

Recognizing and Learning Object Categories

Based on work and slides by R. Fergus, P. Perona, A. Zisserman, A. Efros, J. Ponce, S. Lazebnik, C. Schmid, F. DiMaio, and others

Traditional Problem: Single Object Recognition



Most Objects Exhibit Considerable Intra-Class Variability



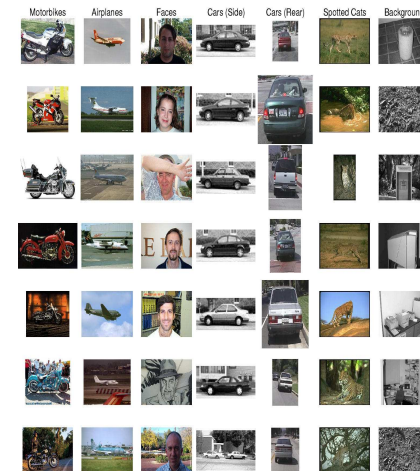
Task: Recognition of object **categories**

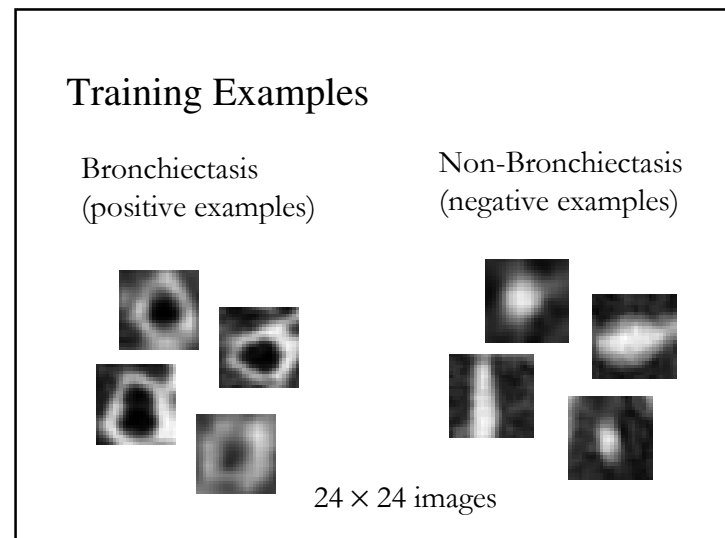
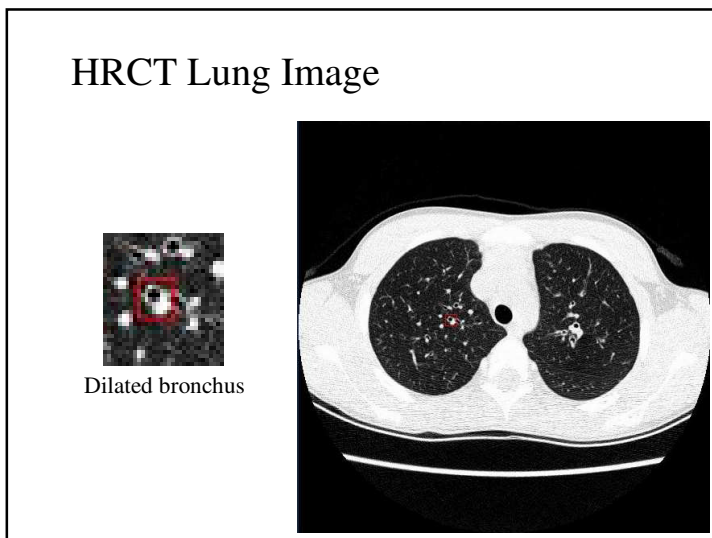
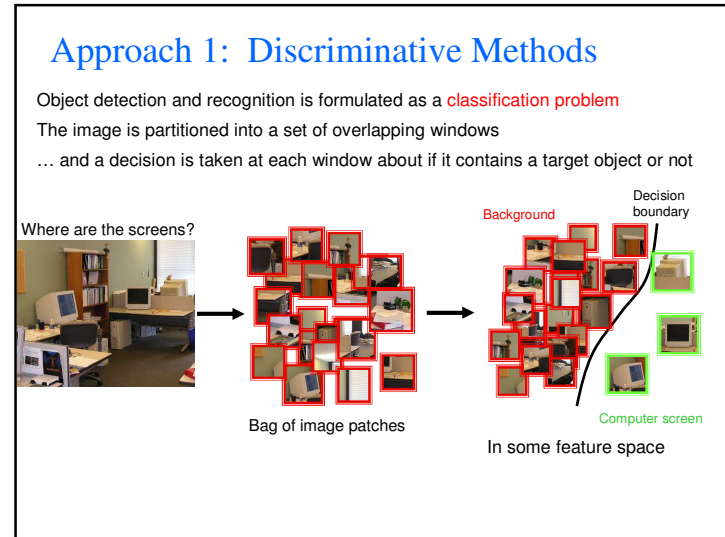
Some object categories

Learn from just examples

Difficulties:








- ⌘ Size variation
- ⌘ Background clutter
- ⌘ Occlusion
- ⌘ Intra-class variation
- ⌘ Viewpoint variation
- ⌘ Illumination variation





Formulation

§ Formulation: binary classification

				...					
Features $x =$	x_1	x_2	x_3	\cdots	x_N	x_{N+1}	x_{N+2}	\cdots	x_{N+M}
Labels $y =$	-1	+1	-1	-1		?	?		?

Training data: each image patch is labeled as containing the object or not

Test data

- Classification function

$$\hat{y} = F(x) \quad \text{Where } F(x) \text{ belongs to some family of functions}$$

- Minimize misclassification error

(Not that simple: we need some guarantees that there will be generalization)

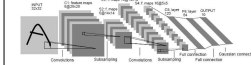
Discriminative Methods

Nearest Neighbor



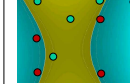
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005
...

Neural Networks



LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
...

Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001
...

Conditional Random Fields



McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003
...

Object categorization: the statistical viewpoint



$$p(\text{zebra} | \text{image})$$

vs.

$$p(\text{no zebra} | \text{image})$$

§ Bayes's rule:

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

Object categorization: the statistical viewpoint

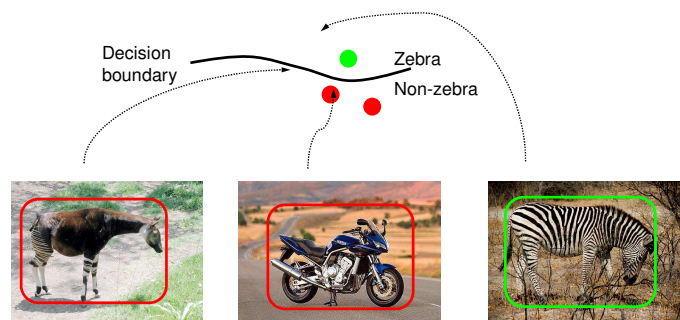
$$\underbrace{\frac{p(\text{zebra} | \text{image})}{p(\text{no zebra} | \text{image})}}_{\text{posterior ratio}} = \underbrace{\frac{p(\text{image} | \text{zebra})}{p(\text{image} | \text{no zebra})}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(\text{zebra})}{p(\text{no zebra})}}_{\text{prior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

§ Discriminative methods model the *posterior*

§ Generative methods model the *likelihood* and *prior*

Discriminative



§ Direct modeling of $\frac{p(\text{zebra} | \text{image})}{p(\text{no zebra} | \text{image})}$



Generative

§ Model $p(\text{image} | \text{zebra})$ and $p(\text{image} | \text{no zebra})$



	$p(\text{image} \text{zebra})$	$p(\text{image} \text{no zebra})$
	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

Constructing models of image content

Basic components: *local features* and *spatial relations*

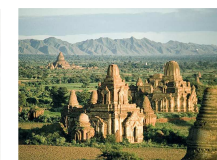
Textures



Objects



Scenes

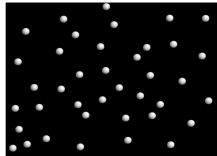


Constructing models of image content

Basic components: *local features* and *spatial relations*



Local model

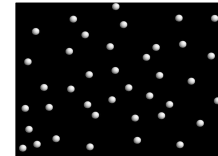


Constructing models of image content

Basic components: *local features* and *spatial relations*



Local model



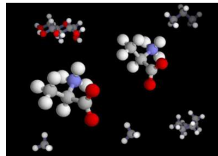
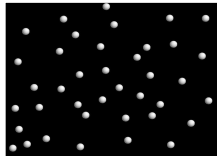
Constructing models of image content

Basic components: *local features* and *spatial relations*



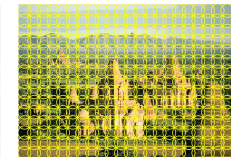
Local model

Semi-local model



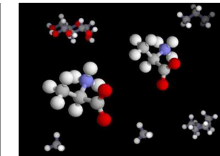
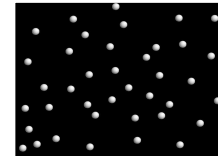
Constructing models of image content

Basic components: *local features* and *spatial relations*



Local model

Semi-local model



Constructing models of image content

Basic components: *local features* and *spatial relations*
(usually appearance)



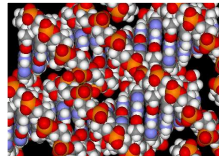
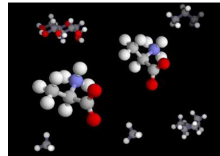
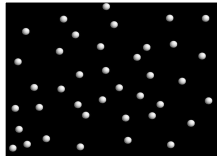
Local model



Semi-local model



Global model



Approach 2: Generative Methods using Bag of Words Models

- § An image is represented by a collection of “visual words” and their corresponding counts given a universal dictionary
- § Object categories are modeled by the distributions of these visual words
- § Although “bag of words” models can use both generative and discriminative approaches, here we will focus on generative models

Object

Bag of ‘words’



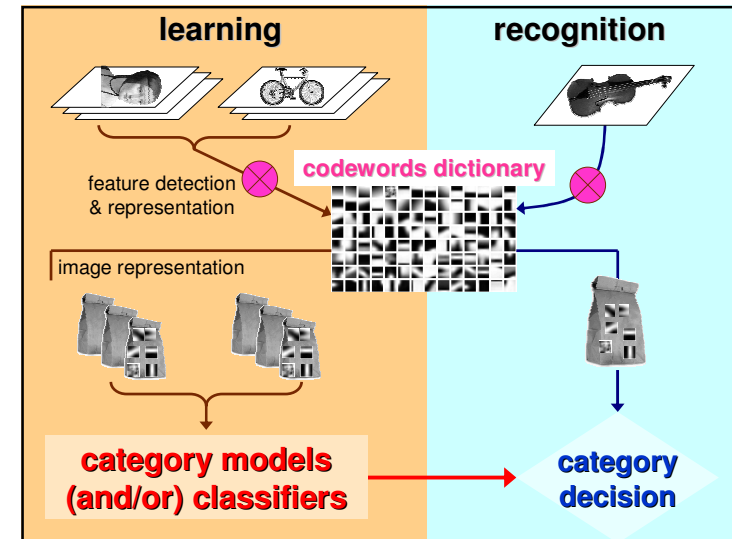
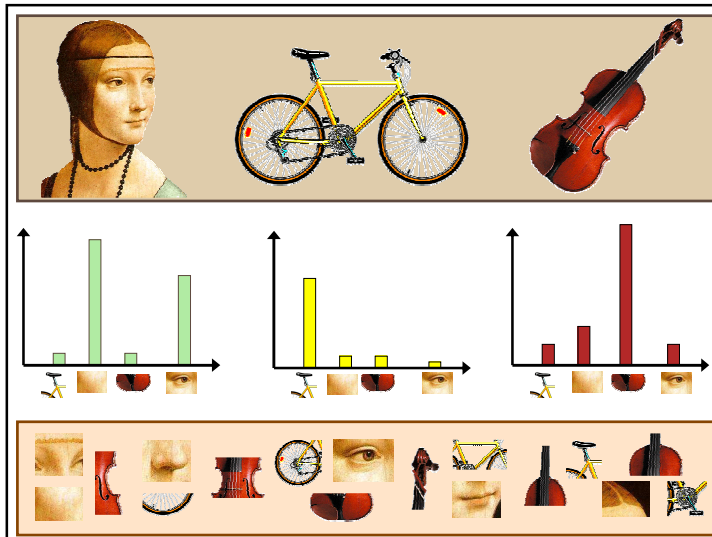
Analogy to documents

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based on the messages that the eyes send to the brain. For a long time, the visual image was considered a simple matter of the centers of the brain. However, the discovery of the visual cortex by Hubel and Wiesel has shown that the visual system is much more complicated than we thought. The visual impulses are processed in the visual cortex, which is composed of many layers of cells. Each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

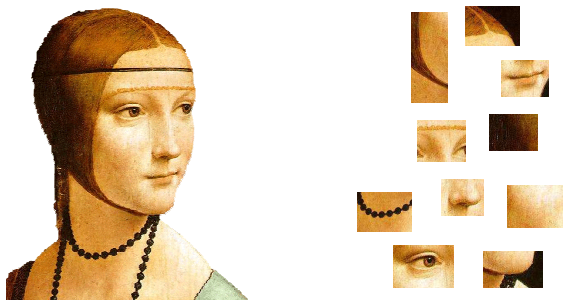
sensory, brain, visual, perception, retinal, cerebral cortex, eye, cell, optical nerve, image Hubel, Wiesel

China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would be created by a predicted 30% increase in exports to \$50bn, compared with \$38bn in 2004. The US trade deficit was \$660bn. The US trade deficit is too high a factor. Xiaochun, the Chinese vice premier, said more to boost exports. The Chinese government wants goods stay in the market. The US trade deficit increased the value of the dollar by 2.1% in July. The US trade deficit within a narrow band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear that it will take time and tread carefully before allowing the yuan to rise further in value.

China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value



1. Feature Detection and Representation



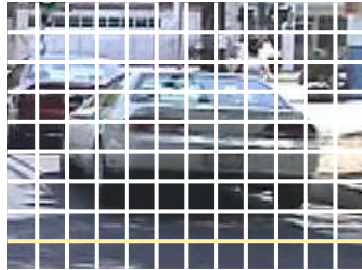
Feature Detection

- § Sliding window
- § Leung et al., 1999
- § Viola et al., 1999
- § Renninger et al. 2002



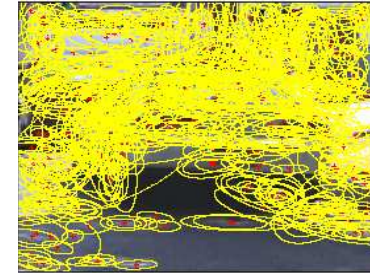
Feature Detection

- § Sliding window
 - § Leung et al., 1999
 - § Viola et al., 1999
 - § Renninger et al., 2002
- § Regular grid
 - § Vogel et al., 2003
 - § Fei-Fei et al., 2005



Feature Detection

- § Sliding window
 - § Leung et al., 1999
 - § Viola et al., 1999
 - § Renninger et al., 2002
- § Regular grid
 - § Vogel et al., 2003
 - § Fei-Fei et al., 2005
- § Interest point detector
 - § Csurka et al., 2004
 - § Fei-Fei et al., 2005
 - § Sivic et al., 2005



Feature Detection

- § Sliding window
 - § Leung et al., 1999
 - § Viola et al., 1999
 - § Renninger et al., 2002
- § Regular grid
 - § Vogel et al., 2003
 - § Fei-Fei et al., 2005
- § Interest point detector
 - § Csurka et al., 2004
 - § Fei-Fei et al., 2005
 - § Sivic et al., 2005
- § Other methods
 - § Random sampling (Ullman et al., 2002)
 - § Segmentation based patches (Barnard et al., 2003)

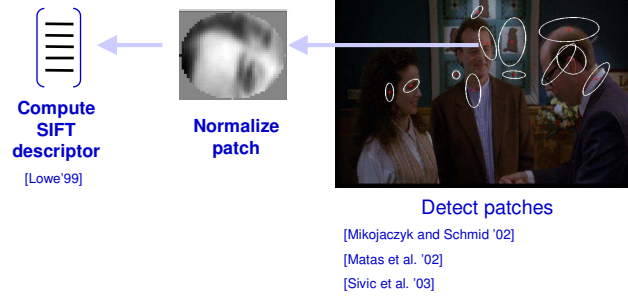
Feature Representation

Visual words, aka textons, aka keypoints:
K-means clustered pieces of the image

- § Various representations:
 - § Filter bank responses
 - § Image Patches
 - § SIFT descriptors

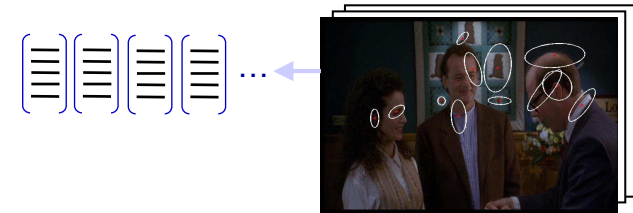
All encode more-or-less the same thing ...

Interest Point Features

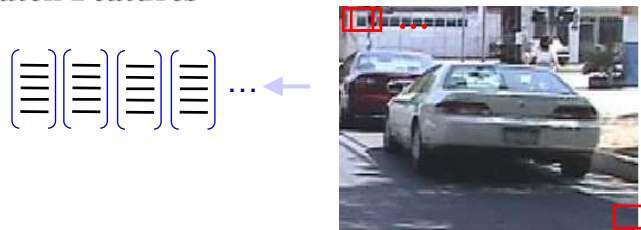


Slide credit: Josef Sivic

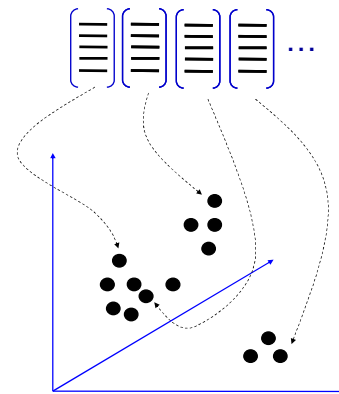
Interest Point Features

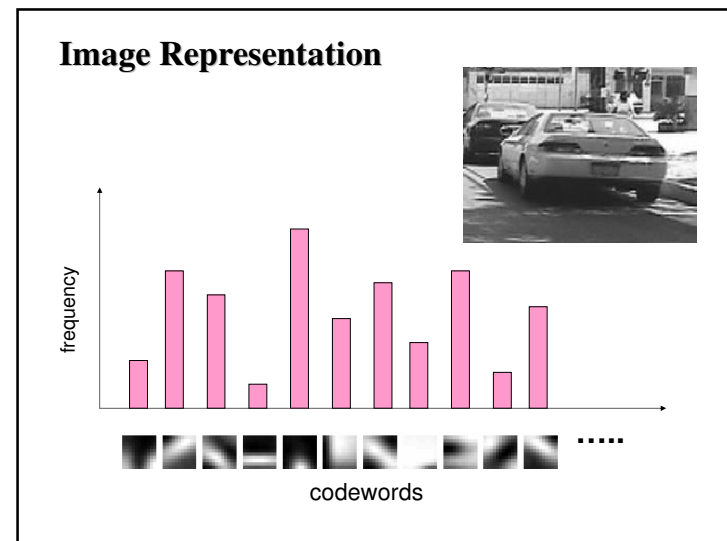
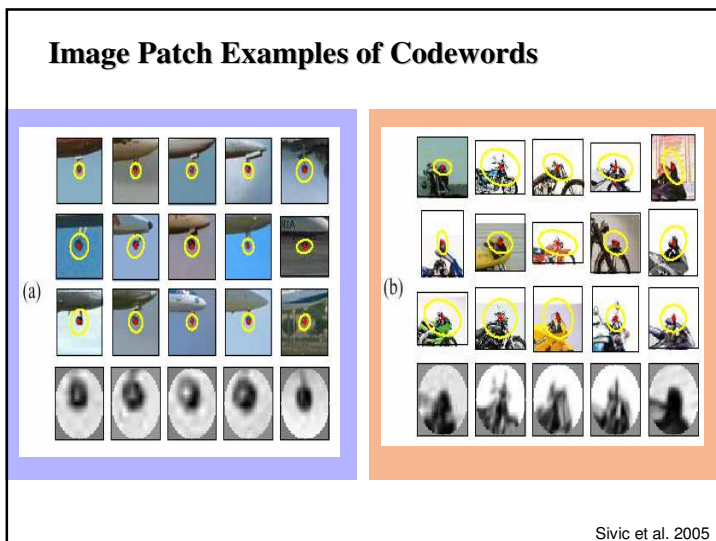
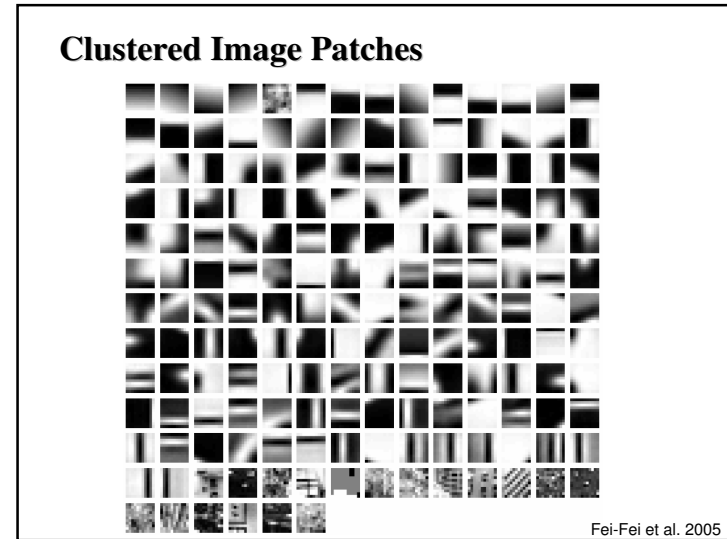
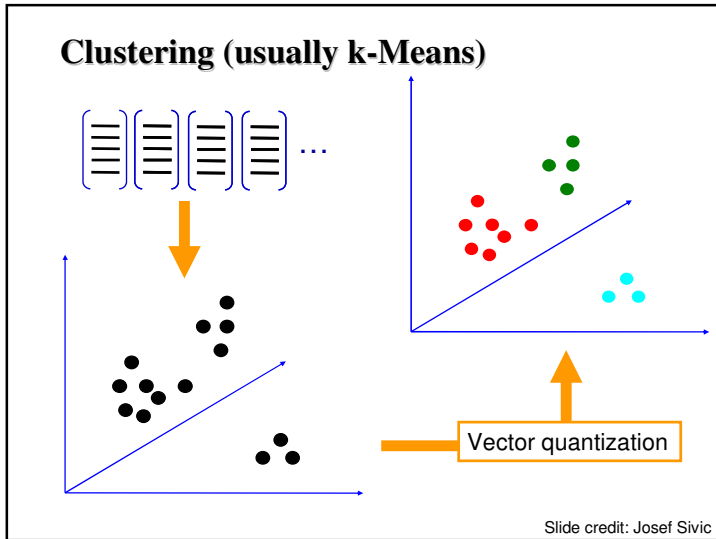


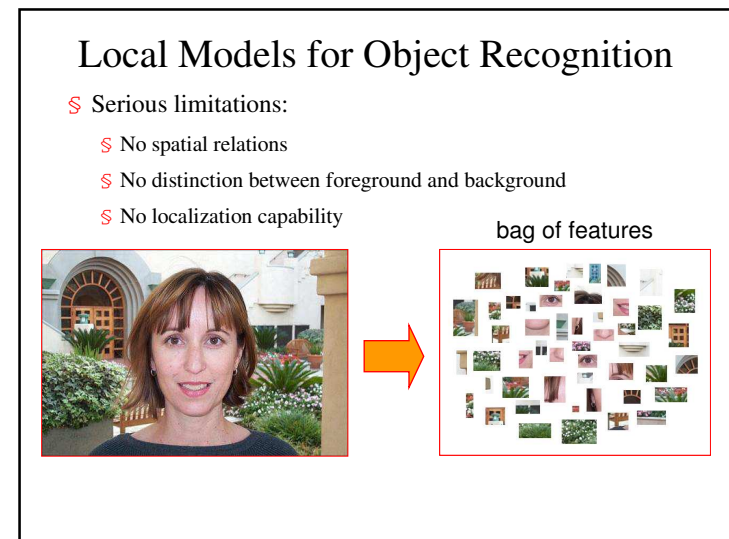
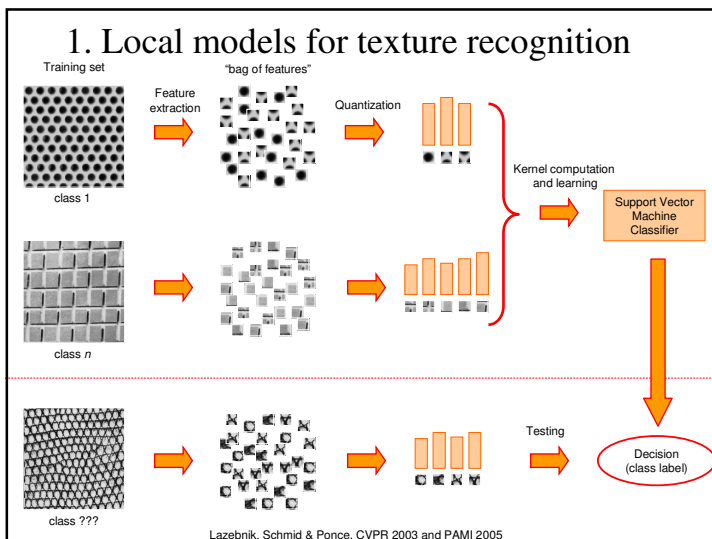
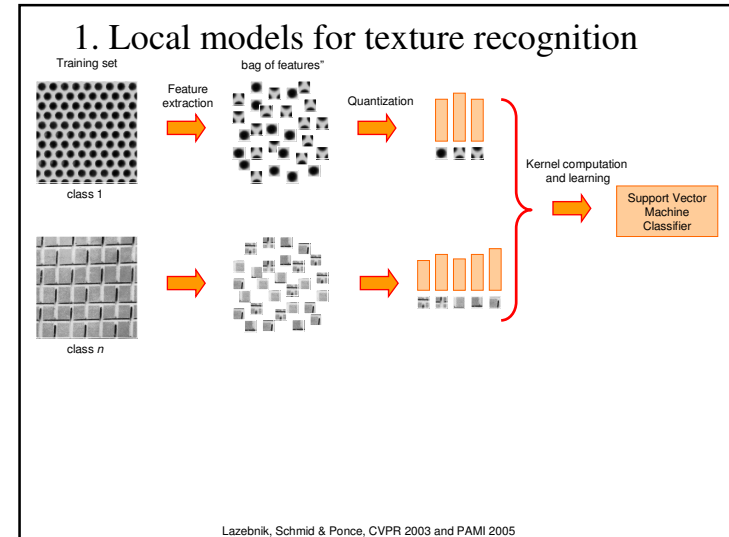
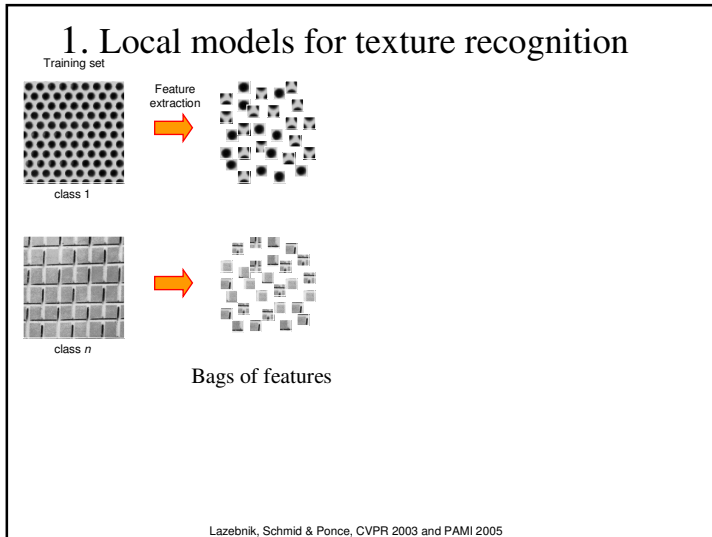
Patch Features



Dictionary Formation







Local Models for Object Recognition

§ Serious limitations:

- § No spatial relations
- § No distinction between foreground and background
- § No localization capability

§ And yet they work!

Caltech6 dataset results

Object vs. background classification, ROC equal error rate



class	ours	other results		
	Zhang et al. (2005)	Willamowski et al. (2004)	Fergus et al. (2003)	
airplanes	98.8	97.1	90.2	
cars (rear)	98.3	98.6	90.3	
cars (side)	95.0	87.3	88.5	
faces	100	99.3	96.4	
motorbikes	98.5	98.0	92.5	
spotted cats	97.0	—	90.0	

bag of features

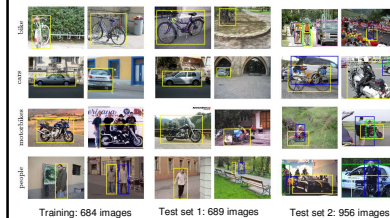
bag of features

constellation model

Local Models for Object Recognition

PASCAL 2005 challenge

<http://www.pascal-network.org/challenges/VOC>



class	test set 1	
	Zhang et al. (2005)	Larlus et al. (2006)
bikes	90.3	93.0
cars	93.0	96.1
motorbikes	96.2	97.7
people	91.6	91.7

class	test set 2	
	Zhang et al. (2005)	Deselaers et al. (2005)
bikes	68.1	66.7
cars	74.1	71.6
motorbikes	79.7	76.9
people	75.3	66.9

Object vs. background classification, ROC equal error rate

§ More comparisons: Xerox7, Graz, Caltech101, ...

- § The simplicity and effectiveness of the bag-of-features method make it a good baseline for evaluating novel approaches and datasets

Object Recognition using Texture

Object Categorization by Learned Universal Visual Dictionary

J. Winn, A. Criminisi and T. Minka

Microsoft Research, Cambridge, UK – <http://research.microsoft.com/vision/cambridge/recognition/>



Learn Texture Model

- Representation:
 - Textons (rotation-variant)
- Clustering
 - K=2000
 - Then clever merging
 - Then fitting histogram with Gaussian
- Training
 - Labeled class data



Results Movie



Simple Works Well

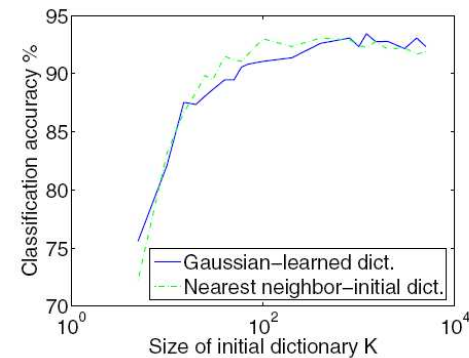


Figure 5: Comparing classification performance for Gaussian class models vs nearest neighbours classification.

Problem with Bag of Words



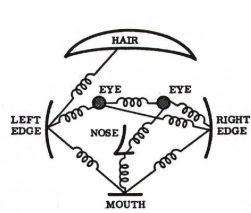
- § All have equal probability for bag-of-words methods
- § Location information is important

Approach 3: Generative Methods using Part-Based Models

- § An object in an image is represented by a collection of parts, characterized by both their visual appearances and locations
- § Object categories are modeled by the appearance and spatial distributions of these characteristic parts
- § Issues for such models include efficient methods for finding correspondences between the object and the scene

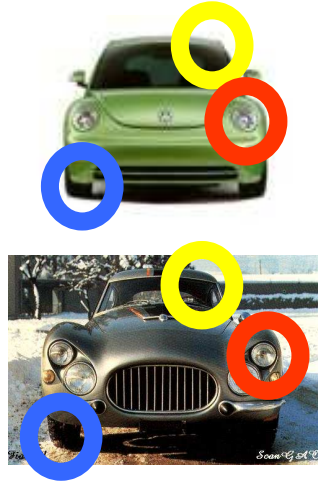


Model: Constellation of Parts



Fischler & Elschlager, 1973

Yuille, 1991
 Brunelli & Poggio, 1993
 Lades, v.d. Malsburg et al. 1993
 Coates, Lanitis, Taylor et al. 1995
 Amit & Geman, 1995, 1999
 Perona et al. 1995, 1996, 1998, 2000
 Felzenszwalb & Huttenlocher, 2000



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

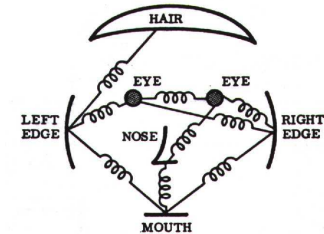
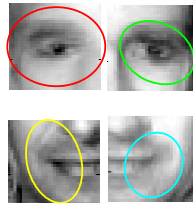


Figure from [Fischler73]

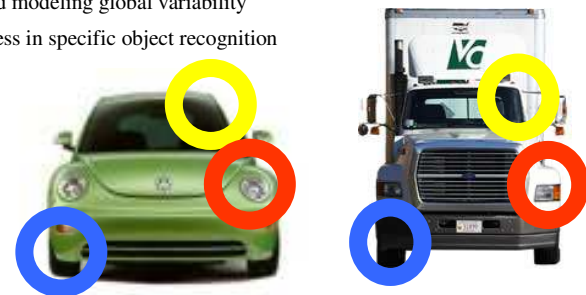
Model Structure

- § Model **shape** using Gaussian distribution on image location between parts and scale of each part
- § Model **appearance** as patches of pixel intensities
- § Represent object class as graph of P image patches with parameters θ



Sparse Representation

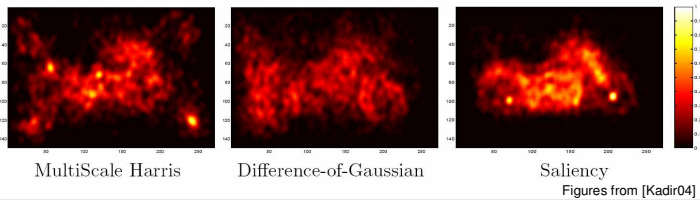
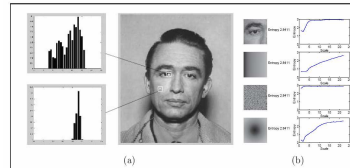
- § + Computationally tractable (10^5 pixels 10^1 -- 10^2 parts)
- § + Generative representation of class
- § + Avoid modeling global variability
- § + Success in specific object recognition



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

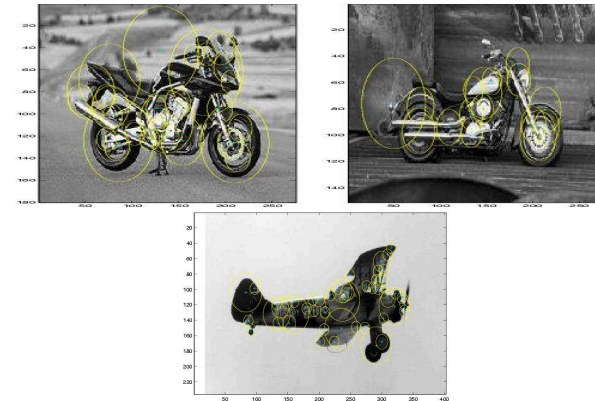
Regions or Pixels?

- § # Regions \ll # Pixels
- § Regions increase tractability but lose information
- § Generally use regions:
 - § Local maxima of interest operators
 - § Can give scale/orientation invariance

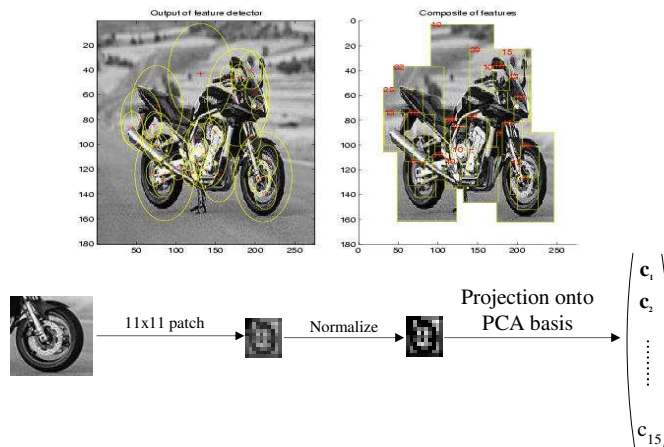


Interest Operator

Kadir and Brady's interest operator
Finds maxima in entropy over scale and location



Representation of Appearance

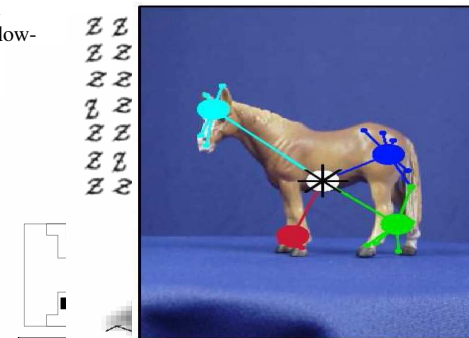


Hierarchical Representations

§ Pixels Pixel groupings Parts Object

- § Multi-scale approach increases number of low-level features

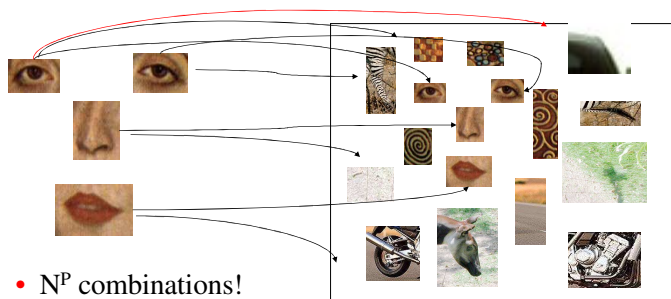
- § [Amit98]
- § [Bouchard05]



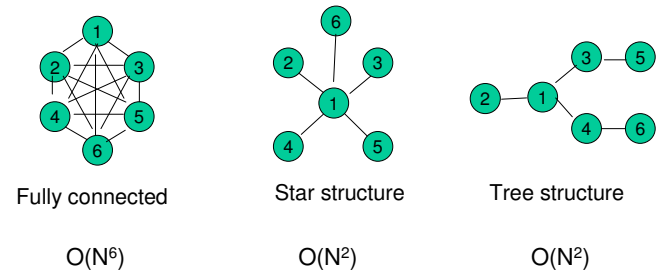
Images from [Amit98,Bouchard05]

The Correspondence Problem

- Model with P parts
- Image with N possible locations for each part



Different Graph Structures



- Sparser graphs cannot capture all interactions between parts

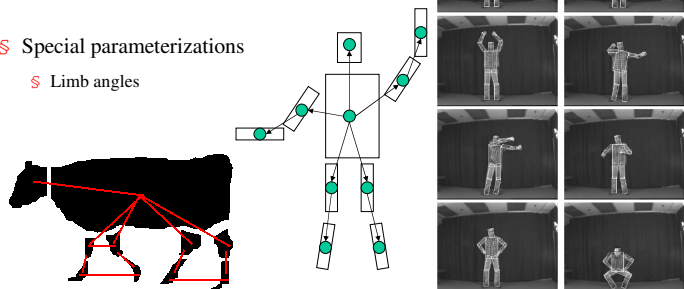
Some Class-Specific Graphs

§ Articulated motion

- § People
- § Animals

§ Special parameterizations

- § Limb angles



Images from [Kumar05, Felzenszwalb05]

Linear-Time Matching Algorithm

§ A *Dynamic Programming* implementation runs in **quadratic time**

§ **Requires tree configuration of parts**

§ Felzenszwalb & Huttenlocher (2000) developed **linear-time** matching algorithm

§ Additional constraint on part-to-part cost function d_{ij}

§ **Basic “Trick”**: Parallelize minimization computation over entire image using a Generalized Distance Transform

Distance Transforms

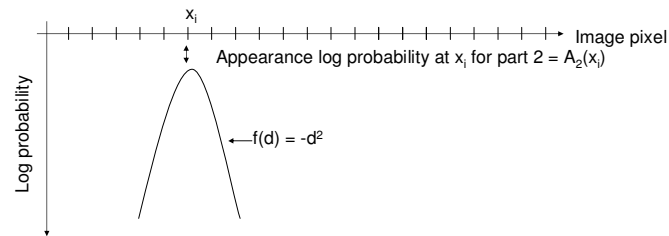
Distance transforms

§ $O(N^2P)$ $O(NP)$ for tree structured models

How it works

§ Assume location model is Gaussian (i.e. e^{-d^2})

§ Consider a two part model with $\mu=0$, $\sigma=1$ on a 1-D image



Model



Distance Transforms 2

§ For each position of landmark part, find best position for part 2

§ Finding most probable x_i is equivalent finding maximum over set of offset parabolas

§ Upper envelope computed in $O(N)$ rather than obvious $O(N^2)$ via distance transform [Felzenszwalb and Huttenlocher '05]

§ Add $A_L(x)$ to upper envelope (offset by μ) to get overall probability map

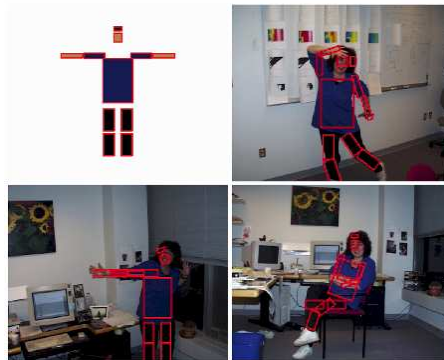
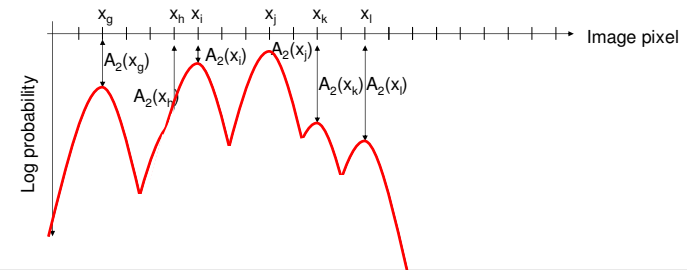
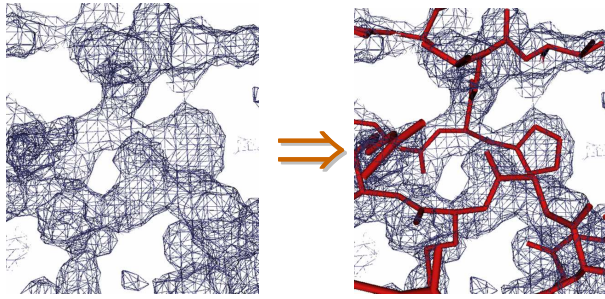


Figure from "Efficient Matching of Pictorial Structures," P. Felzenszwalb and D. Huttenlocher, *Proc. Computer Vision and Pattern Recognition Conf.*, 2000

Using Pictorial Structures to Identify Proteins in X-ray Crystallographic Electron Density Maps

Frank DiMaio
Jude Shavlik
George N. Phillips, Jr.

Task Overview



Given

- Electron density for a region in a protein
- Protein's *topology*

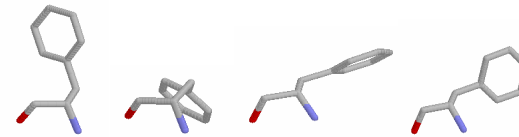
Find

- Atomic positions of individual atoms in the density map

Pictorial Structures for Map Interpretation

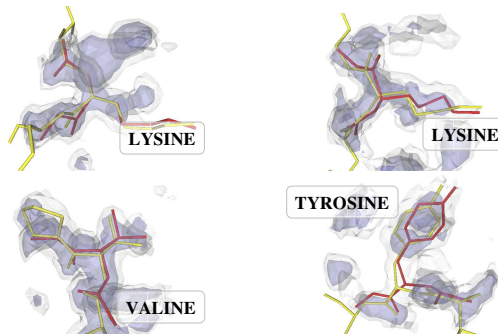
Basic Idea: Build pictorial structure that is able to model *all configurations of a molecule*

- § Each part in “collection of parts” corresponds to an **atom**
- § Model has **low-cost conformation** for **low-energy states** of the molecule



Results

§ **PREDICTED** vs. **ACTUAL**



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
 - § Condition on illumination of landmark part



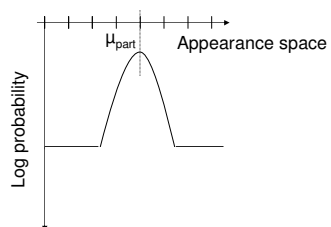
Representation of Occlusion

§ Explicit

§ Additional match of each part to missing state

§ Implicit

§ Truncated minimum probability of appearance



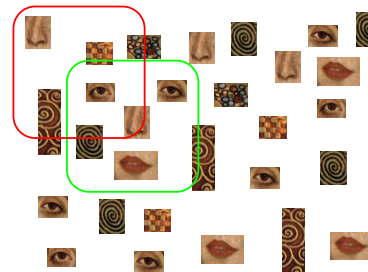
Representation of Background Clutter

§ Explicit model

§ Generative model for clutter as well as foreground object

§ Use a sub-window

§ At correct position, no clutter is present



Object Categorization: The Statistical Viewpoint



$$p(\text{zebra} | \text{image})$$

vs.

$$p(\text{no zebra} | \text{image})$$

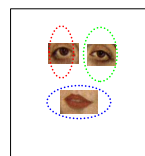
§ Bayes's rule:

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

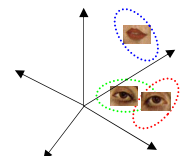
Generative Probabilistic Model

Object model

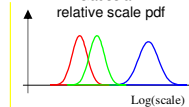
Gaussian shape pdf



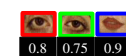
Gaussian part appearance pdf



Gaussian relative scale pdf



Prob. of detection

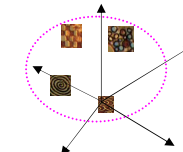


Background clutter model

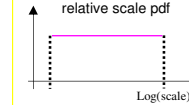
Uniform shape pdf



Gaussian appearance pdf



Uniform relative scale pdf



Poisson pdf on # detections

Model Structure

- Assume prior ratio is known or learned
- Find values for parameters θ that maximizes the likelihood ratio

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

- H is the set of all valid correspondences of image features to model parts, so $|H| = O(N^P)$ in general
- Factor the likelihood to simplify computation (using Chain Rule)

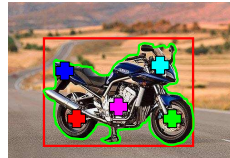
Learning

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

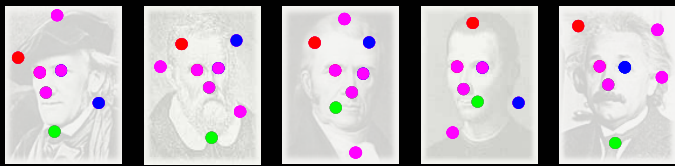
Learning using EM

- 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



Learning procedure

- Find regions & their location & appearance
- Initialize model parameters
- Use EM algorithm and iterate to convergence:
 - E-step: Compute assignments for which regions belong to which part (red, green and blue dots)
 - M-step: Update model parameters
- Try to maximize likelihood – consistency in shape & appearance



Recognition

- § For each of P parts, run template over all locations in image
- § Detect local maxima, giving possible locations of each part
- § Given learned model, find maximum likelihood ratio of $p(\mathbf{X}, \mathbf{S}, \mathbf{A} | \theta) / p(\mathbf{X}, \mathbf{S}, \mathbf{A} | \theta_{bg})$ for all possible correspondences – $O(N^2P)$ where N = number of locations of each part in image
- § If greater than a threshold, signify object detected

Experimental Procedure

Two series of experiments:

1. Scale variant (using pre-scaled images)
2. Scale invariant

$P = 6-7$
 $N = 20-30$
20-30 parameters/part
10-15 PCA features

Datasets:

§ Motorbikes, Faces, Spotted cats, Airplanes, Cars from behind and side

§ **200 - 800 images**



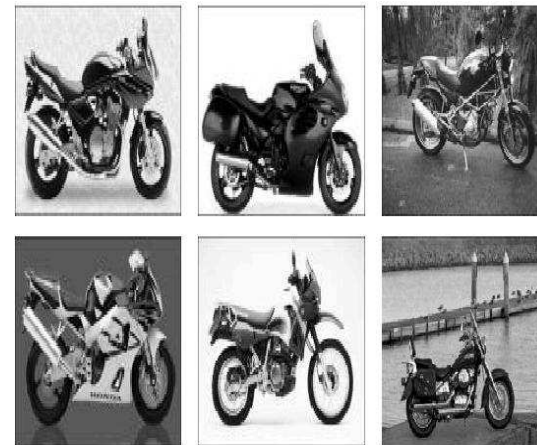
Training

- § 50% images
- § No identification of object within image

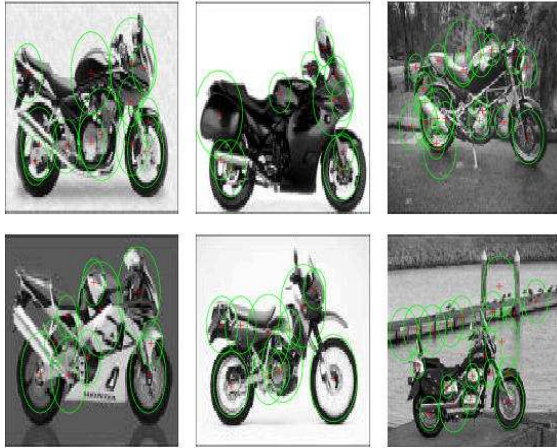
Testing

- § 50% images
- § Simple object present/absent test
- § ROC equal error rate computed, using background set of images

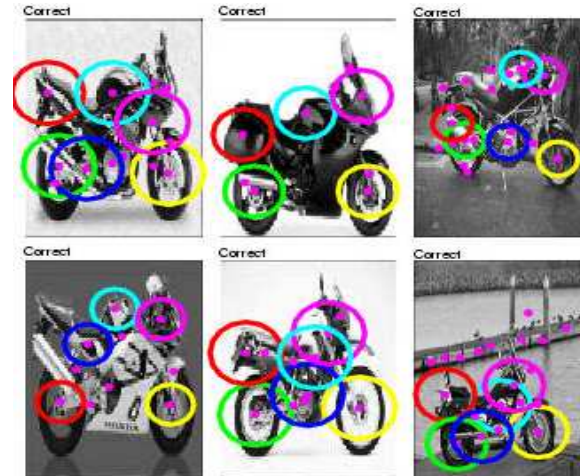
Motorbikes: Input Images



Motorbikes: Features Detected



Motorbikes: Max Likelihood Result



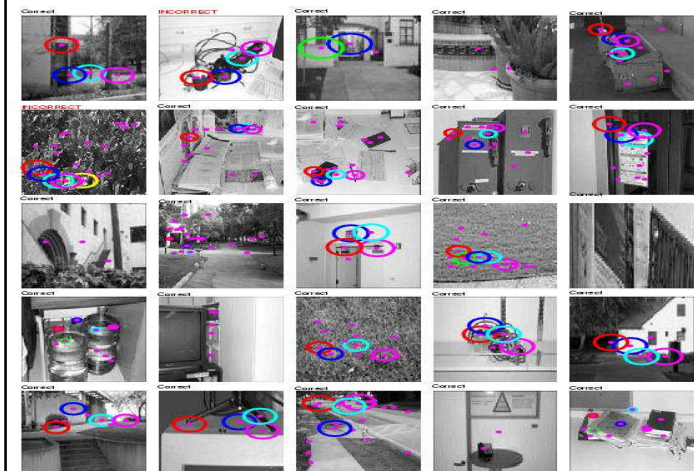
Equal error rate: 7.5%

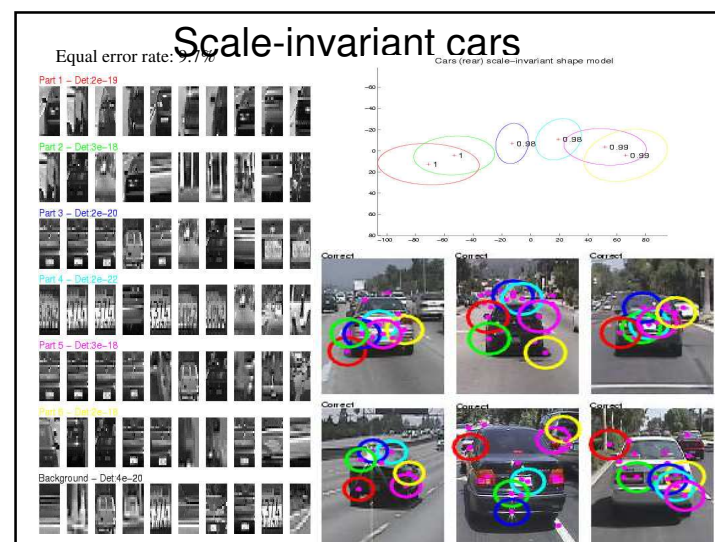
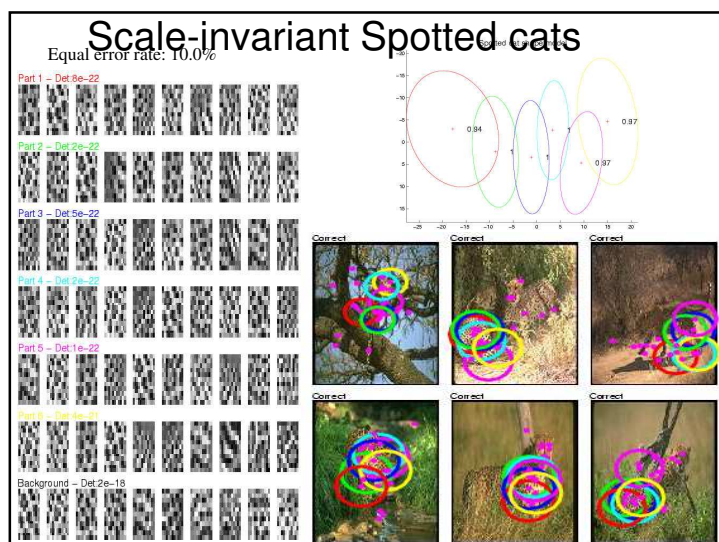
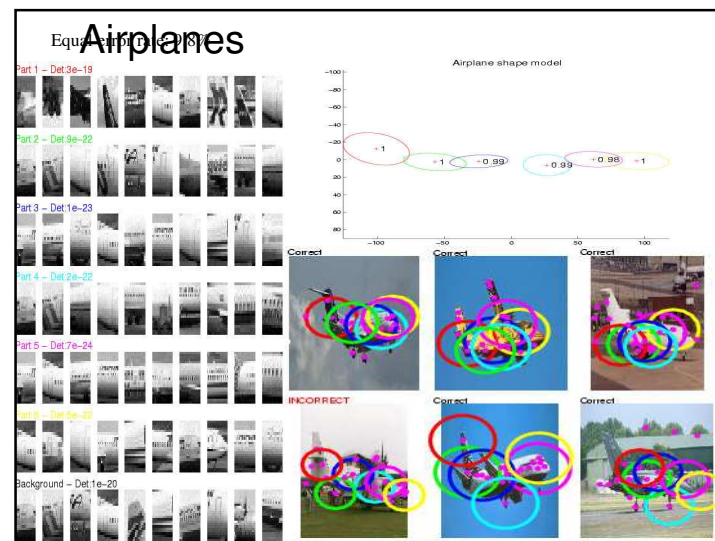
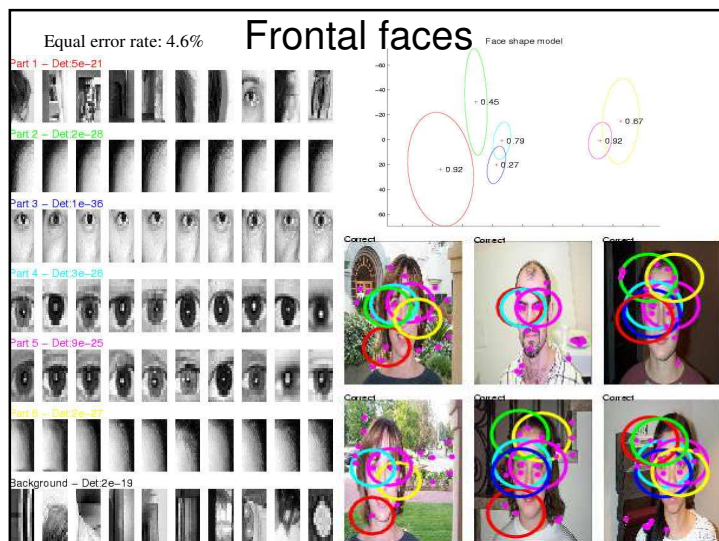
Motorbikes

Shape Model

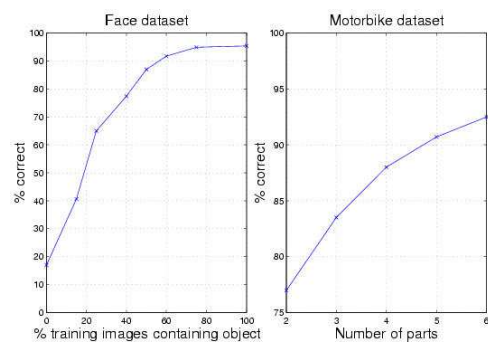


Background images





Robustness of algorithm



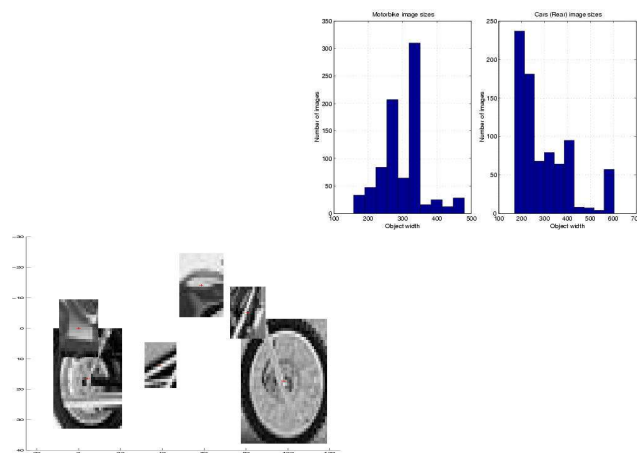
ROC equal error rates

Pre-scaled data (identical settings):

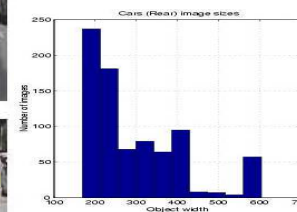
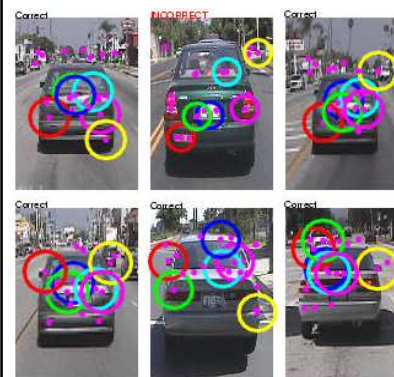
Dataset	Total size of dataset	~ Object width (pixels)	Model			
			Motorbikes	Faces	Airplanes	Spotted Cats
Motorbikes	800	200	92.5	50	51	56
Faces	435	300	33	96.4	32	32
Airplanes	800	300	64	63	90.2	53
Spotted Cats	200	80	48	44	51	90.0

Scale-invariant learning and recognition:

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



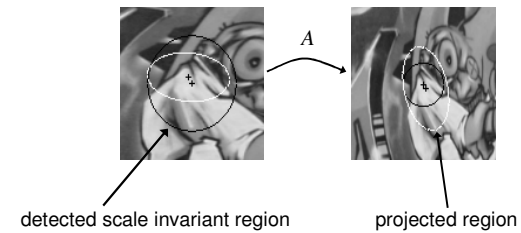
Scale-invariant cars





Adding Viewpoint Invariance

§ Locally approximated by an affine transformation



Affine-Invariant Patches

Lindeberg & Garding (1997); Mikolajczyk & Schmid (2002);
Tell & Carlsson (2000); Tuytelaars & Van Gool (2002)



Idea:
3D objects are never planar
in the large, but they are
always planar in the small



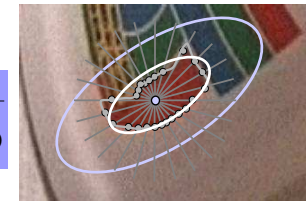
Representation: Local
invariants and their
spatial layout

Intensity-based Method for Detecting Affine-Invariant Interest Points

Tuytelaars et al., 2000

1. Search for intensity extrema
2. Observe intensity profile along rays
3. Search for maximum of invariant
function $f(t)$ along each ray
4. Connect local maxima
5. Fit ellipse
6. Double ellipse size

$$f(t) = \frac{\text{abs}(I_0 - I)}{\max\left(\frac{\int \text{abs}(I_0 - I) dt}{t}, d\right)}$$



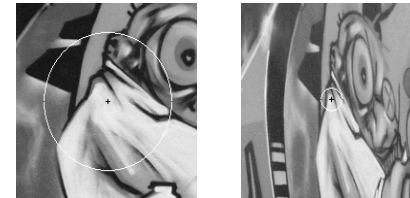
Affine Invariant Harris Interest Points

- § Localization & scale influence affine neighborhood
- § => affine invariant Harris points (Mikolajczyk & Schmid'02)
- § Iterative estimation of these parameters
 - § **localization** – local maximum of the Harris measure
 - § **scale** – automatic scale selection with the Laplacian
 - § **affine neighborhood** – normalization with second moment matrix
- § Repeat estimation until convergence
- § Initialization with multi-scale interest points

Affine invariant Harris points

- § Iterative estimation of localization, scale, neighborhood

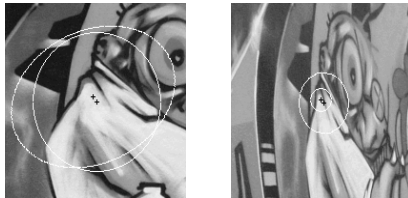
Initial points



Affine invariant Harris points

- § Iterative estimation of localization, scale, neighborhood

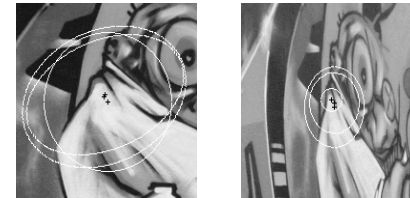
Iteration #1



Affine invariant Harris points

- § Iterative estimation of localization, scale, neighborhood

Iteration #2



Affine invariant Harris points

§ Initialization with multi-scale interest points



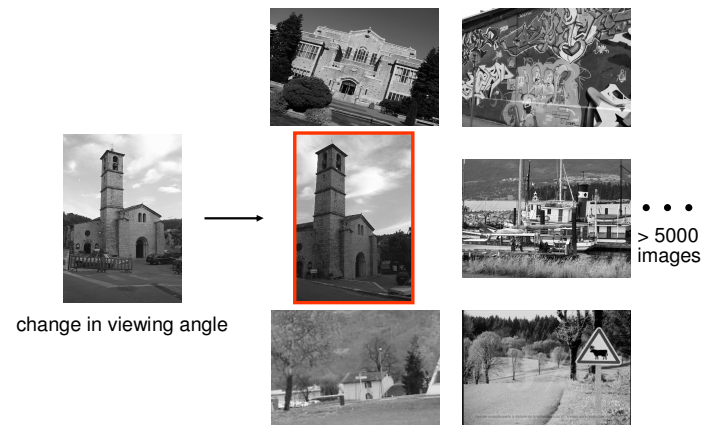
§ Iterative modification of location, scale and neighborhood



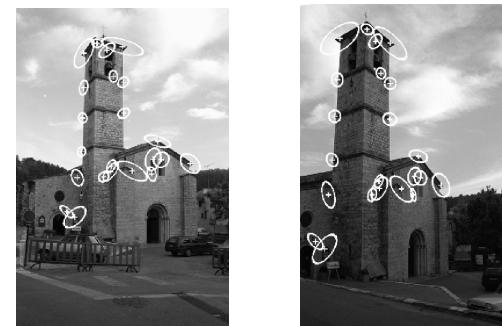
Affine Invariant Interest Point Detection



Application: Image Retrieval

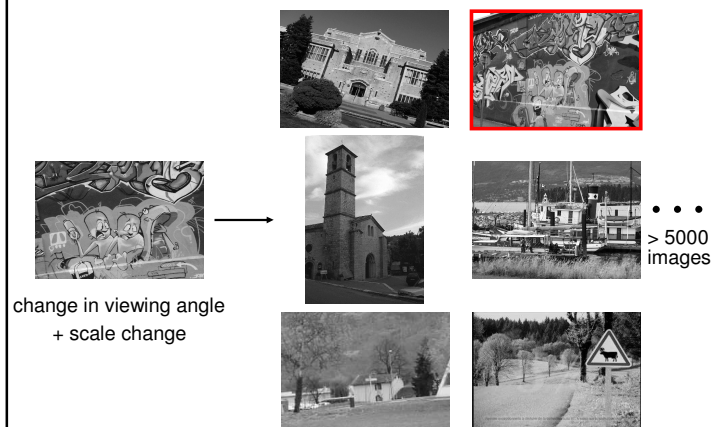


Matches



22 correct matches

Application: Image Retrieval



Matches



33 correct matches



Figure 2: Model gallery: sample input images and renderings of the corresponding models.

Jean Ponce¹, Svetlana Lazebnik¹, Fredrick Rothganger¹, Cordelia Schmid²



Figure 3: Object recognition experiments. The three rows of this figure show (respectively) input images, model patches matched to these images, and recognized models rendered in their estimated pose. Note that the teddy bear in the leftmost column is in a pose quite different from those used to acquire its model. Also note the significant amount of clutter and occlusion in each image.

Jean Ponce¹, Svetlana Lazebnik¹, Fredrick Rothganger¹, Cordelia Schmid²

Application: Photo Tourism

- § <http://phototour.cs.washington.edu/>
- § Detect and match local patch features across images of a scene taken by many different people and found via shared image databases such as Flickr

Photo Tourism

Exploring photo collections in 3D

Noah Snavely Steven M. Seitz Richard Szeliski
University of Washington Microsoft Research

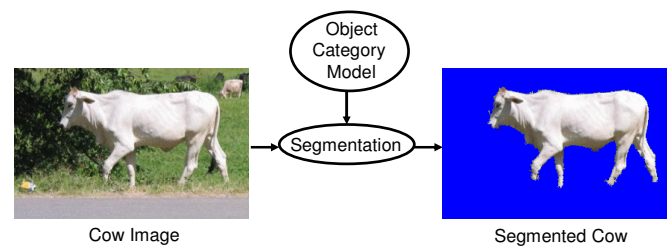
SIGGRAPH 2006

Probabilistic Parts and Structure Models Summary

- § Correspondence problem
- § Efficient methods for large # parts and # positions in image
- § Challenge to get representation with desired invariance
- § Minimal supervision
- § Future directions:
 - § Multiple views
 - § Approaches to learning
 - § Multiple category training

Combining Segmentation and Recognition

- § Example: Given an image and object category, segment the object



Segmentation should (ideally) be

- shaped like the object, e.g., cow-like
- obtained efficiently in an unsupervised manner
- able to handle self-occlusion