Constellation Models for Object Recognition

Amir Reza Saffari Azar Alamdari

Institute for Theoretical Computer Science,

Graz University of Technology

amir@igi.tugraz.at

Contents

- Introduction
- Constellation Models
- Scale-Invariant Constellation Model
- Heterogeneous Star Model

Introduction

- What is ***Object?***
- What is ***Object Recognition***?

Introduction (cont.)

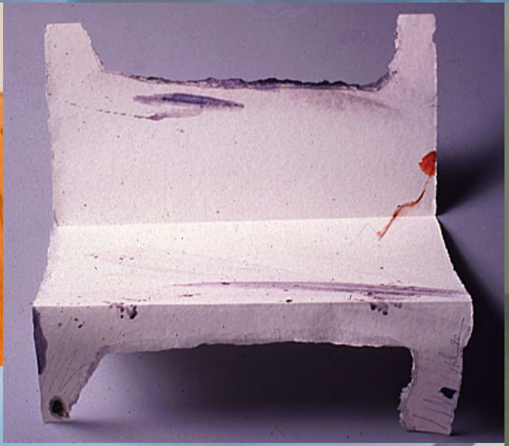
- Variations: such as scaling, translation, rotation, deformation, noise, occlusion, background.
- Object recognition system must be ***Invariant*** to these transformations.

Introduction (cont.)

- ***Classes of Objects*** are collection of objects that are similar (also known as ***Category***).
- ***Object Categorization:*** Invariance to interclass variations.
- Invariance, ***Generalization***.

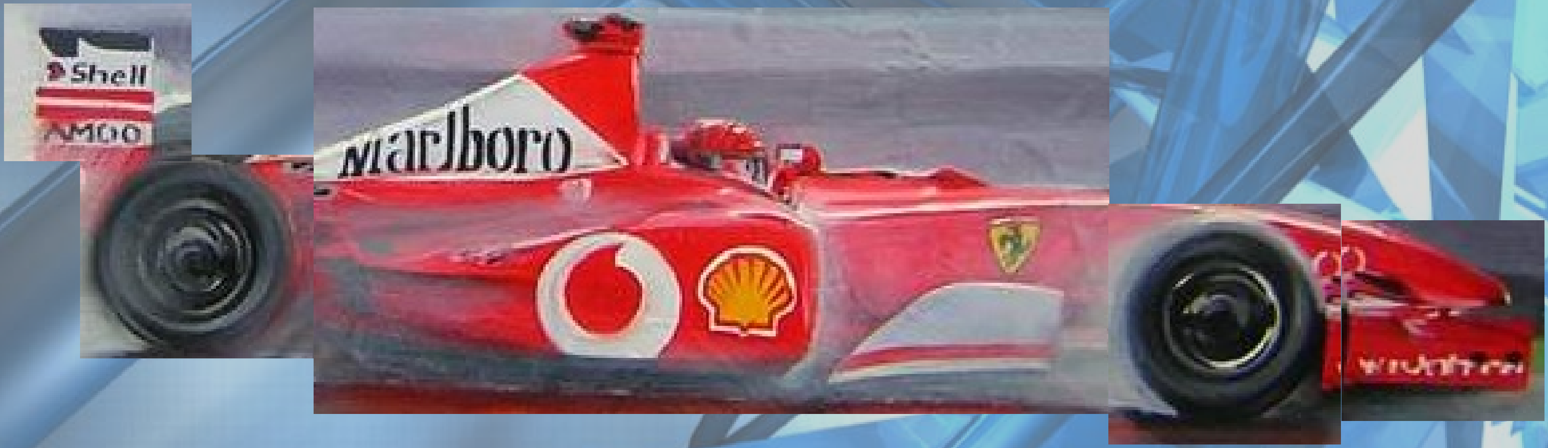
Motivation

- What makes objects within a class similar?
- Abstract higher-level definitions.



Motivation (cont.)

- Intermediate-level: Visually similar.
- Decomposing objects into ***parts***.



Part Properties

1. The parts of one object correspond to parts in other objects belonging to the same class.
2. Corresponding parts are similar; more so than objects belonging to the same class are similar.
3. Detectors can be devised that can detect the part with some degree of success.
4. A part has geometrical properties, such as a spatial position, often a scale, an orientation, a spatial extent, a velocity.

Part Properties (cont.)

Part Properties (cont.)

Signal Variability

- Variability due to ***Absence of Features***
- Variability due to ***Deformation***
- Variability due to ***Pose Changes***
- Variability due to ***Background Clutter***
- Variability due to ***Lighting***

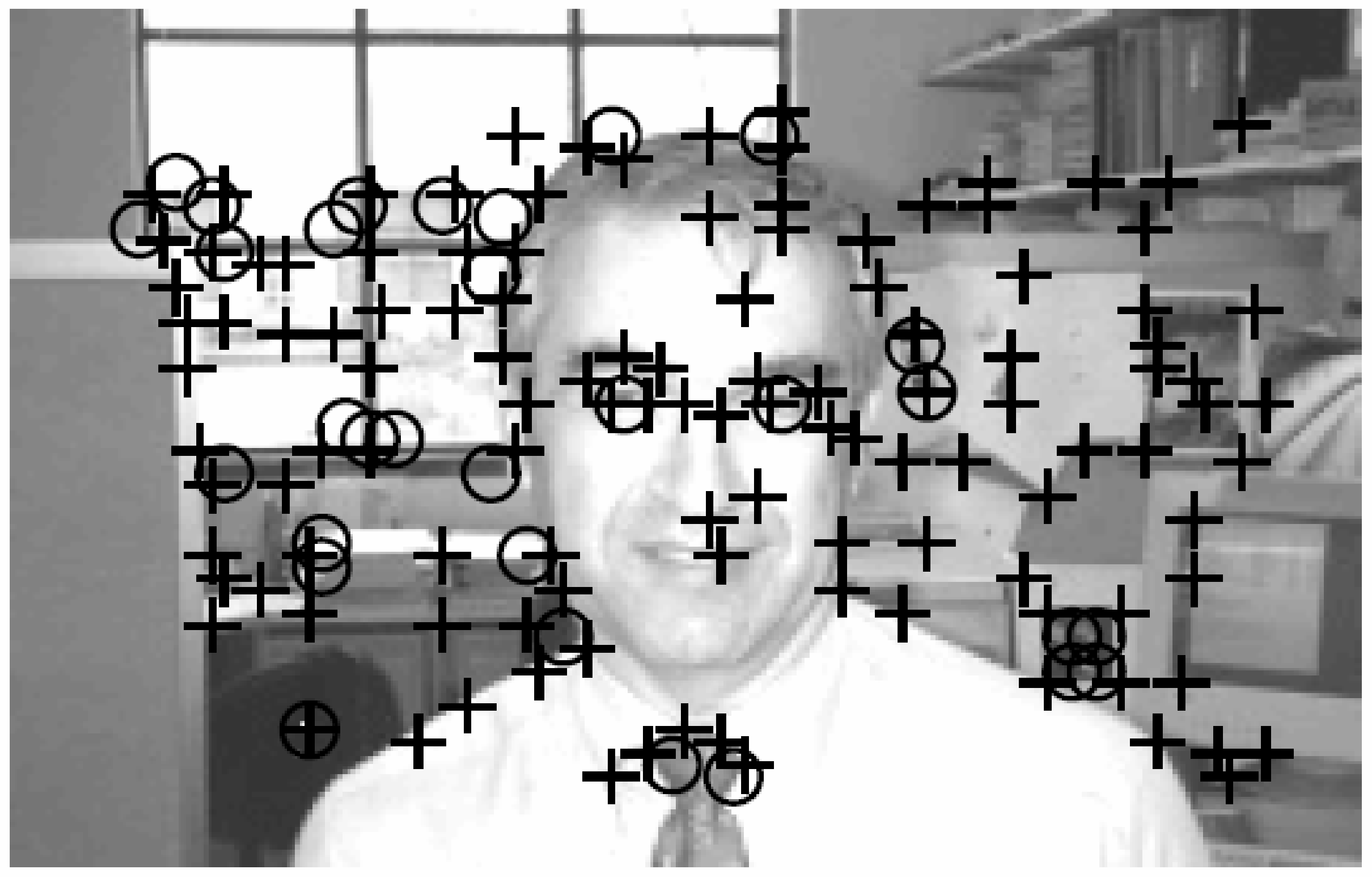
Problem Formulation

- Problem: Given a new image decide whether an instance of an object class is present or absent in the image.
- The only supervision is that an image has or doesn‘t has an object, without any segmentation and localization: weakly supervised.

Decomposition into Parts

- Transform entire image into a collection of parts.
- Some detected parts might correspond to an instance of target object (*foreground*), others are from background or false detection (*background*).
- Assume that there are *T* different type of parts.

Decomposition into Parts (cont.)



Decomposition into Parts (cont.)

- The position for all parts:

$$\mathbf{X}^{o} = \begin{bmatrix} x_{11}, x_{12}, \ldots, x_{1N_1} \\ x_{21}, x_{22}, \ldots, x_{2N_2} \\ \vdots \\ x_{T1}, x_{T2}, \ldots, x_{TN_T} \end{bmatrix}$$

Hypothesis Construction

- Assume object is composed of *F* parts:

$$F \geq T$$

- *Hypothesis*, **h**, is a vector of length *F*, indicating part indices for foreground object: $h_i = j$ shows that point x_{ij} is a point of foreground.
- If an object part is not contained in foreground the corresponding point in **h** will be zero.

Model Description

- From *maximum a posterior probability* (MAP), classification can be viewed as calculating the ratio:

$$R = \frac{p(\text{Object} \mid \mathbf{X}^o)}{p(\text{No Object} \mid \mathbf{X}^o)} = \frac{p(\mathbf{X}^o \mid \text{Object})\, p(\text{Object})}{p(\mathbf{X}^o \mid \text{No Object})\, p(\text{No Object})}$$

Model Description (cont.)

$$\frac{p(\mathbf{X}^o \mid \text{Object})}{p(\mathbf{X}^o \mid \text{No Object})} = \frac{\sum_{\mathbf{h}} p(\mathbf{X}^o, \mathbf{h} \mid \text{Object})}{p(\mathbf{X}^o, \mathbf{h}_0 \mid \text{No Object})}$$

$$p(\mathbf{X}^o, \mathbf{h} \mid O) = p(\mathbf{X}^o \mid \mathbf{h}, \mathbf{n})\, p(\mathbf{h} \mid \mathbf{n}, \mathbf{d})\, p(\mathbf{n})\, p(\mathbf{d} \mid O)$$

Learning

- Part detectors:
  1. Apply interest point detectors to image.
  2. Do an unsupervised clustering over all parts.
  3. Eliminate all small clusters.
  4. Compute centers of clusters as part detector template.
  5. Apply feature selection method to reduce the number of parts into acceptable range.

Learning (cont.)

- Learning of model parameters: *expectation maximization* (EM) method.
- The EM algorithm is designed to estimate model parameters according to some observations, but in situations where some necessary data is missing.

Summary of Basics

- The model explicitly accounts for shape variations.
- The randomness in presence and absence of features due to false detections or occlusion is approached in a principled manner.
- It explicitly accounts for image clutter.
- The model presents unsupervised automatic way to detect and learn feature extractors.

Scale-Invariant Model

- In feature extraction step, for any given image, extract *N* interesting features with locations **X**, scales **S**, and appearance **A**.
- Assume that object has *P* parts.

Decision

- Bayesian decision:

$$R = \frac{p(\text{Object} \mid \mathbf{X}, \mathbf{S}, \mathbf{A})}{p(\text{No Object} \mid \mathbf{X}, \mathbf{S}, \mathbf{A})}$$

$$= \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{A} \mid \text{Object})\, p(\text{Object})}{p(\mathbf{X}, \mathbf{S}, \mathbf{A} \mid \text{No Object})\, p(\text{No Object})}$$

$$\approx \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{A} \mid \theta)\, p(\text{Object})}{p(\mathbf{X}, \mathbf{S}, \mathbf{A} \mid \theta_{\text{bg}})\, p(\text{No Object})}$$

Likelihood Factorization

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

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

- **h** is a *hypothesis.*

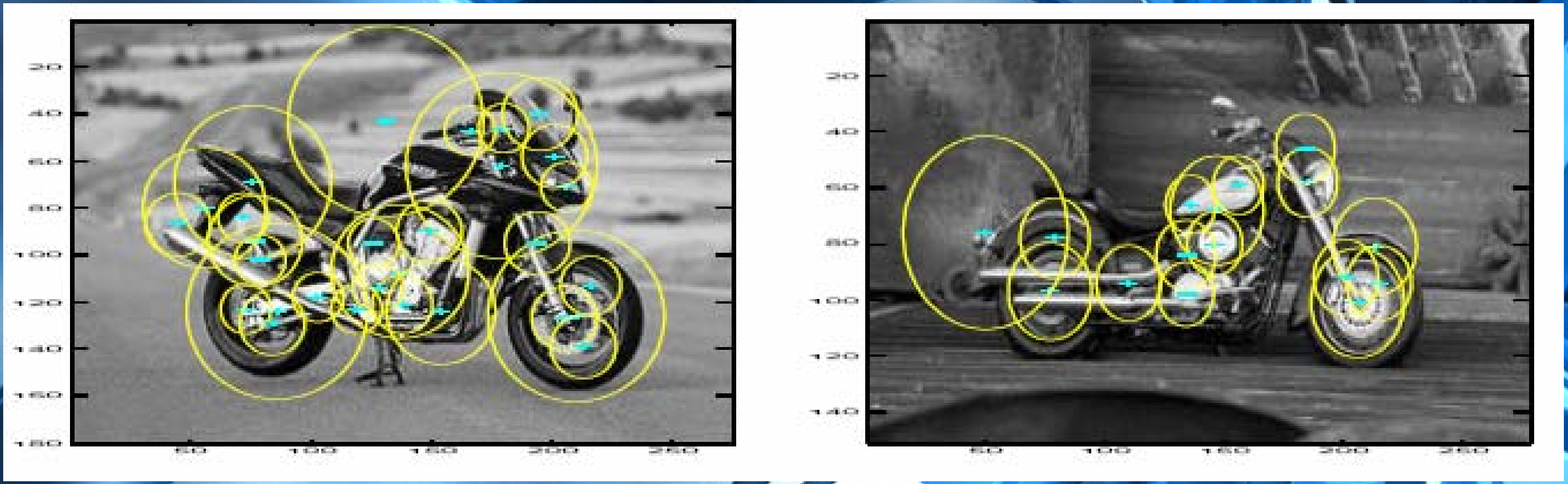
Feature Detection

- Features are found using the detector of Kadir & Brady.
- In this method first a histogram of intensities, *P(I)*, in a circular region of radius *s* (scale) around each pixel is made, then the local maxima of entropy of this histogram, *H(s)*, is considered as feature scale.

Feature Detection (cont.)

- The saliency (importance) of each feature is computed as $H\frac{dP}{ds}$, after appropriate normalization for scale.
- The *N* regions with highest saliency over image provide the features. Each feature has information about the position and scale.
- The saliency measure is designed to be invariant due to scale.

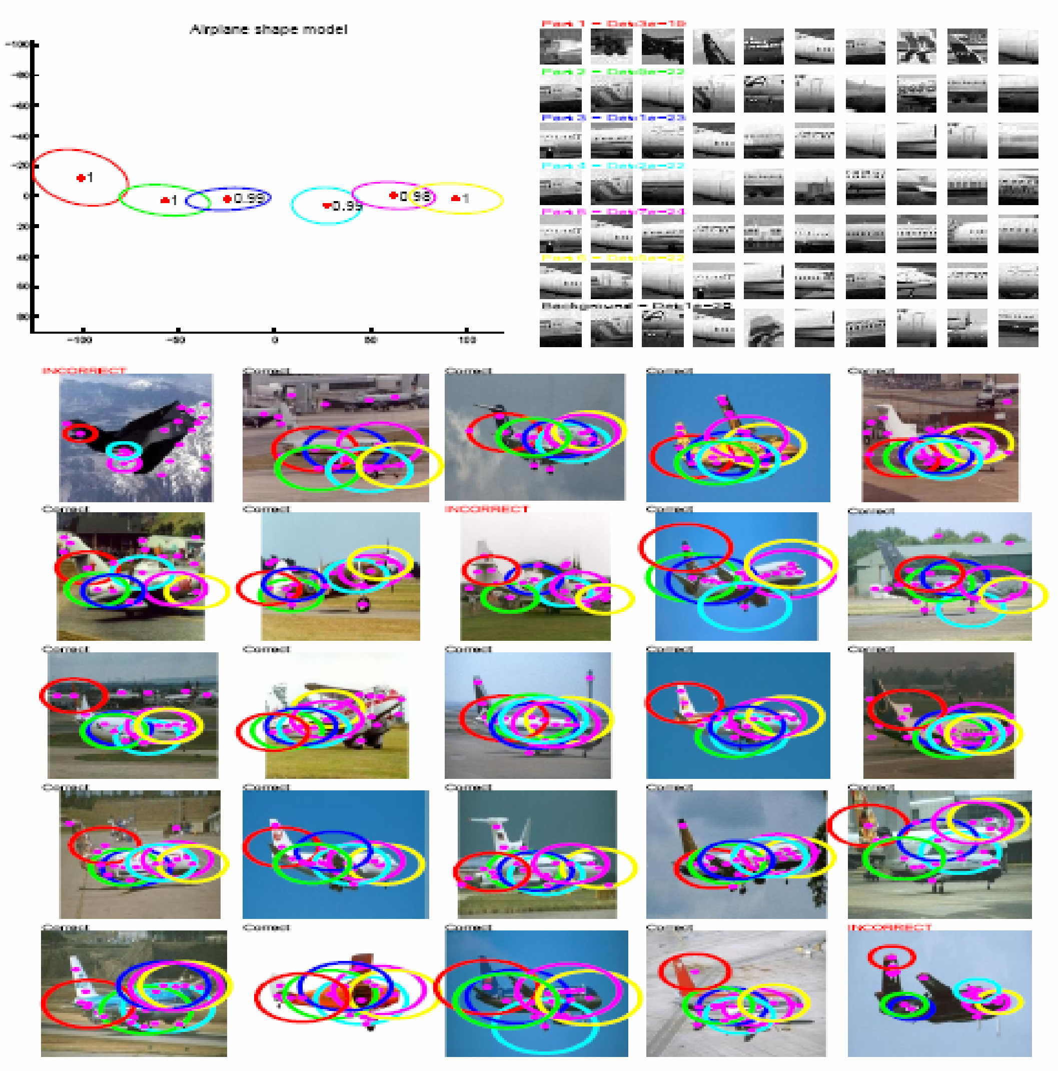
Feature Detection (cont.)



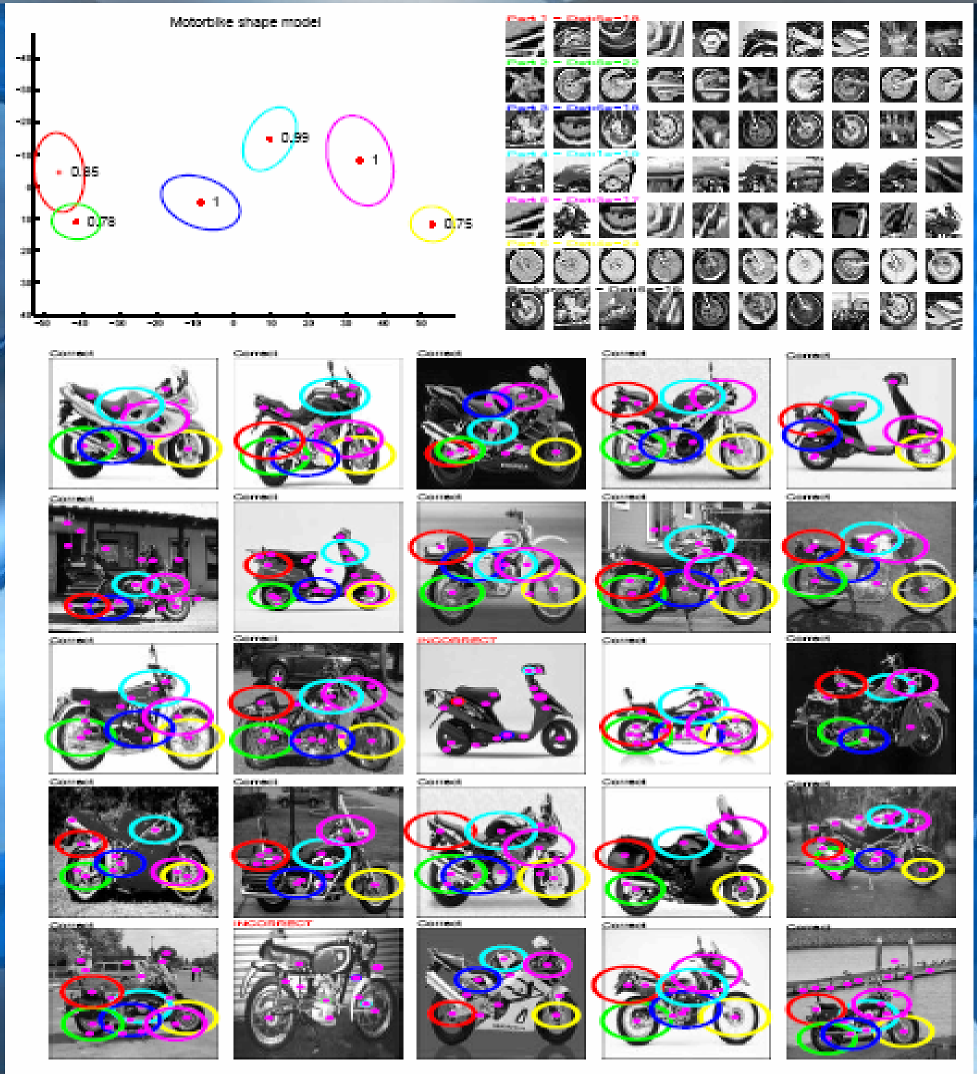
Feature Representation

- Each region is cropped from image and is rescaled to a small size.
- In the learning stage, a PCA is applied to all collected patches from all images. Then the first *k* principal components are selected and represented as appearance, **A**.

Experimental Results



Experimental Results (cont.)



Problems

- The joint nature of the shape model results in an exponential explosion in computational cost, limiting the number of parts and regions per image that can be handled. For *N* feature detections, and *P* model parts the complexity for both learning and recognition is $O(N^P)$.

Problems (cont.)

- Only one type of interest operator (a region operator) was used, making the model very sensitive to the nature of the class.
- The model has many parameters resulting in overfitting unless a large number of training images (typically 200+) are used.

Solutions

- In the new model, both in learning and recognition has a lower complexity than the constellation model. This enables both the number of parts and the number of detected features to be increased substantially.
- It is heterogeneous and is able to make the optimum selection of feature types (here from a pool of three, including curves).

Heterogeneous Star Model

- A Heterogeneous Star Model (HSM) which maintains the simple training aspect of the constellation model, and also, like the constellation model, gives a localization for the recognized object is proposed. The model is translation and scale invariant both in learning and in recognition.

Star Model

- Assume model has *P* parts and parameters θ .
- From each image, extarct *N* features with locations $\mathbf{X}$, scales $\mathbf{S}$ and descriptors $\mathbf{D}$.
- Joint density:

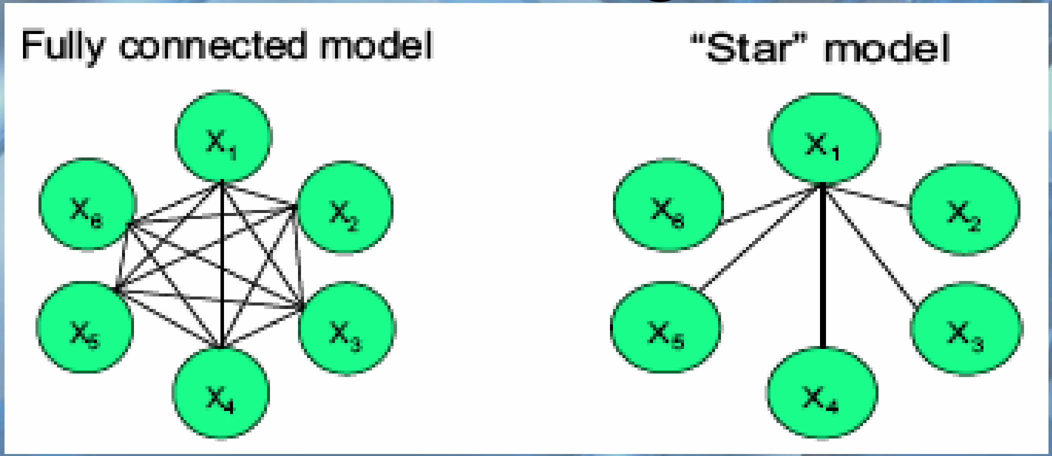
$$p(\mathbf{X},\mathbf{D},\mathbf{S},\mathbf{h}\,|\,\theta) = \underbrace{p(\mathbf{D}\,|\,\mathbf{h},\theta)}_{Appearance}\underbrace{p(\mathbf{X}\,|\,\mathbf{S},\mathbf{h},\theta)}_{Rel.Locations}\underbrace{p(\mathbf{S}\,|\,\mathbf{h},\theta)}_{Rel.Scale}\underbrace{p(\mathbf{h}\,|\,\theta)}_{Other}$$

Star Model (cont.)

- In Constellation model, the appearance models for each part is assumed independent but the relative location of the model parts is represented by a joint Gaussian density. While this provides the most thorough description, it makes the location of all parts dependent on one another.

Star Model (cont.)

- In this model a simplified configuration model in which the location of the model part is conditioned on the location of a *landmark* part is proposed. Under this model the non-landmark parts are independent of one another given the landmark.



Star Model (cont.)

- Relative location:

$$p(\mathbf{X} \mid \mathbf{S}, \mathbf{h}, \theta) = p(\mathbf{x}_L \mid h_L) \prod_{j \neq L} p(\mathbf{x}_j \mid \mathbf{x}_L, s_L, h_j, \theta_j)$$

- As a result dependecies between parts in density function is reduced considerabily, so computations with much more parts and features are possible.

Star Model (cont.)

- In practical terms, one can achieve translation invariance by subtracting the location of the landmark part from the non-landmark ones. Scale invariance is achieved by dividing the location of the non-landmark parts by the locally measured scale of the landmark part.

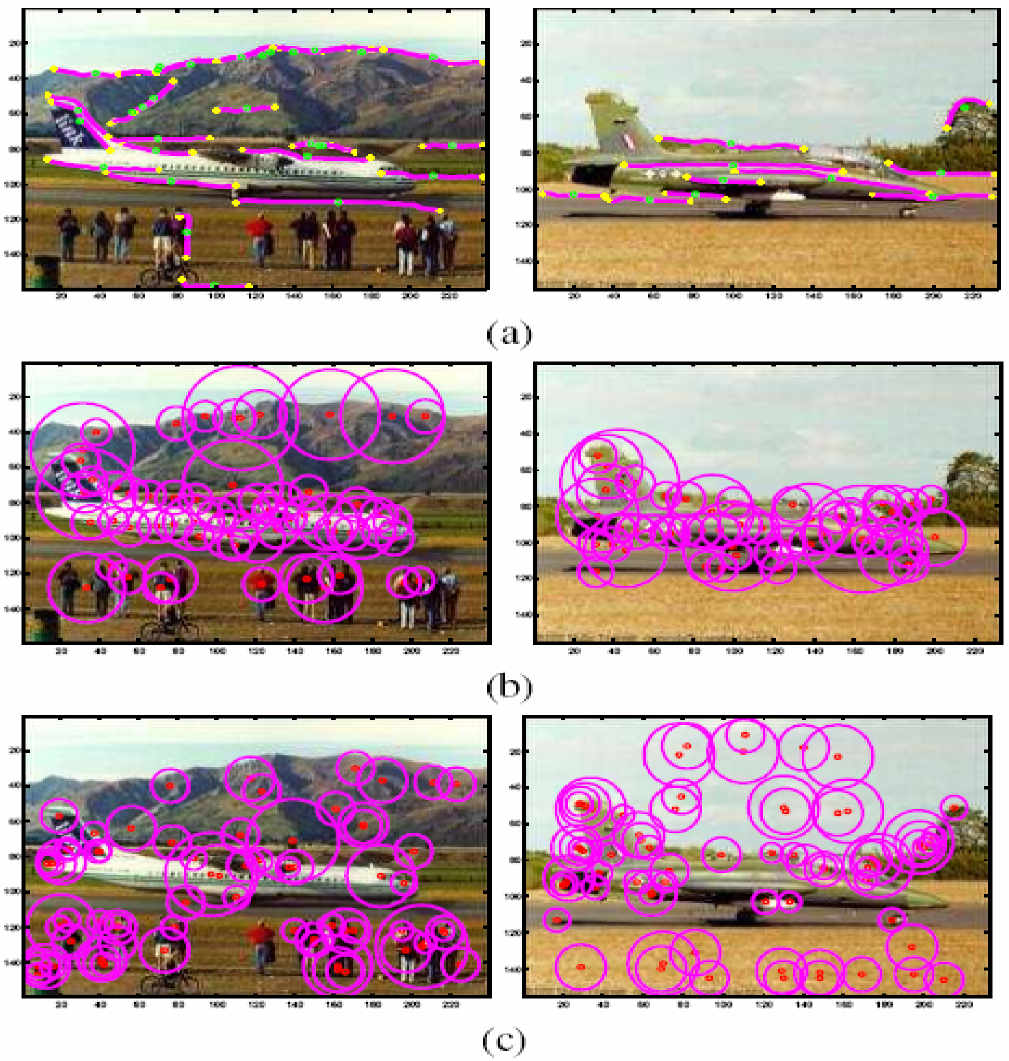
Limitations of Star Model

- Occlusion of landmark part can not be handled.
- Assumption: Landmark part is always visible.
- With large number of features, *N*, the chance of landmark detection increases.

Feature Detectors

- The models can utilize a combination of different features detectors, the optimal selection being made automatically.
- Kadir & Brady, multi-scale Harris and Curves.
- Kadir & Brady favors circular regions; multi-scale Harris prefers interest points, and curves locate the outline of the object.

Feature Detectors (cont.)



(a)

(b)

(c)

Feature Representation

- Each feature operator gives both location and scale of the region. Each region is cropped from image, and rescaled to a *k*k* patch.
- Its gradient is computed and then normalized.
- Then PCA is applied and first *d* principal component is chosen.

Feature Representation (cont.)

- Two additional measurements are made for each gradient-patch: its unnormalized energy and the reconstruction error between the point in the PCA basis and the original gradient-patch.
- Each region is thus represented by a vector of length $d + 2$.

Learning

- The EM algorithm is used for learning of model parameters.
- Different combination of feature operators are tested and chosen using a validation set.
- Kadir & Brady (KB); multi-Scale Harris (MSH); Curves (C); KB + MSH; KB + C; MSH + C; KB + MSH + C.

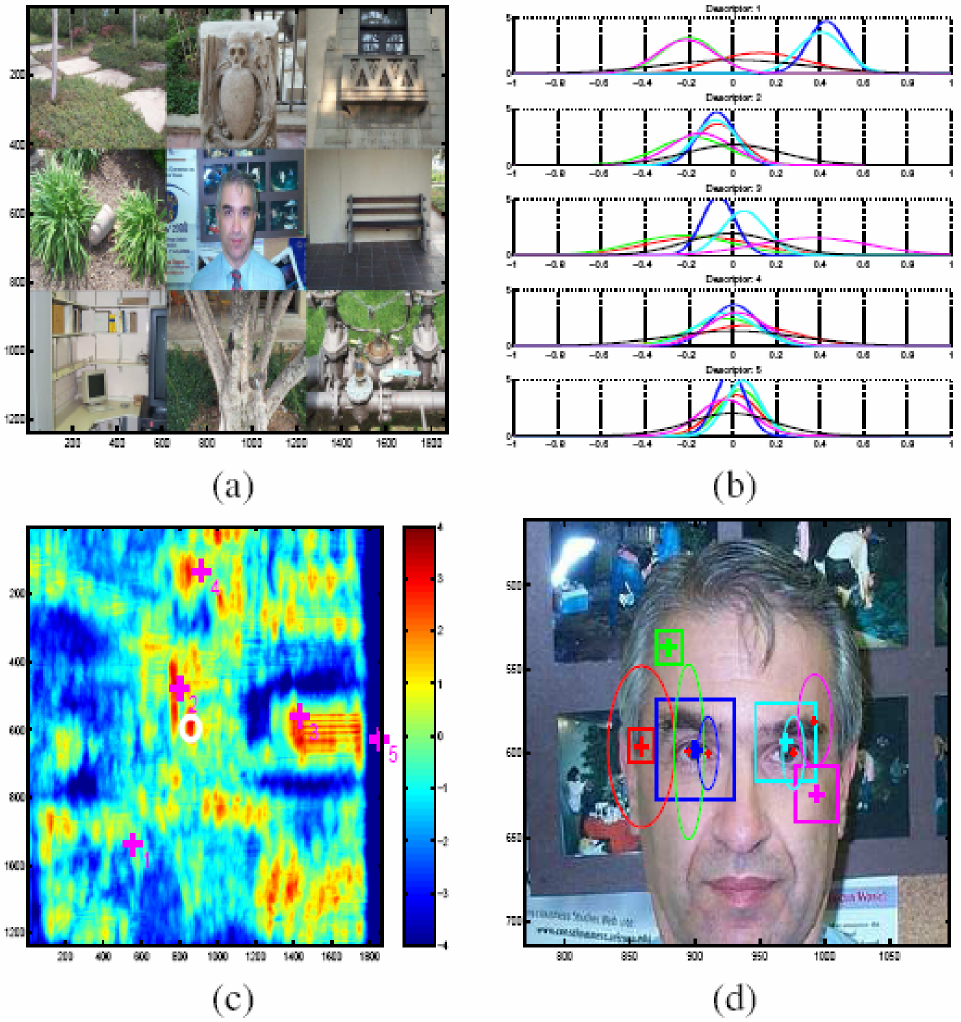
Recognition

- Recognition using features.
- Exhaustive recognition without features: this method can be used only in recognition and replaces feature detection part with searching the appearance densities exhaustively over the entire image (and at different scales). At each location and scale, the likelihood ratio for each part is calculated.

Recognition (cont.)

- In more detail, each PCA basis vector is convolved with the image (employing appropriate normalizations), so projecting every patch in the image into the PCA basis.
- For a given model, the likelihood ratio of each part's appearance density to the background density is then computed at every location, giving a likelihood-ratio map over the entire image for that part.

Recognition (cont.)



(a)

(b)

(c)

(d)

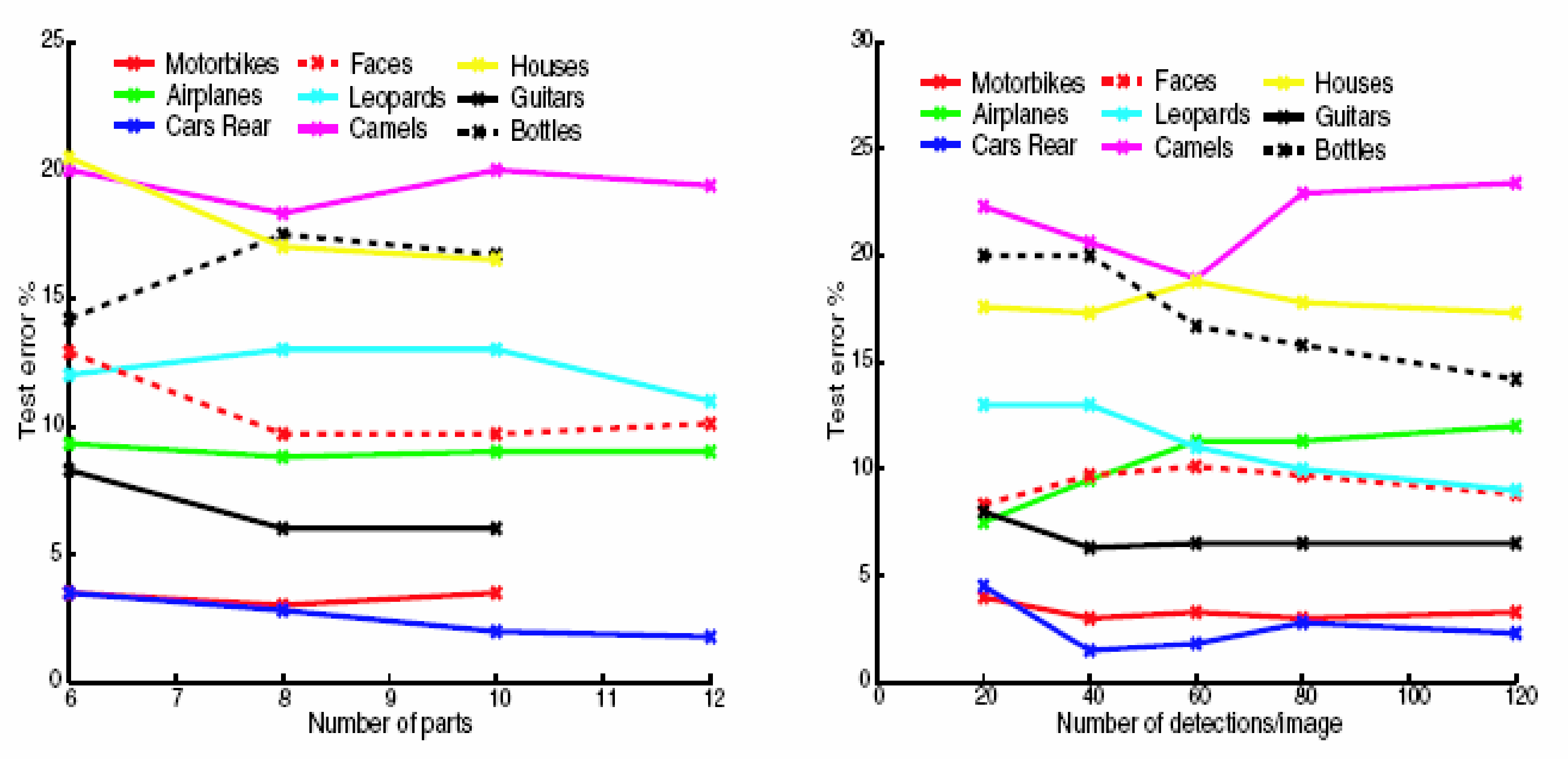
Experimental Results

- The comparison of HSM to the fully connected model.
- The effect of increasing the number of parts and detections/image.
- The difference between feature-based and exhaustive recognition.
- Datasets: Airplanes, Bottles, Camels, Cars (rear), Faces, Guitars, Houses, Leopards, Motorbikes.

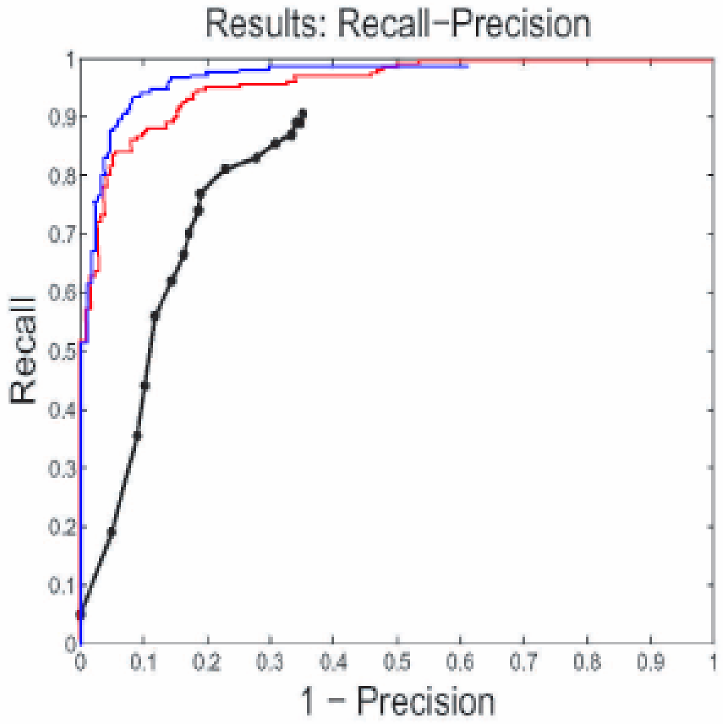
HSM vs. Constellation

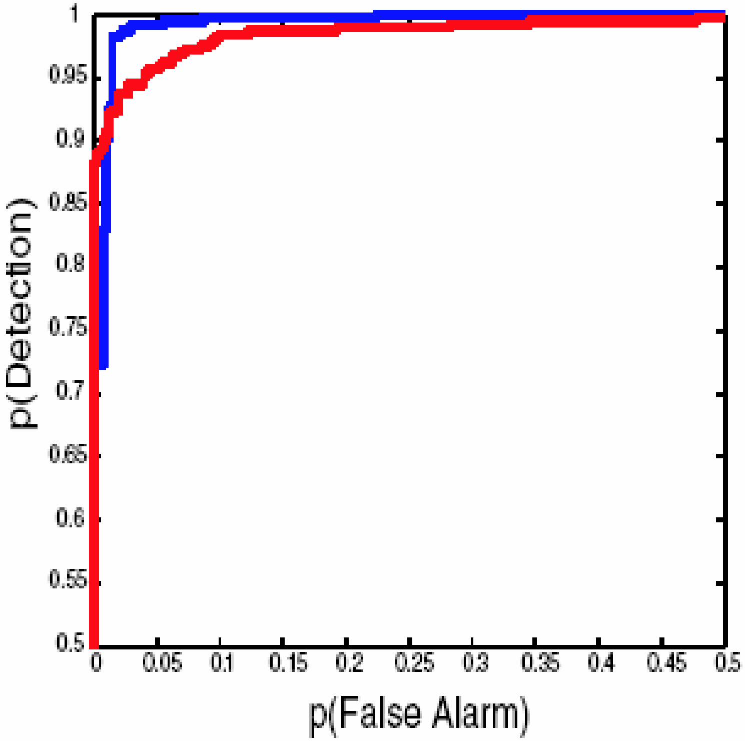
	Total size	Full model	Star model
Dataset	of dataset	test error (%)	test error (%)
Airplanes	800	6.4	6.8
Bottles	247	23.6	27.5
Camels	350	23.0	25.7
Cars (Rear)	900	15.8	12.3
Faces	435	9.7	11.9
Guitars	800	7.6	8.3
Houses	800	19.0	21.1
Leopards	200	12.0	15.0
Motorbikes	900	2.7	4.0

Number of parts



Exhaustive Search





References

1. M. Weber (2000) ***Unsupervised Learning of Models for Object Recognition***, PhD Thesis, California Institute of Technology, Pasadena, CA.
2. M. Weber, M. Welling, P. Perona (1999) ***Unsupervised Learning of Models for Recognition***, *6th Annual Joint Symposium on Neural Computation, (JNSC).*
3. R. Fergus, P. Perona, and A. Zisserman (2003) ***Object Class Recognition by Unsupervised Scale-Invariant Learning*** *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
4. R. Fergus, P. Perona, A. Zisserman (2005) **A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition,** *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*