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.

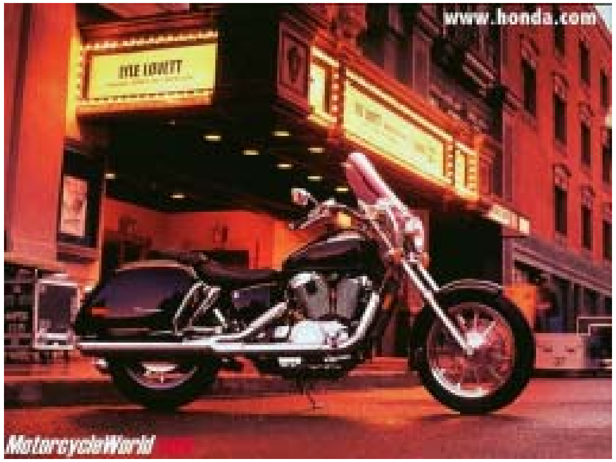
Object Class Recognition by Unsupervised Scale-Invariant Learning

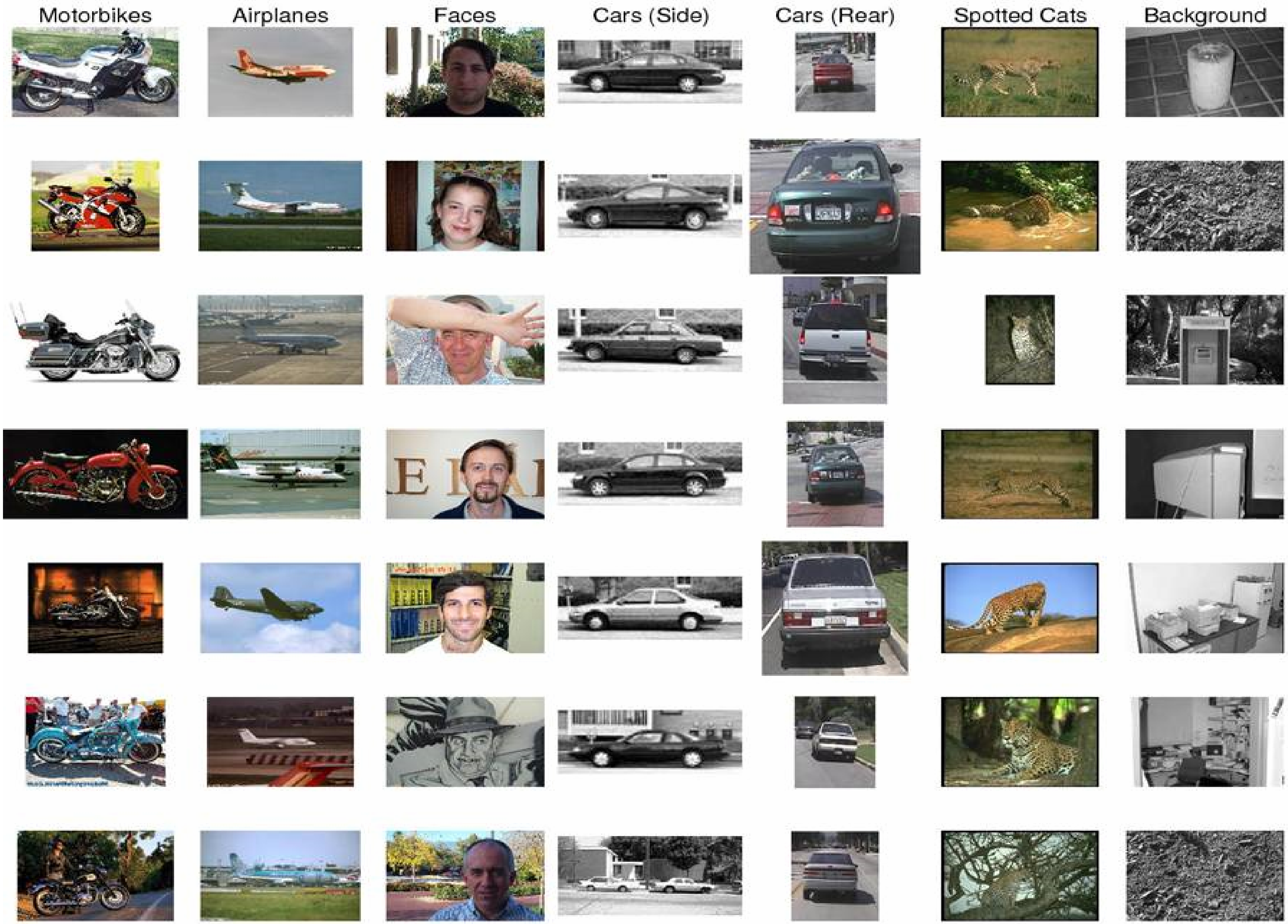
R. Fergus, P. Perona, and A. Zisserman
Oxford University and Caltech

CVPR 2003
won the best student paper award

Goal:

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




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

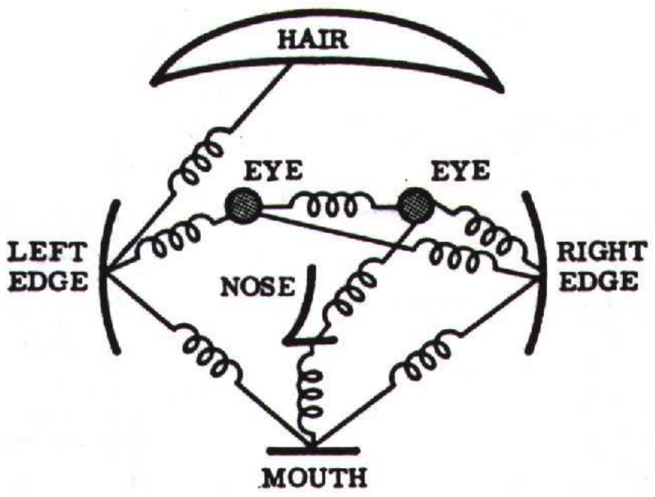
Approach

- An object is a random 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

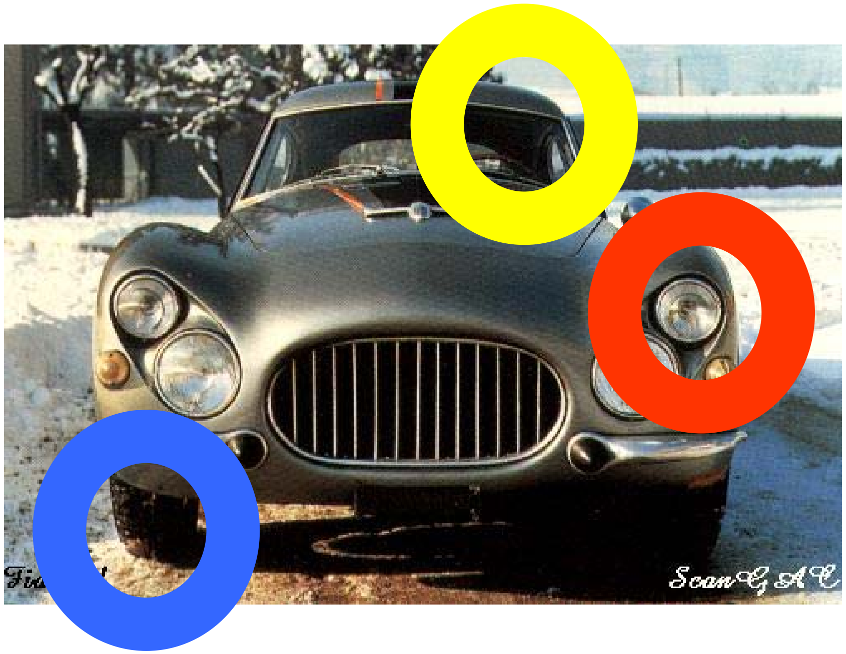
- Model
  - Generative Probabilistic Model including Location, Scale, and Appearance of Parts
- Learning
  - Estimate Parameters Via EM Algorithm
- Recognition
  - Evaluate Image Using Model and Threshold

Model: Constellation Of Parts



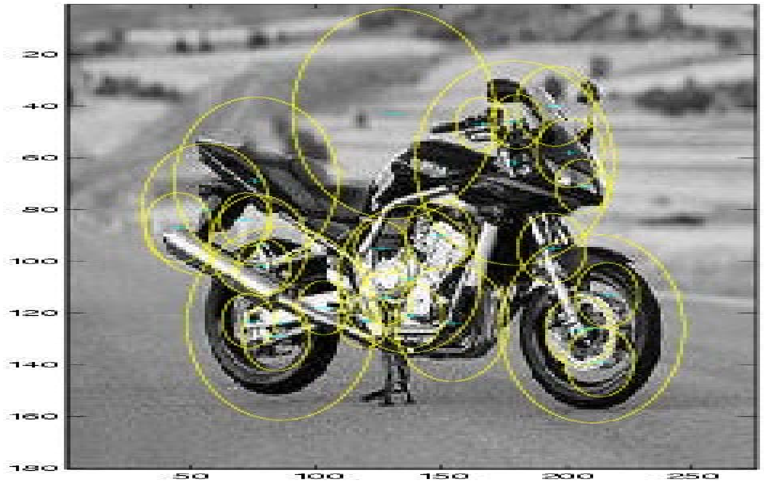
Fischler & Elschlager, 1973

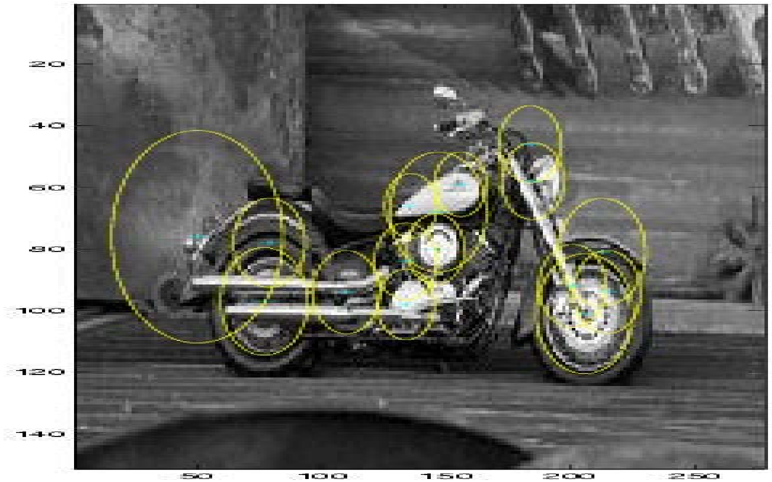
Yuille, 91
Brunelli & Poggio, 93
Lades, v.d. Malsburg et al. 93
Cootes, Lanitis, Taylor et al. 95
Amit & Geman, 95, 99
Perona et al. 95, 96, 98, 00

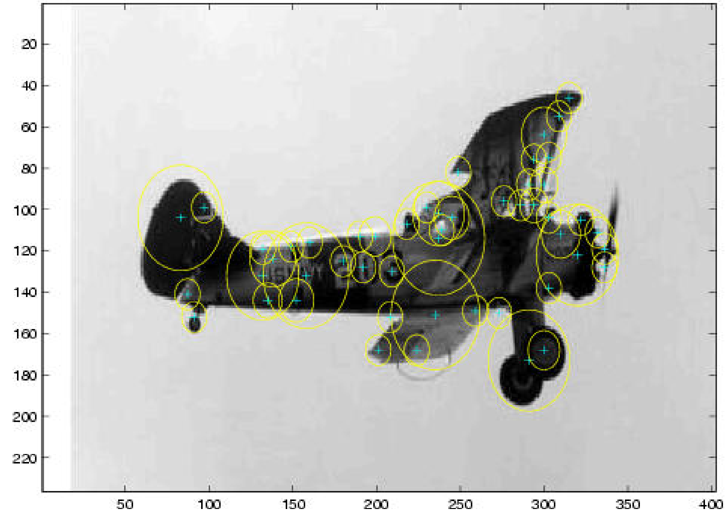


Parts Selected by Interest Operator

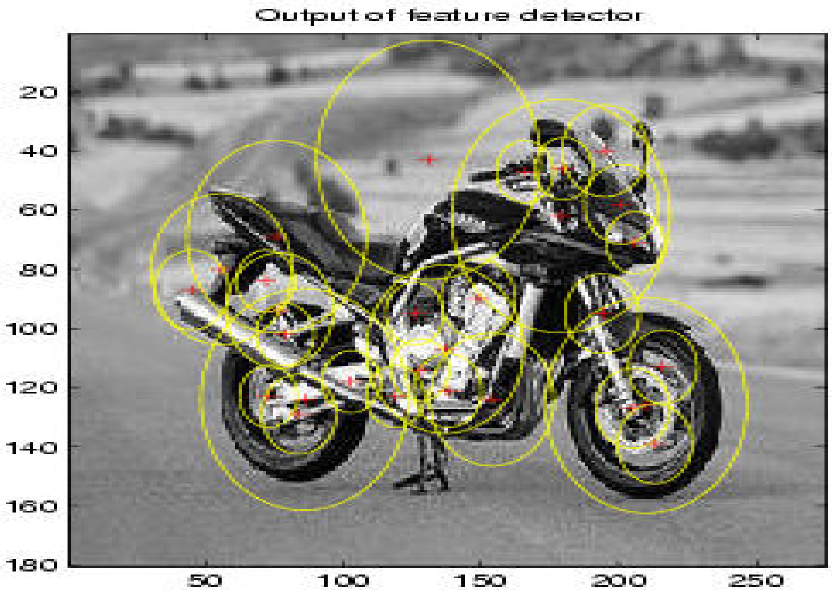
Kadir and Brady's Interest Operator.
Finds Maxima in Entropy Over Scale and Location

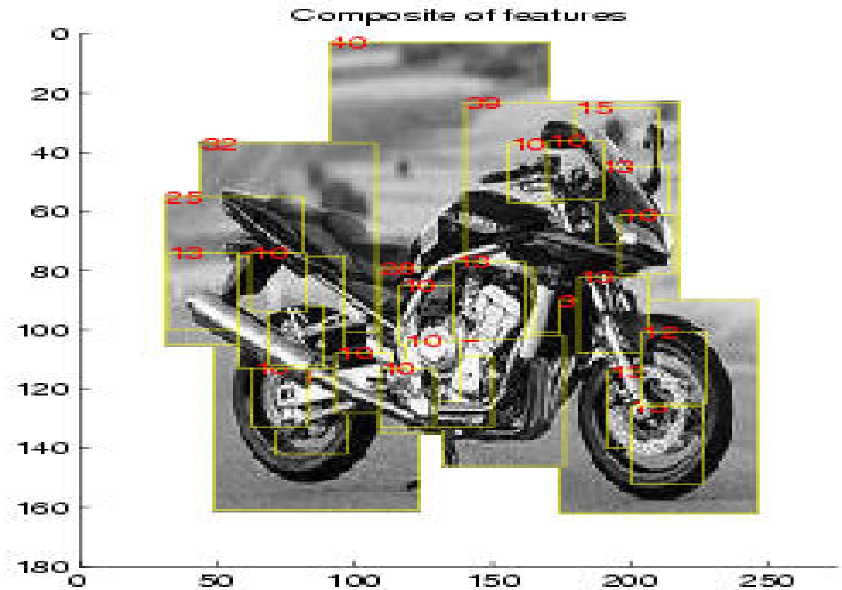




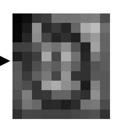


Representation of Appearance

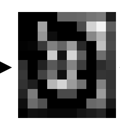




11x11 patch



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.

Generative Probabilistic Model

Top-Down Formulation

Bayesian Decision Rule

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

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

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

R is the likelihood ratio.

θ is the maximum likelihood value of the parameters of the object and θ_{bg} of the background.

h is the hypothesis as to which P of the N features in the image are the object, implemented as a vector of length P with values from 0 to N indicating which image feature corresponds to each object feature.

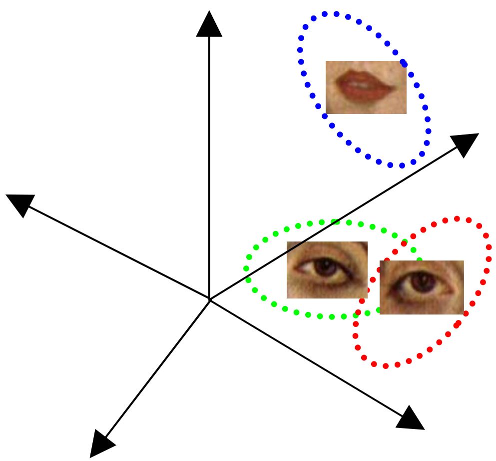
H is the set of all hypotheses; Its size is $O(N^P)$.

Appearance

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

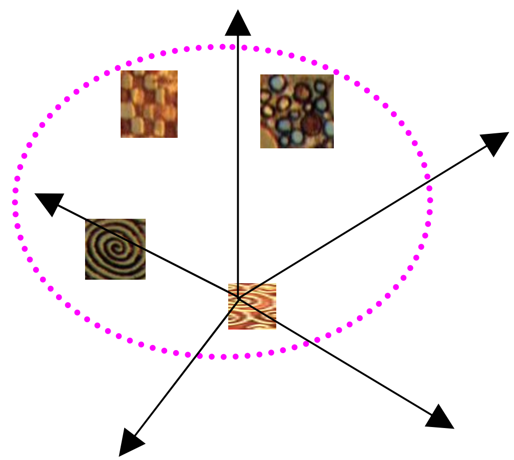
Background model has mean cbg and covariance Vbg.

Gaussian Part Appearance PDF



Object

Guausian Appearance PDF

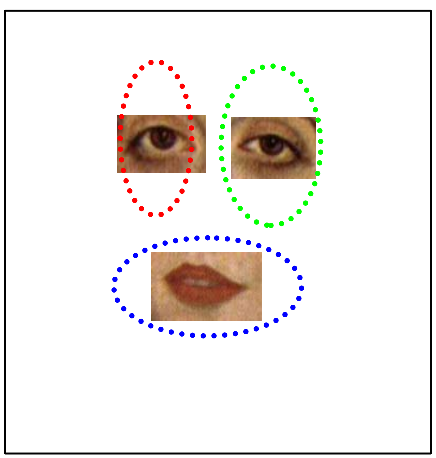


Background

Shape as Location

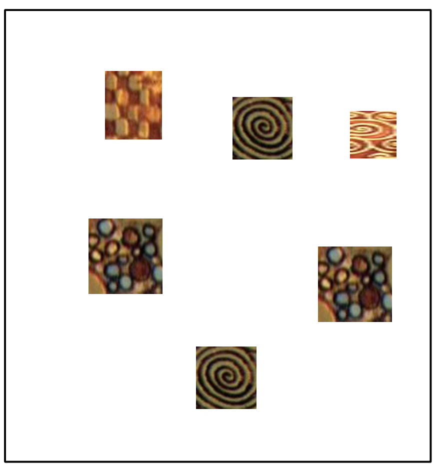
Object shape is represented by a joint Gaussian density of the locations (X) of features within a hypothesis transformed into a scale-invariant space.

Gaussian Shape PDF



Object

Uniform Shape PDF

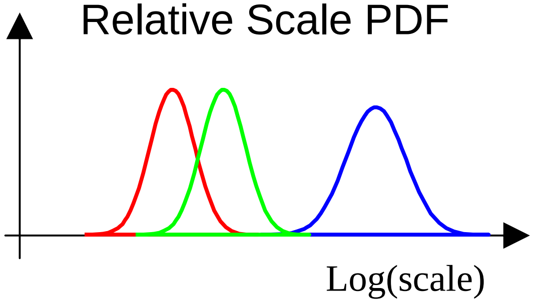


Background

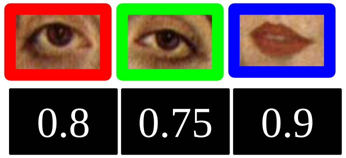
Scale

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

There are 3 terms used:

- First term: Poisson distribution (mean M) models the number of features in the background.
- Second term: (constant) 1/(number of combinations of f_t features out of a total of N_t)
- Third term: gives probability for possible occlusion patterns.

Learning

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

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

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

Recognition

Make This:

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

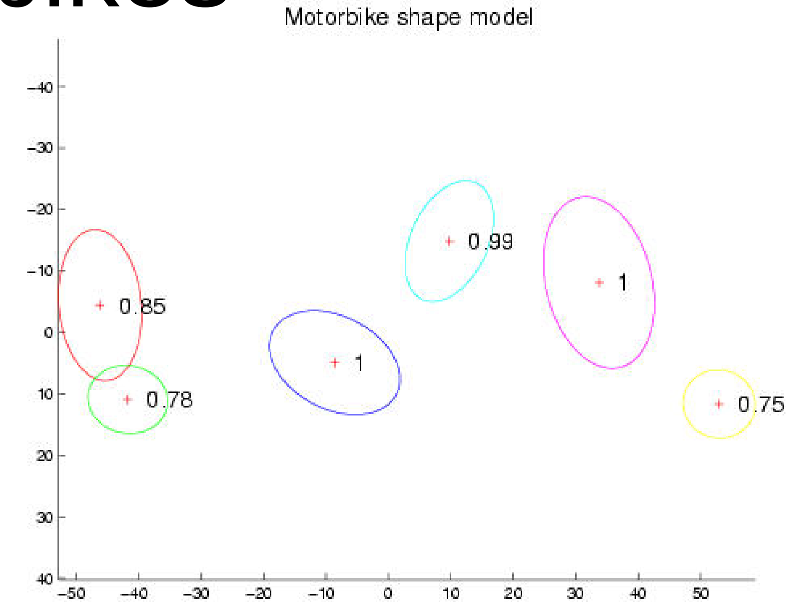
Greater Than Threshold

RESULTS

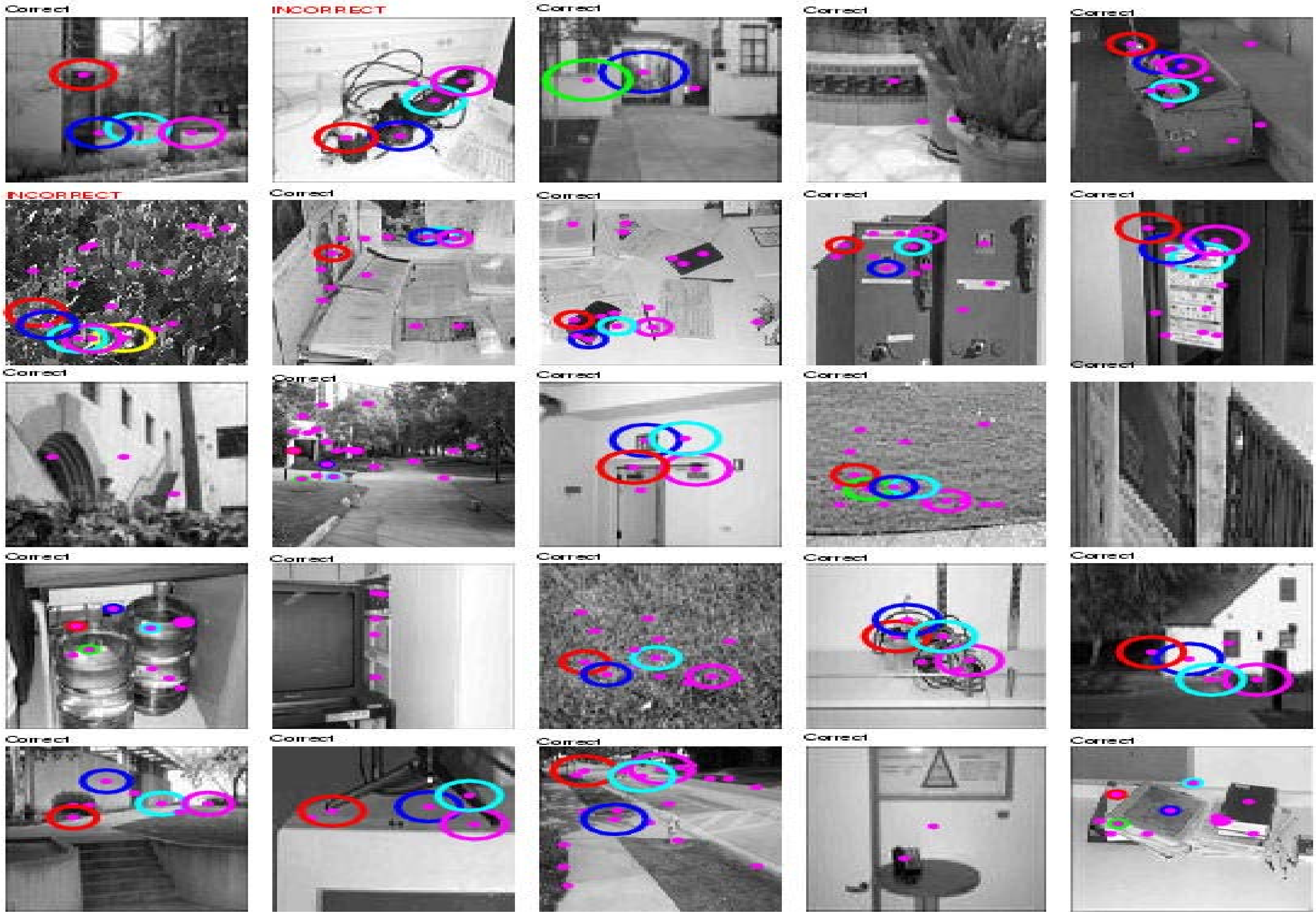
- Initially tested on the Caltech-4 data set
  - motorbikes
  - faces
  - airplanes
  - cars
- Now there is a much bigger data set: the Caltech-101 http://www.vision.caltech.edu/archive.html

Motorbikes

Equal error rate: 7.5%

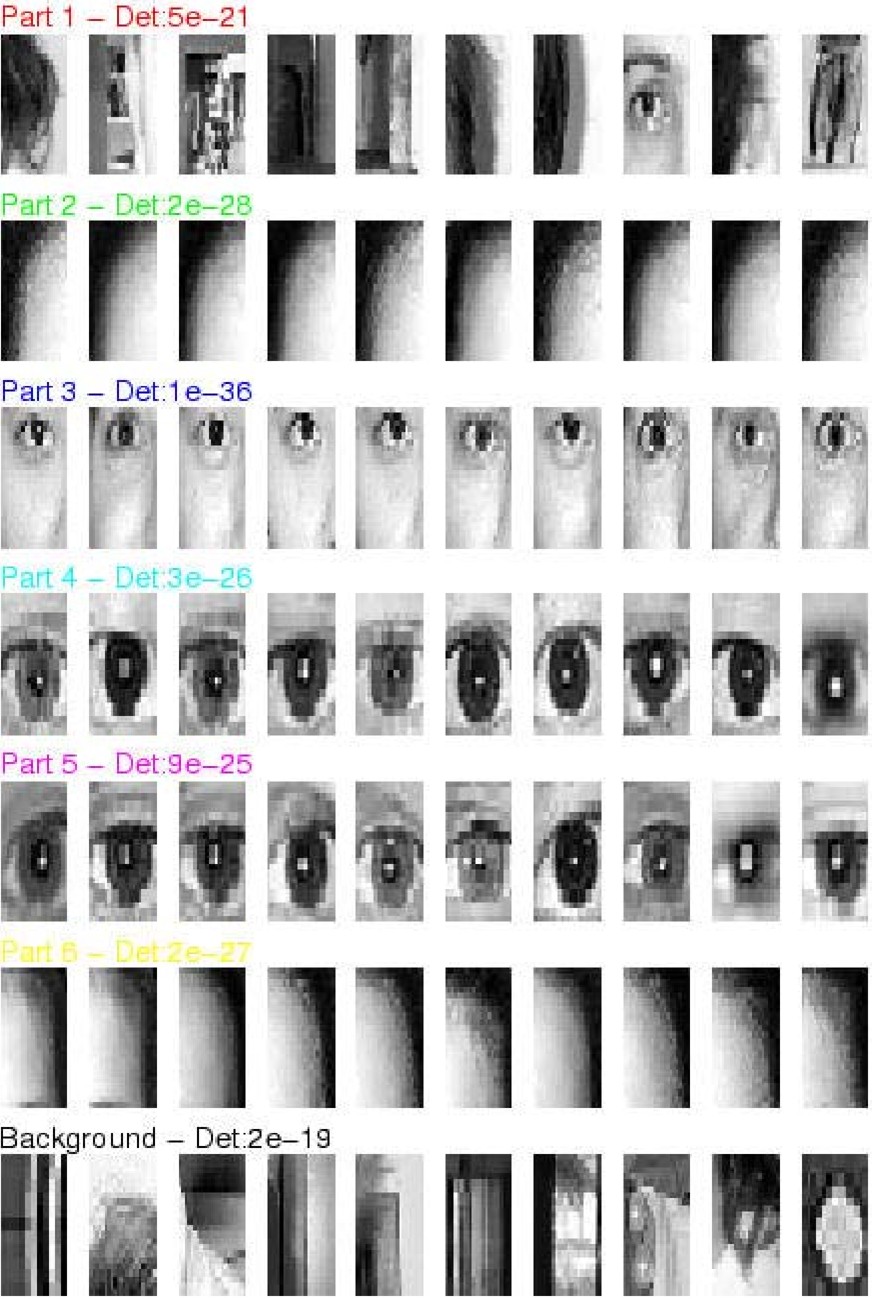


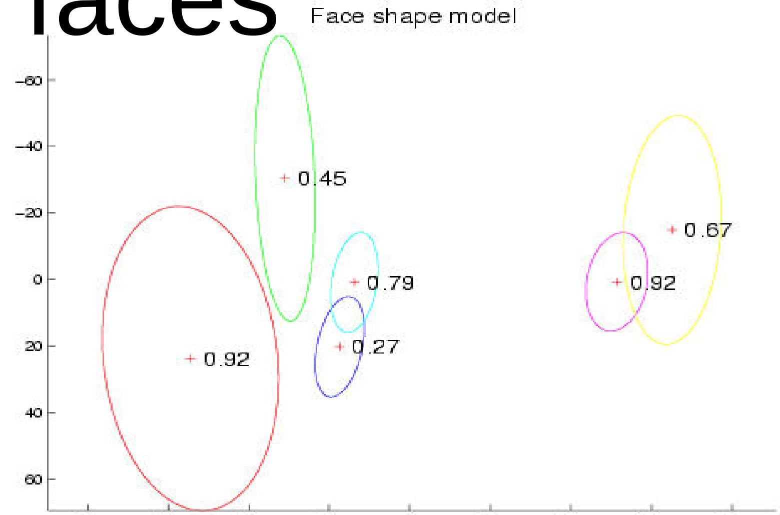
Background Images



Frontal faces

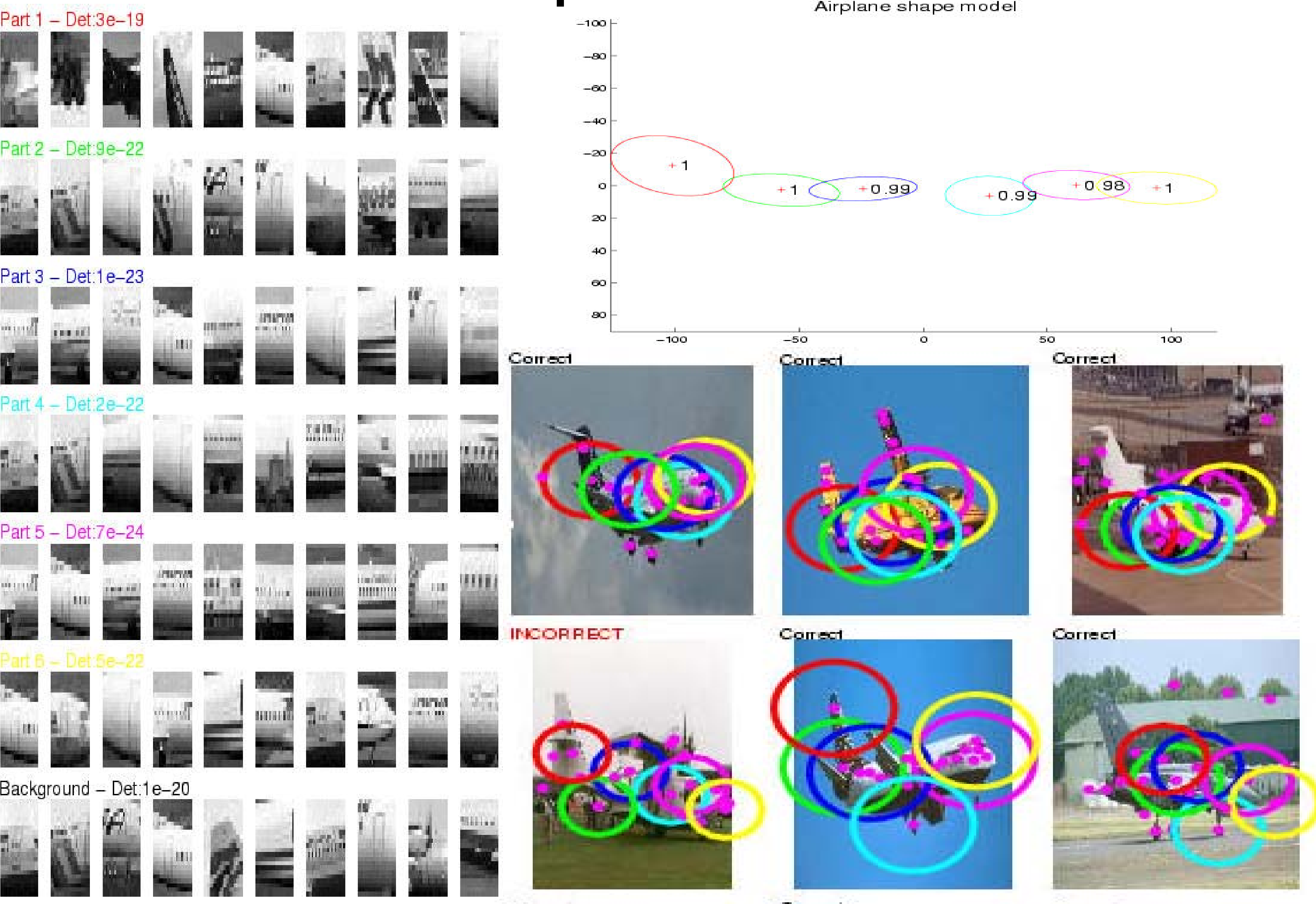
Equal error rate: 4.6%





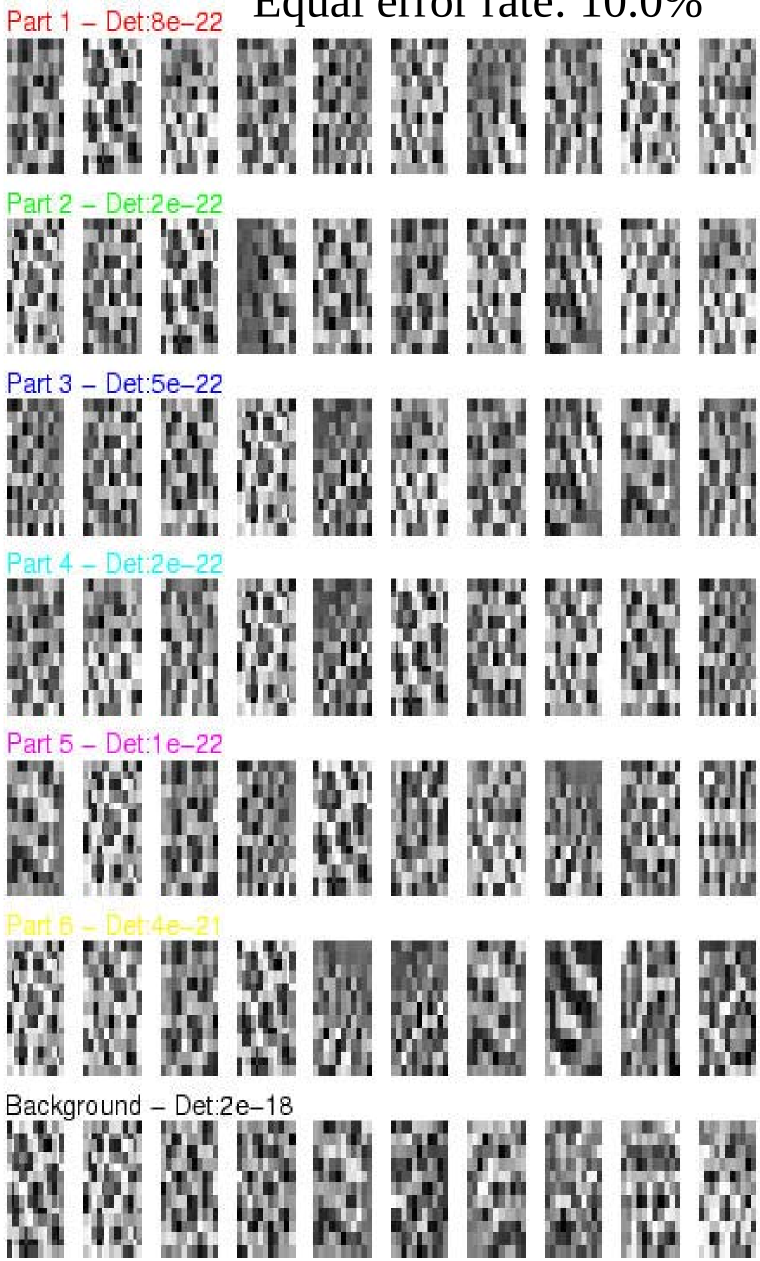
Airplanes

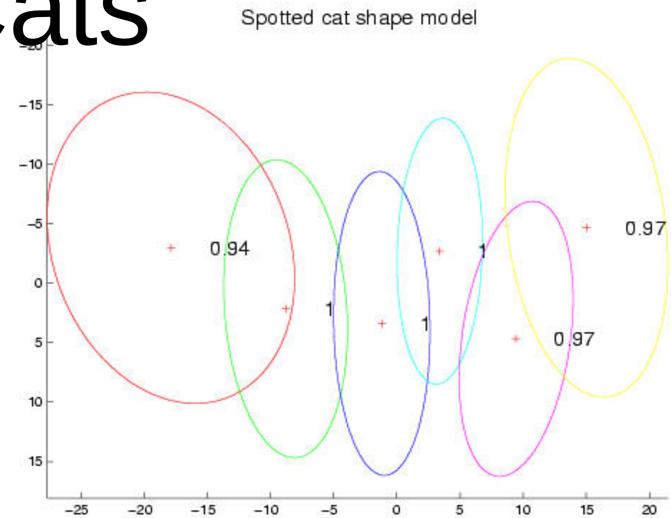
Equal error rate: 9.8%

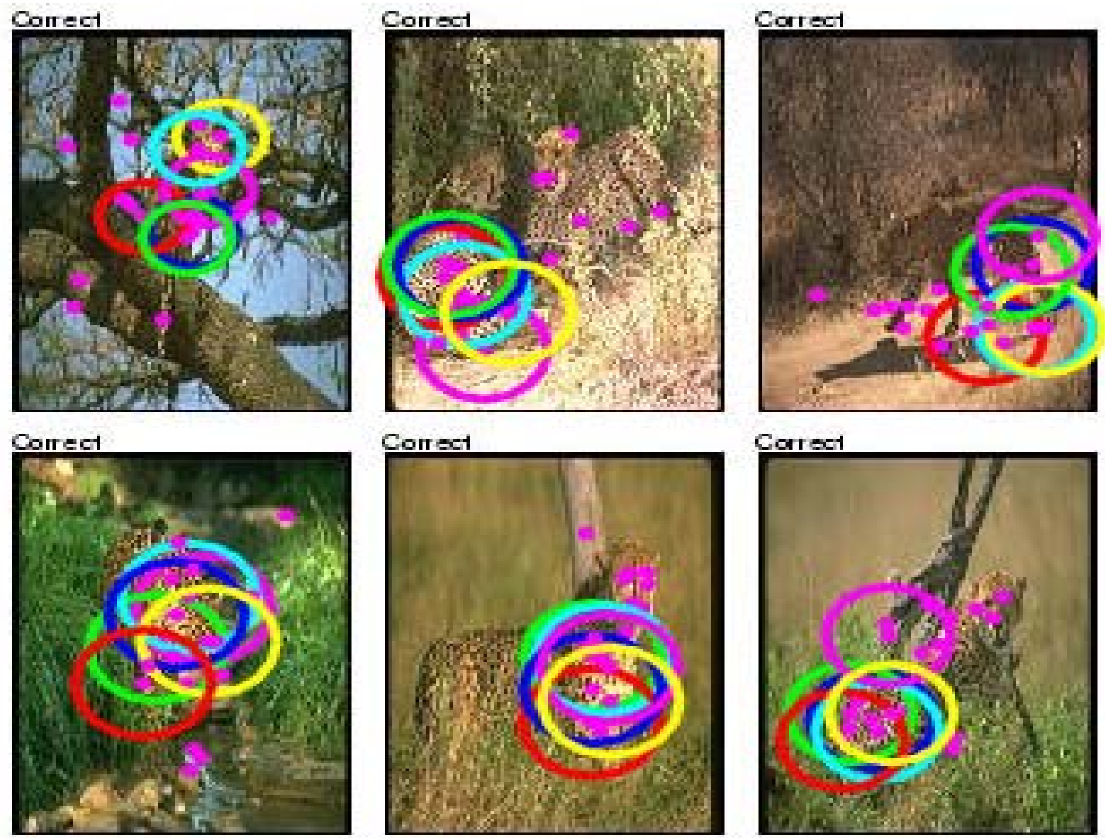


Scale-Invariant Cats

Equal error rate: 10.0%

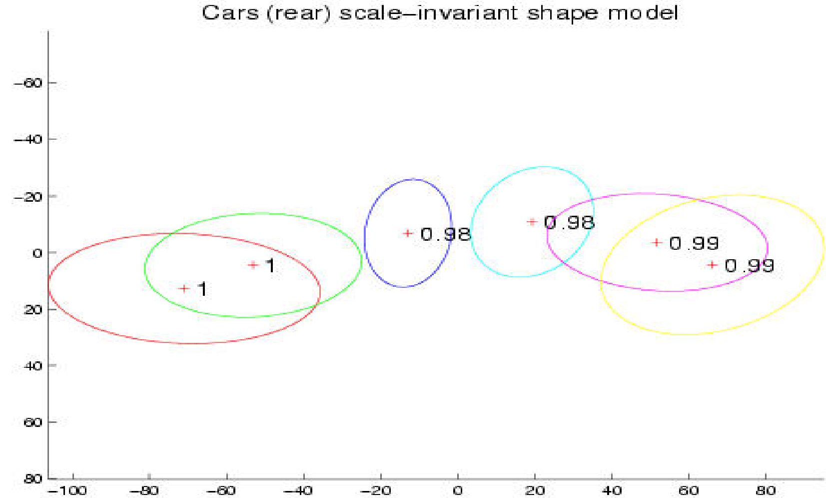


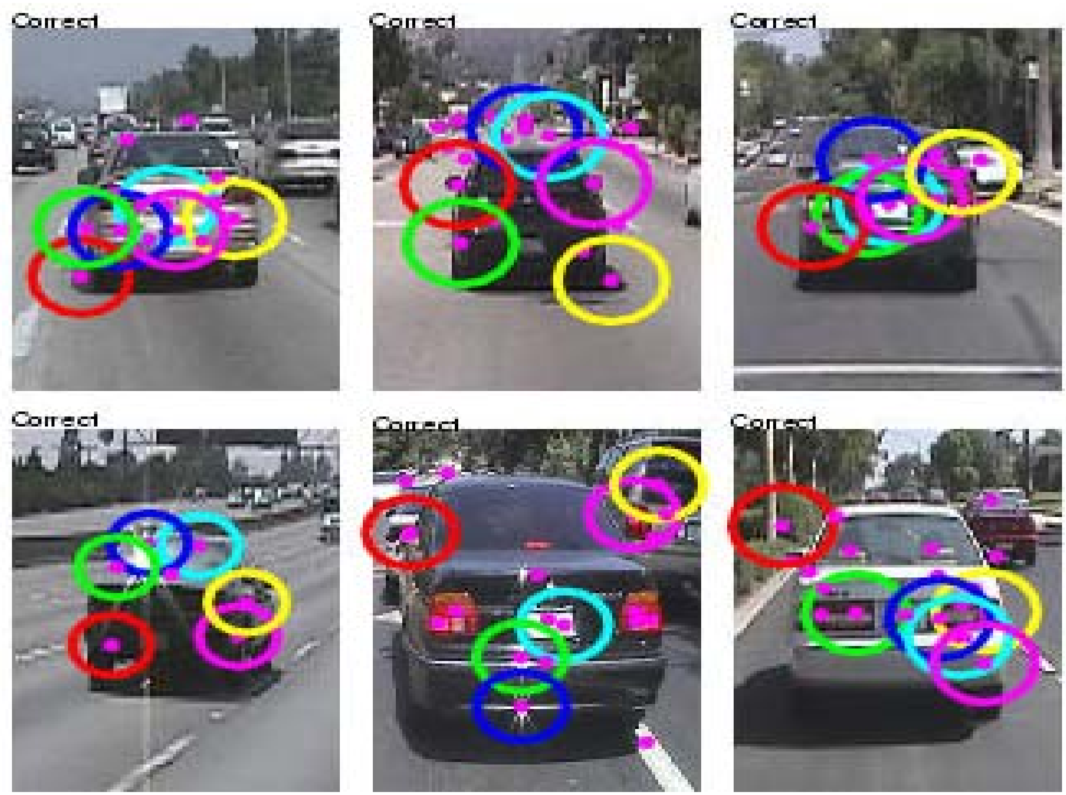




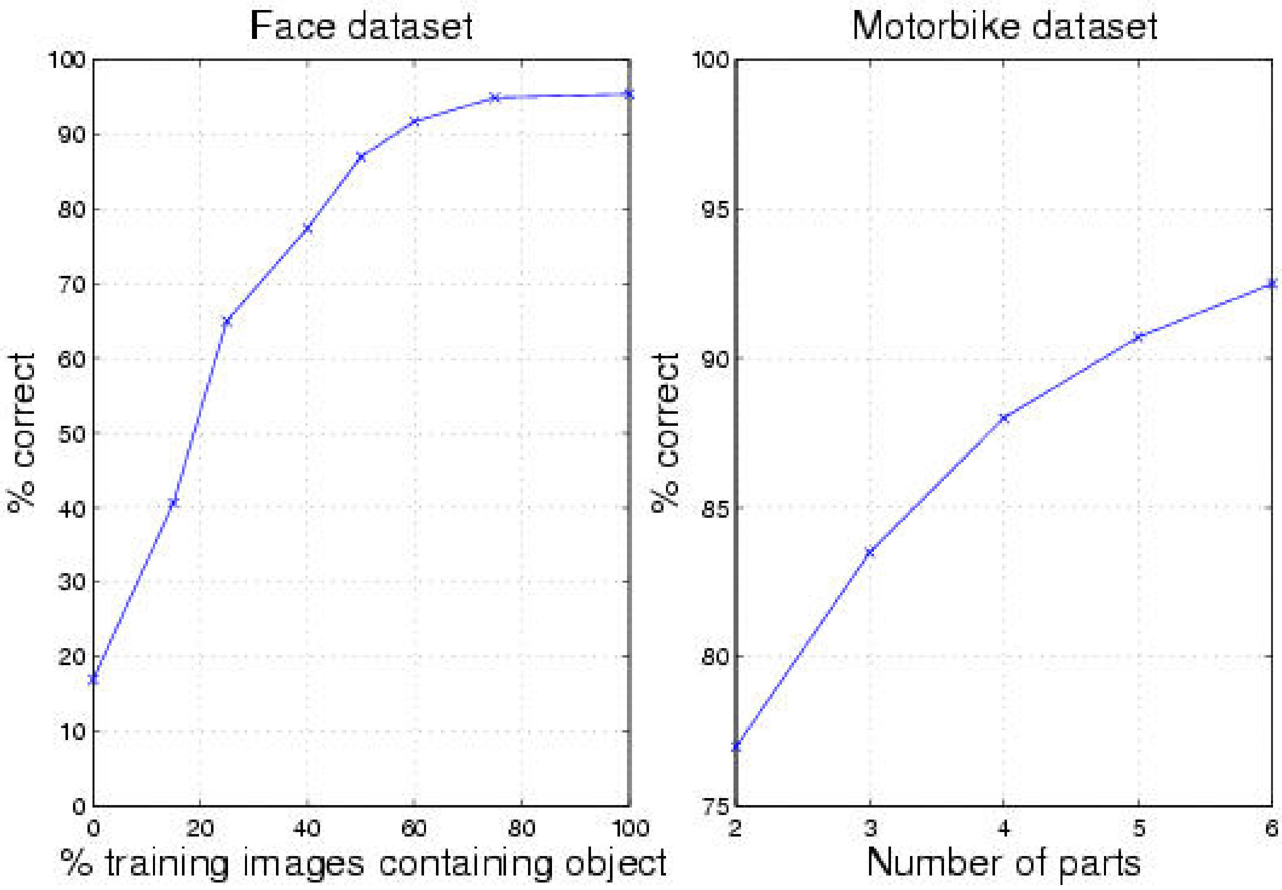
Scale-Invariant cars

Equal error rate: 9.7%





Robustness of Algorithm



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]

ROC equal error rates

Scale-Invariant Learning and Recognition:

	Total size	Object size	Pre-scaled	Unscaled
Dataset	of dataset	range (pixels)	performance	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