

# Object & scene classification for large image collections



Anna Bosch Rué<sup>1</sup>

Andrew Zisserman<sup>2</sup>

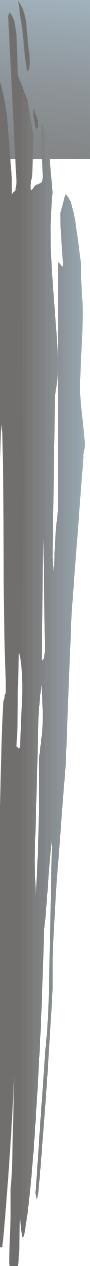
Xavier Muñoz<sup>1</sup>

<sup>1</sup>Arquitectura i Tecnologia de Computadors  
Universitat de Girona



<sup>2</sup>Visual Geometry Group  
University of Oxford

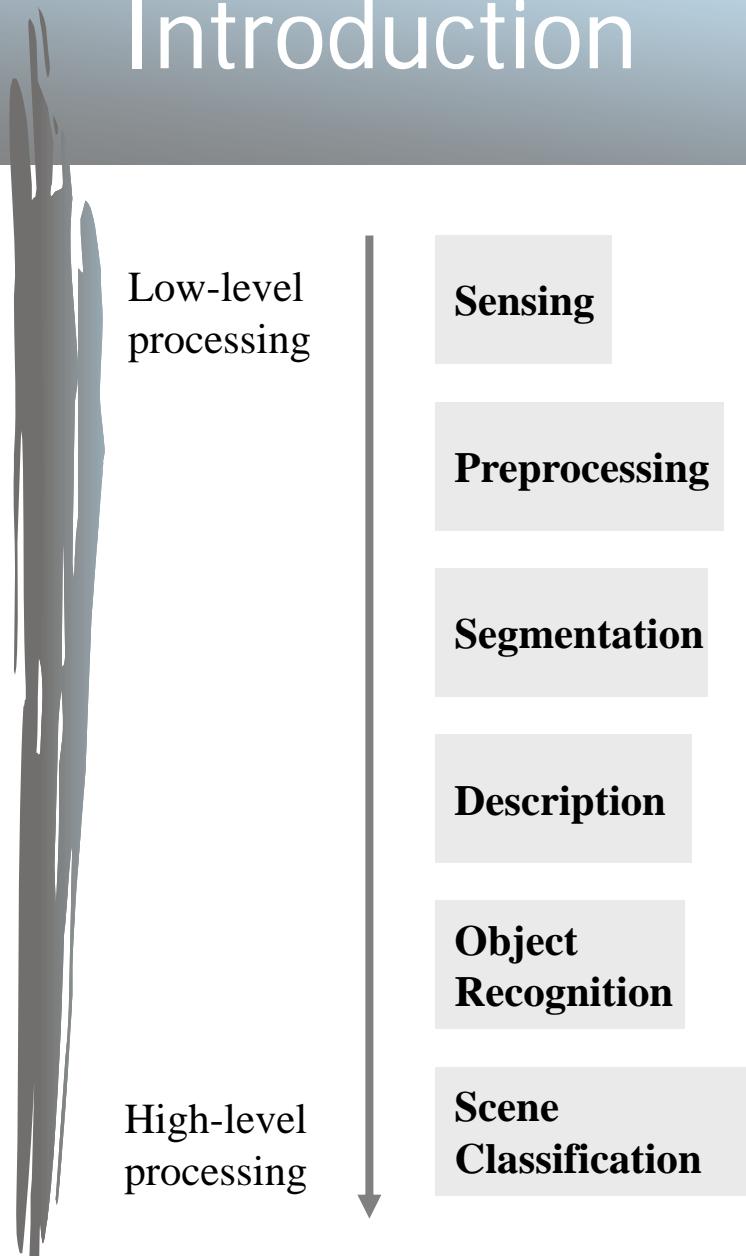




# INTRODUCTION

# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions



# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

Low-level processing

High-level processing

**Sensing**



**Preprocessing**

**Segmentation**

**Description**

**Object  
Recognition**

**Scene  
Classification**



Lens  
focus



shutter



CCD

# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

Low-level processing

Sensing



Preprocessing

Segmentation



Description

Object  
Recognition

Scene  
Classification

High-level processing

# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

Low-level processing

High-level processing

Sensing

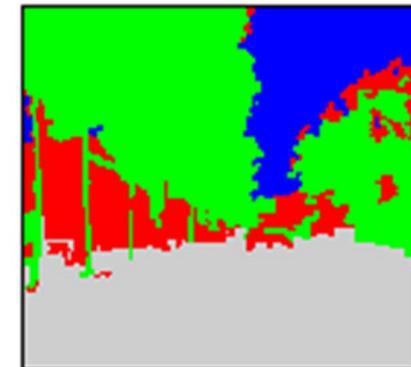
Preprocessing

Segmentation

Description

Object  
Recognition

Scene  
Classification



# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

Low-level processing

Sensing

Preprocessing

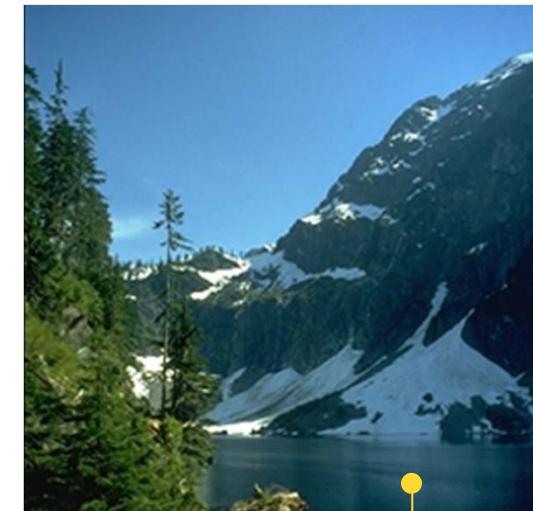
Segmentation

Description

Object  
Recognition

Scene  
Classification

High-level processing



Colour 1	<input type="text"/>
Colour 2	<input type="text"/>
.	.
Texture n	<input type="text"/>

# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

Low-level processing

Sensing

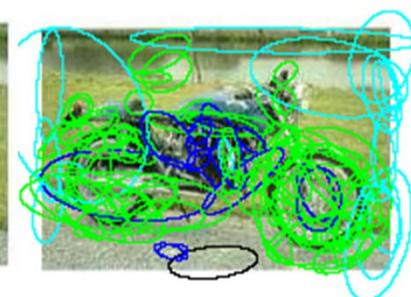


Preprocessing



Segmentation

Description

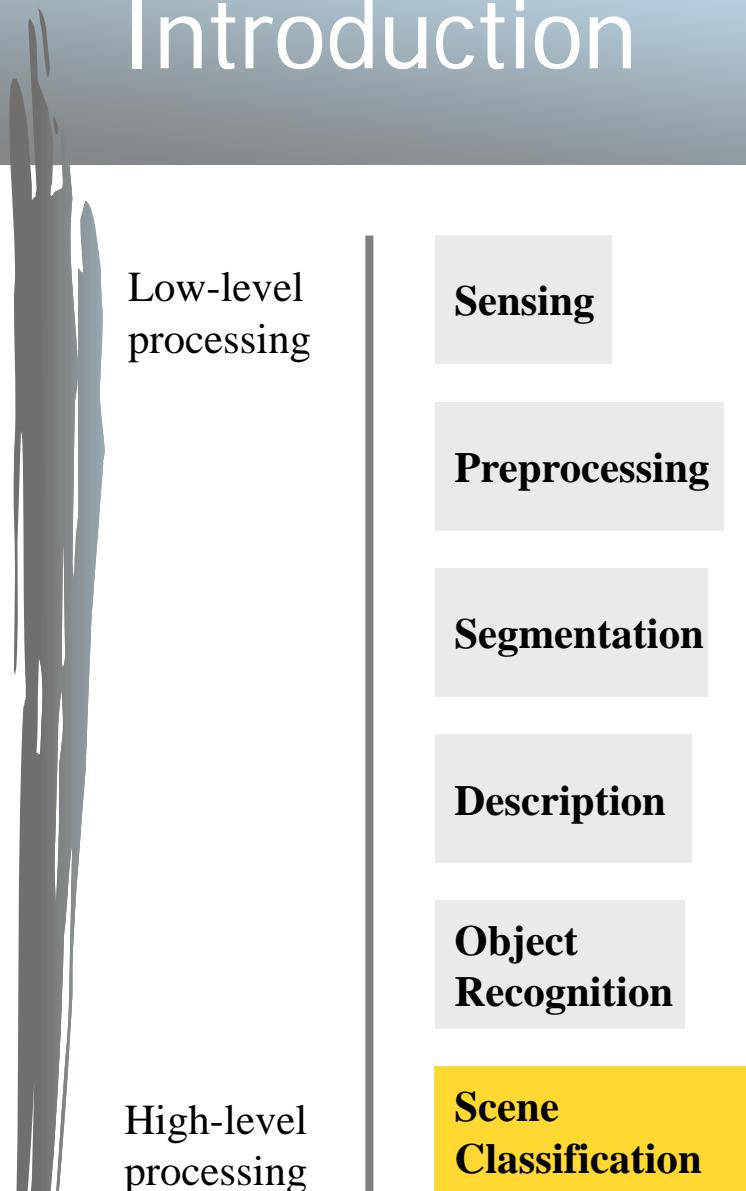


Object  
Recognition

Scene  
Classification

# Introduction

- 1. Introduction
  - 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions



This scene contains: sea and sky → Coast scene



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

# Goals

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

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

**TOPIC DISCOVERY** pLSA over bag-of-words

**SCENE CLASSIFICATION** KNN or SVM



Mountains

Sand

Sea

Vegetation

Sky



COAST SCENE

# Key Points

## 1. Introduction

- 1.1 Computer Vision
- 1.2 Scene Classification

## 1.3 Key points

- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

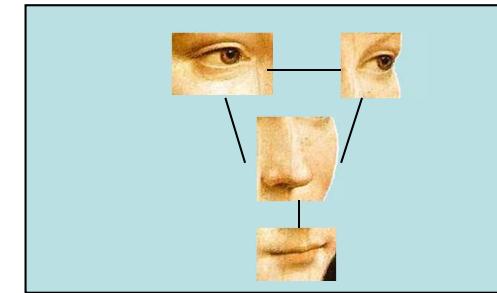
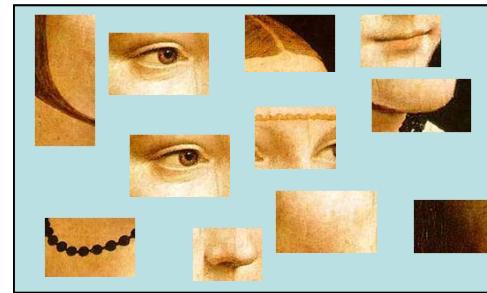
- **Representation**
  - How to represent an object category
- **Learning**
  - How to form the classifier, given training data
- **Recognition**
  - How the classifier is to be used on novel data

# Representation

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

- Appearance only or location and appearance



- Invariances

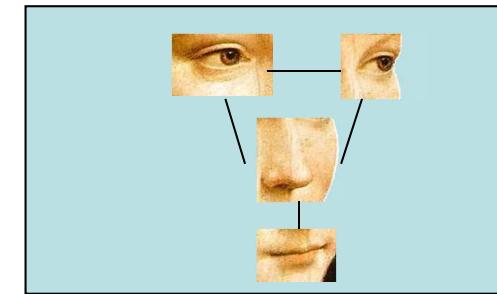
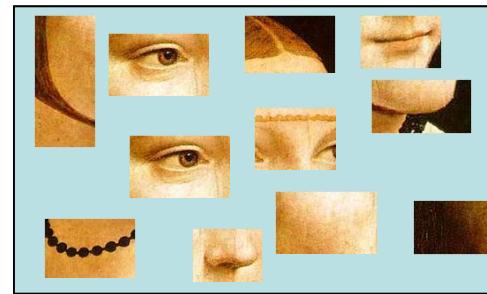
- View point
- Illumination
- Occlusion
- Scale
- Intra-class variation
- etc.

# Representation

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

- Appearance only or location and appearance



- Invariances

**View point**  
Illumination  
Occlusion  
Scale  
Intra-class variation  
etc.

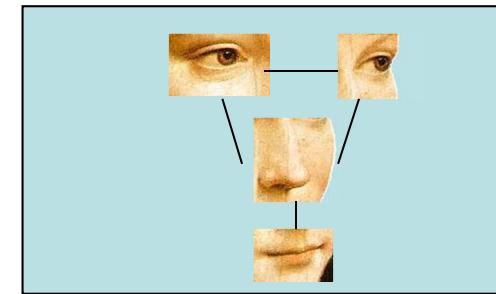
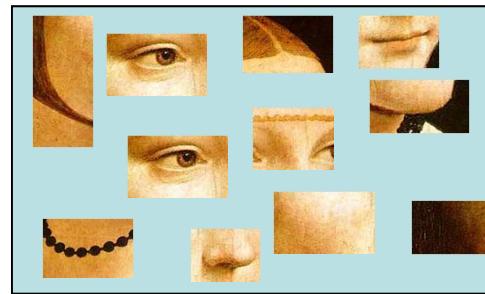


# Representation

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

- Appearance only or **location** and appearance



- Invariances

View point  
**Illumination**  
Occlusion  
Scale  
Intra-class  
variation  
etc.

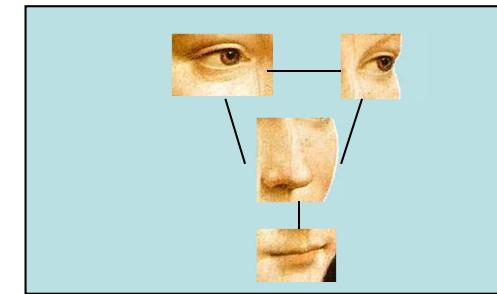
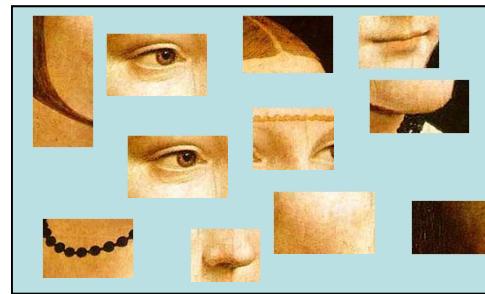


# Representation

## 1. Introduction

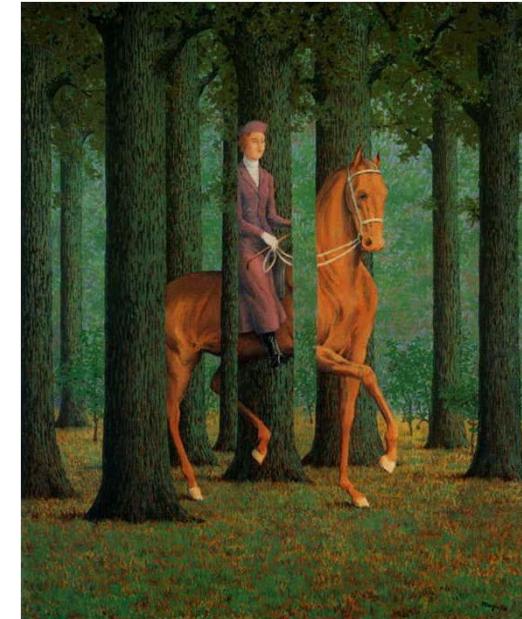
- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

- Appearance only or **location** and appearance



- Invariances

View point  
Illumination  
**Occlusion**  
Scale  
Intra-class variation  
etc.



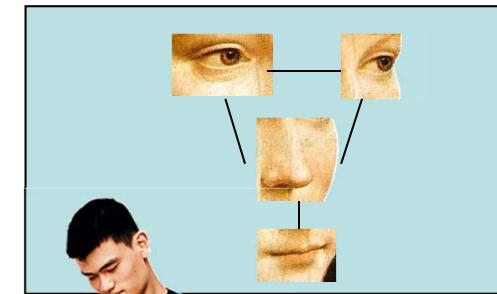
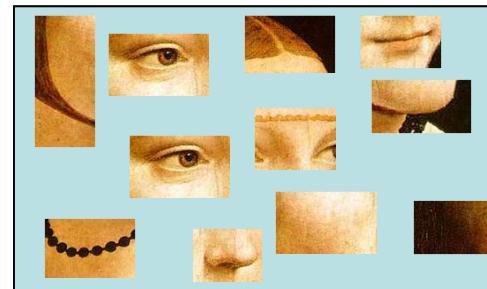
Magritte, 1957

# Representation

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

- Appearance only or **location** and appearance



- Invariances

View point  
Illumination  
Occlusion  
**Scale**  
Intra-class  
variation  
etc.

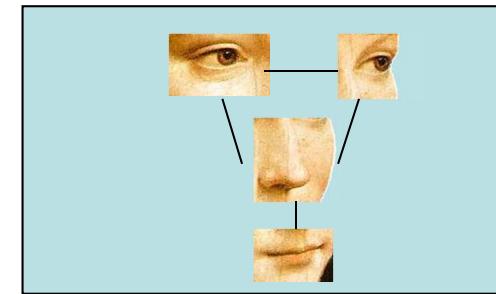
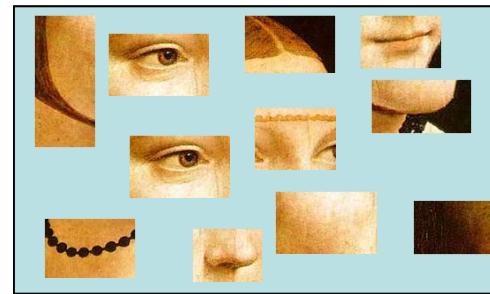


# Representation

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

- Appearance only or **location** and appearance



- Invariances

View point  
Illumination  
Occlusion  
Scale  
**Intra-class variation**  
etc.



# Learning: statistical viewpoint

## 1. Introduction

1.1 Computer Vision  
1.2 Scene Classification

1.3 Key points

2. Bow  
3. Spatial Information  
4. Geometric Information  
5. Merging Features  
6. Conclusions

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

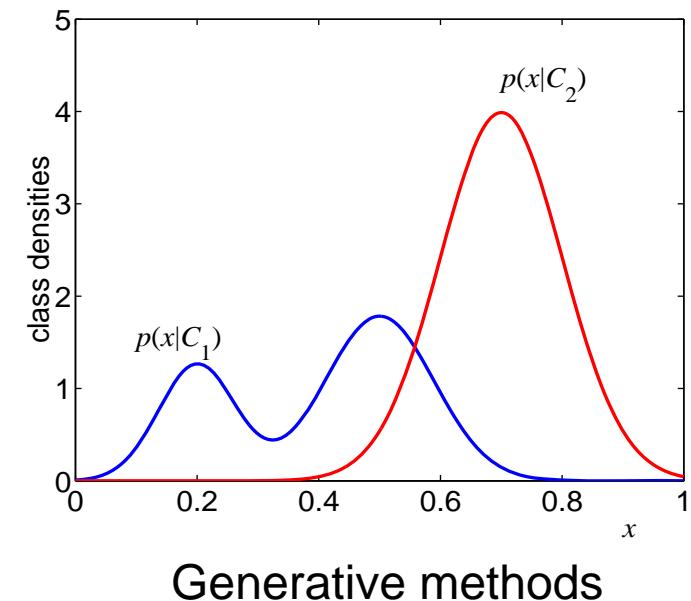
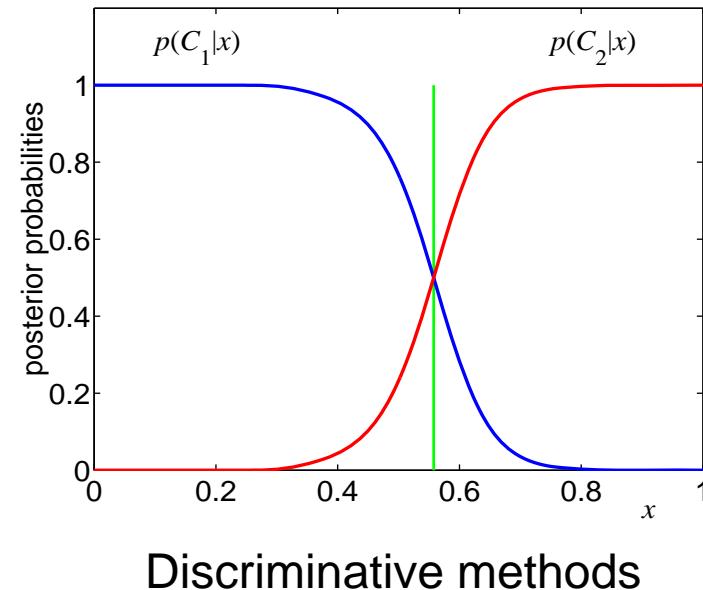
- **Discriminative methods** model posterior
- **Generative methods** model likelihood and prior

# Learning: statistical viewpoint

## 1. Introduction

- 1.1 Computer Vision
- 1.2 Scene Classification
- 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

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



# Learning: Discriminative

## Direct modeling of

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

### 1. Introduction

1.1 Computer Vision

1.2 Scene Classification

1.3 Key points

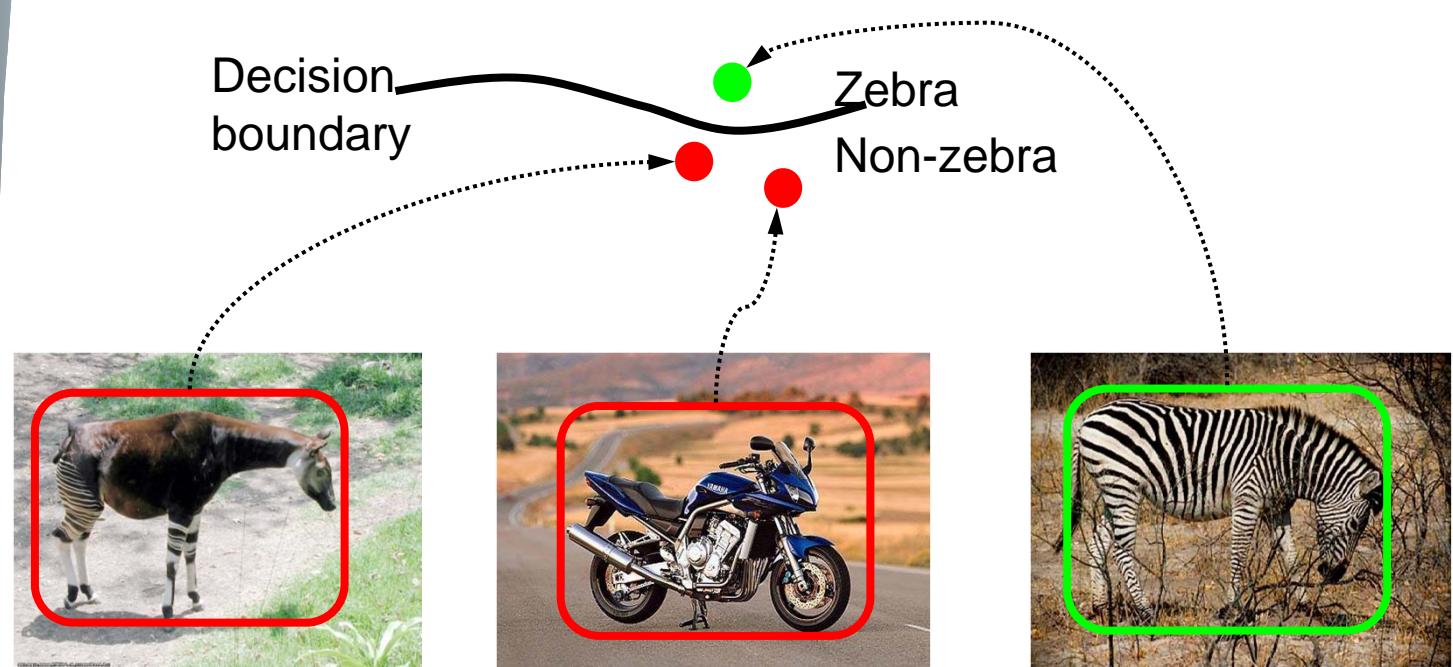
### 2. Bow

3. Spatial Information

4. Geometric Information

5. Merging Features

6. Conclusions

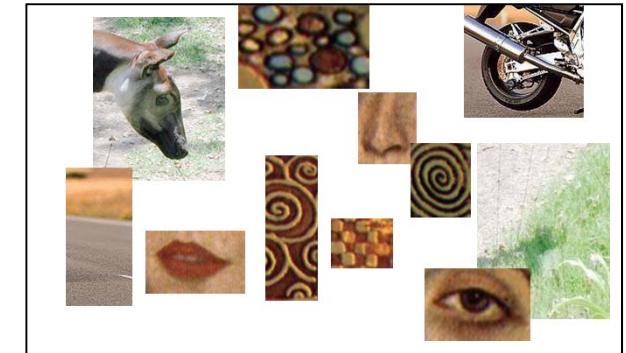
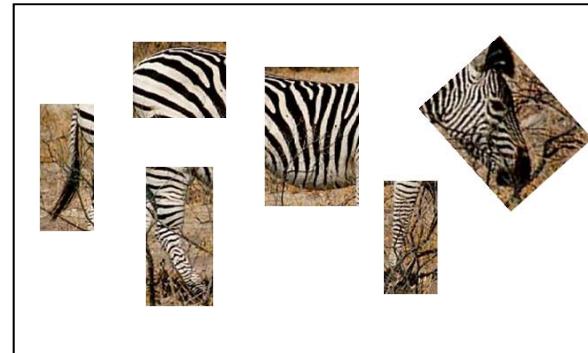


# Learning: Generative Model

## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

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



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



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

Low

Middle

High

Middle → Low

# Learning

## 1. Introduction

1.1 Computer Vision

1.2 Scene Classification

1.3 Key points

2. Bow

3. Spatial Information

4. Geometric Information

5. Merging Features

6. Conclusions

Hand and code encapsulated data

Supervised Learning

Unsupervised Learning



CODE

# Learning

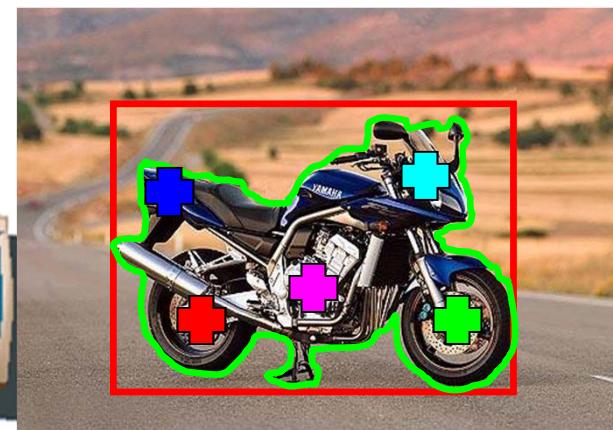
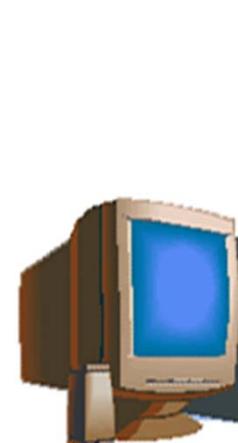
## 1. Introduction

- 1.1 Computer Vision
  - 1.2 Scene Classification
  - 1.3 Key points
2. Bow
3. Spatial Information
4. Geometric Information
5. Merging Features
6. Conclusions

Hand and code encapsulated data

Supervised Learning

Unsupervised Learning



# Learning

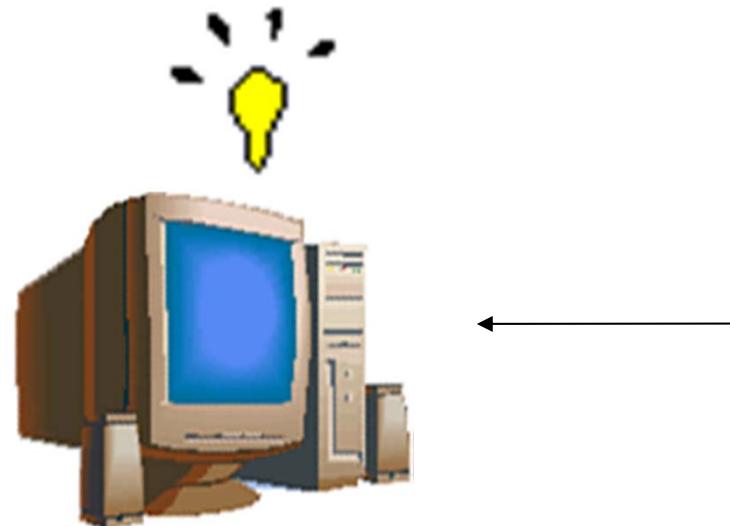
## 1. Introduction

- 1.1 Computer Vision
- 1.2 Scene Classification
- 1.3 Key points
- 2. Bow
- 3. Spatial Information
- 4. Geometric Information
- 5. Merging Features
- 6. Conclusions

Hand and code encapsulated data

Supervised Learning

Unsupervised Learning



# Recognition

## 1. Introduction

1.1 Computer Vision

1.2 Scene Classification

1.3 Key points

## 2. Bow

3. Spatial Information

4. Geometric Information

5. Merging Features

6. Conclusions

- Scale / orientation range to search over
- Speed





# Bag-of-words

# Origin: document retrieval

1. Introduction

2. Bow

2.1 Origin

2.2 Docs to images

2.3 Representation

2.4 Overview

2.5 Vocabulary

2.6 Learning

2.7 Classification

2.8 Results

2.9 Conclusions

3. Spatial Information

4. Geometric features

5. Merging Features

4. Conclusions

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 essentially on the visual sensations that reach the brain from the eyes. The visual sensations are thought to be the result of the visual system's action. At the point by which the visual sensations reach the cerebral cortex, they have been transformed into perceptions. Through the process of perception, we now know more about the world around us. Perception is a more complex process than the visual system. The visual system consists of various cell layers of the retina. These cells have been able to extract information from the visual message about the image received. The visual system undergoes a step-wise analysis in a system of neurons. The neurons are stored in columns. In this system each neuron has its specific function and is responsible for a specific detail in the pattern of the retinal image.

Scientific Document

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

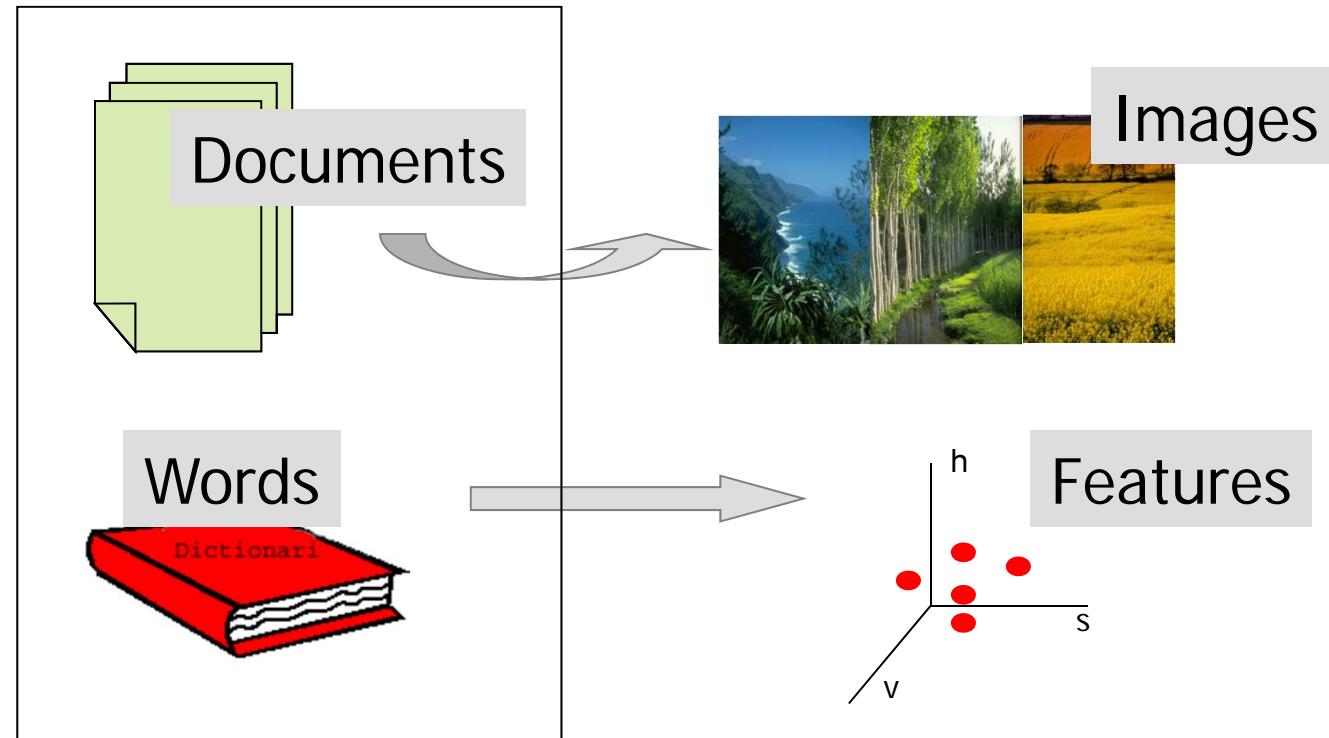
China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Ministry of Commerce said the surplus would rise because it expected 30% jump in exports and a 10% to 18% rise in imports. The ministry also said that China's deliberate policy of keeping the yuan undervalued was likely to continue. It said that the Chinese government had deliberately allowed the value of the yuan to appreciate by 18% in July and permitted it to fluctuate 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 its time and tread carefully before allowing the yuan to rise further in value.

Economics Document

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

# From documents to images

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



# Image representation

1. Introduction

## 2. Bow

2.1 Origin

2.2 Docs to images

2.3 Representation

2.4 Overview

2.5 Vocabulary

2.6 Learning

2.7 Classification

2.8 Results

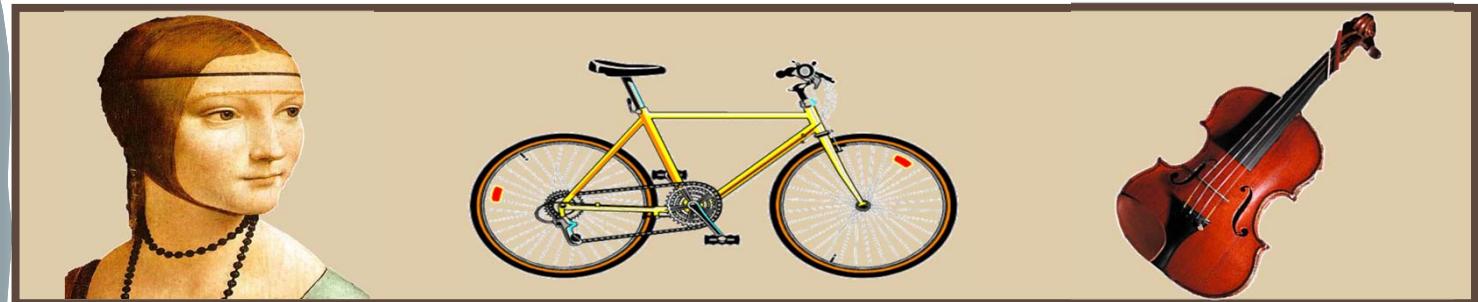
2.9 Conclusions

3. Spatial Information

4. Geometric features

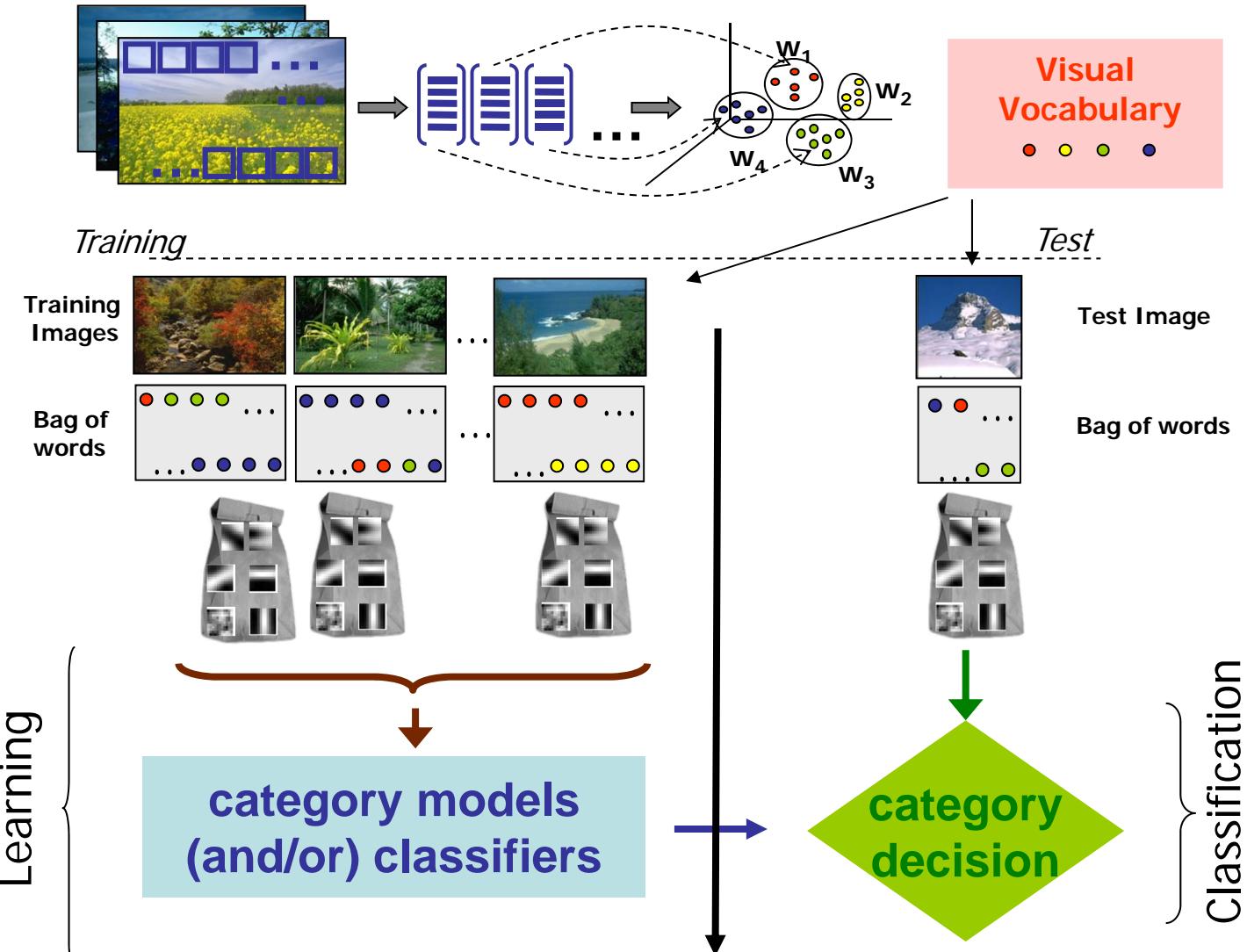
5. Merging Features

4. Conclusions



# Overview

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



# Visual Vocabulary

1. Introduction

## 2. Bow

2.1 Origin

2.2 Docs to images

2.3 Representation

2.4 Overview

## 2.5 Vocabulary

2.6 Learning

2.7 Classification

2.8 Results

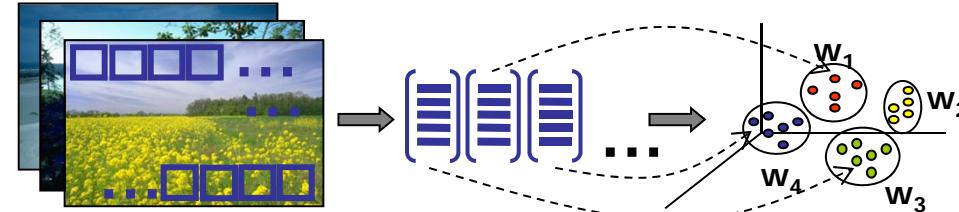
2.9 Conclusions

3. Spatial Information

4. Geometric features

5. Merging Features

4. Conclusions

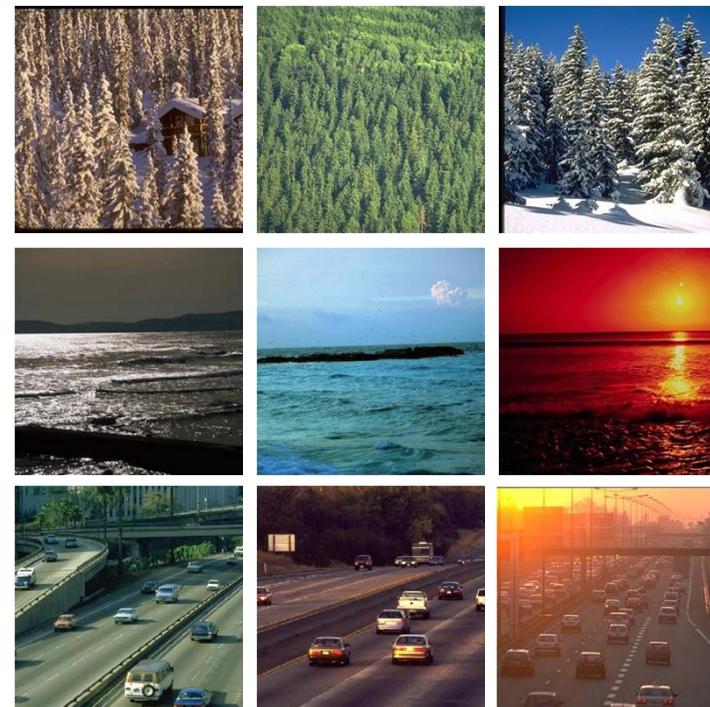


**Visual Vocabulary**



# Feature detection & representation

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



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

1. Introduction

## 2. Bow

2.1 Origin

2.2 Docs to images

2.3 Representation

2.4 Overview

## 2.5 Vocabulary

2.6 Learning

2.7 Classification

2.8 Results

## 2.9 Conclusions

3. Spatial Information

4. Geometric features

5. Merging Features

4. Conclusions

SPARSE PATCHES – Affine covariant regions (harris & hessian)

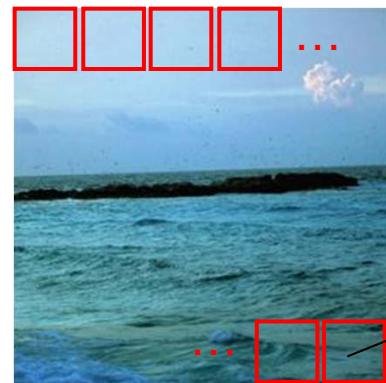


# Feature detection & representation

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions

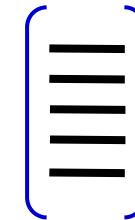
## DENSE PATCHES

### Textons



Parameters: N – size of patch

M – distance between patches



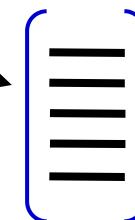
Row reorder gray values  
And form a vector of  
Size  $N^2$

## SIFT



Parameters: r – radii of patch

M – distance between patches



128- SIFT descriptor

# Feature detection & representation

1. Introduction

## 2. Bow

2.1 Origin

2.2 Docs to images

2.3 Representation

2.4 Overview

2.5 Vocabulary

2.6 Learning

2.7 Classification

2.8 Results

2.9 Conclusions

3. Spatial Information

4. Geometric features

5. Merging Features

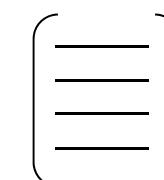
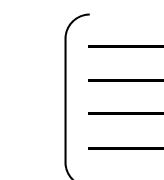
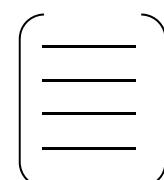
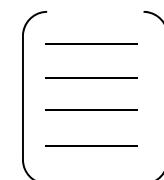
4. Conclusions

## USING COLOUR

Gray

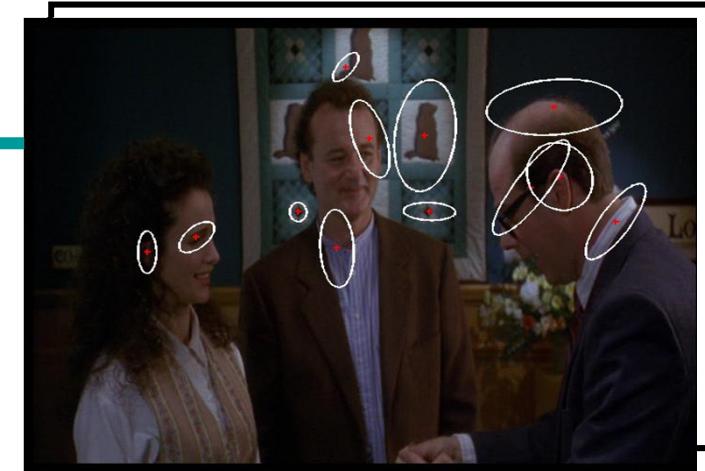
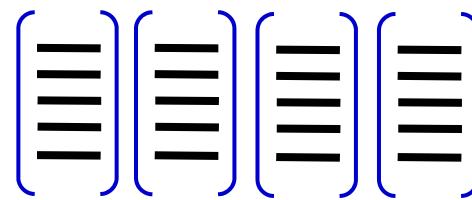


Colour (HSV) → descriptors are computed for each component



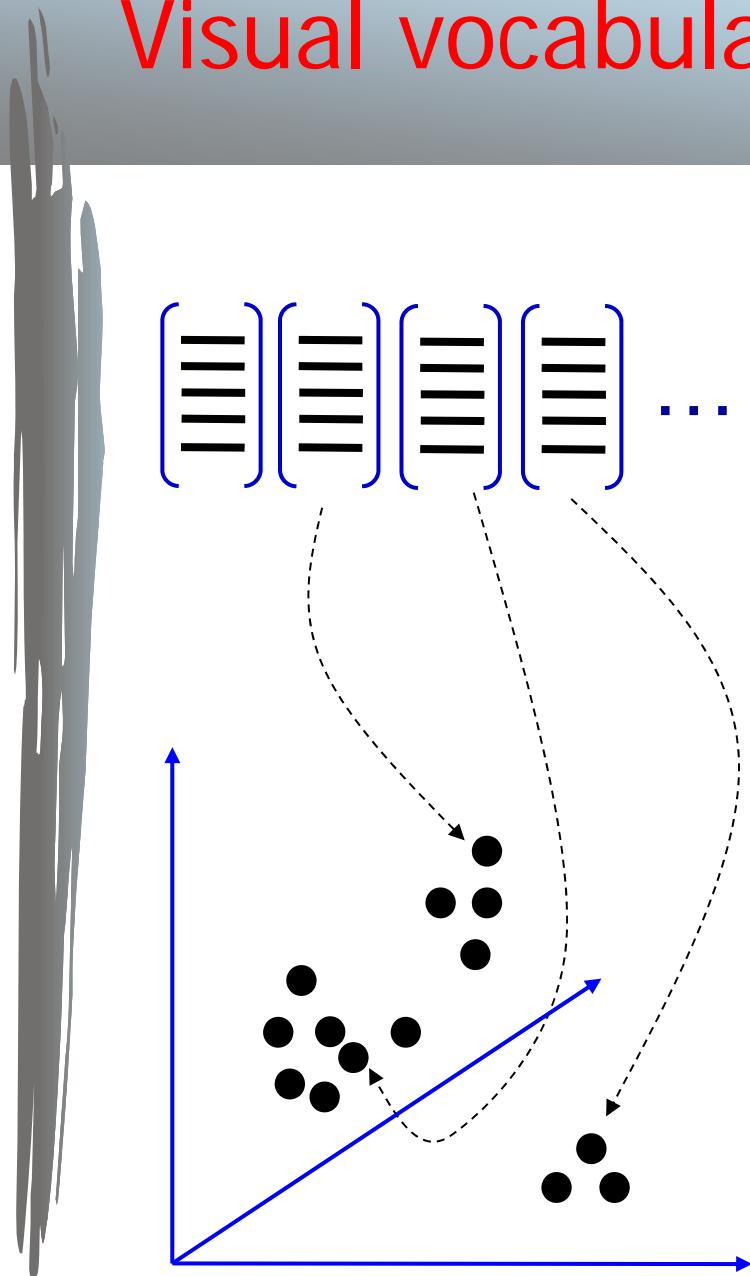
# Visual vocabulary formation

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary**
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



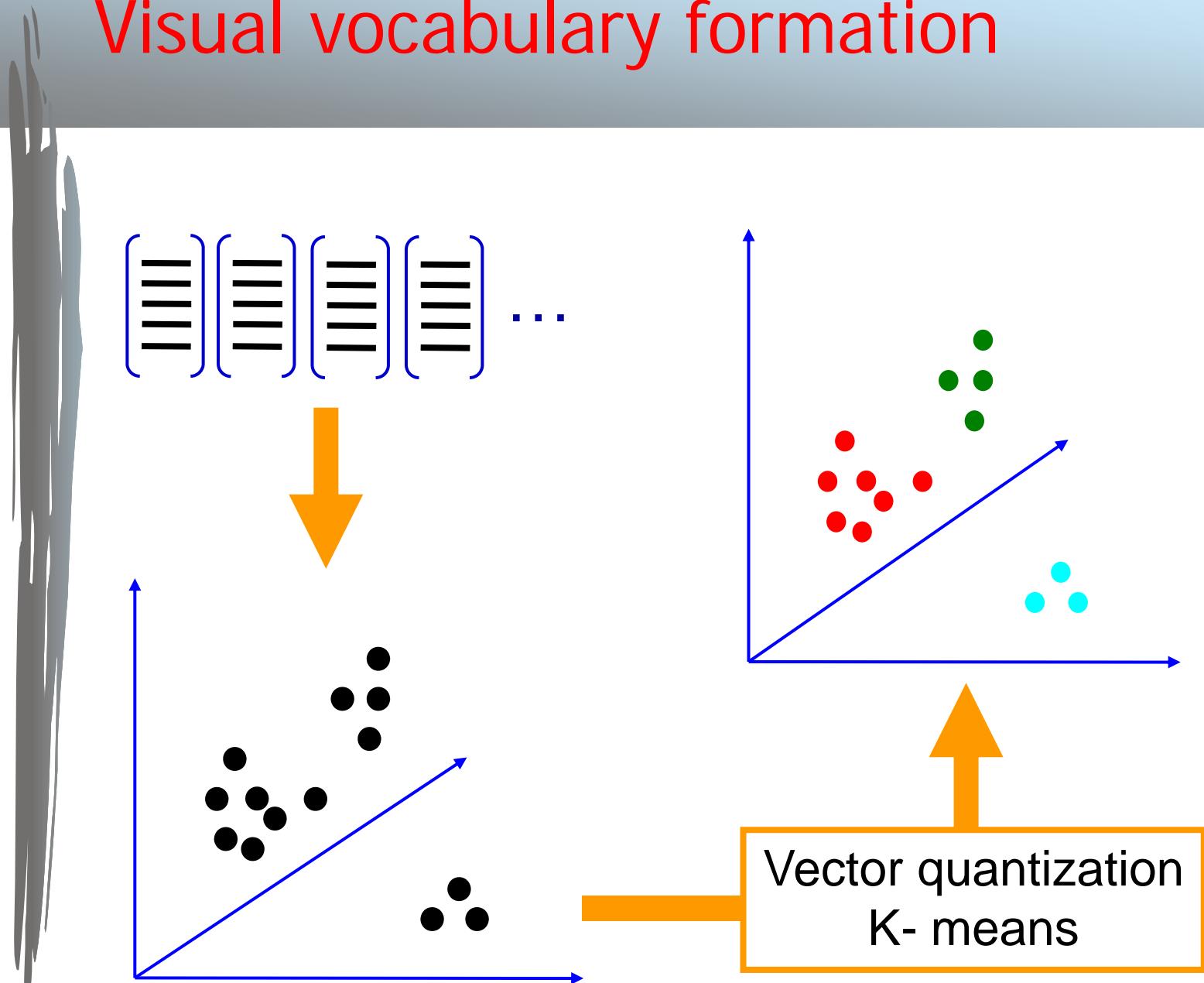
# Visual vocabulary formation

- 1. Introduction
- 2. Bow**
- 2.1 Origin
- 2.2 Docs to images
- 2.3 Representation
- 2.4 Overview
- 2.5 Vocabulary**
- 2.6 Learning
- 2.7 Classification
- 2.8 Results
- 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



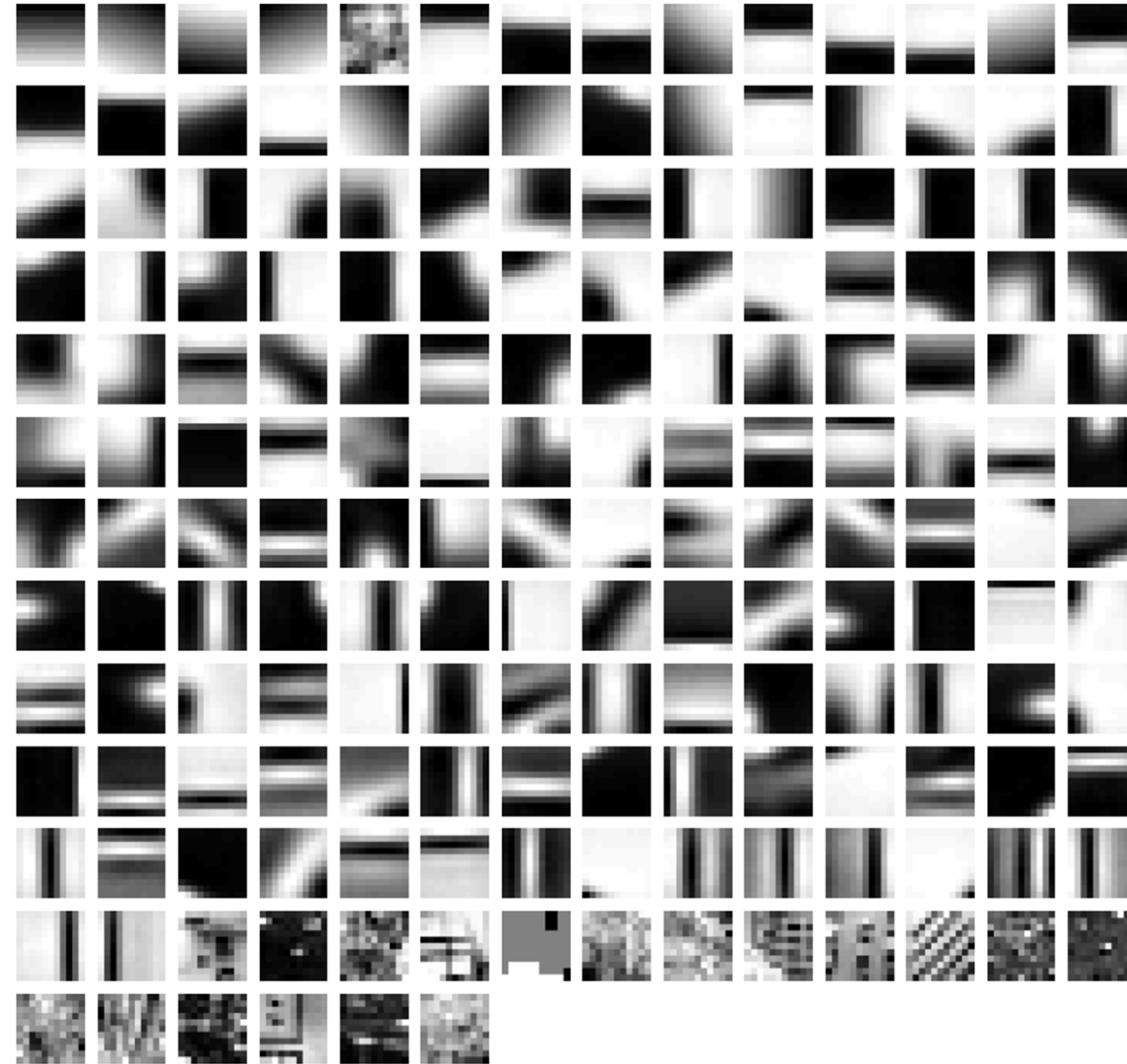
# Visual vocabulary formation

- 1. Introduction
- 2. Bow**
- 2.1 Origin
- 2.2 Docs to images
- 2.3 Representation
- 2.4 Overview
- 2.5 Vocabulary**
- 2.6 Learning
- 2.7 Classification
- 2.8 Results
- 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



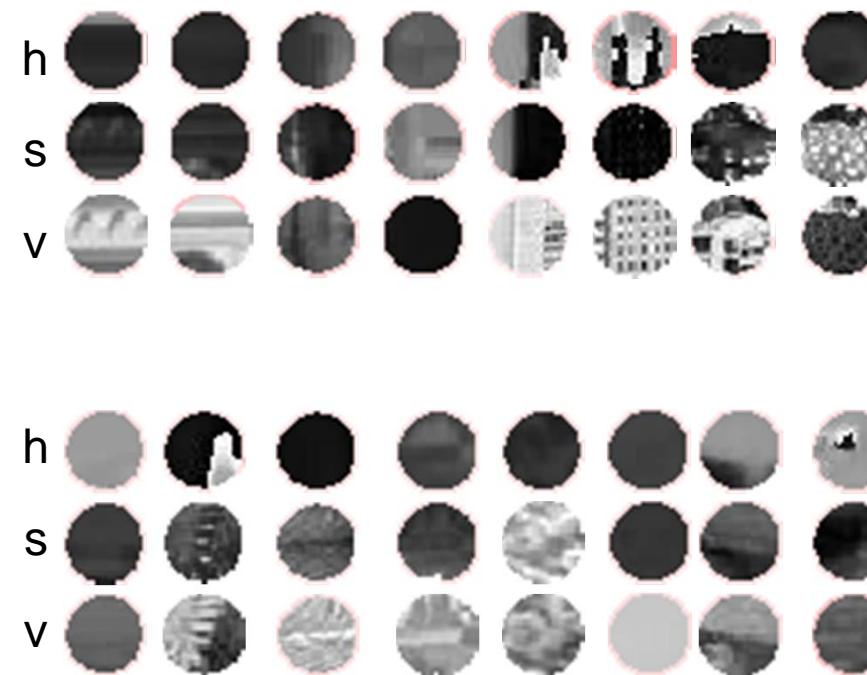
# Examples of visual words

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary**
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



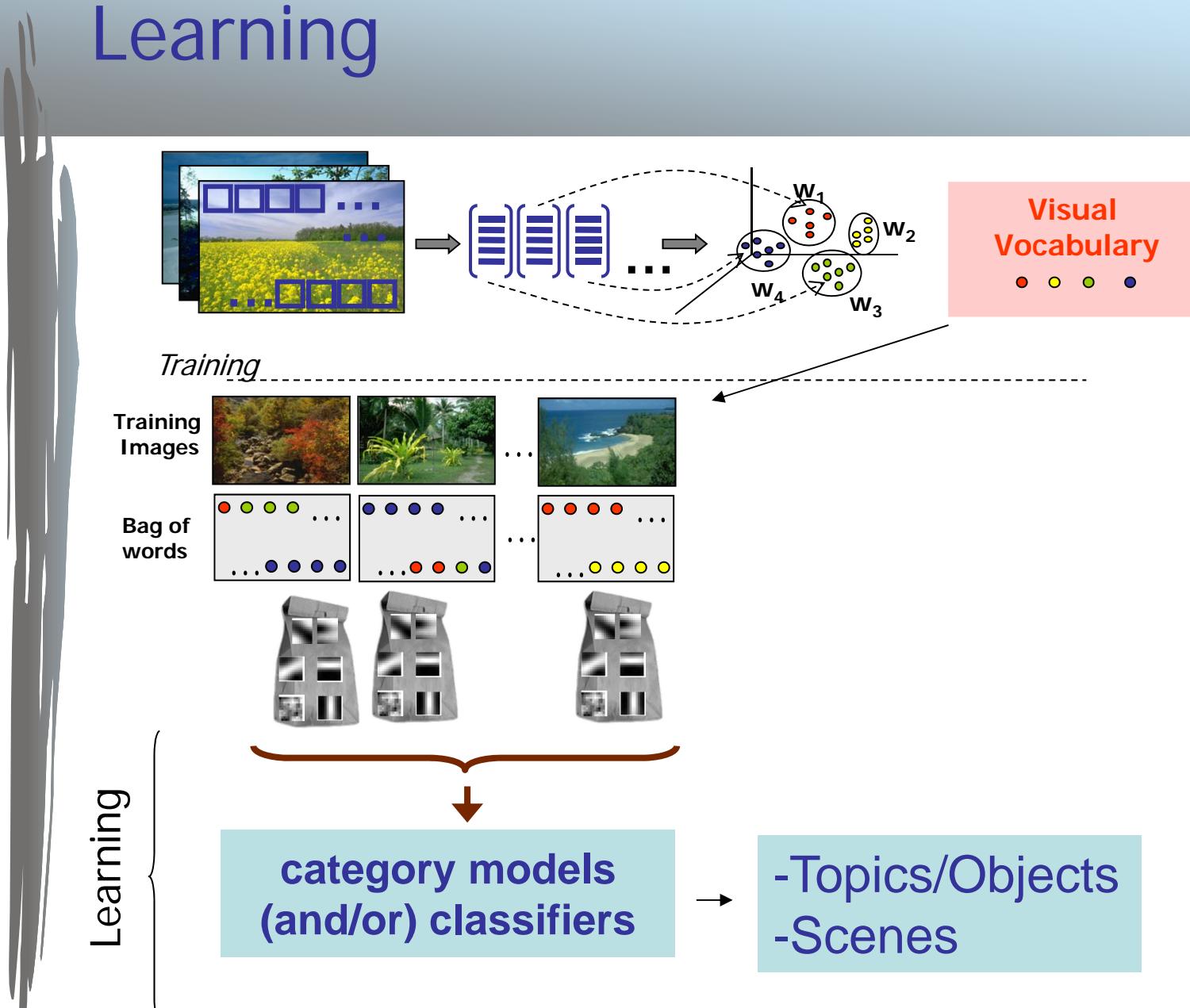
# Examples of visual words

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary**
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



# Learning

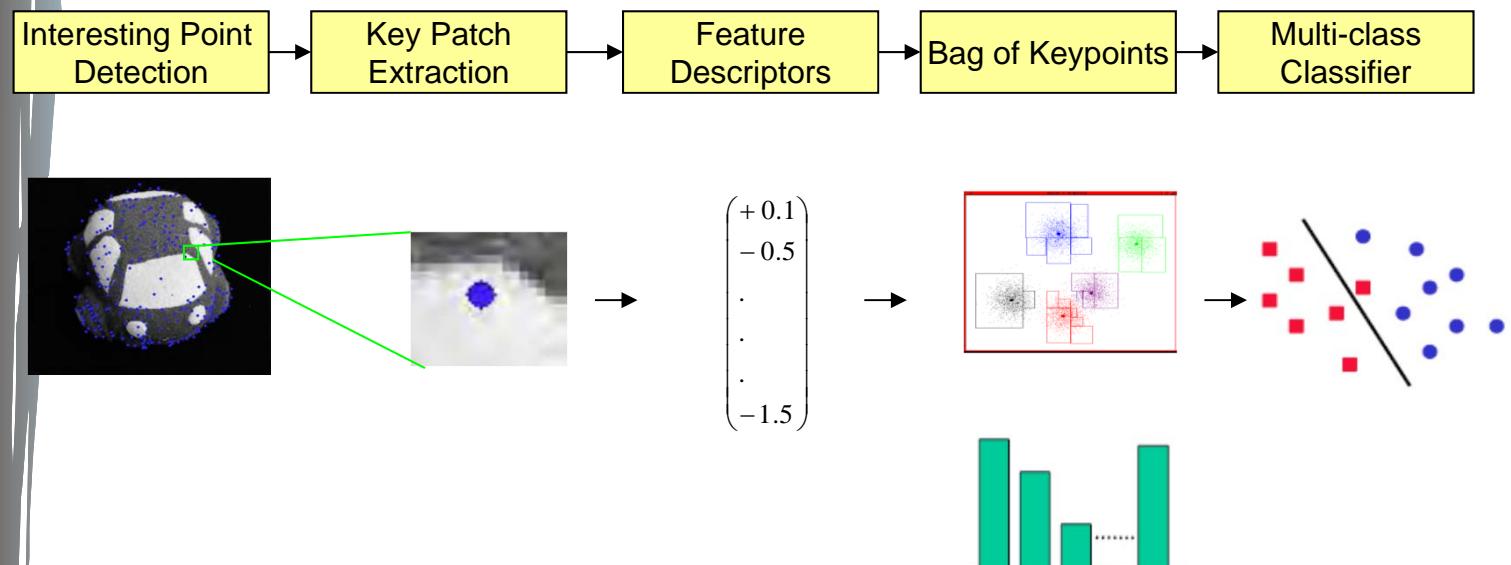
- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning**
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



# Learning

## SPARSE BAG OF WORDS APPROACH

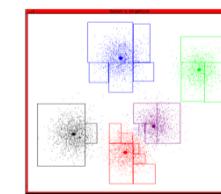
- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



# Learning

## SPARSE BAG OF WORDS APPROACH

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions

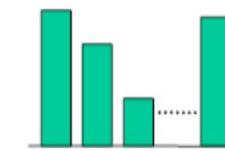


- Construction of a vocabulary
  - Kmeans clustering → find “centroids”  
(on all the descriptors we find from a set of training images)
  - Define a “vocabulary” as a set of “centroids”, where every centroid represents a “word”

# Learning

## SPARSE BAG OF WORDS APPROACH

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions

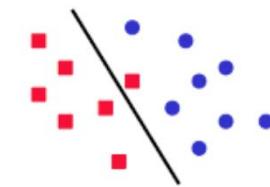


- **Histogram**
  - Counts the number of occurrences of different *visual words* in each image

# Learning

## SPARSE BAG OF WORDS APPROACH

- 1. Introduction
- 2. Bow**
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions



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

# Generative Learning

- 1. Introduction
- 2. Bow
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions

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



- 1. Introduction
- 2. Bow
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions

- 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

1. Introduction
<b>2. Bow</b>
2.1 Origin
2.2 Docs to images
2.3 Representation
2.4 Overview
2.5 Vocabulary
2.6 Learning
2.7 Classification
2.8 Results
2.9 Conclusions
3. Spatial Information
4. Geometric features
5. Merging Features
4. Conclusions

- **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

- 1. Introduction
- 2. Bow
  - 2.1 Origin
  - 2.2 Docs to images
  - 2.3 Representation
  - 2.4 Overview
  - 2.5 Vocabulary
  - 2.6 Learning
  - 2.7 Classification
  - 2.8 Results
  - 2.9 Conclusions
- 3. Spatial Information
- 4. Geometric features
- 5. Merging Features
- 4. Conclusions

- [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/>