Multiclass object recognition Sharing parts and transfer learning

Sharat Chikkerur

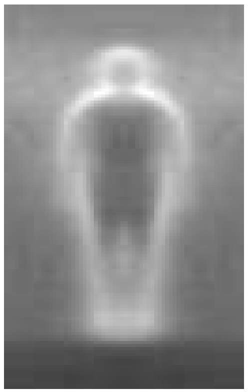
Outline

- Historical perspective and motivation

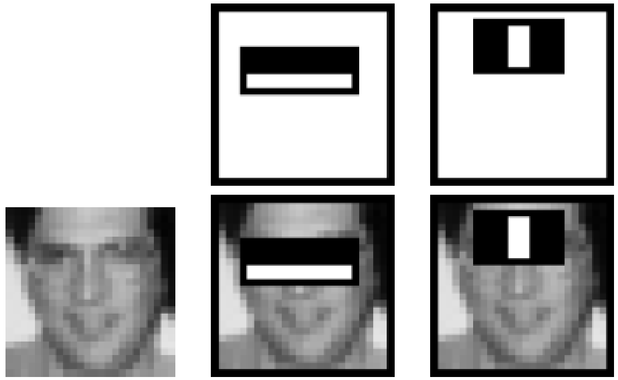
- Discriminative approach
  - A. Torralba, K. Murphy, W. Freeman, Sharing visual features for multiclass and multiview object detection, IEEE PAMI 2007

- Bayesian approach
  - (Prelude) R. Fergus, P. Perona, A. Zisserman, Object recognition by unsupervised scale-invariant learning, CVPR 03
  - L. Fei-Fei, R. Fergus and P. Perona. One-Shot learning of object categories. PAMI, 2006

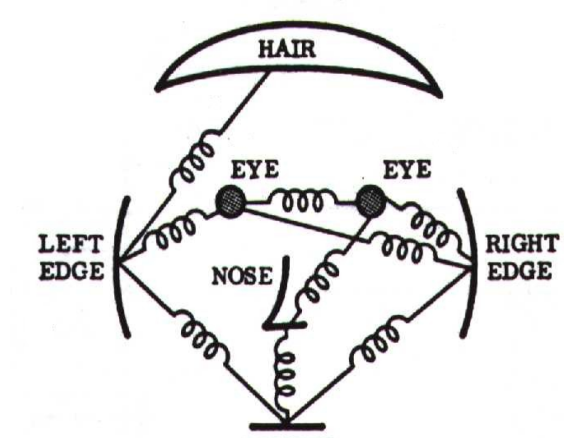
Perspective: Template vs parts



[Dalal & Triggs 05]



[Viola & Jones 01]



[Fischler73]

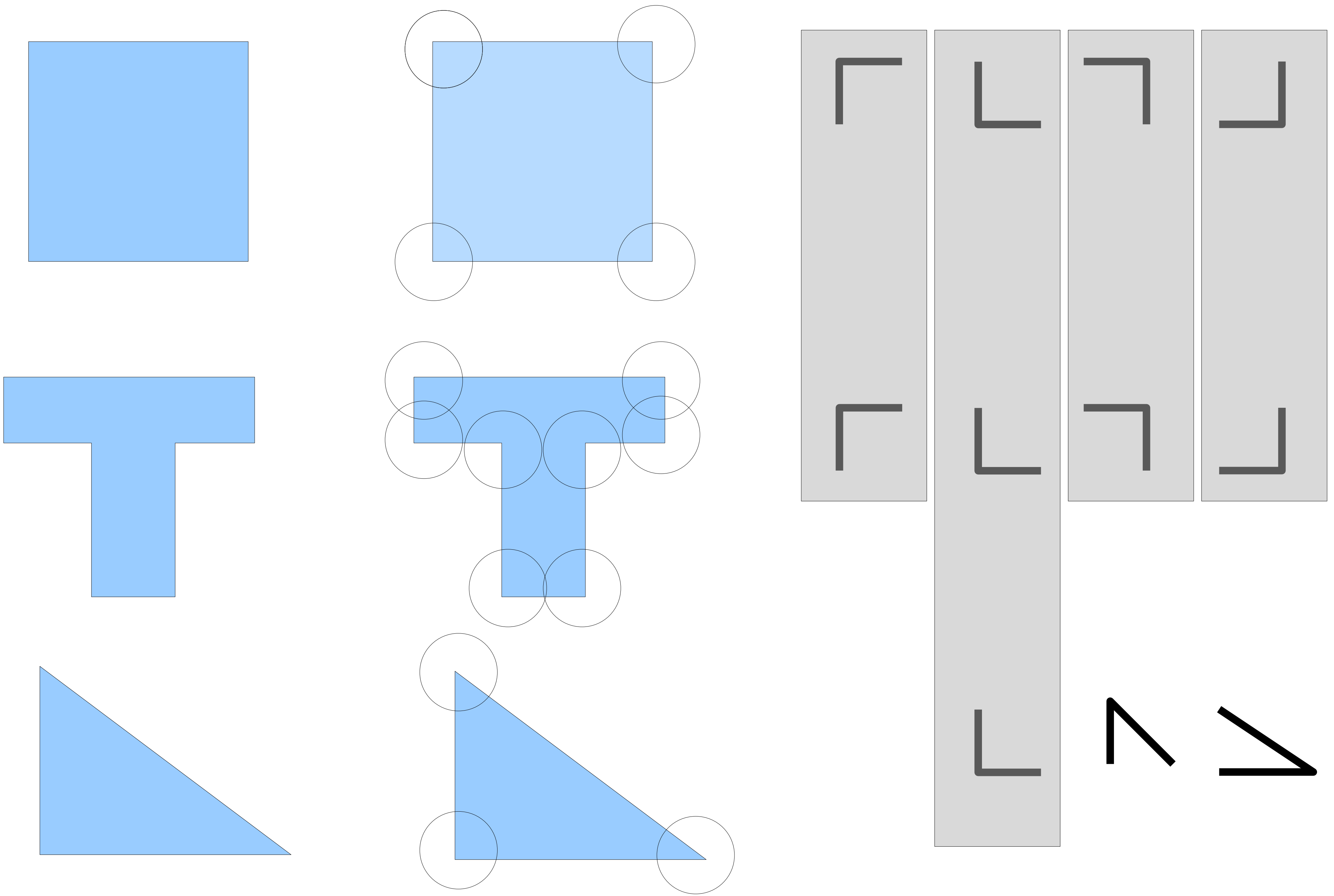
[Fergus 03]

- Dense representation
- Useful for rigid objects
- Less robust
- Appearance only
- Objects share features

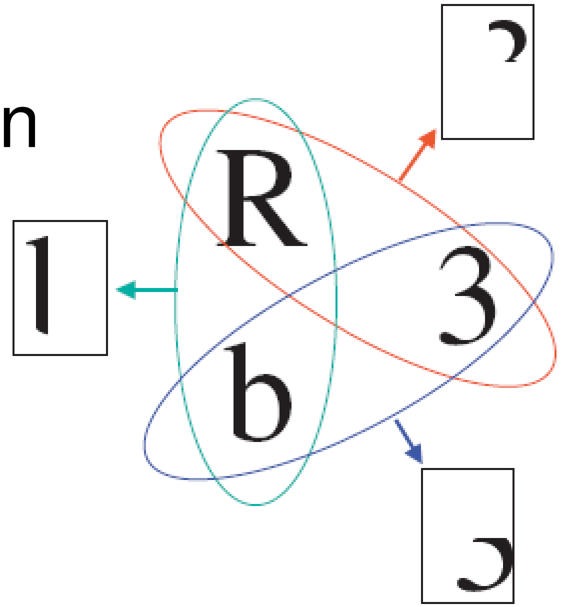
- Sparse representation
- Rigid and articulate objects
- More robust
- Appearance and shape
- Objects share parts

Summary: Part-based representation make more sense!

Motivation: Sharing parts



- Benefits
  - Learning is faster
    - Features are reused
    - Time complexity ~ O(log n) instead of O(n)
  - Better generalization
    - Individual parts share training data across classes
    - Robust to inter-class variation
- Challenges
  - Identity of shared parts/classes unknown
  - Sharing may not follow tree structure
  - Exhaustive search ~ $O(2^P)$



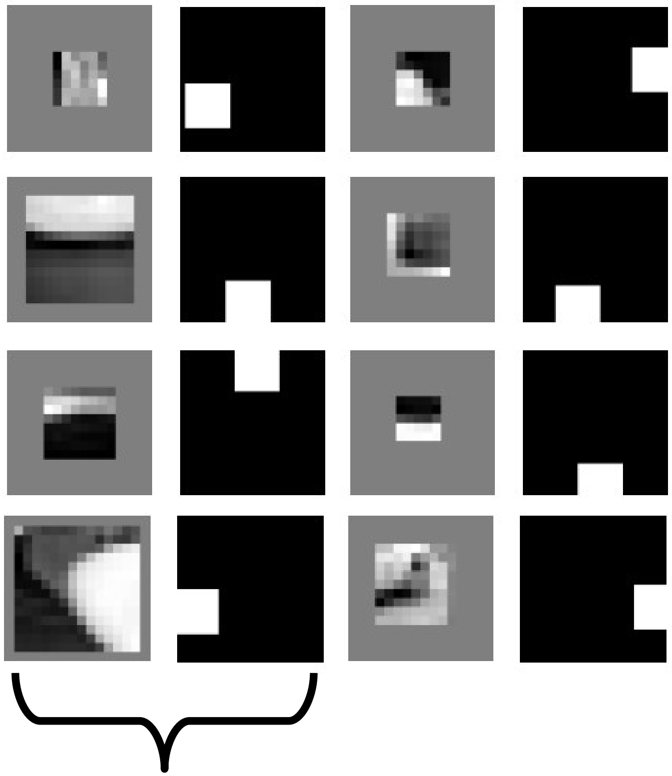
How do you share parts?

- Create a universal dictionary of parts
  - Serre et al 07 (HMAX), Ke and Sukhtankar (PCA SIFT)
- Learn the shared dictionary of parts
  - Discriminative
    - Embed sharing into optimization
      - Discriminative dictionary (Marial et al 08)
      - Joint boosting (Torralba et al 07)
  - Generative
    - Use unlabeled data to learn **prior**
      - Constellation model (Fei-Fei et al 06)

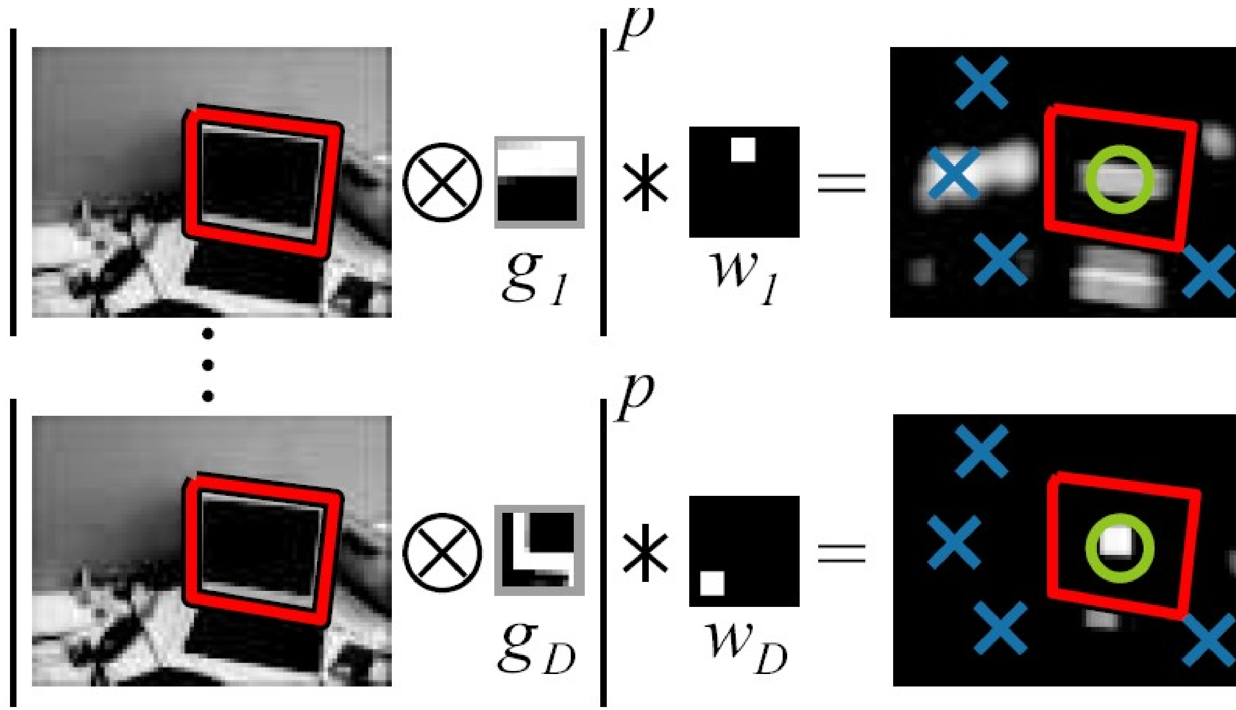
Discriminative approach

A. Torralba, K. Murphy, W. Freeman, Sharing visual features for multiclass and multiview object detection, IEEE PAMI 2007

Recap: Part representation



Feature (Appearance + Position)



Recap: Boosting

- An additive model for combining weak classifiers

$$H(v) = \sum_{m=1}^{M} h_m(v) \qquad J = E\left[e^{-zH(v)}\right]$$

- Weak classifier: $h_m(v_i) = a\delta(v_i^f > \theta) + b\delta(v_i^f \le \theta)$
- Algorithm:

1) Initialize the weights $w_i = 1$ and set $H(v_i) = 0$, $i = 1..N$.
2) Repeat for $m = 1, 2, \ldots, M$
   a) Fit stump: $h_m(v_i) = a\delta(v_i^f > \theta) + b\delta(v_i^f \le \theta)$
   b) Update class estimates for examples $i = 1, \ldots, N$: $H(v_i) := H(v_i) + h_m(v_i)$
   c) Update weights for examples $i = 1, \ldots, N$: $w_i := w_i e^{-z_i h_m(v_i)}$

Choosing a weak classifier

1) Initialize the weights $w_i = 1$ and set $H(v_i) = 0$, $i = 1..N$.
2) Repeat for $m = 1, 2, \ldots, M$
   a) $\boxed{\text{Fit stump: } h_m(v_i) = a\delta(v_i^f > \theta) + b\delta(v_i^f \leq \theta)}$
   b) Update class estimates for examples $i = 1, \ldots, N$: $H(v_i) := H(v_i) + h_m(v_i)$
   c) Update weights for examples $i = 1, \ldots, N$: $w_i := w_i e^{-z_i h_m(v_i)}$

- For each feature
  - Evaluate the weighted error $J_{wse} = \sum_{i=1}^{N} w_i (z_i - h_m(v_i))^2$
- Pick the feature with minimum error

Joint Boosting

- An additive model that jointly optimizes for all classes

$$H(v,c) = \sum_{m=1}^{M} h_m(v,c) \qquad J = \sum_{c=1}^{C} E\left[e^{-z^c H(v,c)}\right]$$

- Weak classifier:

$$h_m(v,c) = \begin{cases} a_S & \text{if } v_i^f > \theta \text{ and } c \in \boxed{S(n)} \\ b_S & \text{if } v_i^f \leq \theta \text{ and } c \in S(n) \\ k_S^c & \text{if } c \notin S(n) \end{cases}$$

$$\begin{pmatrix} H(v,1) \\ H(v,2) \\ H(v,3) \\ H(v,4) \\ H(v,5) \end{pmatrix} = \begin{pmatrix} h_1(v,1) \\ h_1(v,2) \\ h_1(v,3) \\ h_1(v,4) \\ h_1(v,5) \end{pmatrix} + \begin{pmatrix} h_2(v,1) \\ h_2(v,2) \\ h_2(v,3) \\ h_2(v,4) \\ h_2(v,5) \end{pmatrix} \cdots\cdots$$

1) Initialize the weights $w_i^c = 1$ and set $H(v_i, c) = 0$, $i = 1..N$, $c = 1..C$.

2) Repeat for $m = 1, 2, \ldots, M$

   a) Repeat for $n = 1, 2, \ldots, 2^C - 1$

      i) Fit shared stump:

$$h_m^n(v_i, c) = \begin{cases} a_S & \text{if } v_i^f > \theta \text{ and } c \in S(n) \\ b_S & \text{if } v_i^f \leq \theta \text{ and } c \in S(n) \\ k^c & \text{if } c \notin S(n) \end{cases}$$

      ii) Evaluate error

$$J_{wse}(n) = \sum_{c=1}^{C} \sum_{i=1}^{N} w_i^c (z_i^c - h_m^n(v_i, c))^2$$

   b) Find best subset: $n^* = \arg\min_n J_{wse}(n)$.

   c) Update the class estimates

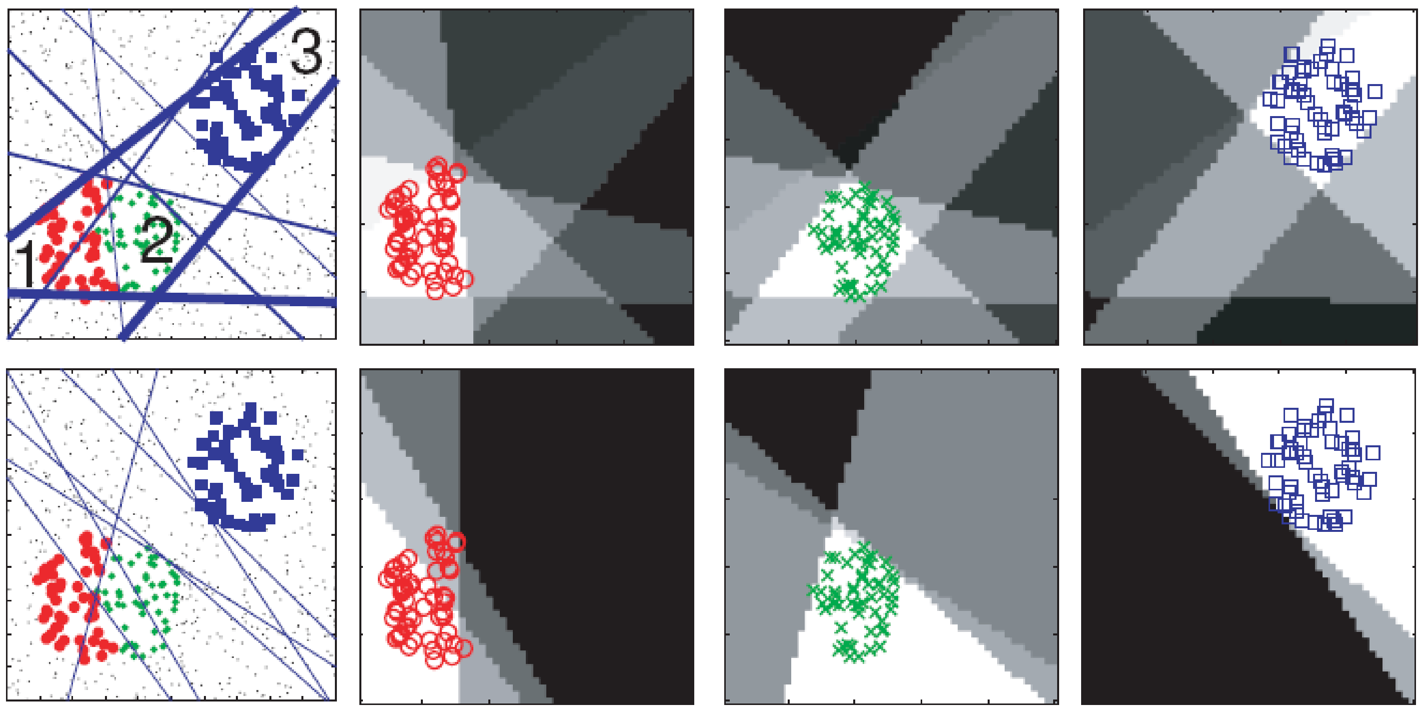
$$H(v_i, c) := H(v_i, c) + h_m^{n^*}(v_i, c)$$

   d) Update the weights

$$w_i^c := w_i^c e^{-z_i^c h_m^{n^*}(v_i, c)}$$

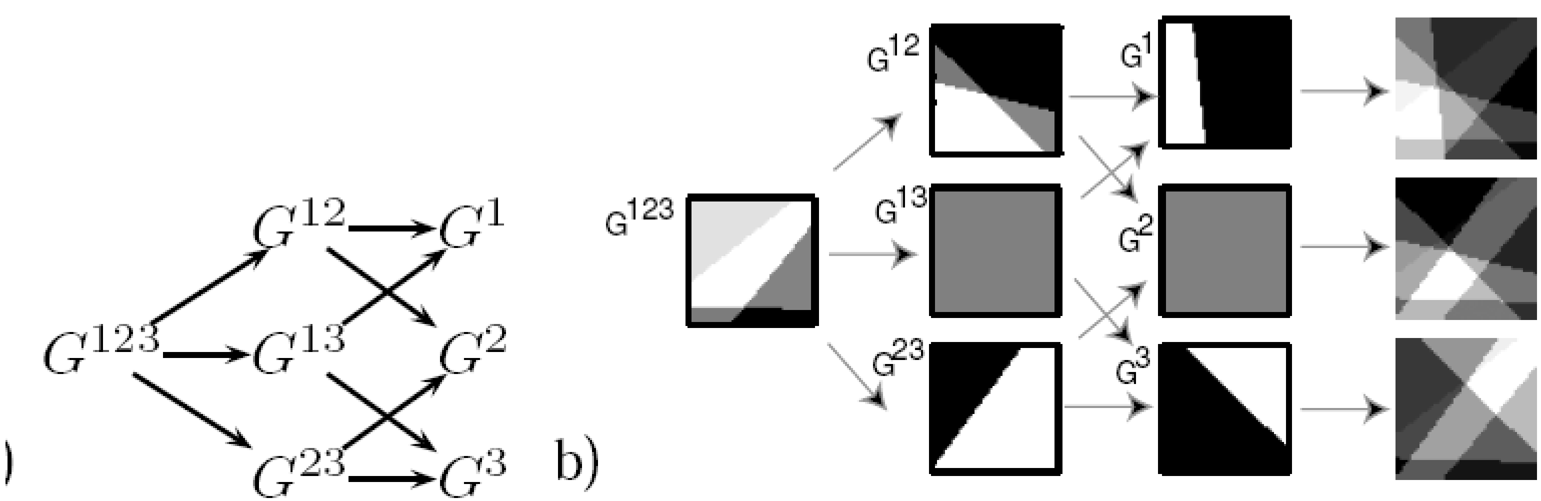
Vector valued

Example



Joint boosting

Independent

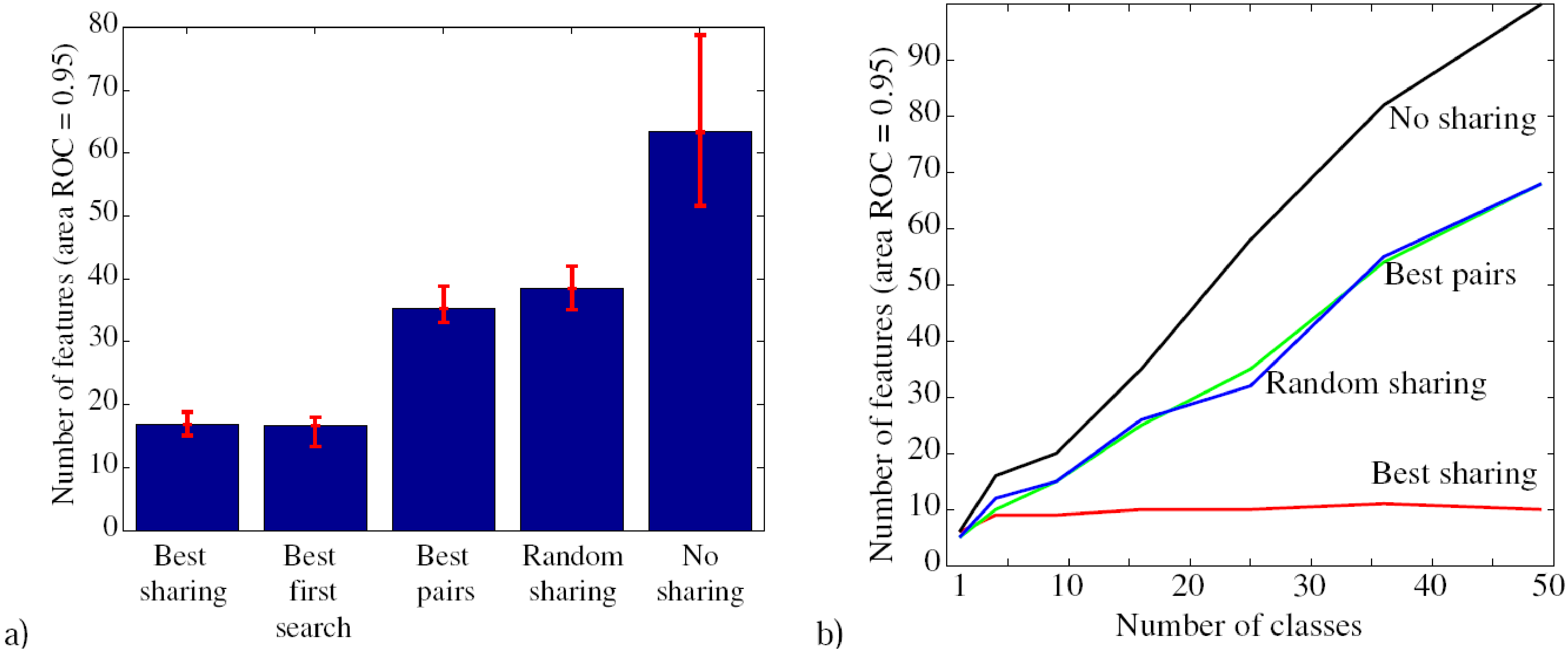


Feature sharing

Greedy approach

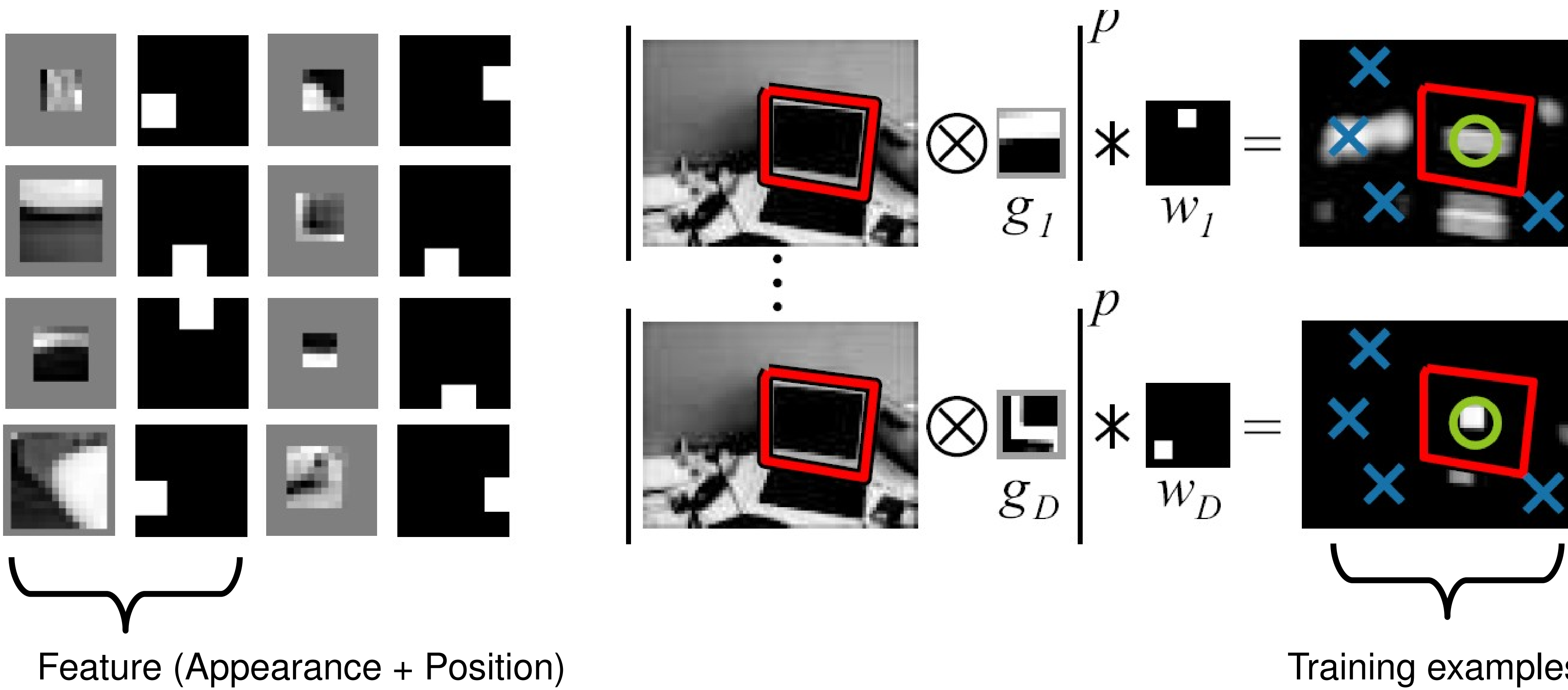
- Exhaustive search of all classes ~ $O(2^C)$
- Greedy approach
  - Select the class with best reduction in error
  - Insert next class with lowest error
  - Continue till all classes are selected
  - Select the best member from the set
  - Complexity ~ $\mathbf{O(C^2)}$

Typical behavior



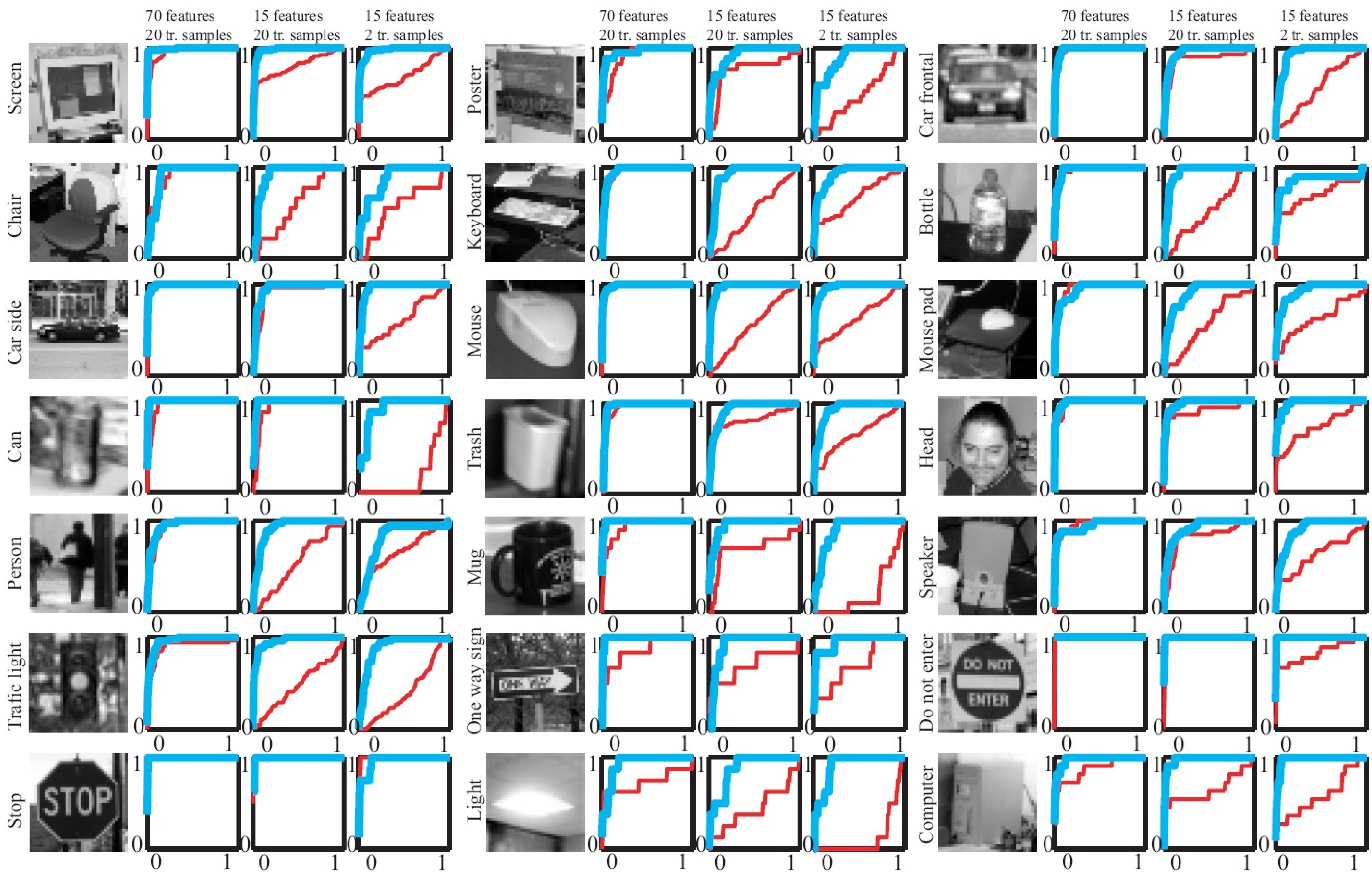
- Independent features/ pairs ~ O(N)
- Shared features ~ O(log N)

Application: Object categorization

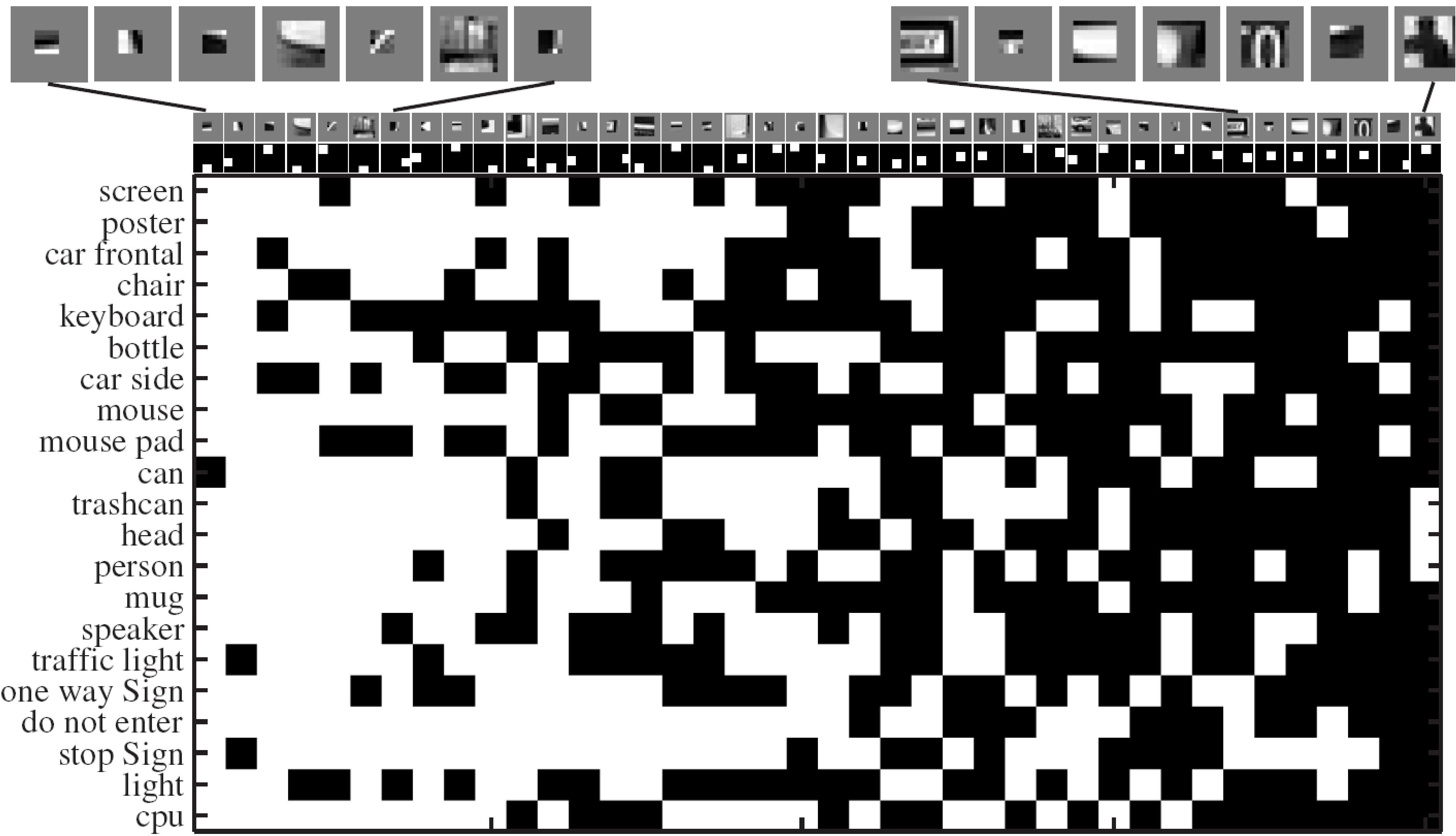


•Data: 21 object categories
•2000 candidate features (extracted by random sampling)
•50 training examples per category

Object categorization: Performance



Object categorization: Shared features



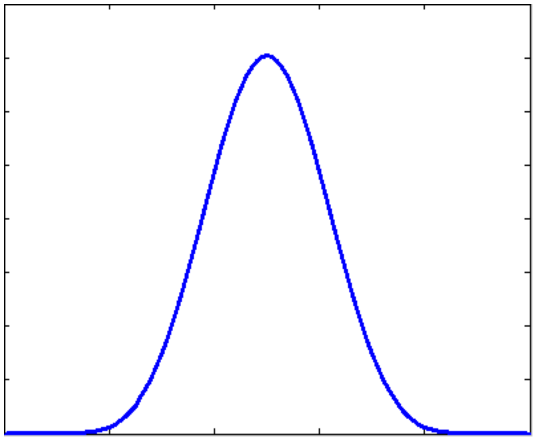
Summary

- Joint boosting allows learning of shared parts (even non-tree structures)
- Learning time reduces from O(N) to O(log N)
  - Allows scaling to large number of categories
- Reduces training sample size (per class)
- Useful for multi-class as well as multi-view recognition
- Wish-list?
  - Automatic scale selection for features
  - Handling occlusion

Bayesian approach

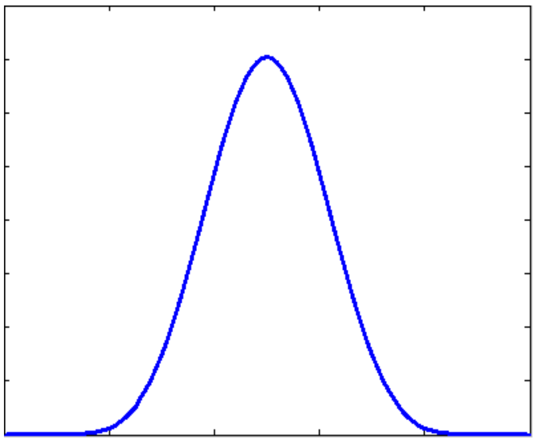
Bayes 101: Coin tossing

- MLE
  - Let p : probability of heads
  - Data : we observed H heads and T tails.
  - Inference: What is the chance of next head?
    - $P(\text{Head}) = p = H/(H+T)$
- Bayesian
  - Let p: probability of heads (unknown!), $p \sim f(p)$
  - $P(\text{Head}) = \int P(\text{Head}|p)\, f(p)\, dp$
  - Data: we observed H heads and T tails
  - $p \sim f(p|D)$, still not fixed!
    - $P(\text{Head}|D) = \int P(\text{Head}|p)\, f(p|D)\, dp$



Learning parameters: Conjugate priors

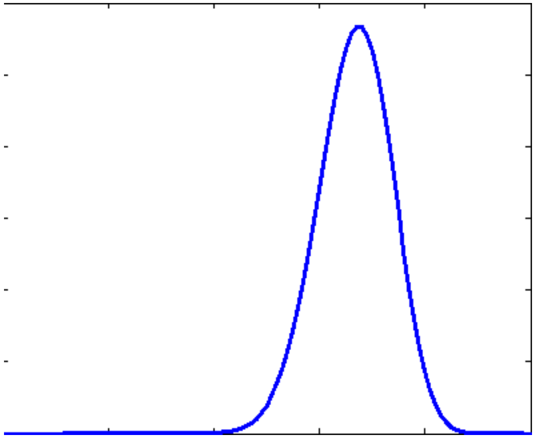
- Conjugate prior: Functional form of the prior and posterior distribution are identical
- With no data, we assume that the coin is likely to be fair
- Uncertainty based on hyper-parameters



Assumption "shared Knowledge"

p(h)~ B(a,b)

- After we observe data
  - D= (H-heads, T-tails)
  - the uncertainty in h is altered



Learning

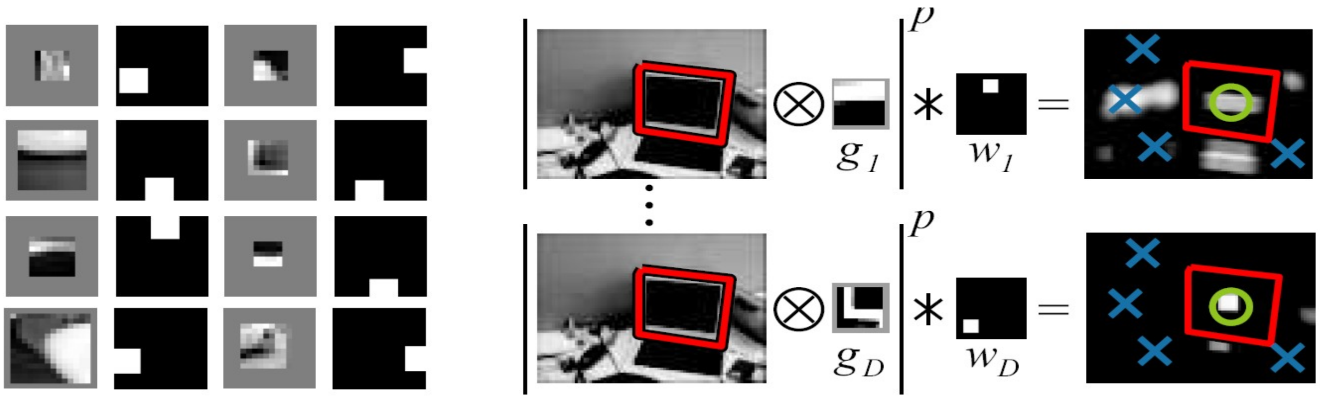
p(h|D) ~ B(a+H,b+T)

Transfer learning

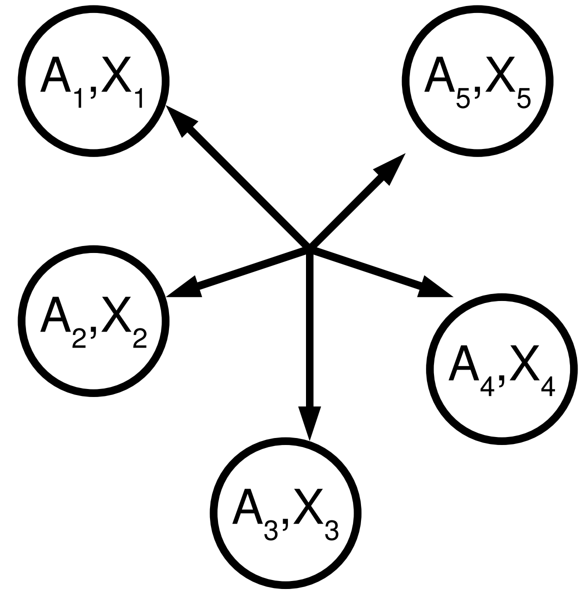
- Discriminative
  - Given data: Learn shared parameters
  - New data : Use all old parameters (+ new)
- Bayesian
  - Given data: Learn priors (“assumptions”)
  - New data : Update priors

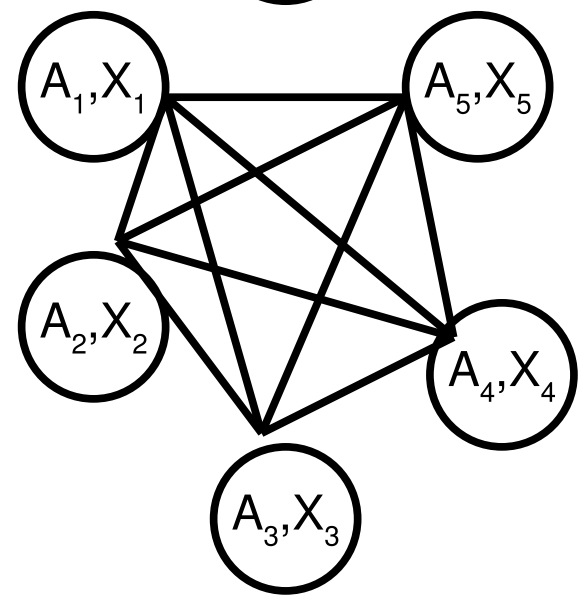
A prelude
Object class recognition by unsupervised scale-invariant learning

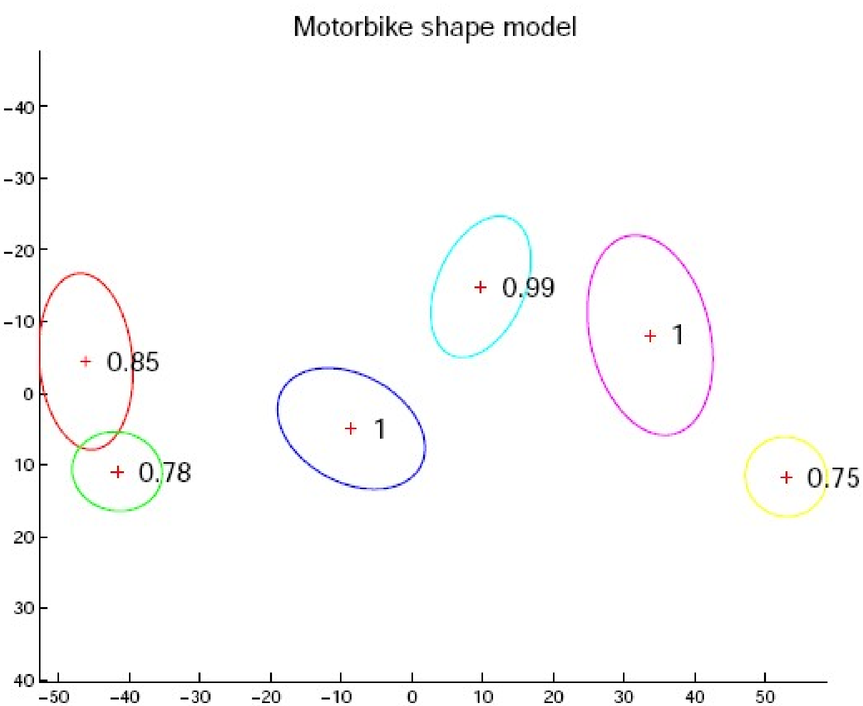
Constellation model



Torralba et al. ~100 parts







Fergus et al < 10 parts

Generative model/Bayesian detection

- Generative model for shape, appearance and scale

Latent variable

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

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

- **H** encodes the mapping from part (P) to interest point (N)
- Example: N=10,P-4, h=[0 3 4 5] or h=[3 1 2 10], |h| ~ O(N^P)
- Bayesian detection:

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

MLE approximation

Factorization

- Appearance

$$\frac{p(\mathbf{A}|\mathbf{X},\mathbf{S},\mathbf{h},\theta)}{p(\mathbf{A}|\mathbf{X},\mathbf{S},\mathbf{h},\theta_{bg})} = \prod_{p=1}^{P}\left(\frac{\mathrm{G}(\mathbf{A}(h_p)|\mathbf{c}_p,V_p)}{\mathrm{G}(\mathbf{A}(h_p)|\mathbf{c}_{bg},V_{bg})}\right)^{d_p}$$

- Shape

$$\frac{p(\mathbf{X}|\mathbf{S},\mathbf{h},\theta)}{p(\mathbf{X}|\mathbf{S},\mathbf{h},\theta_{bg})} = \mathrm{G}(\mathbf{X}(\mathbf{h})|\boldsymbol{\mu},\Sigma)\,\alpha^{f}$$

- Scale

$$\frac{p(\mathbf{S}|\mathbf{h},\theta)}{p(\mathbf{S}|\mathbf{h},\theta_{bg})} = \prod_{p=1}^{P}\mathrm{G}(\mathbf{S}(h_p)|t_p,U_p)^{d_p}\,r^{f}$$

- Occlusion

$$\frac{p(\mathbf{h}|\theta)}{p(\mathbf{h}|\theta_{bg})} = \frac{p_{Poiss}(n|M)}{p_{Poiss}(N|M)}\,\frac{1}{{}^{n}C_r(N,f)}\,p(\mathbf{d}|\theta)$$

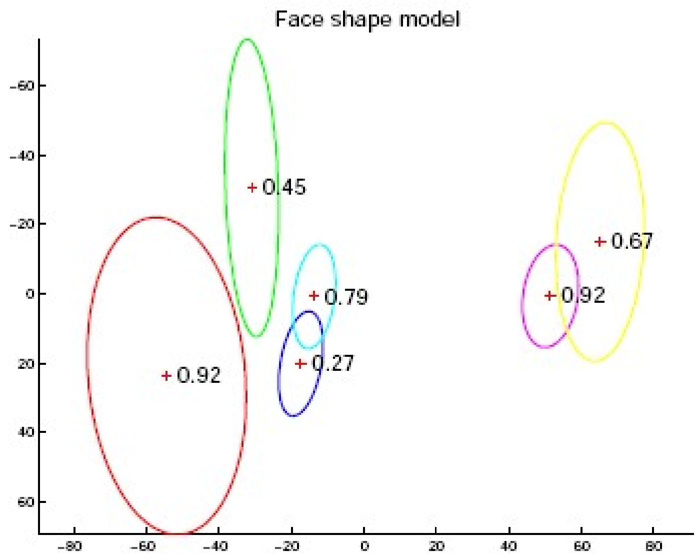
Representation and learning

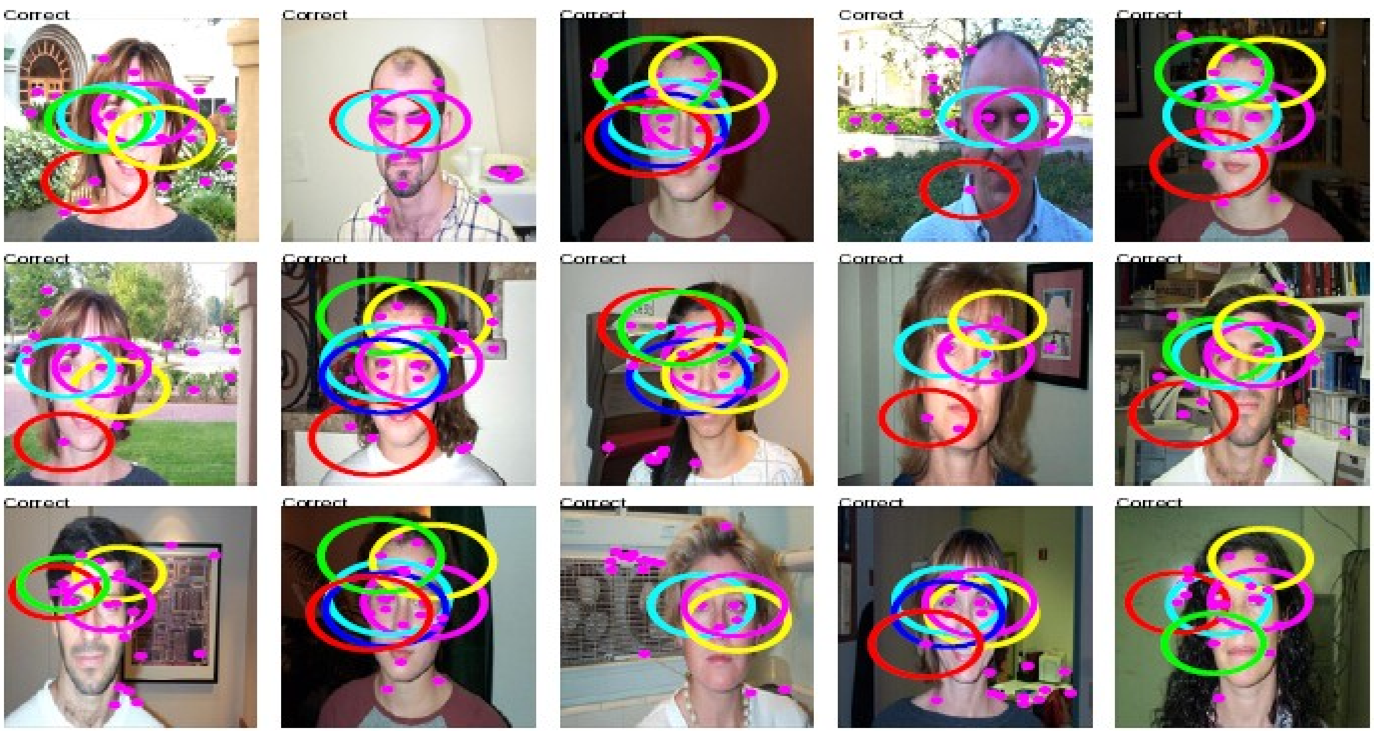
- Position/Shape
  - Candidate part locations are obtained using Kadir-Brady interest point detector
- Appearance
  - Modeled using 11x11 pixels around the interest point (PCA used for reducing dimension)
- Learning (EM)

$$\theta = \{\overbrace{\boldsymbol{\mu}, \Sigma}^{\text{X}}, \overbrace{\mathbf{c}, V}^{\text{A}}, \overbrace{M, p(\mathbf{d}|\theta)}^{\text{h}}, \overbrace{t, U}^{\text{S}}\}$$

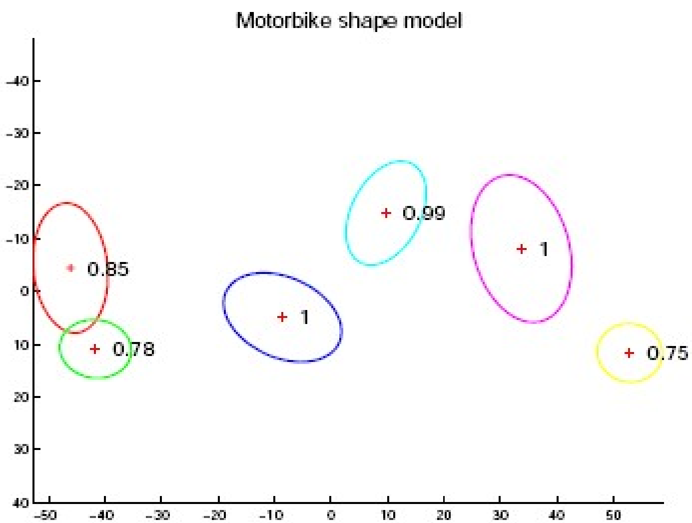
$$\hat{\theta}_{ML} = \underset{\theta}{arg\,max}\ p(\mathbf{X}, \mathbf{S}, \mathbf{A}|\,\theta)$$

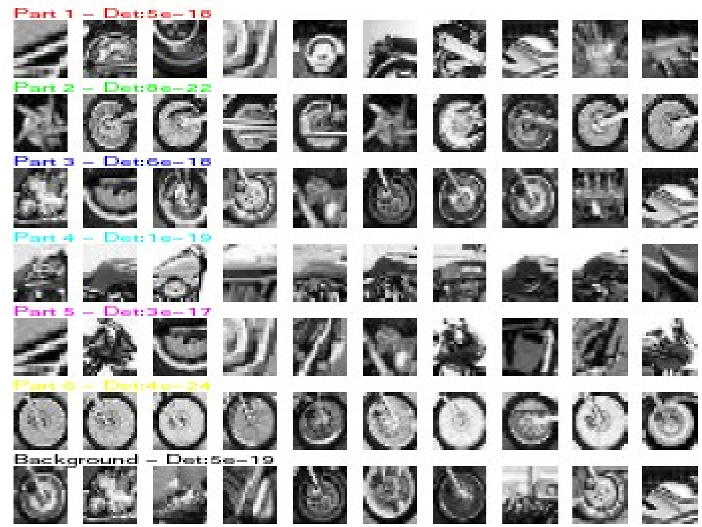
Example: faces

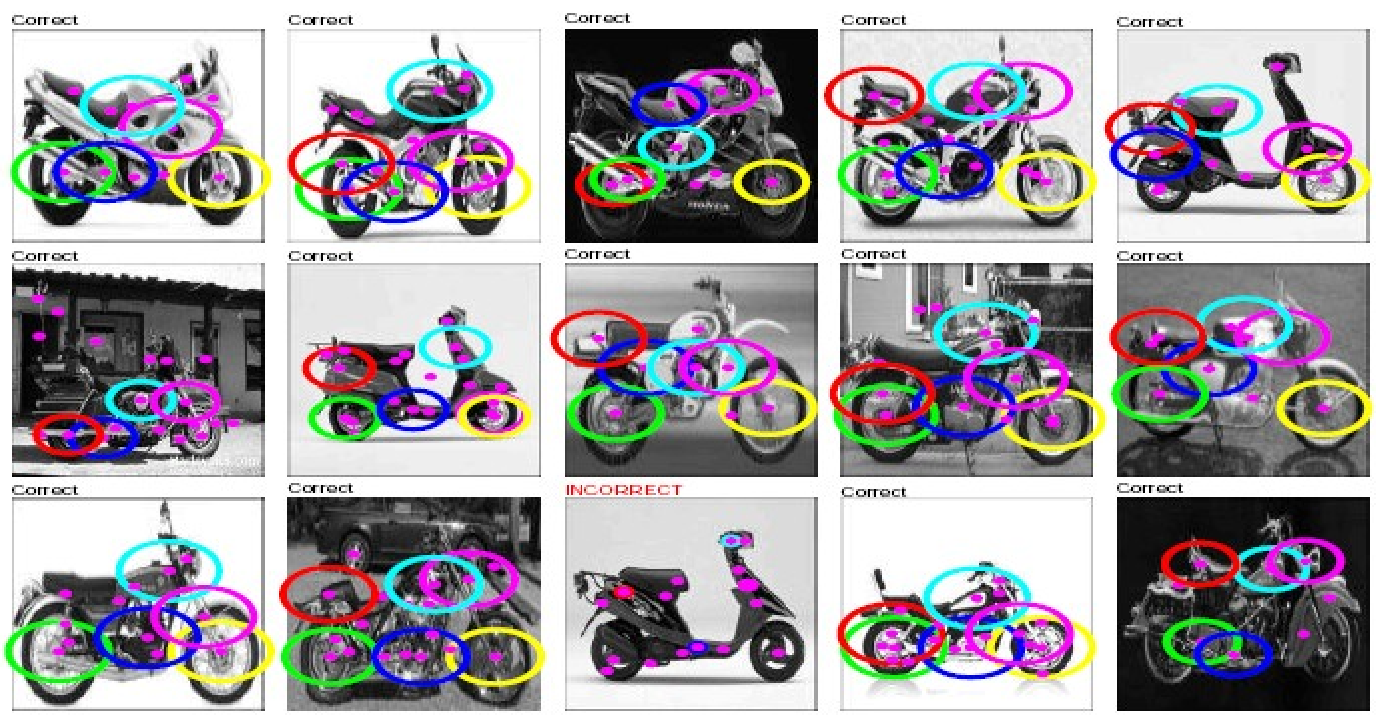




Example: motorbikes







Comparison (Caltech 4)

- Models are class-specific

Dataset	Total size of dataset	Object width (pixels)	Motorbike model	Face model	Airplane model	Cat model
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

- Models are robust to scale variation

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

Bayesian approach

L. Fei-Fei, R. Fergus and P. Perona. One-Shot learning of object categories. PAMI, 2006

Bayesian approach

- Fergus et al (I = X,S,A)

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

MLE approximation

- Fei-Fei et al.

$$
R \propto \frac{\int p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta},\mathcal{O}_{fg})p(\boldsymbol{\theta}|\mathcal{X}_t,\mathcal{A}_t,\mathcal{O}_{fg})\;d\boldsymbol{\theta}}{\int p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta}_{bg},\mathcal{O}_{bg})p(\boldsymbol{\theta}_{bg}|\mathcal{X}_t,\mathcal{A}_t,\mathcal{O}_{bg})\;d\boldsymbol{\theta}_{bg}}
$$

$$
= \boxed{\frac{\int p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{X}_t,\mathcal{A}_t,\mathcal{O}_{fg})\;d\boldsymbol{\theta}}{\int p(\mathcal{X},\mathcal{A}|\boldsymbol{\theta}_{bg})p(\boldsymbol{\theta}_{bg}|\mathcal{X}_t,\mathcal{A}_t,\mathcal{O}_{bg})\;d\boldsymbol{\theta}_{bg}}}.
$$

Parameter integration

Generative model for shape and appearance

- Foreground object (integrate over all hypotheses)

Latent variable

$$
\begin{aligned}
p(\mathcal{X}, \mathcal{A} | \boldsymbol{\theta}) &= \sum_{w=1}^{\Omega} \sum_{\mathbf{h} \in H} p(\mathcal{X}, \mathcal{A}, \boxed{\mathbf{h}, w} | \boldsymbol{\theta}) \\
&= \sum_{w=1}^{\Omega} p(w | \boldsymbol{\pi}) \sum_{\mathbf{h} \in H} \underbrace{p(\mathcal{A} | \mathbf{h}, \boldsymbol{\theta}_w^{\mathcal{A}})}_{Appearance} \underbrace{p(\mathcal{X} | \mathbf{h}, \boldsymbol{\theta}_w^{\mathcal{X}})}_{Shape} p(\mathbf{h} | \boldsymbol{\theta}_w)
\end{aligned}
$$

- In the paper, Ω=1 (Ω>1 can handle pose variation)
- Background (has a single null hypothesis)

$$
\begin{aligned}
p(\mathcal{X}, \mathcal{A} | \boldsymbol{\theta}_{bg}) &= p(\mathcal{X}, \mathcal{A}, \mathbf{h}_0 | \boldsymbol{\theta}_{bg}) \\
&= p\Big(\mathcal{A} | \mathbf{h}_0, \boldsymbol{\theta}_{bg}^{\mathcal{A}}\Big) p\Big(\mathcal{X} | \mathbf{h}_0, \boldsymbol{\theta}_{bg}^{\mathcal{X}}\Big) p(\mathbf{h}_0 | \boldsymbol{\theta}_{bg})
\end{aligned}
$$

Factorization

- Appearance

$$\frac{p(\mathbf{A}|\mathbf{X},\mathbf{S},\mathbf{h},\theta)}{p(\mathbf{A}|\mathbf{X},\mathbf{S},\mathbf{h},\theta_{bg})} = \prod_{p=1}^{P} \left(\frac{\mathrm{G}(\mathbf{A}(h_p)|\mathbf{c}_p, V_p)}{\mathrm{G}(\mathbf{A}(h_p)|\mathbf{c}_{bg}, V_{bg})} \right)^{d_p}$$

Fergus et al.

$$\frac{p\left(\mathcal{A}|\mathrm{h},\boldsymbol{\theta}_w^{\mathcal{A}}\right)}{p\left(\mathcal{A}|\mathrm{h}_0,\boldsymbol{\theta}_{bg}^{\mathcal{A}}\right)} = \prod_{p=1}^{P} \frac{\mathcal{G}\left(\mathcal{A}(\boldsymbol{h}_p)|\boldsymbol{\mu}_{p,w}^{\mathcal{A}}, \boldsymbol{\Gamma}_{p,w}^{\mathcal{A}}\right)}{\mathcal{G}\left(\mathcal{A}(\boldsymbol{h}_p)|\boldsymbol{\mu}_{bg}^{\mathcal{A}}, \boldsymbol{\Gamma}_{bg}^{\mathcal{A}}\right)}$$

Fei Fei et al.

- Shape

$$\frac{p(\mathbf{X}|\mathbf{S},\mathbf{h},\theta)}{p(\mathbf{X}|\mathbf{S},\mathbf{h},\theta_{bg})} = \mathrm{G}(\mathbf{X}(\mathbf{h})|\boldsymbol{\mu},\Sigma)\,\alpha^f$$

Fergus et al.

$$\frac{p\left(\mathcal{X}|\mathrm{h},\boldsymbol{\theta}_w^{\mathcal{X}}\right)}{p\left(\mathcal{X}|\mathrm{h}_0,\boldsymbol{\theta}_{bg}^{\mathcal{X}}\right)} = \alpha^{P-1}\,\mathcal{G}\left(\mathcal{X}(\mathrm{h})|\boldsymbol{\mu}_w^{\mathcal{X}}, \boldsymbol{\Gamma}_w^{\mathcal{X}}\right)$$

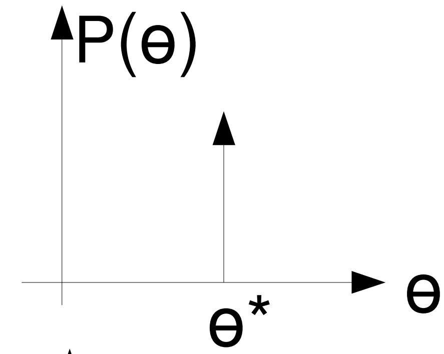
Fei Fei et al.

- **Scale and occlusion are not modeled**

Comparison

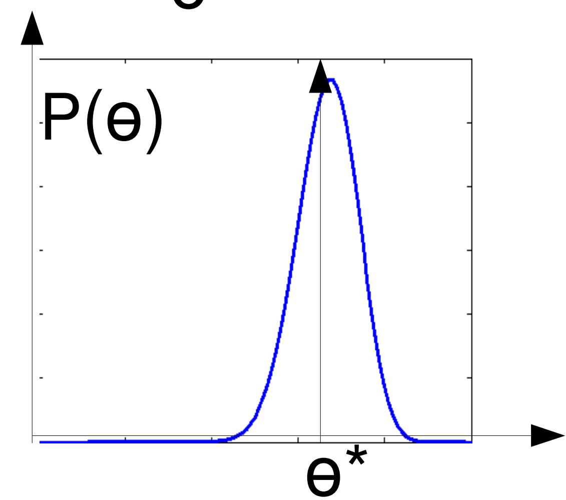
- MLE

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{\mathrm{ML}} = \underset{\theta}{argmax}\, p(\mathcal{X}_t, \mathcal{A}_t|\boldsymbol{\theta})$$



- MAP

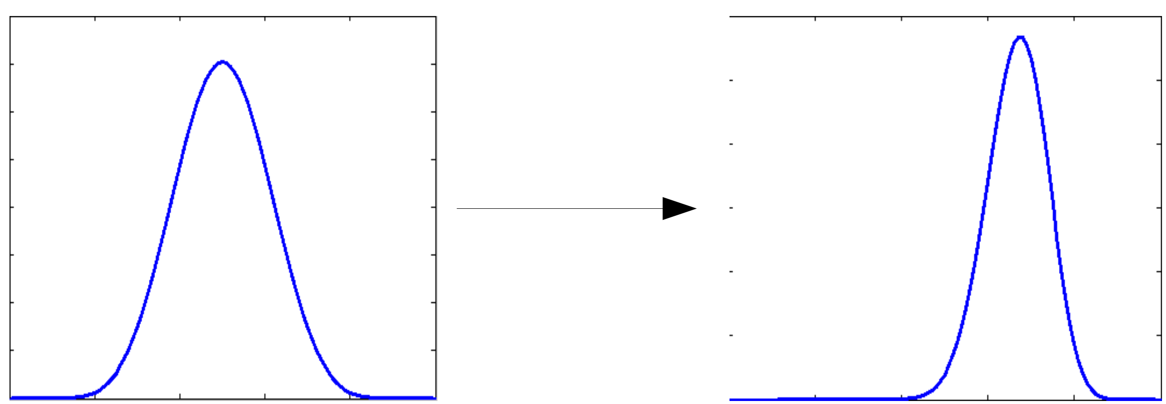
$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{\mathrm{MAP}} = \underset{\theta}{argmax}\, p(\mathcal{X}_t, \mathcal{A}_t|\boldsymbol{\theta})p(\boldsymbol{\theta})$$



- Bayesian

$$\bar{\boldsymbol{\theta}} = \{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Gamma}^{\mathcal{X}}, \boldsymbol{\Gamma}^{\mathcal{A}}\}$$

$$p(\boldsymbol{\theta}) \longrightarrow p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})$$



Conjugate priors

- Parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}^{\mathcal{X}}, \boldsymbol{\mu}^{\mathcal{A}}, \boldsymbol{\Gamma}^{\mathcal{X}}, \boldsymbol{\Gamma}^{\mathcal{A}}\}$
- Priors

$$p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t) = p(\boldsymbol{\pi}) \prod_{\omega} p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) p(\boldsymbol{\mu}_{\omega}^{\mathcal{A}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{A}}) p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{A}})$$

$$p(\boldsymbol{\pi}) = \mathcal{D}ir(\lambda_{\omega} \boldsymbol{I}_{\Omega}) \qquad p(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) = \mathcal{W}(\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}|a_{\omega}^{\mathcal{X}}, \boldsymbol{B}_{\omega}^{\mathcal{X}}) \qquad p(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{\Gamma}_{\omega}^{\mathcal{X}}) = \mathcal{G}(\boldsymbol{\mu}_{\omega}^{\mathcal{X}}|\boldsymbol{m}_{\omega}^{\mathcal{X}}, \beta_{\omega}^{\mathcal{X}} \boldsymbol{\Gamma}_{\omega}^{\mathcal{X}})$$

Dirichlet **Wishart** **Normal**

- Hyper-parameters $\{\lambda_{\omega}, a_{\omega}, \boldsymbol{B}_{\omega}, \boldsymbol{m}_{\omega}, \beta_{\omega}\}$
- Closed form solution

$$p(\mathcal{X}, \mathcal{A}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg}) = \int p(\mathcal{X}, \mathcal{A}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{X}_t, \mathcal{A}_t, \mathcal{O}_{fg})\, d\boldsymbol{\theta}$$

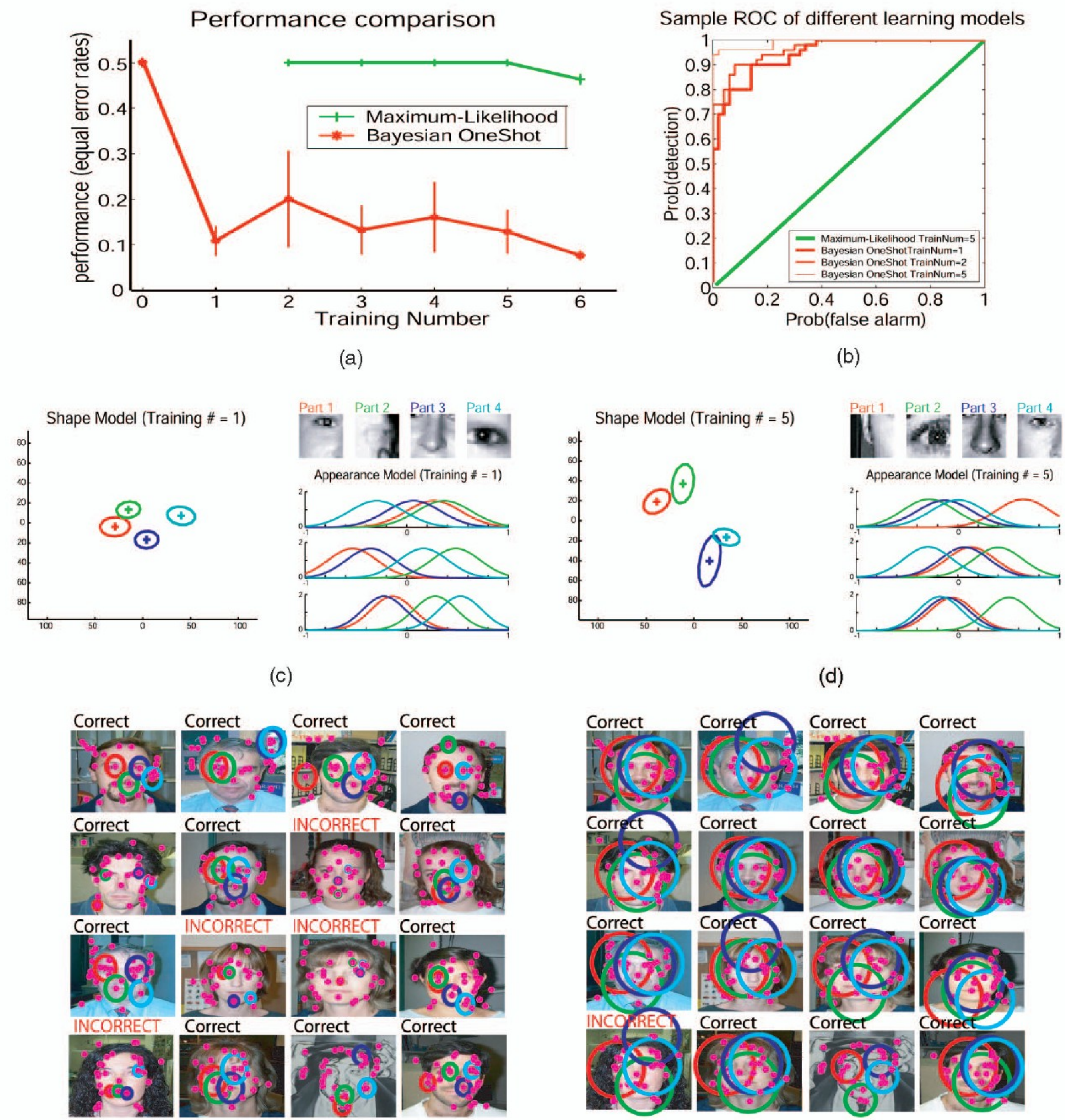
$$\sum_{\omega=1}^{\Omega} \sum_{h=1}^{|H|} \tilde{\pi}_{\omega}\, \mathcal{S}(\mathcal{X}_h|\, g_{\omega}^{\mathcal{X}}, \mathbf{m}_{\omega}^{\mathcal{X}}, \boldsymbol{\Lambda}_{\omega}^{\mathcal{X}})\; \mathcal{S}\Big(\mathcal{A}_h|\, g_{\omega}^{\mathcal{A}}, \mathbf{m}_{\omega}^{\mathcal{A}}, \boldsymbol{\Lambda}_{\omega}^{\mathcal{A}}\Big),$$

$$\text{where } g_{\omega} = a_{\omega} + 1 - d \text{ and } \boldsymbol{\Lambda}_{\omega} = \frac{\beta_{\omega} + 1}{\beta_{\omega} g_{\omega}} \boldsymbol{B}_{\omega} \text{ and } \tilde{\pi}_{\omega} = \frac{\lambda_{\omega}}{\sum_{\omega'} \lambda_{\omega'}}$$

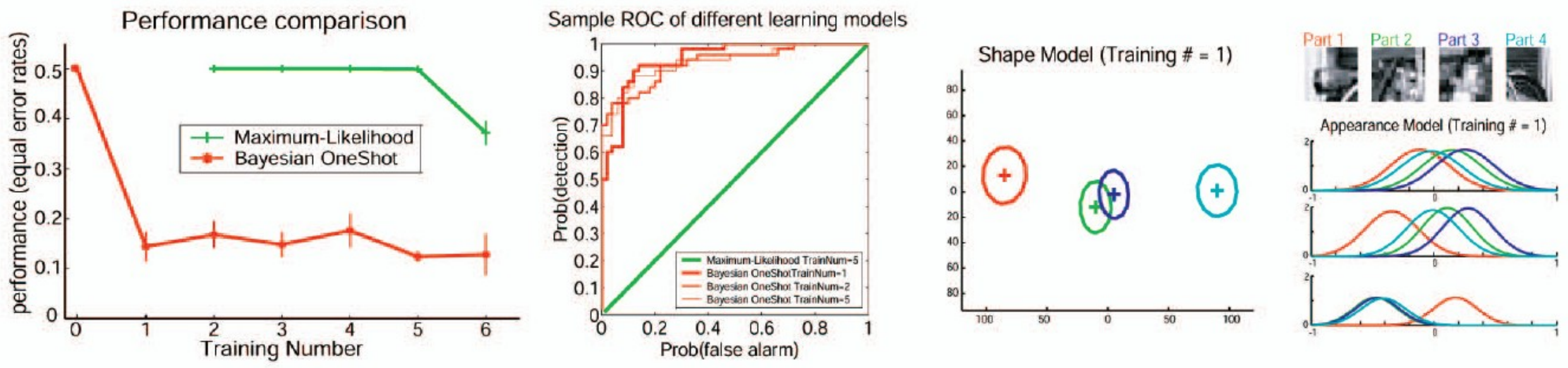
Learning

- $p(x|\theta) = \sum p(x,h|\theta)p(h|\theta)$, h-unknown, but 'convenient'
- Regular EM
  - E-step: Estimate $p(h|x,\theta^n)$, $Q(\theta)=E_h\{ \log(p(x,h|\theta)) \mid h\}$
  - (Usually available in closed form)
  - M-step: $\Theta^{n+1} = \text{argmax } Q(\theta)$
- Variational (EM)
  - Getting $p(h|x,\theta^n)$ is hard-no closed form
  - $p(h|x,\theta^n) \sim q(x)$, approximate the posterior

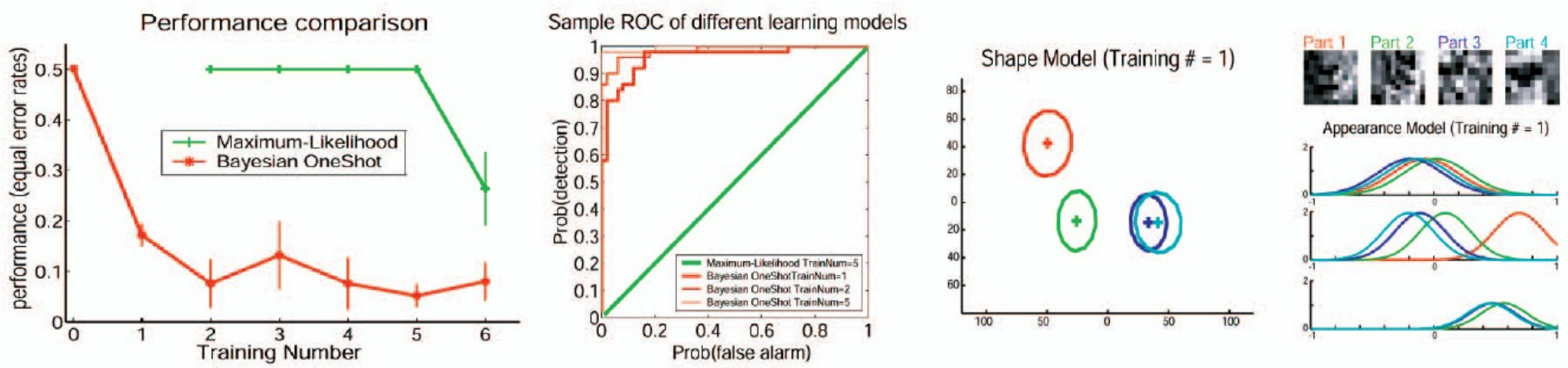
Performance: Caltech 4



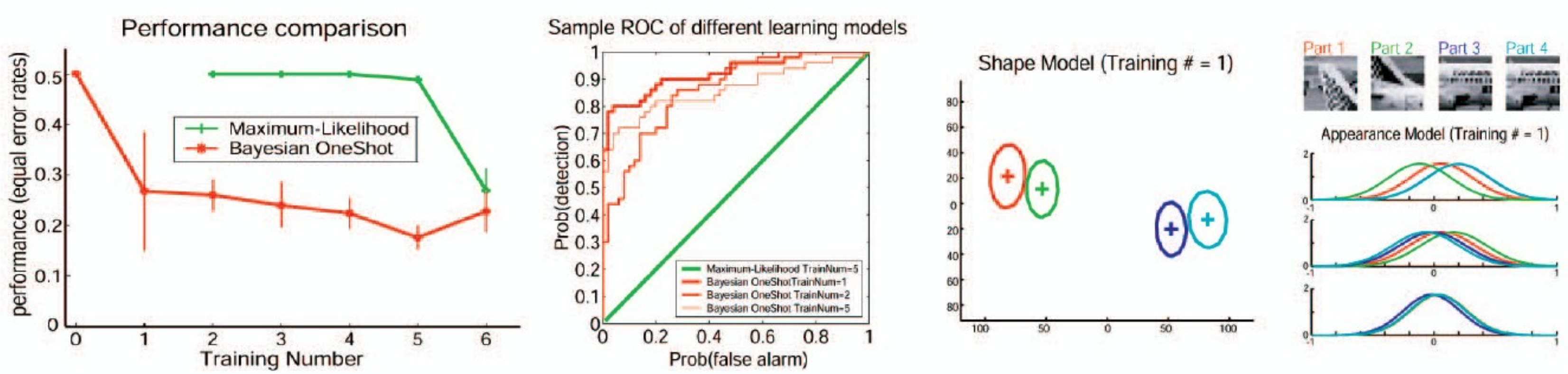
Caltech 4 (cont.)



(a)

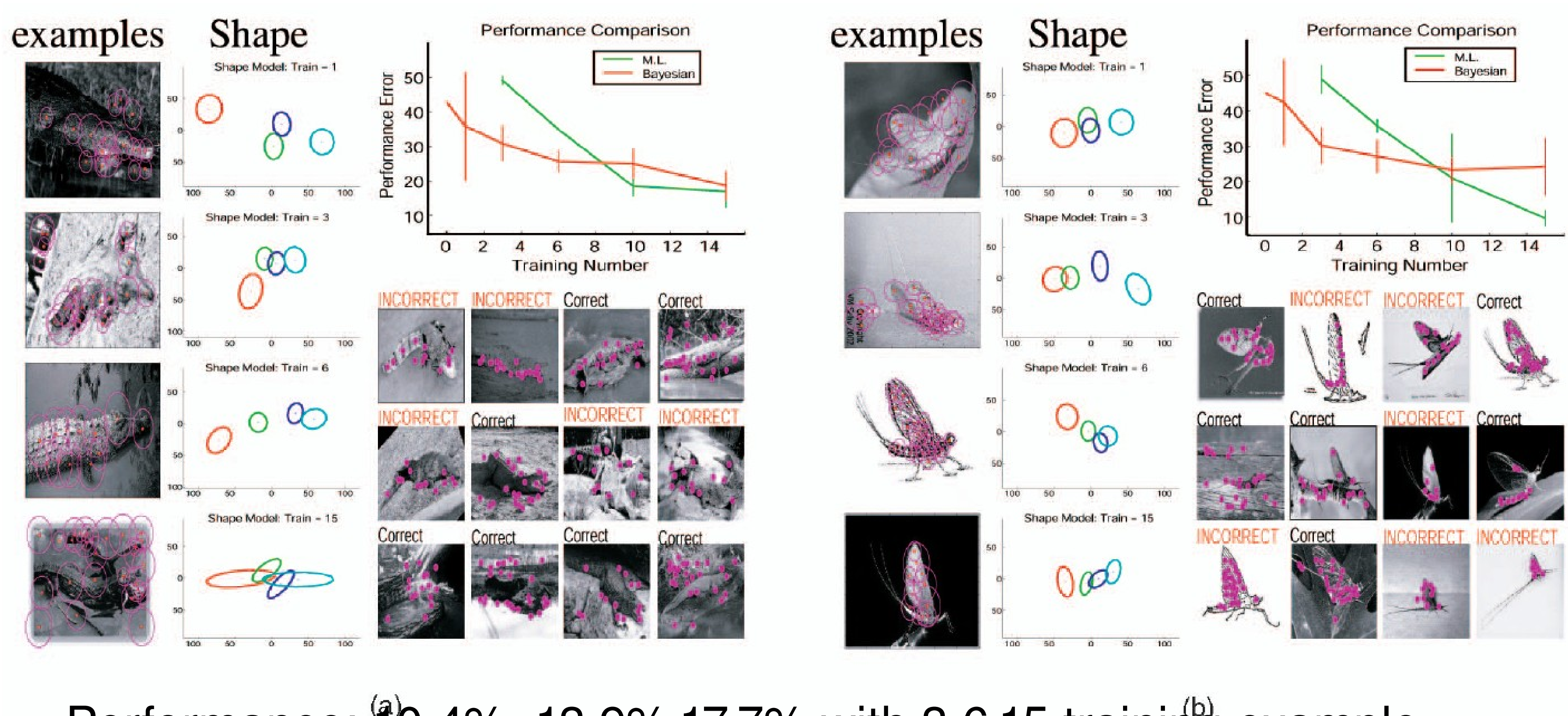


(b)



(c)

Performance : Caltech 101



- Performance: 10.4%, 13.9%,17.7% with 3,6,15 training example
- State-of-the-art : > 60%

Comparison

Authors	Categories	# Categories	# Training images	Framework	Hand alignment	Segmented
Fei-Fei *et al.*	Assorted	101	1-5	Gen.	N	N
Fergus *et al.* [14]	Assorted	6	> 100	Gen.	N	N
Weber *et al.* [40]	Cars, Faces	2	> 100	Gen. + Disc.	N	N
Viola & Jones [37]	Faces	1	$\sim 10{,}000$	Disc.	Y	Y
Schneiderman & Kanade [35]	Cars	1	2,000	Disc.	Y	N
Rowley *et al.* [33]	Cars	1	500	Disc.	Y	N
Amit *et al.* [2]	Faces, Characters	3	300	Gen.	Y	Y
LeCun *et al.* [25]	Digits	10	60,000	Disc.	N	Y
LeCun *et al.* [26]	Assorted	5	$\sim$300,000	Disc.	Y	N

Summary

- Transfer learning in a Bayesian setting
- Recipe = learning priors on given data+ updating priors on new data
- Good results with just 1~5 training examples (compared to MLE approaches)
- Learning is hard (computationally)
- Wishlist
  - Handling multiple objects within the image.

Thank You!