# One-Shot Learning of Object Categories

Li Fei-Fei, Rob Fergus, Peitro Perona

Presented by: Cobi Cario

# Main issues in this article

- A method using a constellation model to learn and recognize Object categories
- A method that try and succeed in learning from a small training set (1-5) images for each category

# Categories Algorithm seen so far

- All Used a bag of words methods

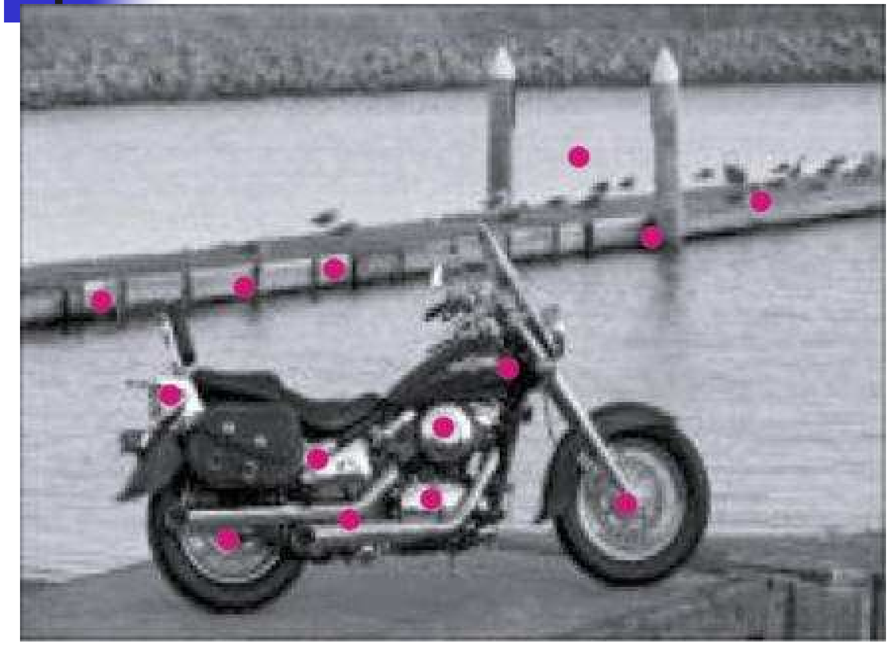
## Weaknesses:

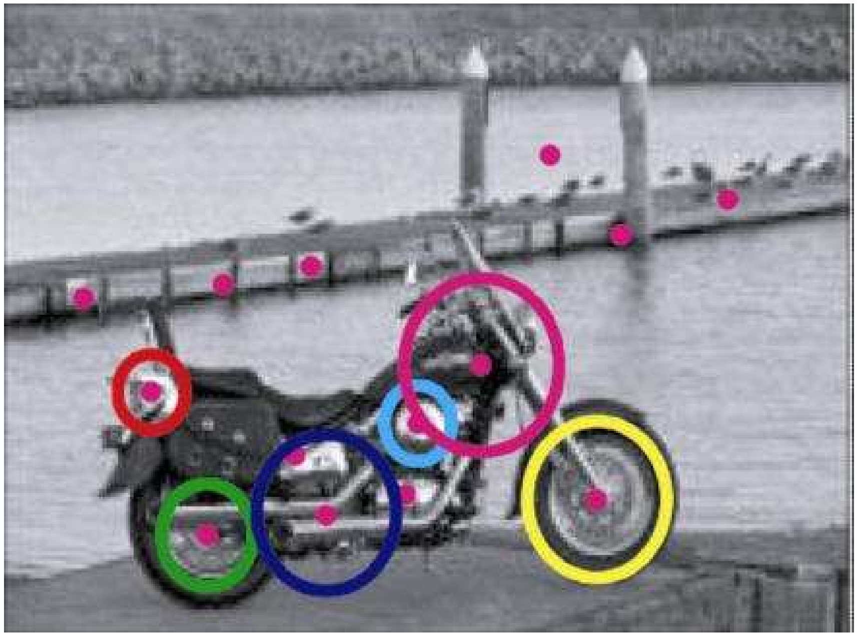
- All features has the same probability
- Location and shape knowledge is lost

# Constellation Model

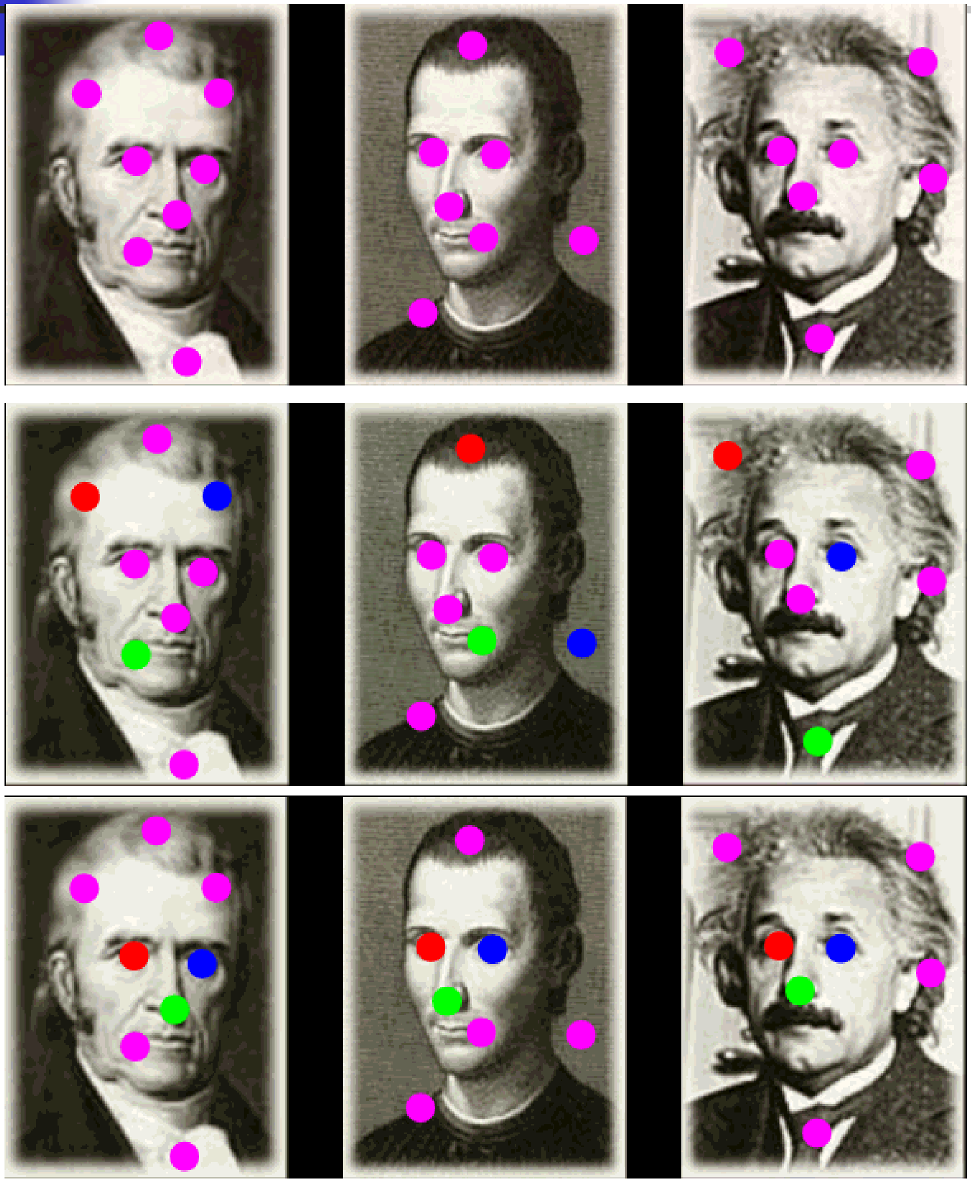
- Collect features from the image (include their location data)
- Choose P of the features as the object feature (choosing a hypothesis)
- Calculate the probability the object is in the picture using both the features appearance and relative location

# Constellation Model (cont.)

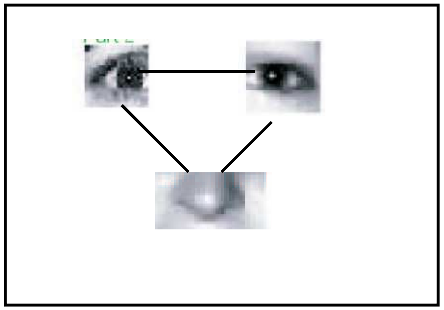




# Constellation Model – Is there a face



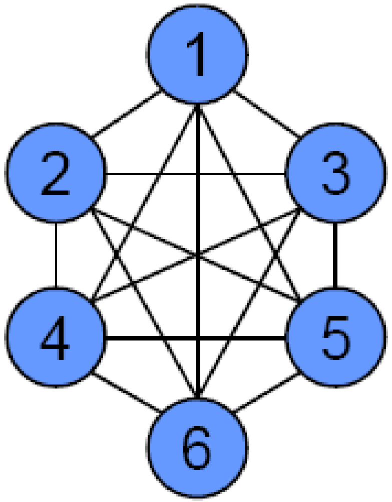
Faces learned model



We are looking to know

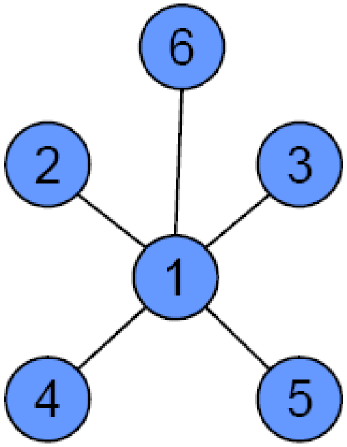
$$P(O_{fg} \mid X, A, h, \theta)$$

# Mathematical Approach – using Graphical Models



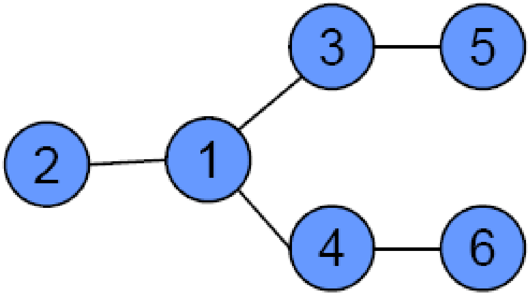
Fully connected

$O(N^P)$



Star structure

O($N^2$)



Tree structure

O($N^2$)

- When the graph connections depends on both the location (shape) and appearance

# Paper Approach

- Learning Categories from few (1-5) training samples
- Using only few (3) learned in advance categories
- Avoid using hand alignment on the training samples

# Bayesian Approach

- One versus all method
- Compare the probability there is an object in the image with the probability it is only background

$$R = \frac{\mathrm{P}(O_{fg} \mid I, I_t)}{\mathrm{P}(O_{bg} \mid I, I_t)} = \frac{\mathrm{P}(I \mid I_t, O_{fg})\mathrm{P}(O_{fg})}{\mathrm{P}(I \mid I_t, O_{bg})\mathrm{P}(O_{bg})}$$

$$R \geq THRESHOLD$$

# Bayesian Approach

$$R = \frac{\mathrm{P}(O_{fg} \mid I, I_t)}{\mathrm{P}(O_{bg} \mid I, I_t)} = \frac{\mathrm{P}(I \mid I_t, O_{fg})\mathrm{P}(O_{fg})}{\mathrm{P}(I \mid I_t, O_{bg})\mathrm{P}(O_{bg})}$$

$$= \frac{\mathrm{P}(O_{fg})}{\mathrm{P}(O_{bg})} \frac{\int \mathrm{P}(I \mid \theta)\mathrm{P}(\theta \mid I_t, O_{fg})\partial\theta}{\int \mathrm{P}(I \mid \theta_{bg})\mathrm{P}(\theta \mid I_t, O_{bg})\partial\theta_{bg}}$$

$$\propto \frac{\int \mathrm{P}(I \mid \theta)\mathrm{P}(\theta \mid I_t, O_{fg})\partial\theta}{\int \mathrm{P}(I \mid \theta_{bg})\mathrm{P}(\theta_{bg} \mid I_t, O_{bg})\partial\theta_{bg}}$$

# Representation of an Image

- Each image will be modeled as the set of features extracted from it
- Divide the features data to Shape and Appearance
- This leads us to

$$\propto \frac{\int \mathrm{P}(X, A \mid \theta)\mathrm{P}(\theta \mid X_t, A_t, O_{fg})\partial\theta}{\int \mathrm{P}(X, A \mid \theta_{bg})\mathrm{P}(\theta_{bg} \mid X_t, A_t, O_{bg})\partial\theta_{bg}}$$

# The likelihood $\mathrm{P}(X, A \mid \theta)$

- Using a constellation model, while h is an index of P (4) features which assume to be the object feature

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

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

- Under the assumption that Shape and Appearance are independent

# Background likelihood

- Assumption: the background model is fixed (Only one teta exist) The integral collapses
- There is only 1 background hypothesis, the null hypothesis
- We get for the background likelihood:

$$\mathrm{P}(X, A | \theta_{bg}) = \mathrm{P}(X, A, h_0 | \theta_{bg})$$

$$= P(X | h_0, \theta_{bg}^X) P(A | h_0, \theta_{bg}^A) P(h_0, \theta_{bg})$$

# Calculation of R

- Maximum Likelihood approach
- Maximum A posteriori approach
- Conjugate Densities

# Maximum Likelihood approach

Mission: Calculating $\int \mathrm{P}(X,A|\theta)\mathrm{P}(\theta|X_t,A_t,O_{fg})\partial\theta$

Assuming the following probability is highly peaked $\mathrm{P}(\theta|X_t,A_t,O_{fg})$

$$\theta^* : \delta(\theta-\theta^*)$$

This cause the integral to collapse which leaves us with

$$\mathrm{P}(X_t,A_t|\theta)$$

And using an ML approach we get

$$\theta^* = \theta^{ML} = \arg\max_{\theta} \mathrm{P}(X_t,A_t|\theta)$$

# Maximum A Posteriori Approach

- We use the same assumption and again we want to calculate $\theta^*$
- Only this time we have a prior knowledge $P(\theta)$
- So we can calculate:

$$\theta^* = \theta^{MAP} = \arg\max_{\theta} \mathrm{P}(X_t, A_t \mid \theta)\mathrm{P}(\theta)$$

# Conjugate Densities Approach

- Assume that $P(\theta \mid X_t, A_t, O_{fg})$ has a specific parametric form
- Which creates the integral

$$\int P(X, A \mid \theta) P(\theta \mid X_t, A_t, O_{fg}) \partial\theta$$

to have a close form solution

# Conjugate Densities Approach (Cont.)

- Actually this means choosing the following distributions density such as everything needed to be calculated and learned is achievable

$\mathrm{P}(X, A \mid \theta)$ - Is countable

$\mathrm{P}(\theta \mid X_t, A_t, O_{fg})$ - Can be learned

$\mathrm{P}(\theta)$ - Is countable

$\int \mathrm{P}(X, A \mid \theta) \mathrm{P}(\theta \mid X_t, A_t) \partial \theta$ - All the above generate this to be countable (close-form)

# Learning Using a Conjugate density

- Given few training samples which assumes to contain an Object of the category
- Features are extracted from the whole image
- And using a variation of EM (VBEM) learn $\mathrm{P}(\theta \mid X_t, A_t, O_{fg})$
- When ‘h’ is the hidden variable

# Implementation

# Implementation - General

- Gray scale images were used
- Experiments made on 101 Caltech data set (which was created for this assignment)
- Due to complexity issues only 4 features were used in each hypothesis

# Implementation - Features

- Kadir and Brady detectors were used
- This finds features that are salient over location and scale
- Location X was the center of each feature
- Features were scale to 11X11 pixel patch
- The Appearance was calculated as the first 10 principals over a fixed, trained PCA (trained over all features of background category)

# Prior calculation

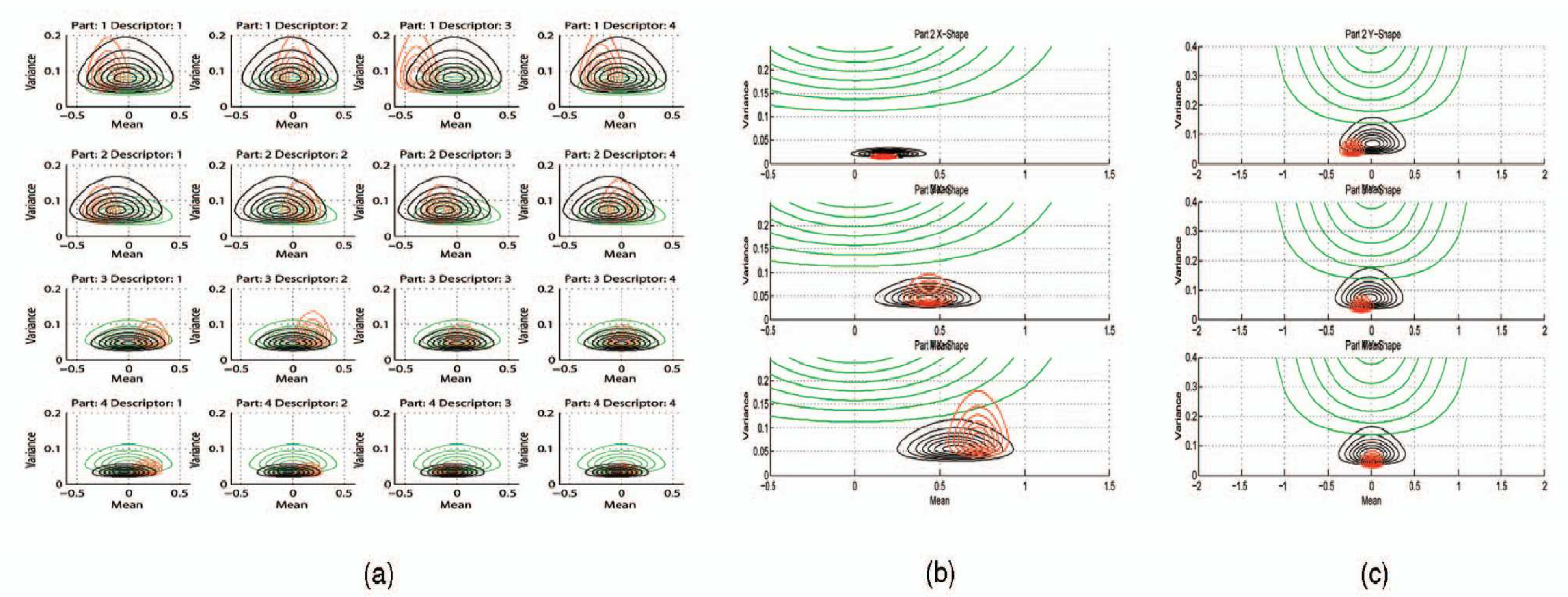
- Idea is that we can get the prior from the categories we’ve learned so far
- Assuming together they can create good prior for new categories
- Prior was calculated from 3 classes learned using ML method
- 30 models were calculated for $\theta$

# Experiments Settings

- From each category a fixed set of 50 images were selected
- From them 1-6 images were defined as the training images and the rest were test data
- Also 50 images were taken as background test images
- ML model was calculated for comparing

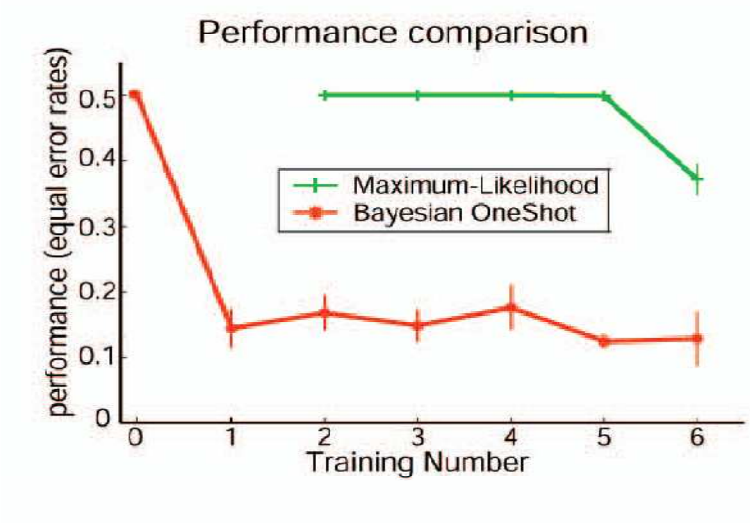
# Experiments Results

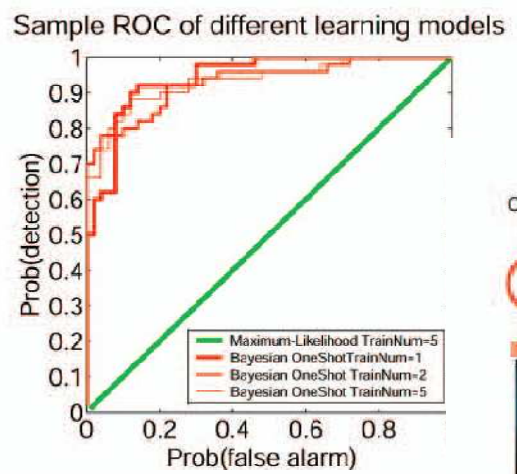
# Posterior Learning

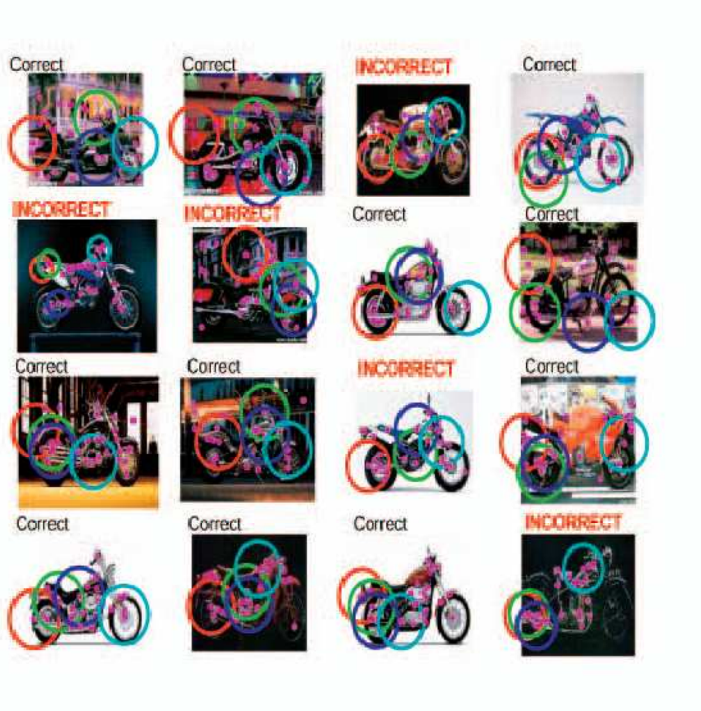


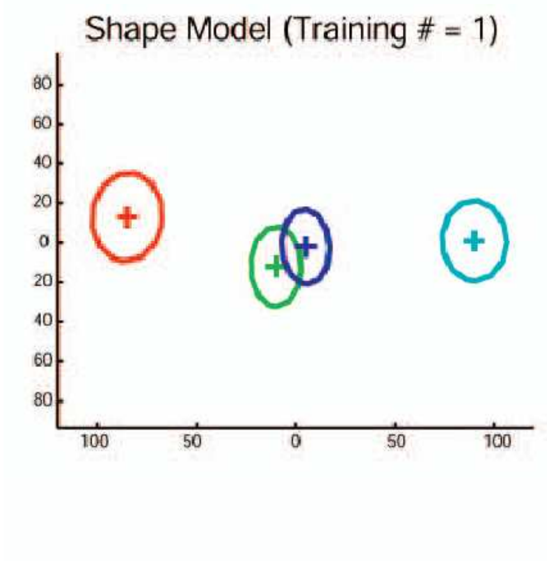
The learning process. (a) Appearance parameter space, showing the mean and variance distributions for each of the models' four parts for the first four descriptors. The parameter densities are colored as follows: black for the prior, green for the initial posterior density, and red for the density after 30 iterations of Bayesian One-Shot, when convergence is reached. (b) X component of the shape term for each of the model parts. (c) Y component of shape. Note that, in both (b) and (c), only the variance terms along the diagonal are visualized—not the covariance terms. This figure is best viewed in color with magnification.

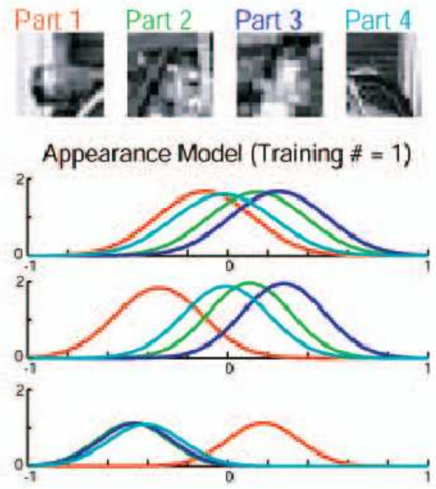
# Recognition Results-Over 1 train image



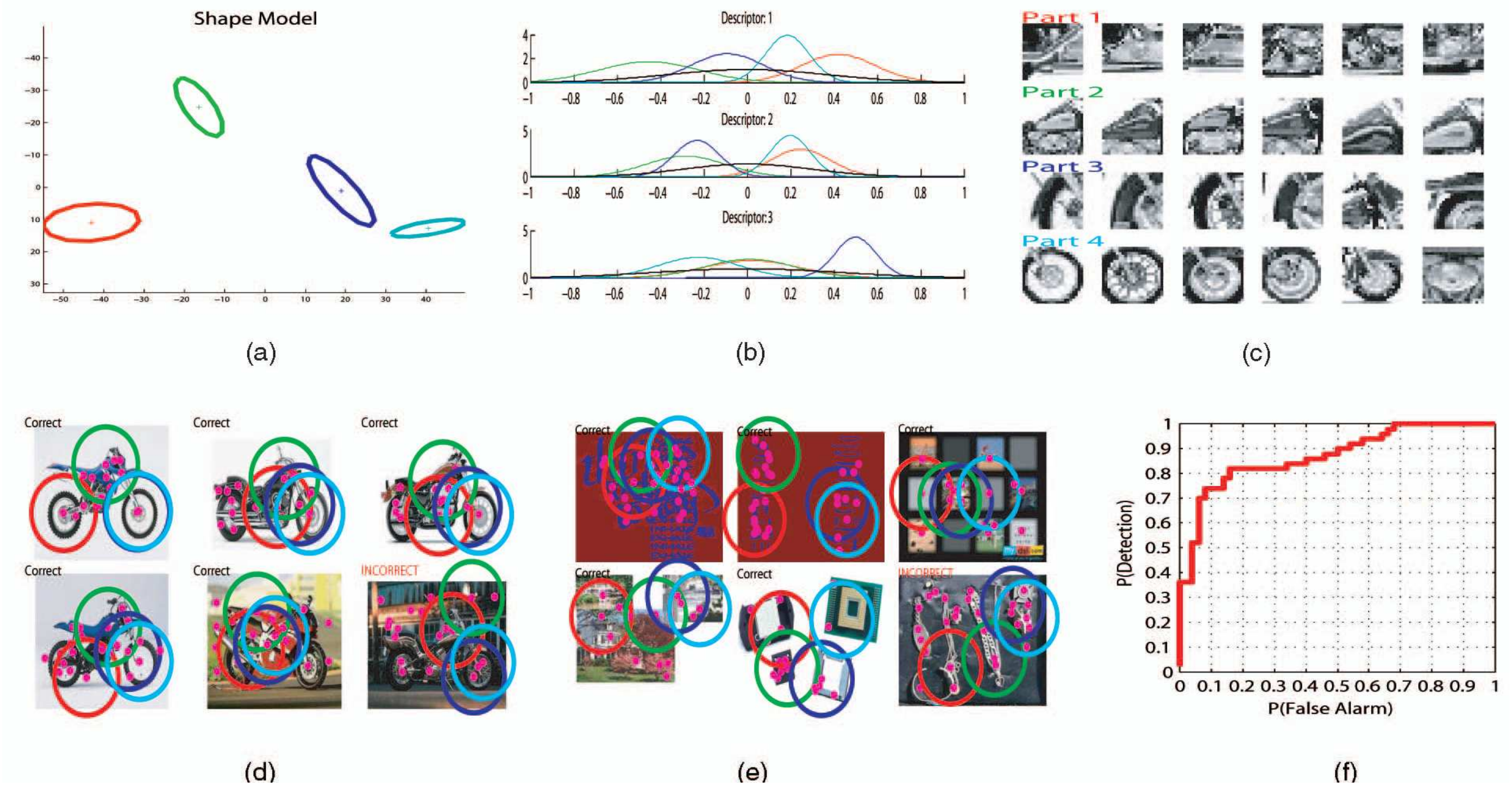








# Recognition results - over 6 train images



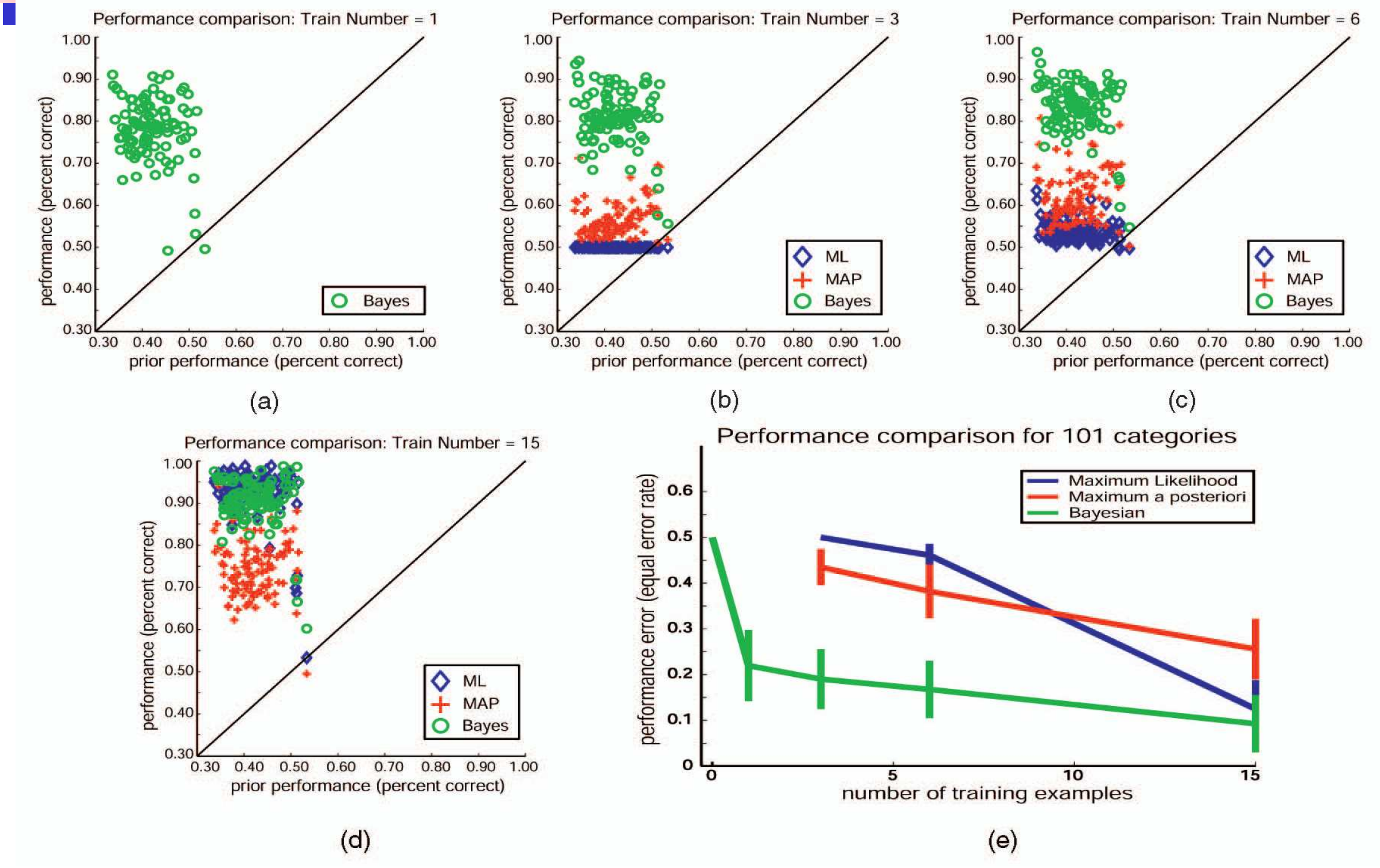
(a) (b) (c)

(d) (e) (f)

# ML and MAP vs. Beyesian Approach

- A question arises: What is the effect of the prior in the bayesian learning?
- This can be checked by comparing to the ML and MAP methods
- The ML totally lacks the prior
- The MAP is set using the same prior of the bayesian

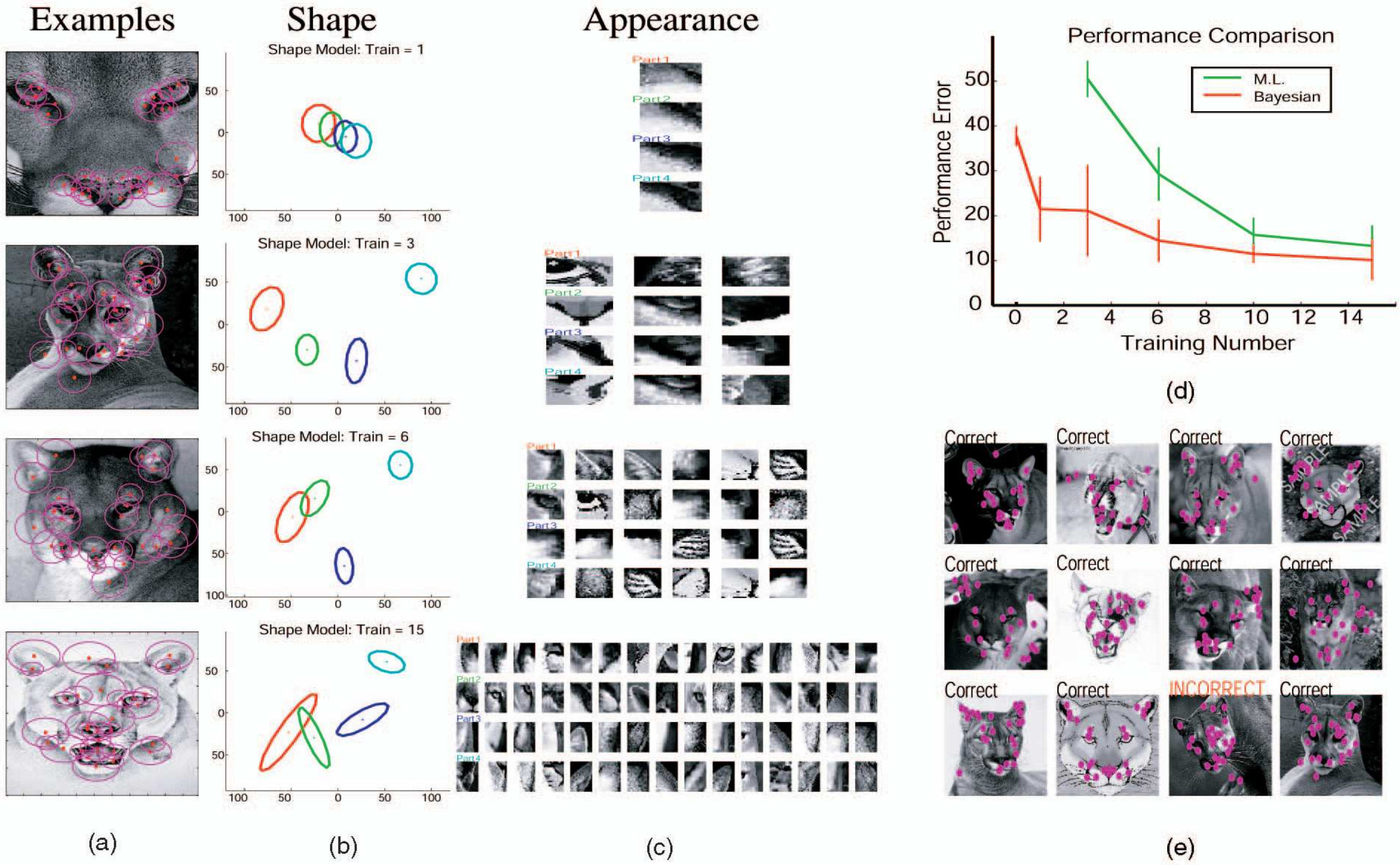
# ML/MAP vs. Bayesian



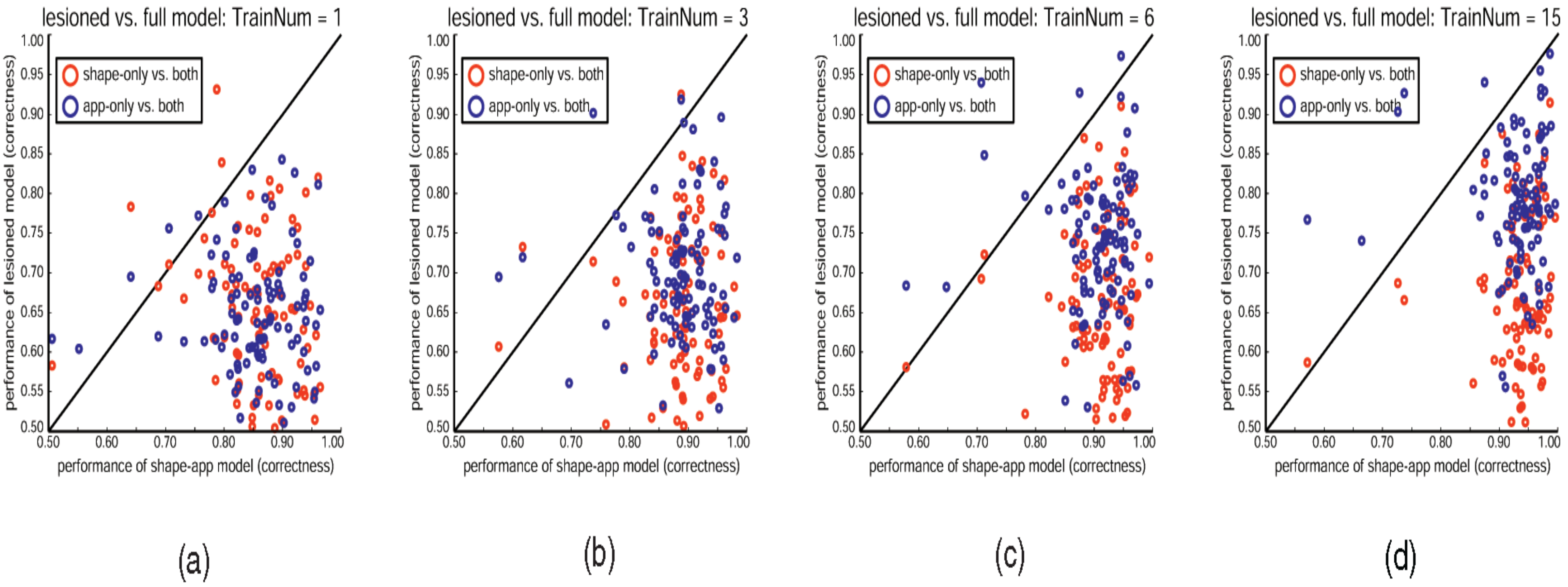
(a) (b) (c)

(d) (e)

# ML/MAP vs. Bayesian



# Shape/Appearance Only vs. Shape-Appearance



# Discrimination between all 101 categories

- Using a winner takes all method
- For each image the most successful (biggest R) category was selected
- Results were:
- 3 – 10.4%     6 – 13.9%       15 – 17.7%
- 1 percent is the random decision result

# Future Work

- Research the prior by using checking the effect of more complex priors
- The effect of similarity to “prior” categories on the results
- Other models that uses a prior knowledge
- Using an incremental model on the prior

# Appearance likelihood calculation

- Assuming independency between features
- Assuming a gaussian distribution over each feature

$$\mathrm{P}(A \mid h, \theta^A) = \prod_{p=1}^{P} g(A(h_p) \mid \mu_p^A, \Gamma_p^A) \prod_{j=1,\, j/h}^{N} g(A(j) \mid \mu_{bg}^A, \Gamma_{bg}^A)$$

background $$\mathrm{P}(A \mid h_0, \theta_{bg}^A) = \prod_{j=1}^{N} g(A(j) \mid \mu_{bg}^A, \Gamma_{bg}^A)$$ Const for a given image

Finally we get $$\frac{\mathrm{P}(A \mid h, \theta^A)}{\mathrm{P}(A \mid h_0, \theta_{bg}^A)} = \prod_{p=1}^{P} \frac{g(A(h_p) \mid \mu_p^A, \Gamma_p^A)}{g(A(h_p) \mid \mu_{bg}^A, \Gamma_{bg}^A)}$$

# Shape likelihood calculation

- Using joint gaussian distribution
- Using the left most feature as a landmark we create an invariant space
- Use a uniform density for the object’s position

$$\mathrm{P}(X \mid h, \theta^{X}) = \alpha^{-1} g(X(h) \mid \mu^{X}, \Gamma^{X}) \alpha^{-(N-P)}$$

background $\mathrm{P}(X \mid h_0, \theta^{X}_{bg}) = \alpha^{-N}$

$$\frac{\mathrm{P}(X \mid h, \theta^{X})}{\mathrm{P}(X \mid h_0, \theta^{X}_{bg})} = \alpha^{P-1} g(X(h) \mid \mu^{X}, \Gamma^{X})$$

# Implementation of the Conjugate Densities

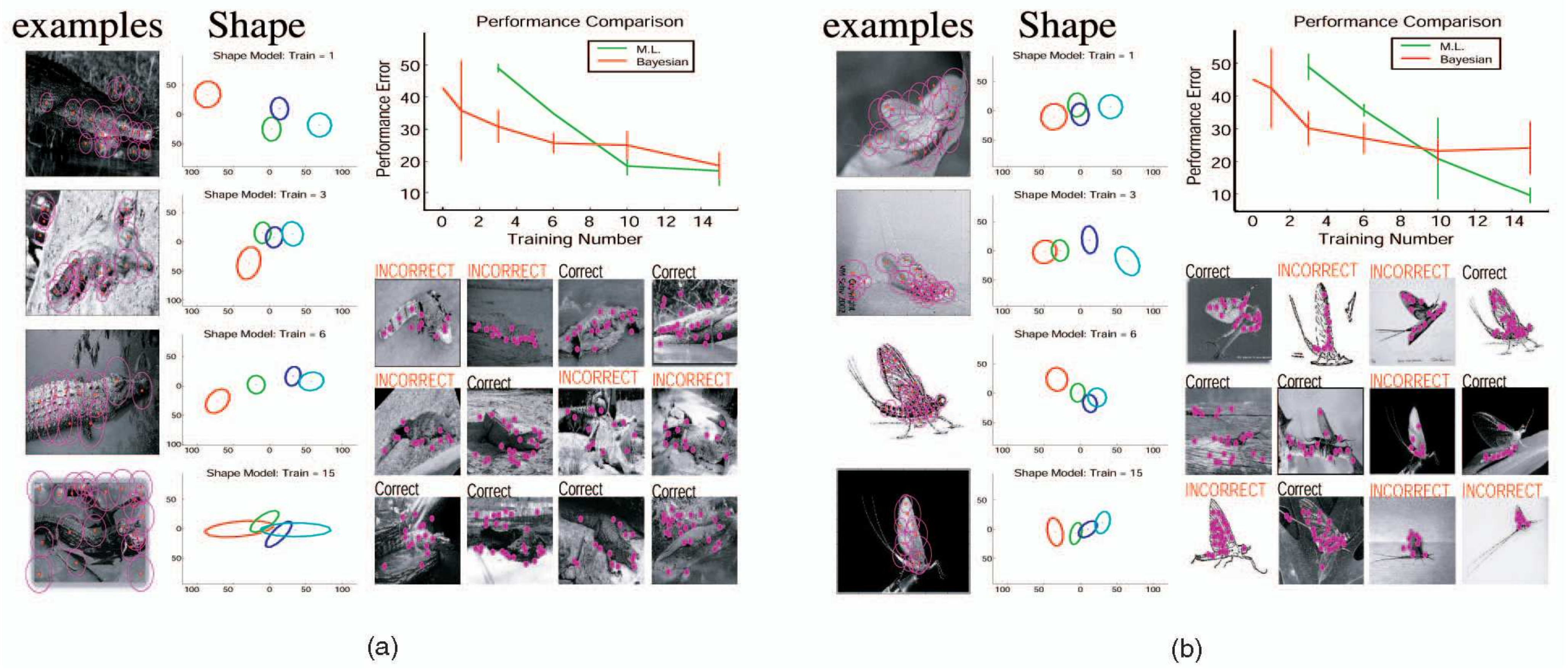
$$\mathrm{P}(X, A \mid \theta) = \sum_{h=1}^{H} \mathrm{P}(X(h) \mid \mu^X, \Gamma^X)\mathrm{P}(A(h) \mid \mu^A, \Gamma^A)$$

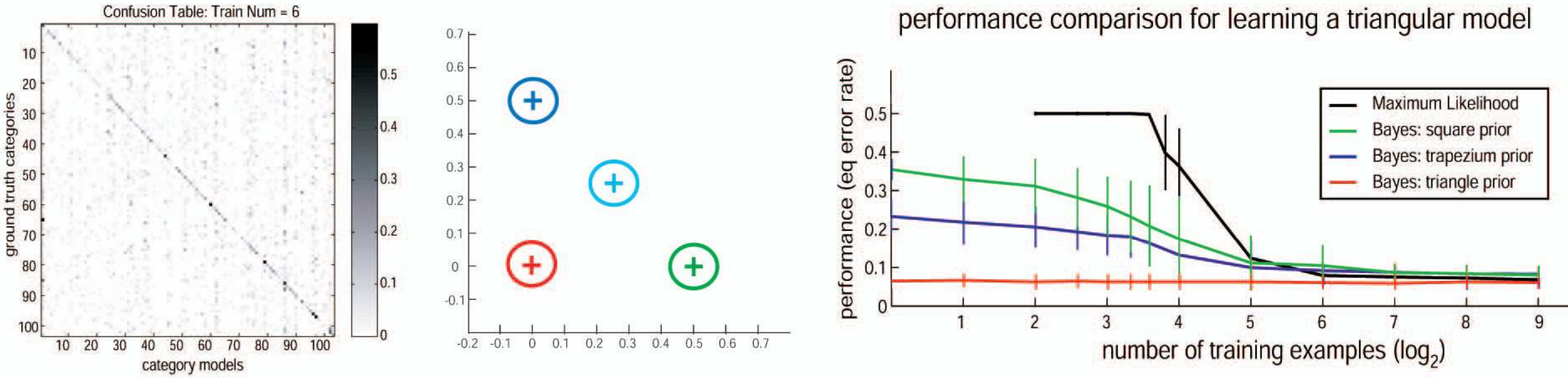
$$\mathrm{P}(\theta \mid X_t, A_t, O_{fg}) = \mathrm{P}(\pi)\mathrm{P}(\mu^X \mid \Gamma^X)\mathrm{P}(\Gamma^X)\mathrm{P}(\mu^A \mid \Gamma^A)\mathrm{P}(\Gamma^A)$$

$\mathrm{P}(\theta)$ - When choosing wisely $\mathrm{P}(\pi), \mathrm{P}(\Gamma^X)$ we get this to be Normal-Wishart distribution

$\int \mathrm{P}(X, A \mid \theta)\mathrm{P}(\theta \mid X_t, A_t)\partial\theta$ - On this setting this term become a multimodal Student's T distribution

# ML/MAP vs. Bayesian – bad category




Confusion Table: Train Num = 6
ground truth categories
category models
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
performance comparison for learning a triangular model
performance (eq error rate)
number of training examples ($\log_2$)
Maximum Likelihood
Bayes: square prior
Bayes: trapezium prior
Bayes: triangle prior
0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
9

(a)

(b)

(c)