Object Recognition by Parts

- Object recognition started with line segments.

  - Roberts recognized objects from line segments and junctions.

  - This led to systems that extracted linear features.

  .

  - CAD-model-based vision works well for industrial.

- An “appearance-based approach” was first developed for face recognition and later generalized up to a point.

- The new interest operators have led to a new kind of recognition by “parts” that can handle a variety of objects that were previously difficult or impossible.

Object Class Recognition by Unsupervised Scale-Invariant Learning

R. Fergus, P. Perona, and A. Zisserman
Oxford University and Caltech

CVPR 2003

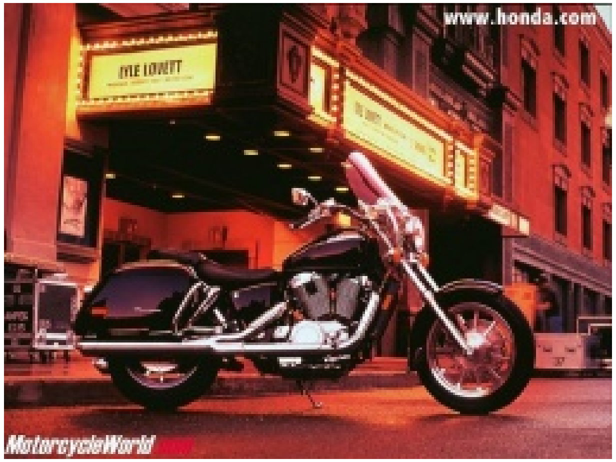
won the best student paper award

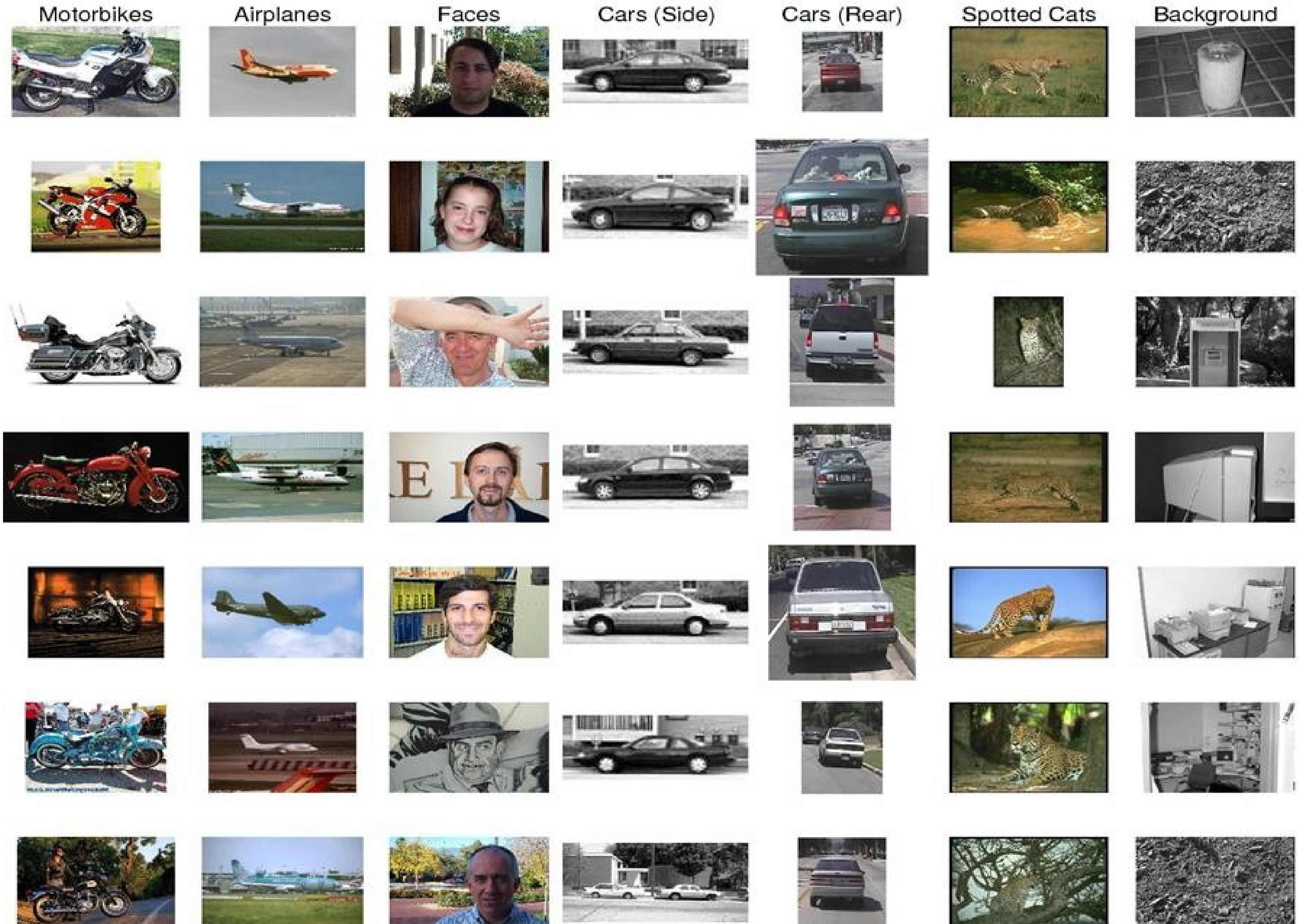
CVPR 2013

won the best 10-year award

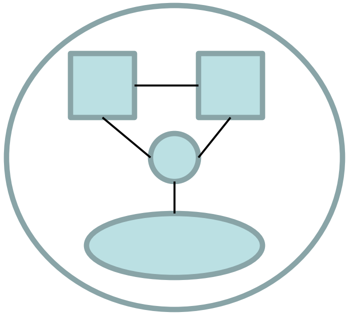
Goal:

- Enable Computers to Recognize Different Categories of Objects in Images.




Motorbikes
Airplanes
Faces
Cars (Side)
Cars (Rear)
Spotted Cats
Background

Approach

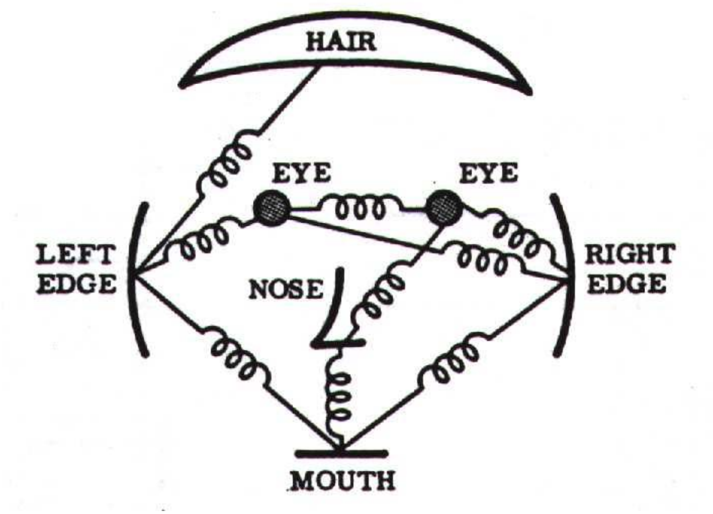


- An object is a constellation of parts (from Burl, Weber and Perona, 1998).
- The parts are detected by an interest operator (Kadir's).
- The parts can be recognized by appearance.
- Objects may vary greatly in scale.
- The constellation of parts for a given object is learned from training images

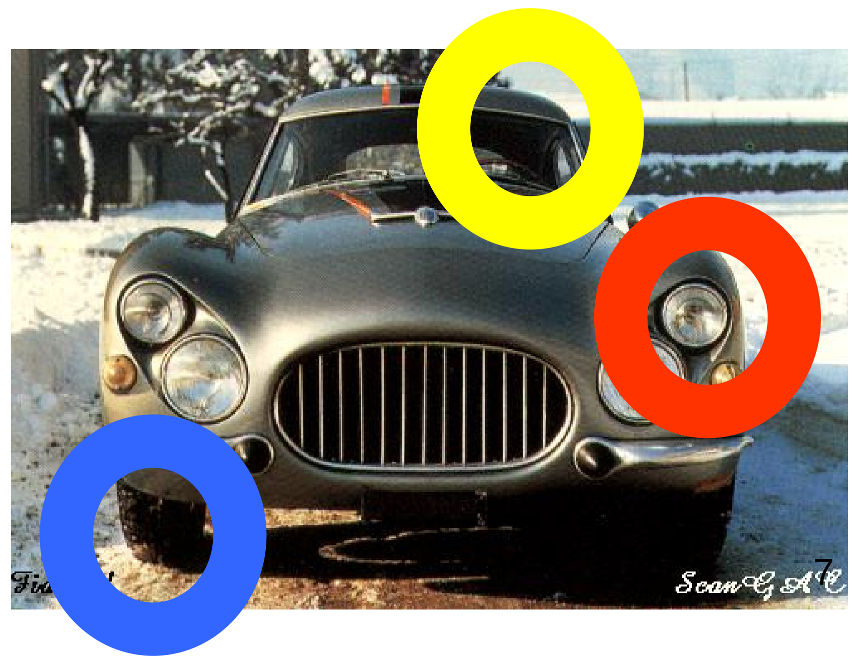
Components

- Model
  – Generative Probabilistic Model including Location, Scale, and Appearance of Parts
- Learning
  – Estimate Parameters Via EM Algorithm
- Recognition
  – Evaluate Image Using Model and Threshold

Model: Constellation Of Parts



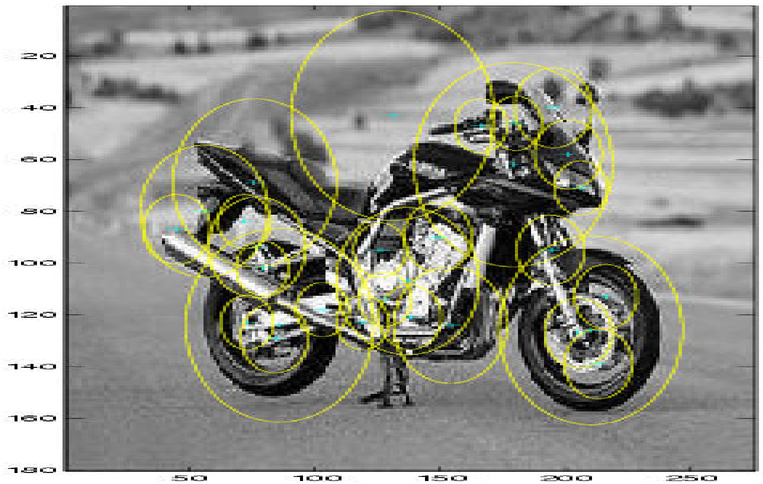
Fischler & Elschlager, 1973

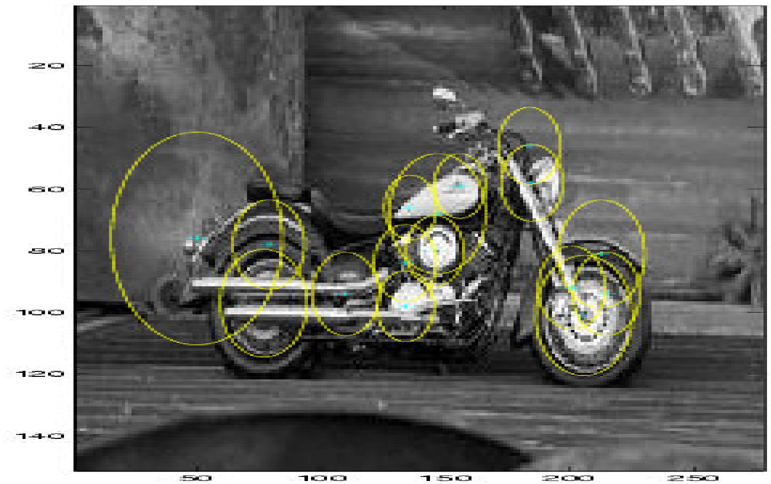


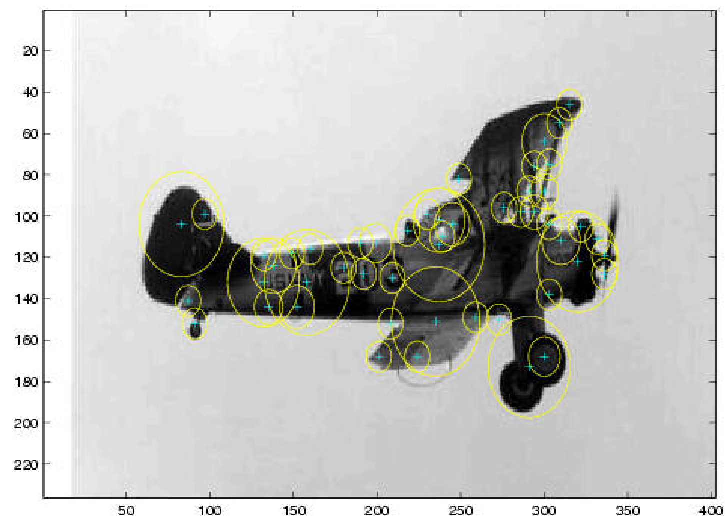
Yuille, '91
Brunelli & Poggio, '93
Lades, v.d. Malsburg et al. '93
Cootes, Lanitis, Taylor et al. '95
Amit & Geman, '95, '99
Perona et al. '95, '96, '98, '00

Parts Selected by Interest Operator

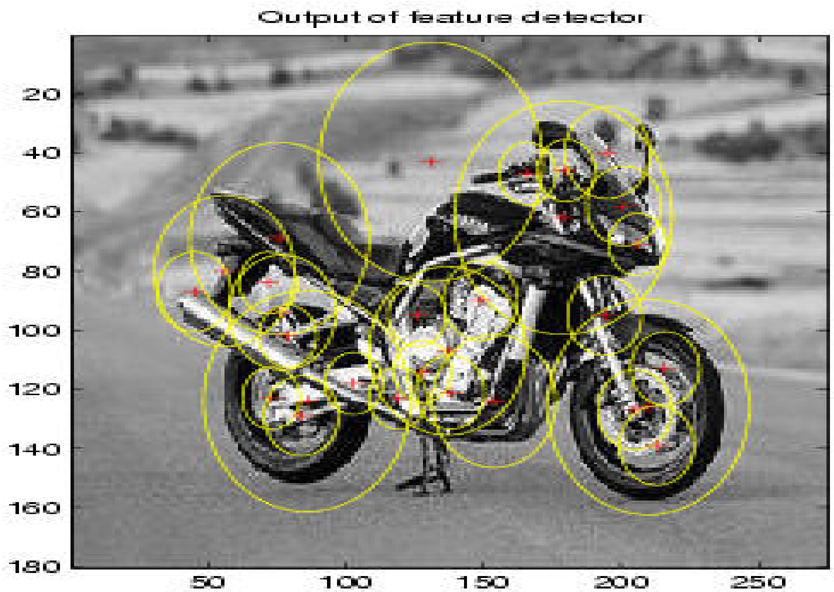
Kadir and Brady's Interest Operator.
Finds Maxima in Entropy Over Scale and Location

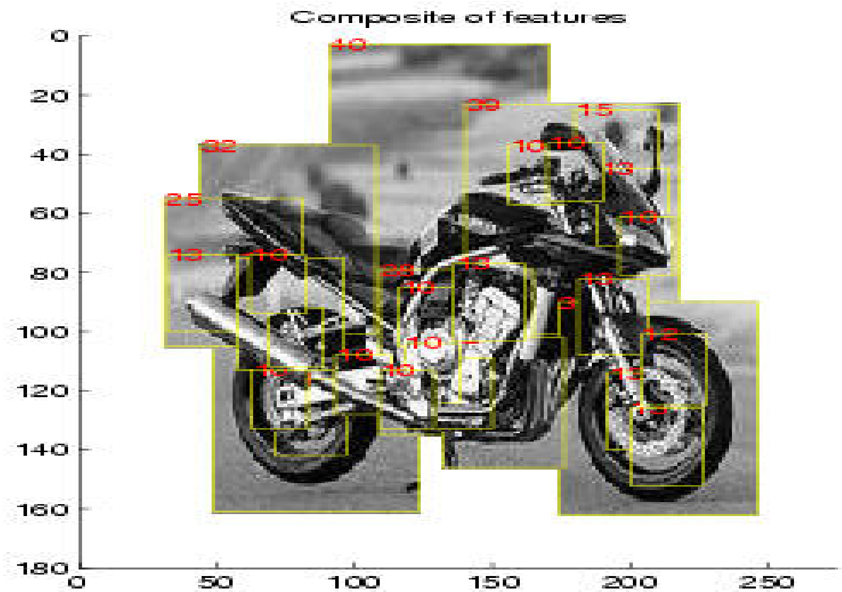




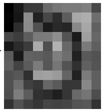


Representation of Appearance

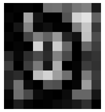




11x11 patch →



Normalize →



Projection onto PCA basis →

$$\begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ c_{15} \end{pmatrix}$$

121 dimensions was too big, so they used PCA to reduce to 10-15.

Learning a Model

- An object class is represented by a generative model with P parts and a set of parameters θ.

- Once the model has been learned, a decision procedure must determine if a new image contains an instance of the object class or not.

- Suppose the new image has N interesting features with locations X, scales S and appearances A.

Probabilistic Model

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

$$\sum_{\mathbf{h} \in H} \underbrace{p(\mathbf{A}|\mathbf{X}, \mathbf{S}, \mathbf{h}, \theta)}_{Appearance} \underbrace{p(\mathbf{X}|\mathbf{S}, \mathbf{h}, \theta)}_{Shape} \underbrace{p(\mathbf{S}|\mathbf{h}, \theta)}_{Rel.\ Scale} \underbrace{p(\mathbf{h}|\theta)}_{Other}$$

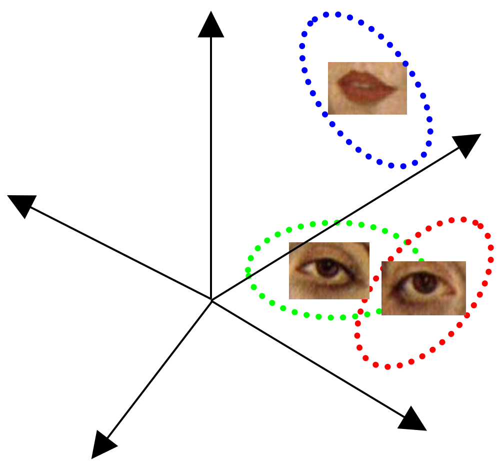
- X is a description of the shape of the object (in terms of locations of parts)
- S is a description of the scale of the object
- A is a description of the appearance of the object
- θ is the (maximum likelihood value of) the parameters of the object
- h is a hypothesis: a set of parts in the image that might be the parts of the object
- H is the set of all possible hypotheses for that object in that image.
- For N features in the image and P parts in the object, its size is $O(N^P)$

Appearance

The appearance (A) of each part p has a Gaussian density with mean c_p and covariance V_P.

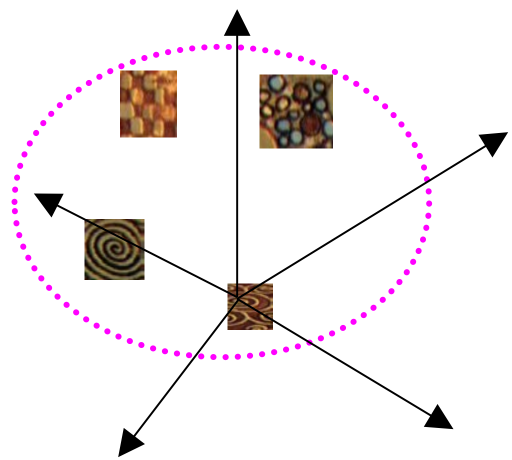
Background model has mean c_{bg} and covariance V_{bg}.

Gaussian Part Appearance PDF



Object

Guausian Appearance PDF

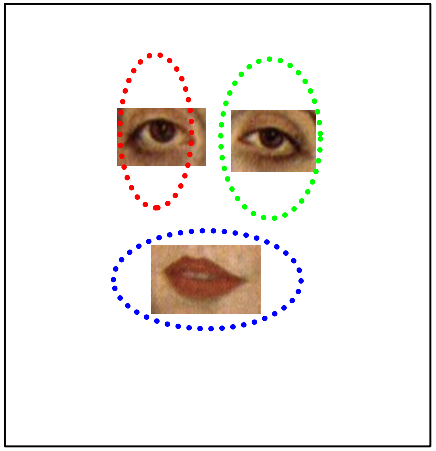


Background

Shape as Location

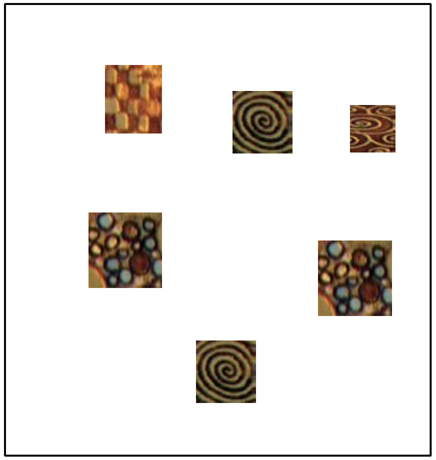
Object shape is represented by a joint Gaussian density of the locations (X) of features within a hypothesis transformed into a scale-invariant space.

Gaussian Shape PDF



Object

Uniform Shape PDF

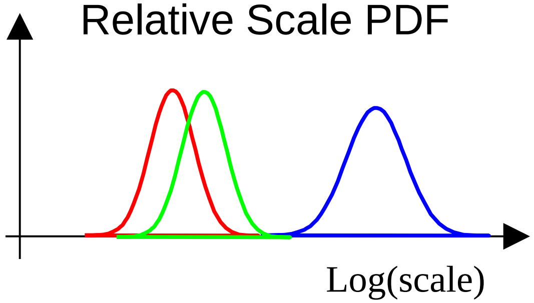


Background

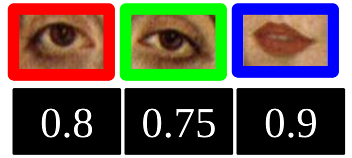
Scale

The relative scale of each part is modeled by a Gaussian density with mean t_p and covariance U_p.

Gaussian
Relative Scale PDF



Prob. of detection



Occlusion and Part Statistics

This was very complicated and turned out to not work well and not be necessary, in both Fergus's work and other subsequent works.

Learning

- Train Model Parameters Using EM:
  - Optimize Parameters
  - Optimize Assignments
  - Repeat Until Convergence

$$\theta = \{\underbrace{\boldsymbol{\mu}, \Sigma}_{\text{location}}, \underbrace{\mathbf{c}, V}_{\text{appearance}}, \underbrace{M, p(\mathbf{d}|\theta)}_{\text{occlusion}}, \underbrace{t, U}_{\text{scale}}\}$$

$$\hat{\theta}_{ML} = arg\,\underset{\theta}{max}\; p(\mathbf{X}, \mathbf{S}, \mathbf{A}|\,\theta)$$

Recognition

Make this likelihood ratio:

$$\begin{aligned} R &= \frac{p(\text{Object}|\mathbf{X},\mathbf{S},\mathbf{A})}{p(\text{No object}|\mathbf{X},\mathbf{S},\mathbf{A})} \\ &= \frac{p(\mathbf{X},\mathbf{S},\mathbf{A}|\text{Object})\; p(\text{Object})}{p(\mathbf{X},\mathbf{S},\mathbf{A}|\text{No object})\; p(\text{No object})} \\ &\approx \frac{p(\mathbf{X},\mathbf{S},\mathbf{A}|\,\theta)\; p(\text{Object})}{p(\mathbf{X},\mathbf{S},\mathbf{A}|\theta_{bg})\; p(\text{No object})} \end{aligned}$$

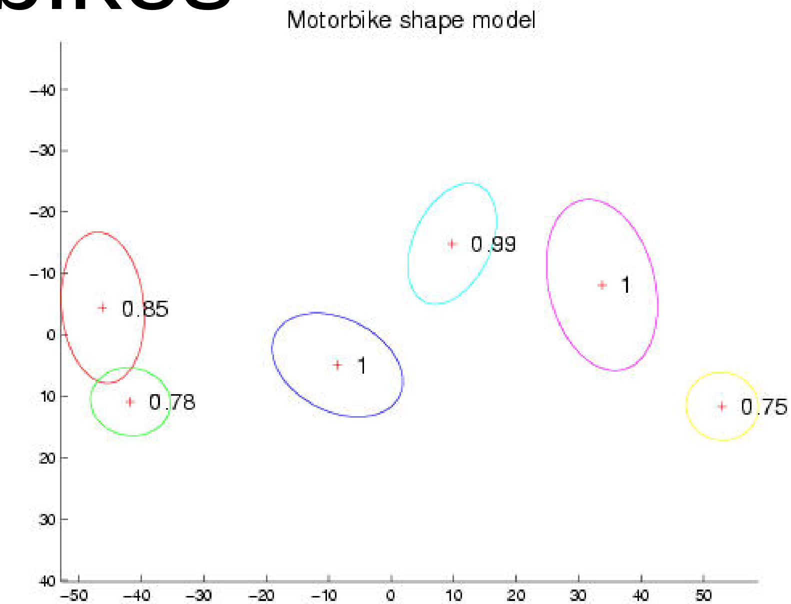
greater than a threshold.

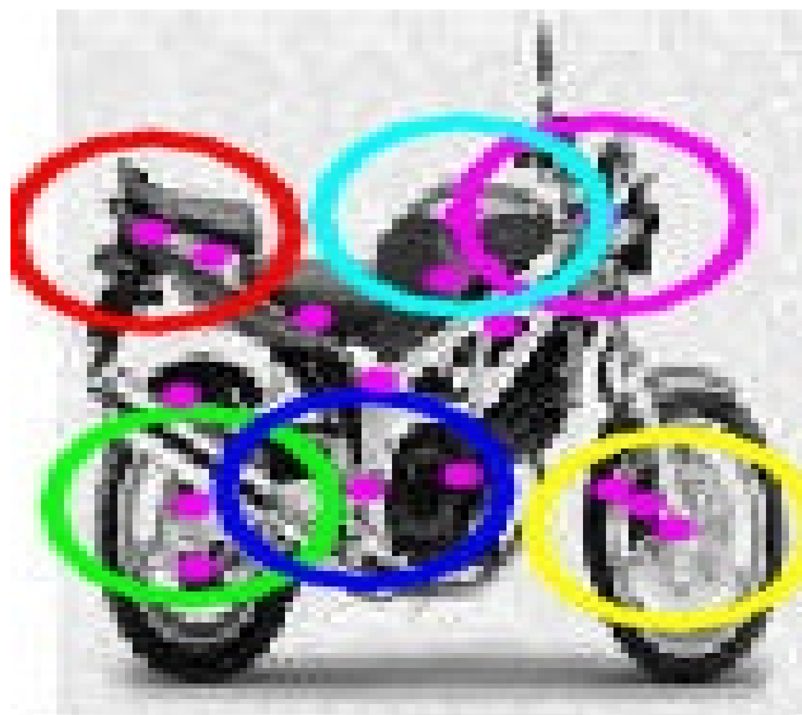
RESULTS

- Initially tested on the Caltech-4 data set
  - motorbikes
  - faces
  - airplanes
  - cars
- Now there is a much bigger data set: the Caltech-101 http://www.vision.caltech.edu/archive.html

Motorbikes

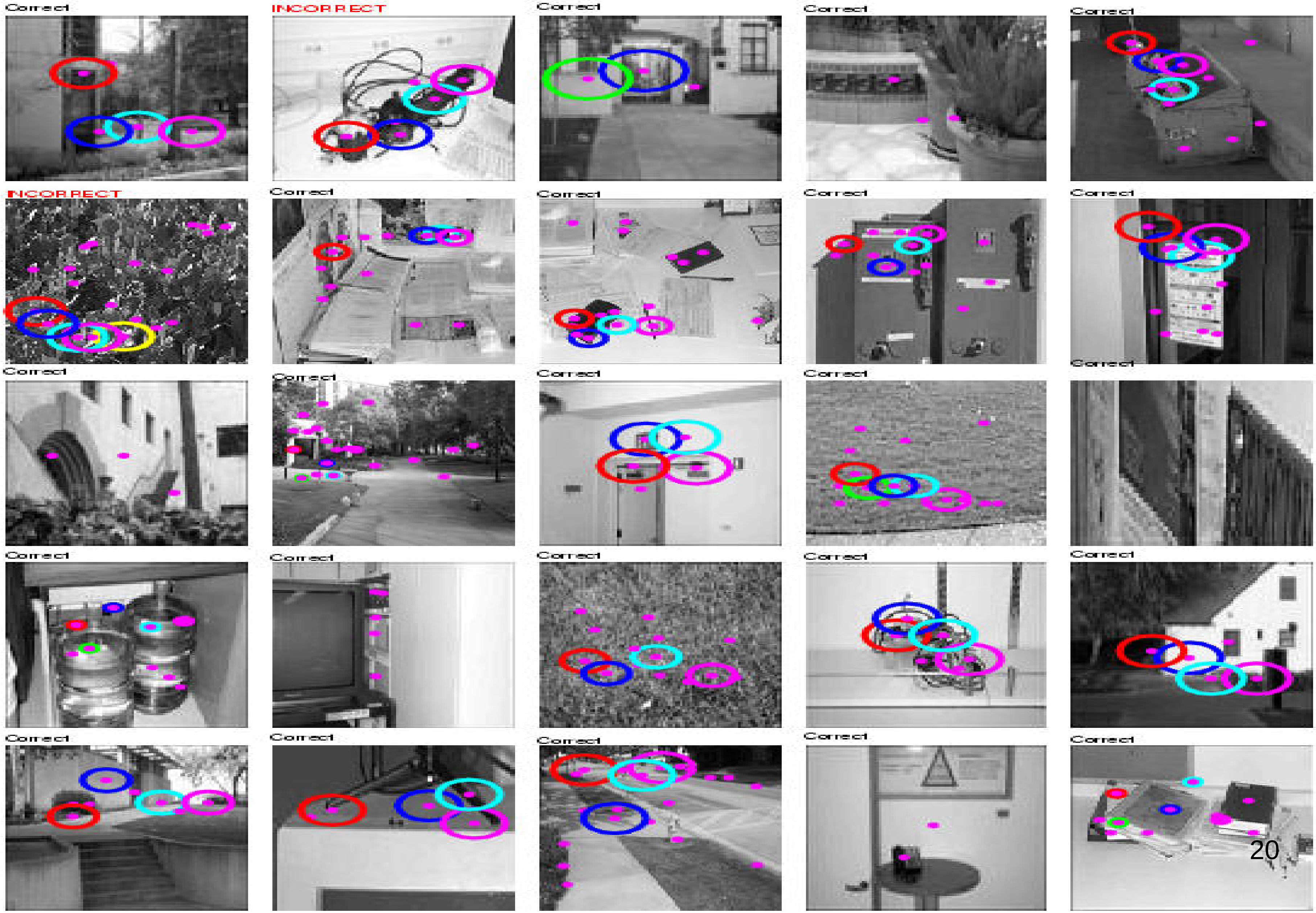
Equal error rate: 7.5%





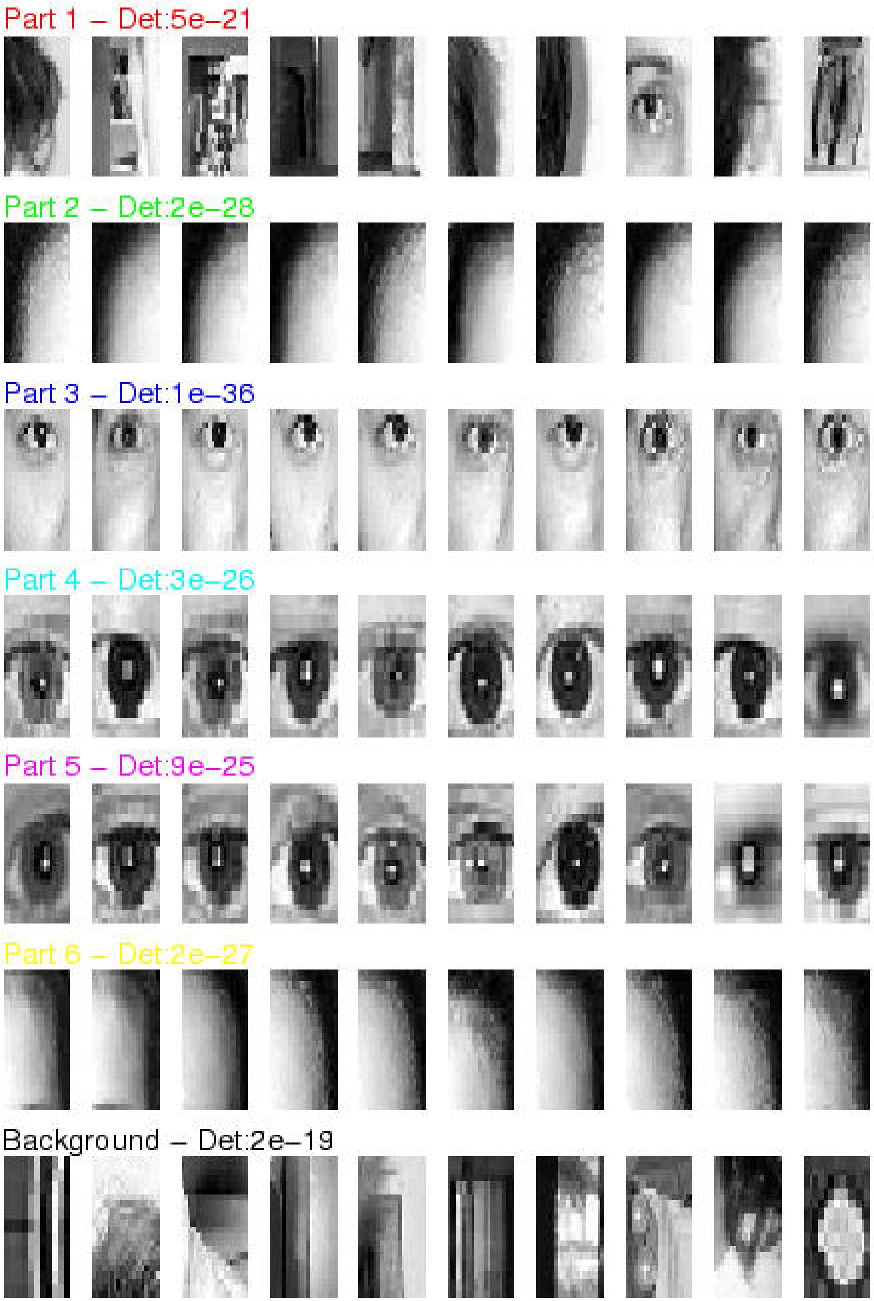
Background Images

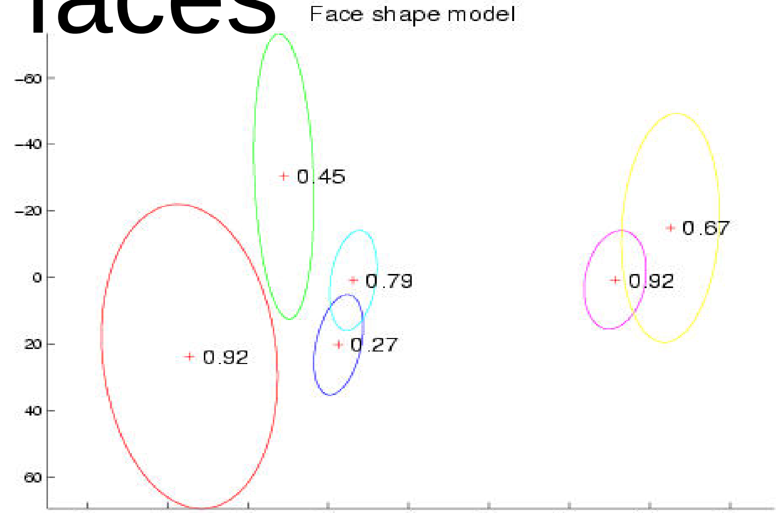
It learns that these are NOT motorbikes.



Frontal faces

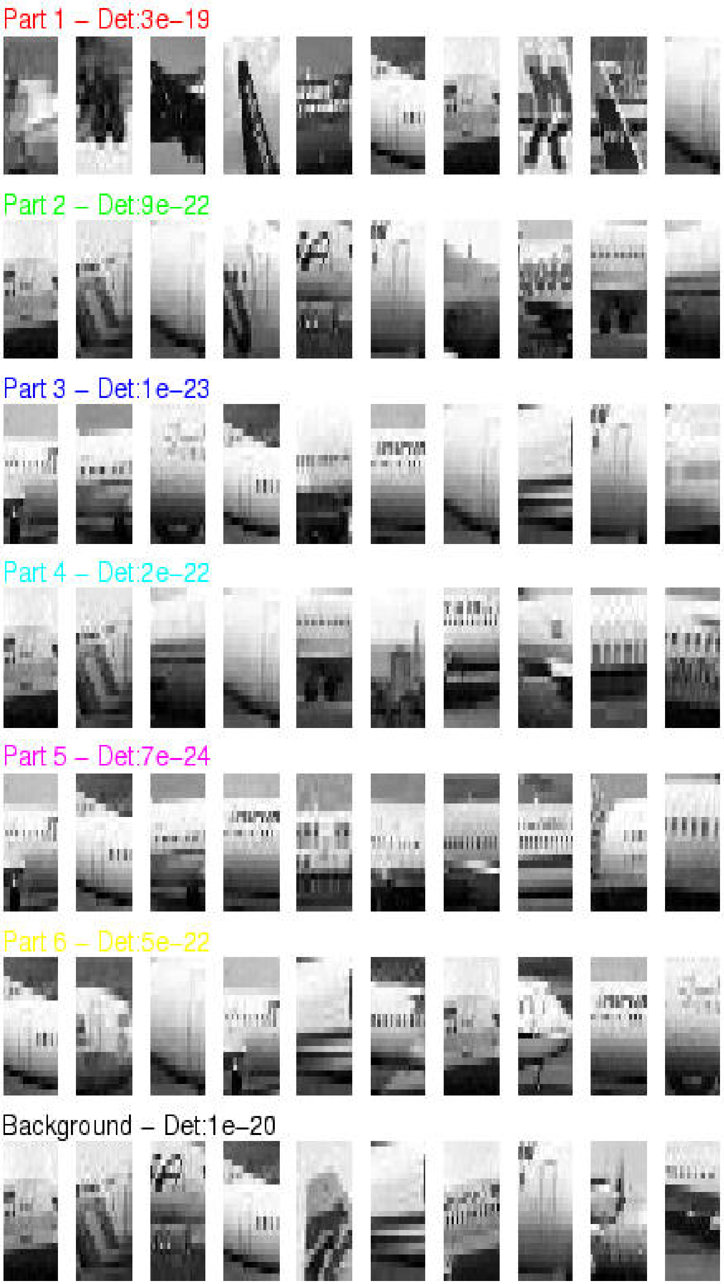
Equal error rate: 4.6%

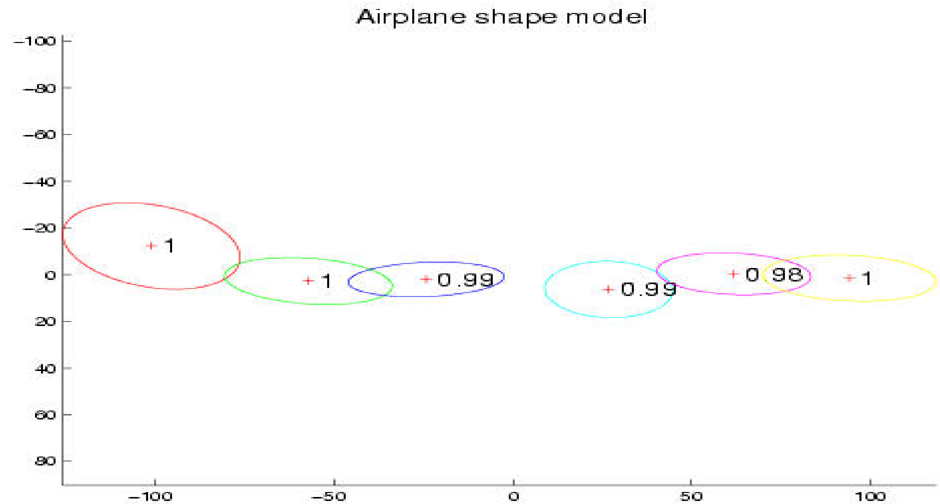


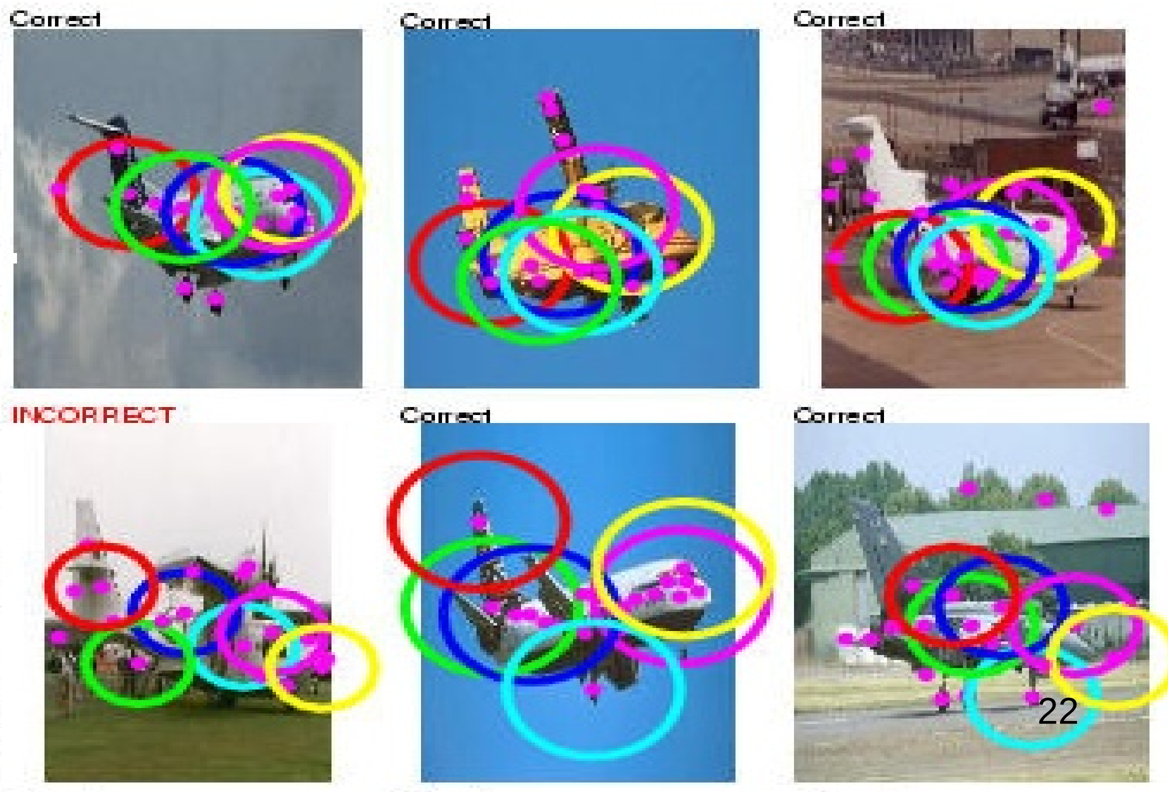


Airplanes

Equal error rate: 9.8%

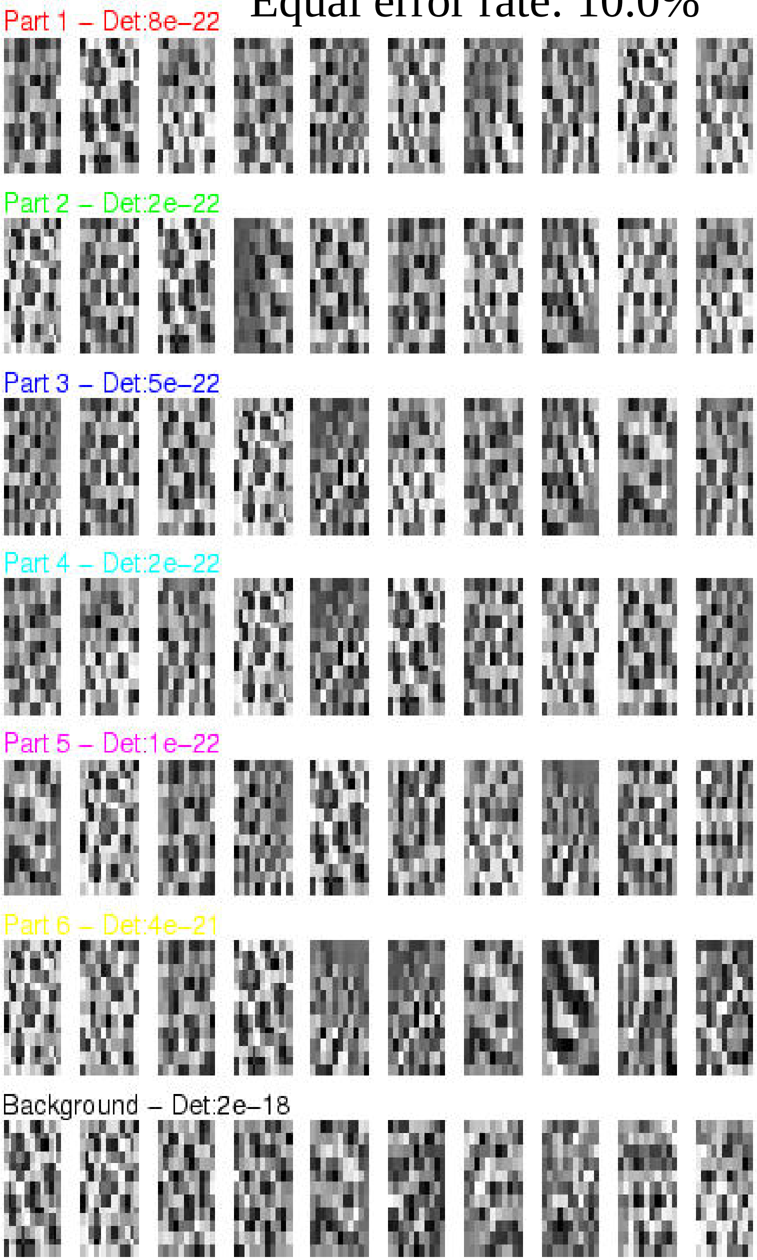


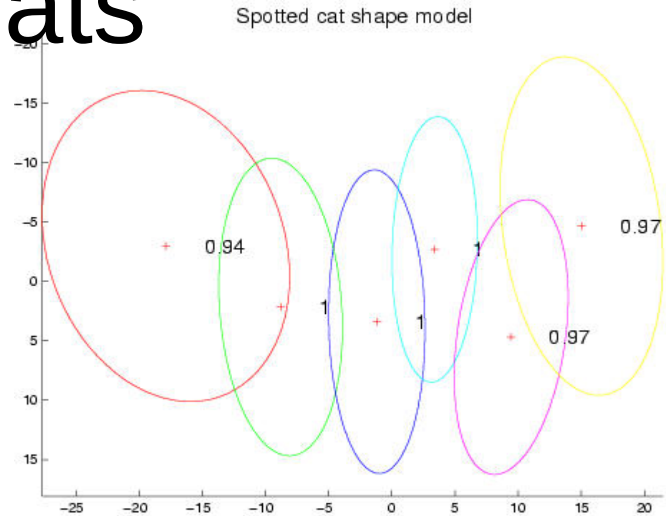


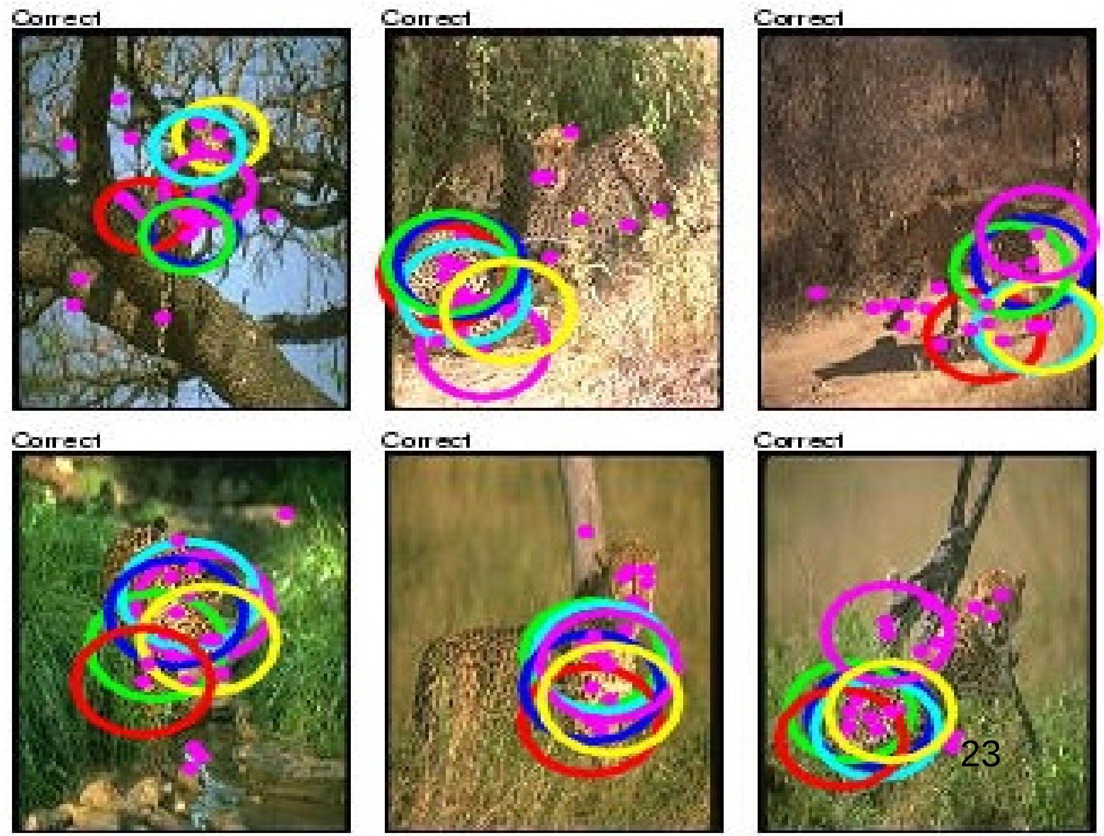


Scale-Invariant Cats

Equal error rate: 10.0%

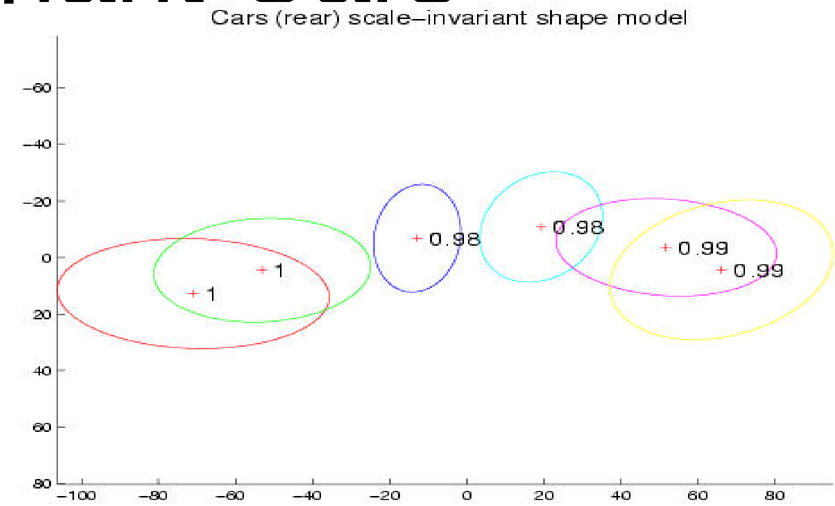






Scale-Invariant Cars

Equal error rate: 9.7%



Accuracy

Initial Pre-Scaled Experiments

Dataset	Ours	Others	Ref.
Motorbikes	92.5	84	[17]
Faces	96.4	94	[19]
Airplanes	90.2	68	[17]
Cars(Side)	88.5	79	[1]