# Recognizing and Learning Object Categories

Based on work by R. Fergus, P. Perona, A. Zisserman, J. Ponce, S. Lazebnik, C. Schmid, F. DiMaio, and others

## Traditional Problem: Single Object Recognition


## Most Objects Exhibit Considerable Intra-Class Variability

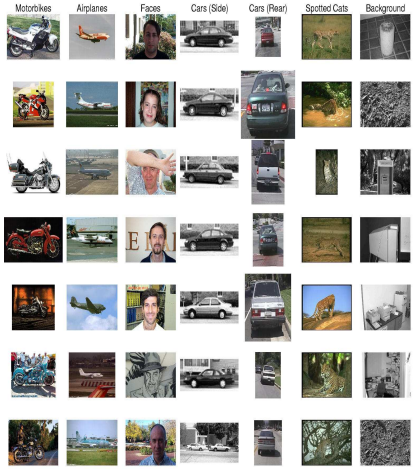
Task: Recognition of object categories

## Some object categories

Learn from just examples

Difficulties:

- Size variation
- Background clutter
- Occlusion
- Intra-class variation
- Viewpoint variation
- Illumination variation



## Chairs

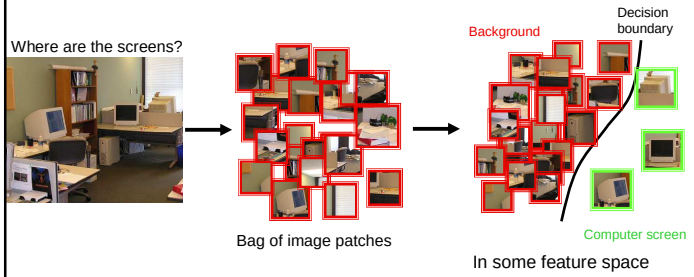
Related by function, not form

## Approach 1: Discriminative Methods

Object detection and recognition is formulated as a classification problem

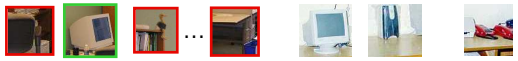
The image is partitioned into a set of overlapping windows

… and a decision is taken at each window about if it contains a target object or not




## Formulation

§ Formulation: binary classification



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Features | x = | $x_1$ | $x_2$ | $x_3$ | ⋯ | $x_N$ | $x_{N+1}$ | $x_{N+2}$ | ⋯ | $x_{N+M}$ |
| Labels | y = | -1 | +1 | -1 | | -1 | ? | ? | | ? |

Training data: each image patch is labeled as containing the object or not

Test data

- Classification function

$\hat{y} = F(x)$ Where $F(x)$ belongs to some family of functions

- Minimize misclassification error

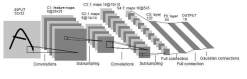
(Not that simple: we need some guarantees that there will be generalization)

## Discriminative Methods

Nearest Neighbor

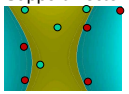

Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005
…

Neural Networks



LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001
…

Conditional Random Fields

McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003
…

## Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

§ Bayes's rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

## Object categorization: the statistical viewpoint

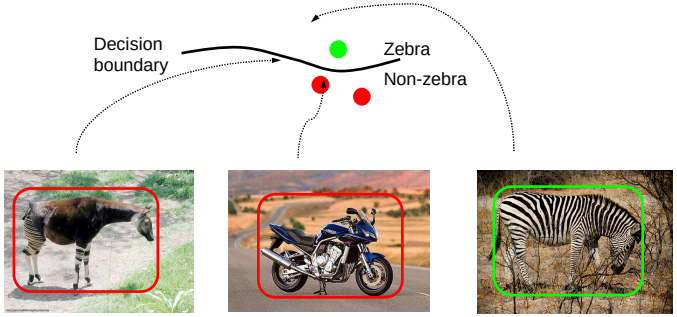
$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

§ **Discriminative methods model posterior**

§ **Generative methods model likelihood and prior**

## Discriminative

§ Direct modeling of $\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$




## Generative

§ Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$

| | $p(image \mid zebra)$ | $p(image \mid no\ zebra)$ |
|---|---|---|
|  | Low | Middle |
|  | High | Middle Low |

## Three main issues

- Representation
  - How to represent an object category
- Learning
  - How to form the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

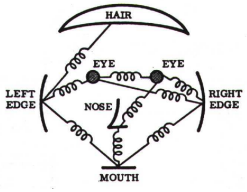
## Approach 2: Generative Methods using Bag of Words Models

- An image is represented by a collection of "visual words" and their corresponding counts given a universal dictionary
- Object categories are modeled by the distributions of these visual words
- Although "bag of words" models can use both generative and discriminative approaches, here we will focus on generative models

## Approach 3: Generative Methods using Part-Based Models

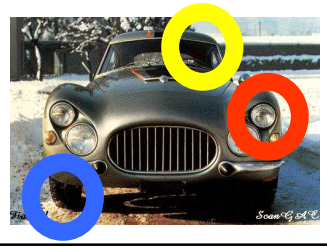
- An object in an image is represented by a collection of parts, characterized by both their visual appearances and locations
- Object categories are modeled by the appearance and spatial distributions of these characteristic parts
- Issues for such models include efficient methods for finding correspondences between the object and the scene

## Model: Constellation of Parts




Fischler & Elschlager, 1973

Yuille, '91
Brunelli & Poggio, '93
Lades, v.d. Malsburg et al. '93
Cootes, Lanitis, Taylor et al. '95
Amit & Geman, '95, '99
Perona et al. '95, '96, '98, '00



## Representation

- Object as set of parts
  - Generative representation
- Model:
  - Relative locations between parts
  - Appearance of part
- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

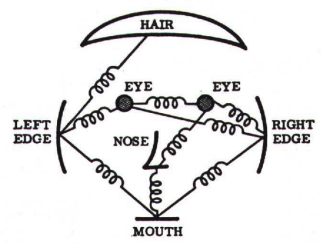
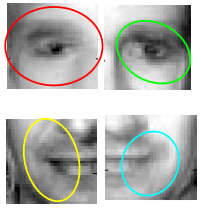



Figure from [Fischler73]

## Model Structure

- Model **shape** using Gaussian distribution on image location between parts and scale of each part
- Model **appearance** as patches of pixel intensities
- Represent object class as graph of $P$ image patches with parameters $\theta$



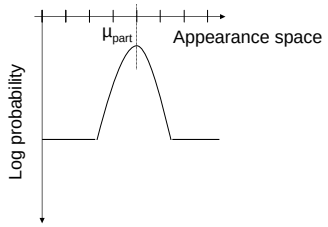
## Sparse representation

- + Computationally tractable ($10^5$ pixels  $10^1$ -- $10^2$ parts)
- + Generative representation of class
- + Avoid modeling global variability
- + Success in specific object recognition

- - Throws away most image information
- - Parts need to be distinctive to separate from other classes

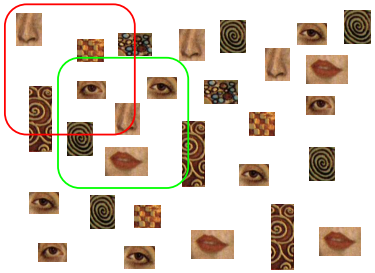
## Representation of Occlusion

- Explicit
  - Additional match of each part to missing state
- Implicit
  - Truncated minimum probability of appearance



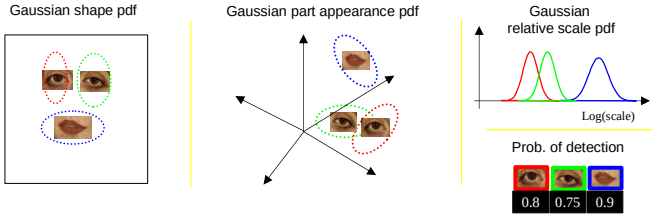
# Representation of Background Clutter

- Explicit model
  - Generative model for clutter as well as foreground object
- Use a sub-window
  - At correct position, no clutter is present

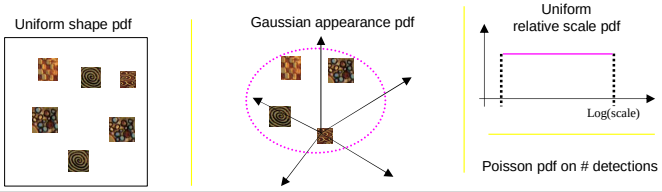


# Generative probabilistic model

Object model

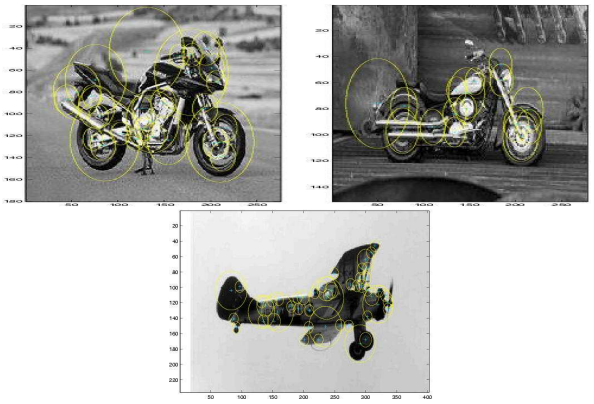



Background clutter model

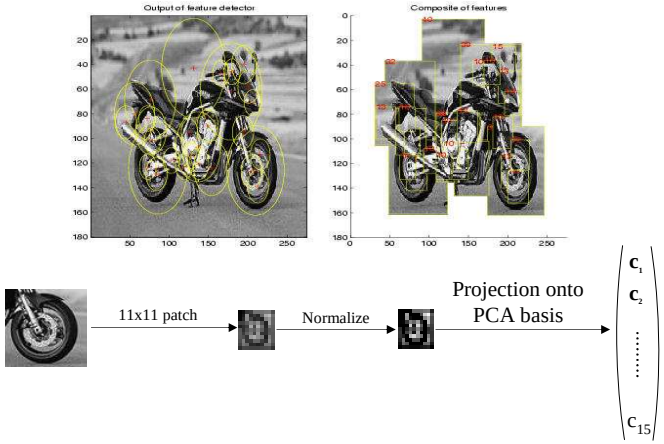



# Interest Operator

Kadir and Brady's interest operator
Finds maxima in entropy over scale and location



# Representation of appearance



## Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

## Model Structure

- Assume prior ratio is known or learned
- Find values for parameters $\theta$ that maximizes the likelihood ratio

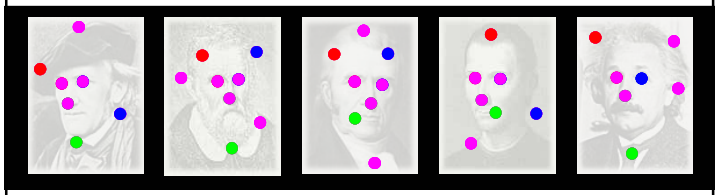
$$p(X, S, A \mid \theta) = \sum_{h \in H} p(X, S, A, h \mid \theta)$$

- H is the set of all valid correspondences of image features to model parts, so $|H| = O(N^P)$
- Factor the likelihood to simplify computation (using Chain Rule)

## Learning

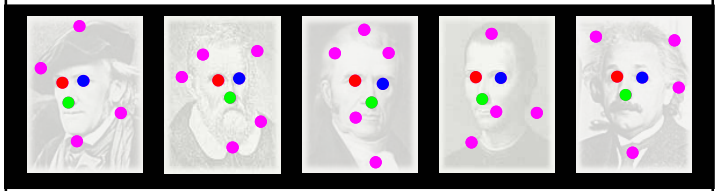
## Learning procedure

- Find regions & their location & appearance
- Initialize model parameters
- Use EM algorithm and iterate to convergence:
  - E-step: Compute assignments for which regions belong to which part (red, green and blue dots)
  - M-step: Update model parameters
- Try to maximize likelihood – consistency in shape & appearance



## Learning using EM

- Task: Estimation of model parameters
- Chicken and Egg type problem, since we initially know neither:
  - Model parameters
  - Assignment of regions to parts
- Let the assignments be a hidden variable and use EM algorithm to learn them and the model parameters



## Recognition

- For each of $P$ parts, run template over all locations in image
- Detect local maxima, giving possible locations of each part
- Given learned model, find maximum likelihood ratio of $p(\mathbf{X},\mathbf{S},\mathbf{A}|\theta)/p(\mathbf{X},\mathbf{S},\mathbf{A}|\theta_{bg})$ for all possible correspondences – $O(N^2P)$ where $N$ = number of locations of each part in image
- If greater than a threshold, signify object detected

## Experimental Procedure

Two series of experiments:

1. Scale variant (using pre-scaled images)
2. Scale invariant

**$P$ = 6-7**
**$N$ = 20-30**
**20-30 parameters/part**
**10-15 PCA features**

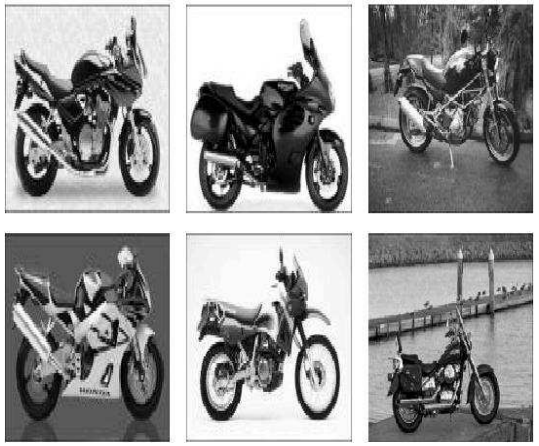
**Datasets**:

- Motorbikes, Faces, Spotted cats, Airplanes, Cars from behind and side
- **200 - 800 images**

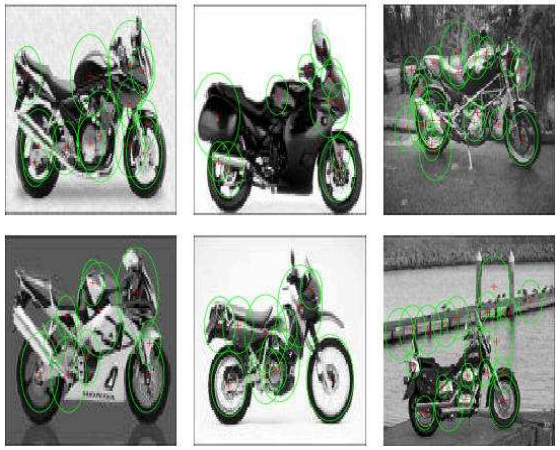
**Training**
- 50% images
- No identifcation of object within image

**Testing**
- 50% images
- Simple object present/absent test
- ROC equal error rate computed, using background set of images

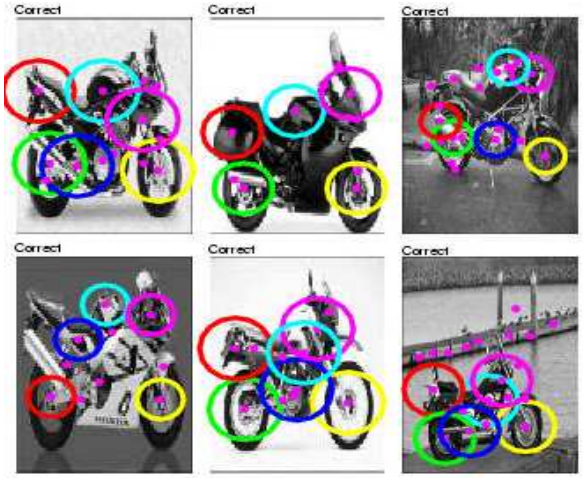
## Motorbikes: Input Images



## Motorbikes: Features Detected



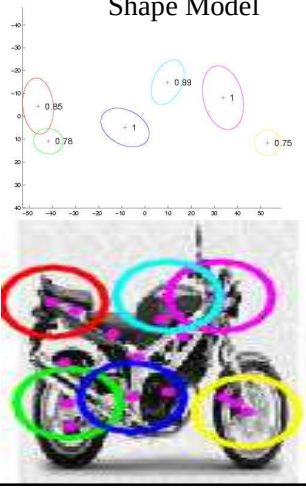
## Motorbikes: Max Likelihood Result



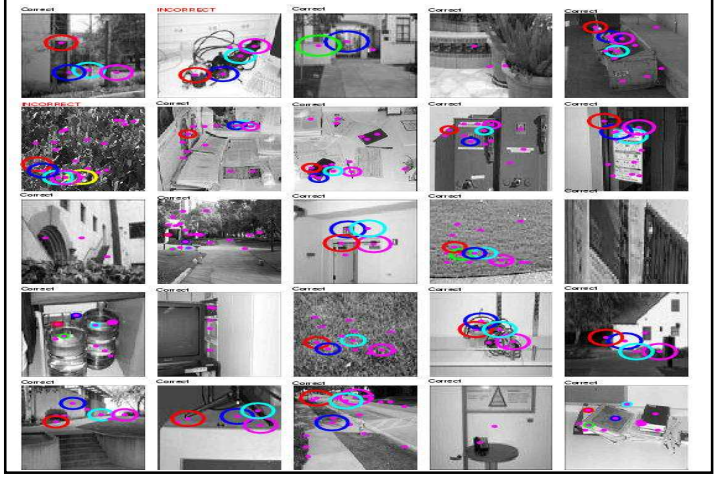

## Motorbikes

Equal error rate: 7.5%


Shape Model

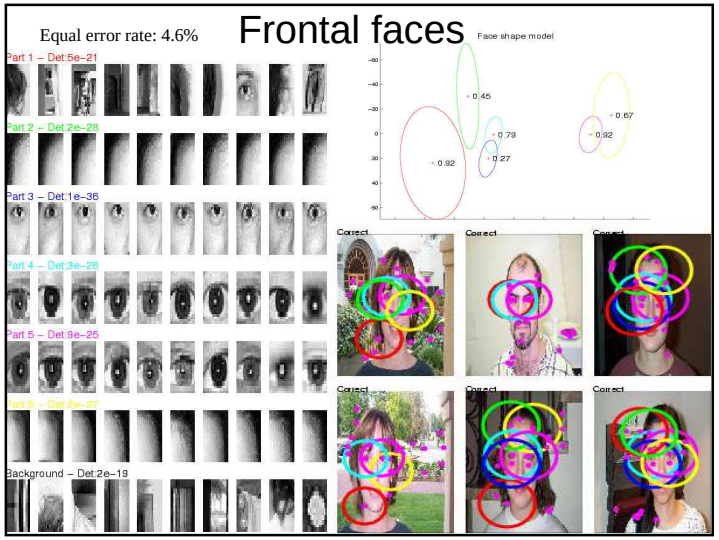



## Background images



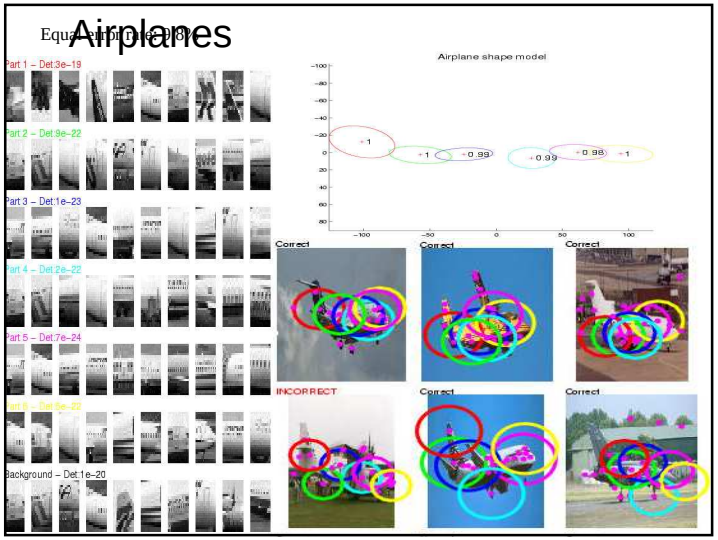
## Frontal faces

Equal error rate: 4.6%



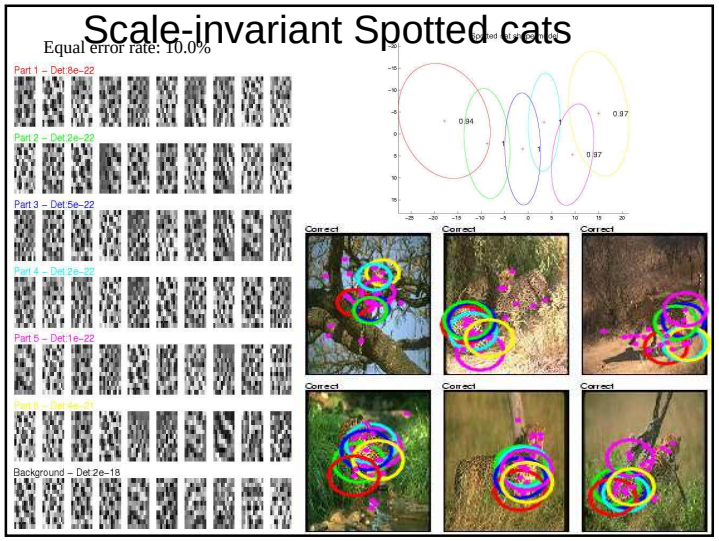

## Airplanes

Equal error rate: 9.8%



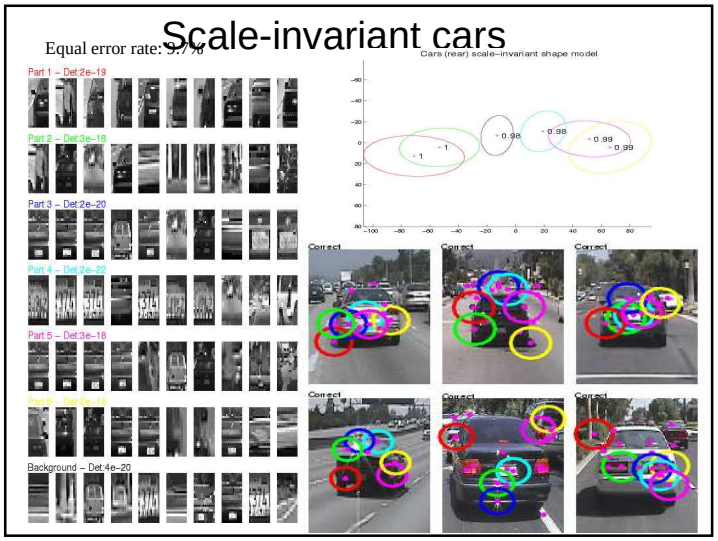

## Scale-invariant Spotted cats

Equal error rate: 10.0%

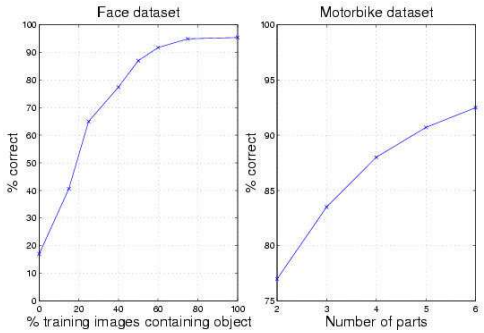



## Scale-invariant cars

Equal error rate: 9.7%



# Robustness of algorithm




# ROC equal error rates

Pre-scaled data (identical settings):

| | | | Model | | | |
|---|---|---|---|---|---|---|
| Dataset | Total size of dataset | ~ Object width (pixels) | Motorbikes | Faces | Airplanes | Spotted Cats |
| Motorbikes | 800 | 200 | 92.5 | 50 | 51 | 56 |
| Faces | 435 | 300 | 33 | 96.4 | 32 | 32 |
| Airplanes | 800 | 300 | 64 | 63 | 90.2 | 53 |
| Spotted Cats | 200 | 80 | 48 | 44 | 51 | 90 .0 |

Scale-invariant learning and recognition:

| Dataset | Total size of dataset | Object size range (pixels) | Pre-scaled performance | Unscaled performance |
|---|---|---|---|---|
| Motorbikes | 800 | 200-480 | 95.0 | 93.3 |
| Airplanes | 800 | 200-500 | 94.0 | 93.0 |
| Cars (Rear) | 800 | 100-550 | 84.8 | 90.3 |

# Probabilistic Parts and Structure Models Summary

- Correspondence problem
- Efficient methods for large # parts and # positions in image
- Challenge to get representation with desired invariance
- Minimal supervision

- Future directions:
  - Multiple views
  - Approaches to learning
  - Multiple category training