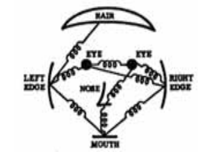
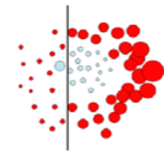
Recognizing and Learning Object Categories: Year 2007

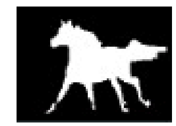
Li Fei-Fei, Princeton
Rob Fergus, MIT
Antonio Torralba, MIT

Agenda

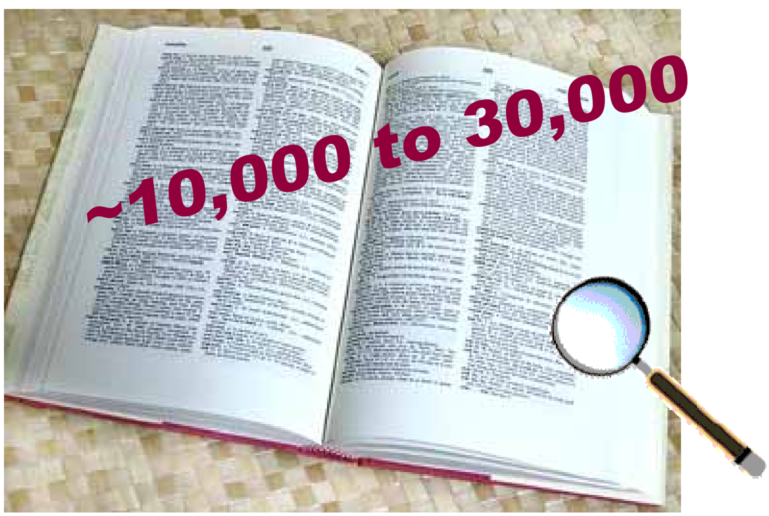
- Introduction
- Bag-of-words models
- Part-based models
- Discriminative methods
- Segmentation and recognition
- Datasets & Conclusions







How many object categories are there?



Biederman 1987

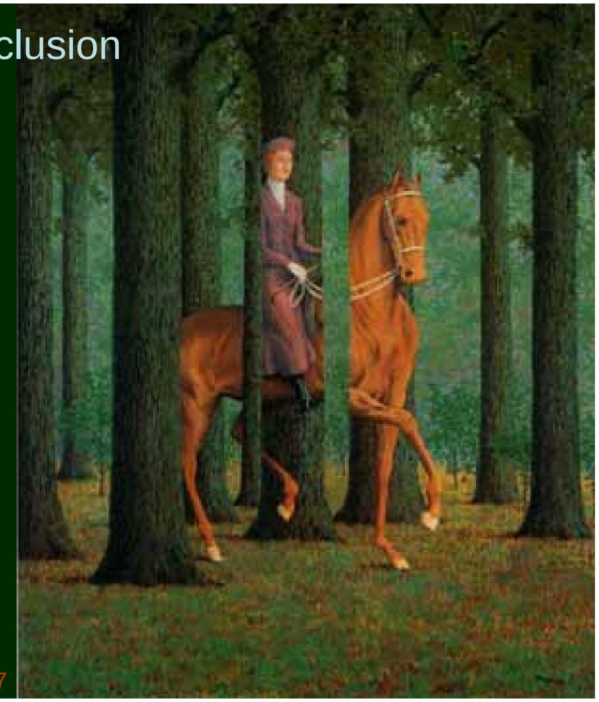
Challenges 1: view point variation

Michelangelo 1475-1564

Challenges 2: illumination

slide credit: S. Ullman

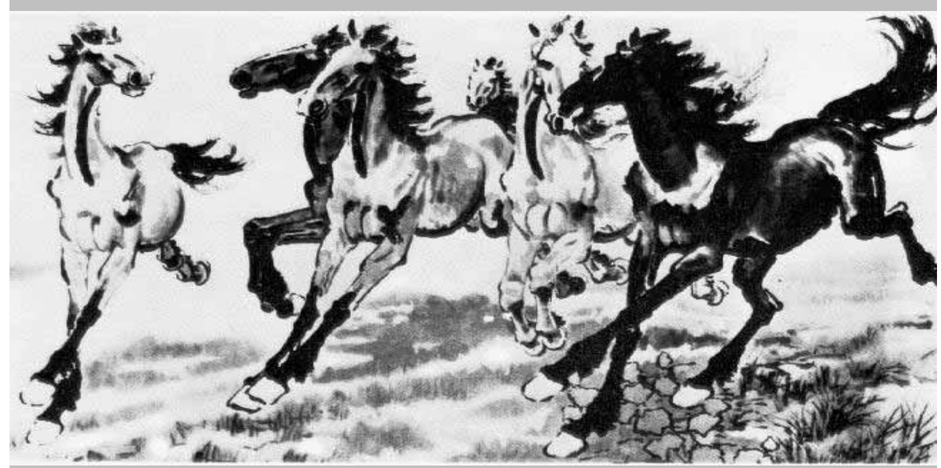
Challenges 3: occlusion



Magritte, 1957

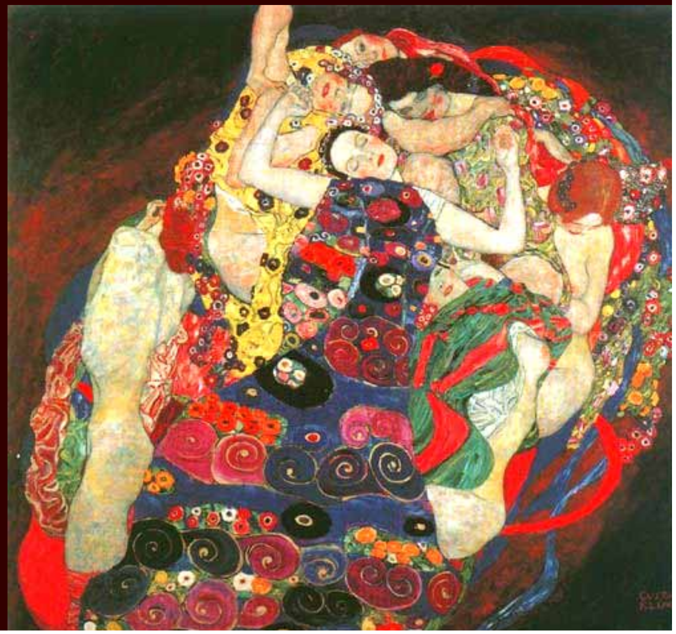
Challenges 4: scale

Challenges 5: deformation



Xu, Beihong 1943

Challenges 6: background clutter



Klimt, 1913

History: single object recognition

History: single object recognition

- Lowe, et al. 1999, 2003
- Mahamud and Herbert, 2000
- Ferrari, Tuytelaars, and Van Gool, 2004
- Rothganger, Lazebnik, and Ponce, 2004
- Moreels and Perona, 2005
- …

Challenges 7: intra-class variation

History: early object categorization

- Turk and Pentland, 1991
- Belhumeur, Hespanha, & Kriegman, 1997
- Schneiderman & Kanade 2004
- Viola and Jones, 2000

- Amit and Geman, 1999
- LeCun et al. 1998
- Belongie and Malik, 2002

- Schneiderman & Kanade, 2004
- Argawal and Roth, 2002
- Poggio et al. 1993

Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

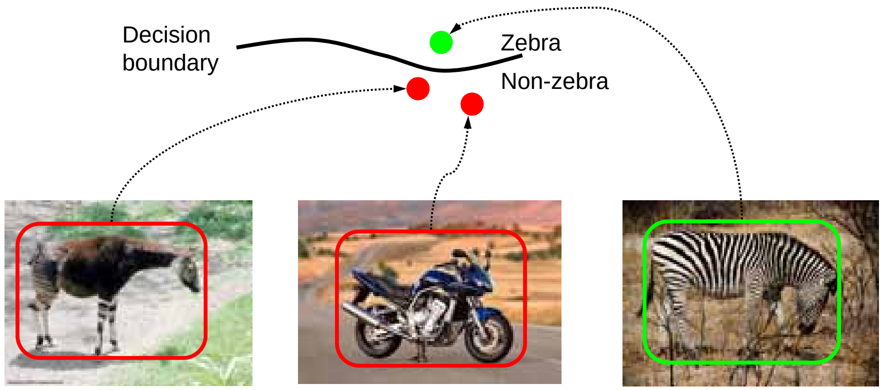
Object categorization: the statistical viewpoint

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

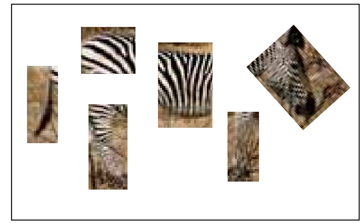
Discriminative

- Direct modeling of $\dfrac{p(zebra \mid image)}{p(no\ zebra \mid image)}$



Generative

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$



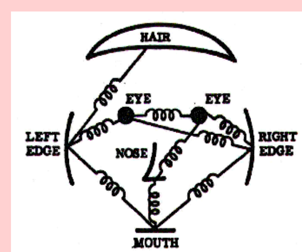
	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

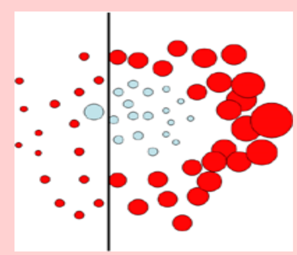
Three main issues

- Representation
  - How to represent an object category
- Learning
  - How to form the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

Representation

- Generative / discriminative / hybrid

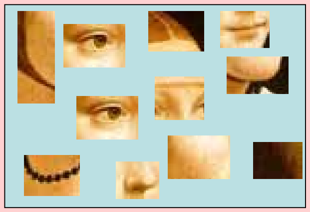


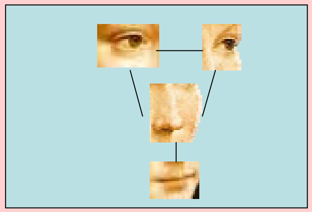


Representation

– Generative / discriminative / hybrid

– Appearance only or location and appearance





Representation

– Generative / discriminative / hybrid

– Appearance only or location and appearance

– Invariances

- View point
- Illumination
- Occlusion
- Scale
- Deformation
- Clutter
- etc.

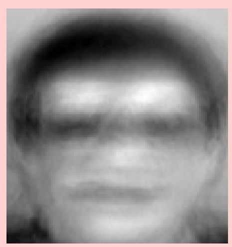
Representation

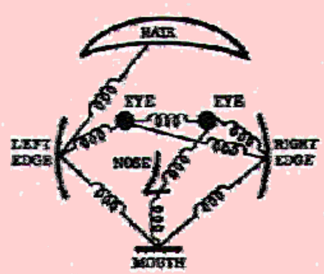
– Generative / discriminative / hybrid

– Appearance only or location and appearance

– invariances

– Part-based or global w/sub-window





Representation

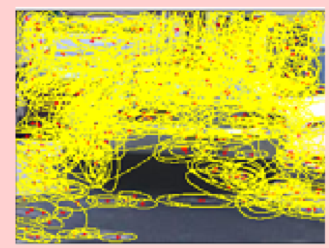
– Generative / discriminative / hybrid

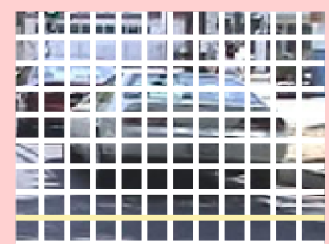
– Appearance only or location and appearance

– invariances

– Parts or global w/sub-window

– Use set of features or each pixel in image





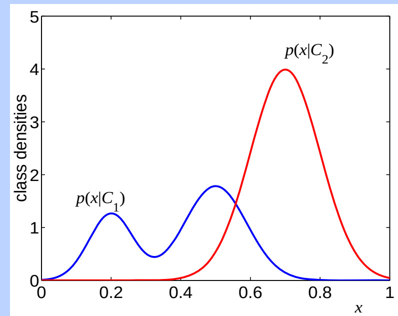
Learning

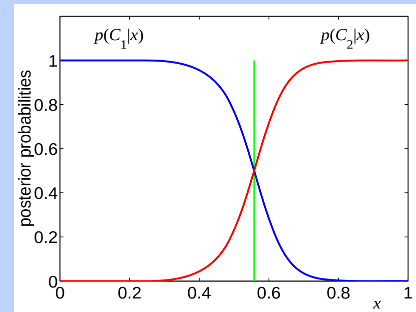
– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)

– Methods of training: generative vs. discriminative





Learning

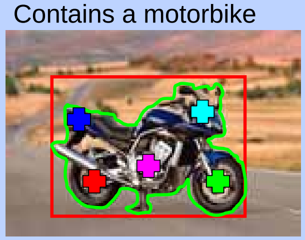
– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)

– What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

– Level of supervision

- Manual segmentation; bounding box; image labels; noisy labels

Contains a motorbike



Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)

– What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

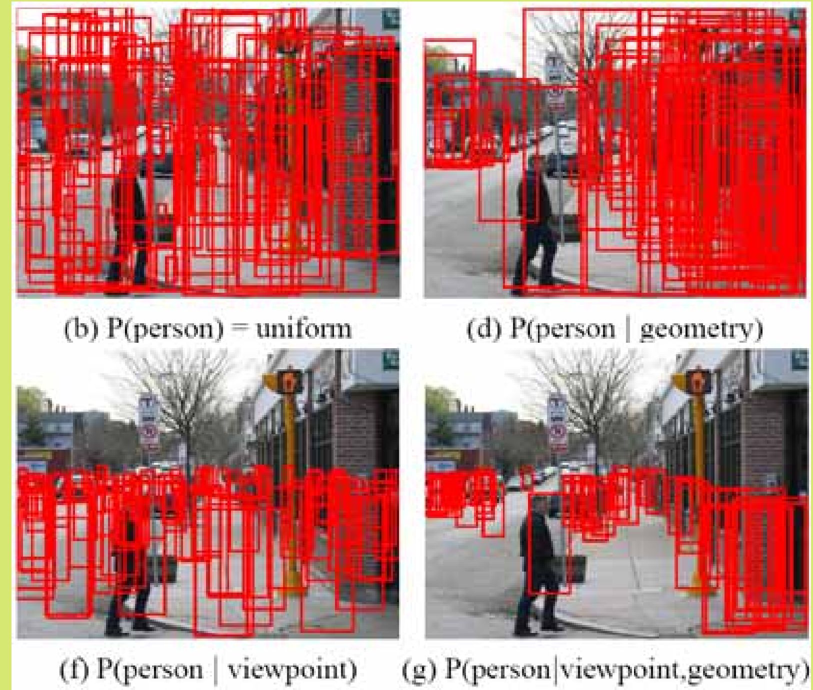
– Level of supervision

- Manual segmentation; bounding box; image labels; noisy labels

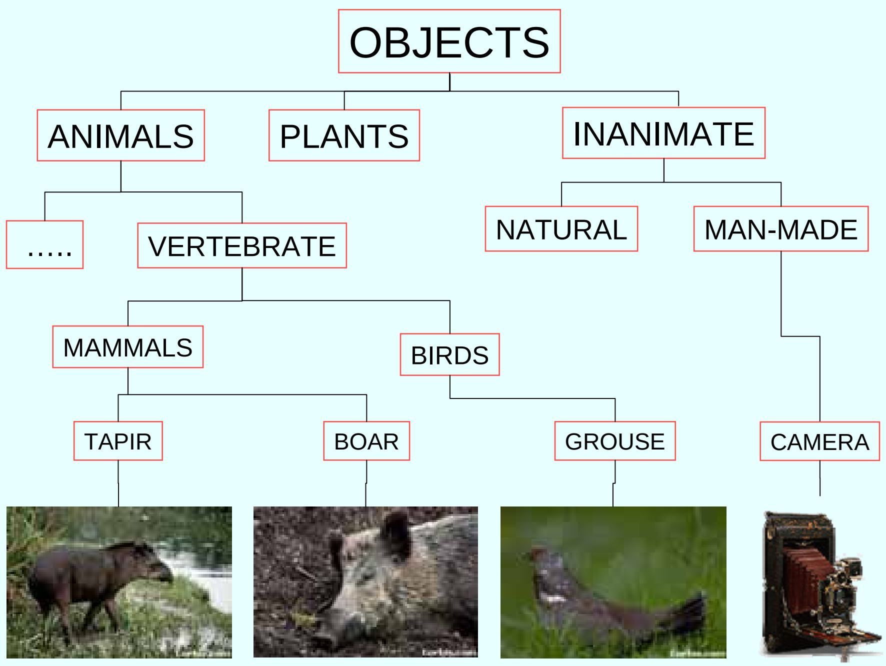
– Batch/incremental (on category and image level; user-feedback)

Recognition

- Scale / orientation range to search over
- Speed
- Context



Hoiem, Efros, Herbert, 2006



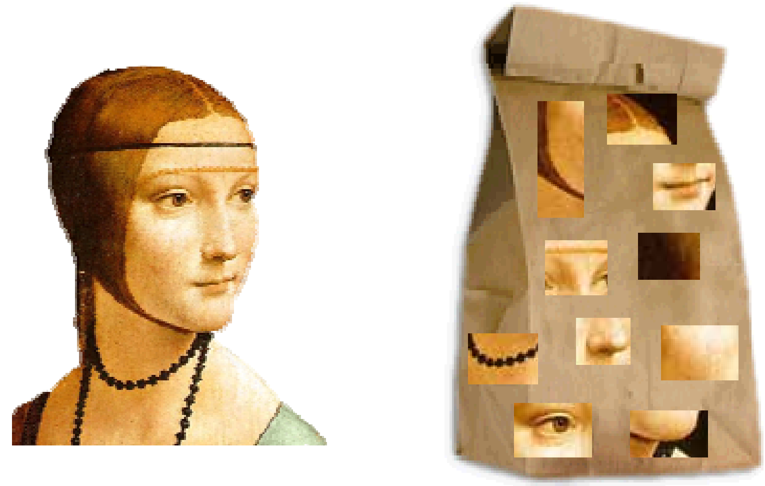
Part 1: Bag-of-words models

by Li Fei-Fei (Princeton)

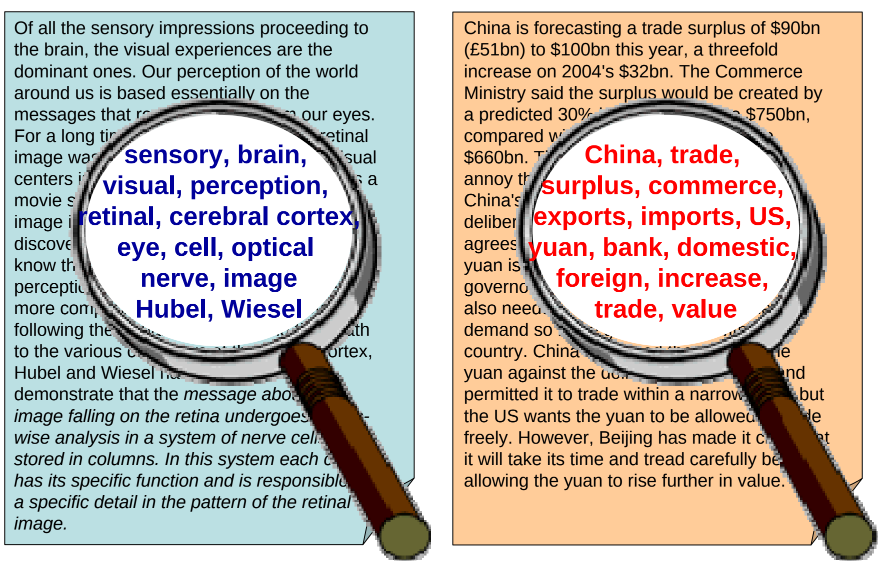
Related works

- Early "bag of words" models: mostly texture recognition
  - Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003;
- Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
  - Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004
- Object categorization
  - Csurka, Bray, Dance & Fan, 2004; Sivic, Russell, Efros, Freeman & Zisserman, 2005; Sudderth, Torralba, Freeman & Willsky, 2005;
- Natural scene categorization
  - Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman & Munoz, 2006

Object → **Bag of 'words'**



Analogy to documents

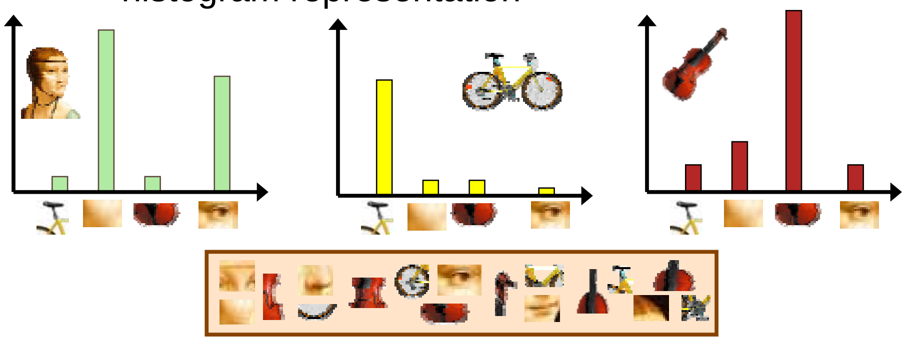


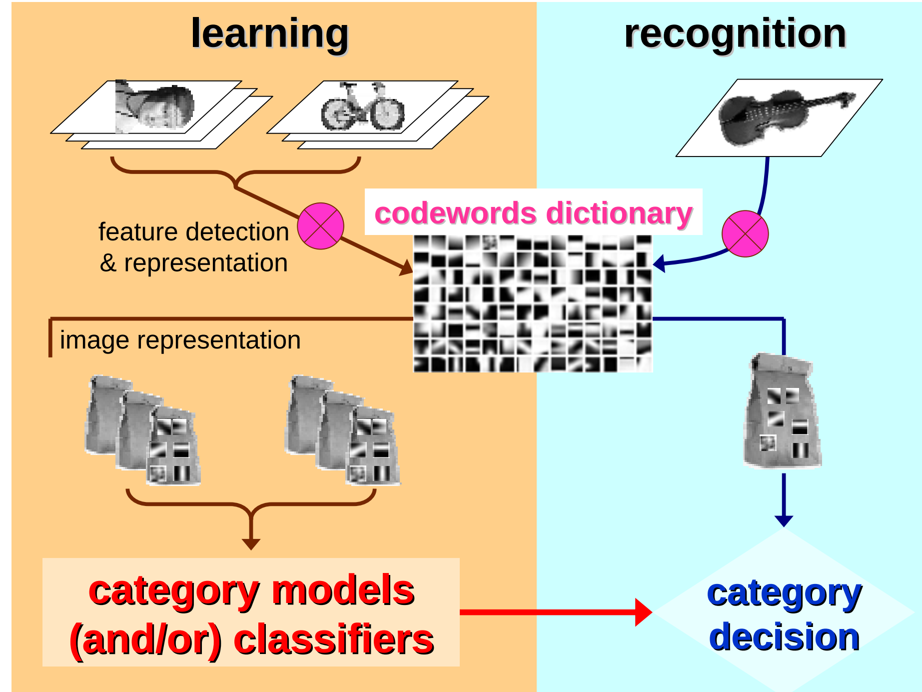
A clarification: definition of "BoW"

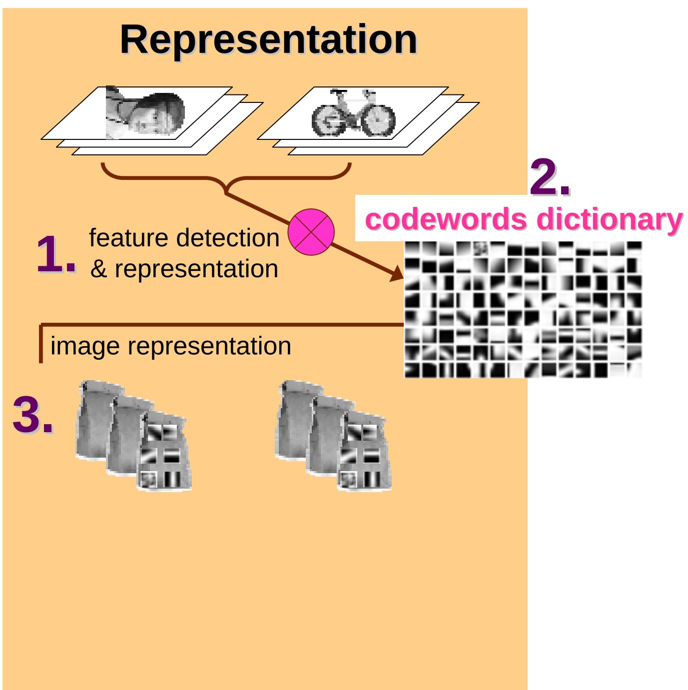
- Looser definition
  - Independent features

A clarification: definition of "BoW"

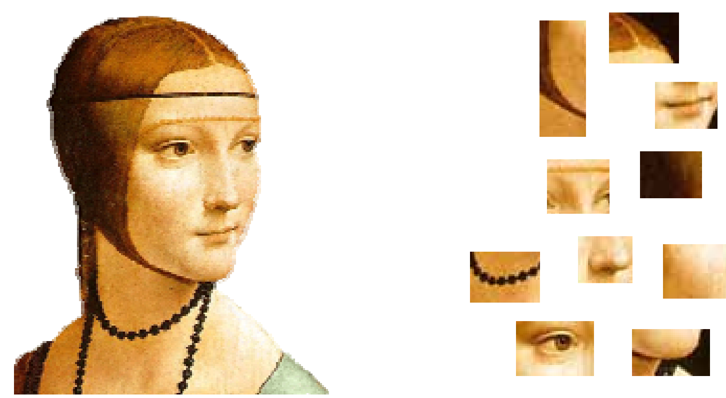
- Looser definition
  - Independent features
- Stricter definition
  - Independent features
  - histogram representation





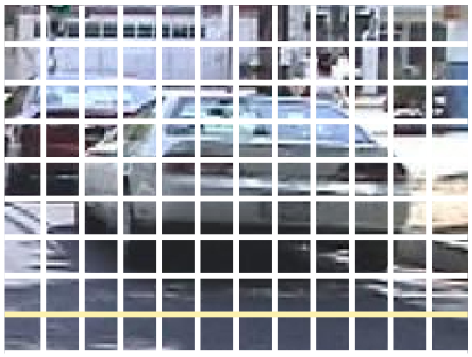


1.Feature detection and representation



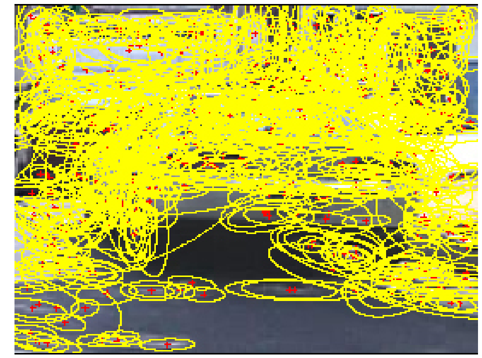
1.Feature detection and representation

- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005



1.Feature detection and representation

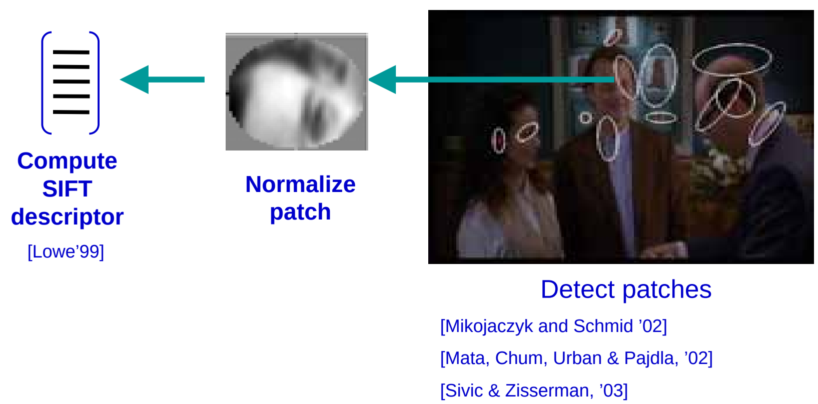
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic, et al. 2005



1.Feature detection and representation

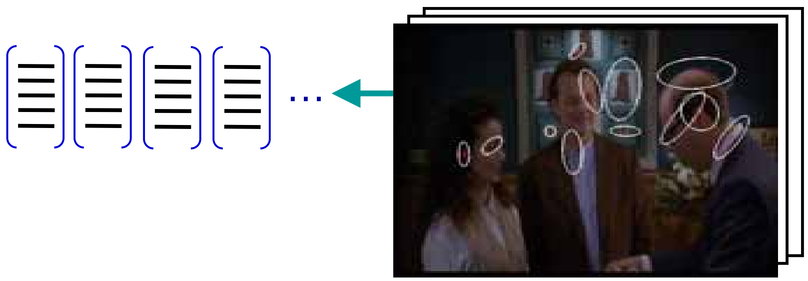
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, Bray, Dance & Fan, 2004
  - Fei-Fei & Perona, 2005
  - Sivic, Russell, Efros, Freeman & Zisserman, 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation based patches (Barnard, Duygulu, Forsyth, de Freitas, Blei, Jordan, 2003)

1.Feature detection and representation

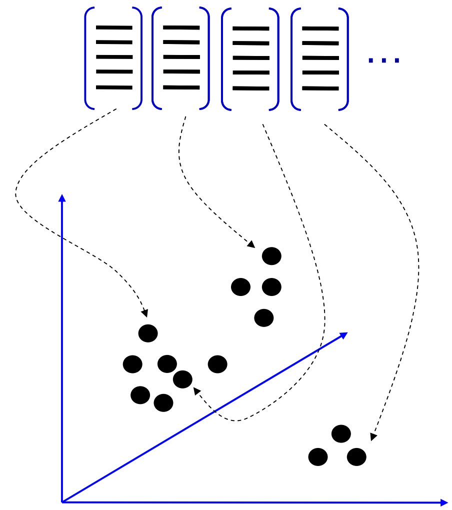


Slide credit: Josef Sivic

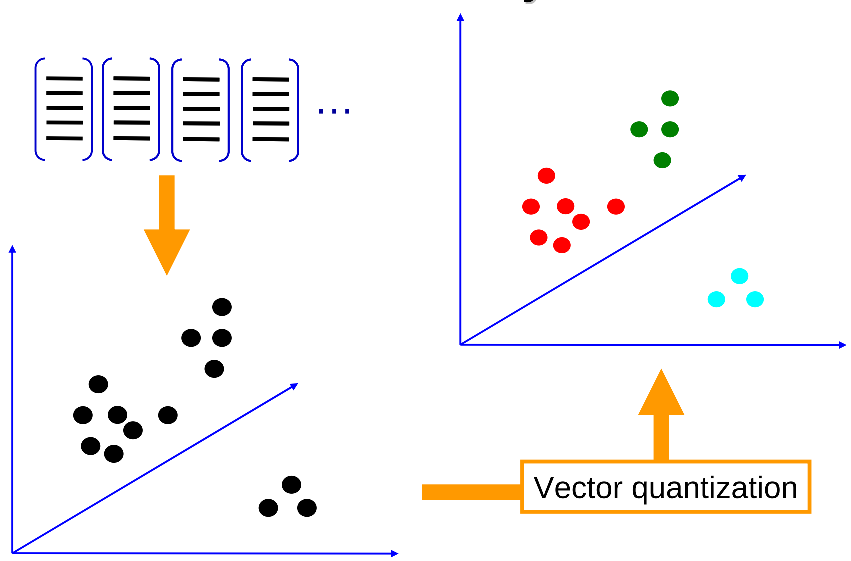
1.Feature detection and representation



2. Codewords dictionary formation

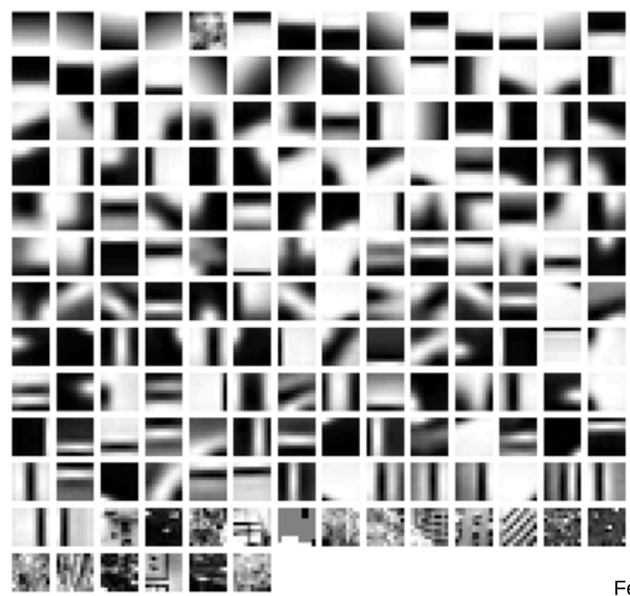


2. Codewords dictionary formation



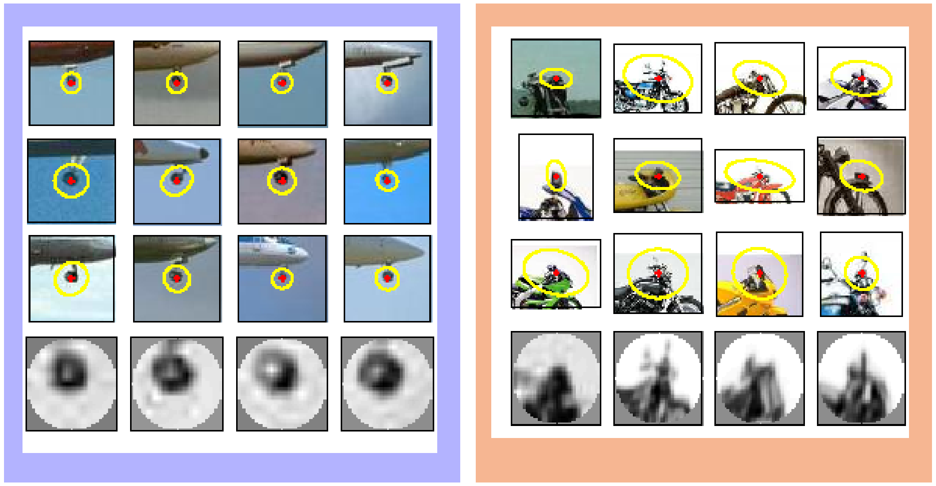
Slide credit: Josef Sivic

2. Codewords dictionary formation



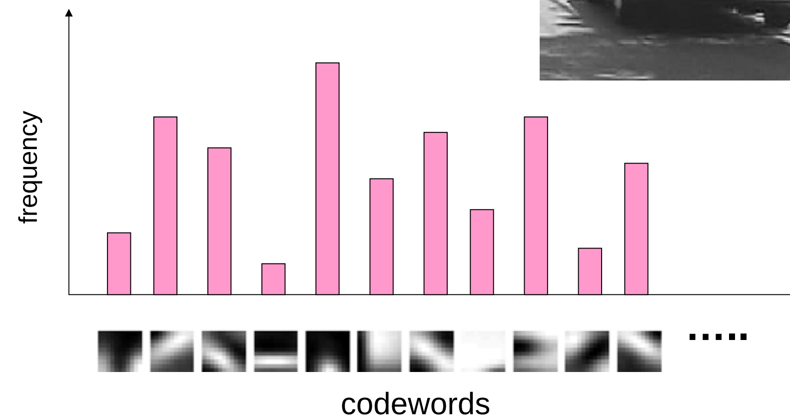
Fei-Fei et al. 2005

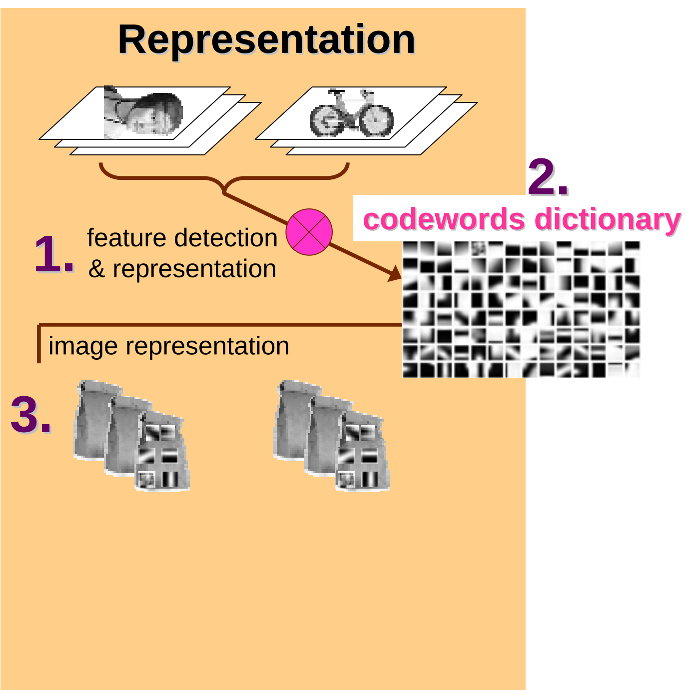
Image patch examples of codewords

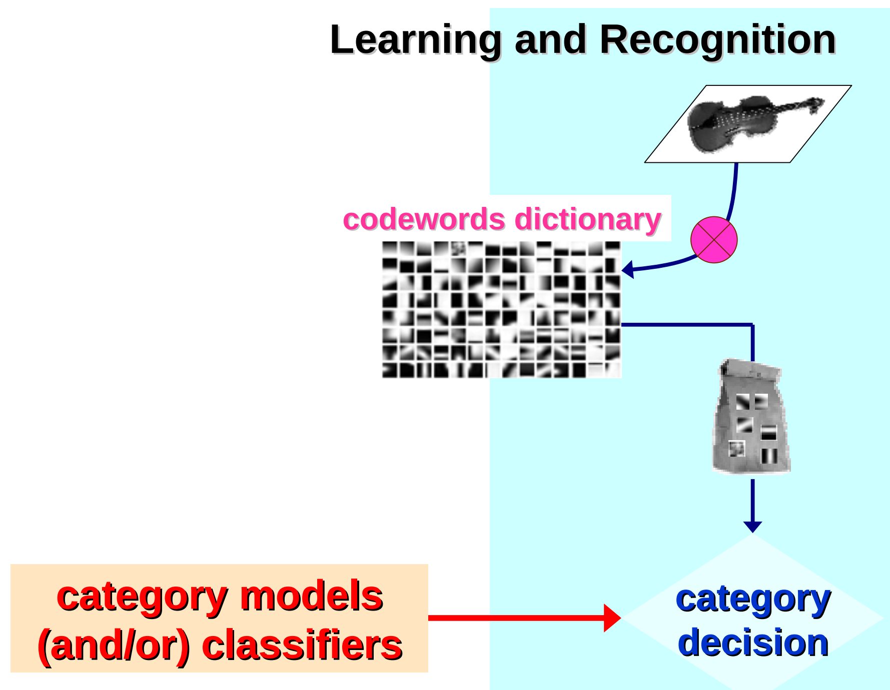


Sivic et al. 2005

3. Image representation



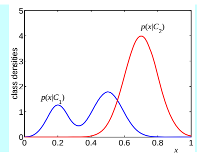




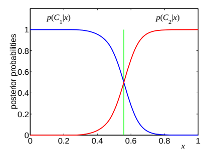
Learning and Recognition

1. Generative method:
   - graphical models

skip – see tutorial Web site



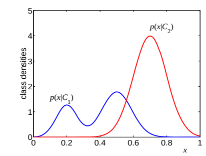
2. Discriminative method:
   - SVM



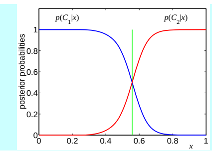
category models (and/or) classifiers

Learning and Recognition

1. Generative method:
   - graphical models

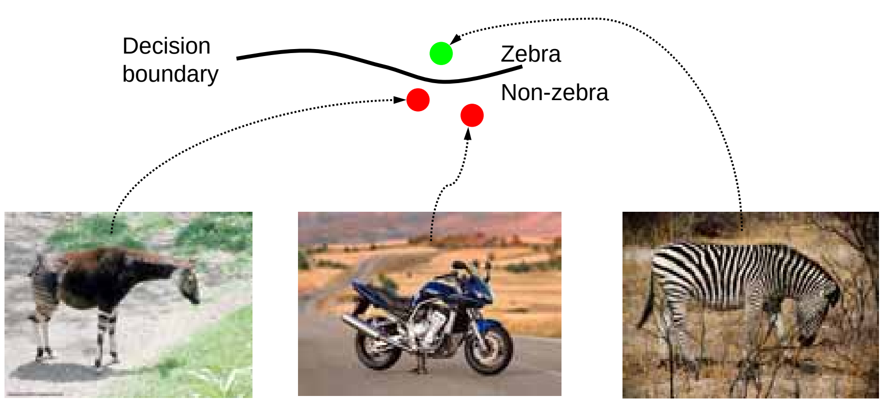


2. Discriminative method:
   - SVM



category models (and/or) classifiers

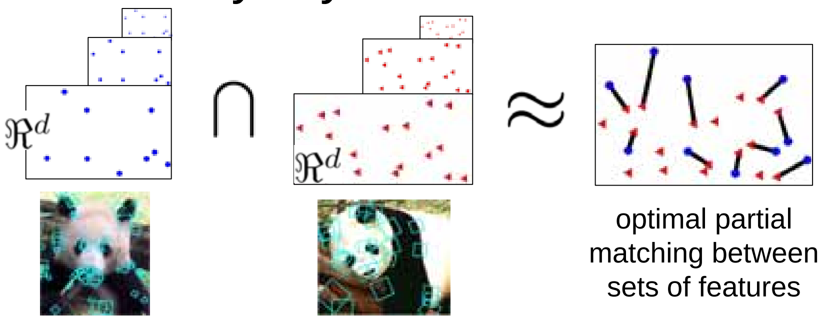
Discriminative methods based on ‘bag of words’ representation



Discriminative methods based on ‘bag of words’ representation

- Grauman & Darrell, 2005, 2006:
  - SVM w/ Pyramid Match kernels
- Others
  - Csurka, Bray, Dance & Fan, 2004
  - Serre & Poggio, 2005

Summary: Pyramid match kernel

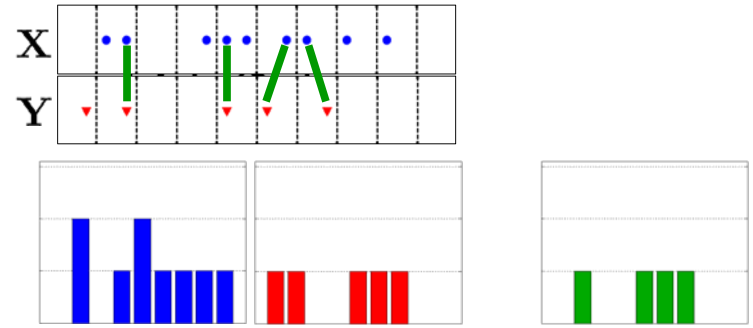


$$K_{\Delta}\left(\Psi(\mathbf{X}), \Psi(\mathbf{Y})\right)$$

Grauman & Darrell, 2005, Slide credit: Kristen Grauman

Pyramid Match (Grauman & Darrell 2005)

Histogram intersection $\mathcal{I}\left(H(\mathbf{X}), H(\mathbf{Y})\right) = \sum_{j=1}^{r} \min\left(H(\mathbf{X})_j, H(\mathbf{Y})_j\right)$



$H(\mathbf{X})$ $\quad$ $H(\mathbf{Y})$ $\quad$ $\mathcal{I}\left(H(\mathbf{X}), H(\mathbf{Y})\right) = 4$

Slide credit: Kristen Grauman

Pyramid Match (Grauman & Darrell 2005)

Histogram intersection $\mathcal{I}\left(H(\mathbf{X}), H(\mathbf{Y})\right) = \sum_{j=1}^{r} \min\left(H(\mathbf{X})_j, H(\mathbf{Y})_j\right)$

$$N_i = \overbrace{\mathcal{I}\left(H_i(\mathbf{X}), H_i(\mathbf{Y})\right)}^{\text{matches at this level}} - \overbrace{\mathcal{I}\left(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})\right)}^{\text{matches at previous level}}$$

Difference in histogram intersections across levels counts *number of new pairs* matched

Slide credit: Kristen Grauman

Pyramid match kernel

$$K_{\Delta}\left(\overbrace{\Psi(\mathbf{X}), \Psi(\mathbf{Y})}^{\text{histogram pyramids}}\right) =$$

$$\sum_{i=0}^{L} \frac{1}{2^i} \left(\underbrace{\mathcal{I}\left(H_i(\mathbf{X}), H_i(\mathbf{Y})\right) - \mathcal{I}\left(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})\right)}_{\text{number of newly matched pairs at level } i} \right)$$

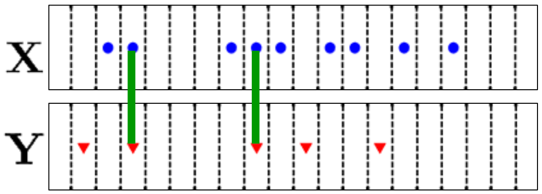
measure of difficulty of a match at level *i*

- Weights inversely proportional to bin size
- Normalize kernel values to avoid favoring large sets

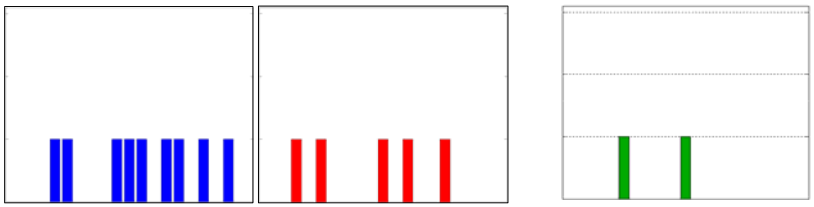
Slide credit: Kristen Grauman

Example pyramid match

Level 0



$$\longrightarrow \begin{aligned} N_0 &= 2 \\ w_0 &= 1 \end{aligned}$$

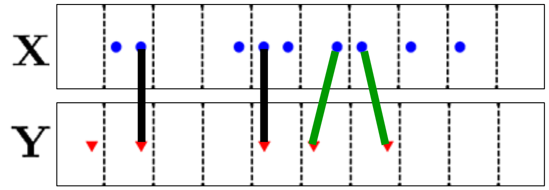


$H_0(\mathbf{X})$ $H_0(\mathbf{Y})$ $\mathcal{I}_0 = 2$

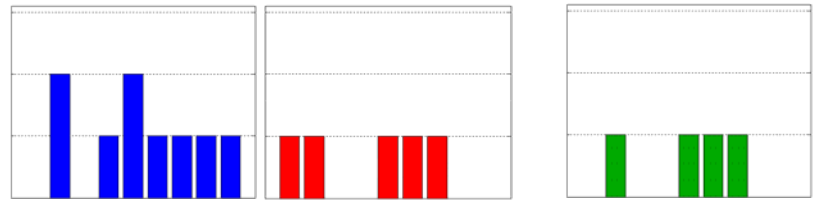
Slide credit: Kristen Grauman

Example pyramid match

Level 1



$$\longrightarrow \begin{aligned} N_1 &= 4 - 2 = 2 \\ w_1 &= \tfrac{1}{2} \end{aligned}$$

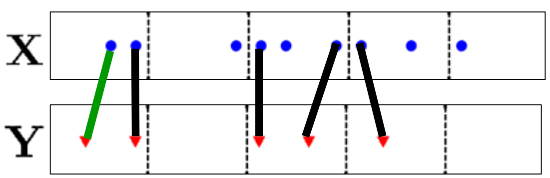


$H_1(\mathbf{X})$ $H_1(\mathbf{Y})$ $\mathcal{I}_1 = 4$

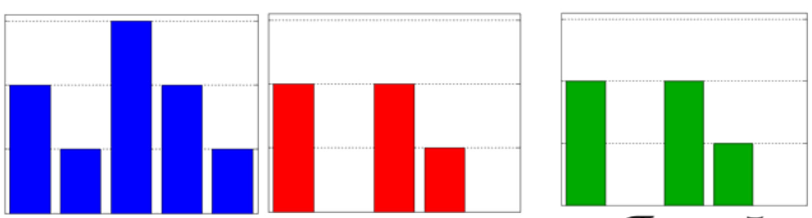
Slide credit: Kristen Grauman

Example pyramid match

Level 2



$$\longrightarrow \begin{aligned} N_2 &= 5 - 4 = 1 \\ w_2 &= \tfrac{1}{4} \end{aligned}$$

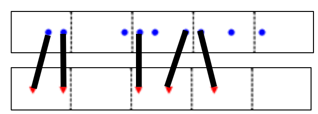


$H_2(\mathbf{X})$ $H_2(\mathbf{Y})$ $\mathcal{I}_2 = 5$

Slide credit: Kristen Grauman

Example pyramid match

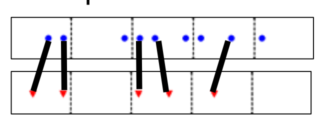
pyramid match



$$K_\Delta = \sum_{i=0}^{L} w_i N_i$$

$$= 1(2) + \tfrac{1}{2}(2) + \tfrac{1}{4}(1) = 3.25$$

optimal match

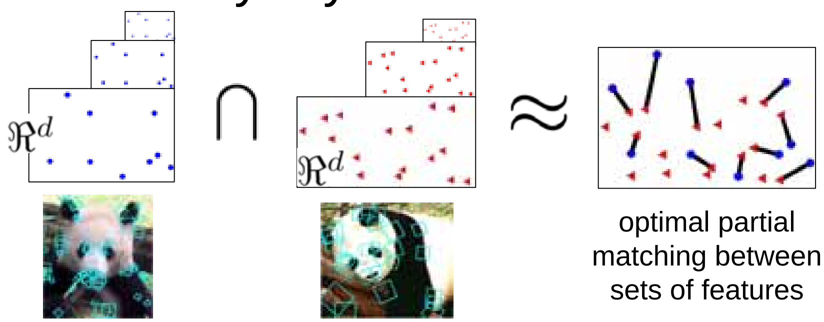


$$K = \max_{\pi:\mathbf{X}\to\mathbf{Y}} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathcal{S}(\mathbf{x}_i, \pi(\mathbf{x}_i))$$

$$= 1(2) + \tfrac{1}{2}(3) = 3.5$$

Slide credit: Kristen Grauman

Summary: Pyramid match kernel

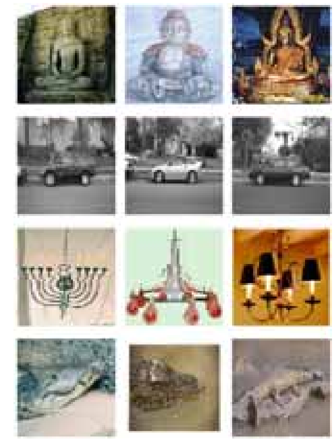


$$K_{\Delta}\left(\Psi(\mathbf{X}), \Psi(\mathbf{Y})\right) = \sum_{i=0}^{L} \underbrace{\frac{1}{2^i}}_{\text{difficulty of a match at level i}} \underbrace{\left(\mathcal{I}\left(H_i(\mathbf{X}), H_i(\mathbf{Y})\right) - \mathcal{I}\left(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})\right)\right)}_{\text{number of new matches at level i}}$$

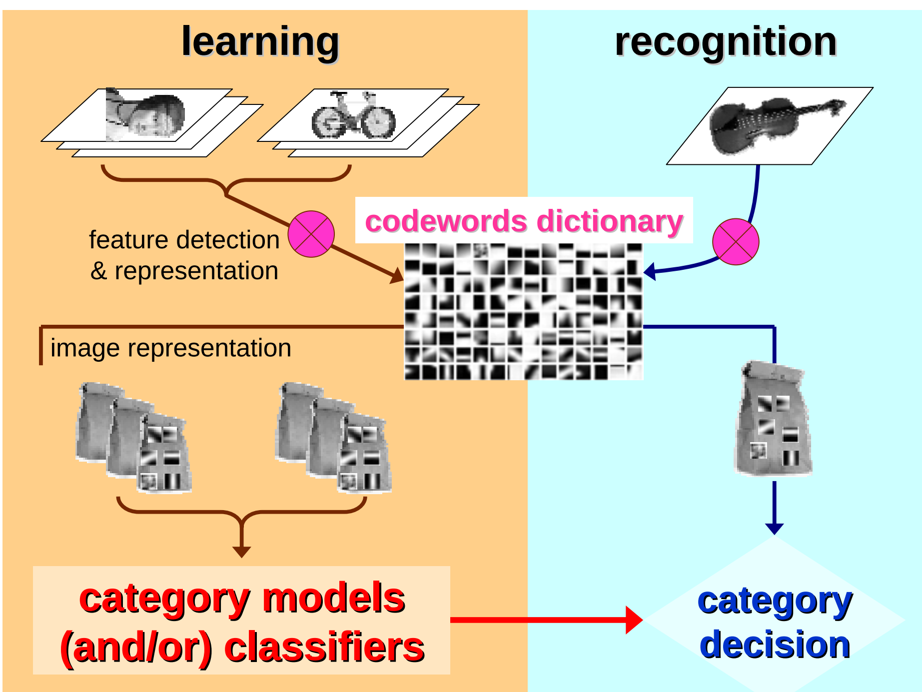
Slide credit: Kristen Grauman

Object recognition results

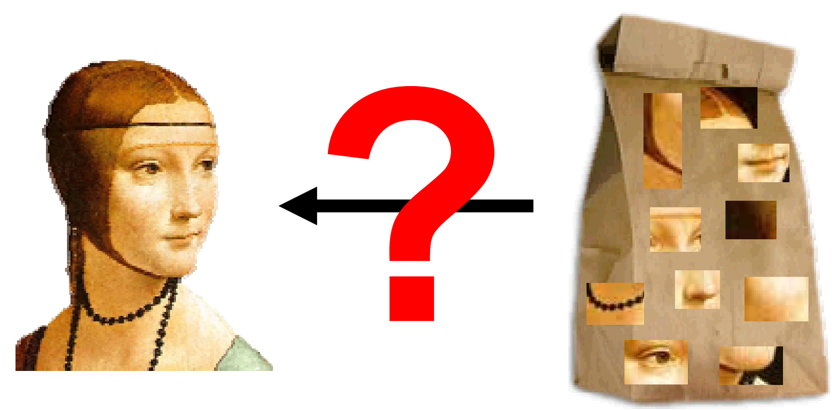
- Caltech objects database 101 object classes
- Features:
  - SIFT detector
  - PCA-SIFT descriptor, *d*=10
- 30 training images / class
- 43% recognition rate (1% chance performance)
- 0.002 seconds per match



Slide credit: Kristen Grauman

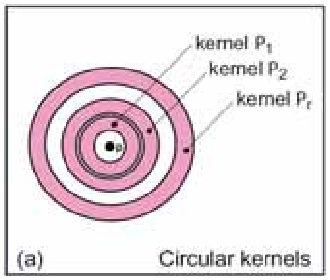


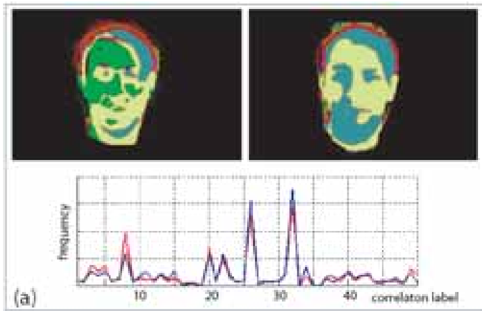
What about spatial info?



What about spatial info?

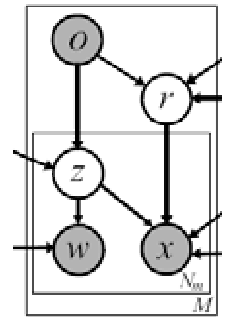
- Feature level
  - Spatial influence through correlogram features: Savarese, Winn and Criminisi, CVPR 2006

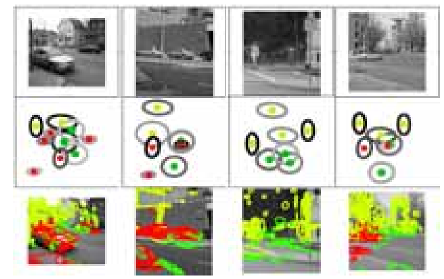




What about spatial info?

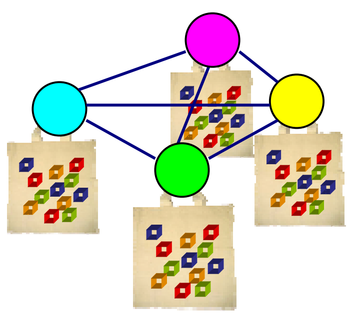
- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007

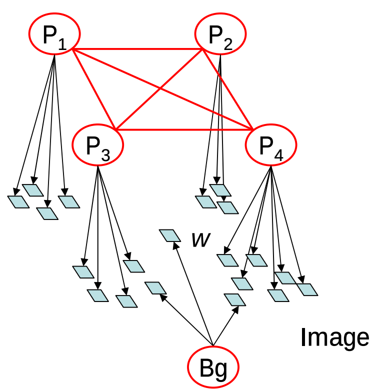




What about spatial info?

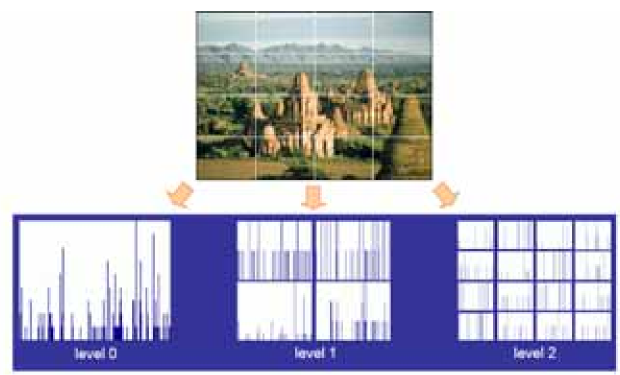
- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007





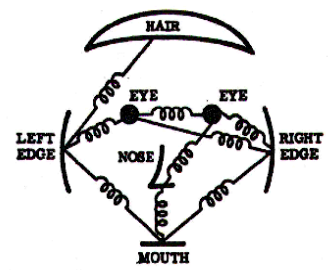
What about spatial info?

- Feature level
- Generative models
- Discriminative methods
  - Lazebnik, Schmid & Ponce, 2006



Weakness of the model

- No rigorous geometric information of the object components
- It's intuitive to most of us that objects are made of parts – no such information
- Not extensively tested yet for
  - View point invariance
  - Scale invariance
- Segmentation and localization unclear



Part 2: part-based models

by Rob Fergus (MIT)

Problem with bag-of-words

- All have equal probability for bag-of-words methods
- Location information is important

Overview of section

- Representation
  - Computational complexity
  - Location
  - Appearance
  - Occlusion, Background clutter
- Recognition

Model: Parts and Structure

Representation

- Object as set of parts
  - Generative representation
- Model:
  - Relative locations between parts
  - Appearance of part
- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

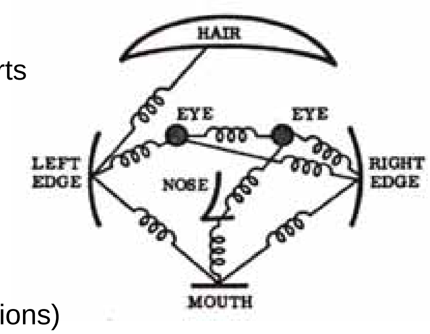
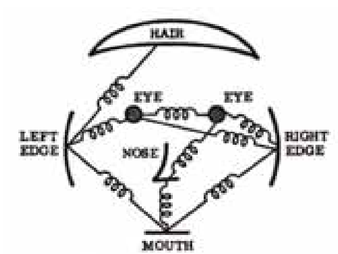


Figure from [Fischler & Elschlager 73]

History of Parts and Structure approaches

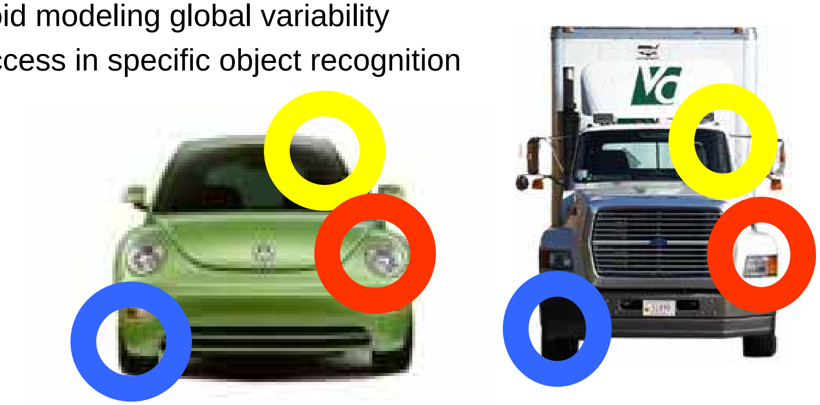
- Fischler & Elschlager 1973



- 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, '03, '04, '05
- Felzenszwalb & Huttenlocher '00, '04
- Crandall & Huttenlocher '05, '06
- Leibe & Schiele '03, '04
- Many papers since 2000

Sparse representation

+ Computationally tractable (10^5 pixels → 10^1 -- 10^2 parts)
+ Generative representation of class
+ Avoid modeling global variability
+ Success in specific object recognition

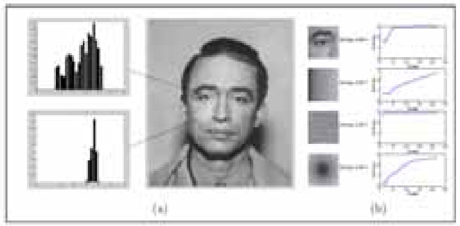


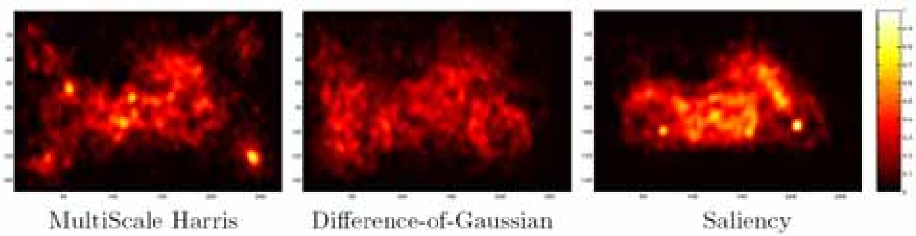
- Throw away most image information
- Parts need to be distinctive to separate from other classes

Region operators

– Local maxima of interest operator function

– Can give scale/orientation invariance

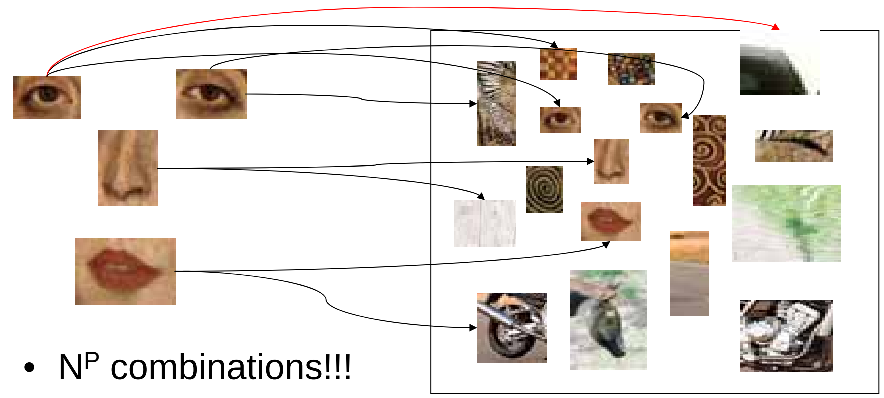




Figures from [Kadir, Zisserman and Brady 04]

The correspondence problem

- Model with P parts
- Image with N possible assignments for each part
- Consider mapping to be 1-1



- N^P combinations!!!

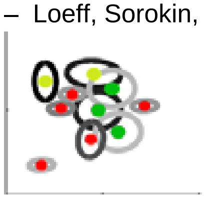
The correspondence problem

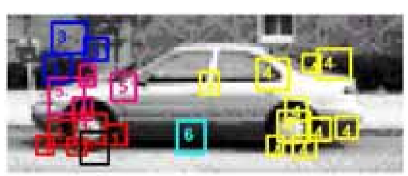
- 1 – 1 mapping
  – Each part assigned to unique feature

As opposed to:

- 1 – Many
  – Bag of words approaches
  – Sudderth, Torralba, Freeman '05
  – Loeff, Sorokin, Arora and Forsyth '05

- Many – 1
  - Quattoni, Collins and Darrell, 04

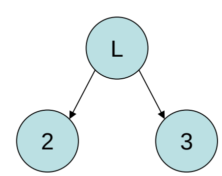




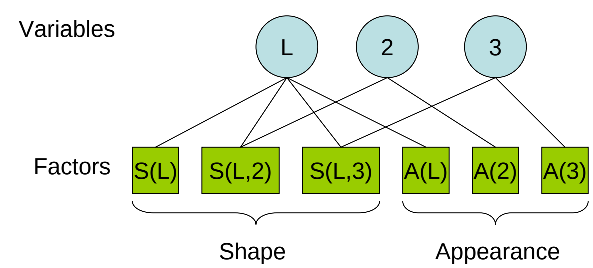
Connectivity of parts

- Complexity is given by size of maximal clique in graph
- Consider a 3 part model
  – Each part has set of N possible locations in image
  – Location of parts 2 & 3 is independent, given location of L
  – Each part has an appearance term, independent between parts.

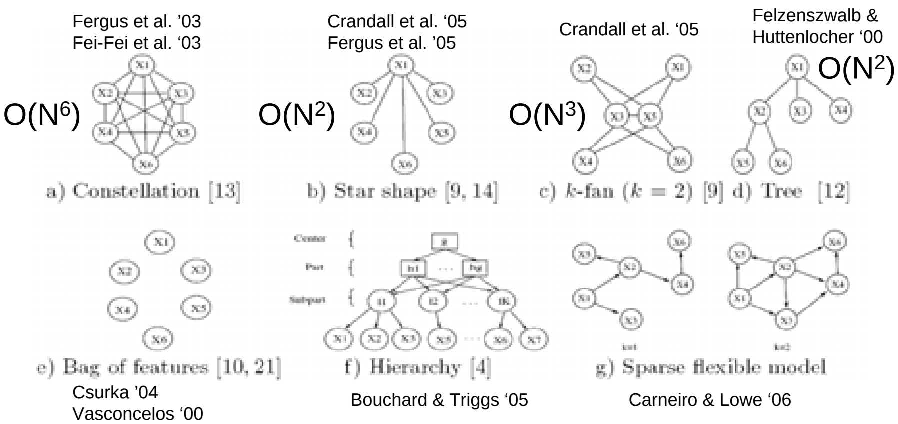
Shape Model



Factor graph



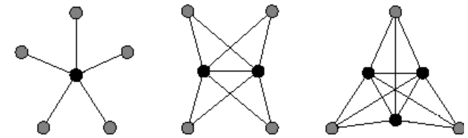
Different connectivity structures

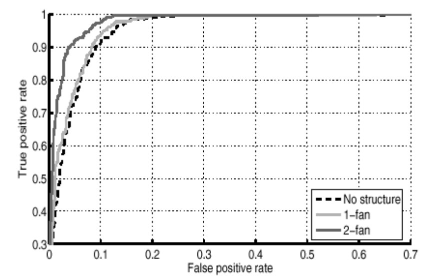


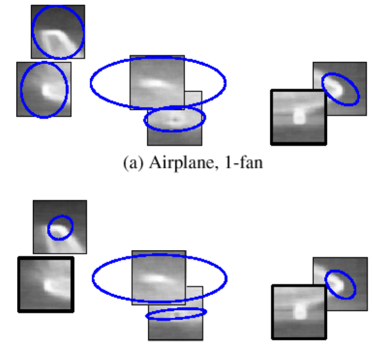
from Sparse Flexible Models of Local Features
Gustavo Carneiro and David Lowe, ECCV 2006

How much does shape help?

- Crandall, Felzenszwalb, Huttenlocher CVPR'05
- Shape variance increases with increasing model complexity
- Do get some benefit from shape





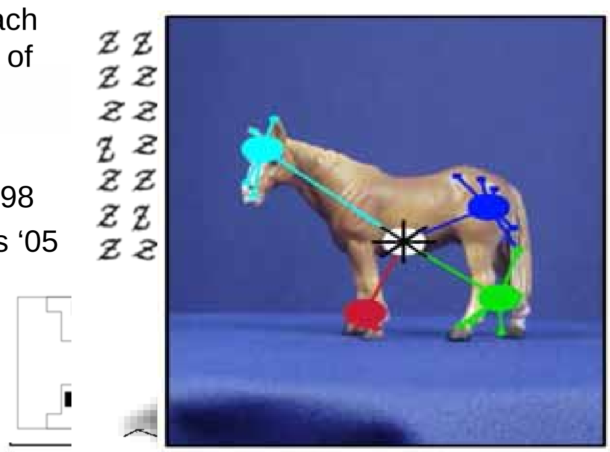


(a) Airplane, 1-fan

(b) Airplane, 2-fan

Hierarchical representations

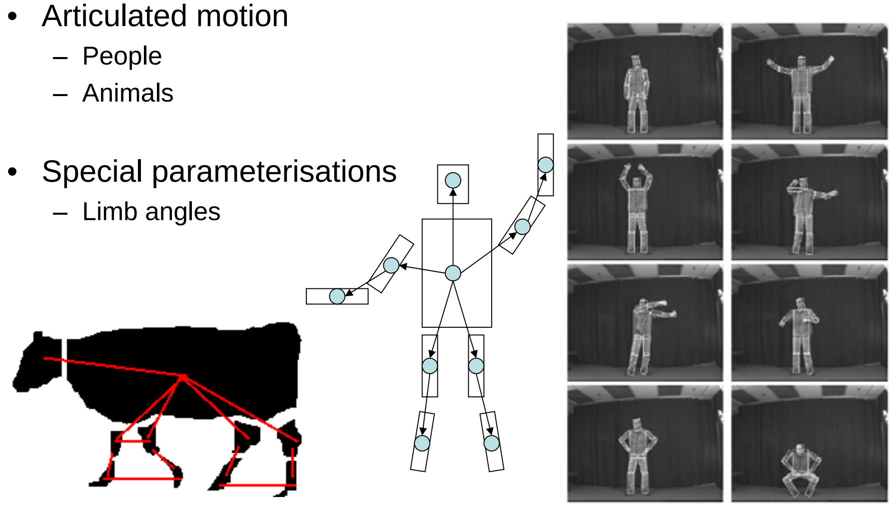
- Pixels → Pixel groupings → Parts → Object
- Multi-scale approach increases number of low-level features
- Amit and Geman '98
- Bouchard & Triggs '05



Images from [Amit98,Bouchard05]

Some class-specific graphs

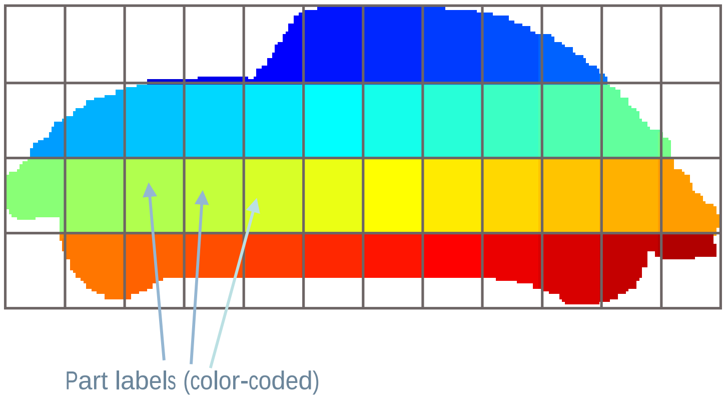
- Articulated motion
  - People
  - Animals
- Special parameterisations
  - Limb angles



Images from [Kumar, Torr and Zisserman 05, Felzenszwalb & Huttenlocher 05]

Dense layout of parts

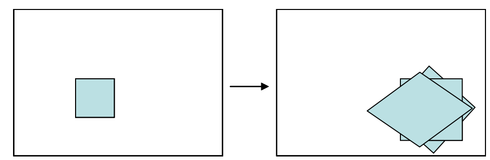
Layout CRF: Winn & Shotton, CVPR ‘06



How to model location?

- Explicit: Probability density functions
- Implicit: Voting scheme

- Invariance
  - Translation
  - Scaling
  - Similarity/affine
  - Viewpoint

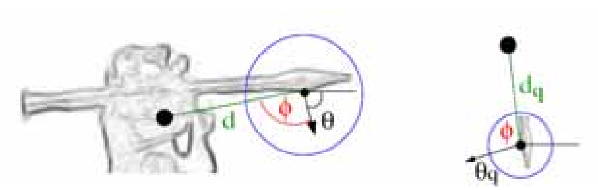


Explicit shape model

- Cartesian
  - E.g. Gaussian distribution
  - Parameters of model, μ and Σ
  - Independence corresponds to zeros in Σ
  - Burl et al. ’96, Weber et al. ‘00, Fergus et al. ’03

$$\mu = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix} \quad \Sigma = \begin{pmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1y_1 & x_1y_2 & x_1y_3 \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2y_1 & x_2y_2 & x_2y_3 \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3y_1 & x_3y_2 & x_3y_3 \\ y_1x_1 & y_1x_2 & y_1x_3 & y_1y_1 & y_1y_2 & y_1y_3 \\ y_2x_1 & y_2x_2 & y_2x_3 & y_2y_1 & y_2y_2 & y_2y_3 \\ y_3x_1 & y_3x_2 & y_3x_3 & y_3y_1 & y_3y_2 & y_3y_3 \end{pmatrix}$$

- Polar
  - Convenient for invariance to rotation



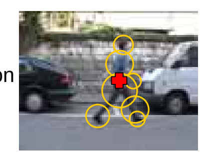
Mikolajczyk et al., CVPR ‘06

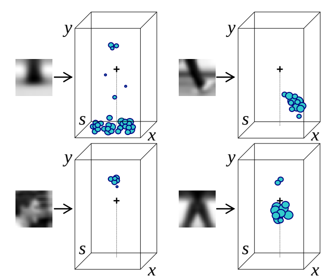
Implicit shape model

- Use Hough space voting to find object
- Leibe and Schiele ’03,’05

Learning

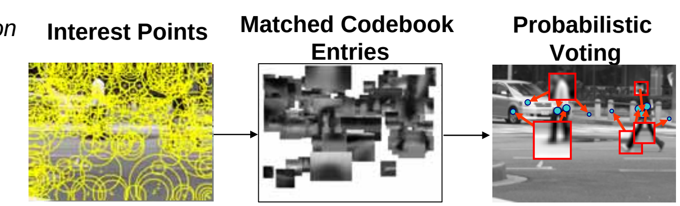
- Learn appearance codebook
  - Cluster over interest points on training images
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
  - Centroid is given



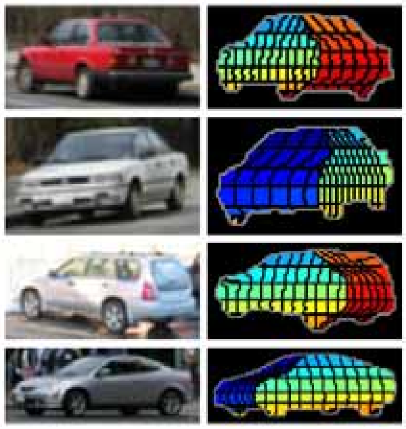


Spatial occurrence distributions

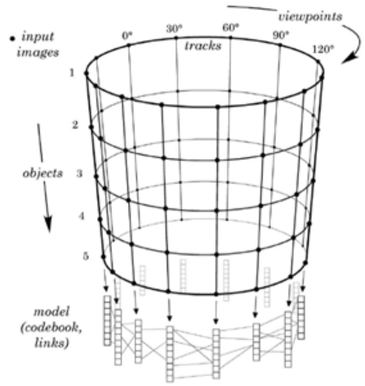
Recognition



Multiple view points



Hoiem, Rother, Winn, 3D LayoutCRF for Multi-View Object Class Recognition and Segmentation, CVPR '07



Thomas, Ferrari, Leibe, Tuytelaars, Schiele, and L. Van Gool. Towards Multi-View Object Class Detection, CVPR 06

Representation of appearance

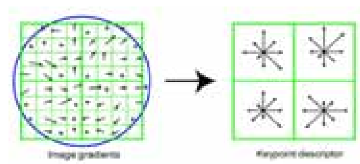
- Needs to handle intra-class variation
  - Task is no longer matching of descriptors
  - Implicit variation (VQ to get discrete appearance)
  - Explicit model of appearance (e.g. Gaussians in SIFT space)
- Dependency structure
  - Often assume each part's appearance is independent
  - Common to assume independence with location

Representation of appearance

- Invariance needs to match that of shape model
- Insensitive to small shifts in translation/scale
  - Compensate for jitter of features
  - e.g. SIFT
- Illumination invariance
  - Normalize out

Appearance representation

- SIFT



- PCA

- Decision trees

[Lepetit and Fua CVPR 2005]

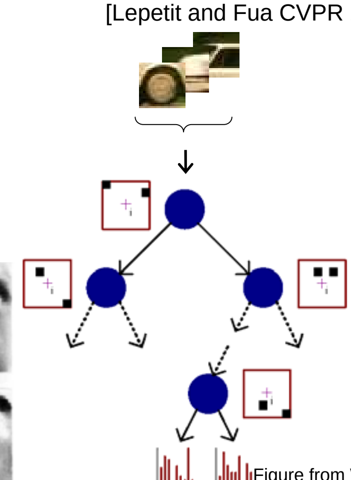
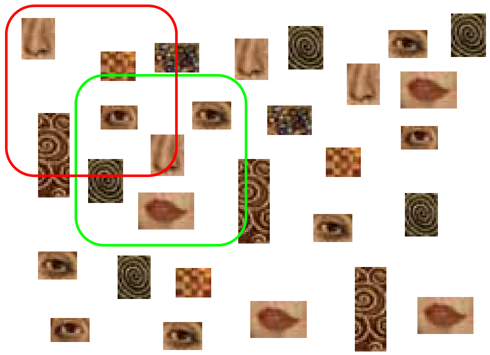


Figure from Winn & Shotton, CVPR '06

Background clutter

- Explicit model
  - Generative model for clutter as well as foreground object
- Use a sub-window
  - At correct position, no clutter is present



What task?

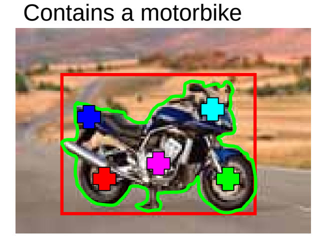
- Classification
  - Object present/absent in image
  - Background may be correlated with object
- Localization / Detection
  - Localize object within the frame
  - Bounding box or pixel-level segmentation

Demo Web Page

Learning situations

- Varying levels of supervision
  - Unsupervised
  - Image labels
  - Object centroid/bounding box
  - Segmented object
  - Manual correspondence (typically sub-optimal)

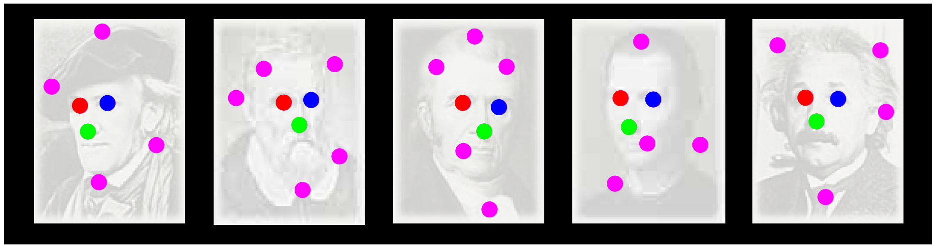
Contains a motorbike



- Generative models naturally incorporate labelling information (or lack of it)
- Discriminative schemes require labels for all data points

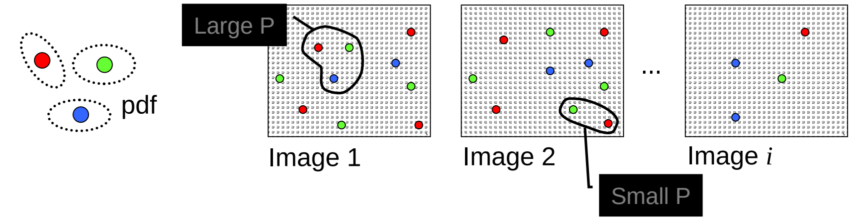
Learning using EM

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

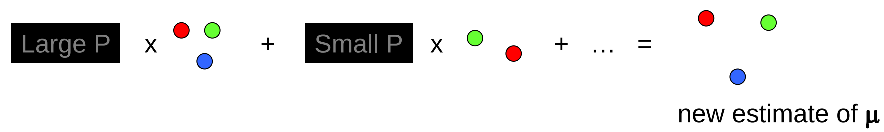


Example scheme, using EM for maximum likelihood learning

1. Current estimate of θ 2. Assign probabilities to constellations

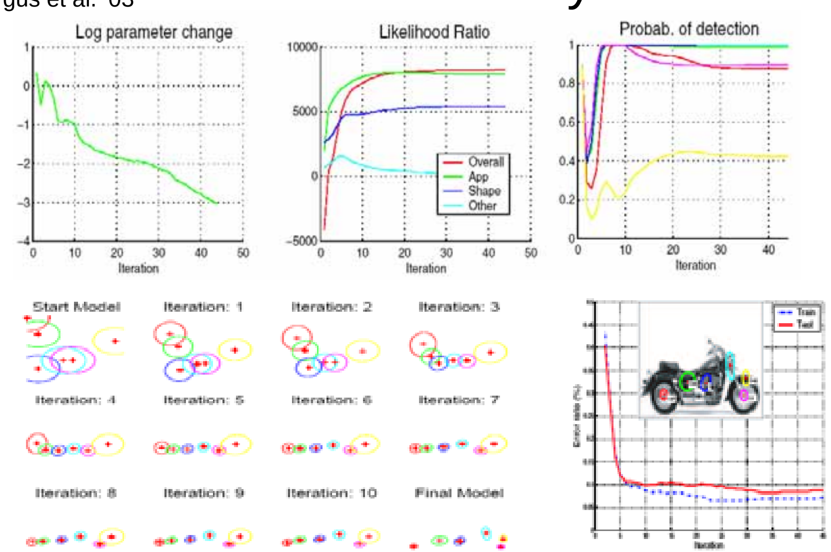


3. Use probabilities as weights to re-estimate parameters. Example: $\boldsymbol{\mu}$



Learning Shape & Appearance simultaneously

Fergus et al. '03



Last part: datasets and object collections

Links to datasets

The next tables summarize some of the available datasets for training and testing object detection and recognition algorithms. These lists are far from exhaustive.

Databases for object localization

CMU/MIT frontal faces	vasc.ri.cmu.edu/idb/html/face/frontal_images cbcl.mit.edu/software-datasets/FaceData2.html	Patches	Frontal faces
Graz-02 Database	www.emt.tugraz.at/~pinz/data/GRAZ_02/	Segmentation masks	Bikes, cars, people
UIUC Image Database	l2r.cs.uiuc.edu/~cogcomp/Data/Car/	Bounding boxes	Cars
TU Darmstadt Database	www.vision.ethz.ch/leibe/data/	Segmentation masks	Motorbikes, cars, cows
LabelMe dataset	people.csail.mit.edu/brussell/research/LabelMe/intro.html	Polygonal boundary	>500 Categories

Databases for object recognition

Caltech 101	www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html	Segmentation masks	101 categories
COIL-100	www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html	Patches	100 instances
NORB	www.cs.nyu.edu/~ylclab/data/norb-v1.0/	Bounding box	50 toys

On-line annotation tools

ESP game	www.espgame.org	Global image descriptions	Web images
LabelMe	people.csail.mit.edu/brussell/research/LabelMe/intro.html	Polygonal boundary	High resolution images

Collections

PASCAL	http://www.pascal-network.org/challenges/VOC/	Segmentation, boxes	various

Collecting datasets (towards $10^{6\text{-}7}$ examples)

- ESP game (CMU)
  Luis Von Ahn and Laura Dabbish 2004
- LabelMe (MIT)
  Russell, Torralba, Freeman, 2005
- StreetScenes (CBCL-MIT)
  Bileschi, Poggio, 2006
- WhatWhere (Caltech)
  Perona et al, 2007
- PASCAL challenge
  2006, 2007
- Lotus Hill Institute
  Song-Chun Zhu et al 2007

Labeling with games

Figure 1. Partners agreeing on an image in the ESP Game. Neither player can see the other's guesses.

Figure 2. Peekaboom. "Peek" tries to guess the word associated with an image slowly revealed by "Boom."

L. von Ahn, L. Dabbish, 2004; L. von Ahn, R. Liu and M. Blum, 2006

Lotus Hill Research Institute image corpus

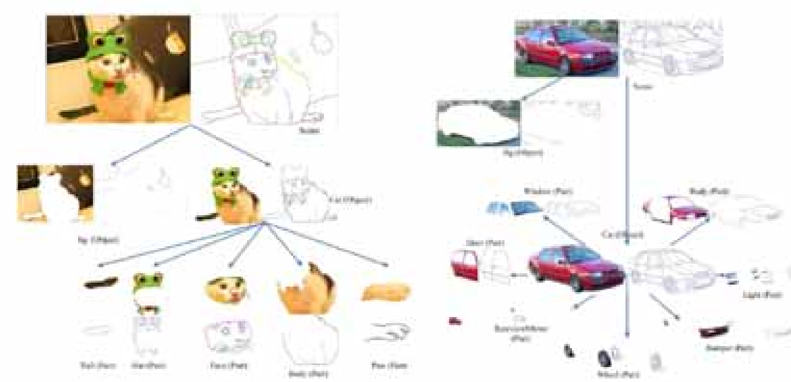


Figure 5: Two examples of the parse trees (cat and car) in the Lotus Hill Research Institute image corpus. From [87].

Z.Y. Yao, X. Yang, and S.C. Zhu, 2007

The PASCAL Visual Object Classes Challenge 2007

The twenty object classes that have been selected are:

Person: person
Animal: bird, cat, cow, dog, horse, sheep
Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

M. Everingham, Luc van Gool , C. Williams, J. Winn, A. Zisserman 2007

LabelMe

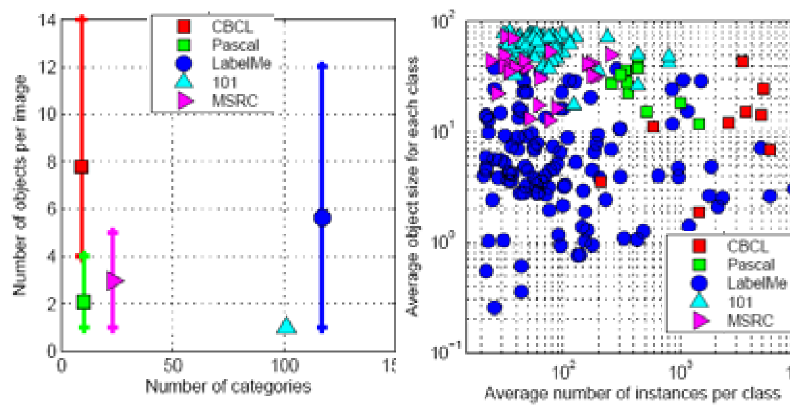
Russell, Torralba, Freman, 2005

Caltech 101 & 256

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

How to evaluate datasets?



How many labeled examples? How many classes? Segments or bounding boxes? How many instances per image? How small are the targets? Variability across instances of the same classes (viewpoint, style, illumination). How different are the images?

How representative of the visual world is? **What happens if you nail it?**

Summary

- Methods reviewed here
  - Bag of words
  - Parts and structure
  - Discriminative methods
  - Combined Segmentation and recognition

- Resources online
  - Slides
  - Code
  - Links to datasets

List properties of ideal recognition system

- Representation
  - 1000's categories,
  - Handle all invariances (occlusions, view point, …)
  - Explain as many pixels as possible (or answer as many questions as you can about the object)
  - fast, robust
- Learning
  - Handle all degrees of supervision
  - Incremental learning
  - Few training images
- …

Thank you