points – although 17 percent of voters are still undecided, according to a poll released October 22, 2002 by the Public Policy Institute of California. Davis is shown speaking to reporters after his debate with Simon in Los Angeles, on Oct. 7. (Jim Ruymen/Reuters)

T. Berg et al.

Slide credit: Kristen Grauman

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?

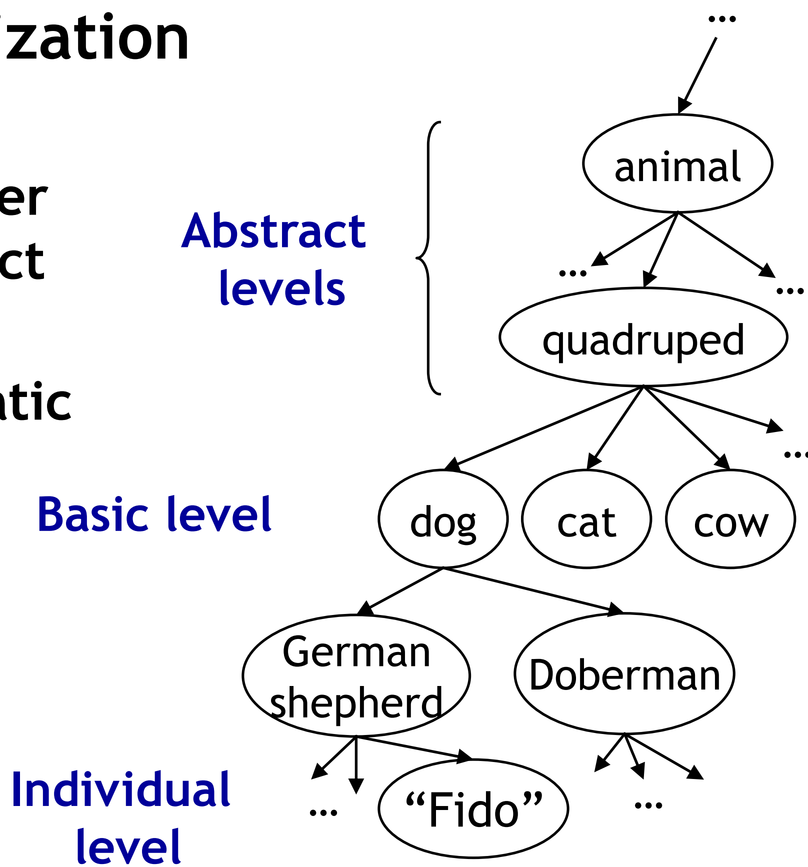
"Fido" | German shepherd | dog | animal | living being

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

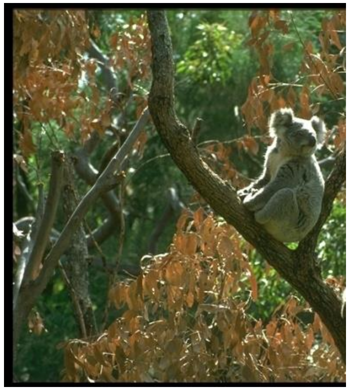
- Basic-level categories in humans seem to be defined predominantly visually.
- There is evidence that humans (usually) start with basic-level categorization *before* doing identification.
  - ⇒ Basic-level categorization is easier and faster for humans than object identification!
  - ⇒ How does this transfer to automatic classification algorithms?



Challenges: robustness

Illumination

Object pose



Clutter

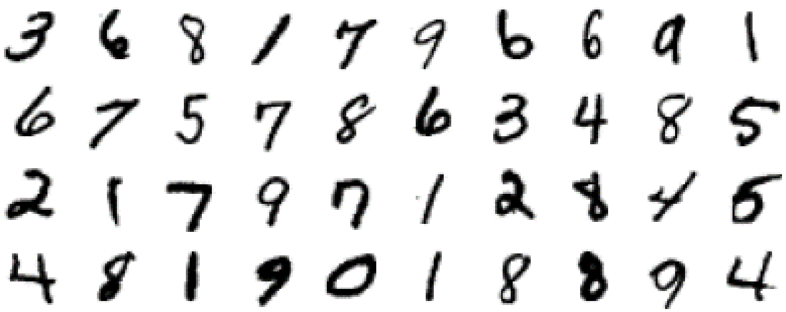
Occlusions

Intra-class appearance

Viewpoint

Slide credit: Kristen Grauman

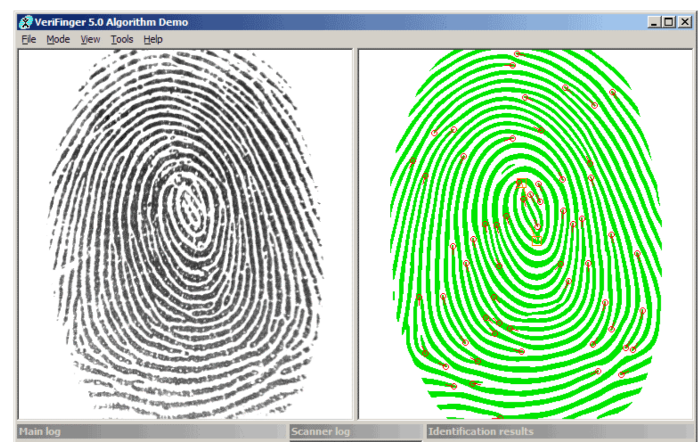
What kinds of things work best today?



Reading license plates, zip codes, checks

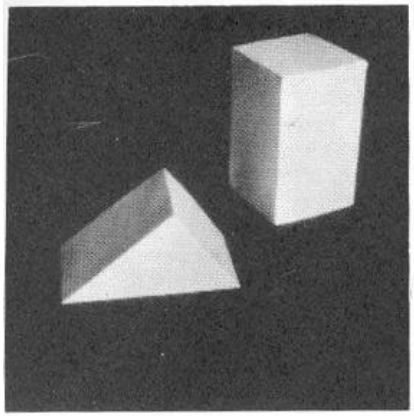
Frontal face detection

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

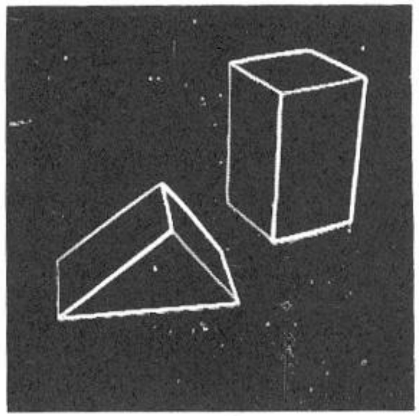


Fingerprint recognition

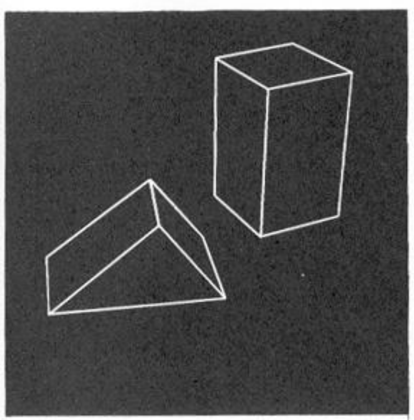
Inputs in 1963…



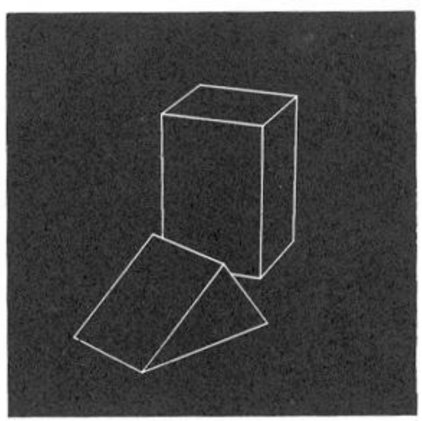
(a) Original picture.



(b) Differentiated picture.



(c) Line drawing.



(d) Rotated view.

L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

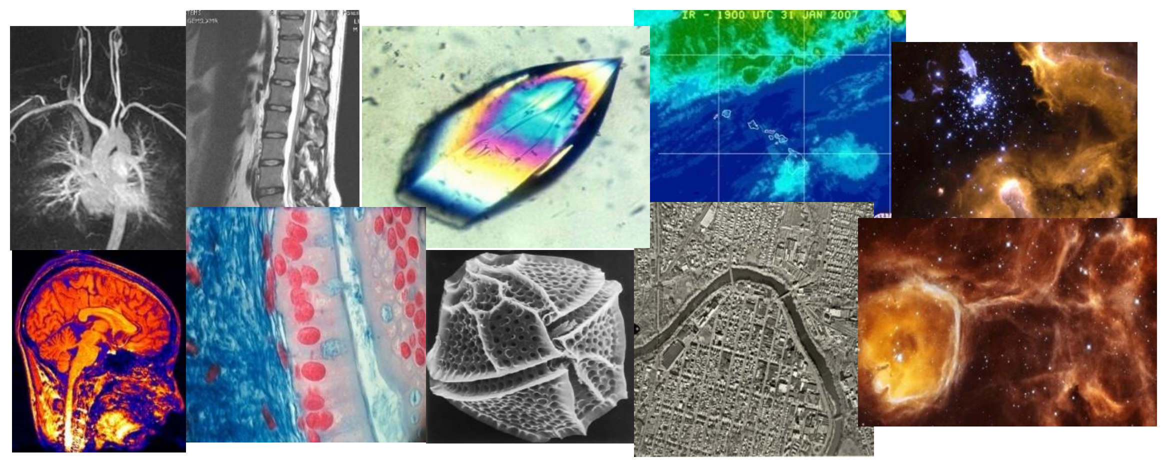
… and inputs today

Personal photo albums

Movies, news, sports

Google Image Search Picasa flickr webshots picsearch YouTube Broadcast Yourself

Surveillance and security



Medical and scientific images

Slide credit; L. Lazebnik

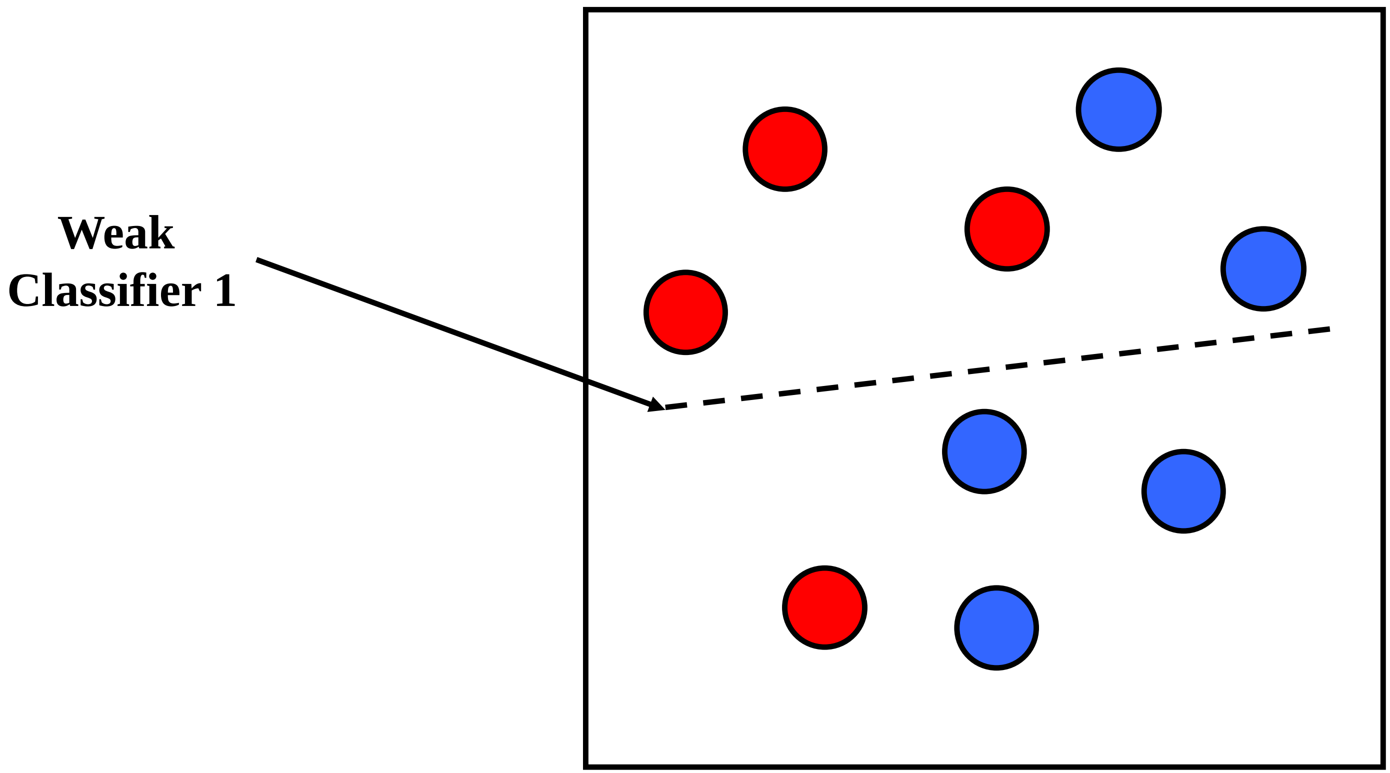
Generic category recognition: basic framework

- Build/train object model
  - Choose a representation
  - Learn or fit parameters of model / classifier
- Generate candidates in new image
- Score the candidates

Not all recognition tasks are suited to features + supervised classification…but what makes a class a good candidate?

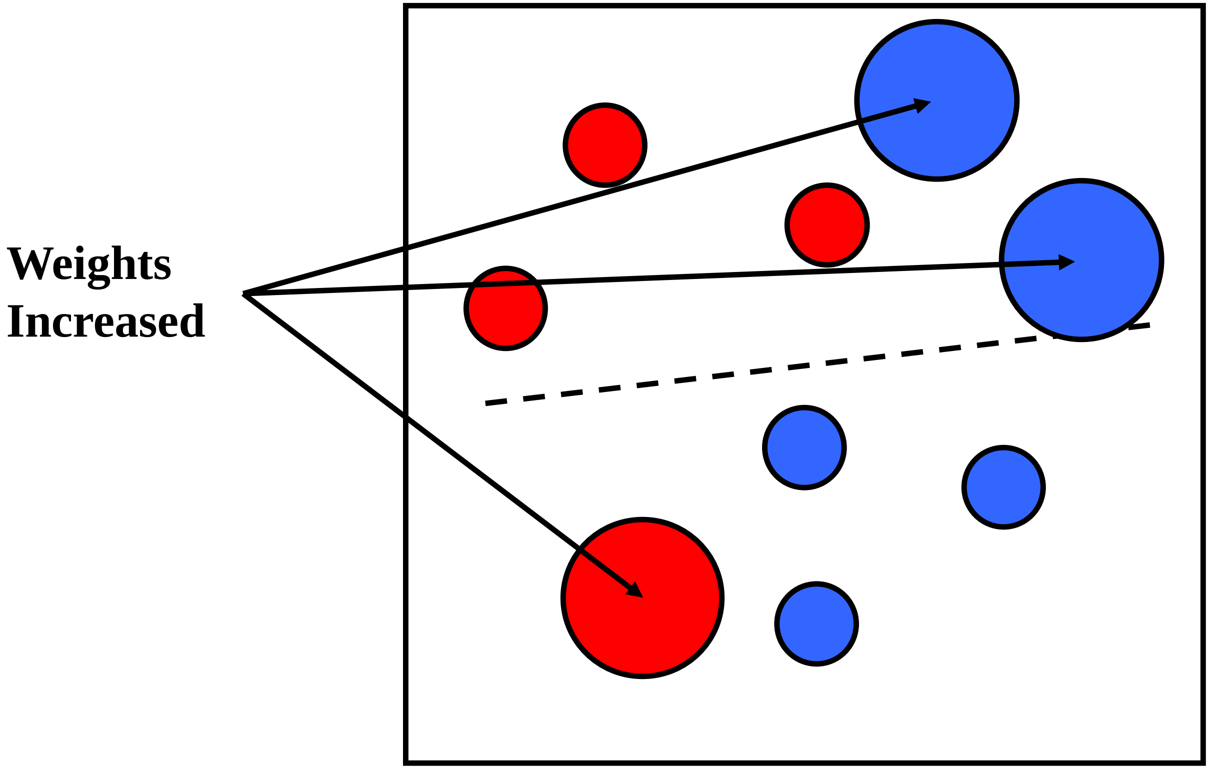
Slide credit: Kristen Grauman

Boosting intuition

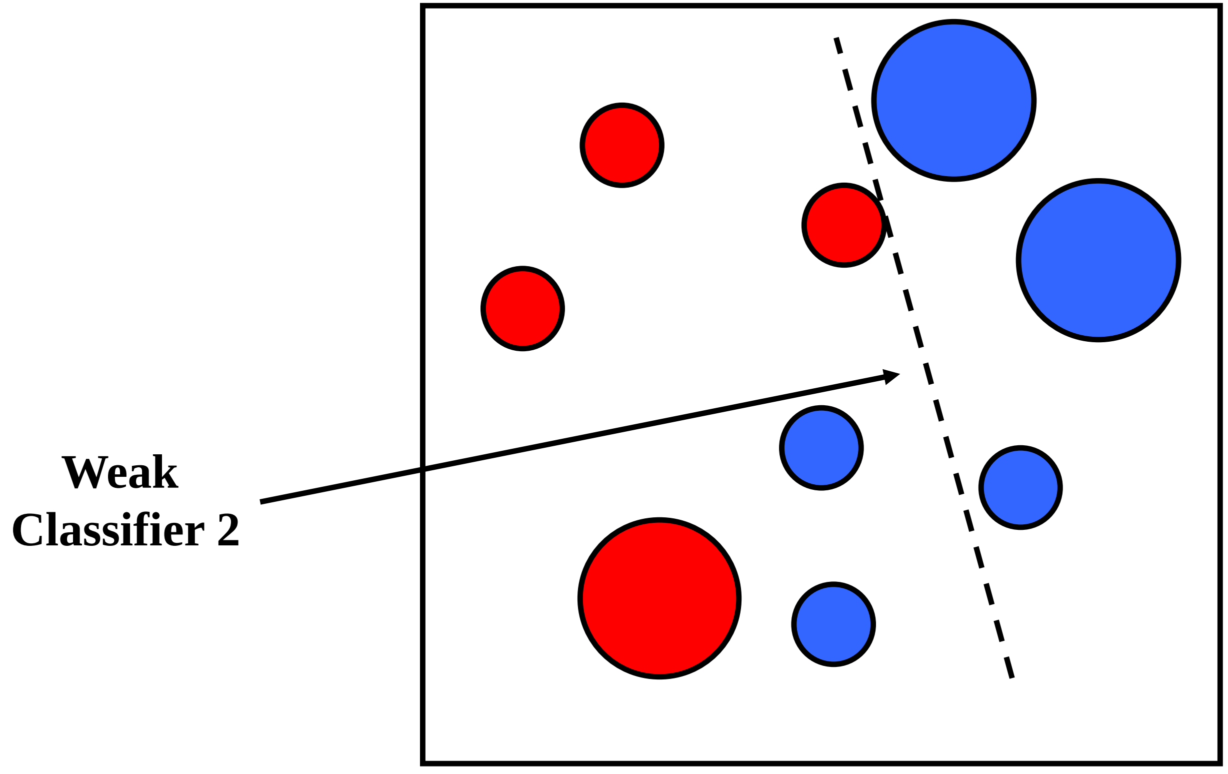


Slide credit: Paul Viola

Boosting illustration

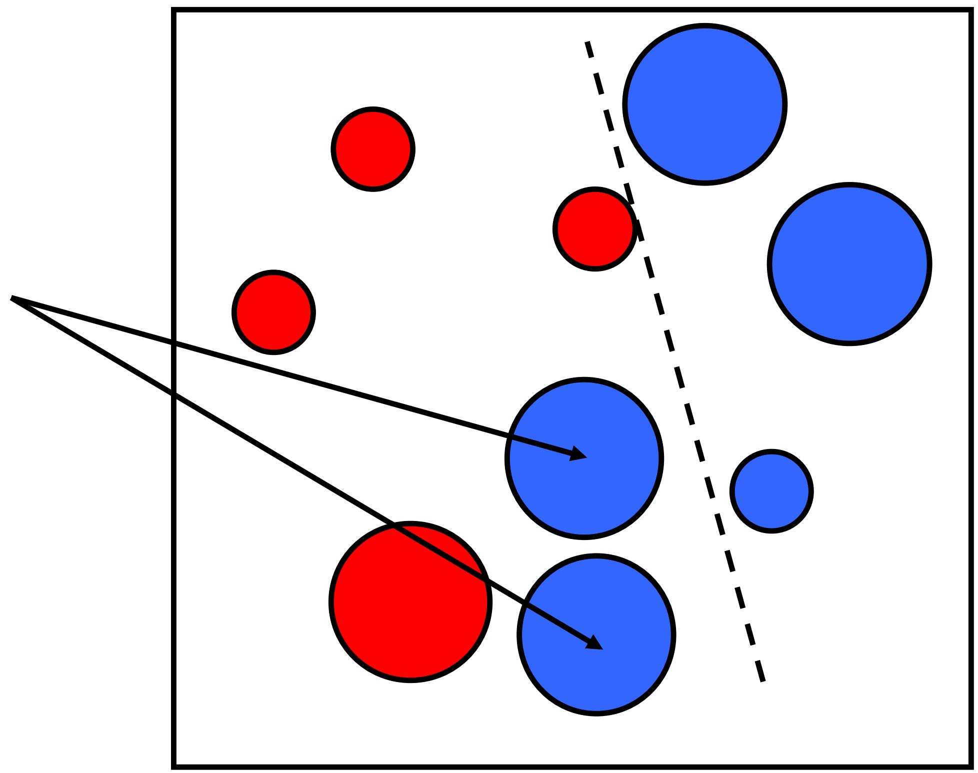


Boosting illustration

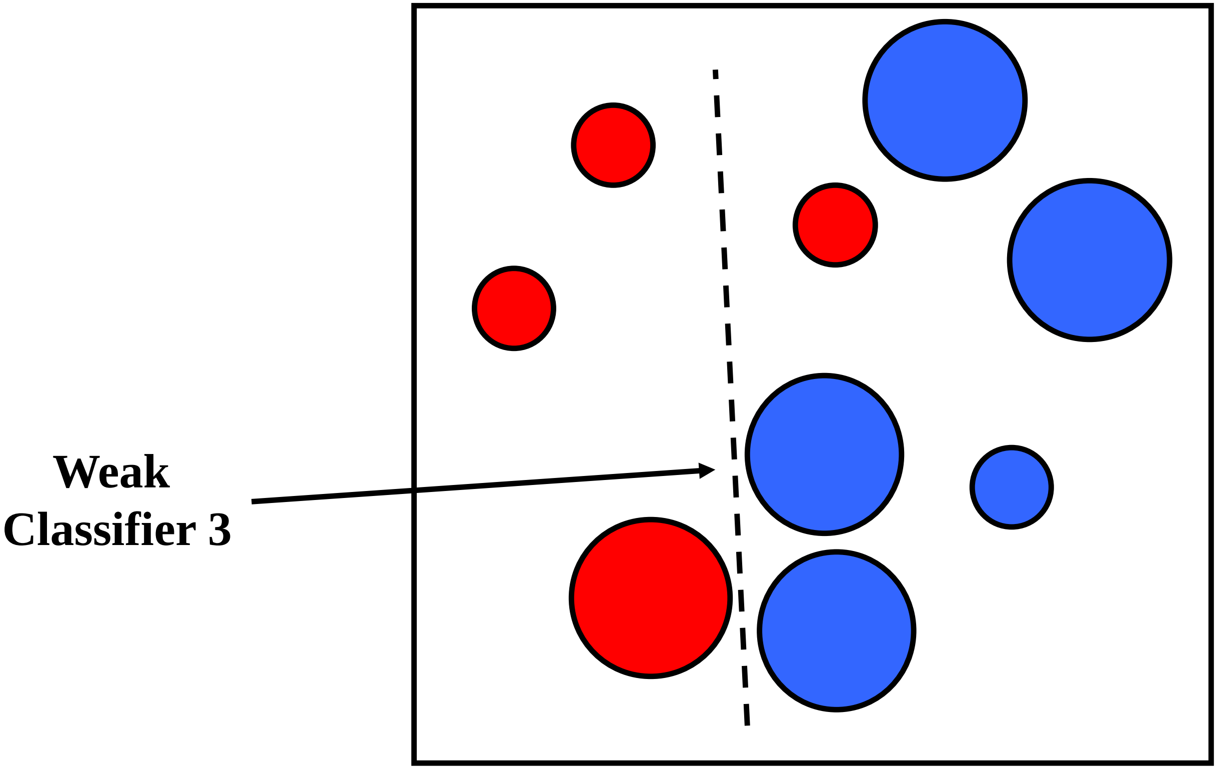


Boosting illustration

Weights Increased

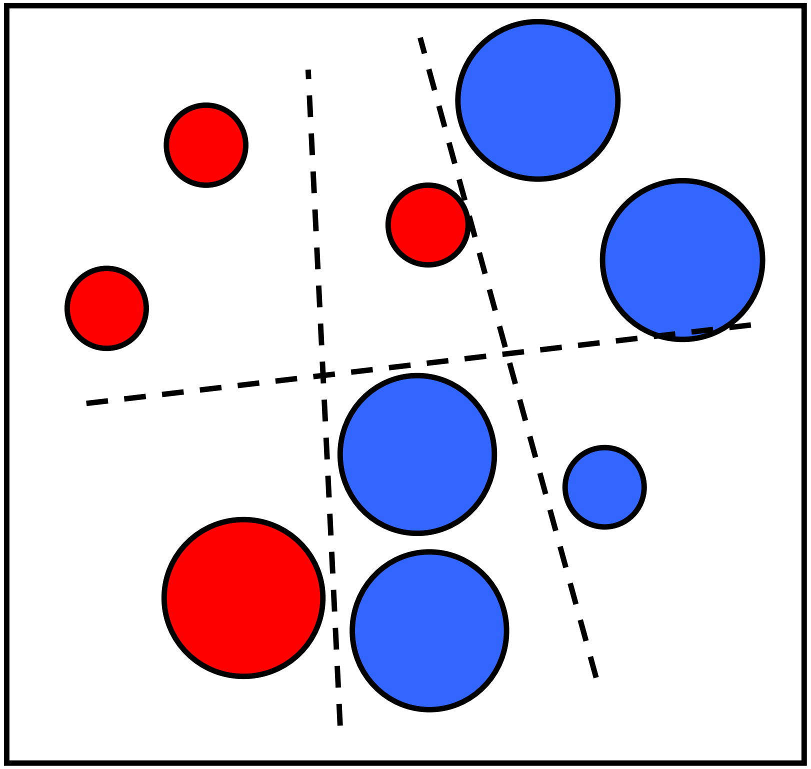


Boosting illustration



Boosting illustration

Final classifier is a combination of weak classifiers



Boosting: training

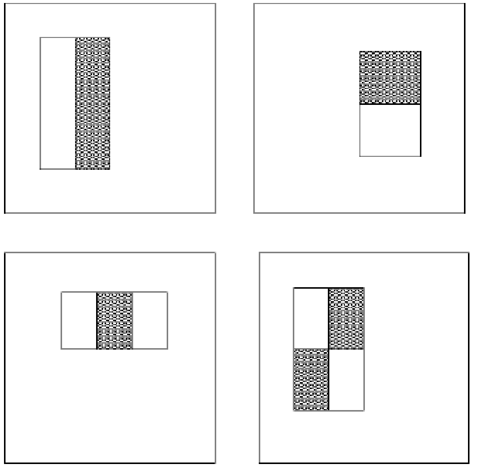
- Initially, weight each training example equally
- In each boosting round:
  - Find the weak learner that achieves the lowest *weighted* training error $\sum_i w_i \left| h_j(x_i) - y_i \right|$.
  - Raise weights of training examples misclassified by current weak learner
- Compute final classifier as linear combination of all weak learners (weight of each learner is directly proportional to its accuracy)

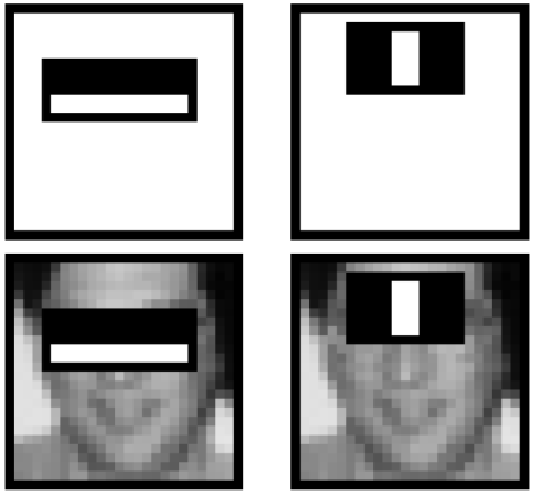
Viola-Jones face detector

Main idea:

- Represent local texture with efficiently computable “rectangular” features within window of interest
- Select discriminative features to be weak classifiers
- Use boosted combination of them as final classifier
- Form a cascade of such classifiers, rejecting clear negatives quickly

Viola-Jones detector: features

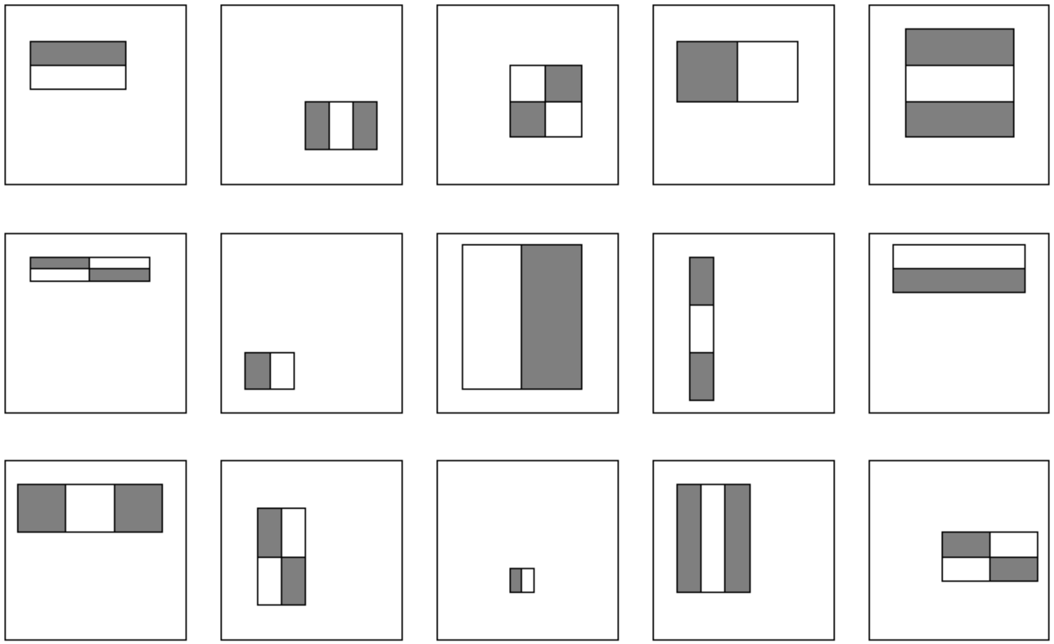




"Rectangular" filters

Feature output is difference between adjacent regions

Viola-Jones detector: features



Considering all possible filter parameters: position, scale, and type:

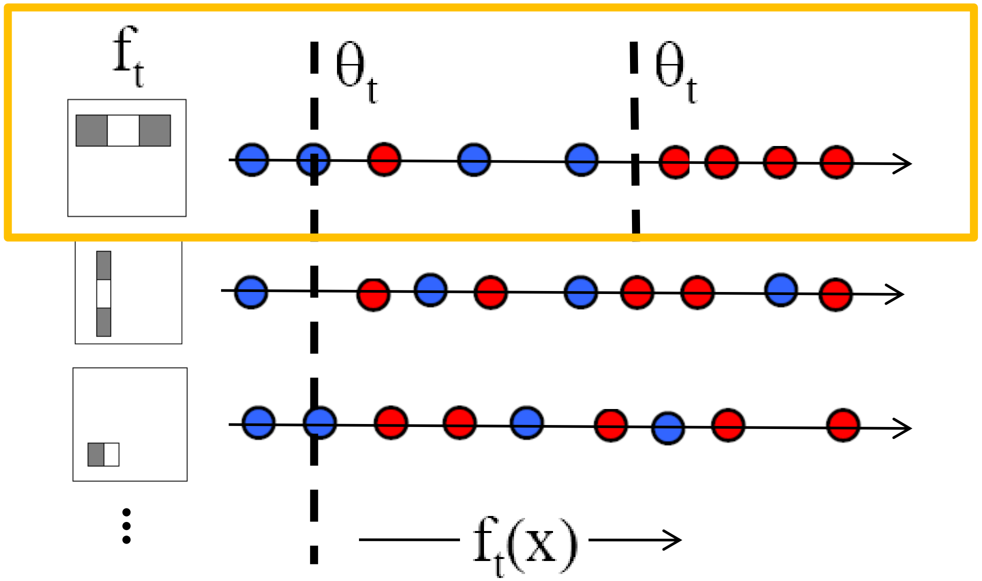
180,000+ possible features associated with each 24 x 24 window

Which subset of these features should we use to determine if a window has a face?

Use boosting both to select the informative features and to form the classifier

Viola-Jones detector: AdaBoost

- **Want to select the single rectangle feature and threshold that best separates positive (faces) and negative (non-faces) training examples, in terms of *weighted* error.**



Outputs of a possible rectangle feature on faces and non-faces.

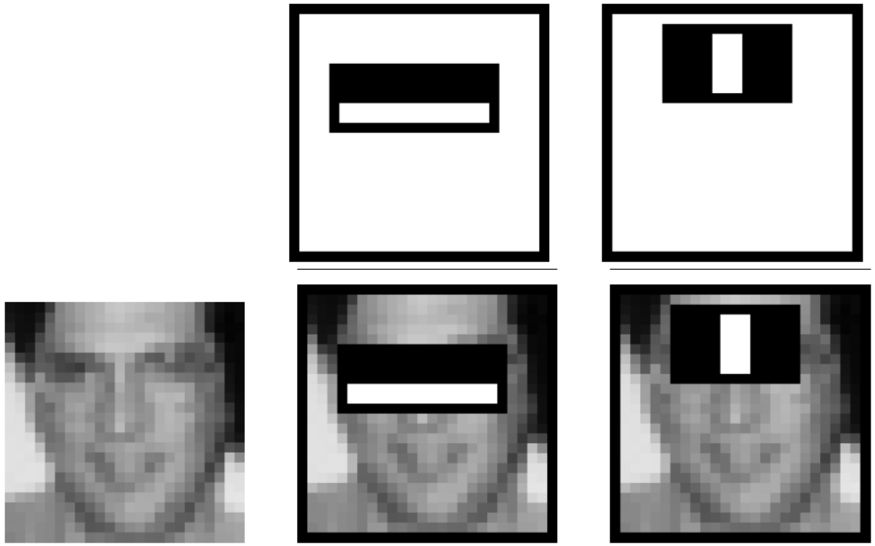
Resulting weak classifier:

$$h_t(x) = \begin{cases} +1 & \text{if } f_t(x) > \theta_t \\ -1 & \text{otherwise} \end{cases}$$

For next round, reweight the examples according to errors, choose another filter/threshold combo.

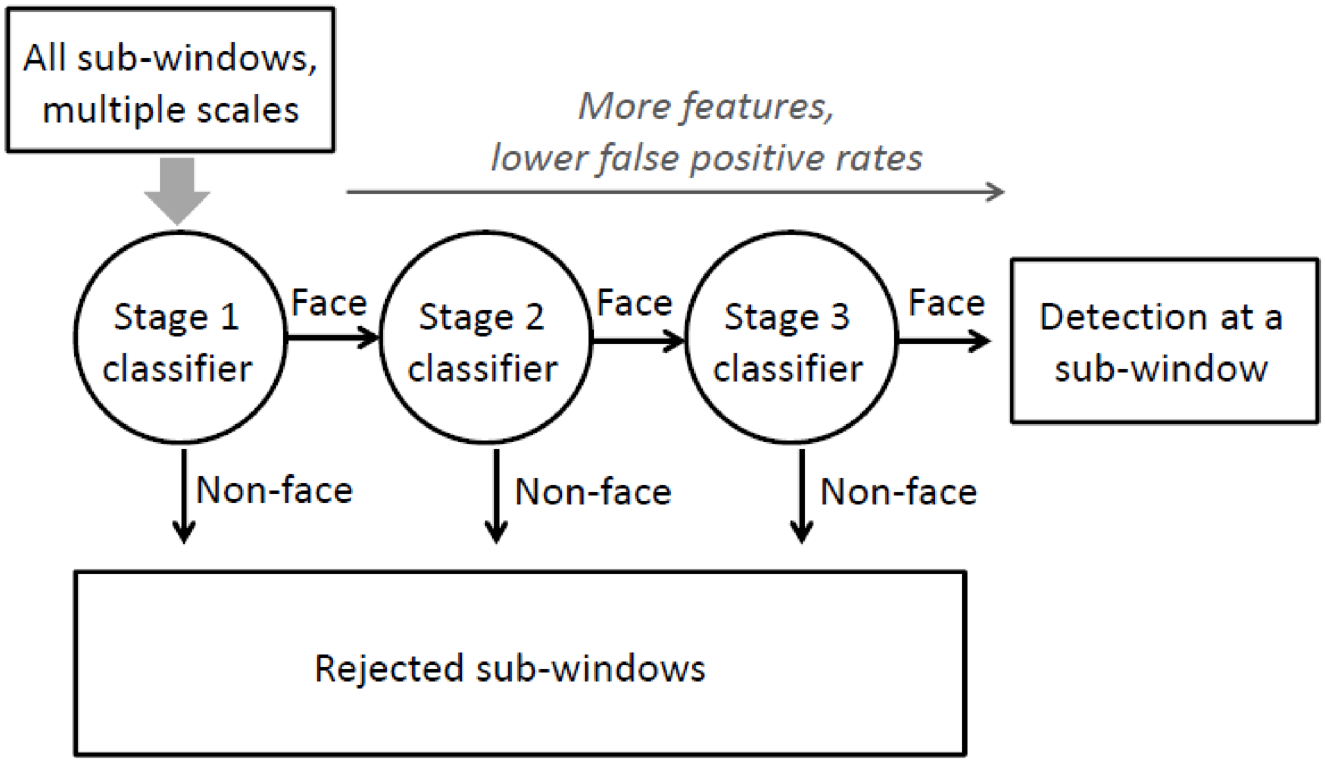
Slide credit: Kristen Grauman

Viola-Jones Face Detector: Results



First two features selected

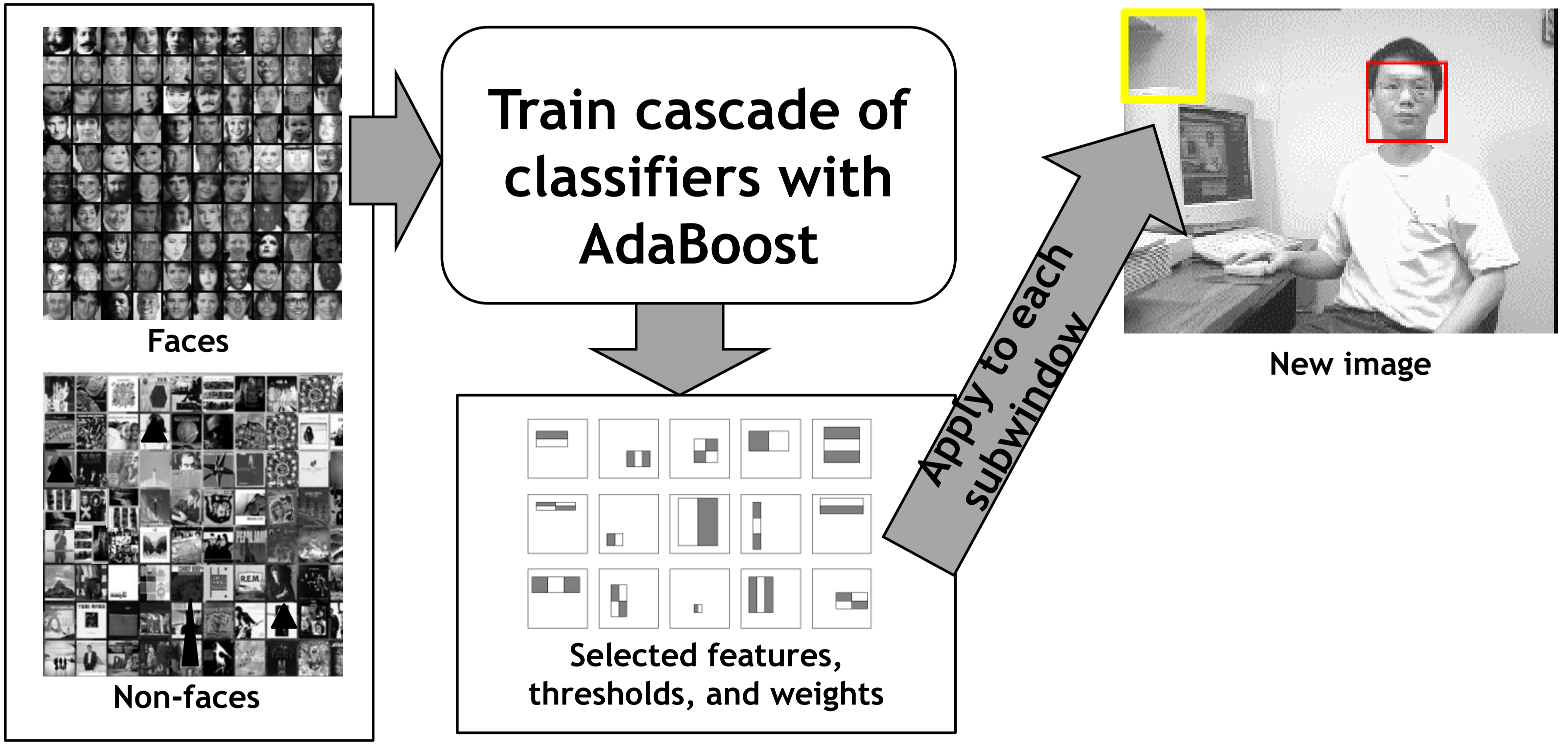
Cascading classifiers for detection



- Form a *cascade* with low false negative rates early on
- Apply less accurate but faster classifiers first to immediately discard windows that clearly appear to be negative

Slide credit: Kristen Grauman

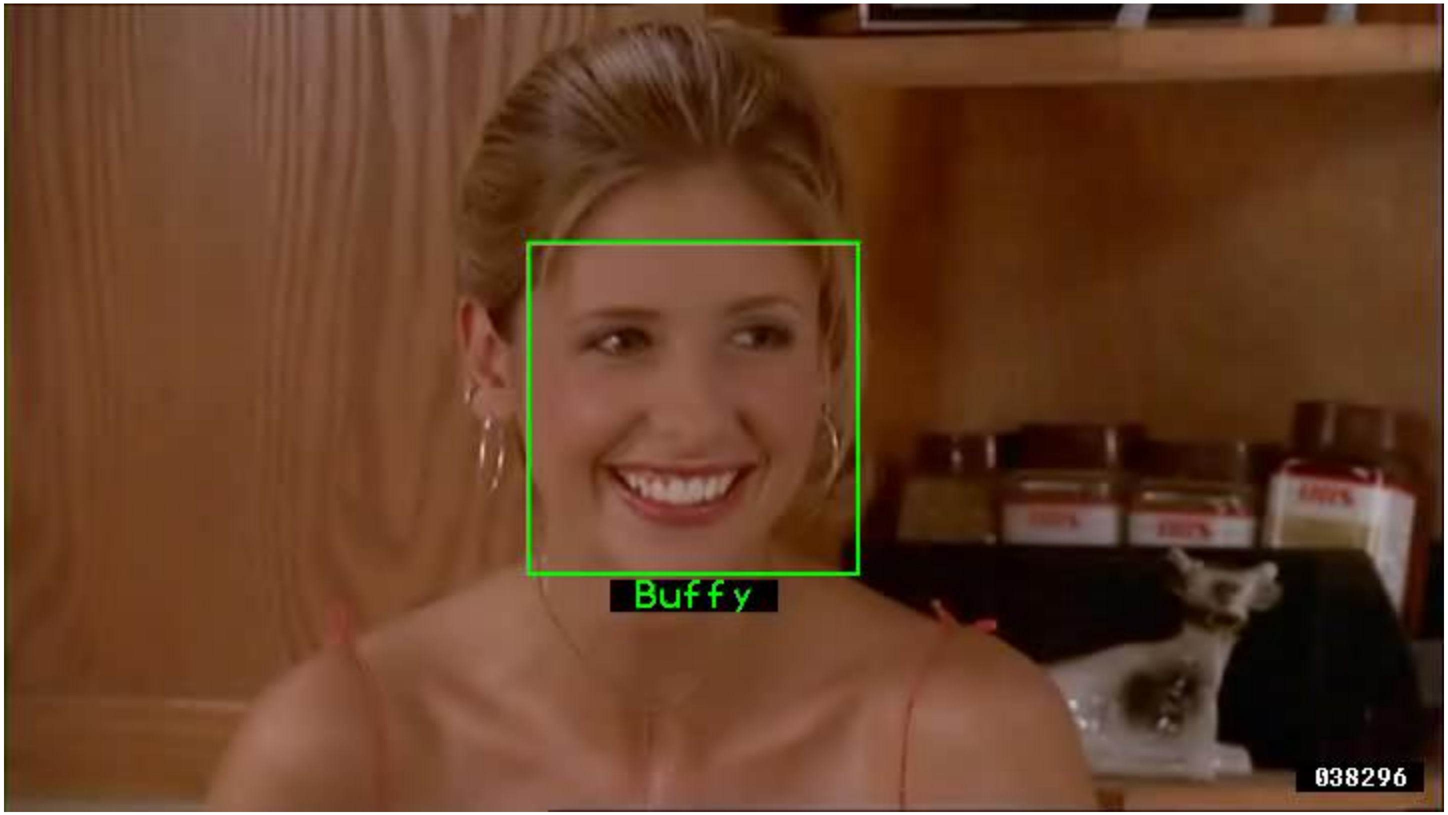
Viola-Jones detector: summary



Train with 5K positives, 350M negatives
Real-time detector using 38 layer cascade
6061 features in all layers

[Implementation available in OpenCV: http://www.intel.com/technology/computing/opencv/]

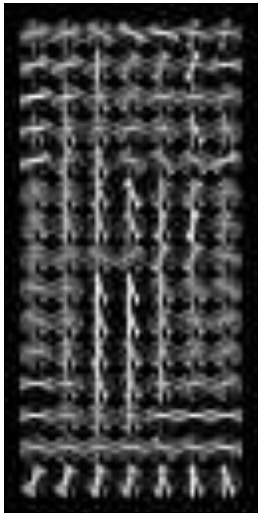
Example using Viola-Jones detector



Frontal faces detected and then tracked, character names inferred with alignment of script and subtitles.

Everingham, M., Sivic, J. and Zisserman, A.
"Hello! My name is... Buffy" - Automatic naming of characters in TV video, BMVC 2006. http://www.robots.ox.ac.uk/~vgg/research/nface/index.html

Person detection with HoG's & linear SVM's

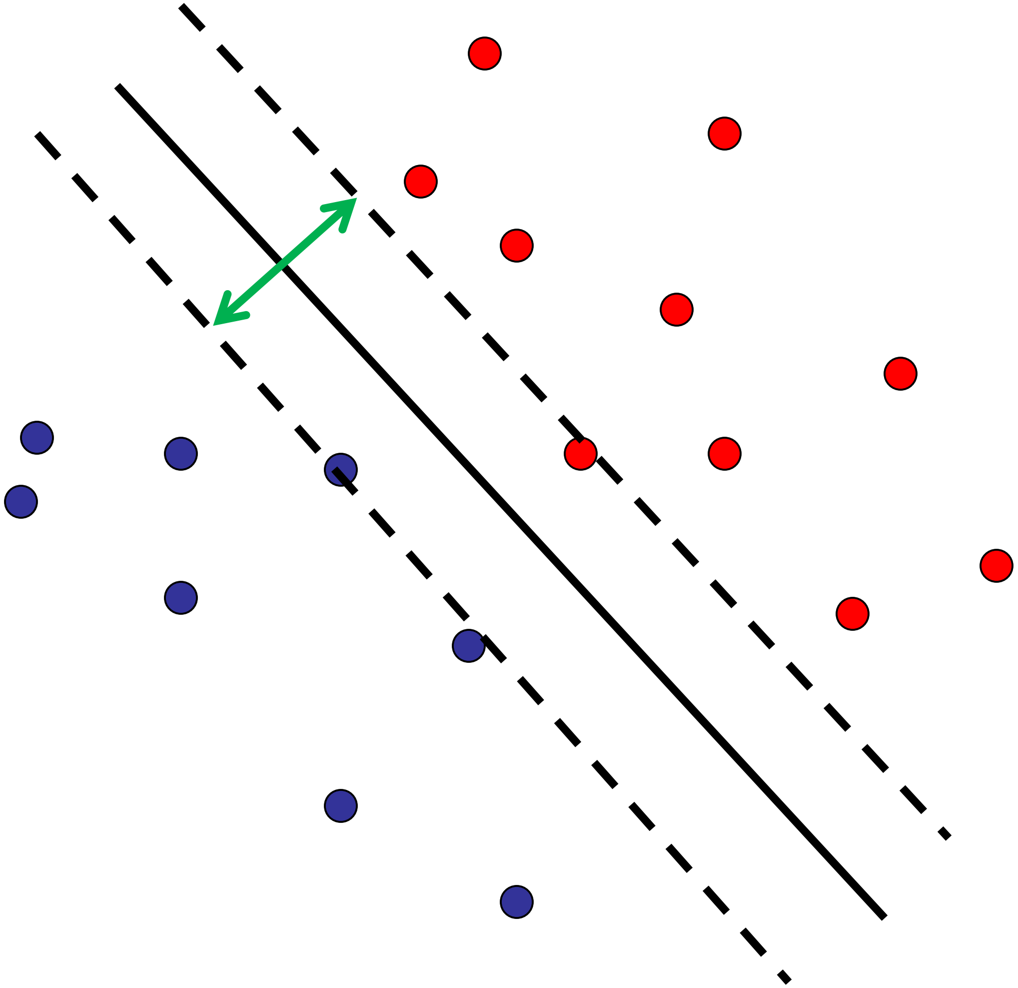


- Map each grid cell in the input window to a histogram counting the gradients per orientation.
- Train a linear SVM using training set of pedestrian vs. non-pedestrian windows.

Code available:
http://pascal.inrialpes.fr/soft/olt/

Dalal & Triggs, CVPR 2005

Support Vector Machines (SVMs)



- Discriminative classifier based on *optimal separating line (for 2d case)*
- Maximize the *margin* between the positive and negative training examples

Person detection with HoG's & linear SVM's

- Histograms of Oriented Gradients for Human Detection, Navneet Dalal, Bill Triggs, International Conference on Computer Vision & Pattern Recognition - June 2005
- http://lear.inrialpes.fr/pubs/2005/DT05/

Multi-class SVMs

- SVM is a binary classifier. What if we have multiple classes?
- **One vs. all**
  - Training: learn an SVM for each class vs. the rest
  - Testing: apply each SVM to test example and assign to it the class of the SVM that returns the highest decision value
- **One vs. one**
  - Training: learn an SVM for each pair of classes
  - Testing: each learned SVM "votes" for a class to assign to the test example

Real-Time Human Pose Recognition in Parts from Single Depth Images

Jamie Shotton, Andrew Fitzgibbon, Mat Cook,
Toby Sharp, Mark Finocchio, Richard Moore,
Alex Kipman, Andrew Blake

CVPR 2011

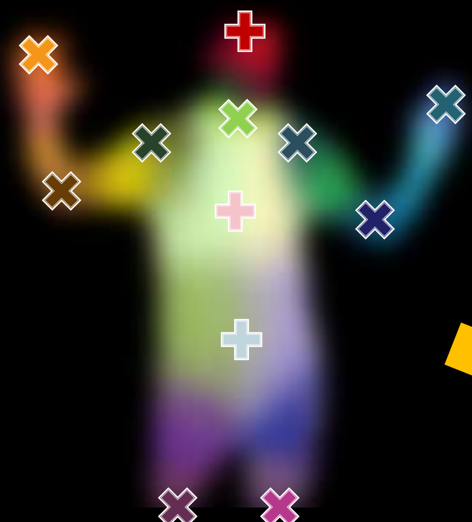
Microsoft Research

XBOX 360

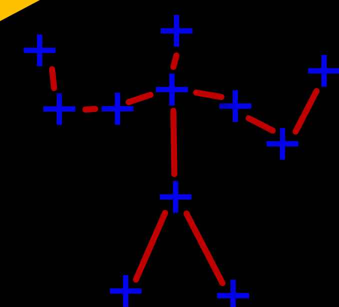
The Kinect pose estimation pipeline

capture
depth image &
remove bg

infer
body parts
per pixel



cluster pixels to
hypothesize
body joint
positions

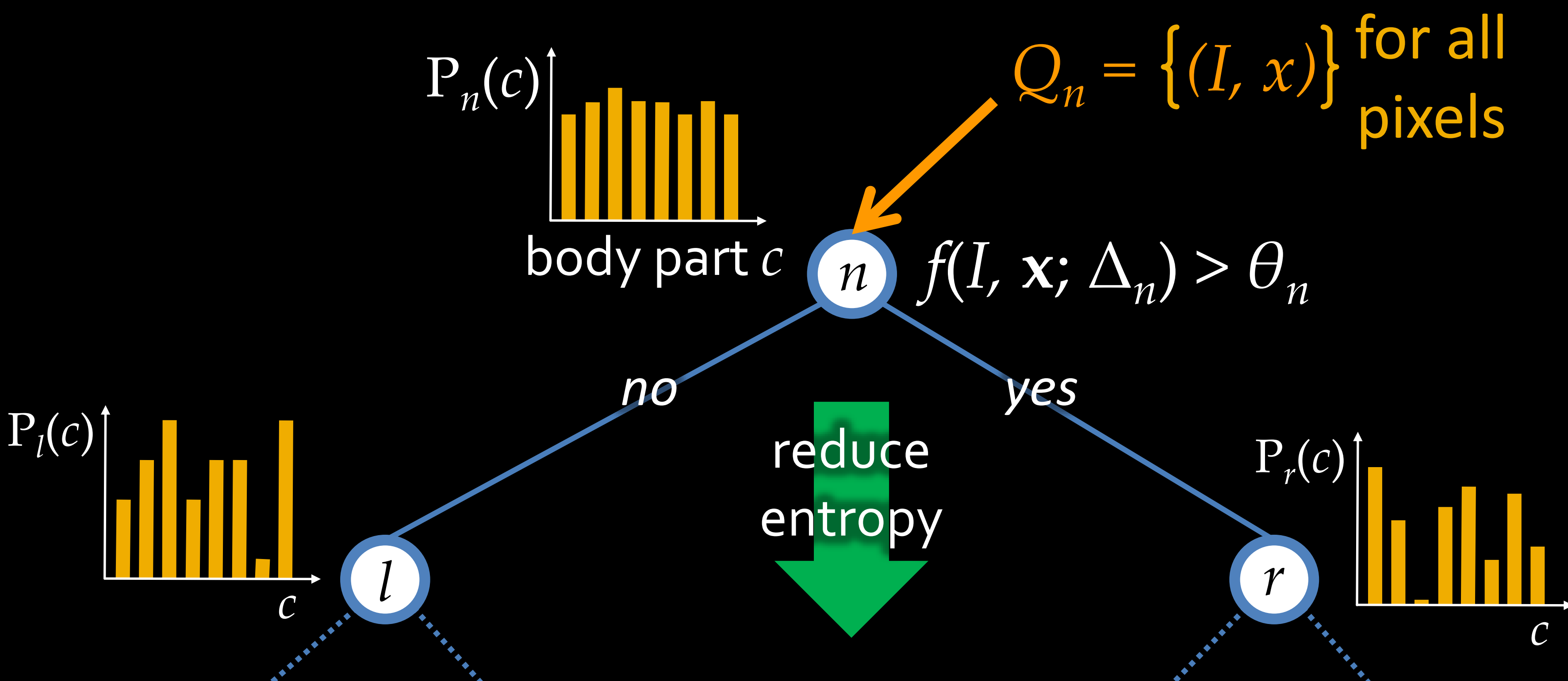


fit model &
track skeleton

Slide credit: Jamie Shotton

Training decision trees

[Breiman *et al.* 84]



Take (Δ, θ) that maximises information gain:

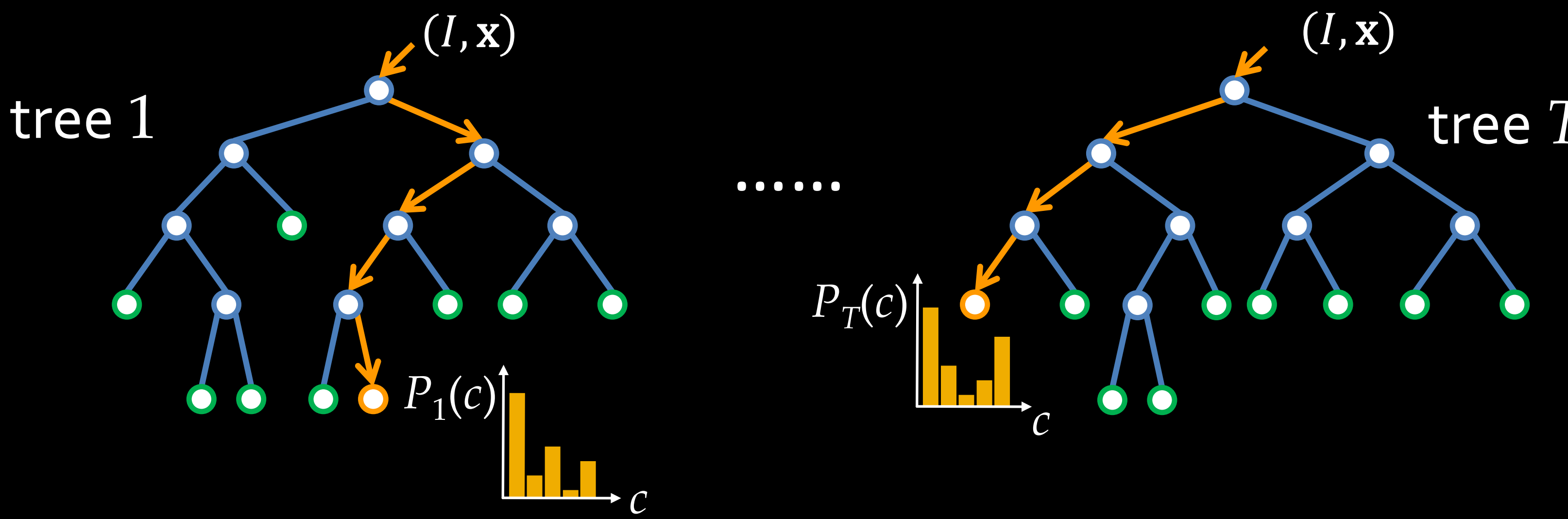
$$\Delta E = -\frac{|Q_{\mathrm{l}}|}{|Q_n|} E(\mathrm{Q_l}) - \frac{|Q_{\mathrm{r}}|}{|Q_n|} E(\mathrm{Q_r})$$

Goal: drive entropy at leaf nodes to zero

Slide credit: Jamie Shotton

Decision forest classifier

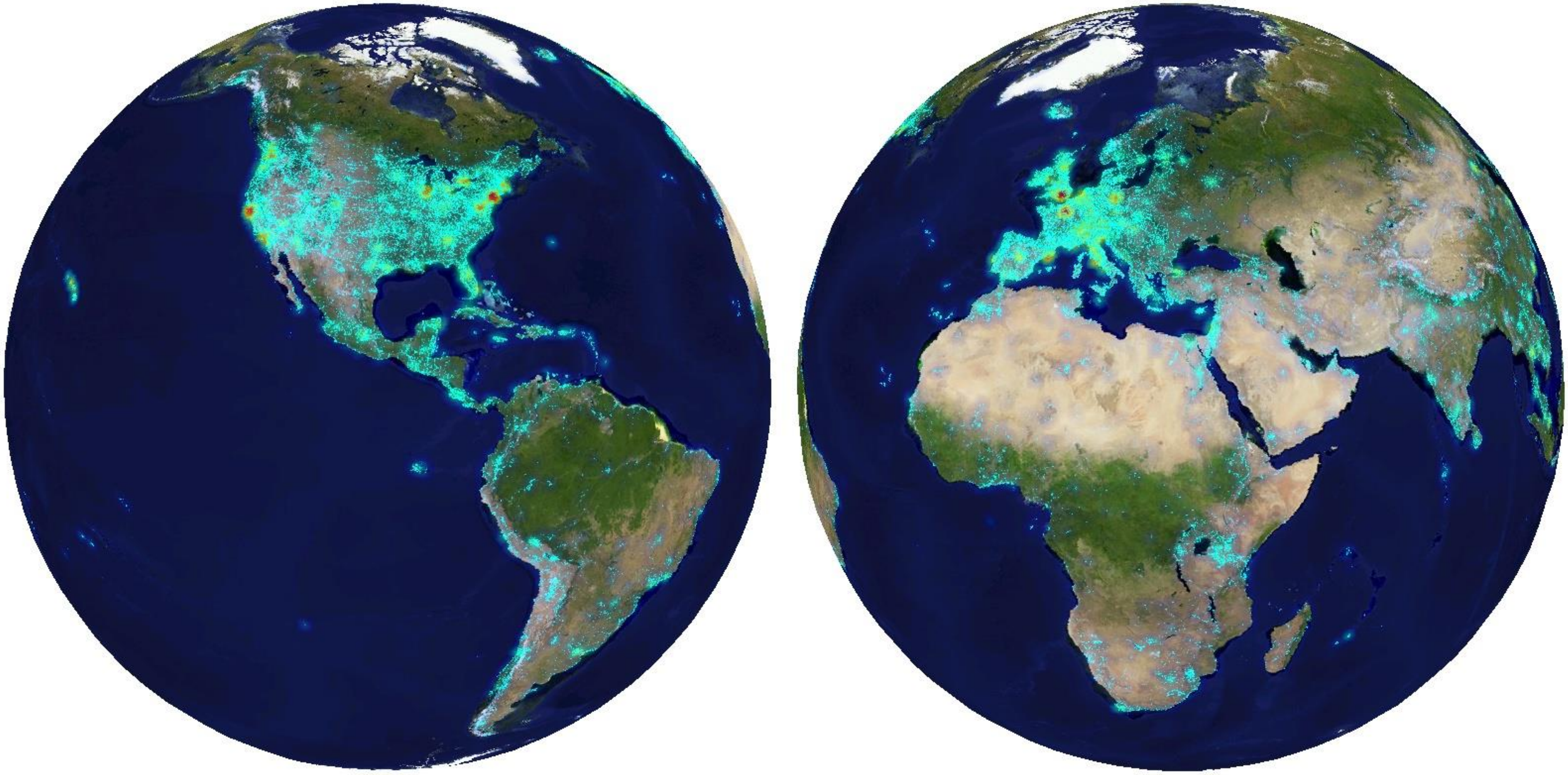
[Amit & Geman 97]
[Breiman 01]
[Geurts *et al.* 06]



- Trained on different random subset of images
  - "bagging" helps avoid over-fitting
- Average tree posteriors $P(c|I,\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,\mathbf{x})$

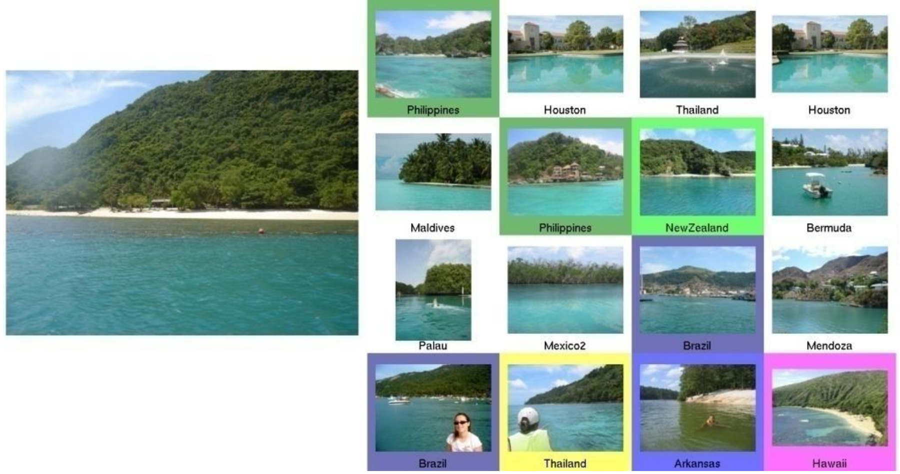
Slide credit: Jamie Shotton

6+ million geotagged photos by 109,788 photographers



Annotated by Flickr users

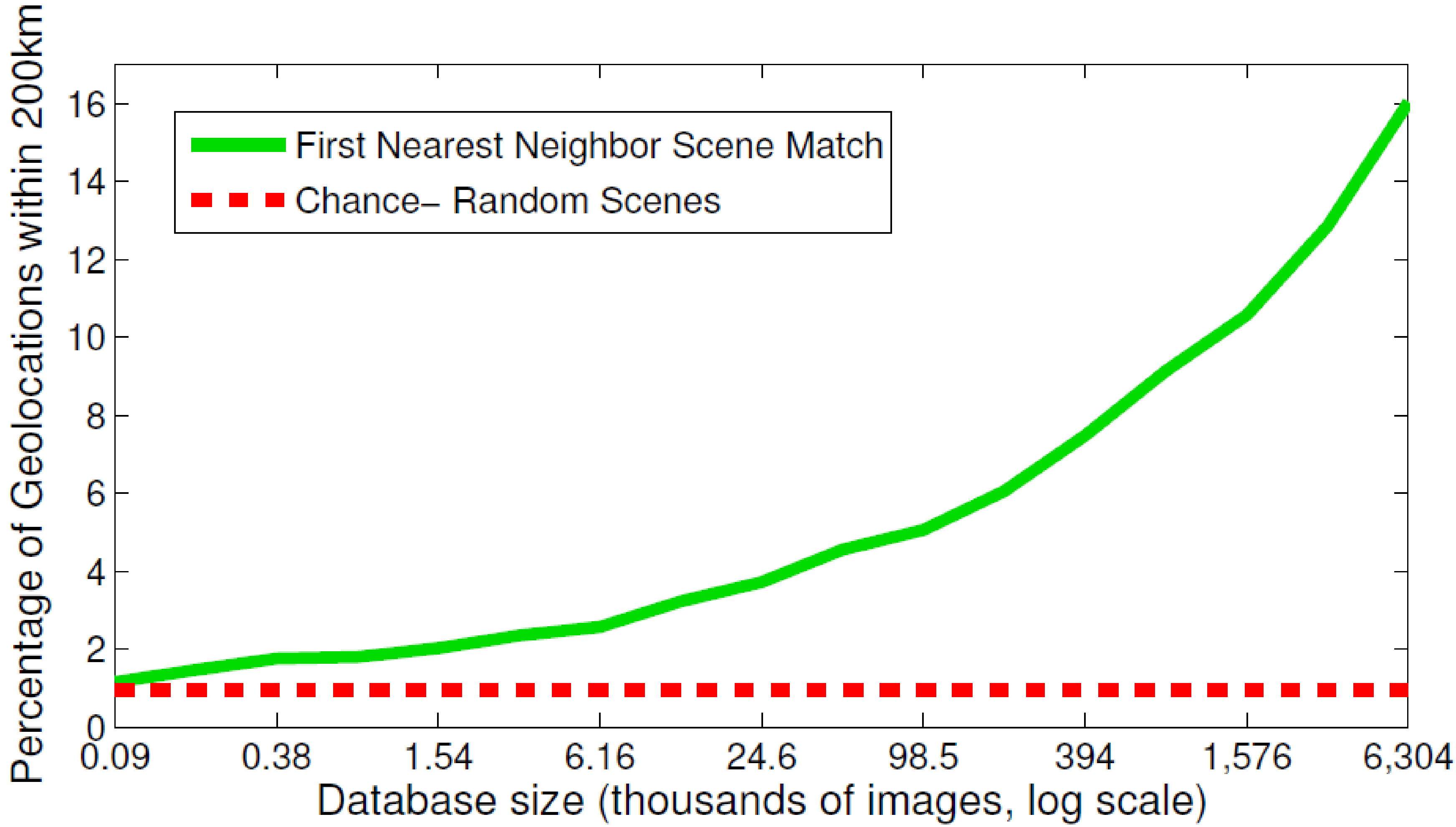
Slide credit: James Hays



Slide credit: James Hays

[Hays and Efros. **im2gps**: Estimating Geographic Information from a Single Image. CVPR 2008.]

The Importance of Data



[Hays and Efros. **im2gps**: Estimating Geographic Information from a Single Image. CVPR 2008.]

Slide credit: James Hays

Summary

- Neural networks
- Boosting
- Decision forests
- Classifier cascades
- Binary classifiers → multi-class

- Visual recognition tasks with supervised classification
  - Variety of features and models
  - Training data quality and/or quantity essential