CMT10[illegible]ual Computing

Object Recognition

Jing Wu

School of Computer Science and Informatics
Cardiff University

Overview

- Object Recognition
  - Overview
  - History
  - What “Works” Today
- Machine Learning Approach for Recognition
  - The Machine Learning Framework
  - Classifiers
    - Nearest neighbour
    - Linear
  - Recognition Task and Supervision
  - Generalization
  - Datasets
- Face Detection and Recognition
  - The Viola/Jones Face Detector
  - Face Recognition

Object Recognition

- Object recognition is the task of finding a given object in an image or a video.
- The object recognition problem can be defined as a labelling problem based on models of known objects.
- Object recognition approaches:
  - Geometric model-based methods
  - Appearance-based methods
  - Feature-based methods

How many Visual Object Categories are there?

Slides adapted from Fei-Fei Li, Rob Fergus, Antonio Torralba, and Jean Ponce

Biederman 1987

How many Visual Object Categories are there?

Visual Object Categories

Specific Recognition Tasks

Scene Categorisation

- outdoor/indoor
  /forest/factory/etc.

Image Annotation/Tagging

- street
- people
- building
- mountain
- …

Object Detection

- **find pedestrians**

Image Parsing

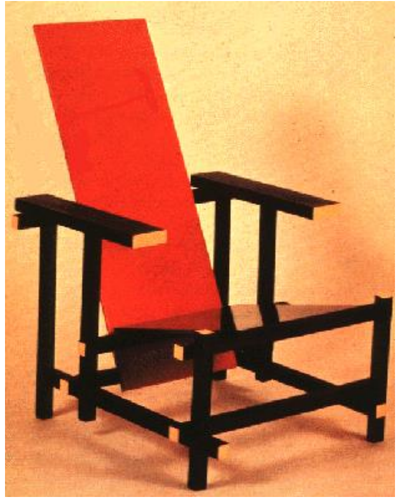
Image Understanding?

Recognition is All about Modelling Variability

- Variability
  - Camera position
  - Illumination
  - Shape parameters

Within-class variations?

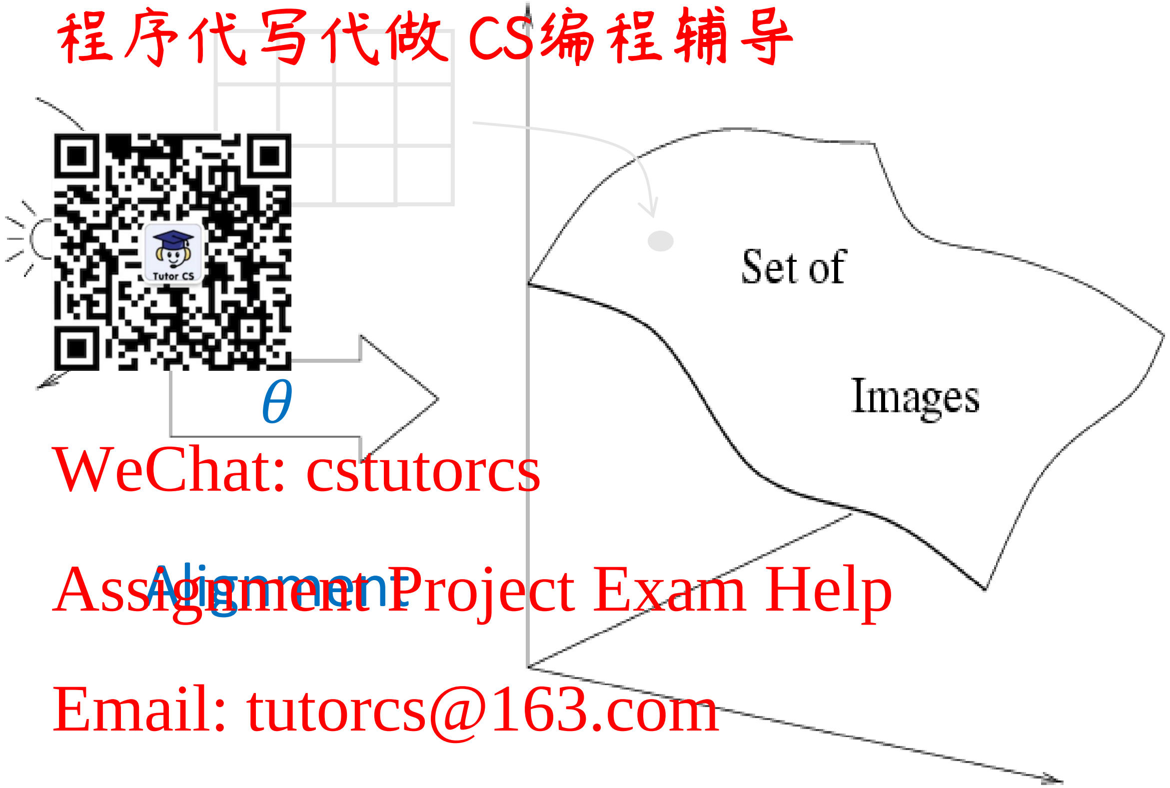
Within-class Variations



History of Ideas in Recognition

- 1960s – early 1990s: the geometric era

Recognition is All about Modelling Variability

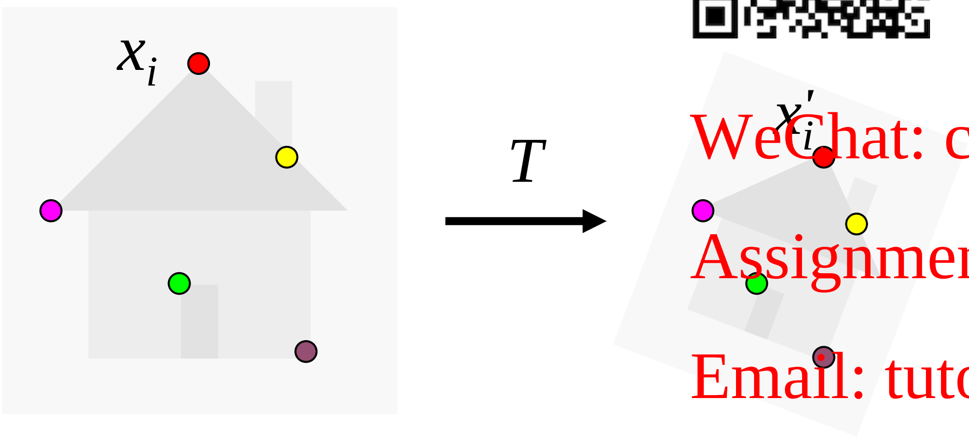


- Variability
  - Camera position
  - Illumination
  - Shape parameters : assumed known

Roberts (1965); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

Alignment

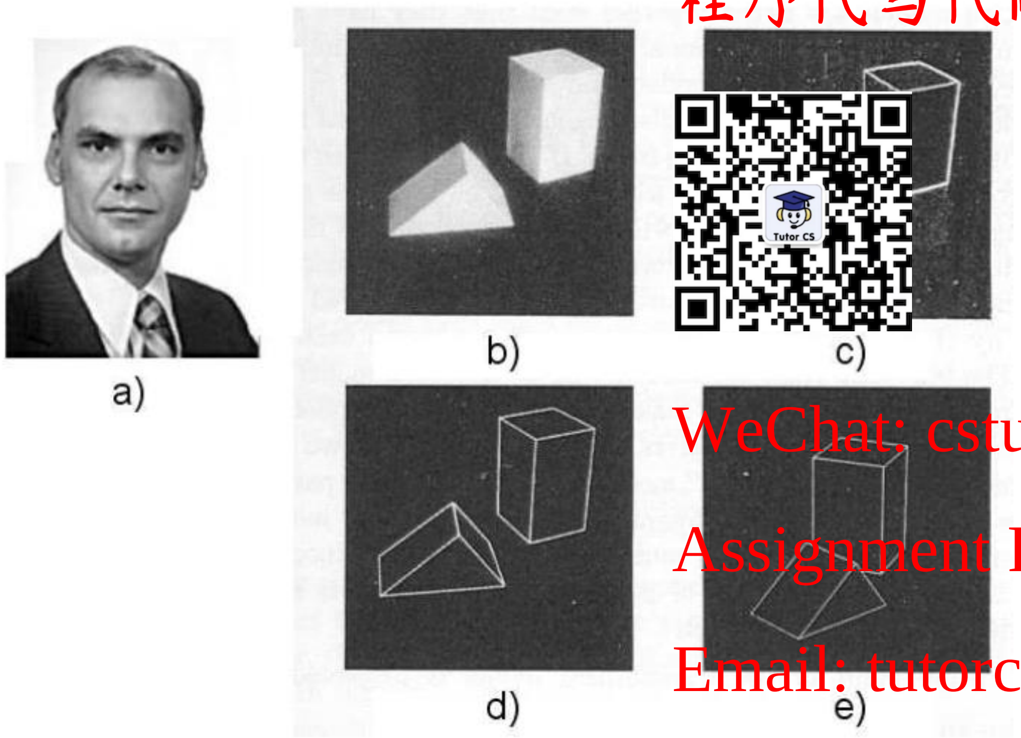
- Alignment: fitting a model to a transformation between pairs of features (matches) in two images



Find transformation T that minimizes

$$\sum_i residual(T(x_i), x_i')$$

Recognition as an Alignment Problem



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

Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - e) are taken from [64] with permission MIT Press.)

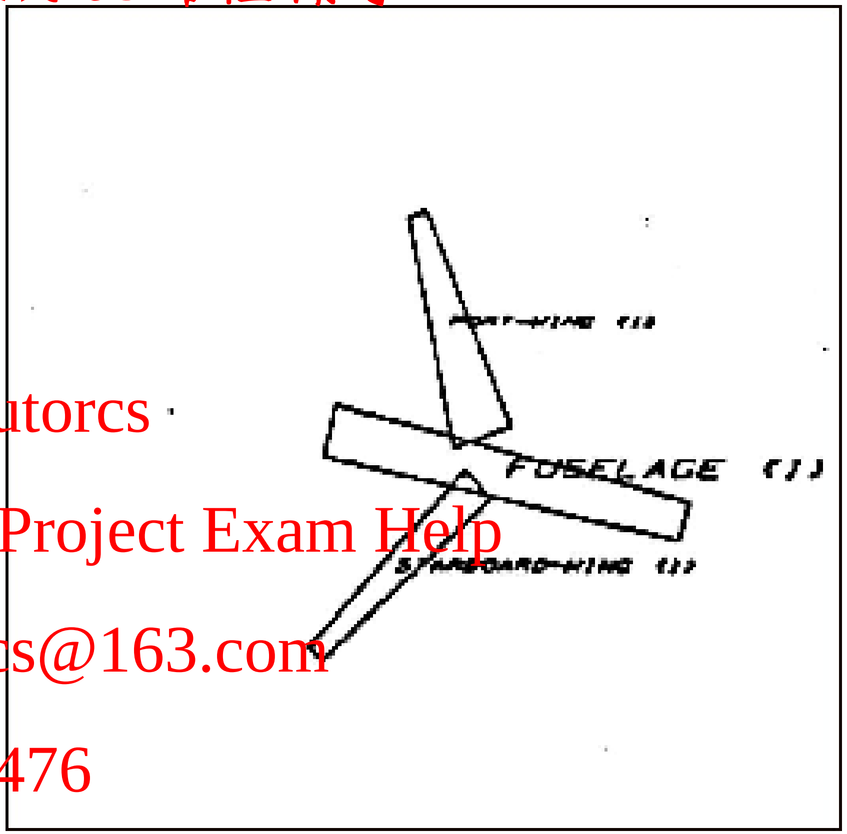
J. Mundy, Object Recognition in the Geometric Era: a Retrospective, 2006

Recognition is All about Modelling Variability

- ~~Variability~~ Invariant to
  - Camera position
  - Illumination
  - Internal parameters

Duda & Hart (1972);
Weiss (1987);
Mundy et al. (1992-94);
Rothwell et al. (1992);
Burns et al. (1993)

Representing and Recognising Object Categories is Harder…



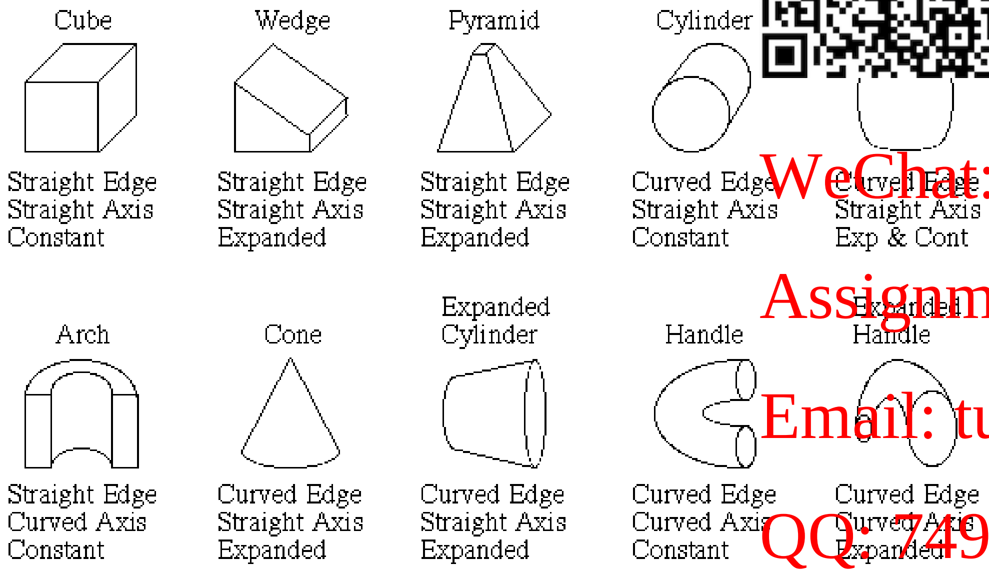
ACRONYM (Brooks and Binford, 1981)

Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

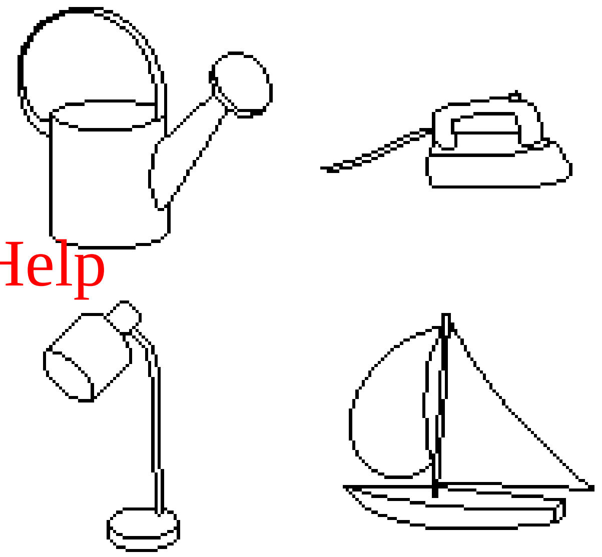
Recognition by Components

Biederman (1987)

Primitives (geons)

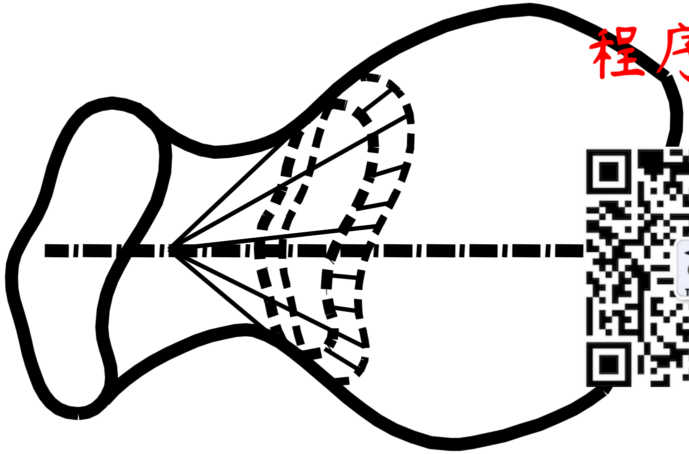


Objects

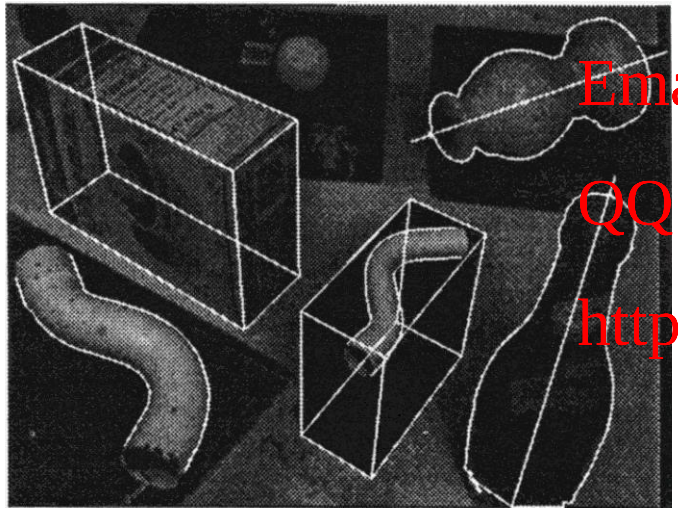


http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

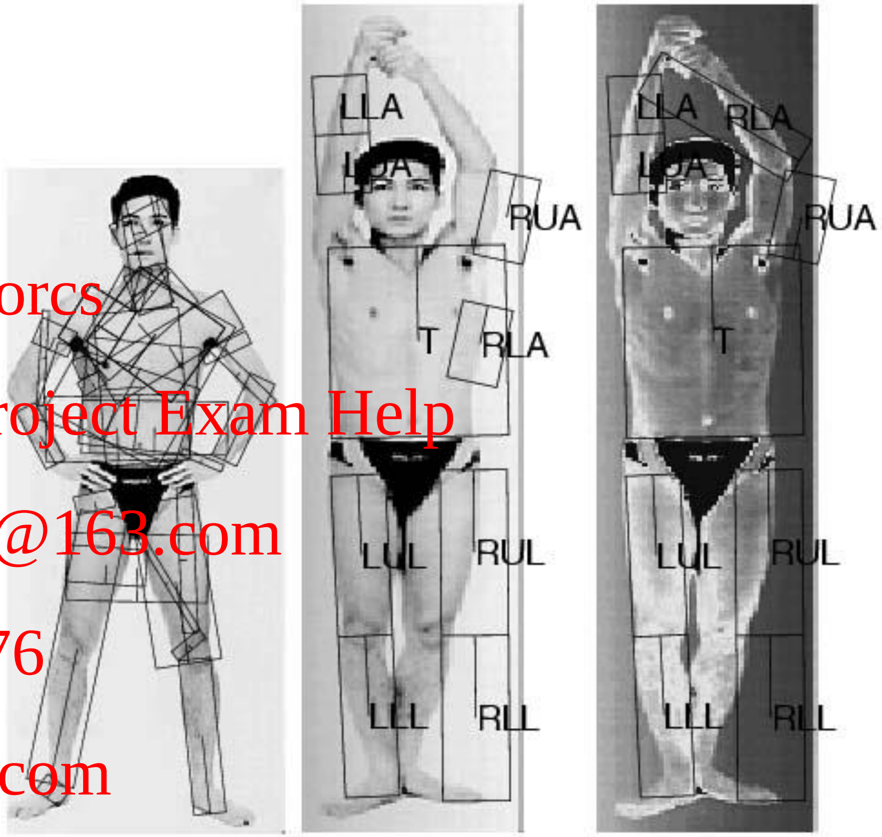
General Shape Primitives



Generalized cylinders
Ponce et al. (1989)



Zisserman et al. (1995)

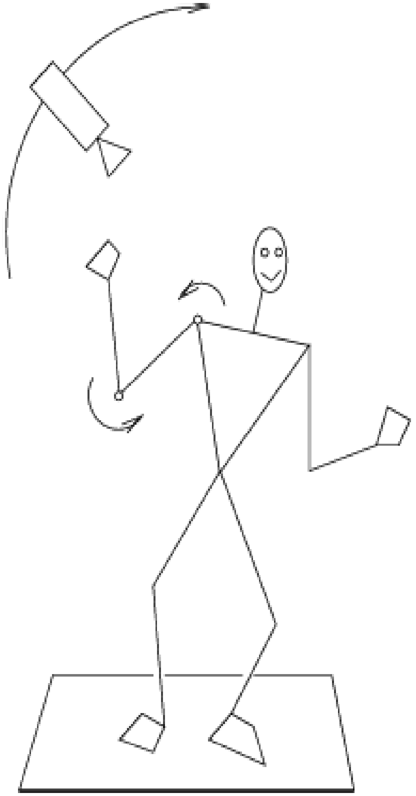


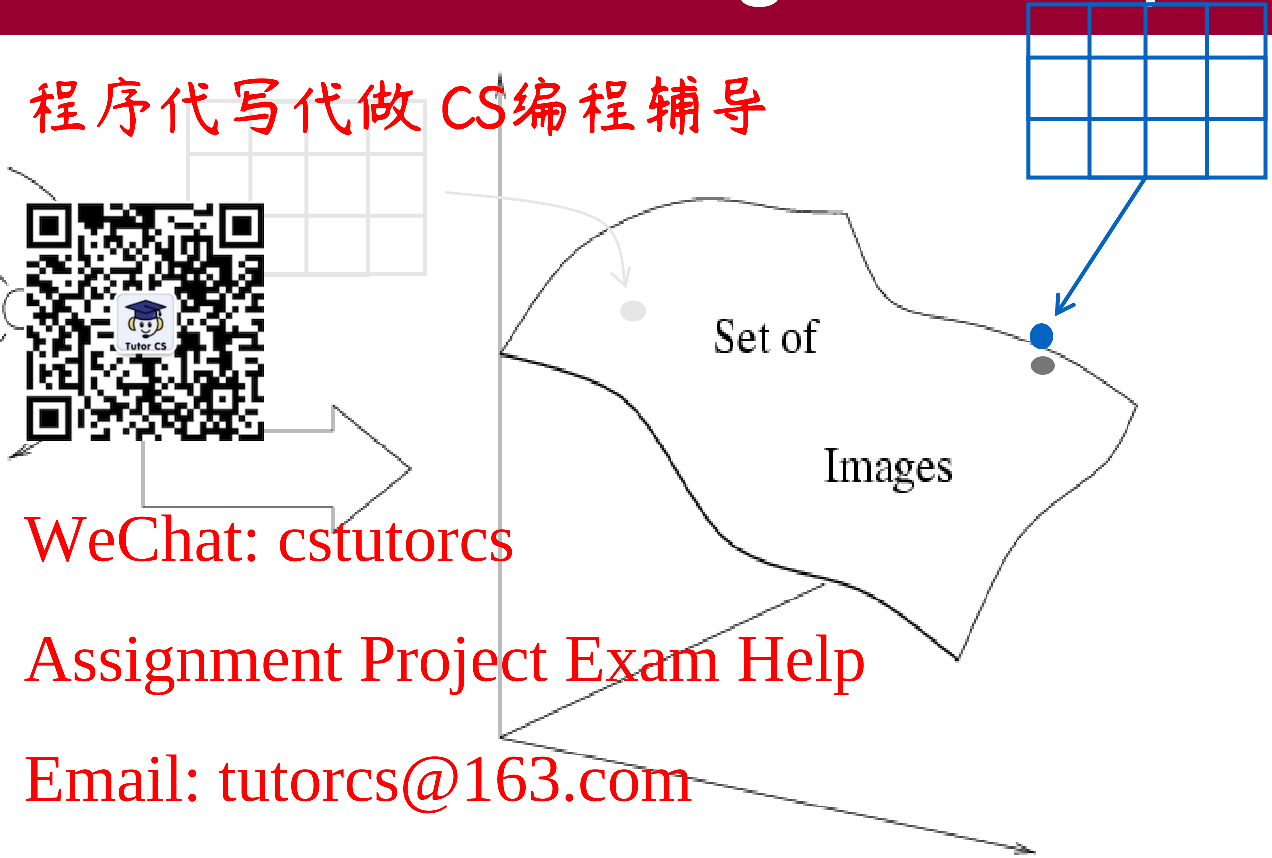
Forsyth (2000)

History of Ideas in Recognition

- 1960s – early 1990s: the geometric era
- 1990s – appearance-based [obscured]

Recognition is All about Modelling Variability

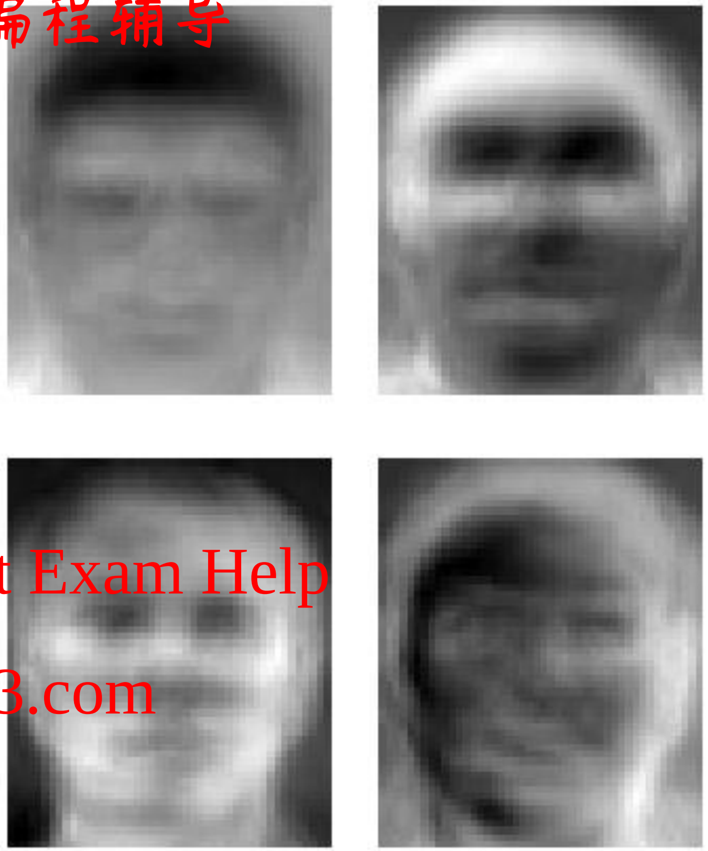




- Empirical models of image variability
- Appearance-based models:

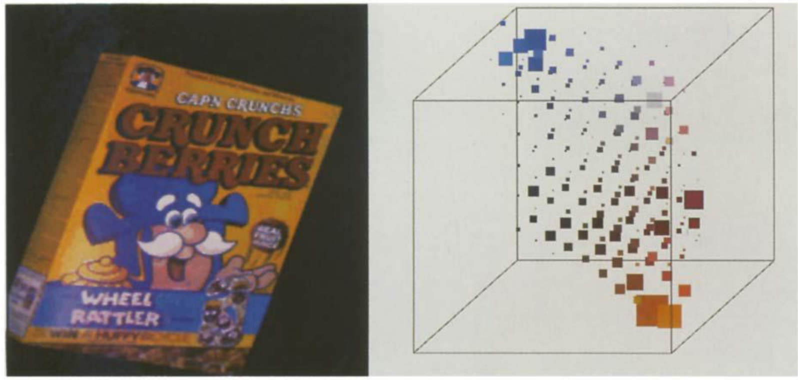
  Turk & Pentland (1991); Murase & Nayar (1995); etc.

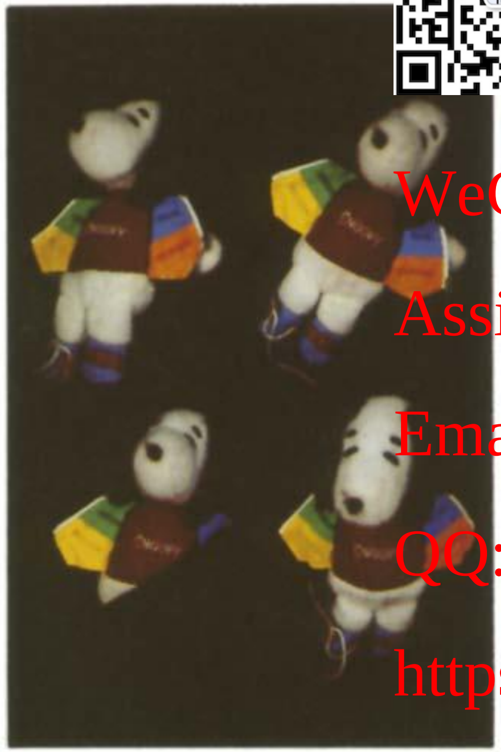
Eigenfaces (Turk & Pentland 1991)

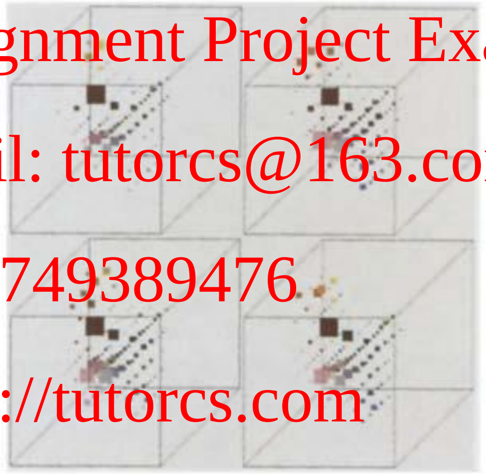


Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

Colour Histograms

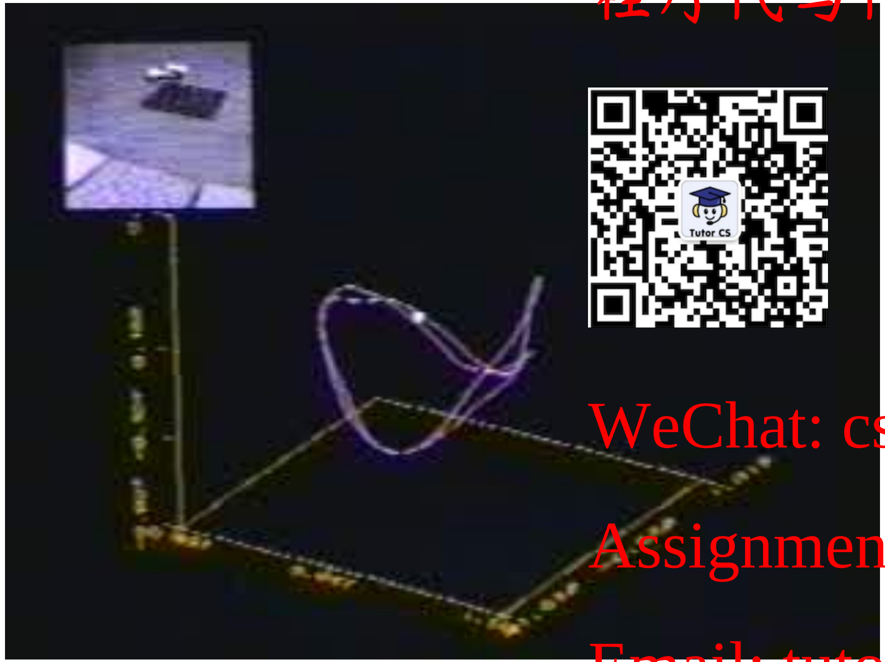






Swain and Ballard, Color Indexing, IJCV 1991.

Appearance Manifolds



H. Murase and S. Nayar, Visual learning and recognition of 3-d objects from appearance, IJCV 1995

Limitations of Global Appearance Models

- Requires global registration of patterns
- Not robust to clutter, occlus… metric transformations

History of Ideas in Recognition

- 1960s – early 1990s: the geometric era
- 1990s – appearance-based
- Mid-1990s – sliding window ches

Sliding Window Approaches

Sliding Window Approaches

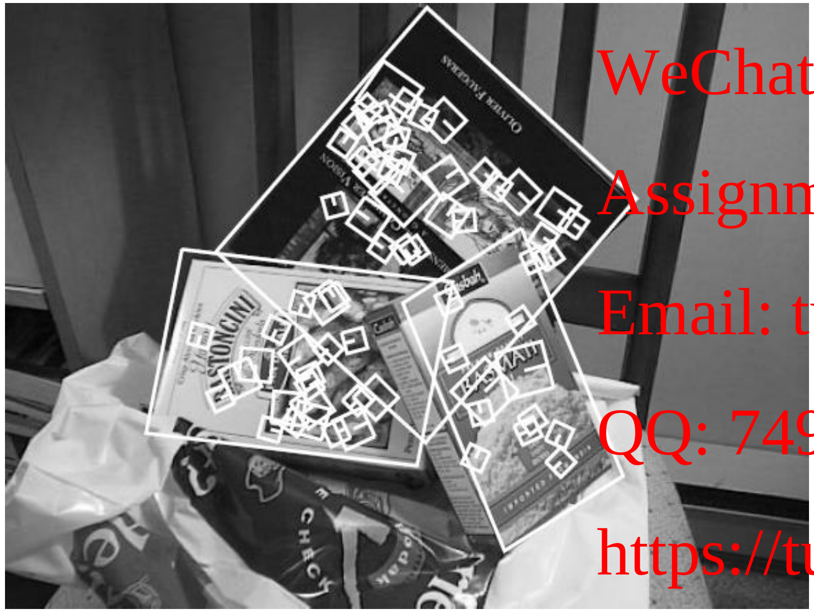
- Turk and Pentland, 1991
- Belhumeur, Hespanha, & Kriegman, 1997
- Schneiderman & Kanade 2004
- Viola and Jones, 2000

- Schneiderman & Kanade, 2004
- Argawal and Roth, 2002
- Poggio et al. 1993

History of Ideas in Recognition

- 1960s – early 1990s: the geometric era
- 1990s – appearance-based
- Mid-1990s – sliding window ches
- Late-1990s – local features

Local features for Object Instance Recognition



D. Lowe (1999, 2004)

Large-Scale Image Search

- Combine local features, indexing, and spatial constraints

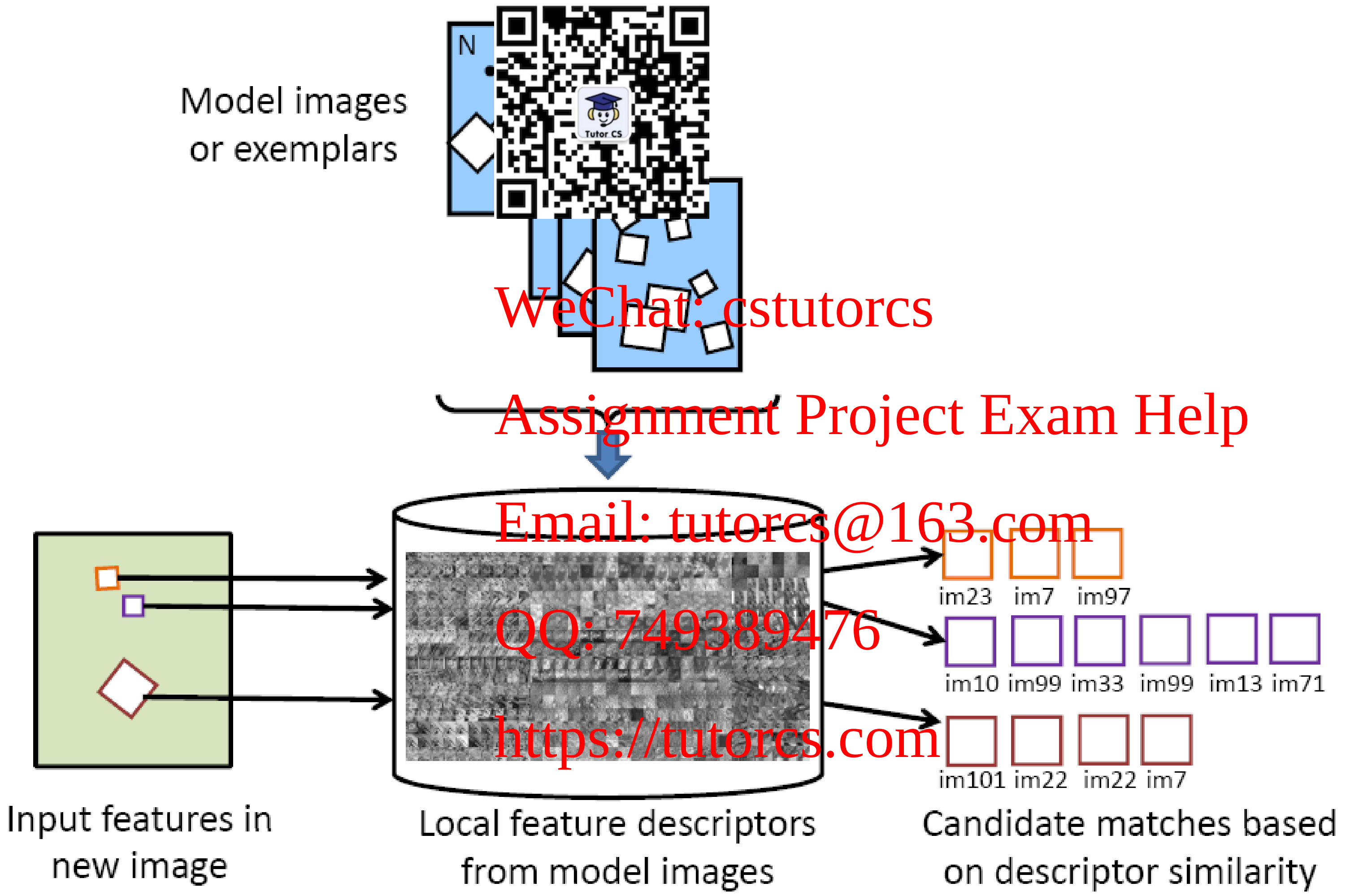


Image credit: K. Grauman and B. Leibe

Large-Scale Image Search

- Combine local features, indexing, and spatial constraints

Philbin et al. '07

History of Ideas in Recognition

- 1960s – early 1990s: the geometric era
- 1990s – appearance-based models
- Mid-1990s – sliding window approaches
- Late-1990s – local features
- Early-2000s – parts-and-shape models

Parts-and-Shape Models

- Model
  - Object as a set of parts
  - Relative locations between p
  - Appearance of parts

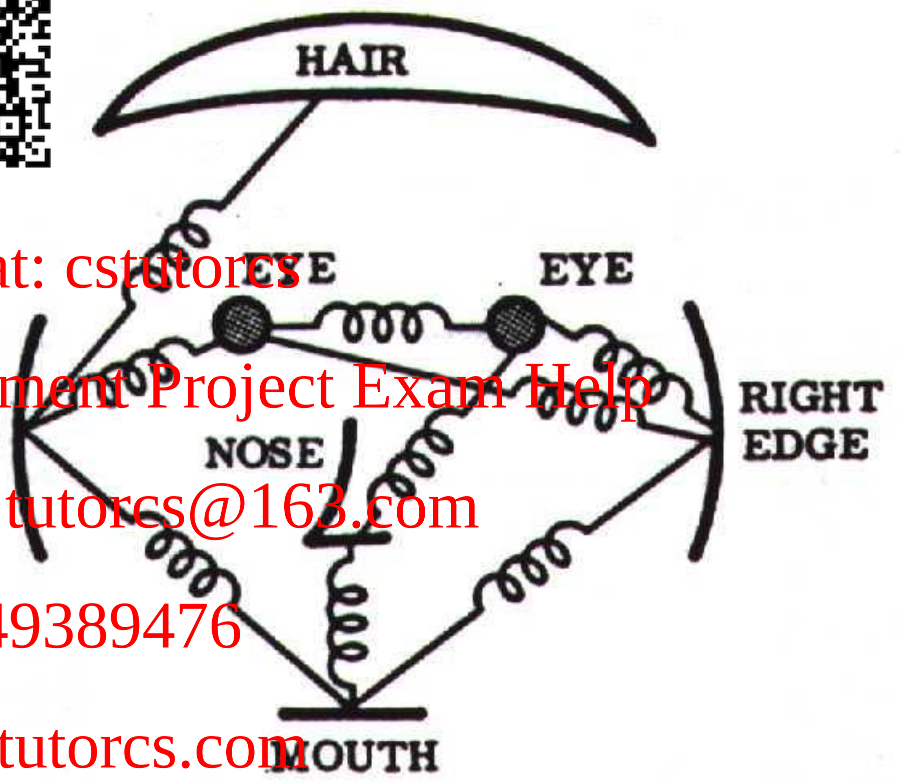


Figure from [Fischler & Elschlager 73]

Constellation Models

Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

Pictorial Structure Models

- Representing people

$$\Pr(P_{tor}, P_{arm}, \ldots | Im) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(Im(P_i))$$

Part Geometry **Part Appearance**

Fisheler and Elschlager(73), Felzenszwalb and Huttenlocher(00)

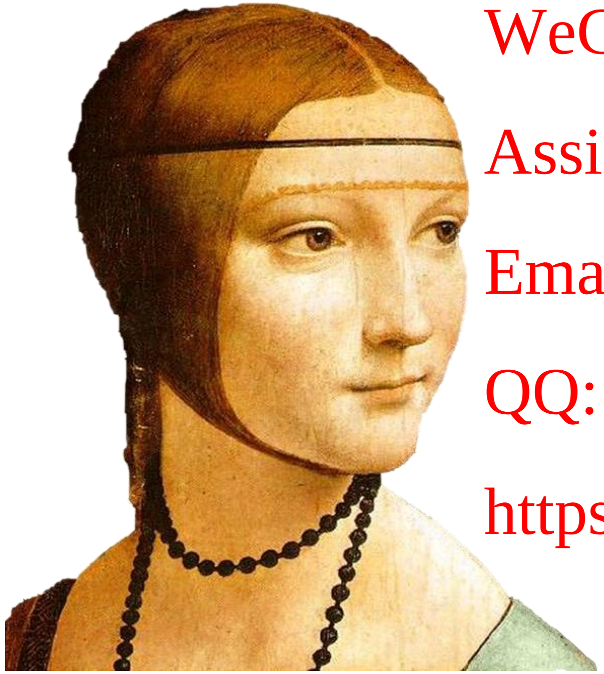
History of Ideas in Recognition

- 1960s – early 1990s: the geometric era
- 1990s – appearance-based [illegible]
- Mid-1990s – sliding window [illegible]ches
- Late-1990s – local features
- Early-2000s – parts-and-shape models
- Mid-2000s – bags of features

Bag-of-Features Model

Object

Bag of ‘words’



Objects as Texture

- All of these are being treated as the same

- No distinction between foreground and background: Scene recognition?

History of Ideas in Recognition

- 1960s – early 1990s: the geometric era
- 1990s – appearance-based m
- Mid-1990s – sliding window ches
- Late-1990s – local features
- Early-2000s – parts-and-shape models
- Mid-2000s – bags of features
- Present trends: combination of local and global methods, data-driven methods, context

Global Scene Descriptors

- The “gist” of a scene: Oliva & Torralba (2001)

http://people.csail.mit.edu/torralba/code/spatialenvelope/

Data Driven Methods

J. Hays and A. Efros, [Scene Completion using Millions of Photographs](), SIGGRAPH 2007

Data Driven Methods

J. Tighe and S. Lazebnik, ECCV 2010

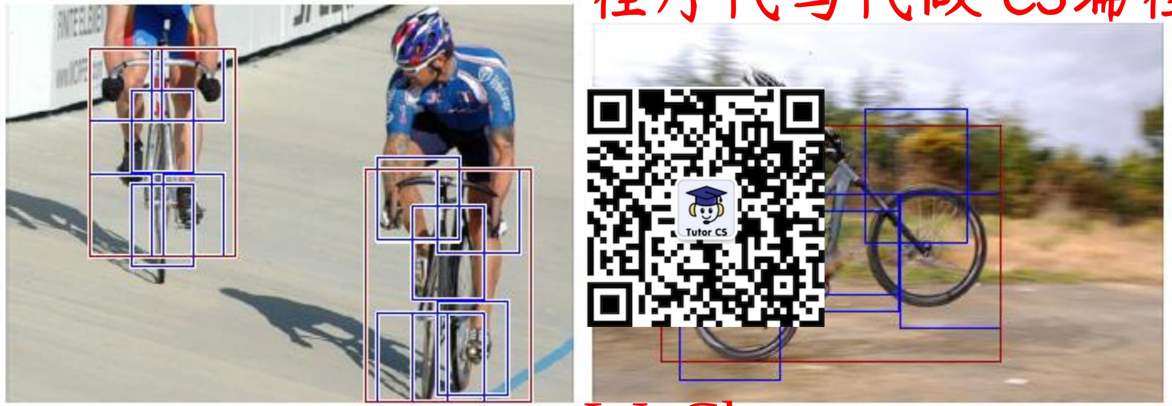
Geometric Context

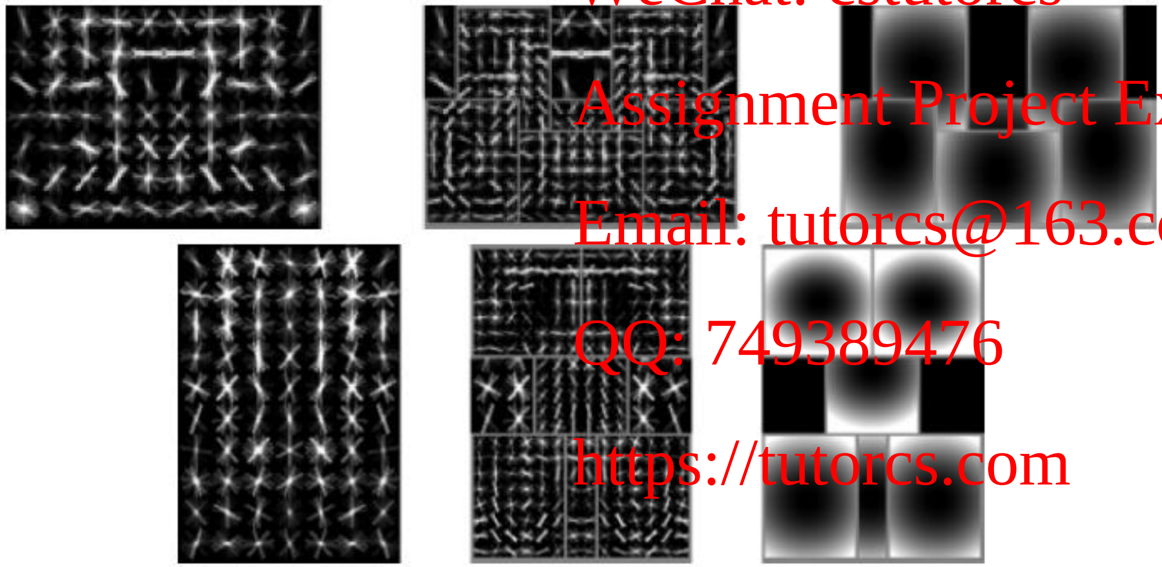
(a) Input image (c) Surface estimate (e) P(viewpoint | objects)

(b) P(person) = uniform (d) P(person | geometry) (f) P(person | viewpoint) (g) P(person|viewpoint,geometry)

D. Hoiem, A. Efros, and M. Herbert. [Putting Objects in Perspective.] CVPR 2006.

Discriminatively Trained Part-based Models





P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, **"Object Detection with Discriminatively Trained Part-Based Models,"** PAMI 2009

What “Works” Today

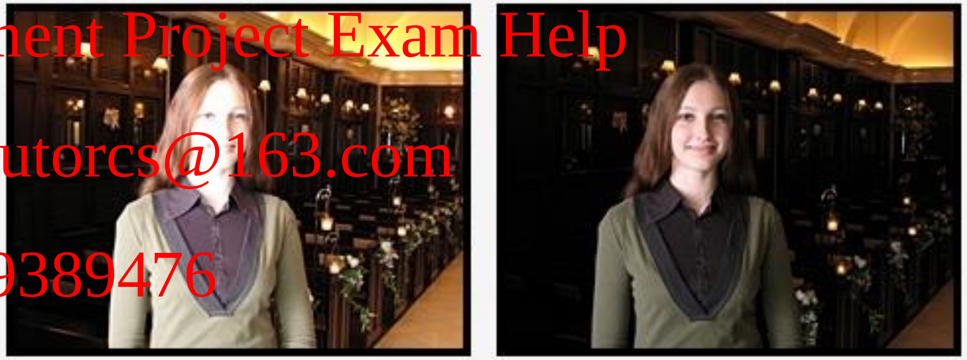
- Reading license plates, postcodes, cheques

What “Works” Today

- Reading license plates, postcodes, cheques
- Fingerprint recognition

What “Works” Today

- Reading license plates, postcodes, cheques
- Fingerprint recognition
- Face detection



[Face priority AE] When a bright part of the face is too bright

What “Works” Today

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

Recognition: A Machine Learning Approach

Slides adapted from Fei-Fei Li, Rob Fergus, Antonio Torralba, Kristen Grauman, and Derek Hoiem

The Machine Learning Framework

- Apply a prediction function to a feature representation of the image to get the desired output

f() = “apple”

f() = “tomato”

f() = “cow”

The Machine Learning Framework

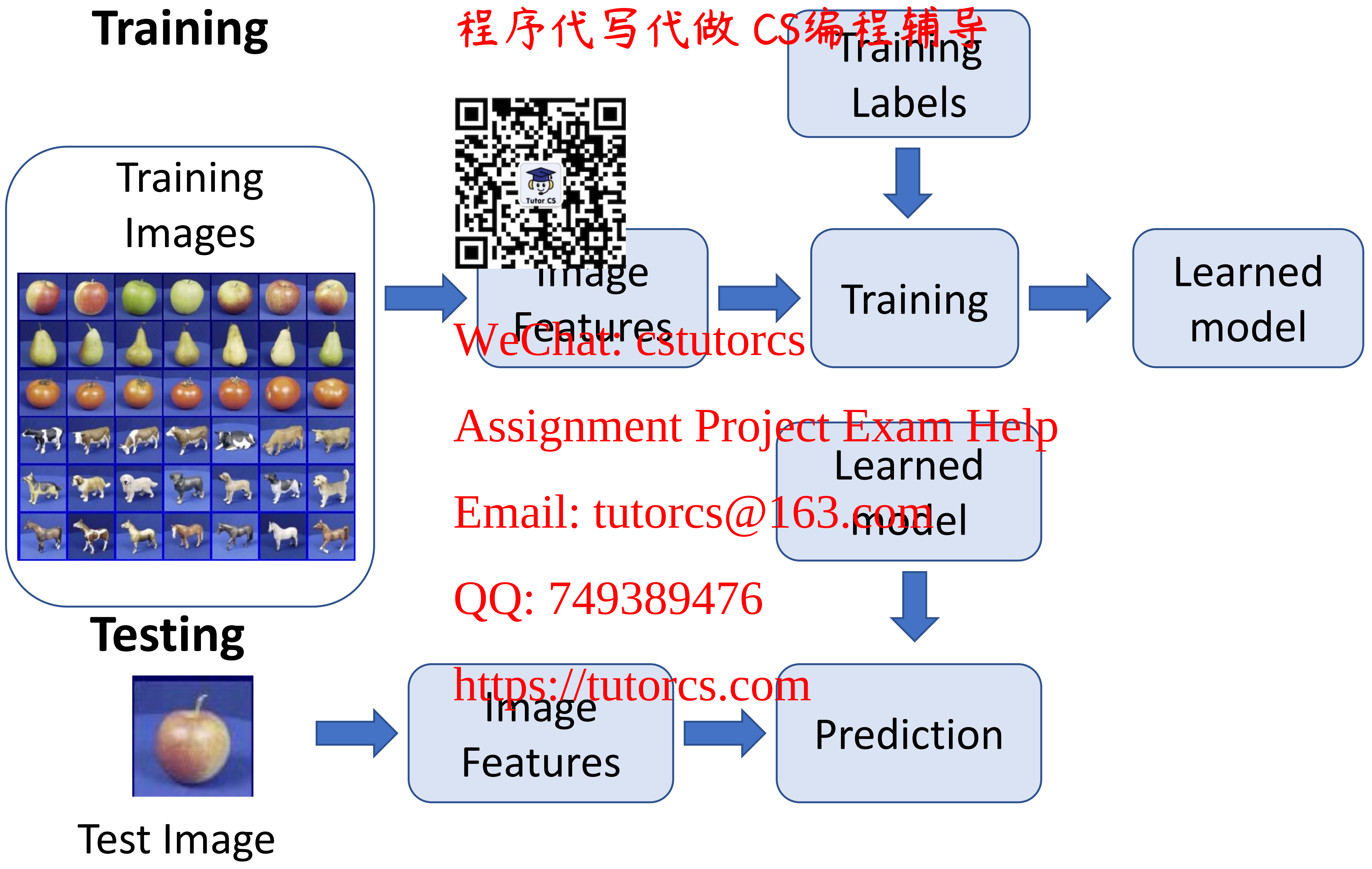
$$\mathbf{y = f(x)}$$

output prediction function Image feature

- Training: given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

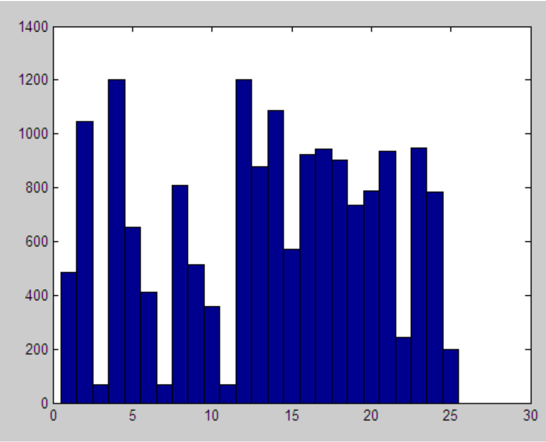
- Testing: apply f to a never seen before *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

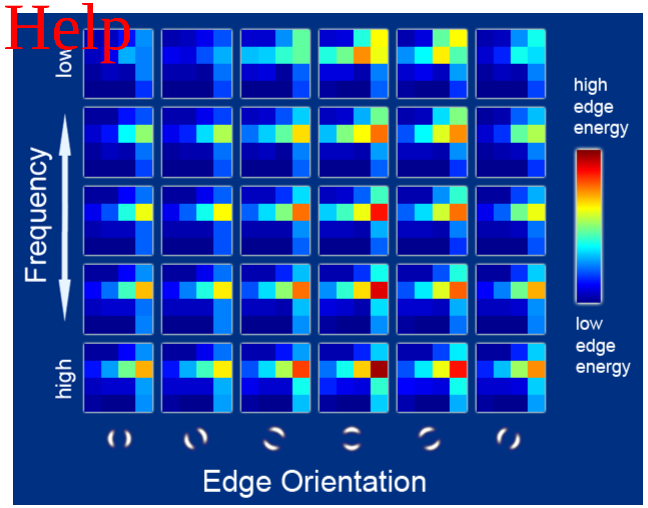
The Machine Learning Framework



Features

- Raw pixels
- Histograms
- GIST Descriptors
- …





Classifier: Nearest Neighbour

f(x) = label of the training sample nearest to x

- All we need is a distance function for the inputs
- No training required!

Classifier: Linear

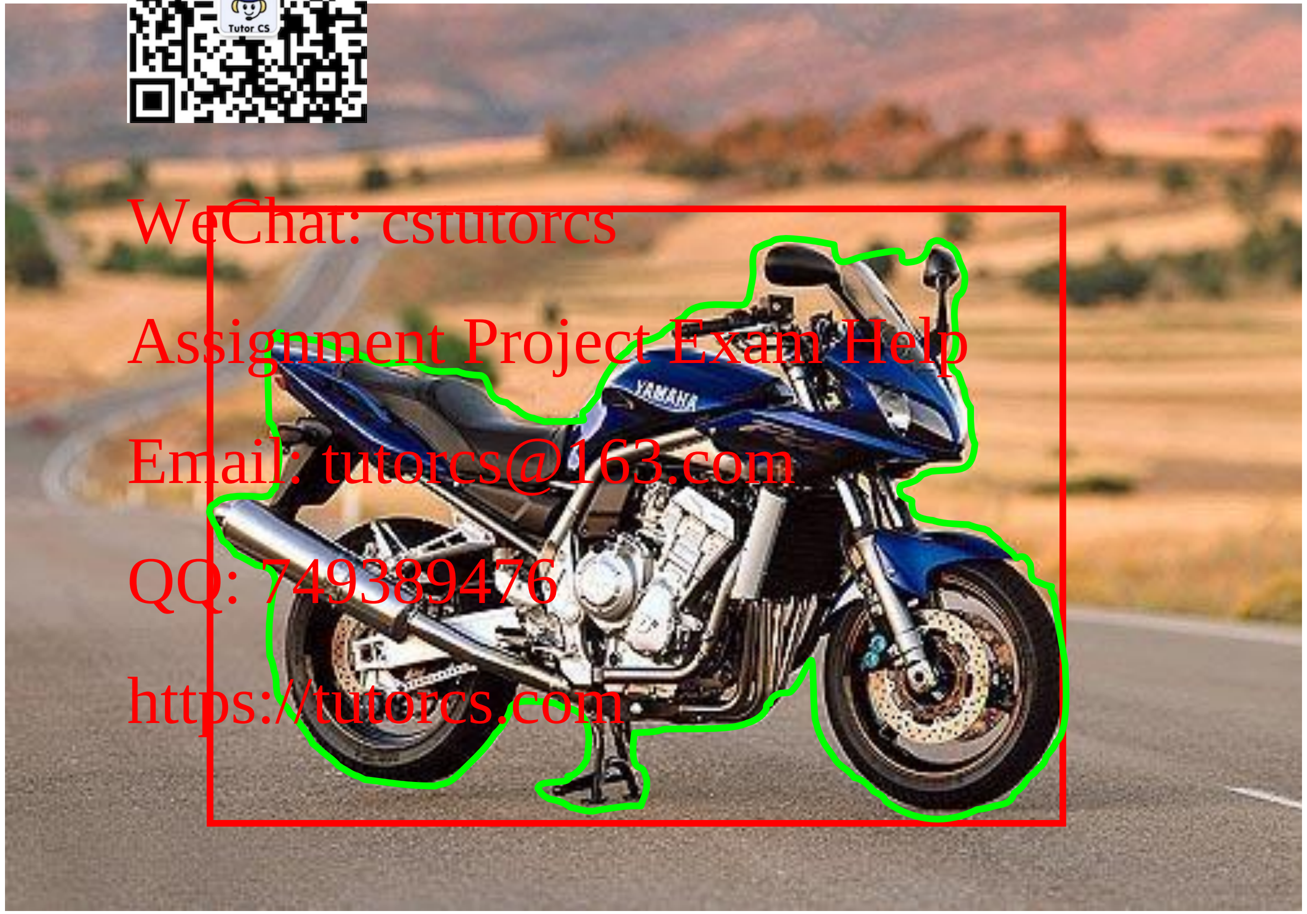
- Find a linear function to separate the classes:

$$f(x) = \mathrm{sgn}(w \cdot x + b)$$

Recognition Task and Supervision

- Images in the training set must be annotated with the "correct answer" that the model is expected to pr

"Contains a motorbike"



Spectrum of Supervision

Definition depends on task

Generalisation

- How well does a learned model generalise from the data it was trained on to a new test set?

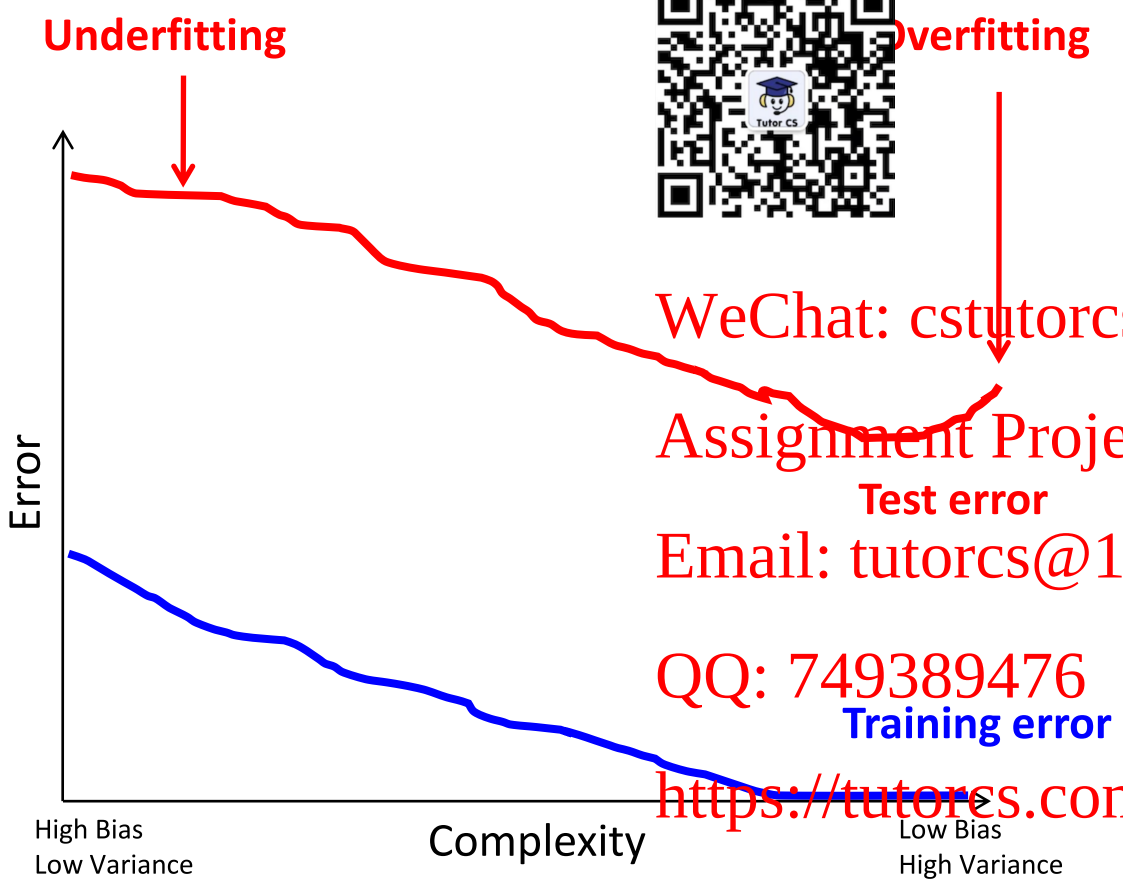
Training set (labels known)

Test set (labels unknown)

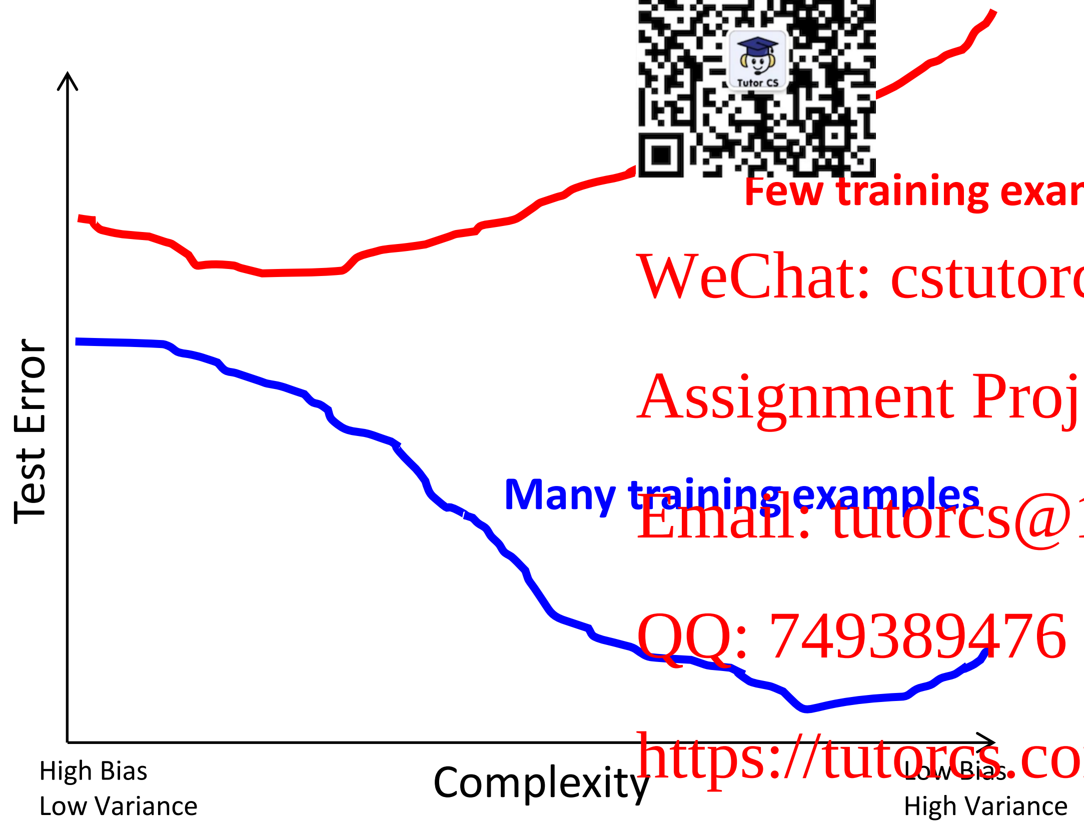
Generalisation

- Components of generalisation error
  - Bias: how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - Variance: how much models learned from different training sets differ from each other
- Underfitting: model is too "simple" to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error
- Overfitting: model is too "complex" and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

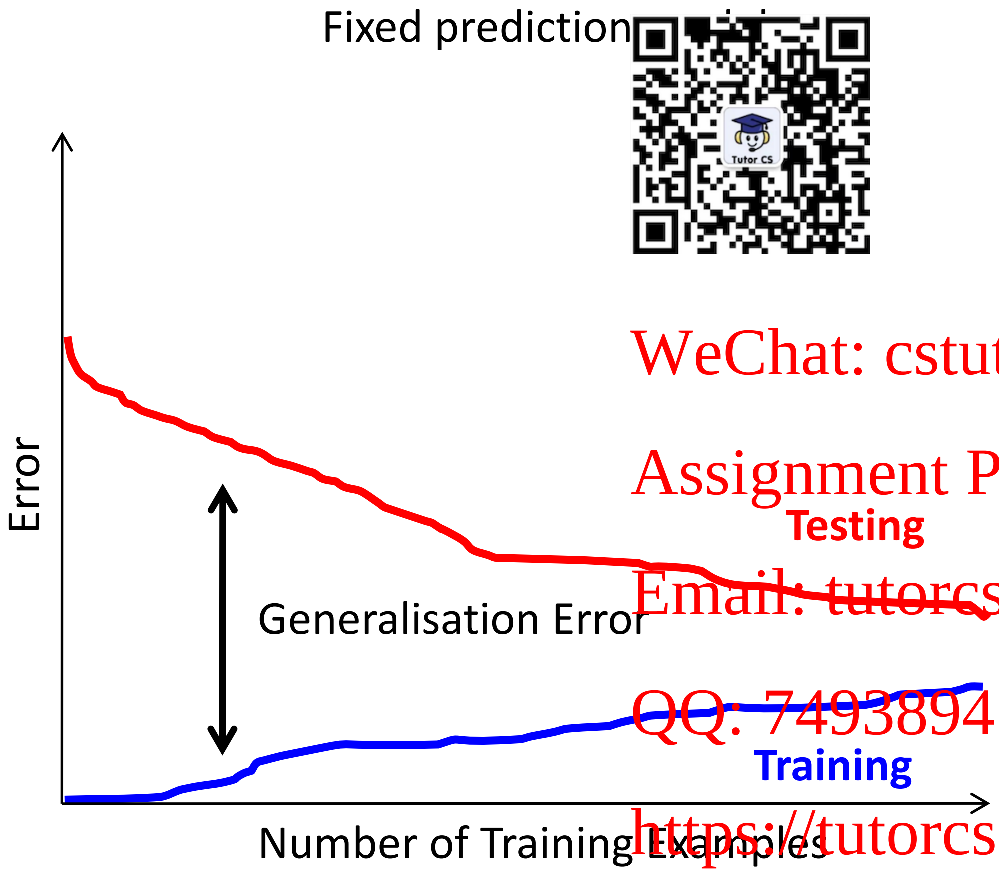
Bias-Variance Tradeoff



Effect of Training Size



Effect of Training Size



Datasets

- Circa 2001: 5 categories, 100s of images per category
- Circa 2004: 101 categories
- Today: up to thousands of categories, millions of images

Caltech 101 & 256

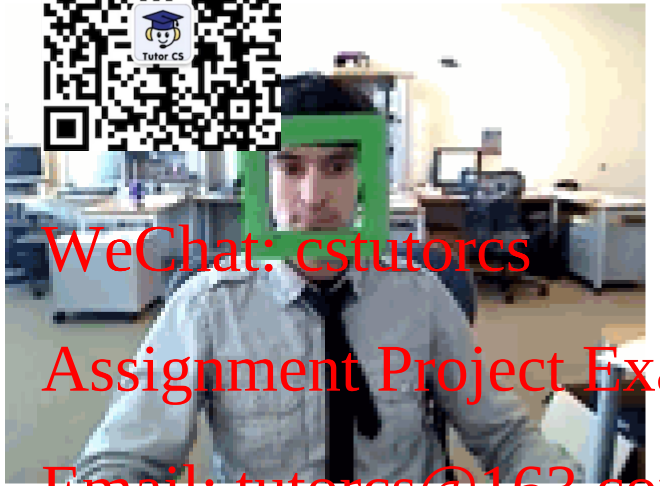
http://www.vision.caltech.edu/Image_Datasets/Caltech101/
http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

Caltech 101: Intraclass Variability

Face Detection



Behold a state-of-the-art face detector!

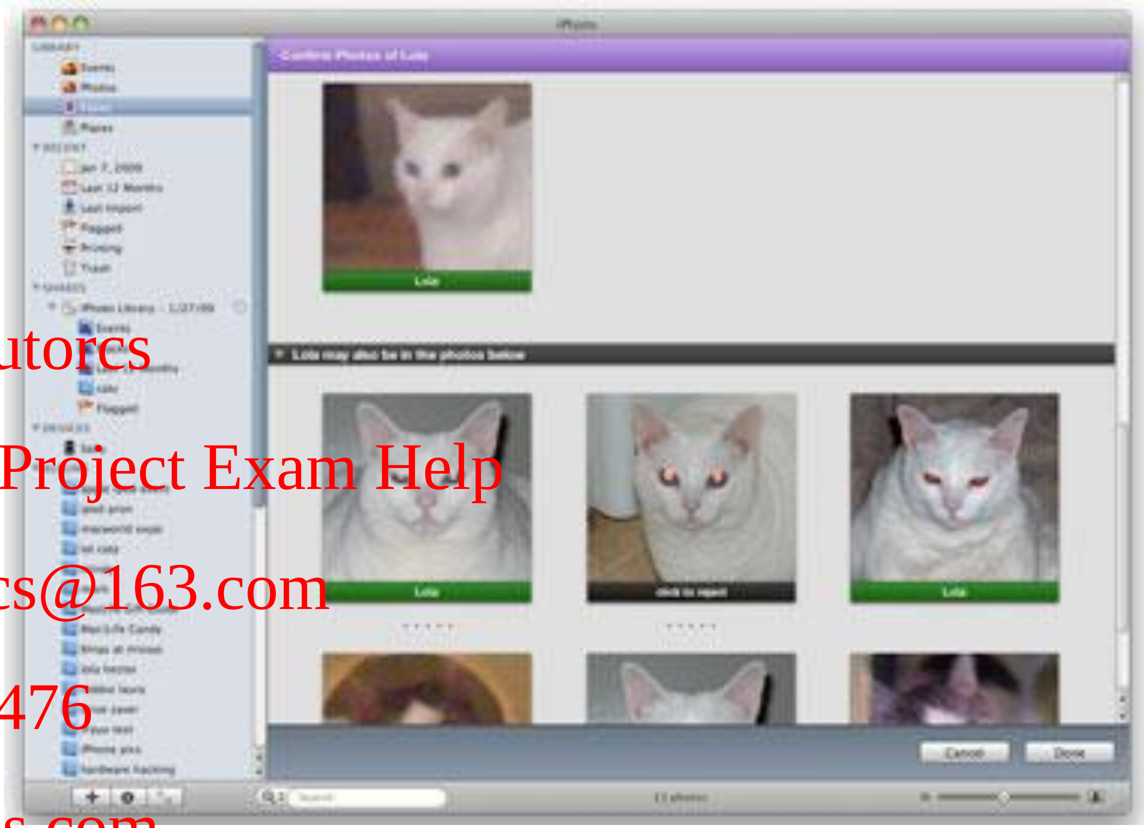
(Courtesy Boris Babenko)

Face Detection and Recognition

Consumer Application: Apple iPhoto

http://www.apple.com/ilife/iphoto/

Consumer Application: Apple iPhoto

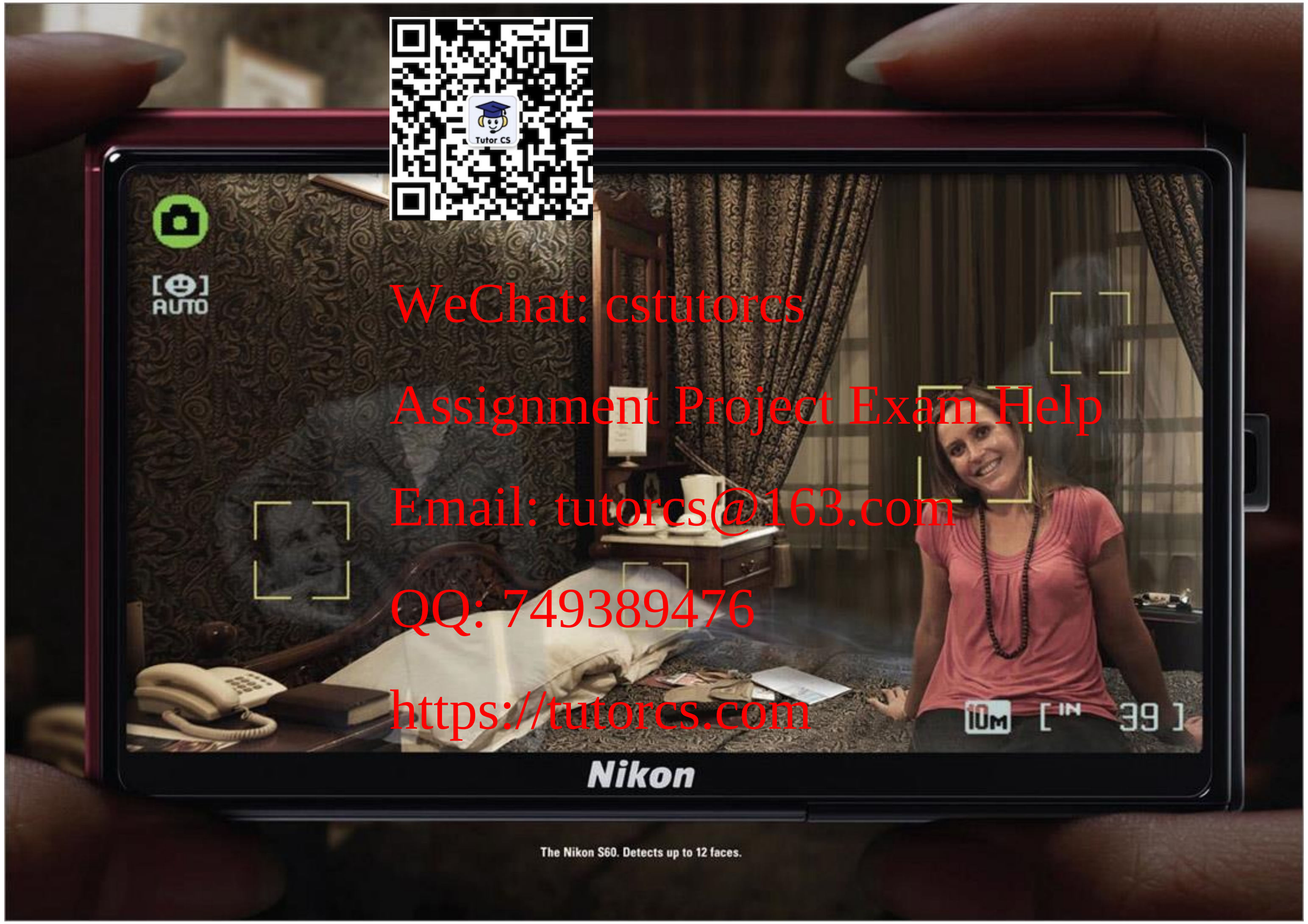


Consumer Application: Apple iPhoto

- Things iPhoto thinks are faces

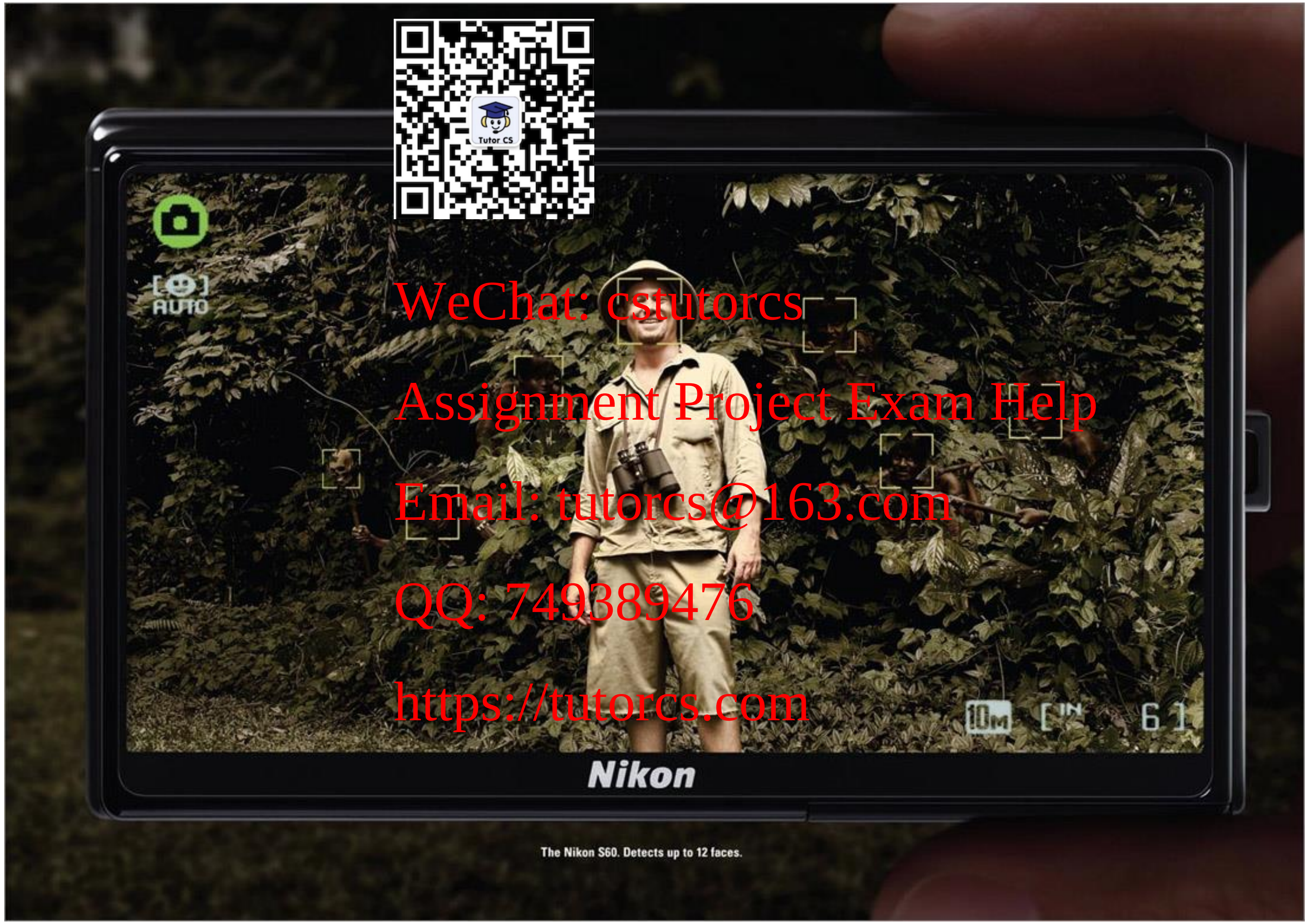
Funny Nikon Ads

"The Nikon S60 detects up to 12 faces."



Funny Nikon Ads

"The Nikon S60 detects up to 12 faces."



Challenges of Face Detection

- Sliding window detector must evaluate tens of thousands of local/scale combinations
- Faces are rare: 0 – 10 per im
  - For computational efficiency, ld try to spend as little time as possible on the non-face windows
  - A megapixel image has ~10^6 pixels and a comparable number of candidate face locations
  - To avoid having a false positive in every image, our false positive rate has to be less than 10^{-6}

The Viola/Jones Face Detector

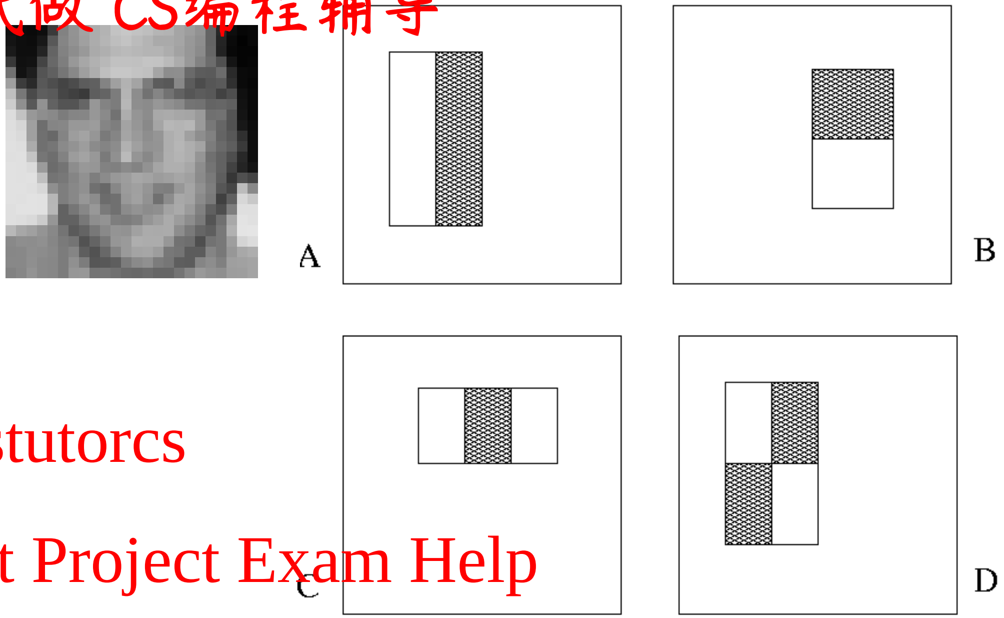
- A seminal approach to real-time object detection
- Training is slow but detection is very fast
- Key ideas
  - Integral images for fast feature evaluation
  - Boosting for feature selection
  - *Attentional cascade* for fast rejection of non-face windows

P. Viola and M. Jones. *Rapid object detection using a boosted cascade of simple features.* CVPR 2001.

P. Viola and M. Jones. *Robust real-time face detection.* IJCV 57(2), 2004.

Image Features

- "Rectangular filters"

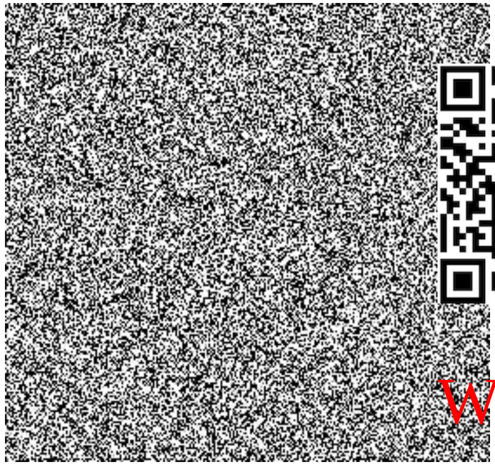


- *Value* = $\sum$ *(pixels in white area)*

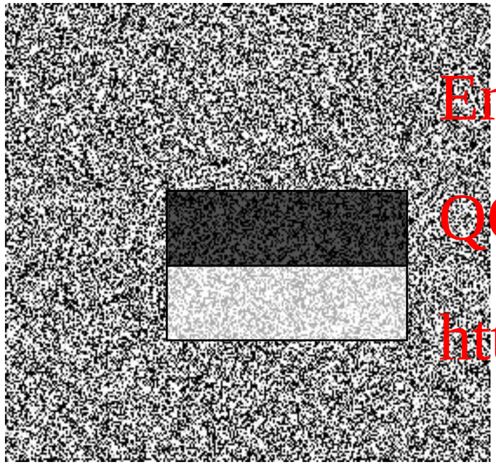
  $-\sum$ *(pixels in black area)*

Example

Source

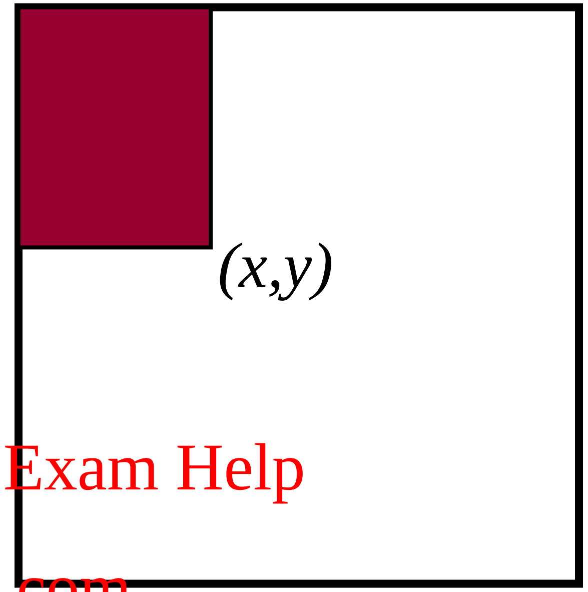


Result



Fast Computation with Integral Images

- The *integral image* computes a value at each pixel *(x,y)* that sum of the pixel values abov to the left of *(x,y)*, inclusive
- This can quickly be computed in one pass through the image



Computing the Integral Image

Computing the Integral Image

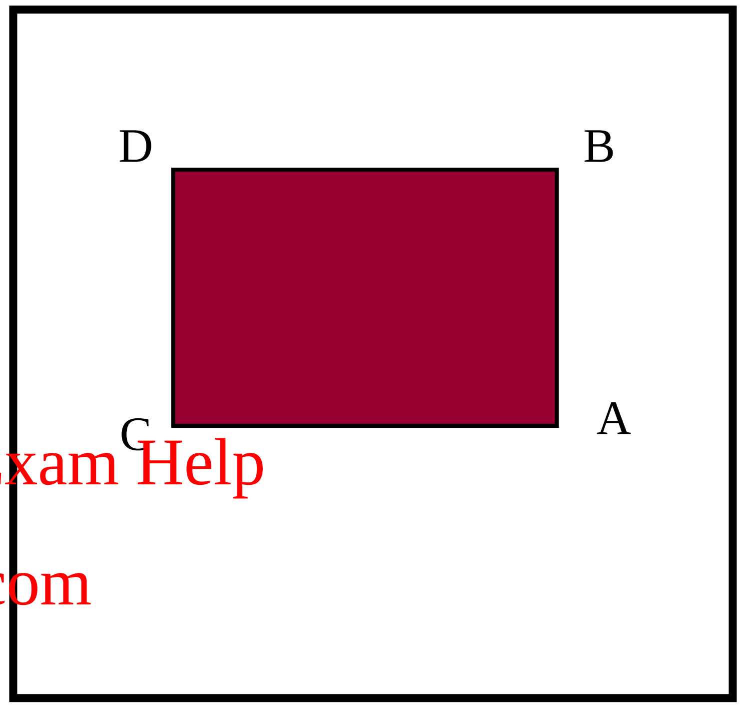
- Cumulative row sum: $s(x, y) = s(x-1, y) + i(x, y)$
- Integral image: $ii(x, y) = ii(x, y-1) + s(x, y)$

Computing Sum within a Rectangle

- Let A,B,C,D be the values of the integral image at the corner of the rectangle
- Then the sum of original image values within the rectangle can be computed as:

  sum = A – B – C + D

- Only 3 additions are required for any size of rectangle!



Example

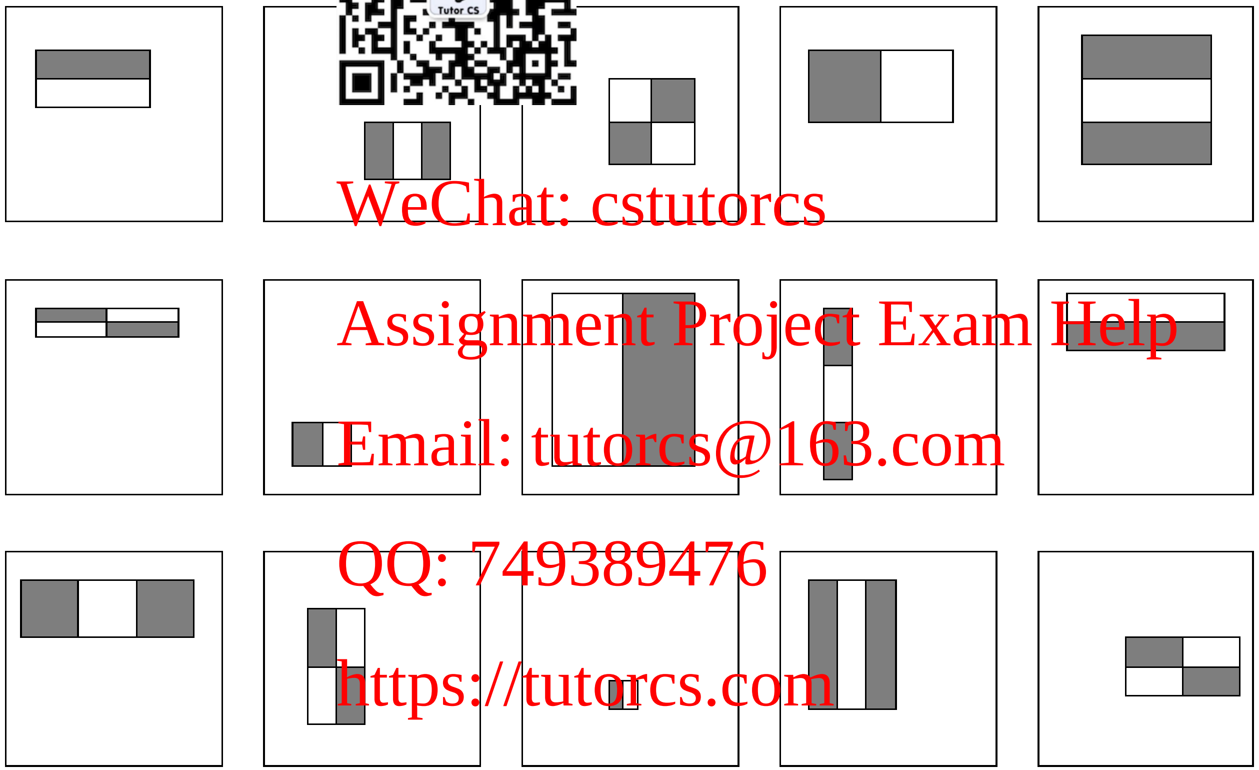
Black = A-B-C+D

White = C-D-E+F

Value = White-Black = -A+B+2C-2D-E+F

Feature Selection

- For a 24x24 detection region, the number of possible rectangle features is ~160,000!



Feature Selection

- For a 24x24 detection region, the number of possible rectangle features is ~160,000!
- At test time, it is impractical to evaluate the entire feature set
- Can we create a good classifier using just a small subset of all possible features?
- How to select such a subset?

Boosting

- Boosting is a classification scheme that combines *weak learners* into a more accurate *ensemble classifier*
- Training procedure
  - Initially, weight each training example equally
  - In each boosting round:
    - Find the weak learner that achieves the lowest *weighted* training error
    - Raise the 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)
    - Exact formulas for re-weighting and combining weak learners depend on the particular boosting scheme (e.g., AdaBoost)

Y. Freund and R. Schapire, A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780, September, 1999.

Boosting for Face Detection

- Define weak learners based on rectangle features

lue of rectangle feature

$$h_t(x) = \begin{cases} 1 & \text{if } \ldots) > p_t \theta_t \\ 0 & \text{otherwise} \end{cases}$$

window

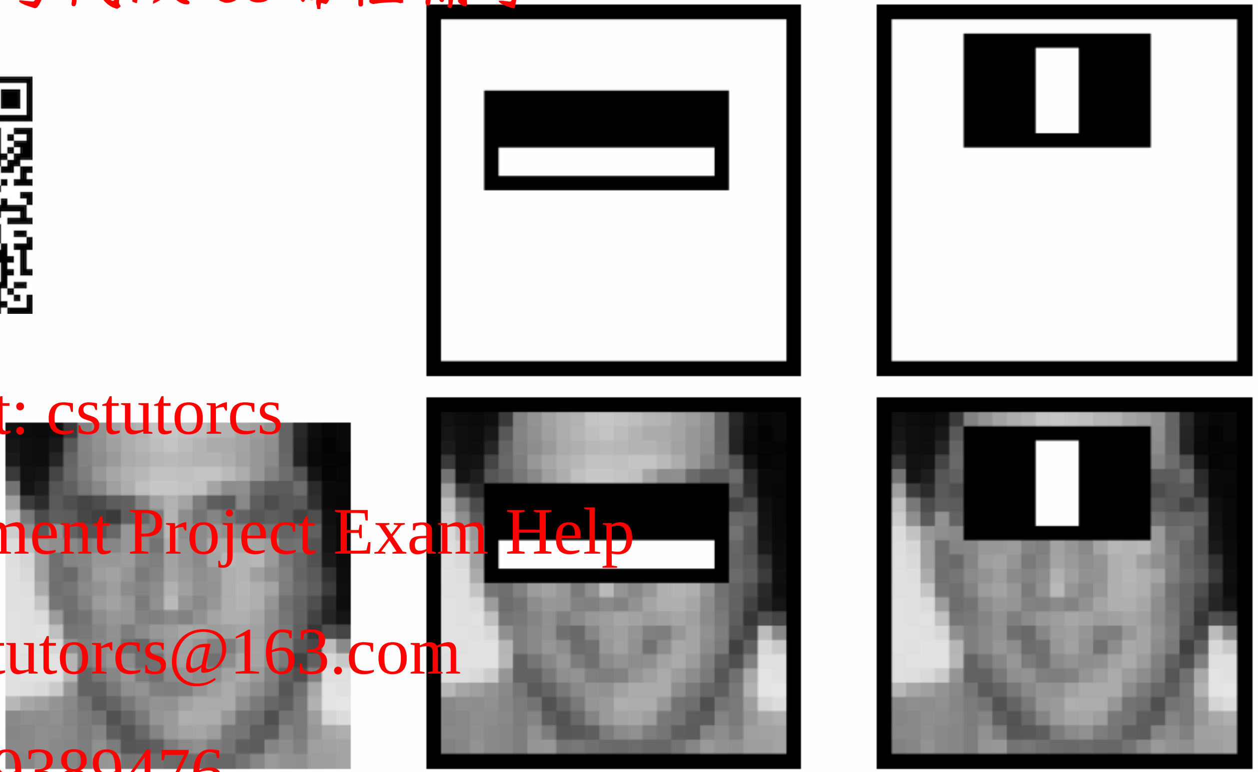
parity

threshold

- For each round of boosting:
  - Evaluate each rectangle filter on each example
  - Select best filter/threshold combination based on weighted training error
  - Reweight examples

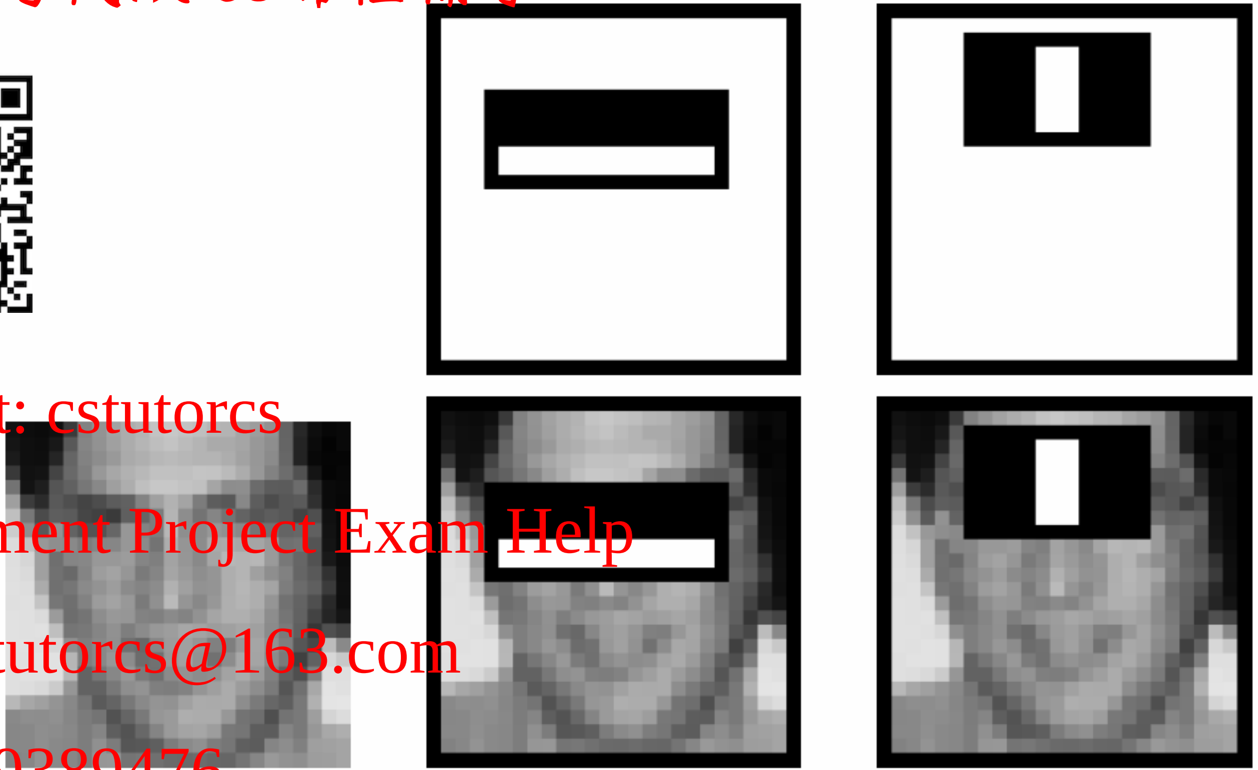
Boosting for Face Detection

- First two features selected by boosting
- This feature combination can yield 100% detection rate and 50% false positive rate



Boosting for Face Detection

- First two features selected by boosting
- This feature combination ca
  yield 100% detection rate a
  50% false positive rate



Boosting for Face Detection

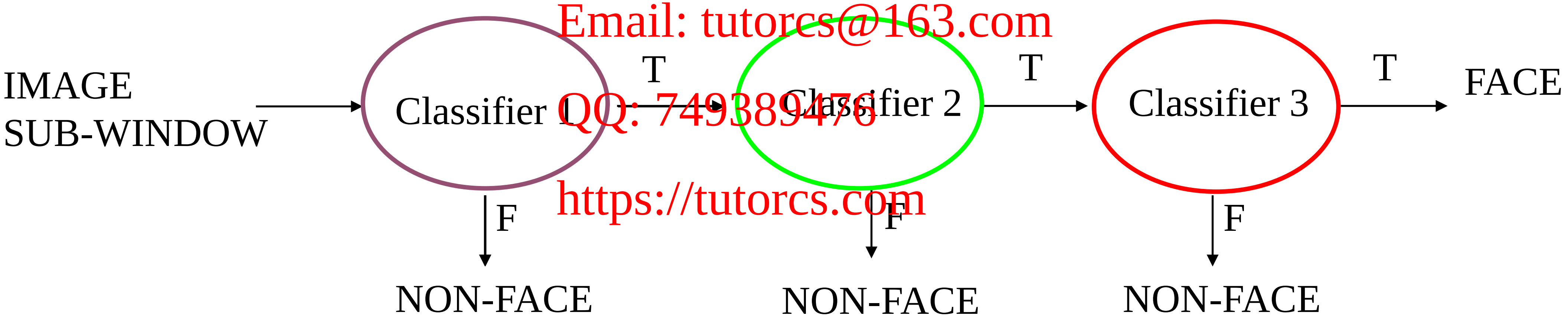
- A 200-feature classifier can yield 95% detection rate and a false positive rate of 1 in 14084

Not good enough!

Receiver operating characteristic (ROC) curve

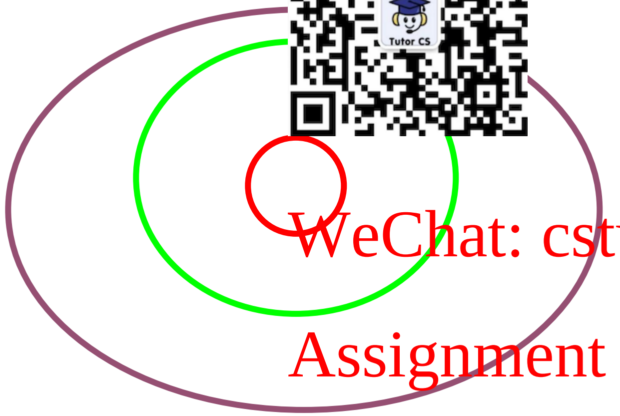
Attentional Cascade

- We start with simple classifiers which reject many of the negative sub-windows while detecting al[illegible] positive sub-windows
- Positive response from the f[illegible]sifier triggers the evaluation of a second (more complex) classifier, a[illegible]
- A negative outcome at any point leads to the immediate rejection of the sub-window

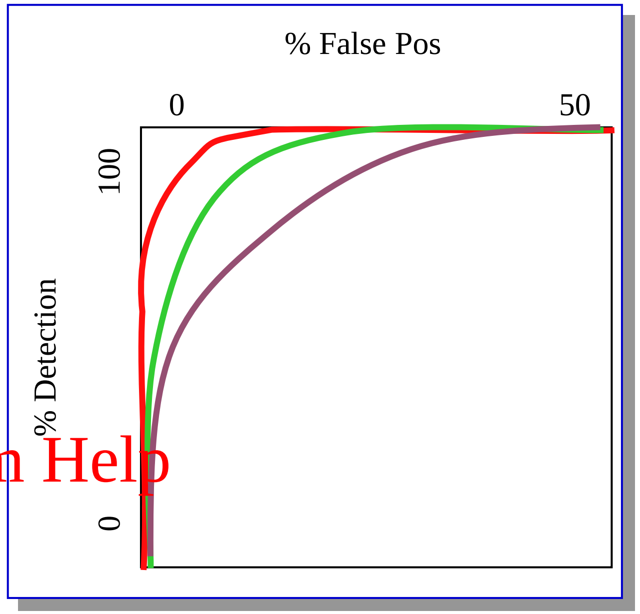


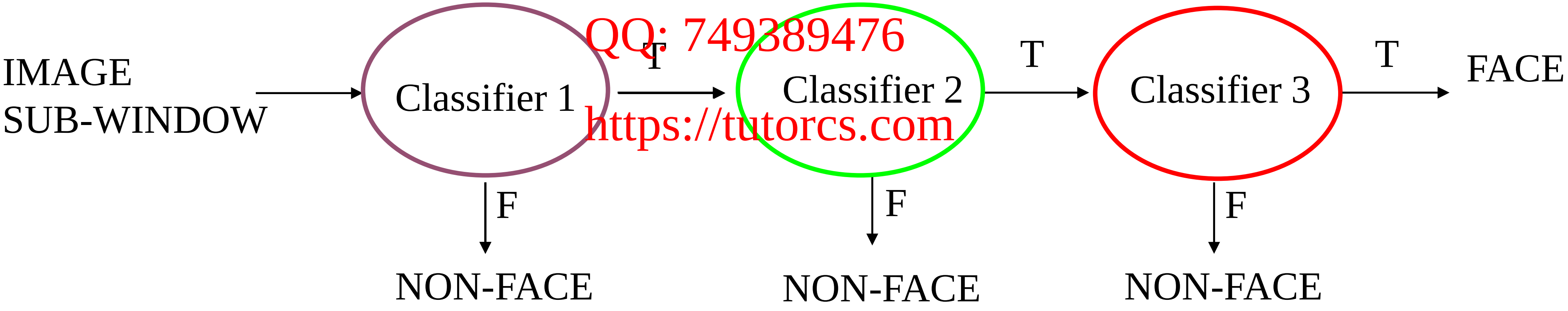
Attentional Cascade

- Chain classifiers that are progressively more complex and have lower false positive rates:



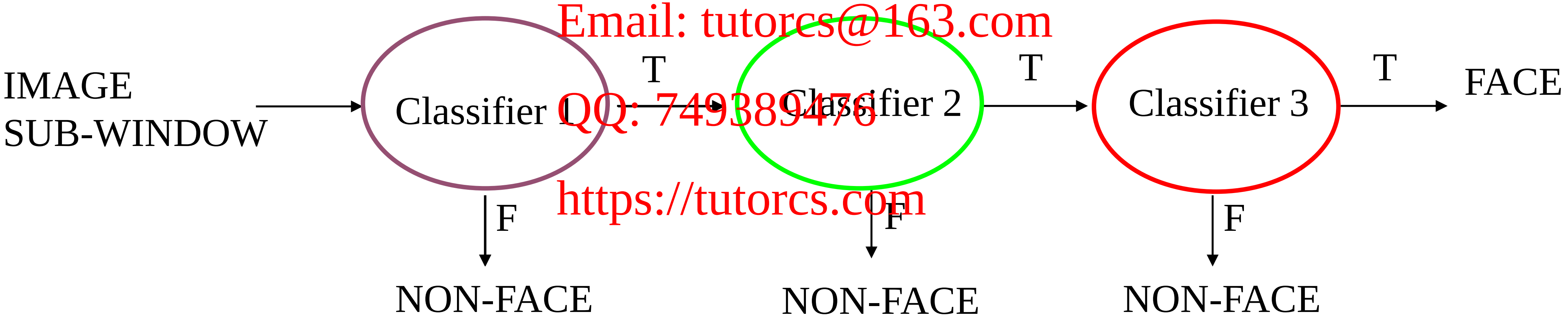
Receiver operating characteristic





Attentional Cascade

- The detection rate and the false positive rate of the cascade are found by multiplying the respective rates of the individual stages
- A detection rate of 0.9 and a false positive rate on the order of 10^{-6} can be achieved by a 10-stage cascade if each stage has a detection rate of 0.99 ($0.99^{10} \approx 0.9$) and a false positive rate of about 0.30 ($0.3^{10} \approx 6\times10^{-6}$)



Training the Cascade

- Set target detection and false positive rates for each stage
- Keep adding features to the current stage until its target rates have been met
  - Need to lower AdaBoost threshold to maximize detection (as opposed to minimizing total classification error)
  - Test on a *validation set*
- If the overall false positive rate is not low enough, then add another stage
- Use false positives from current stage as the negative training examples for the next stage

The Implemented System

- Training data
  - 5000 faces

    all frontal, rescaled to 24x2
  - 300 million non-faces

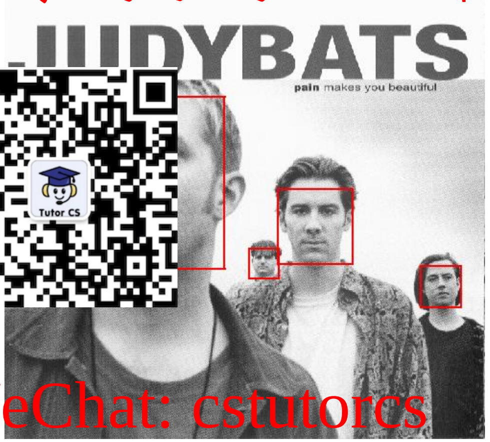
    9500 non-face images
  - Faces are normalized:

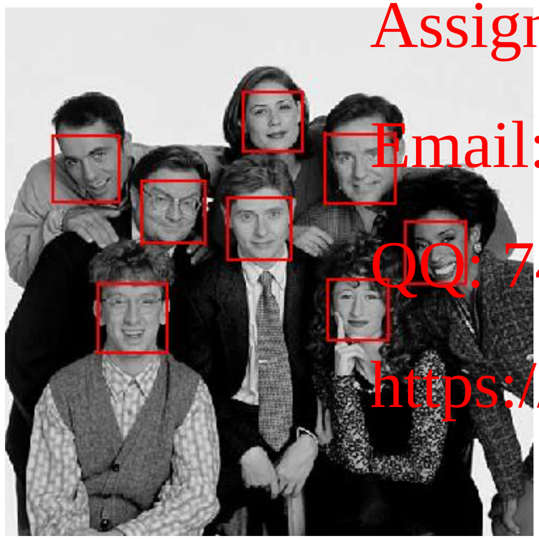
    scale, translation
- Many variations
  - Across individuals
  - Illumination
  - Pose

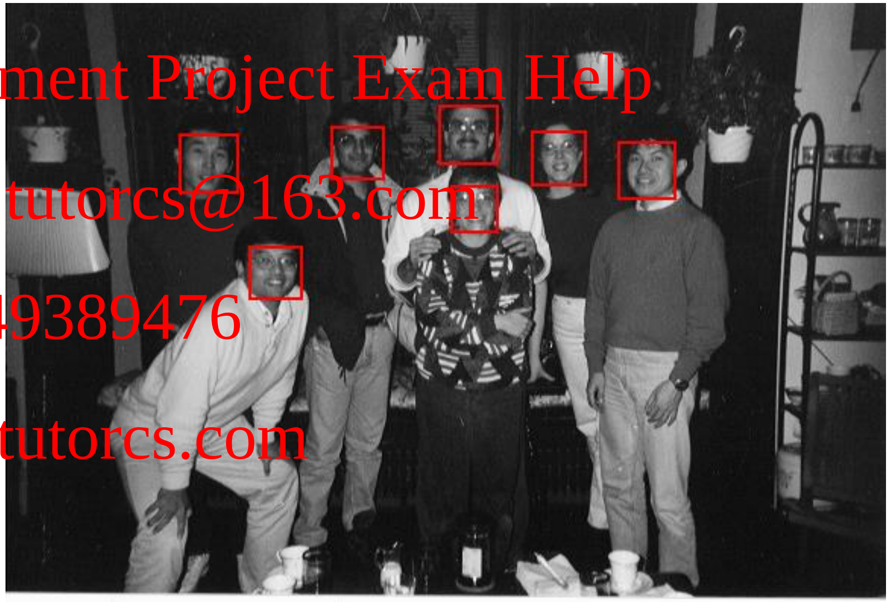
System Performance

- Training time: “weeks” on 466 MHz Sun workstation
- 38 layers, total of 6061 feat
- Average of 10 features evalu r window on test set
- “On a 700 Mhz Pentium III processor, the face detector can process a 384 by 288 pixel image in about .067 seconds”
  - 15Hz
  - 15 times faster than previous detector of comparable accuracy (Rowley et al., 1998)

Output of Face Detector on Test Images

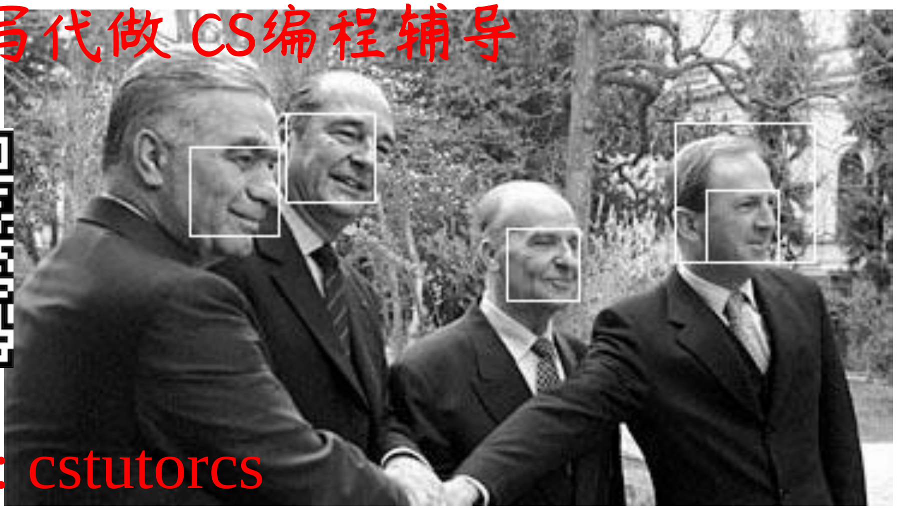






Other Detection Tasks

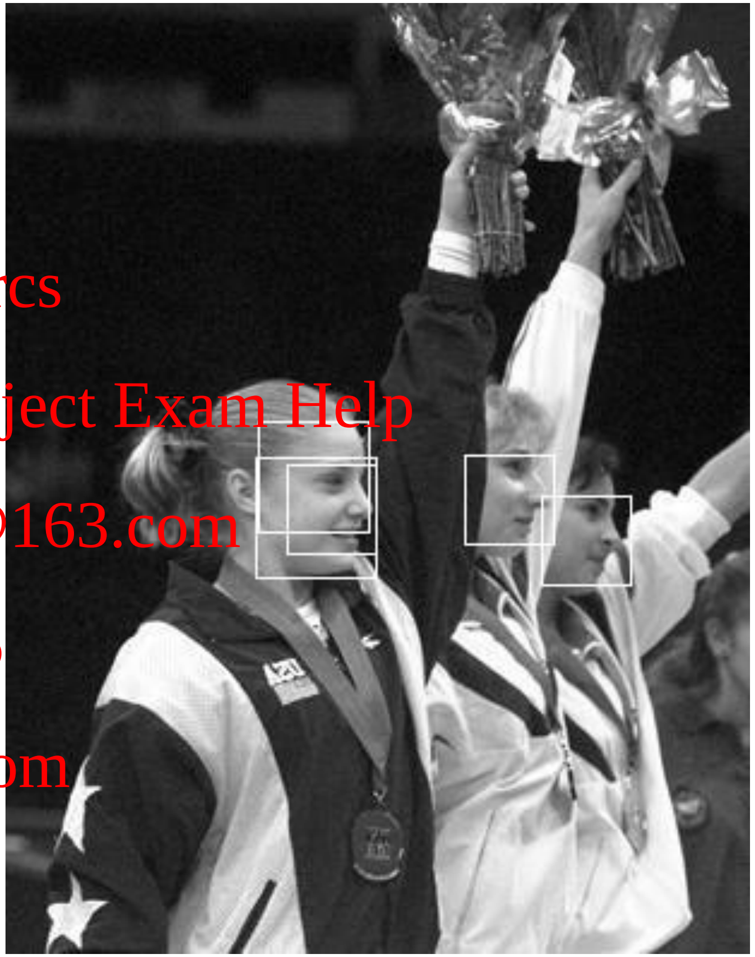
Facial Feature Localization



Profile Detection

Male vs. female

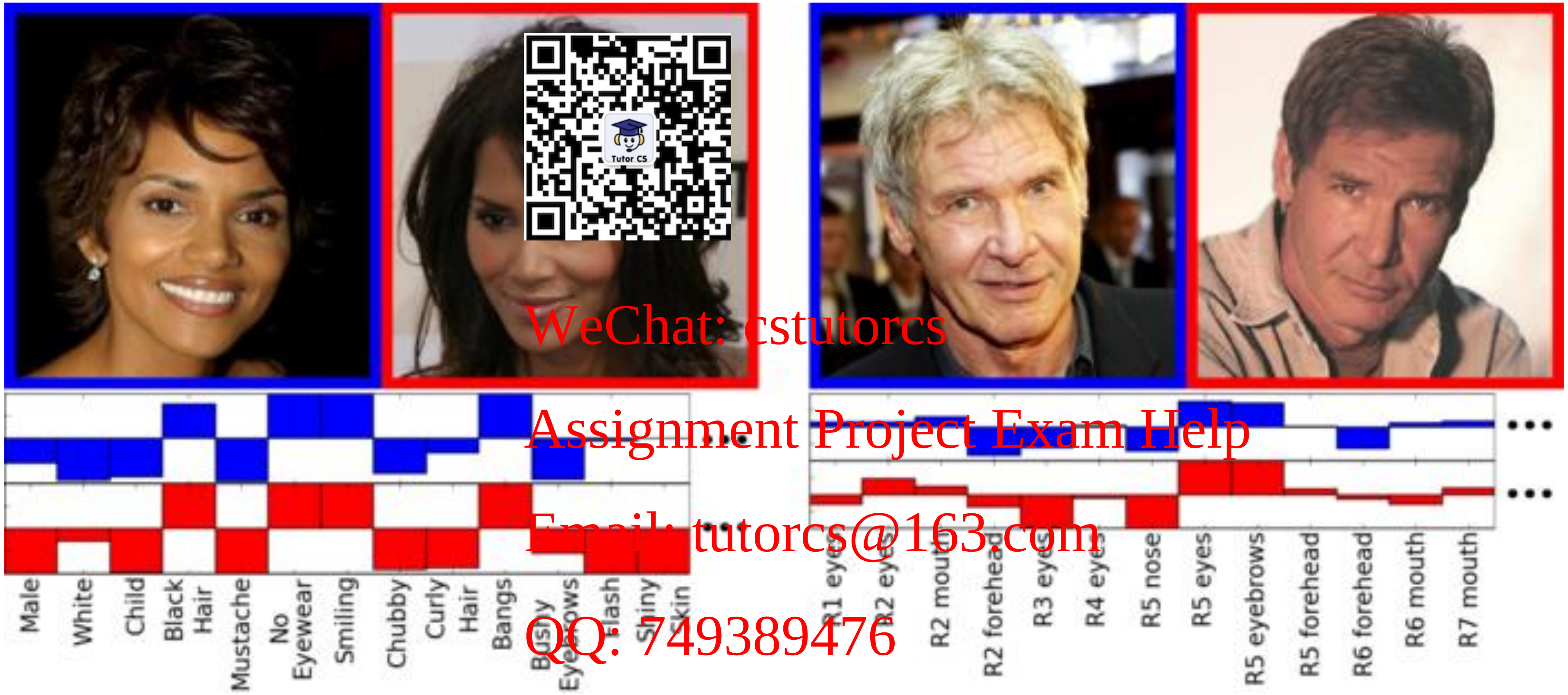
Profile Detection



Viola/Jones Detector Summary

- Rectangle features
- Integral images for fast computation
- Boosting for feature selection
- Attentional cascade for fast rejection of negative windows

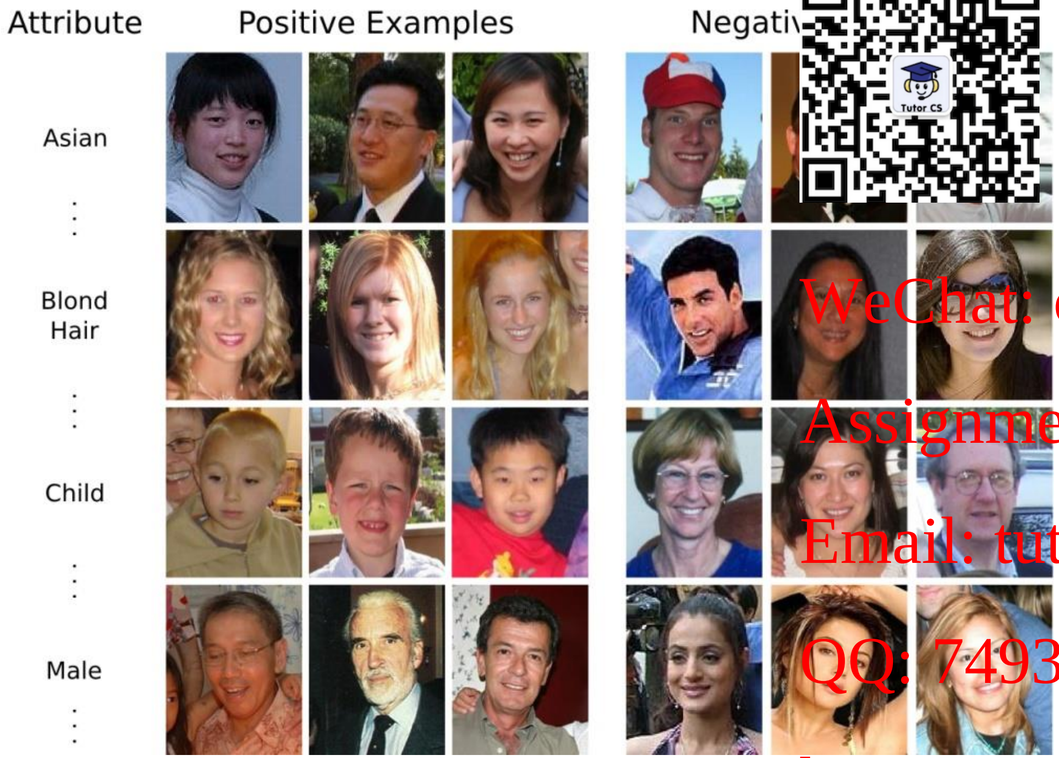
Face Recognition



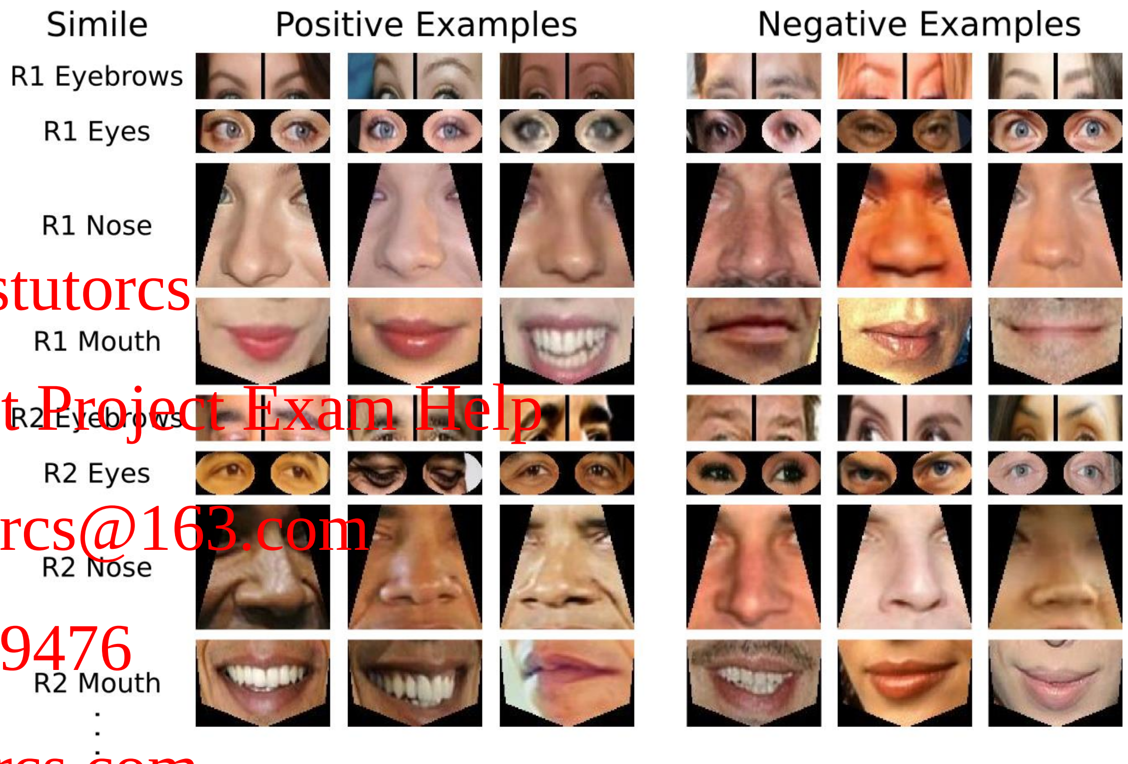
- N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, ["Attribute and Simile Classifiers for Face Verification,"]() ICCV 2009.

Face Recognition

Attributes for training



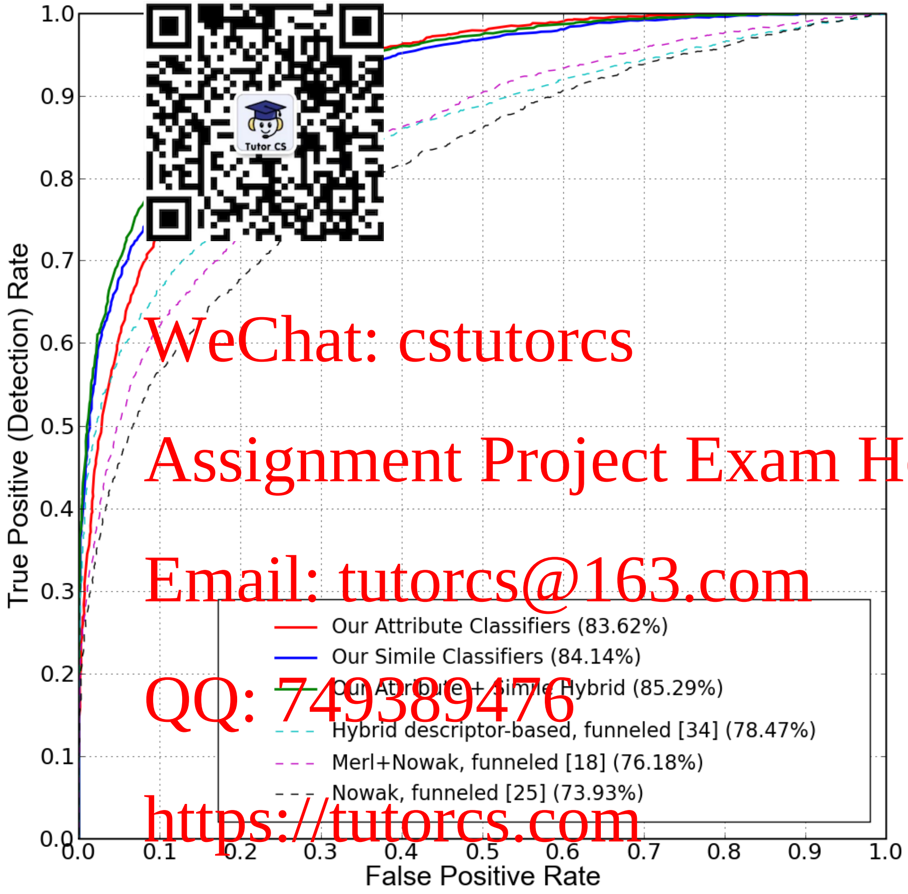
Similes for training



- N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, **"Attribute and Simile Classifiers for Face Verification,"** ICCV 2009.

Face Recognition

- Results on Labeled Faces in the Wild Dataset



- N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," ICCV 2009.

Summary

- What is object recognition?
- Briefly describe the history of [illegible] of object recognition.
- Describe the machine learning framework.
- Describe nearest neighbour and linear classifiers.
- What is the task of face detection and recognition?
- Describe the Viola/Jones Face Detector.