Unsupervised Learning of Models for Object Class Recognition

CS 395T
Object Recognition

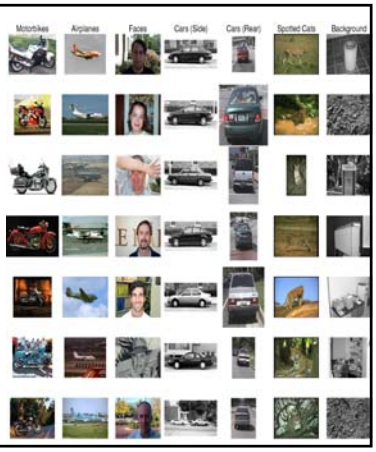
Tuyen Huynh
Prateek Jain

* Slides taken from R. Fergus, P. Perona and A. Zisserman

Goal

- Recognition of object categories
- Unassisted learning

Some object categories



Learn from examples

Difficulties:

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

Problems

- Three problems
  - Segmentation of training images
  - Part selection
  - Estimation of model parameters

→Provide framework to solve these problem automatically

Related Work

- Hierarchical model from edge elements
- Statistical model from shape space densities
- Active appearance models
- Gradient descent on a deformation energy function

→ Require some kinds of labeled in the training images

Approach

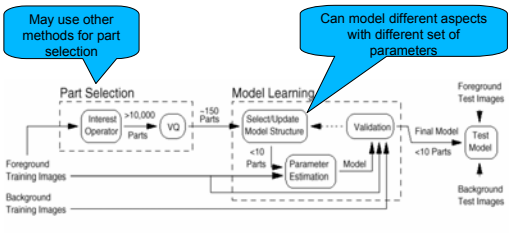
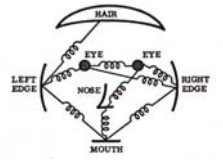
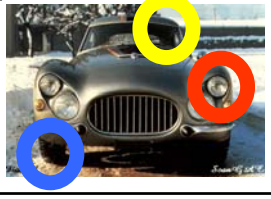


Fig. 2. Block diagram of our method. "Foreground images" are images containing the target objects in cluttered background. "Background images" contain background only.

Model: Constellation of Parts



Object composition by parts and shape



Part selection I

- Interest operator by Forstner
  - corners
  - intersection of 2+ lines
  - center points of circular patterns
- Vector quantization by k-means clustering
  - Retain clusters with at least 10 patterns
- Remove patterns which are similar to others
- Use greedy search to find the most informative parts

Part selection I

Fig. 3. Points of interest (left) identified on a training image of a human face in cluttered background using Förstner's method. Crosses denote corner-type patterns while circles mark circle-type patterns. A sample of the patterns obtained using k-means clustering of small image patches is shown for faces (center) and cars (right). The car images were high-pass filtered before the part selection process. The total number of patterns selected were 81 for faces and 80 for cars.

A generative object model I

- Assume T different types of parts
- Observable data

$$X^0 = \begin{pmatrix} x_{11} & x_{12} & x_{1N_1} \\ x_{21} & x_{22} & x_{2N_2} \\ & & \\ x_{T1} & x_{T2} & x_{TN_T} \end{pmatrix}$$

2-dimensional vector

Row t = All locations of part type t in the image

- Hidden/Missing data:
  - Vector $\mathbf{h}=(h_1, h_2, \ldots, h_T)$ where $h_i \in \{0,1,2,\ldots,N_i\}$
  - Vector $\mathbf{x^m}$: positions of missing parts

A generative object model I

- Joint probability density: $P(X^0,x^m,h)$
- Two auxiliary variables:
  - Binary vector $\mathbf{b}=(b_1,b_2,\ldots,b_F)$ where $b_i = 1$ if $h_i > 0$
  - Vector $\mathbf{n}=(n_1,n_2,\ldots,n_T)$ where n_i is the number of background candidates in row $\mathbf{i}$ of $\mathbf{X^0}$

$\rightarrow P(X^0,x^m,h) = P(X^0, x^m, h, n, b)$

$= P(X^0,x^m|h,n)p(h|n,b)p(n)p(b)$

A generative model I

$$p(\mathbf{n}) = \prod_{t=1}^{T} \frac{1}{n_t!}(M_t)^{n_t} e^{-M_t}$$

M_t: the average number of background of type t per image

$$p(\mathbf{h}|\mathbf{n},\mathbf{b}) = \begin{cases} \frac{1}{\prod_{f=1}^{F} N_f^{h_f}} & \mathbf{h} \in \mathcal{H}(\mathbf{b},\mathbf{n}) \\ 0 & \text{other } \mathbf{h} \end{cases}$$

H(b,n): the set of all hypotheses consitent with b and n

N_f: the total number of dectections of the type of part f

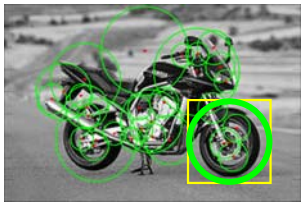
$$p(X^o, \mathbf{x}^m|\mathbf{h},\mathbf{n}) = p_{\mathbf{fg}}(\mathbf{z})\, p_{\mathbf{bg}}(\mathbf{x}_{bg})$$

$z^T=(x^{oT}x^{mT})$: the coordinates of all foreground detections. Modeled by Gaussian(μ,Σ)

$$p_{\mathbf{bg}}(\mathbf{x}_{bg}) = \prod_{t=1}^{T} \frac{1}{A^{n_t}}$$

X_{bg}: the coordinates of all background detections. Modeld by an uniform density.

Part Selection II



- Find regions within image
- Use Kadir and Brady's salient region operator [IJCV '01]

Location

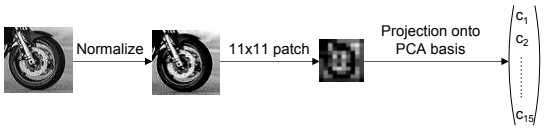
(x,y) coords. of region center

Scale

Diameter of region (pixels)

- Uses gray-scale input
- Finds maxima in entropy over scale and location

Appearance



- Normalize all the images to same size
- Representation in low-dimensional vector space, hence speeds up computation
- Ignores noise, hence makes algorithm robust

Model II

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

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

- Appearance is highest level abstraction
- Occlusion is lowest level abstraction
- h – Hypothesis vector, mapping a feature to each part
- If h (*p*)=0, part *p* is occluded

Model II

- Appearance
  - Each part has Gaussian Density in Appearance Space
  - Independent for each part
  - Gaussian density for background
  - Feature not in hypothesis calculated under background density
  - d_p=0 if part occluded
  - Parameters: $\theta_p^{app} = \{\mathbf{c_p}, V_p\}$ $\theta_{bg}^{app} = \{\mathbf{c_{bg}}, V_{bg}\}$,

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

Model II

- Shape
  - Same as Weber et al.'s model
  - Joint Gaussian density for foreground features
  - Uniform density for background features
  - Parameters: $\theta^{shape} = \{\mu, \Sigma\}$

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

Model II

- Relative Scale
  - Scale with reference to a fixed frame
  - Modeled by Gaussian density
  - Parts assumed independent of each other
  - Parameters $\theta^{scale} = \{t_p, \hat{U}_p\}$

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

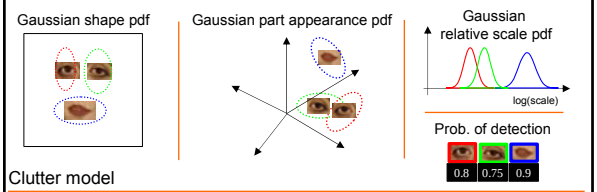
Model II

- Occlusion
  - Same as Weber et al.'s model
  - Number of features detected modeled using Poisson distribution
  - Probability table for all possible occlusion patterns is a parameter

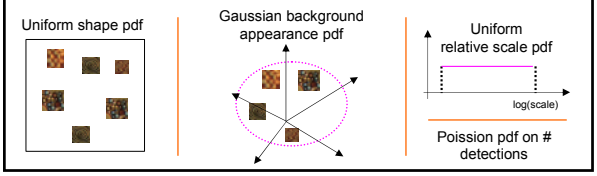
$$\frac{p(\mathbf{h}|\theta)}{p(\mathbf{h}|\theta_{bg})} = \frac{p_{Poiss}(n|M)}{p_{Poiss}(N|M)} \frac{1}{{}^{n}C_{r}(N,f)} p(\mathbf{d}|\theta)$$

Generative probabilistic model

Foreground model — based on Burl, Weber et al. [ECCV '98, '00]



Clutter model

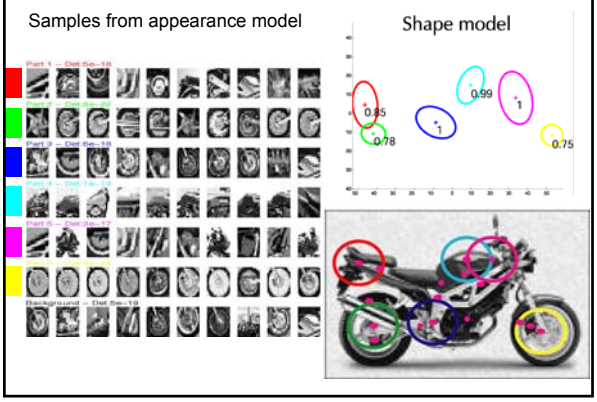


Classification

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

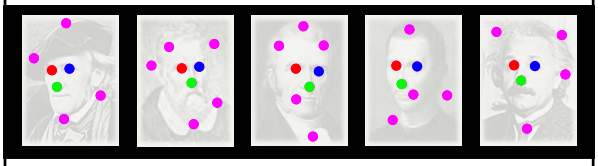
- Priors estimated using training set
- Object iff R>threshold

Motorbikes



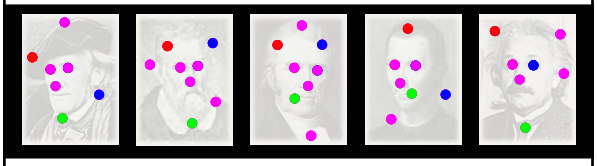
Learning

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



Learning procedure

- Find regions & their location, scale & appearance
- Initialize model parameters
- Use EM and iterate to convergence:
  - E-step: Compute assignments for which regions are foreground / background
  - M-step: Update model parameters
- Trying to maximize likelihood – consistency in shape & appearance



Experiments

Experimental procedure

Two series of experiments:

- Fixed-scale model - Objects the same size (manual normalization)
- Scale-invariant model - Objects between 100 and 550 pixels in width

Datasets

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

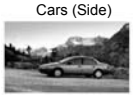
Motorbikes

Airplanes

Frontal Faces

Testing
- 50% images
- Simple object present/absent test

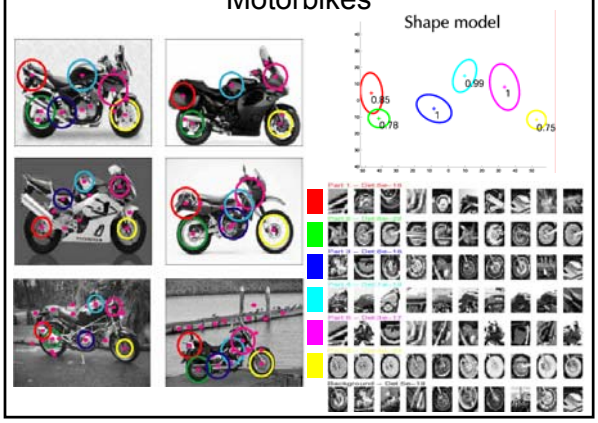
Cars (Side)



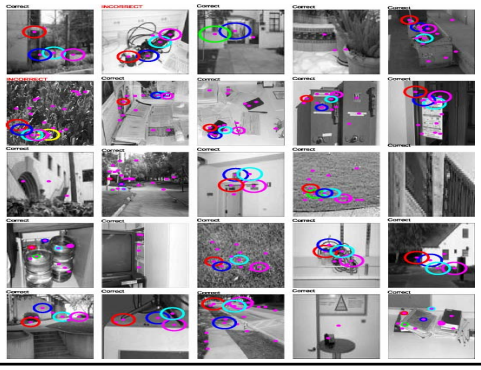
Cars (Rear)

Spotted cats

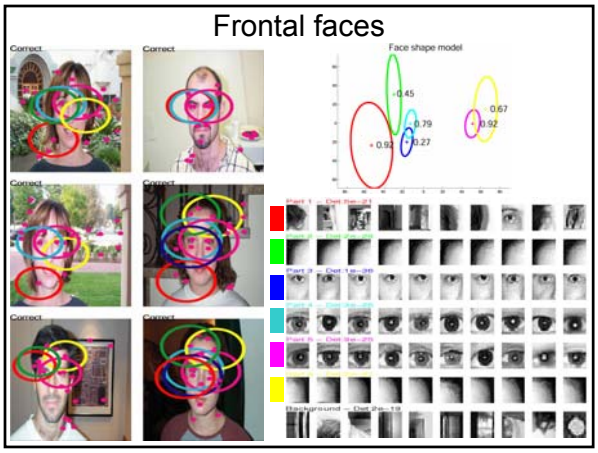
Motorbikes



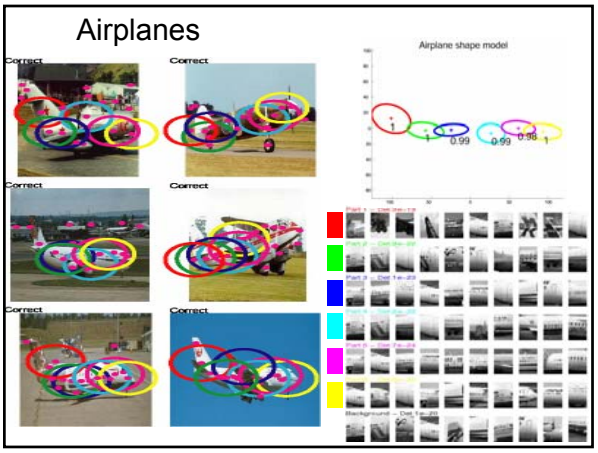
Background images evaluated with motorbike model



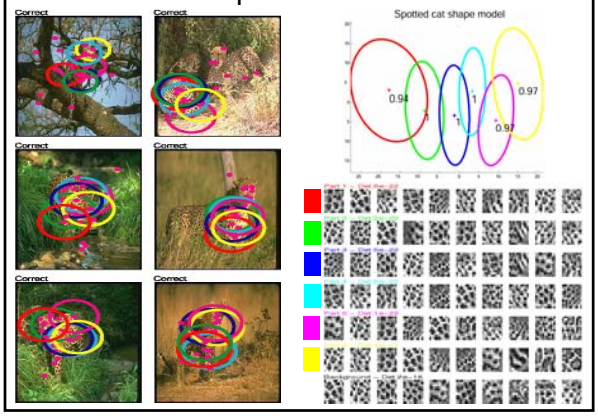
Frontal faces



Airplanes



Spotted cats



Summary of results

Dataset	Fixed scale experiment	Scale invariant experiment
Motorbikes	7.5	6.7
Faces	4.6	4.6
Airplanes	9.8	7.0
Cars (Rear)	15.2	9.7
Spotted cats	10.0	10.0

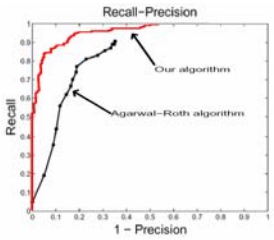
% equal error rate

Note: Within each series, same settings used for all datasets

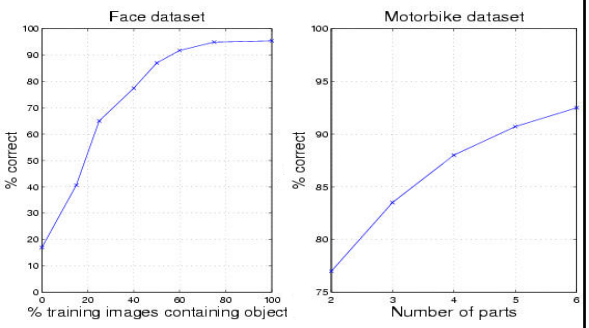
Comparison to other methods

Dataset	Ours	Others	
Motorbikes	7.5	16.0	Weber et al. [ECCV '00]
Faces	4.6	6.0	Weber
Airplanes	9.8	32.0	Weber
Cars (Side)	11.5	21.0	Agarwal Roth [ECCV '02]

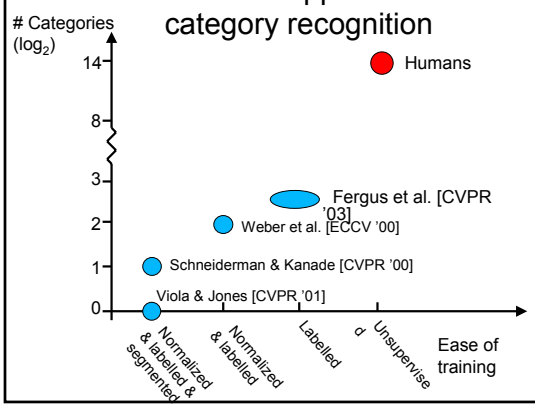
% equal error rate



Robustness of Algorithm



Overview of approaches to category recognition



Summary

- Comprehensive probabilistic model for object classes
- Learn appearance, shape, relative scale, occlusion etc. simultaneously in scale and translation invariant manner
- Same algorithm gives <= 10% error across 5 diverse datasets with identical settings

Limitations → future work

- Very reliant on region detector
  - Different part types (e.g. edgel curves)
- Only learns a single viewpoint
  - Use mixture models
- Need lots of images to learn
  - Bayesian learning - fewer images [ICCV '03 (Fei Fei, Fergus, Perona)]
- Need more through testing