Object & scene classification for large image collections

Anna Bosch Rué[1]

Andrew Zisserman[2] **Xavier Muñoz**[1]

[1]Arquitectura i Tecnologia de Computadors
Universitat de Girona

[2]Visual Geometry Group
University of Oxford

INTRODUCTION

Introduction

Low-level processing

Sensing

Preprocessing

Segmentation

Description

Object Recognition

High-level processing

Scene Classification

Introduction

Low-level processing

Sensing

Preprocessing

Segmentation

Description

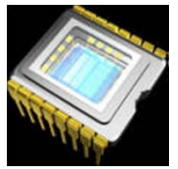
Object Recognition

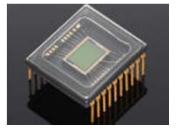
Scene Classification

High-level processing

Lens focus

shutter





CCD

Introduction

Low-level processing

Sensing

Preprocessing

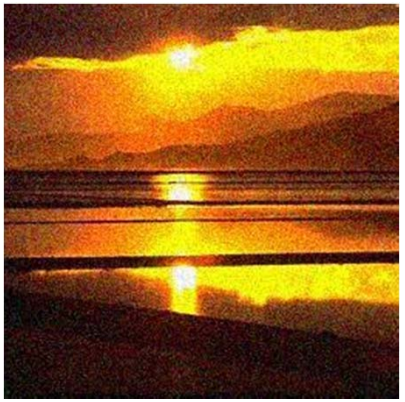
Segmentation

Description

Object Recognition

High-level processing

Scene Classification



Introduction

Low-level processing

Sensing

Preprocessing

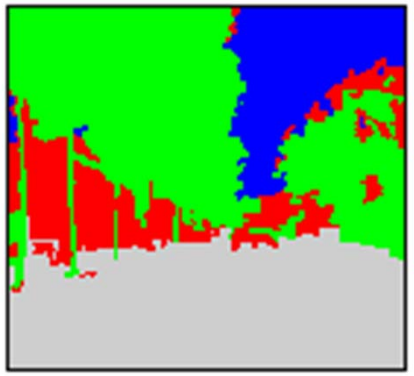
Segmentation

Description

Object Recognition

High-level processing

Scene Classification



Introduction

Low-level processing

Sensing

Preprocessing

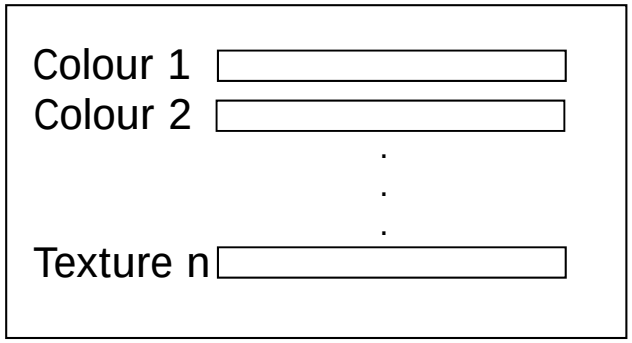
Segmentation

Description

Object Recognition

High-level processing

Scene Classification

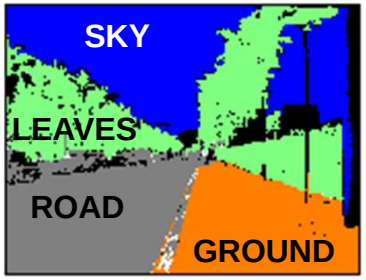


Introduction

Low-level processing

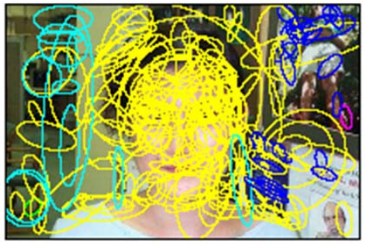
Sensing

Preprocessing



Segmentation

Description



Object Recognition

High-level processing

Scene Classification

Introduction

Low-level processing

Sensing

Preprocessing

Segmentation

Description

Object Recognition

Scene Classification

High-level processing

This scene contains: sea and sky → Coast scene

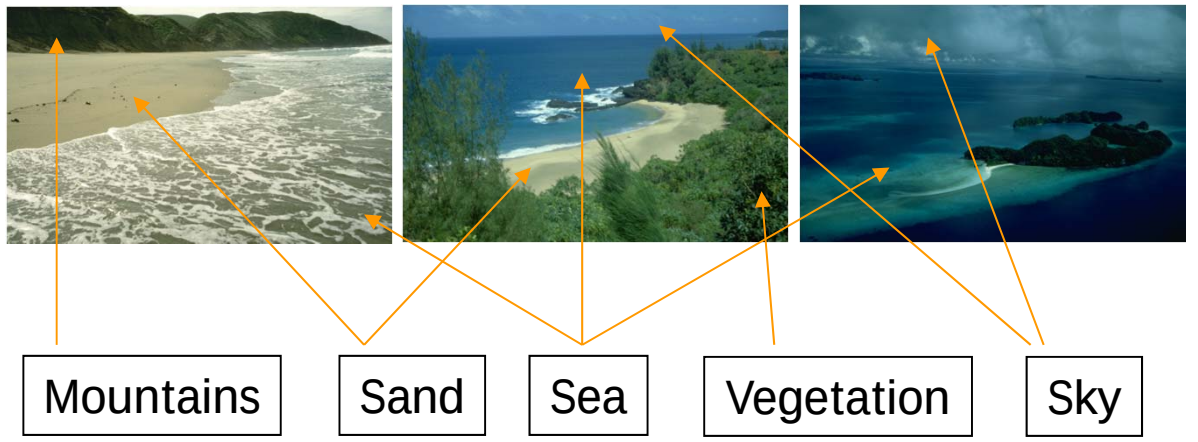
This scene contains: sea, sky, sand and mountains → Coast scene

Goals

GOAL Scene classification using *topic* distribution
Discover *topics* in images.

TOPIC DISCOVERY pLSA over bag-of-words

SCENE CLASSIFICATION KNN or SVM

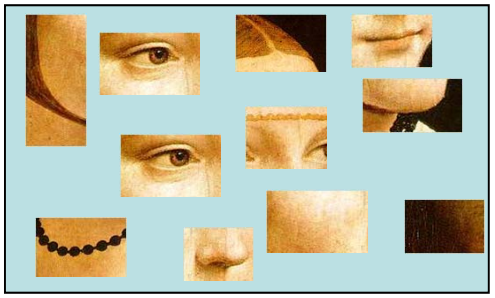


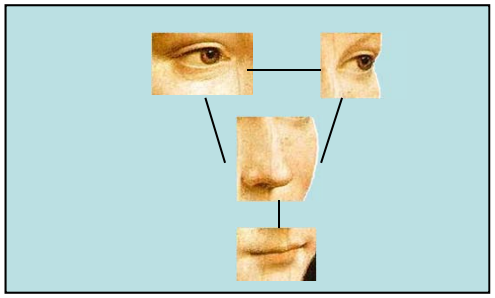
Key Points

- **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

- **Appearance** only or **location** and appearance



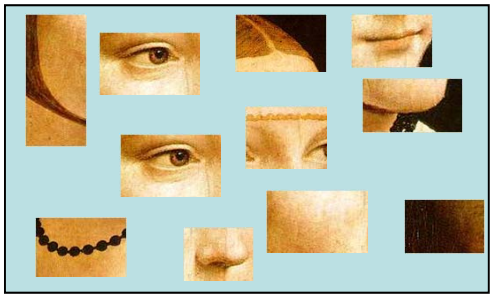


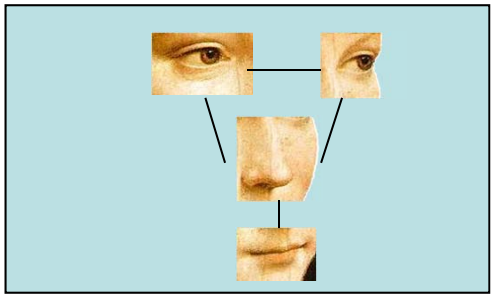
- **Invariances**

View point
Illumination
Occlusion
Scale
Intra-class variation
etc.

Representation

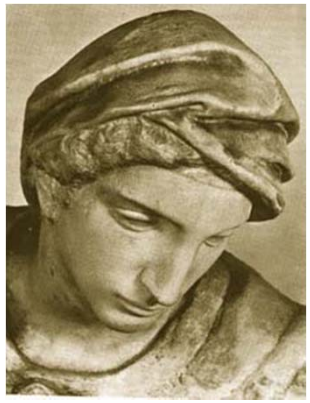
- **Appearance** only or **location** and appearance





- **Invariances**

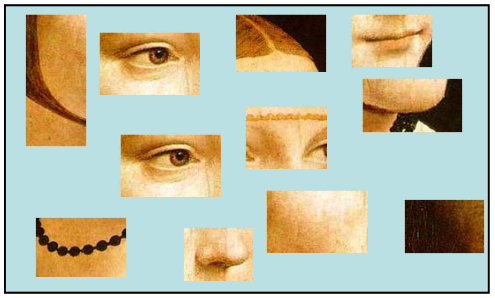
View point
Illumination
Occlusion
Scale
Intra-class variation
etc.

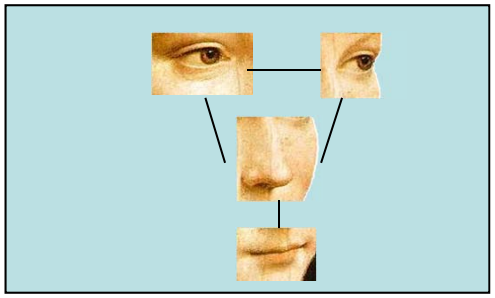


Michelangelo 1475-1564

Representation

- **Appearance** only or **location** and appearance



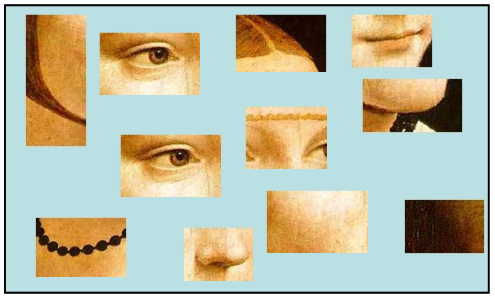


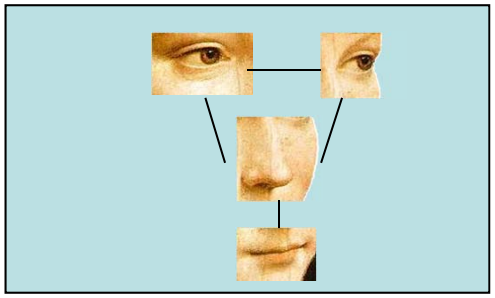
- **Invariances**

View point
Illumination
Occlusion
Scale
Intra-class variation
etc.

Representation

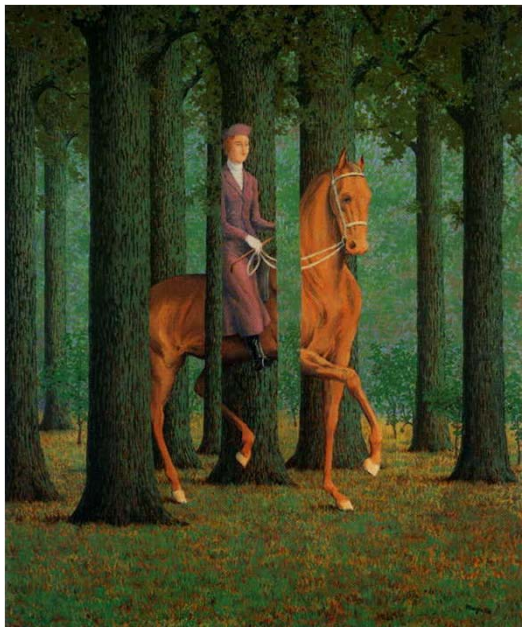
- **Appearance** only or **location** and appearance





- **Invariances**

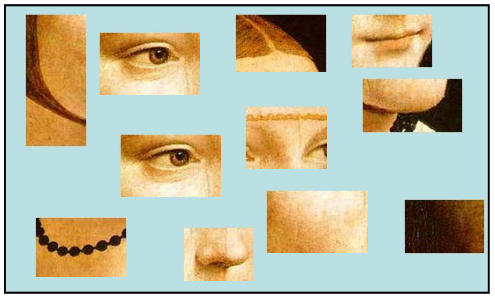
View point
Illumination
Occlusion
Scale
Intra-class variation
etc.

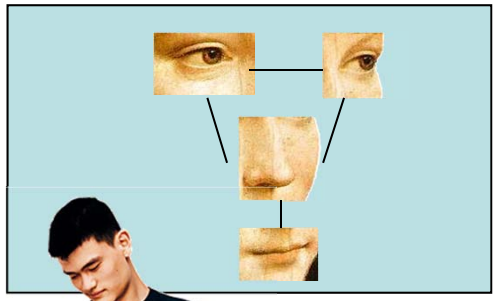


Magritte, 1957

Representation

- **Appearance** only or **location** and appearance



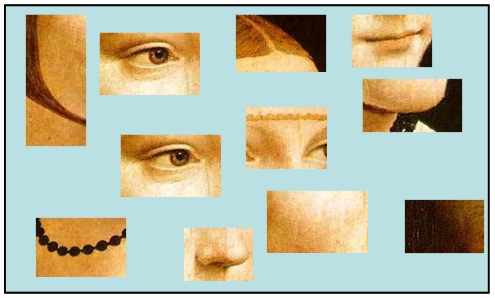


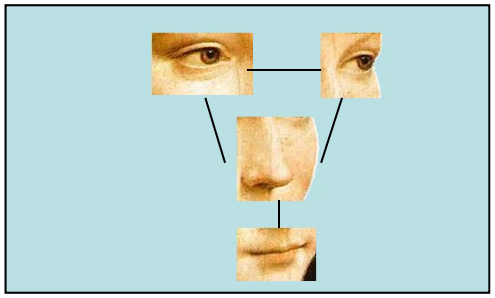
- **Invariances**

View point
Illumination
Occlusion
Scale
Intra-class variation
etc.

Representation

- **Appearance** only or **location** and appearance





- **Invariances**

View point
Illumination
Occlusion
Scale
Intra-class variation
etc.

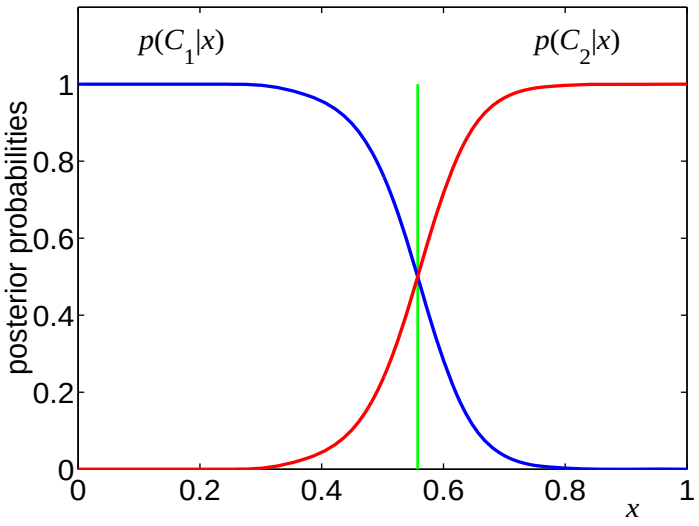
Learning: statistical viewpoint

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

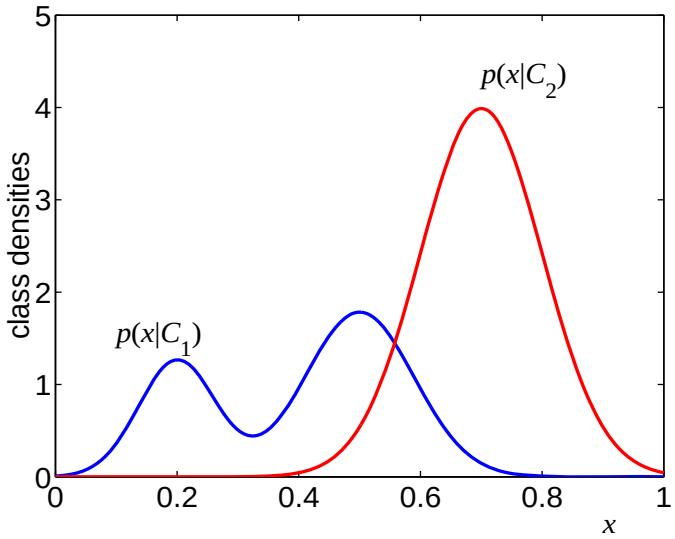
- **Discriminative methods** model posterior
- **Generative methods** model likelihood and prior

Learning: statistical viewpoint

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



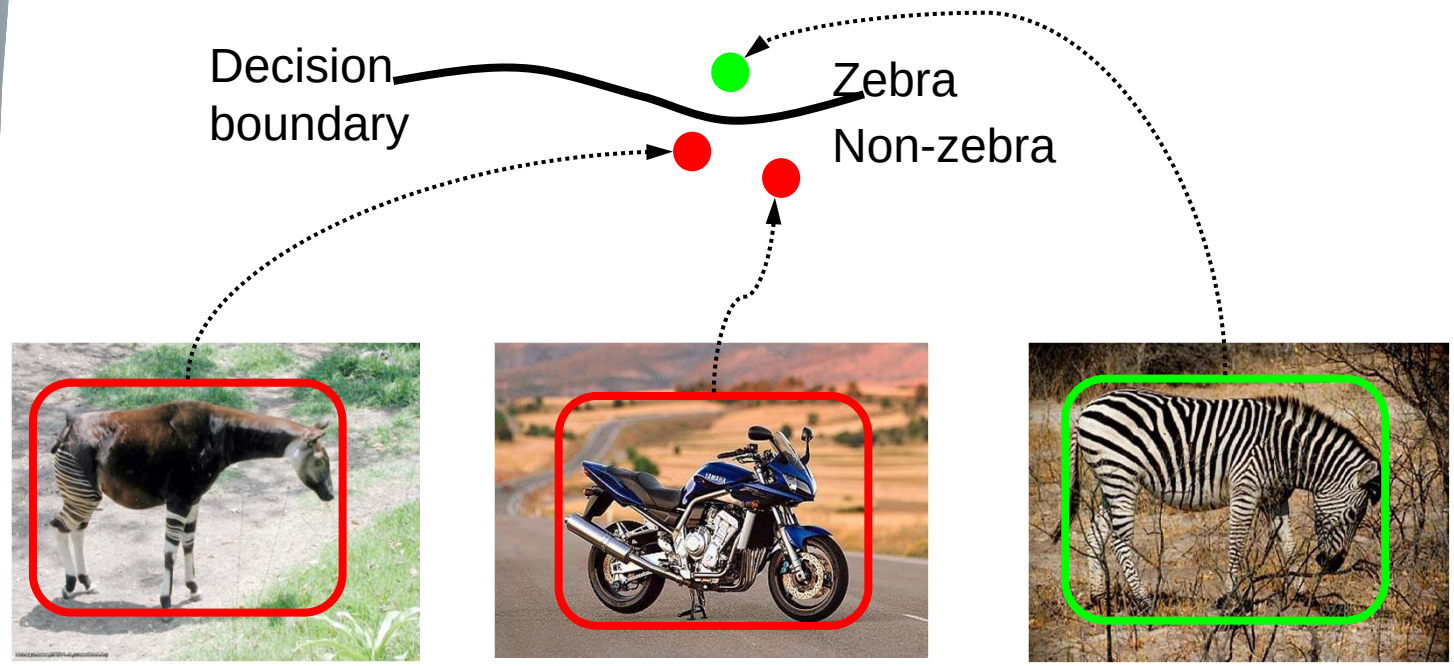
Discriminative methods



Generative methods

Learning: Discriminative

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

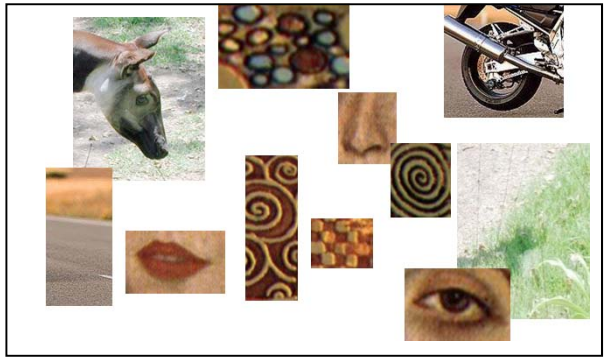


Learning: Generative

Model

$p(image \mid zebra)$

$p(image \mid no\ zebra)$



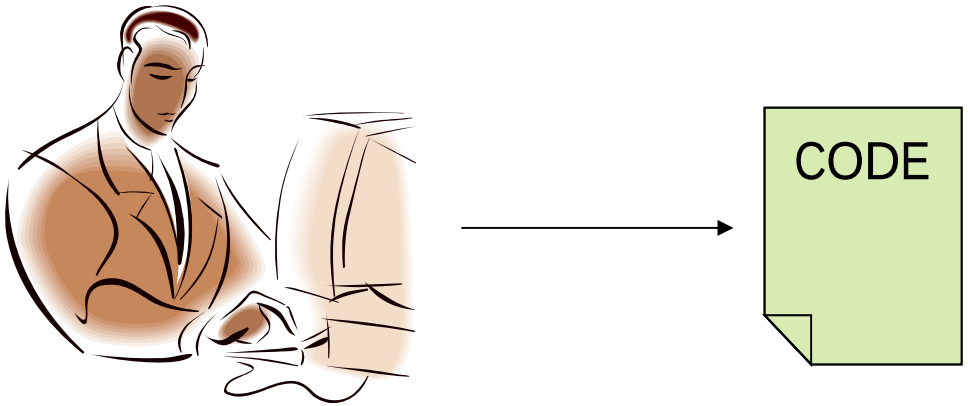
$p(image \mid zebra)$	$p(image \mid no\ zebra)$
Low	Middle
High	Middle→Low

Learning

Hand and code encapsulated data

Supervised Learning

Unsupervised Learning

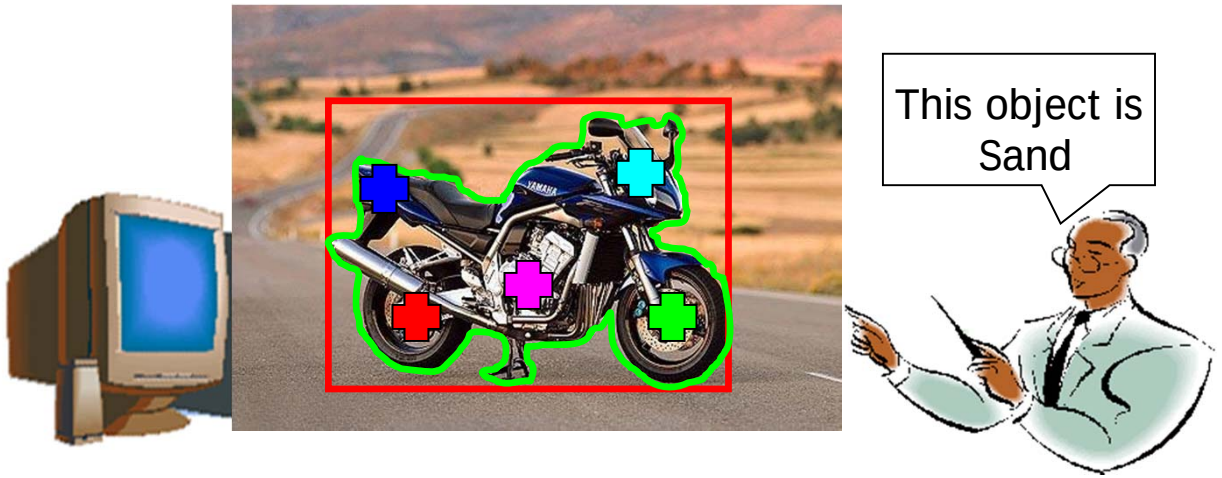


Learning

Hand and code encapsulated data

Supervised Learning

Unsupervised Learning

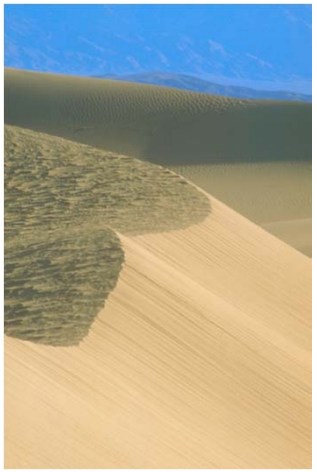


Learning

Hand and code encapsulated data

Supervised Learning

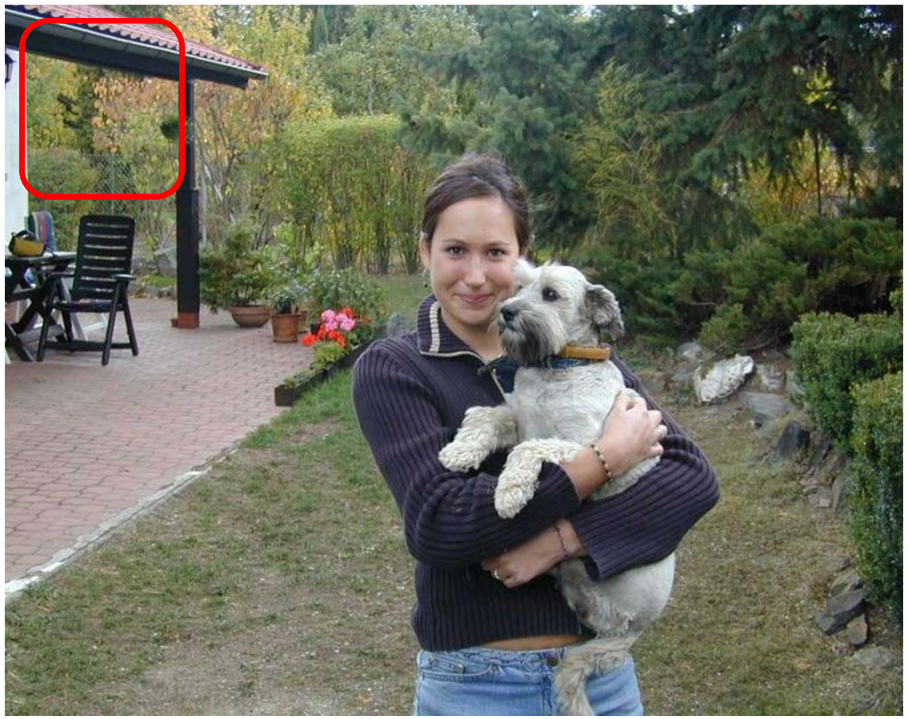
Unsupervised Learning



Recognition

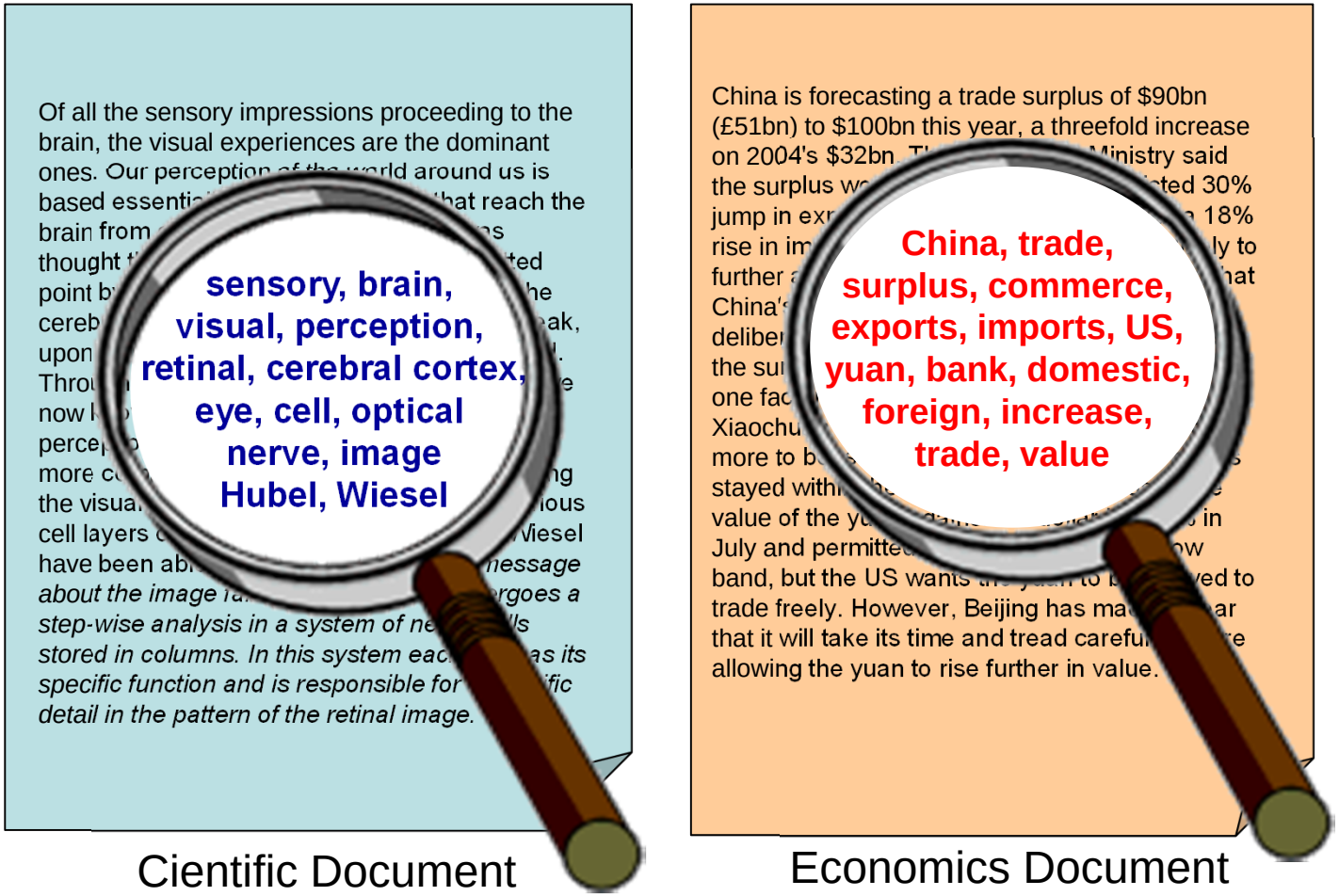
– Scale / orientation range to search over

– Speed



Bag-of-words

Origin: document retrieval



Cientific Document

Economics Document

From documents to images

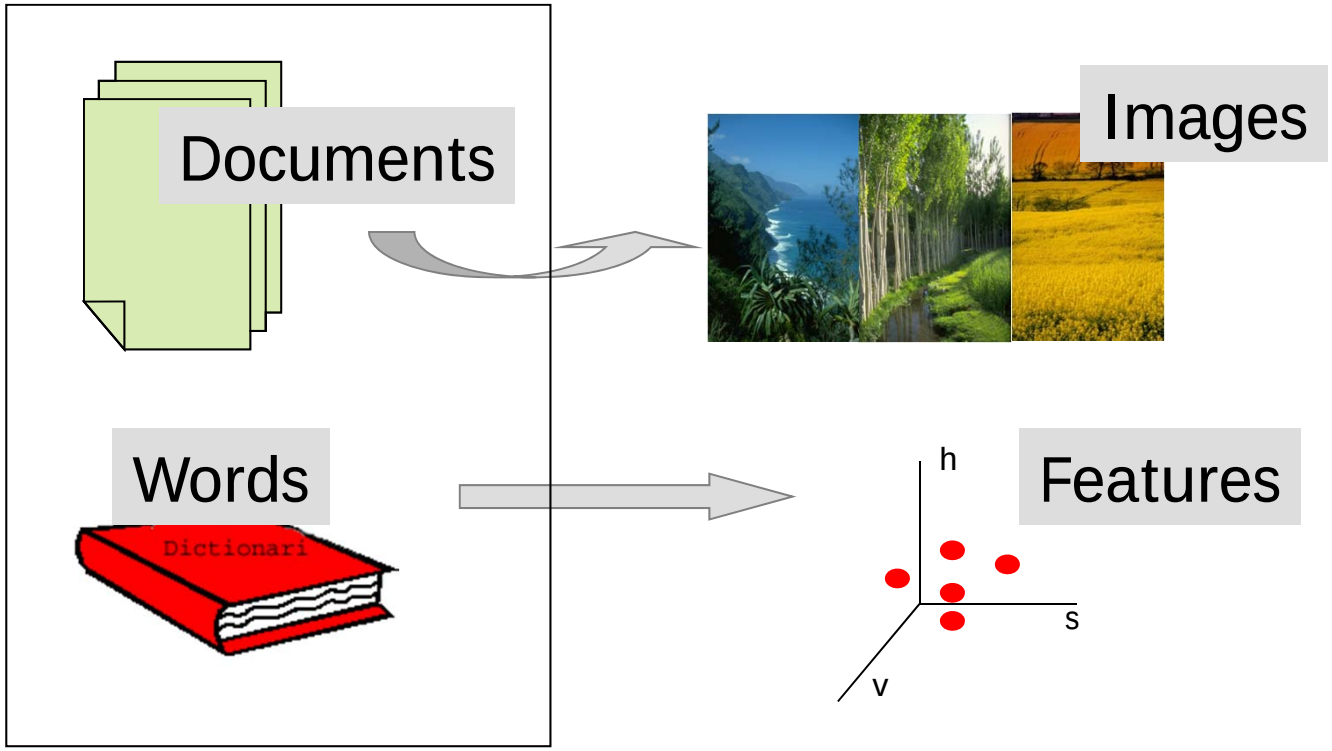
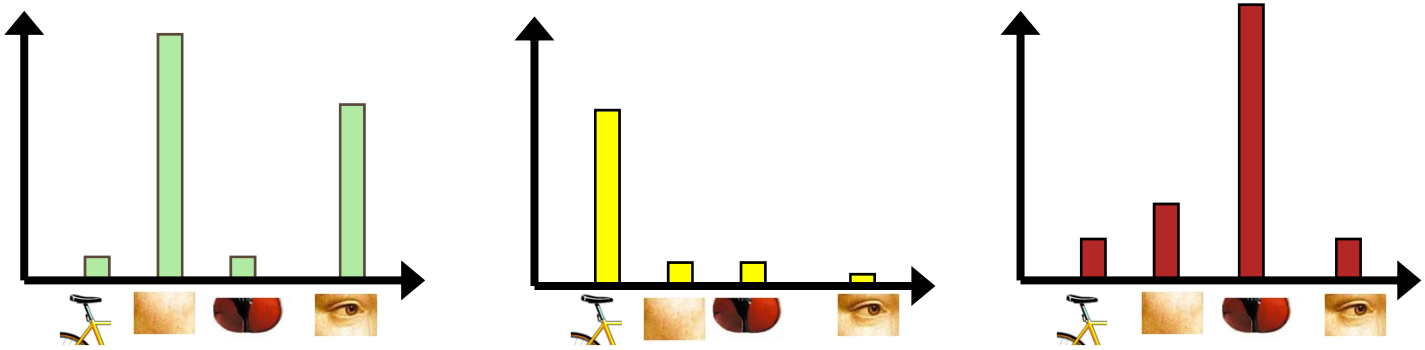
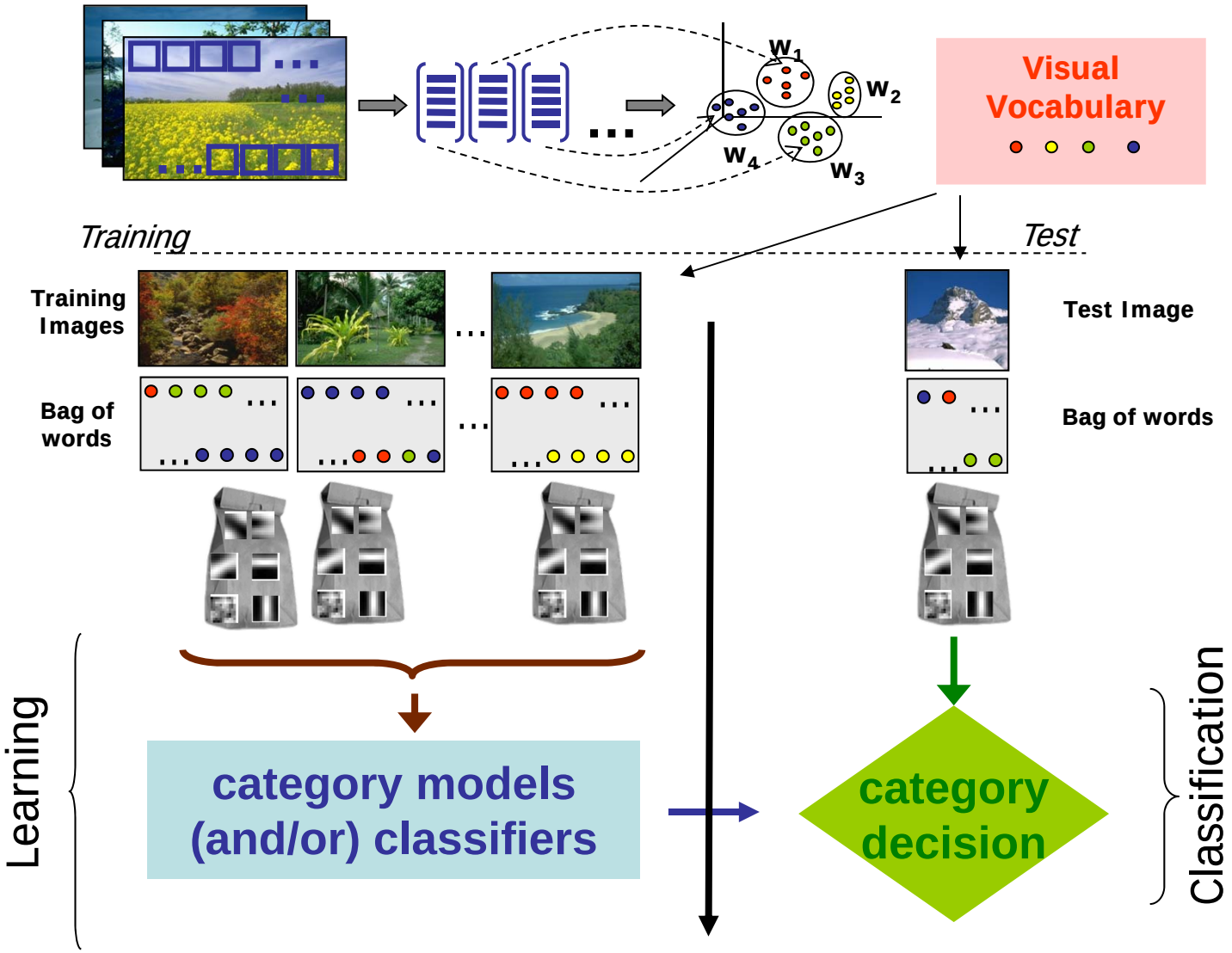


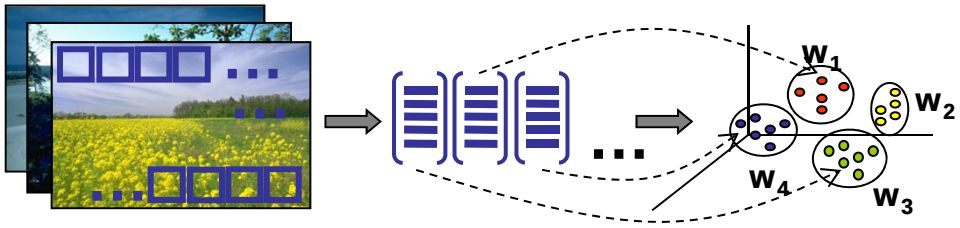
Image representation



Overview

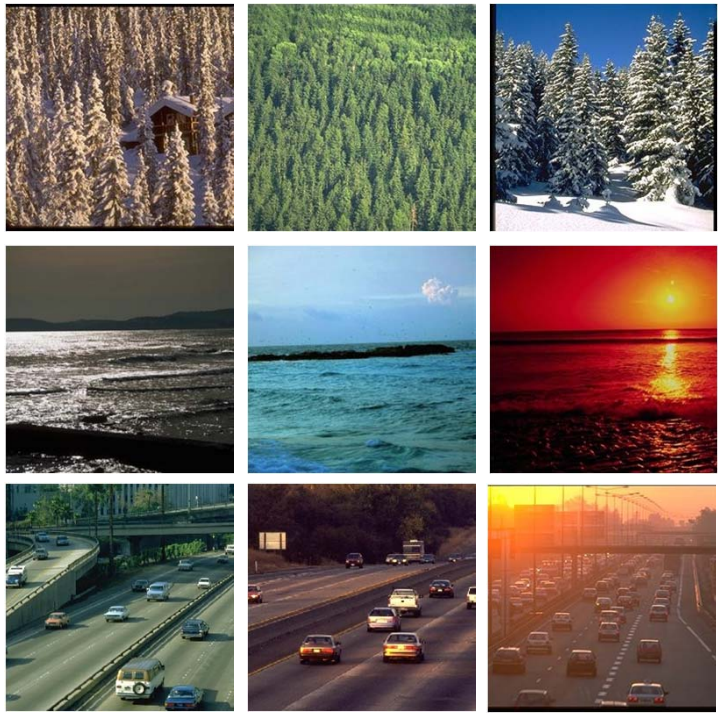


Visual Vocabulary



Visual Vocabulary

Feature detection & representation



Middle 90' until now:

-Textons: good performance in texture classification [Varma'03]

-Features invariant to affine geometric and photometric Transformantions [ICCV'01]

-Scale Invariant Feature Transform (SIFT) [IJCV'04]

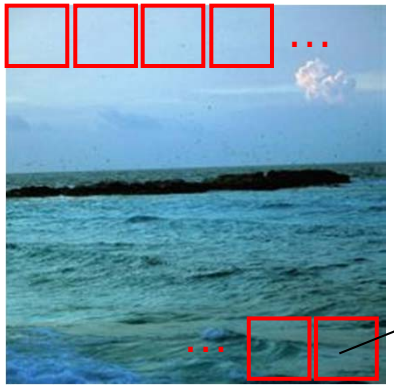
Feature detection & representation

SPARSE PATCHES – Affine covariant regions (harris & hessian)

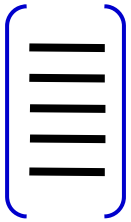
Feature detection & representation

DENSE PATCHES

Textons



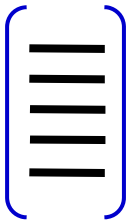
Parameters: N – size of patch
M – distance between patches



Row reorder gray values
And form a vector of
Size N^2

SIFT

Parameters: r – radi of patch
M – distance between patches



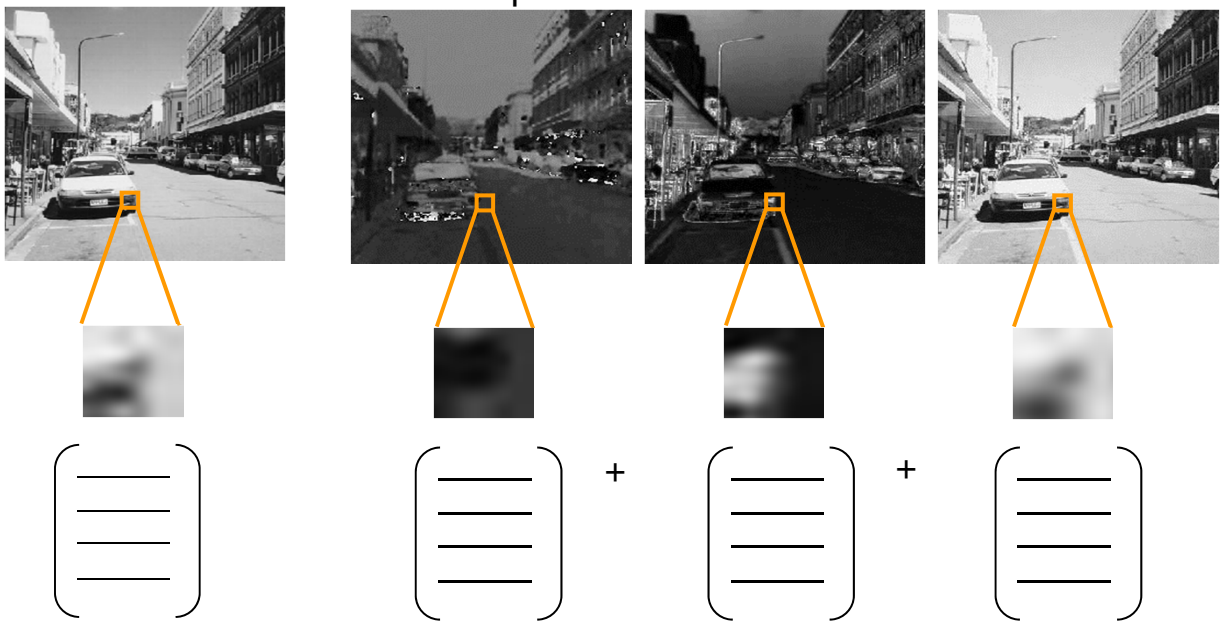
128- SIFT descriptor

Feature detection & representation

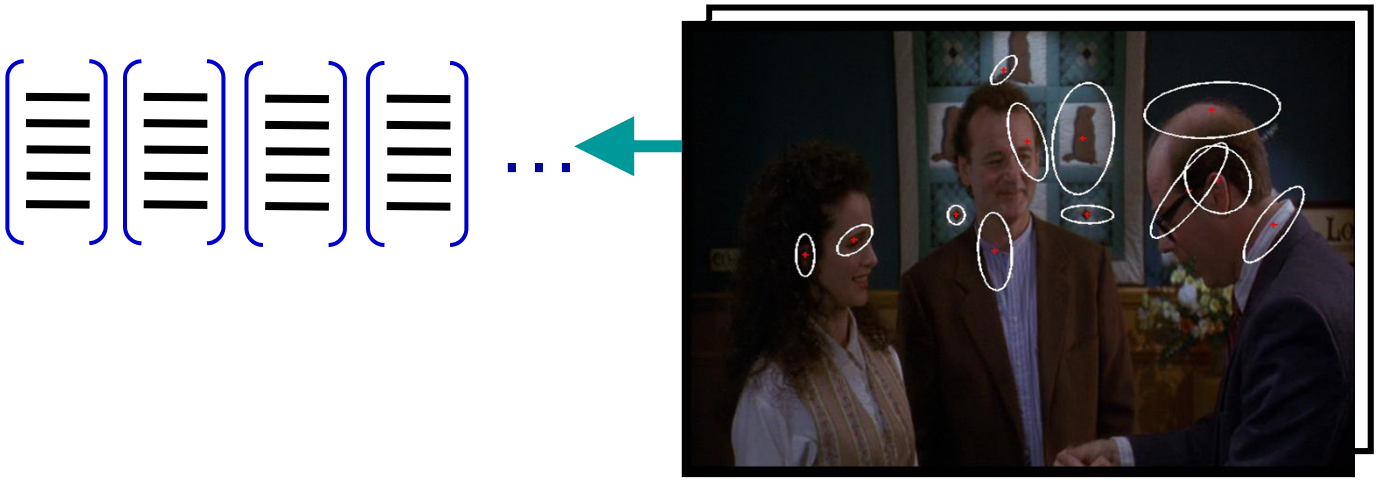
USING COLOUR

Gray

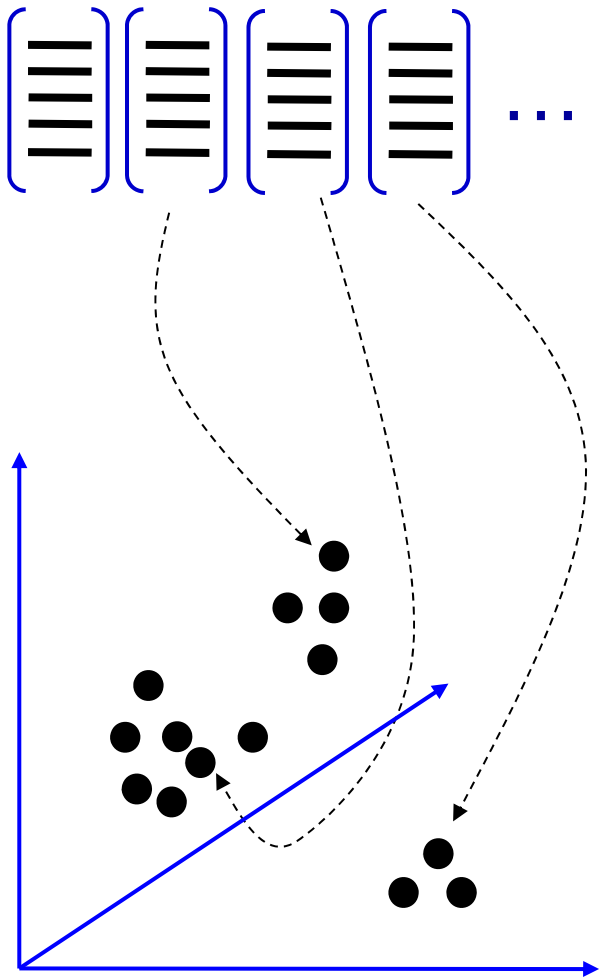
Colour (HSV) → descriptors are computed for each component



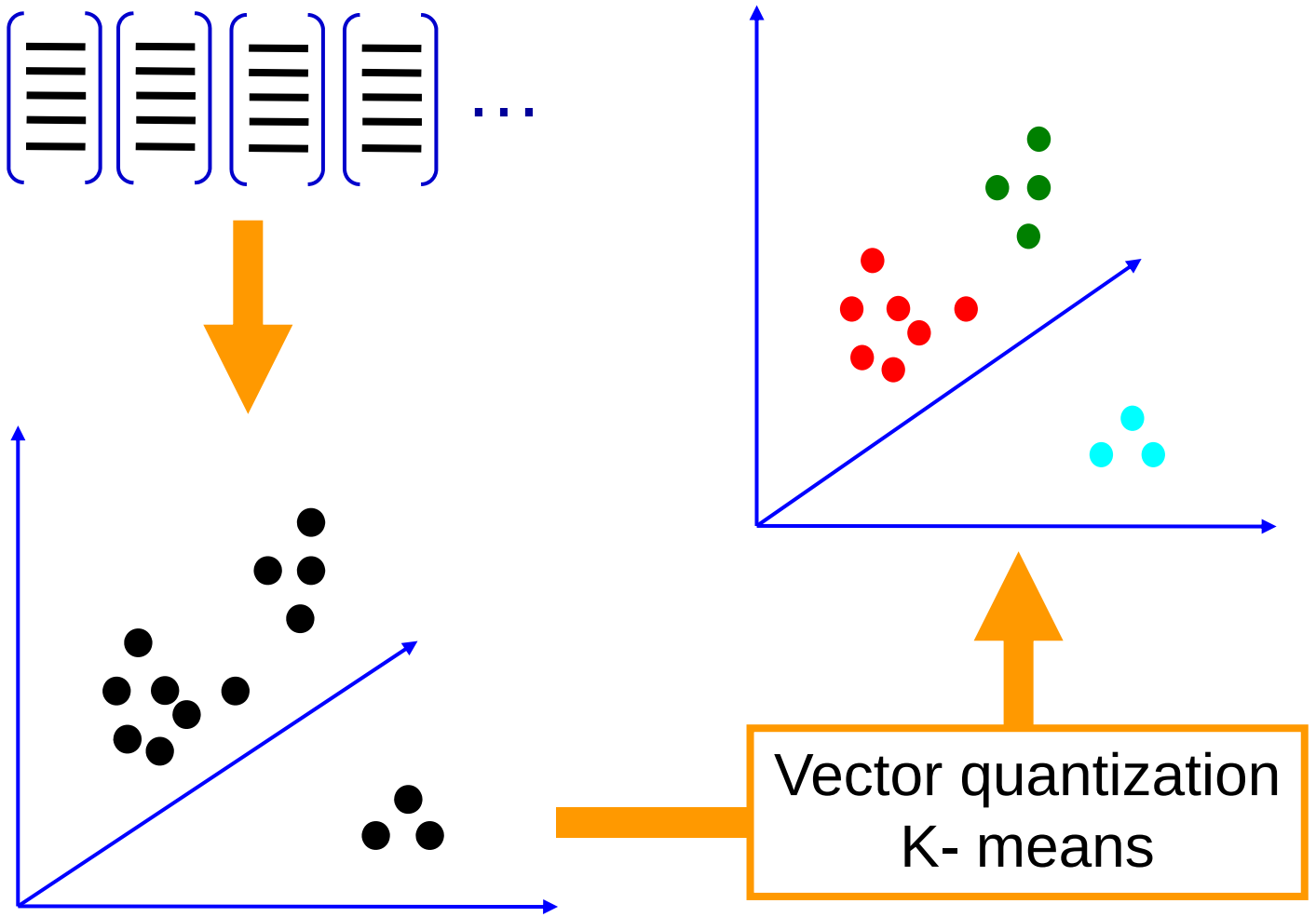
Visual vocabulary formation



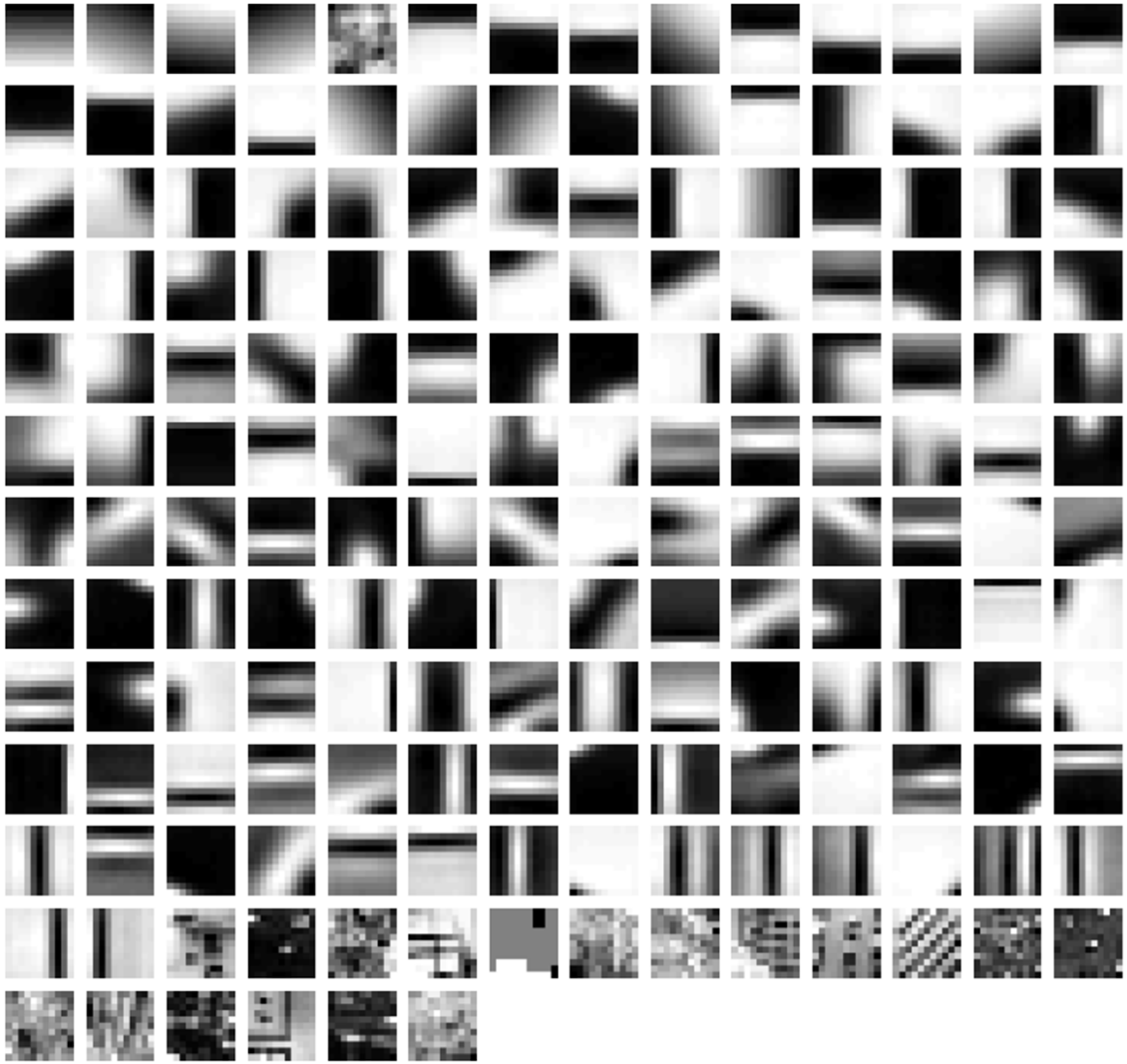
Visual vocabulary formation



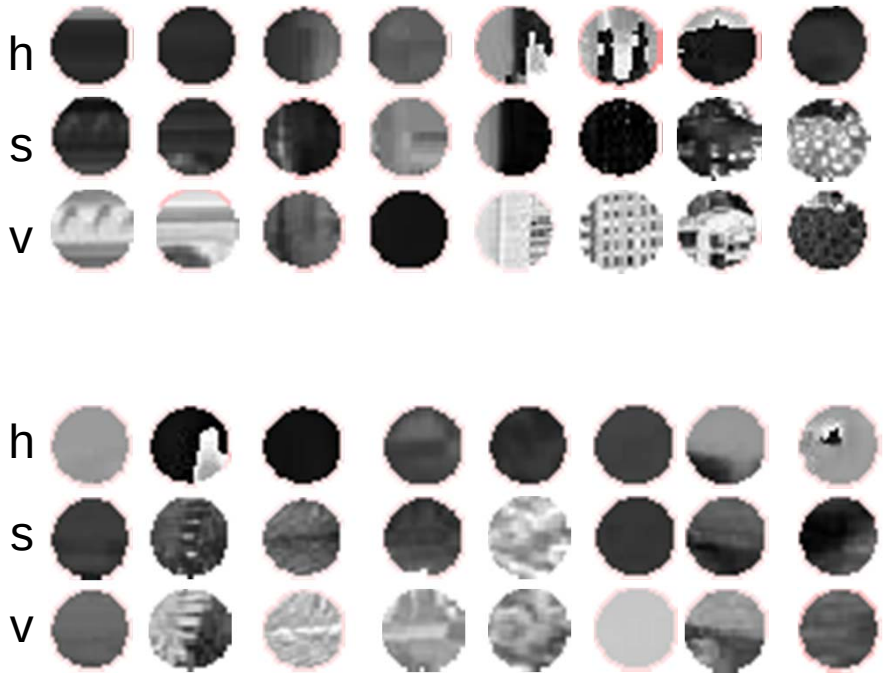
Visual vocabulary formation



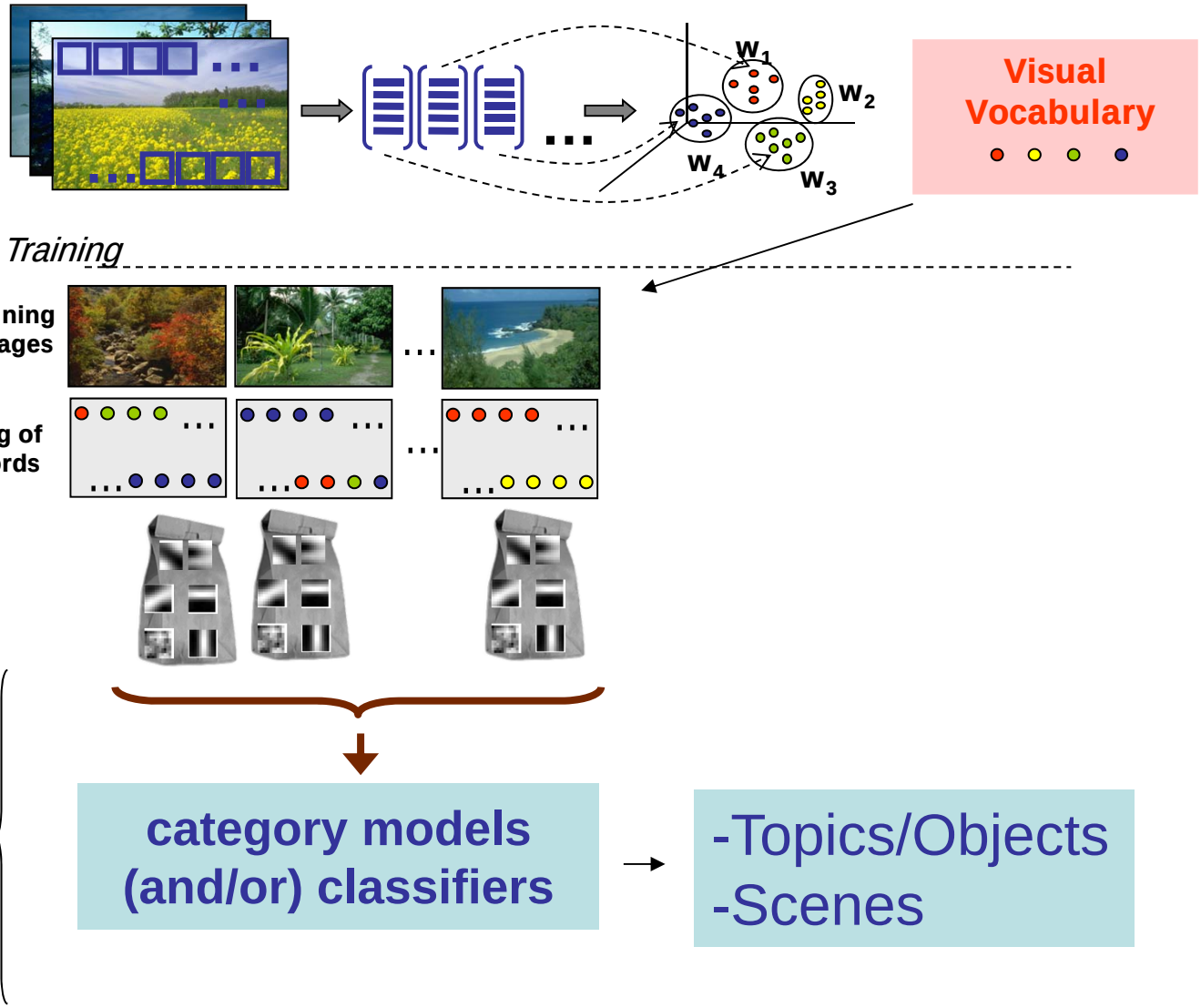
Examples of visual words



Examples of visual words

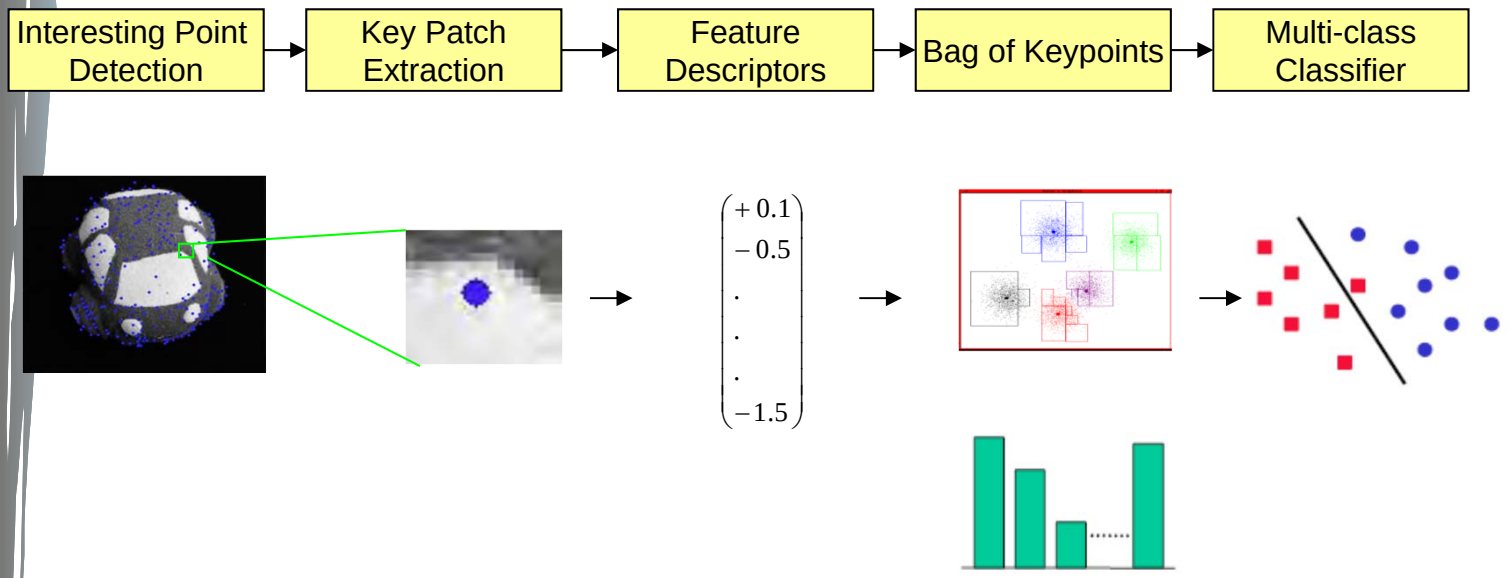


Learning



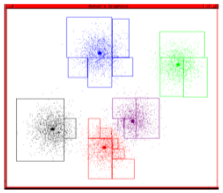
Learning

SPARSE BAG OF WORDS APPROACH



Learning

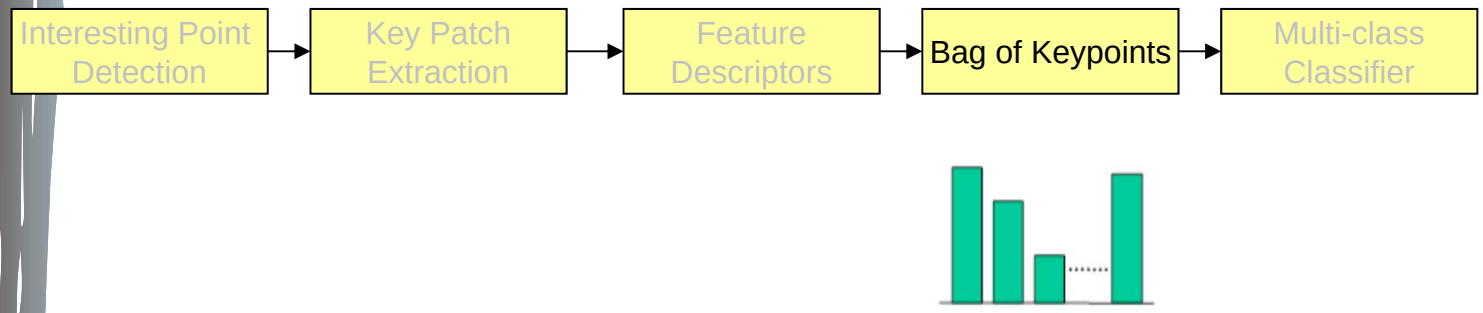
SPARSE BAG OF WORDS APPROACH



- Construction of a vocabulary
  - Kmeans clustering → find “centroids”
    (on all the descriptors we find from a set of training images)
  - Define a “vocabulary” as a set of “centroids”, where every centroid represents a “word”

Learning

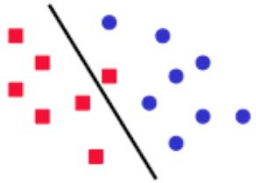
SPARSE BAG OF WORDS APPROACH



- Histogram
  - Counts the number of occurrences of different *visual words* in each image

Learning

SPARSE BAG OF WORDS APPROACH



- The classification using different classifiers
  - NN, K-NN
  - Support Vector Machine (SVM)
  - Adaboost
- What happens with a dense approach?

Generative Learning

1. Naïve Bayes classifier
   - Csurka et al. 2004

2. Hierarchical Bayesian text models
   - pLSA Hoffman 2001
   - LDA Blei et al. 2004

Weakness of the model

- No context information about the **scenes**
- No rigorous geometric information of the **object** components
- It's intuitive to most of us that **objects** are made of parts – no such information

Properties of an 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

Bibliography

- [PAMI 2008] A. Bosch, A. Zisserman, X. Muñoz. *Scene classification using a hybrid generative / discriminative approach*. IEEE Transactions on Pattern Analysis and Machine Intelligence

- [Hofman 2001] Probabilistic Latent Semantic Analysis (pLSA)

- [Tutorial 2007] Some of the slides are extracted from the following link, where you can also find available code:

http://people.csail.mit.edu/torralba/shortCourseRLOC/