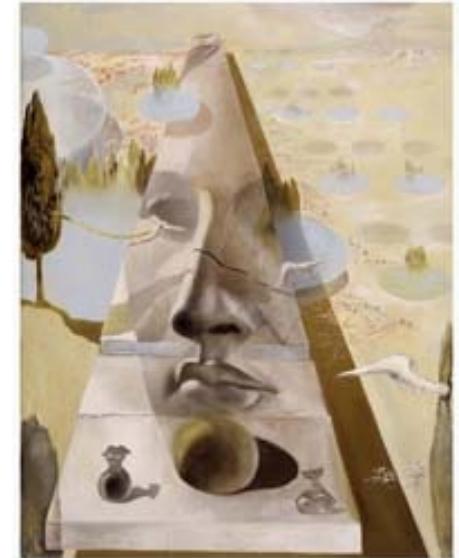


Lecture 11

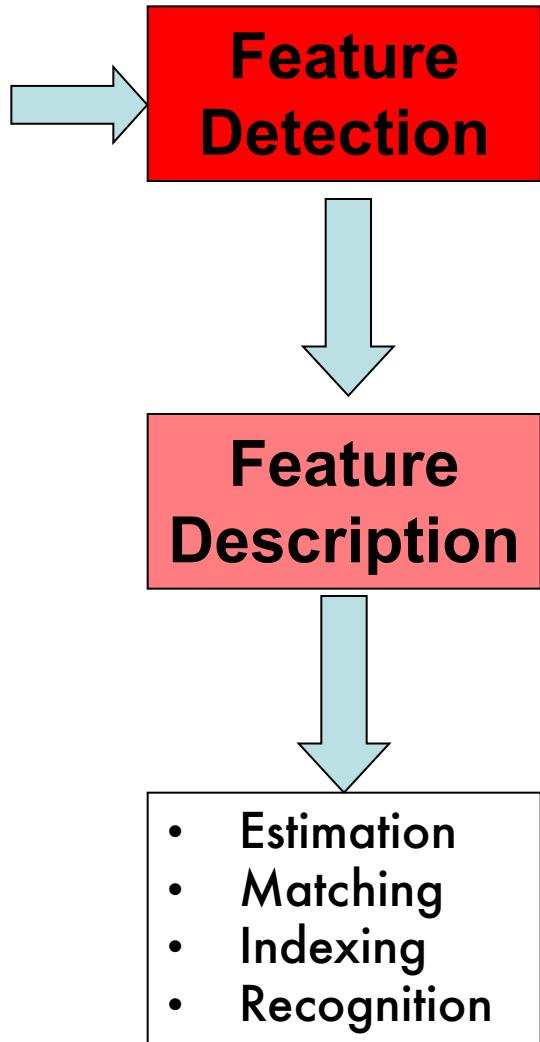
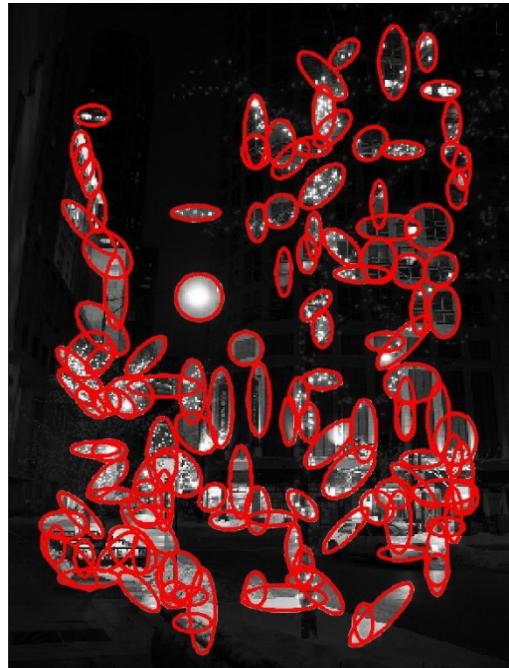
Visual recognition



- An introduction to recognition
- Image classification – the bag of words model

[FP] - Chapters 6 (sec. 6.2)
[FP] - Chapters 16 (sec. 16.1)
[FP] - Chapters 17 (sec. 17.1)

What we have seen for far



e.g. DoG

e.g. SIFT

Deep neural network

What's visual recognition?



Classification:

Does this image contain a building? [yes/no]



Classification:

Is this an beach?

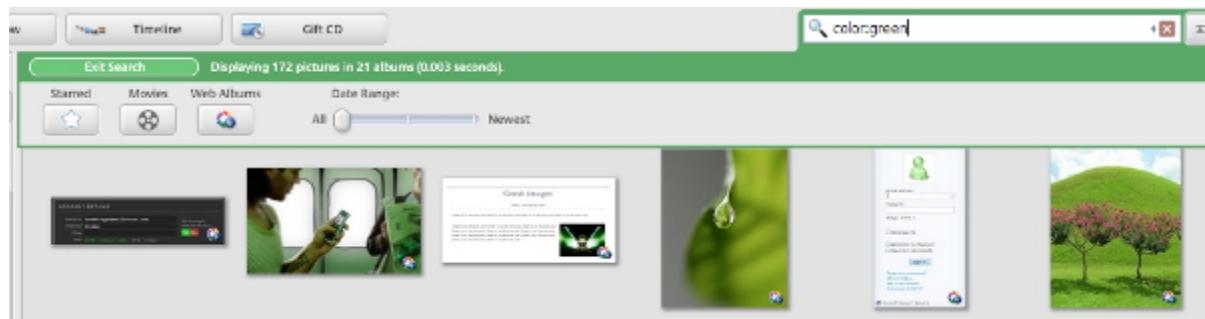


Image Search or Indexing



A screenshot of a Google Images search results page. The search term 'street' is entered in the search bar. The results show six images: 'Street sweeper', 'Street Maintenance', 'Main Street Station', 'SHPO Wayne Donaldson at Main Street ...', 'Lombard Street, words crookedest See ...', and 'Street Bike (BS70 4A) Details'. Each result includes a thumbnail, a caption, and a link.

Organizing photo collections



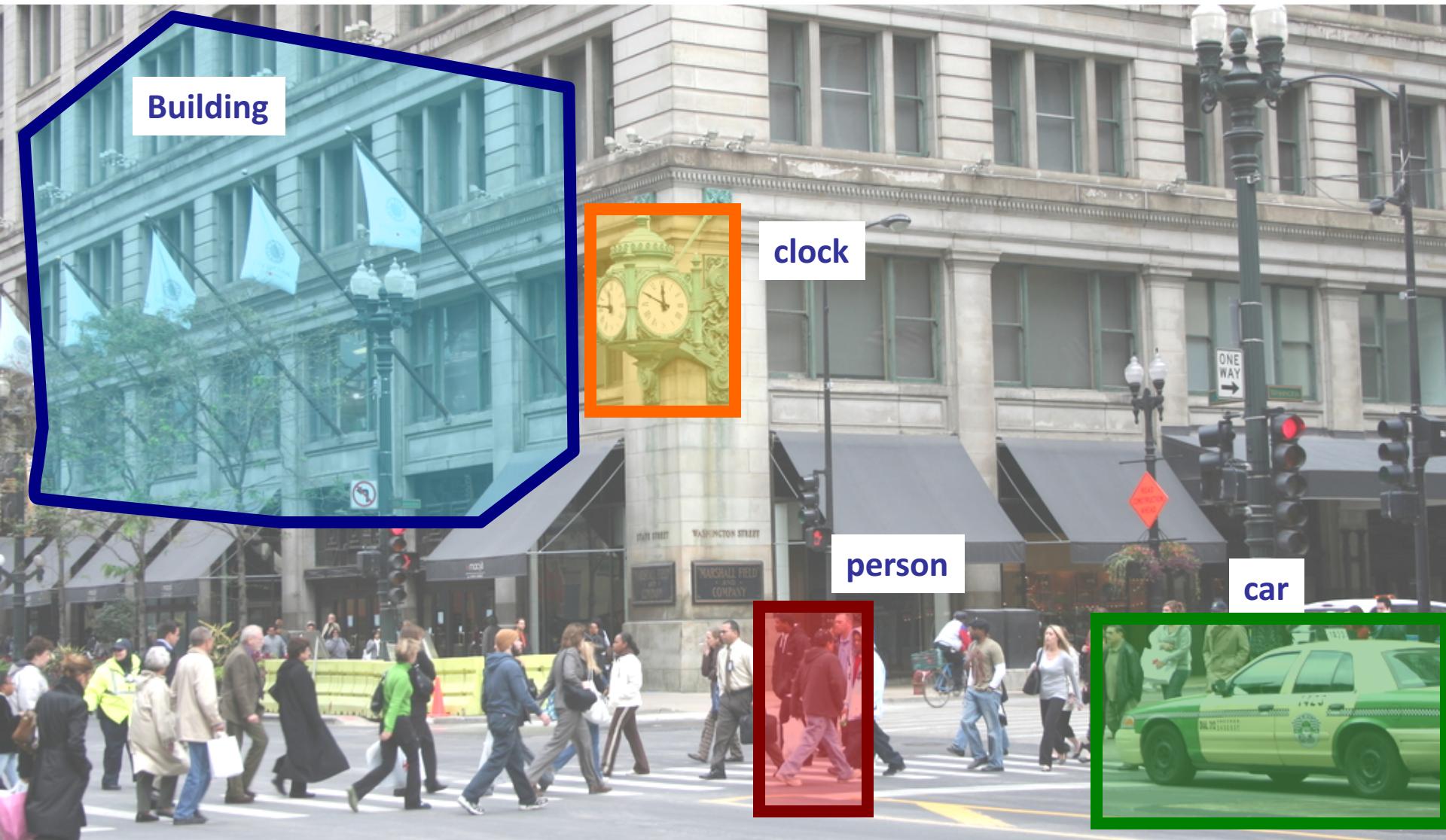
Detection:

Does this image contain a car? [where?]



Detection:

Which object does this image contain? [where?]



Detection:

Accurate localization (segmentation)



Object detection is useful...



Computational photography



Assistive technologies



Surveillance



Security



Assistive driving

Categorization vs Single instance recognition

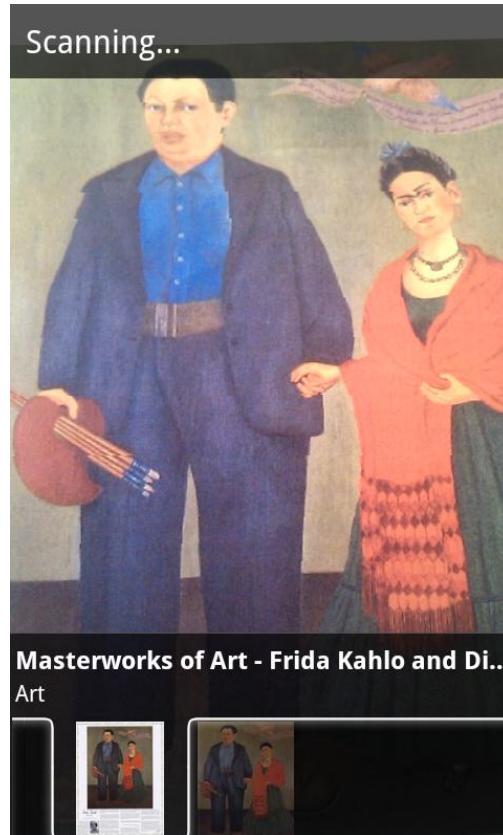
Which building is this? *Marshall Field* building in Chicago



Visual search and landmarks recognition



Google
Goggles



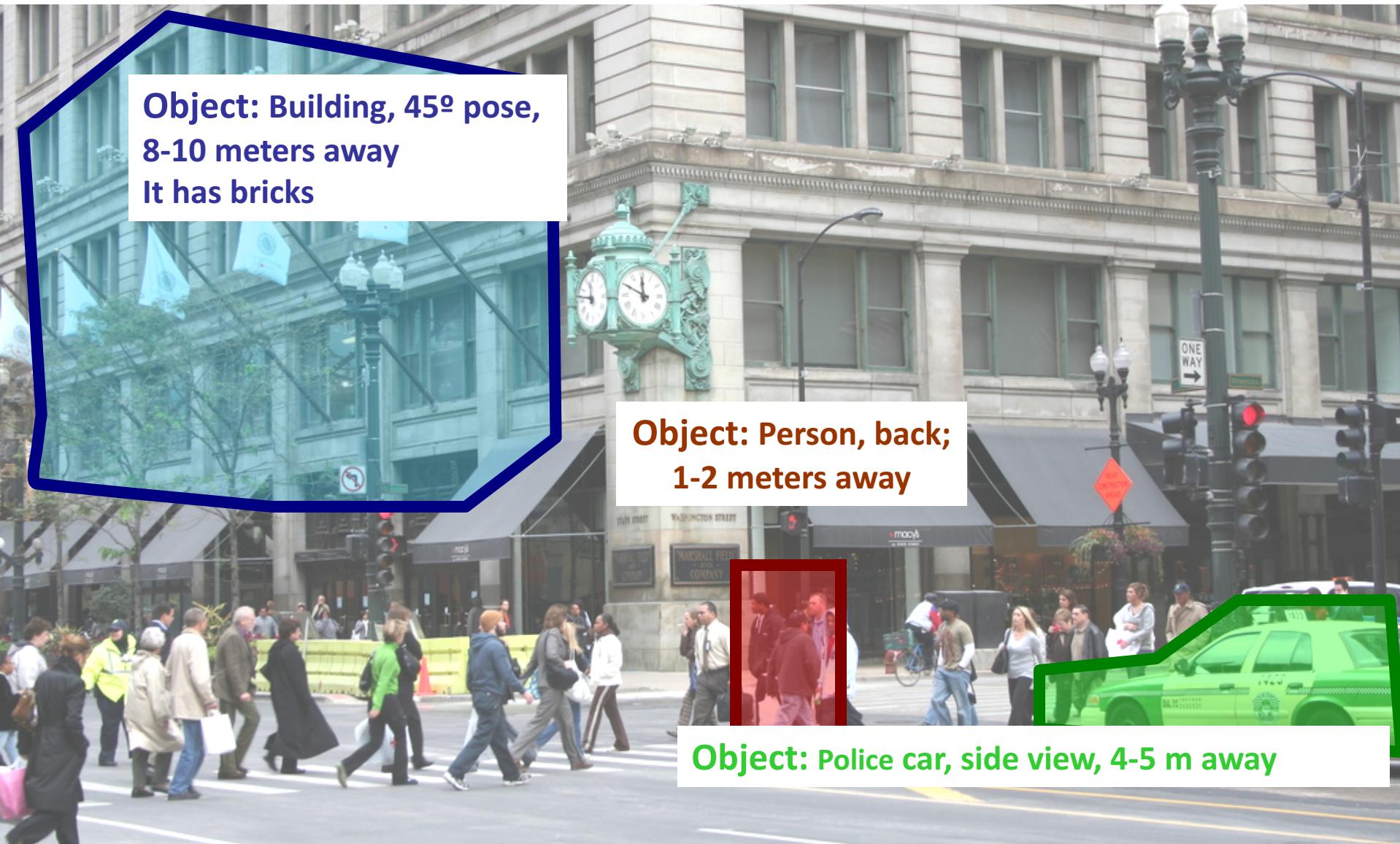
Visual search and landmarks recognition

Where is the crunchy nut?



Detection: Estimating object semantic & geometric attributes

Object: Building, 45° pose,
8-10 meters away
It has bricks



Object: Person, back;
1-2 meters away

Object: Police car, side view, 4-5 m away

Activity or Event recognition

What are these people doing?



Visual Recognition

- Design algorithms that are capable to
 - Classify images or videos
 - Detect and localize objects
 - Estimate semantic and geometrical attributes
 - Classify human activities and events

Why is this challenging?

How many object categories are there?

~10,000 to 30,000



Challenges: viewpoint variation



Michelangelo 1475-1564

slide credit: Fei-Fei, Fergus & Torralba

Challenges: illumination

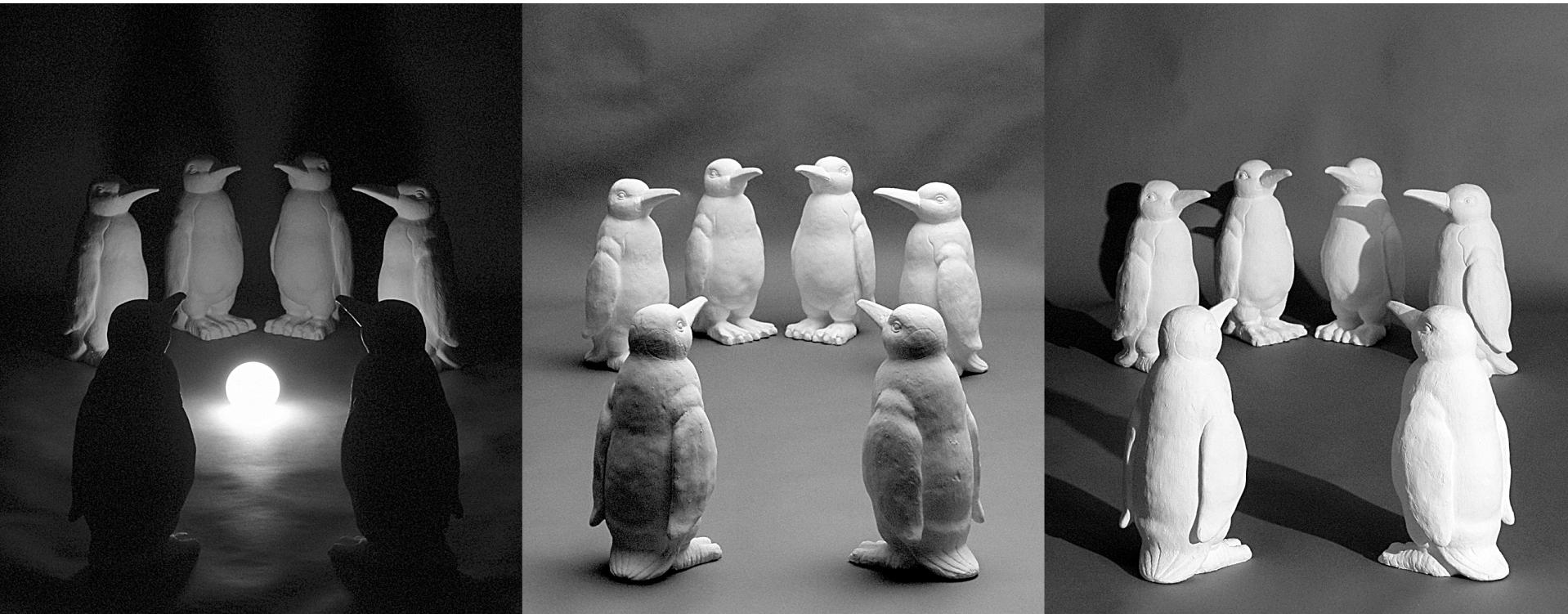


image credit: J. Koenderink

Challenges: scale



slide credit: Fei-Fei, Fergus & Torralba

Challenges: deformation



Challenges: occlusion



Magritte, 1957

slide credit: Fei-Fei, Fergus & Torralba

Challenges: background clutter



Kilmenny Niland. 1995

Challenges: intra-class variation

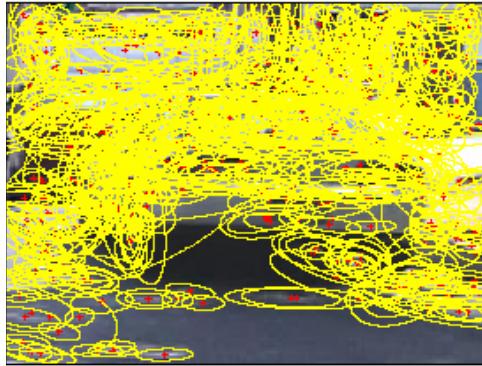


Basic properties

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

Representation

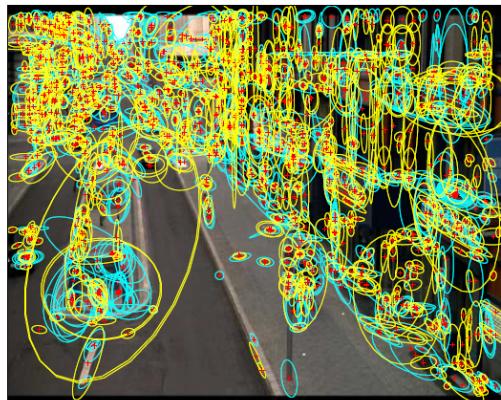
- Building blocks: Sampling strategies



Interest operators



Dense, uniformly



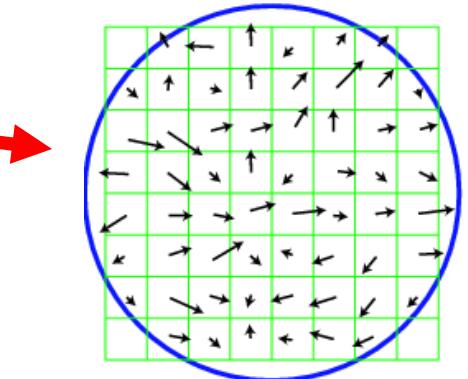
Multiple interest operators



Randomly

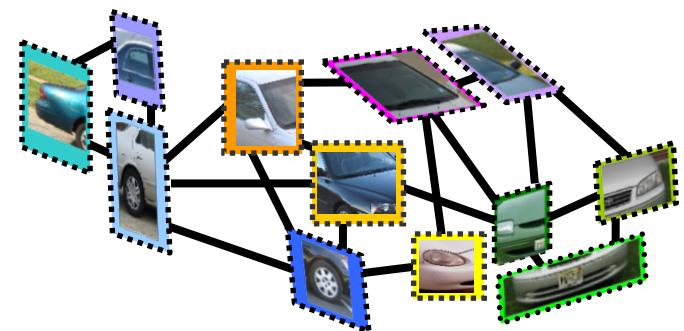
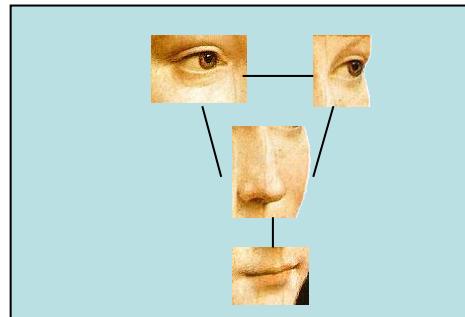
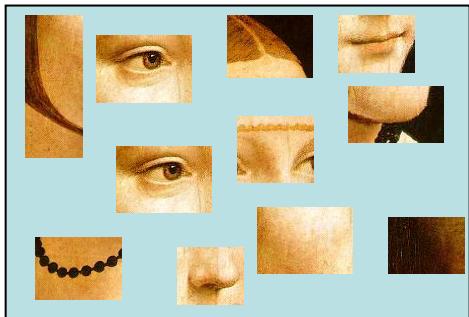
Representation

- Building blocks: Choice of descriptors
[SIFT, HOG, codewords, CNNs feature vectors....]



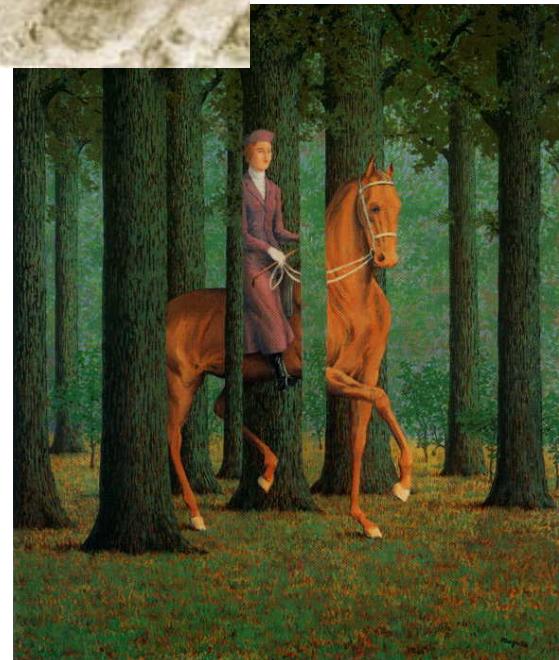
Representation

- Appearance only
- 2D location and appearance
- 3D location and appearance



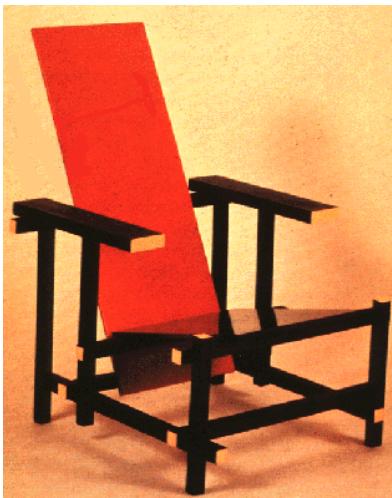
Representation

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



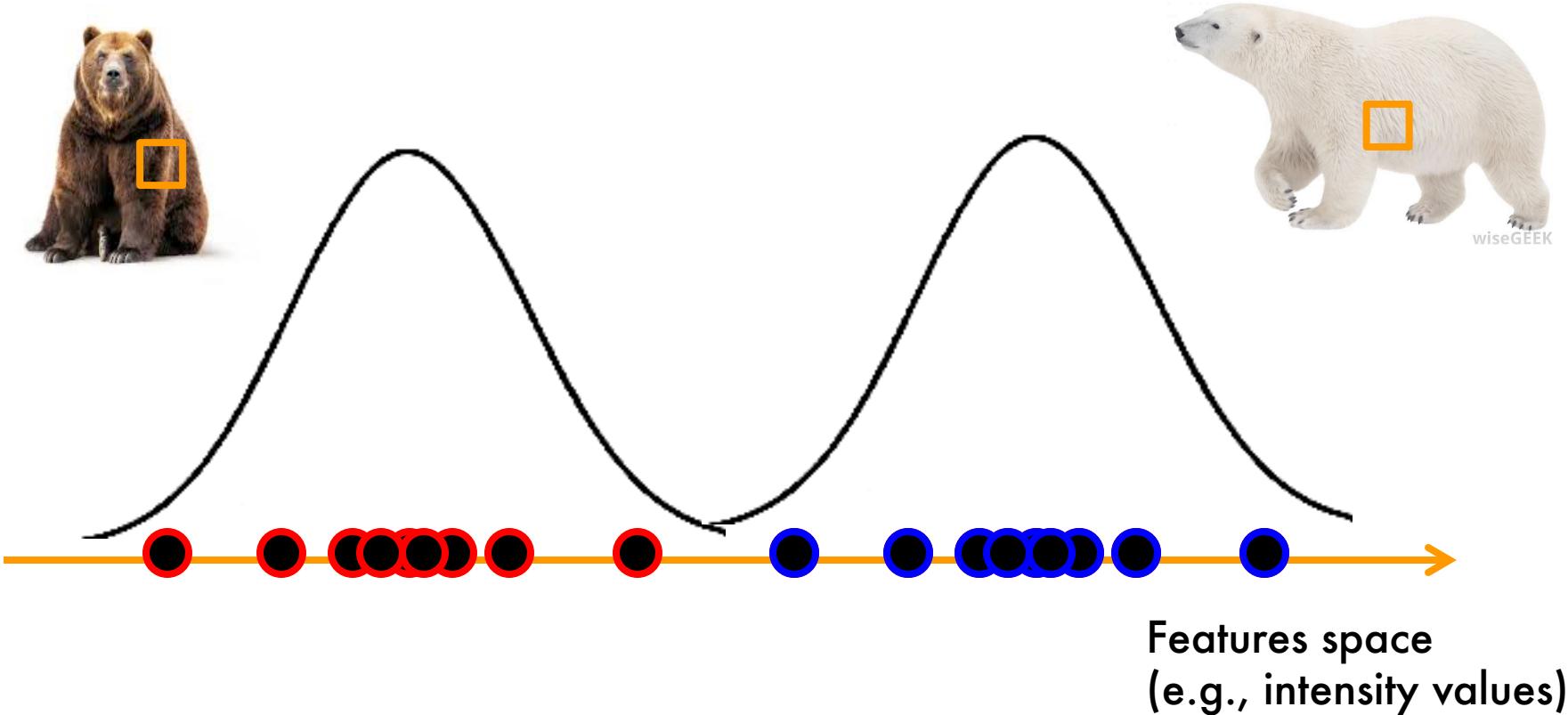
Representation

- How to handle intra-class variability?
 - It is convenient to describe object categories using probabilistic models



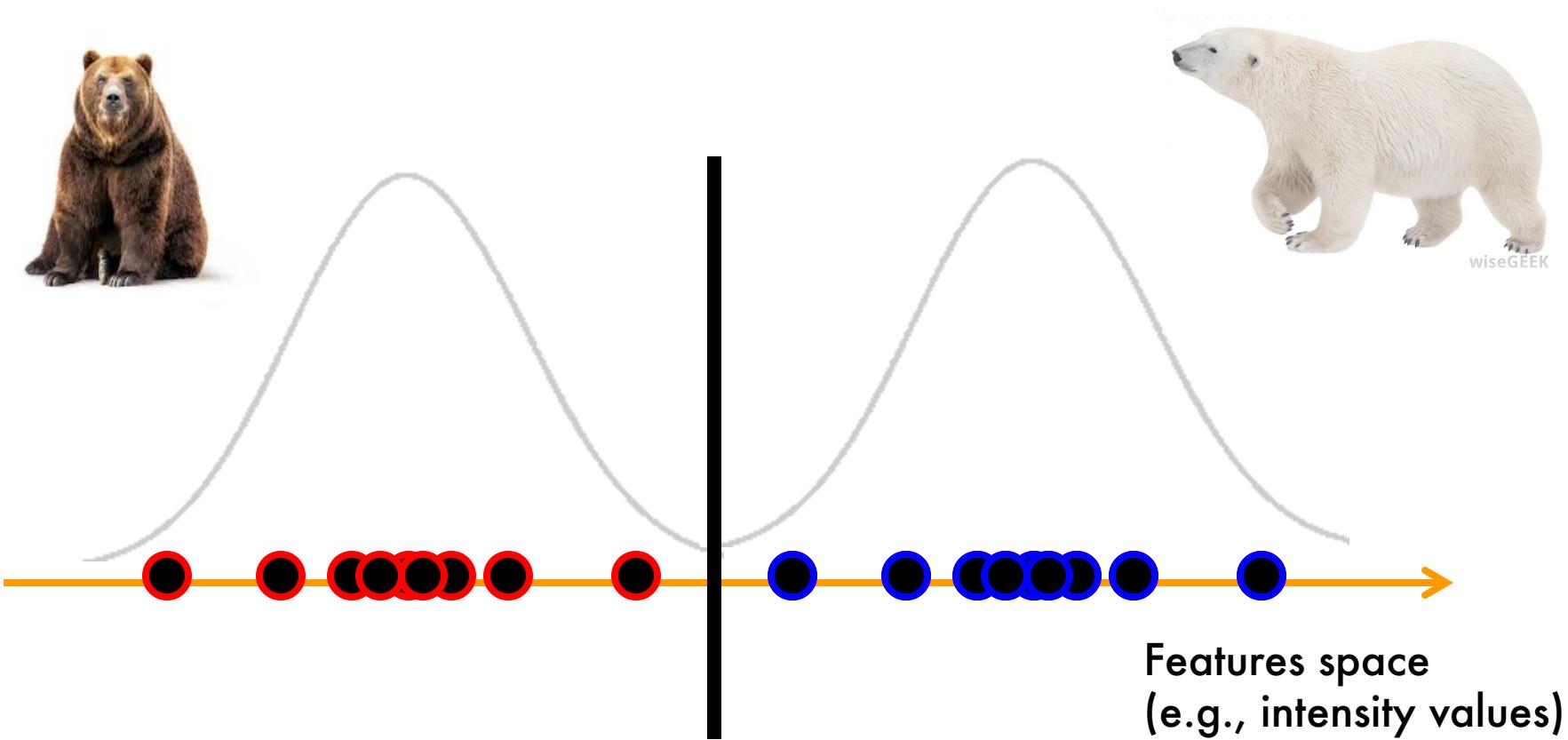
Generative – vs – discriminative

- Generative: Infer a function that can generate (explain) your observations



Generative – vs – discriminative

- Discriminative: Infer a function that can separate (discriminate) your observations

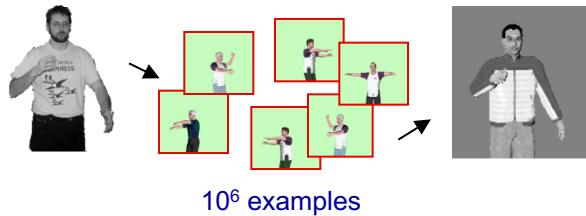


Generative models

- Naïve Bayes classifier
 - Csurka Bray, Dance & Fan, 2004
- Hierarchical Bayesian topic models (e.g. pLSA and LDA)
 - Object categorization: Sivic et al. 2005, Sudderth et al. 2005
 - Natural scene categorization: Fei-Fei et al. 2005
- 2D or 3D Part based models
 - Constellation models: Weber et al 2000; Fergus et al 2003
 - Star models: ISM (Leibe et al 05)
 - Sun, et al., 2009
- Variational Autoencoders (VAEs)
 - Kingma & Welling, 2013
- Generative Adversarial Networks (GANs)
 - Goodfellow et al. 2014

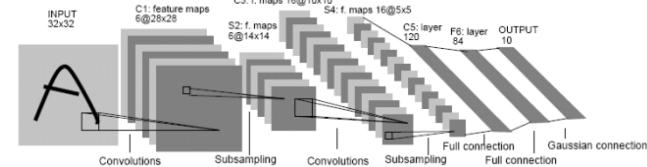
Discriminative models

Nearest neighbor



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

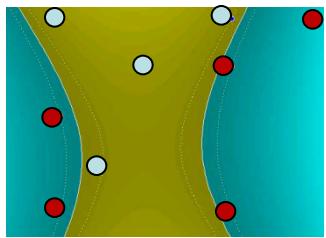
Neural networks



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

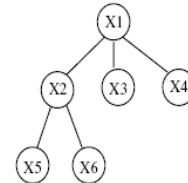
...

Support Vector Machines



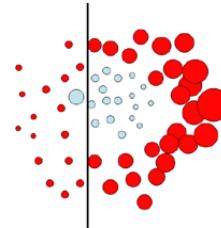
Guyon, Vapnik,
Heisele, Serre, Poggio...

Latent SVM Structural SVM



Felzenszwalb 00
Ramanan 03...

Boosting



Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,...

Basic properties

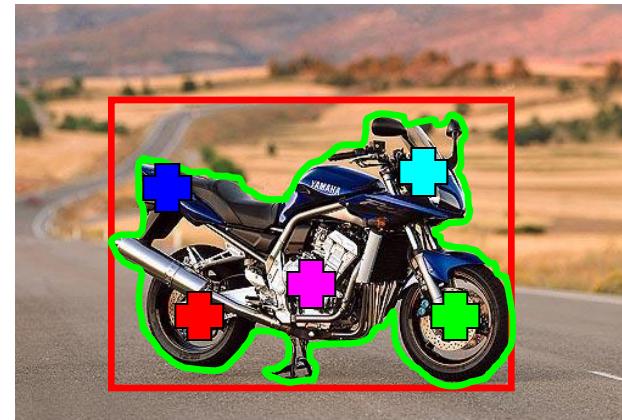
- Representation
 - How to represent an object category; which classification scheme?
- Learning
 - How to learn the classifier, given training data
- Recognition
 - How the classifier is to be used on novel data

Learning

- Learning parameters
- Generative functions or separating functions?

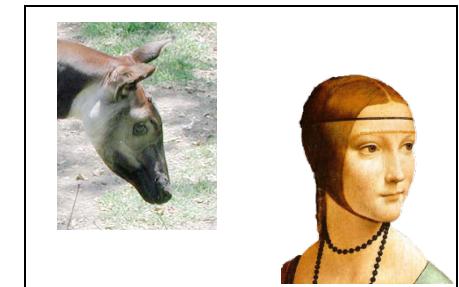
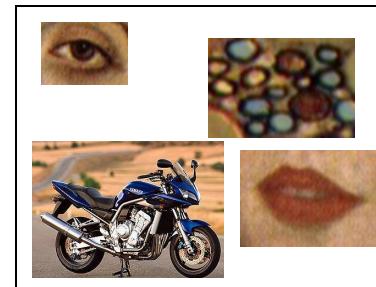
Learning

- Learning parameters
- Generative functions or separating functions?
- Level of supervision
 - Noisy labels; image labels; bounding box; manual segmentation; part annotations
- Batch/incremental
- Priors



Learning

- Learning parameters
- Generative functions or separating functions?
- Level of supervision
 - Noisy labels; image labels; bounding box; manual segmentation; part annotations
- Batch/incremental
- Priors
- Training images:
 - Negative images for discriminative methods
 - Over-fitting

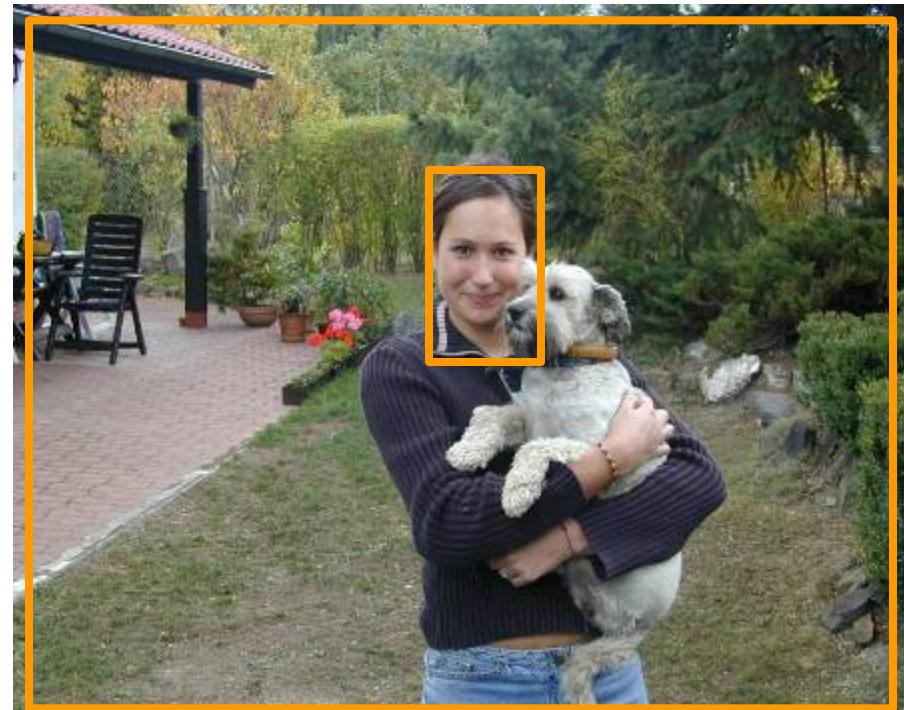


Basic properties

- Representation
 - How to represent an object category; which classification scheme?
- Learning
 - How to learn the classifier, given training data
- Recognition
 - How the classifier is to be used on novel data

Recognition

- Recognition task: classification, detection, etc..



Recognition

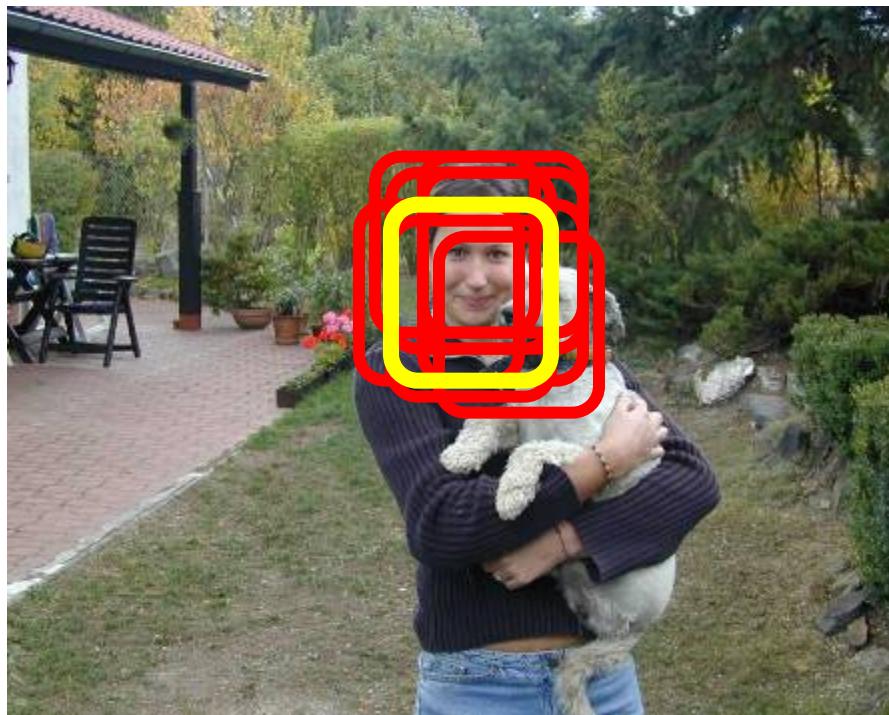
- Recognition task
- Search strategy:
 - Sliding Windows

: Viola, Jones 2001
: Dalal and Bill Triggs, 2005



Non-max suppression

Section 17.1 [FP]



Recognition

- Recognition task
- Search strategy:
 - Sliding Windows

Simple!
But computational
expensive..

Recognition

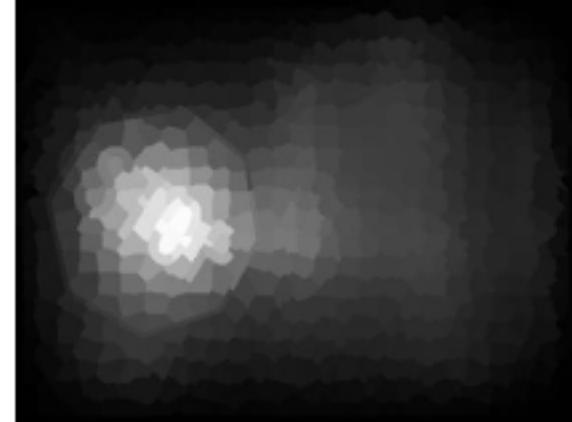
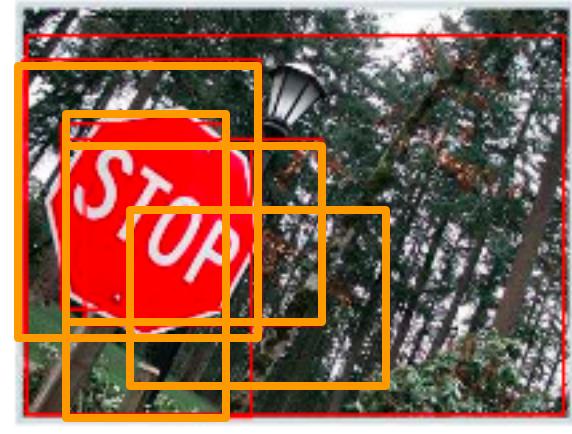
- Recognition task
- Search strategy:
 - Sliding Windows
 - Bottom-up cues (segmentation)



Felzenswalb and Huttenlocher, 2004

Recognition

- Recognition task
- Search strategy:
 - Sliding Windows
 - Bottom-up cues (segmentation)
 - Saliency; regional proposals



Jia & Han, 13
Alexe, et al 10
...

Recognition

- Recognition task
- Search strategy
- Attributes

- Savarese, 2007
- Sun et al 2009
- Liebelt et al., '08, 10
- Farhadi et al 09

**- It has metal
- it is glossy
- has wheels**

- Farhadi et al 09
- Lampert et al 09
- Wang & Forsyth 09



Recognition

- Recognition task
- Search strategy
- Attributes
- Context

Semantic:

- Torralba et al 03
- Rabinovich et al 07
- Gupta & Davis 08
- Heitz & Koller 08
- L-J Li et al 08
- Bang & Fei-Fei 10

Geometric

- Hoiem, et al 06
- Gould et al 09
- Bao, Sun, Savarese 10

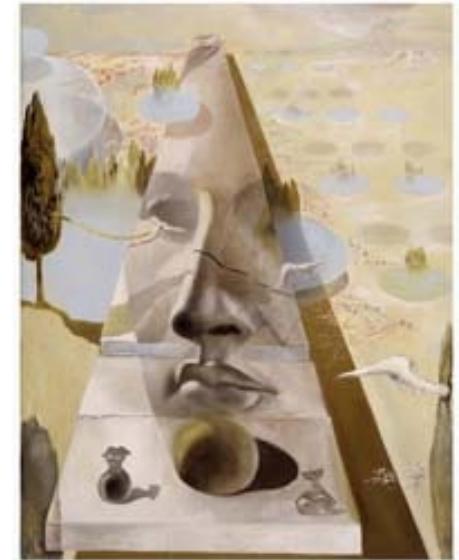


Agenda on recognition

- Image classification (lecture 11, 12, 15)
 - Bag of words representations
- Object detection (lecture 12, 14, 15)
 - 2D object detection
 - 3D object detection
- Scene understanding (lecture 13, 14, 16)

Lecture 11

Visual recognition



- An introduction to recognition
- Image classification – the bag of words model

Bag of words models

- Used for image and object classification
- Designed to handle variability due to:
 - View point
 - Illumination
 - Occlusions
 - Intra-class

Inspired by works on document analysis!

- Early “bag of words” models: mostly texture recognition
 - Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003;
- Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
 - Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004

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 essentially on the messages that reach us from our eyes. For a long time it was believed that the retinal image was processed by visual centers in the brain. In 1960, as a movie screen, a retinal image was discovered in the cerebral cortex, discovered by Hubel and Wiesel. They know that the perception of a visual image is more complex than the simple perception of a retinal image. Following the analysis of the visual pathway to the various components of the cerebral cortex, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes a top-down analysis in a system of nerve cells stored in columns. In this system each column has its specific function and is responsible for a specific detail in the pattern of the retinal image.

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 \$750bn, compared with \$660bn. This will annoy the US, which China's central bank, the People's Bank, deliberately agrees to keep low. The yuan is governed by the central bank, which also needs to demand so much foreign currency from the country. China has agreed to let the value of the yuan against the dollar rise slowly and permitted it to trade 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.

Object

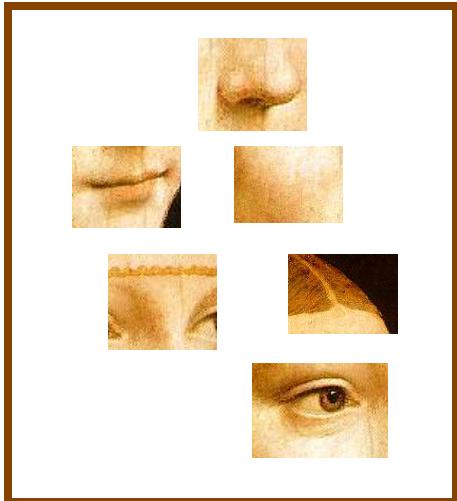
Bag of ‘words’



definition of “BoW”

– Independent features

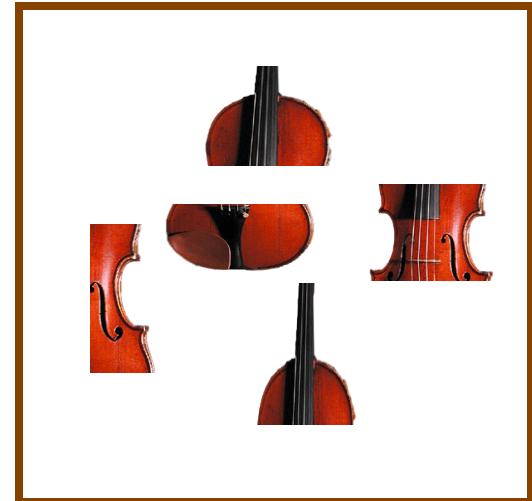
face



bike

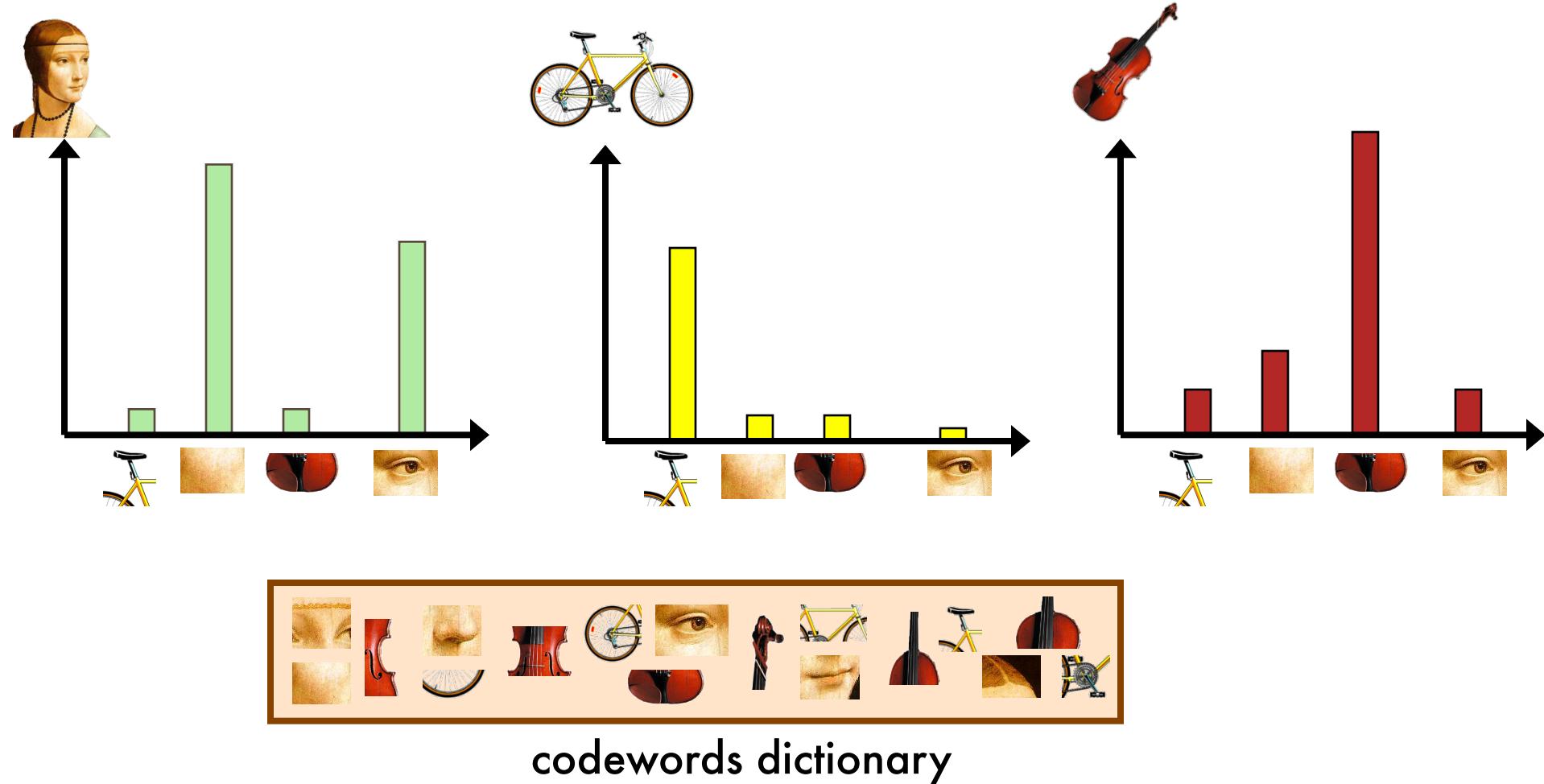


violin

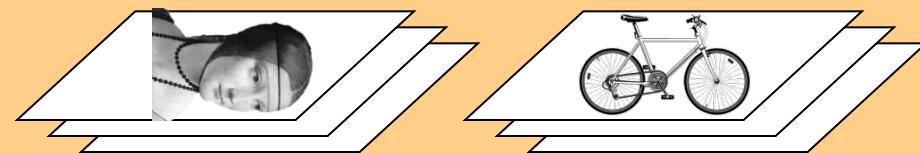


definition of “BoW”

- Independent features
- histogram representation



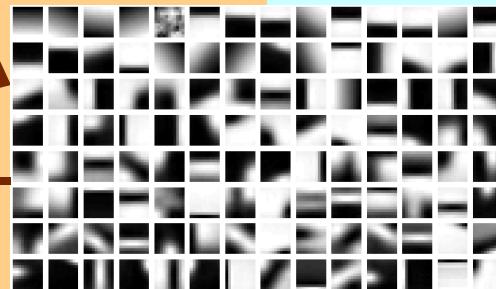
Representation



1. feature detection
& description



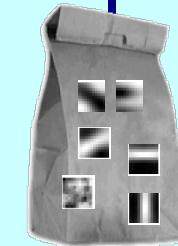
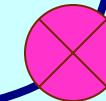
2. codewords dictionary



3. BOW representation



recognition

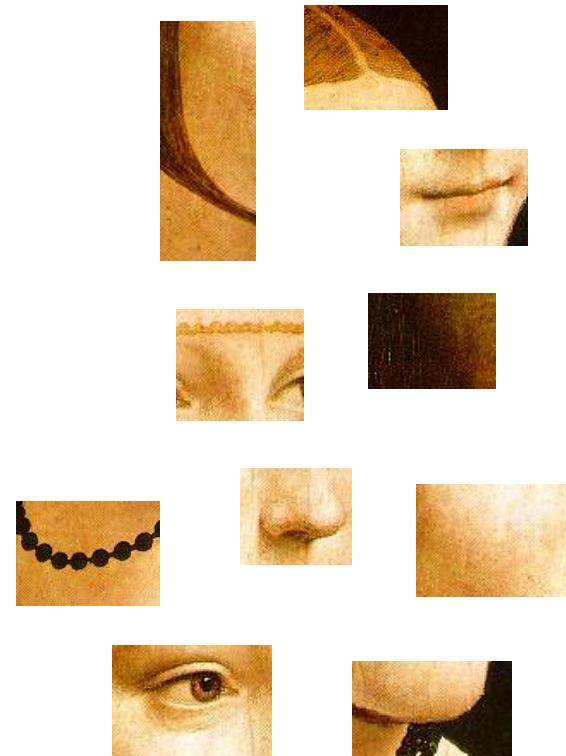


learning

**category models
(and/or) classifiers**

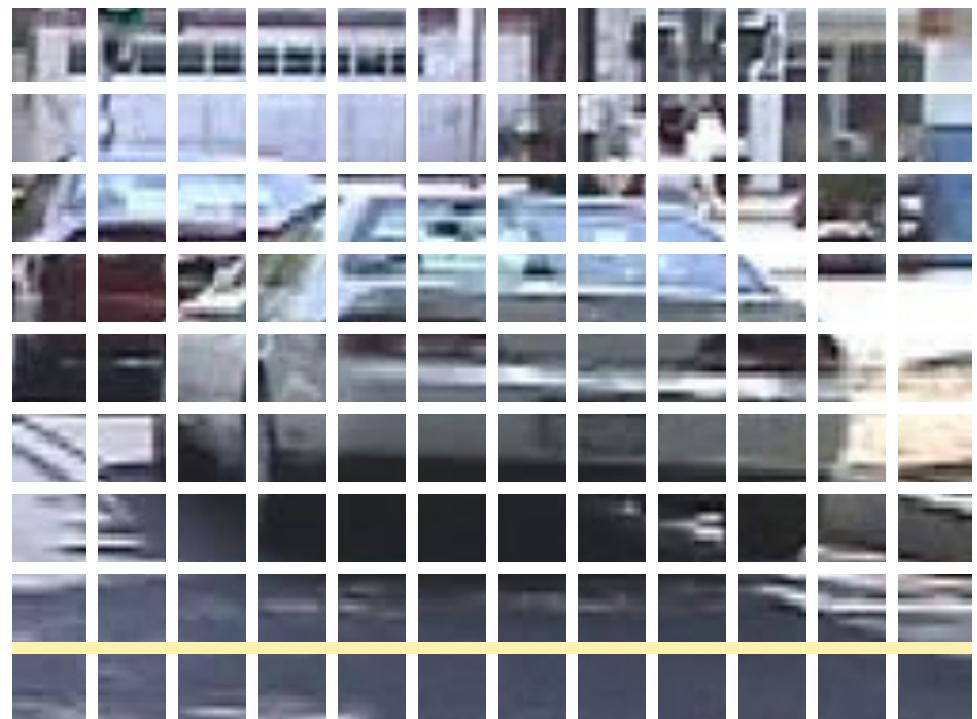
**category
decision**

1. Feature detection and description



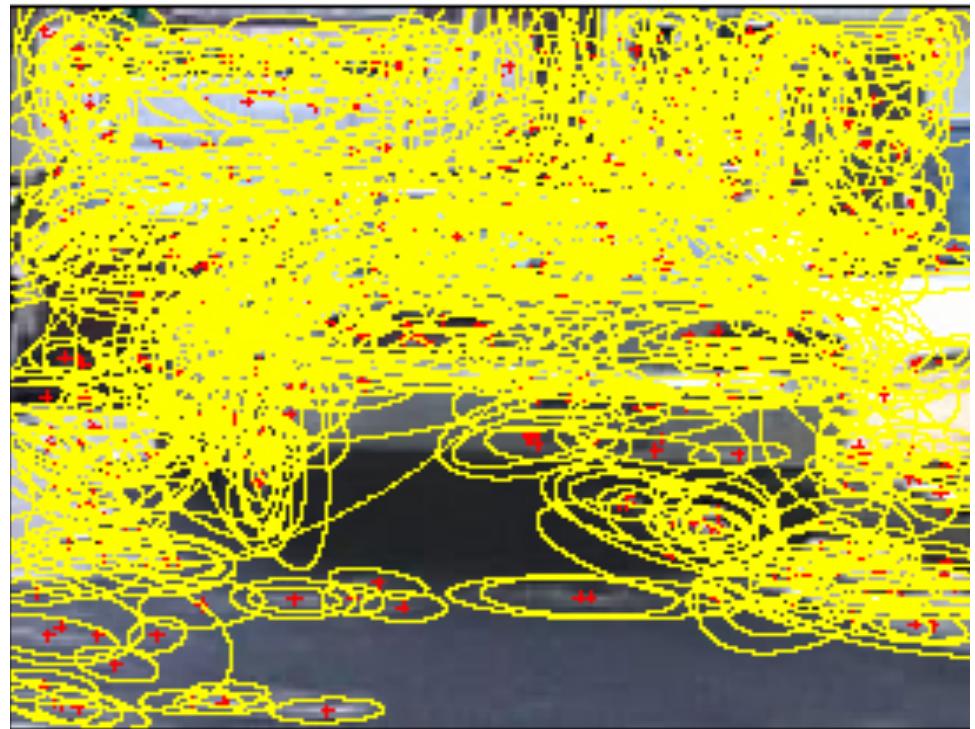
1. Feature detection and description

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



1. Feature detection and description

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



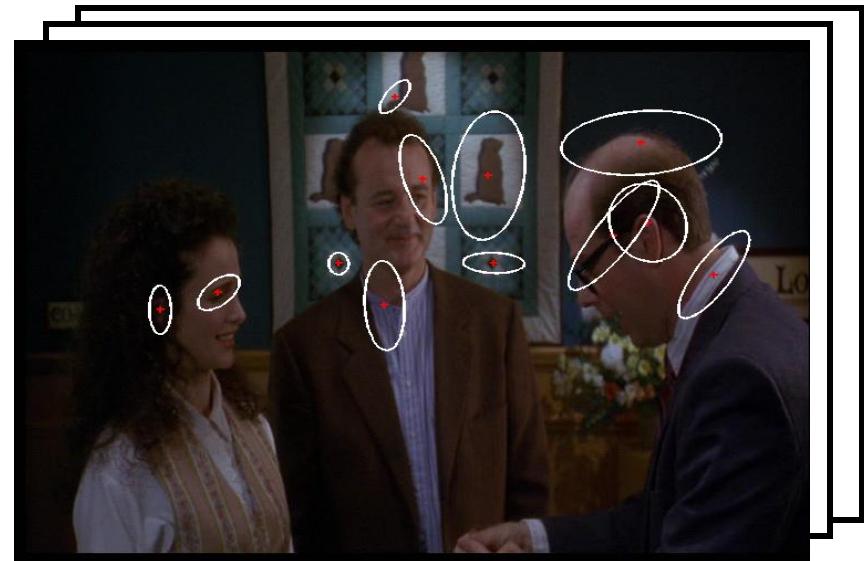
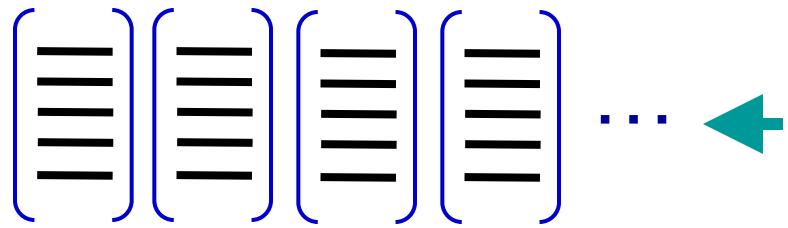
1. Feature detection and description

- Regular grid
 - Vogel & Schiele, 2003
 - Fei-Fei & Perona, 2005
- Interest point detector
 - Csurka, et al. 2004
 - Fei-Fei & Perona, 2005
 - Sivic, et al. 2005
- Other methods
 - Random sampling (Vidal-Naquet & Ullman, 2002)
 - Segmentation based patches (Barnard, Duygulu, Forsyth, de Freitas, Blei, Jordan, 2003)

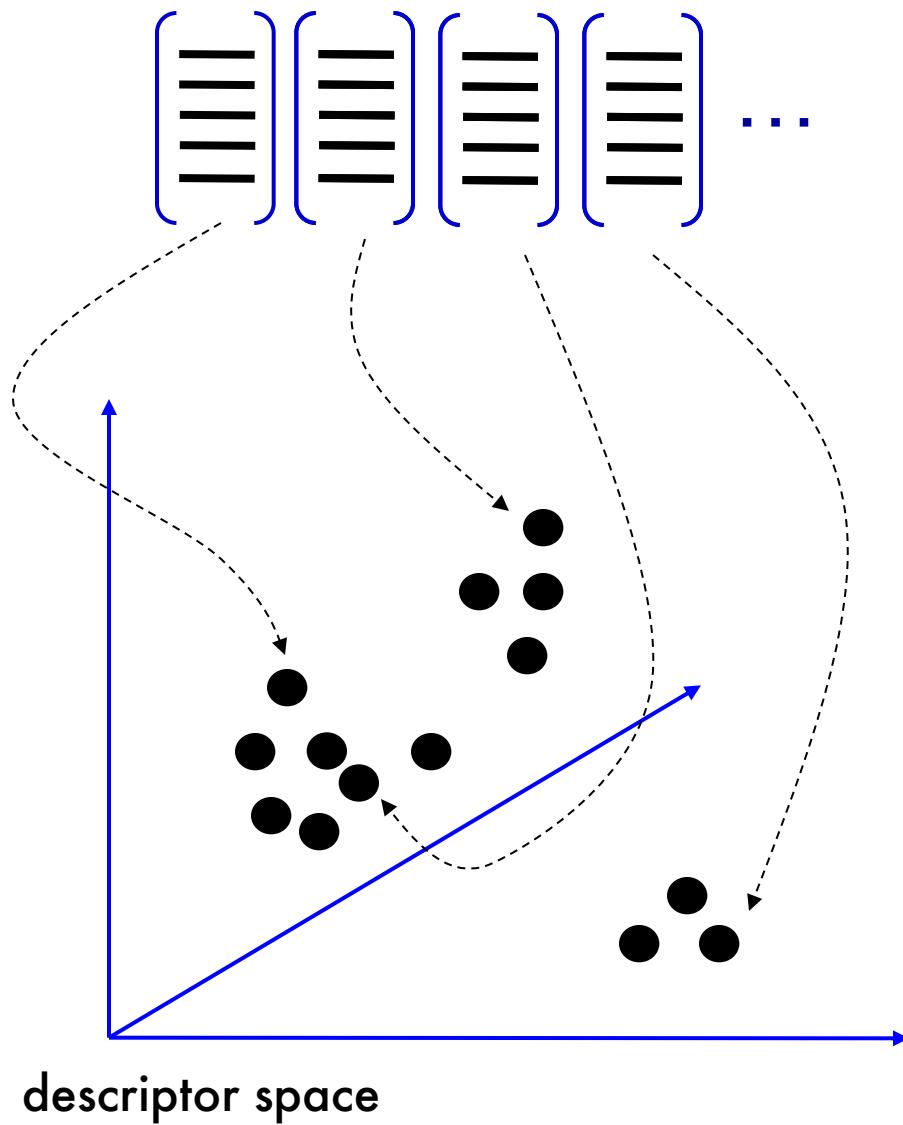
1. Feature detection and description



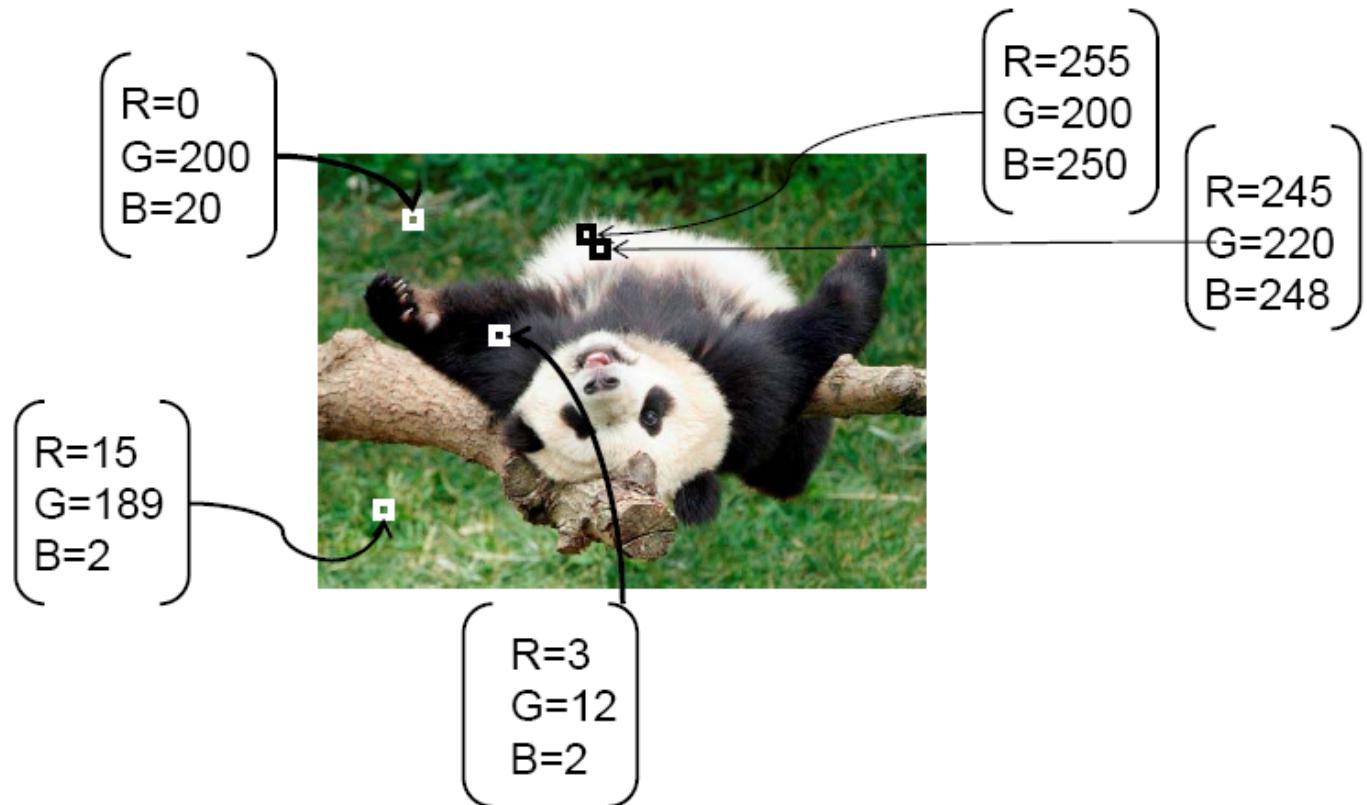
2. Codewords dictionary formation



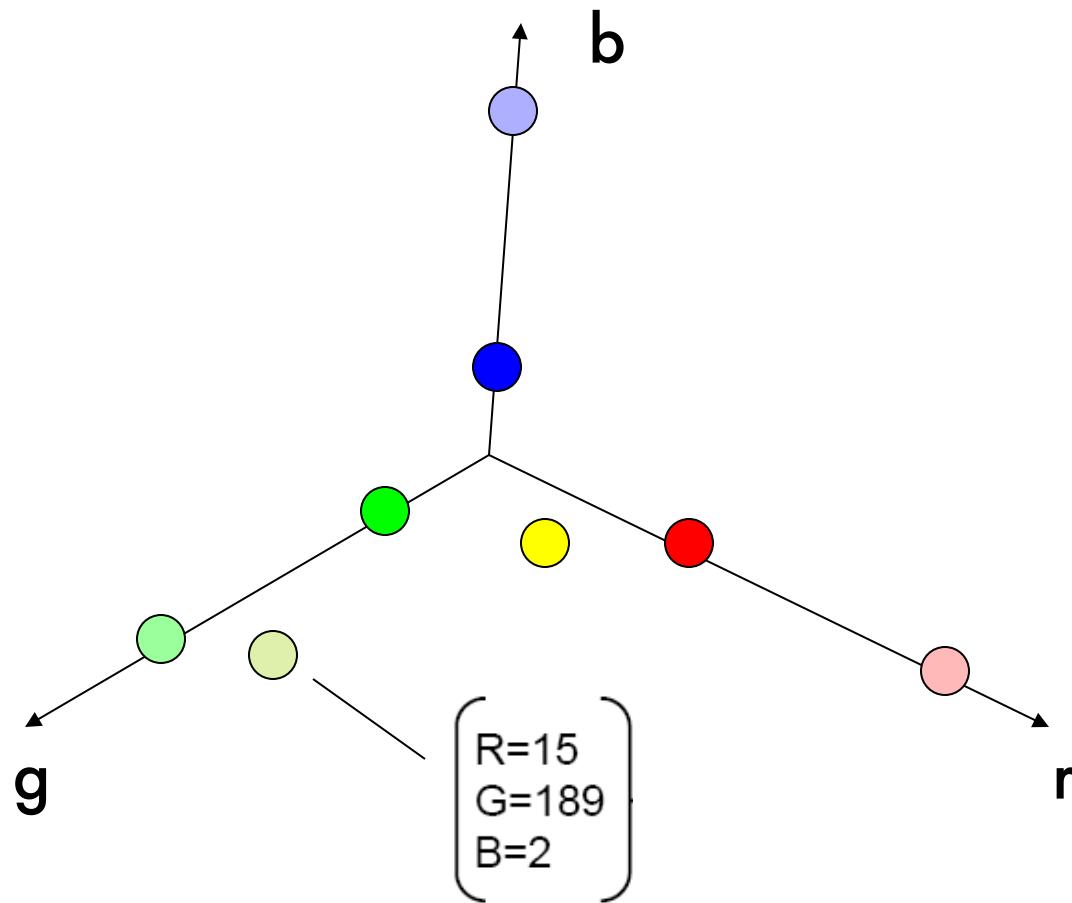
2. Codewords dictionary formation



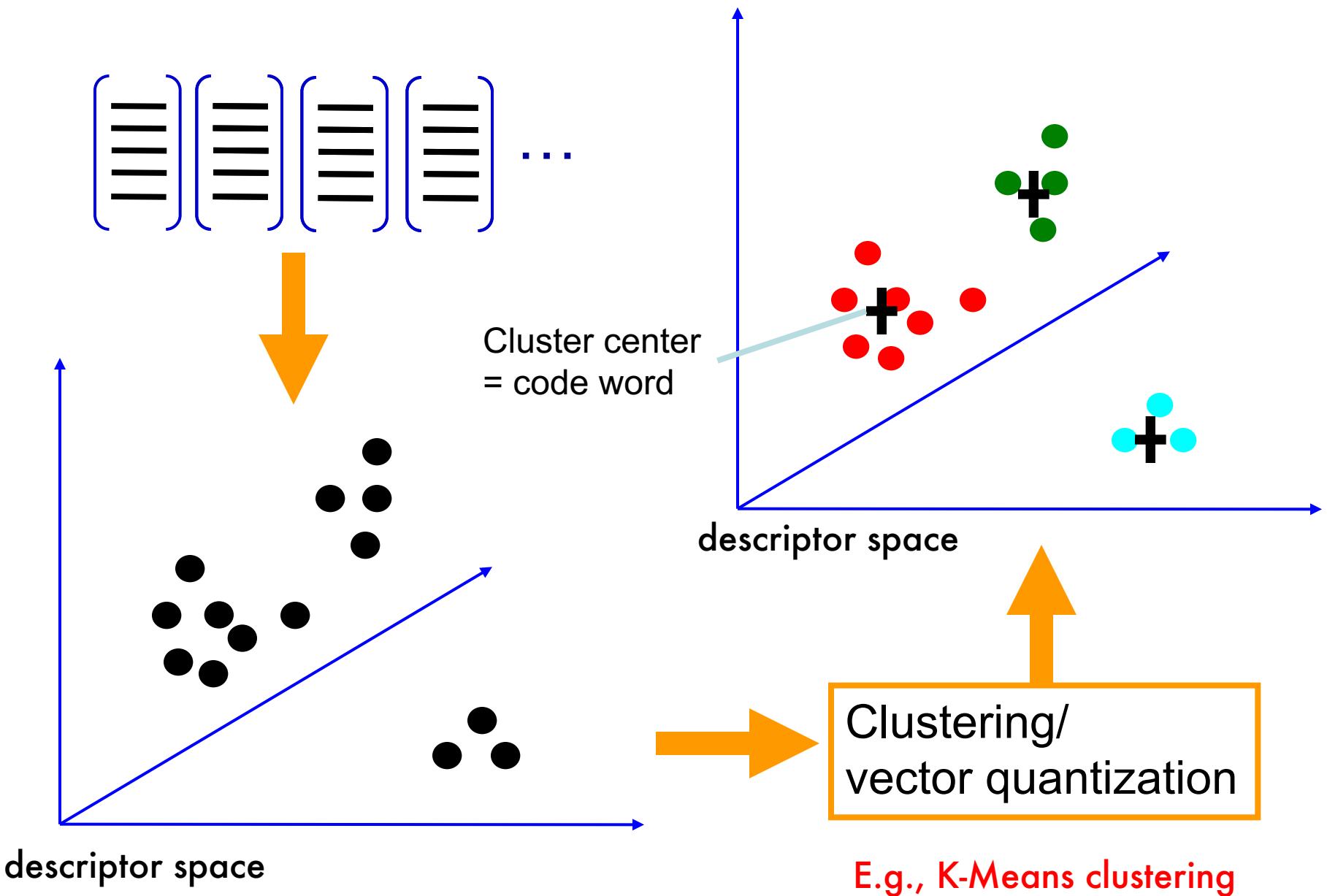
Example: color feature

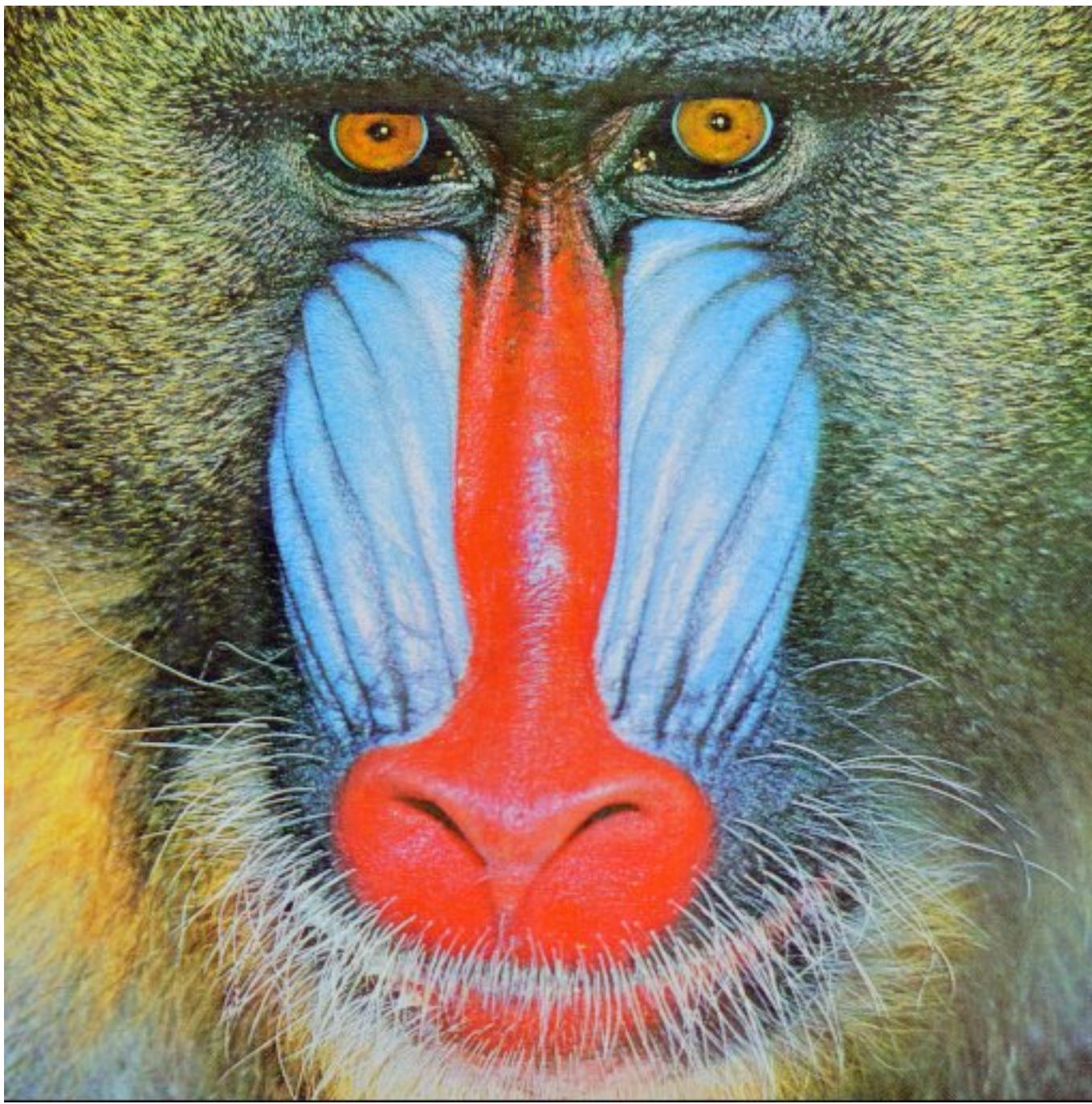


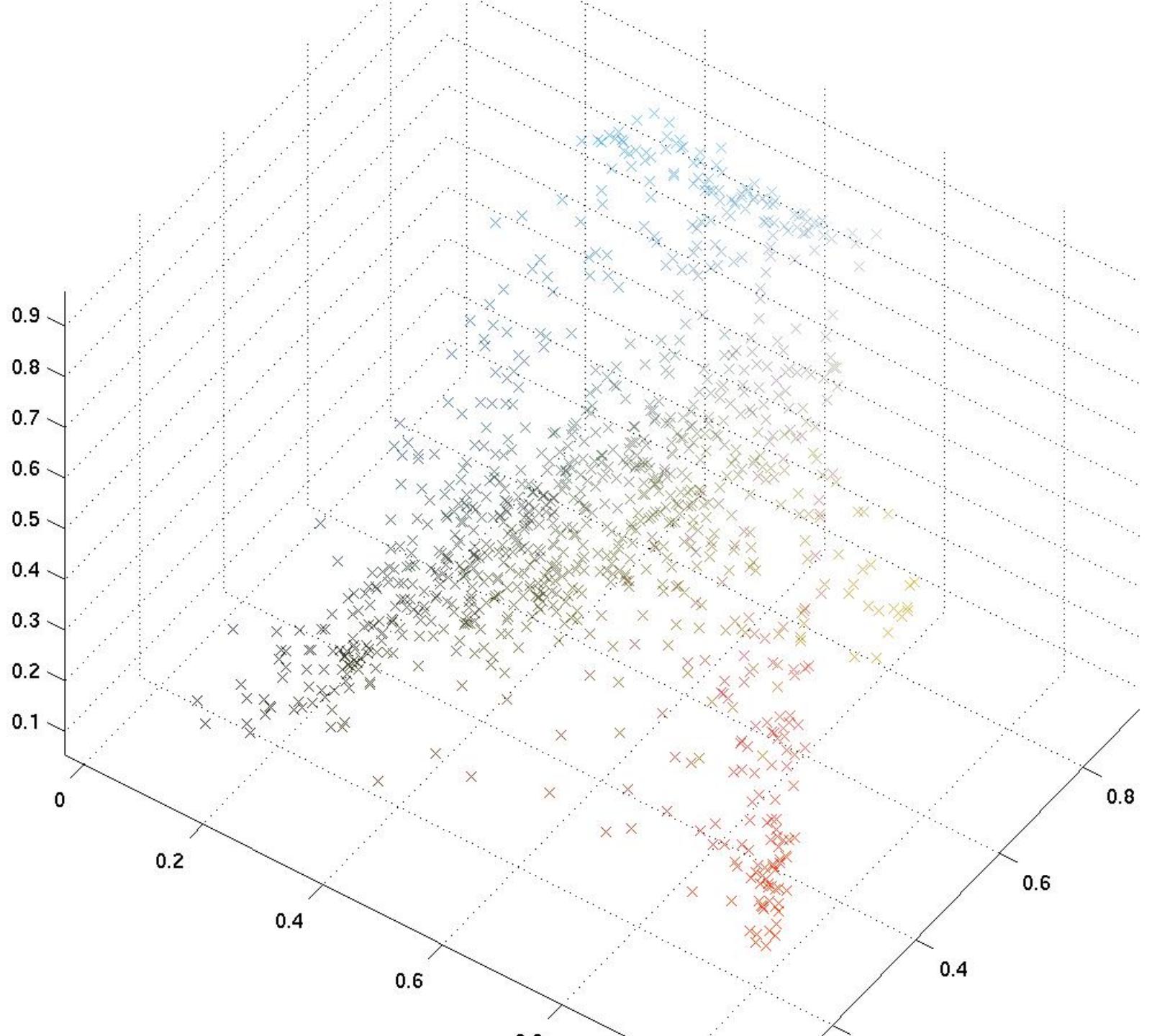
Example: color feature

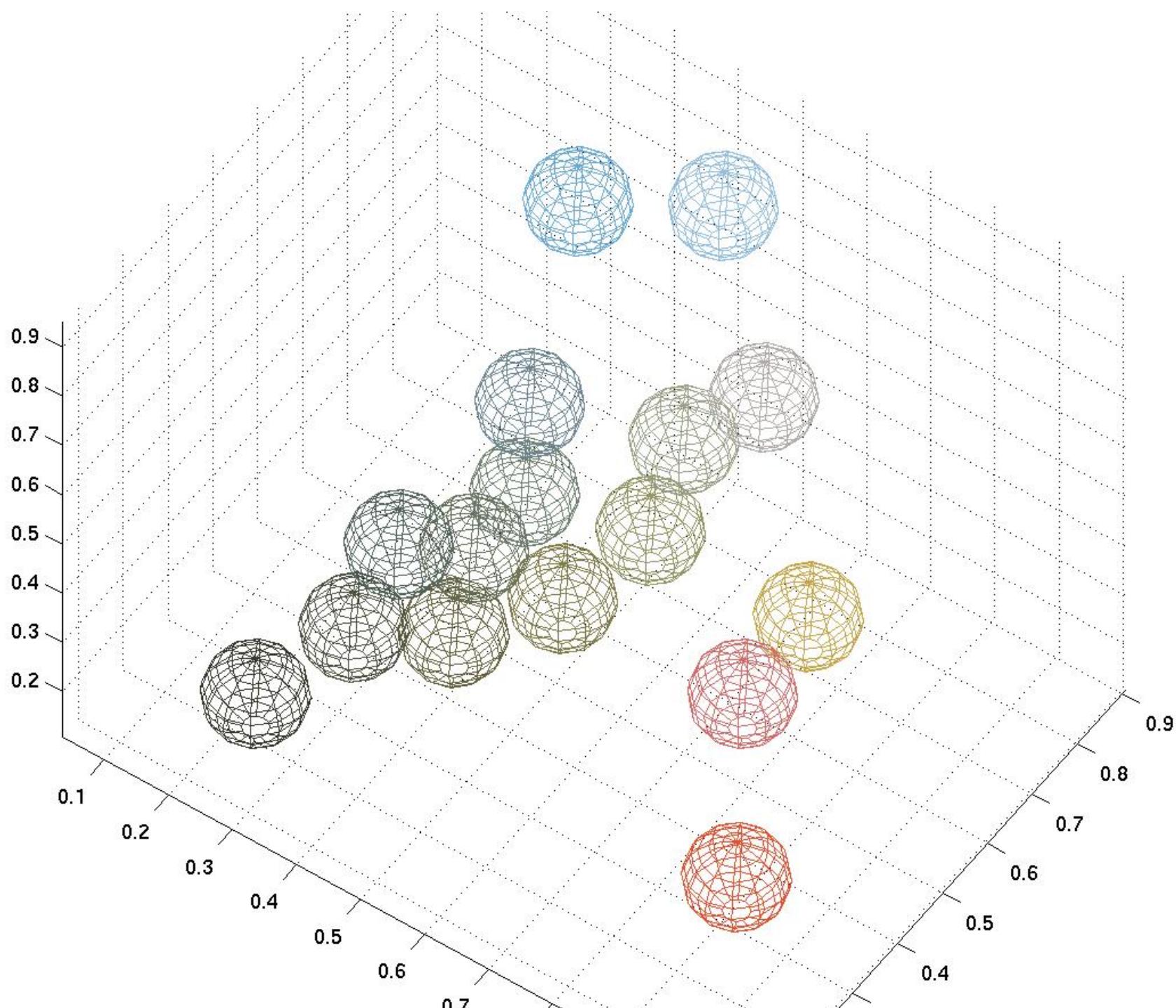


2. Codewords dictionary formation



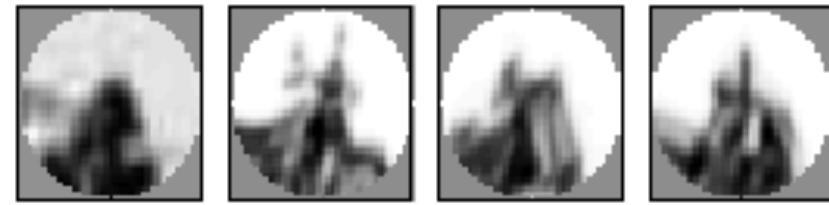
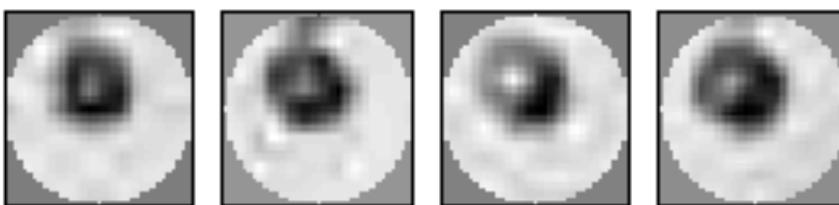
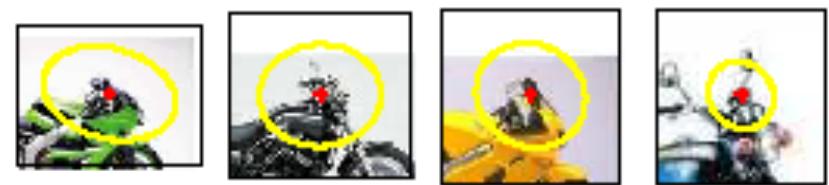




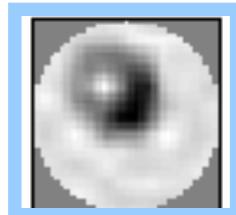


2. Codewords dictionary formation

Image patch examples of codewords

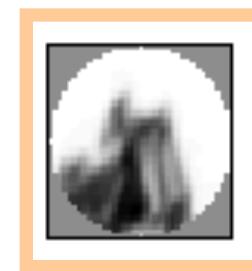


Cluster



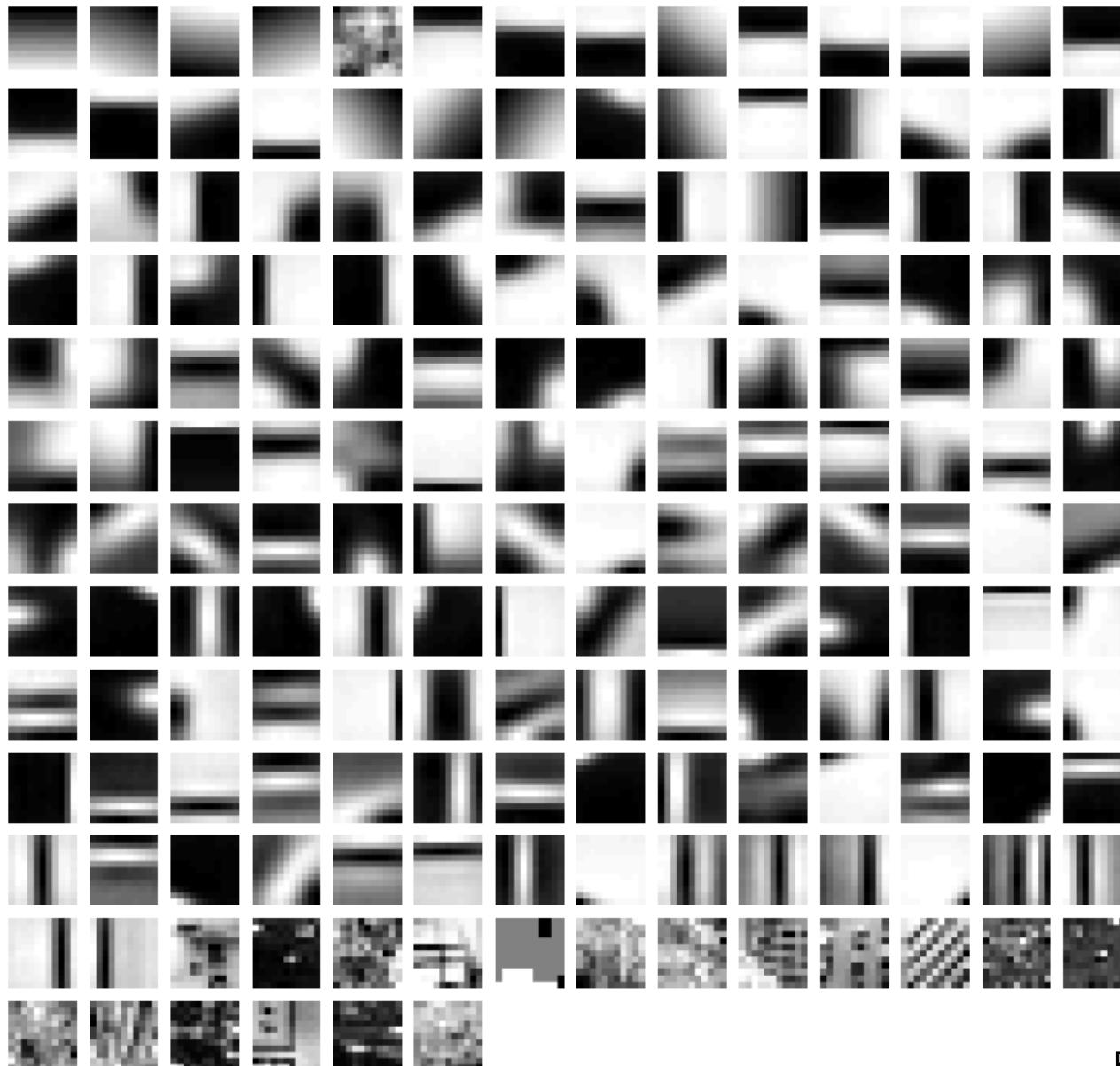
code word

Cluster

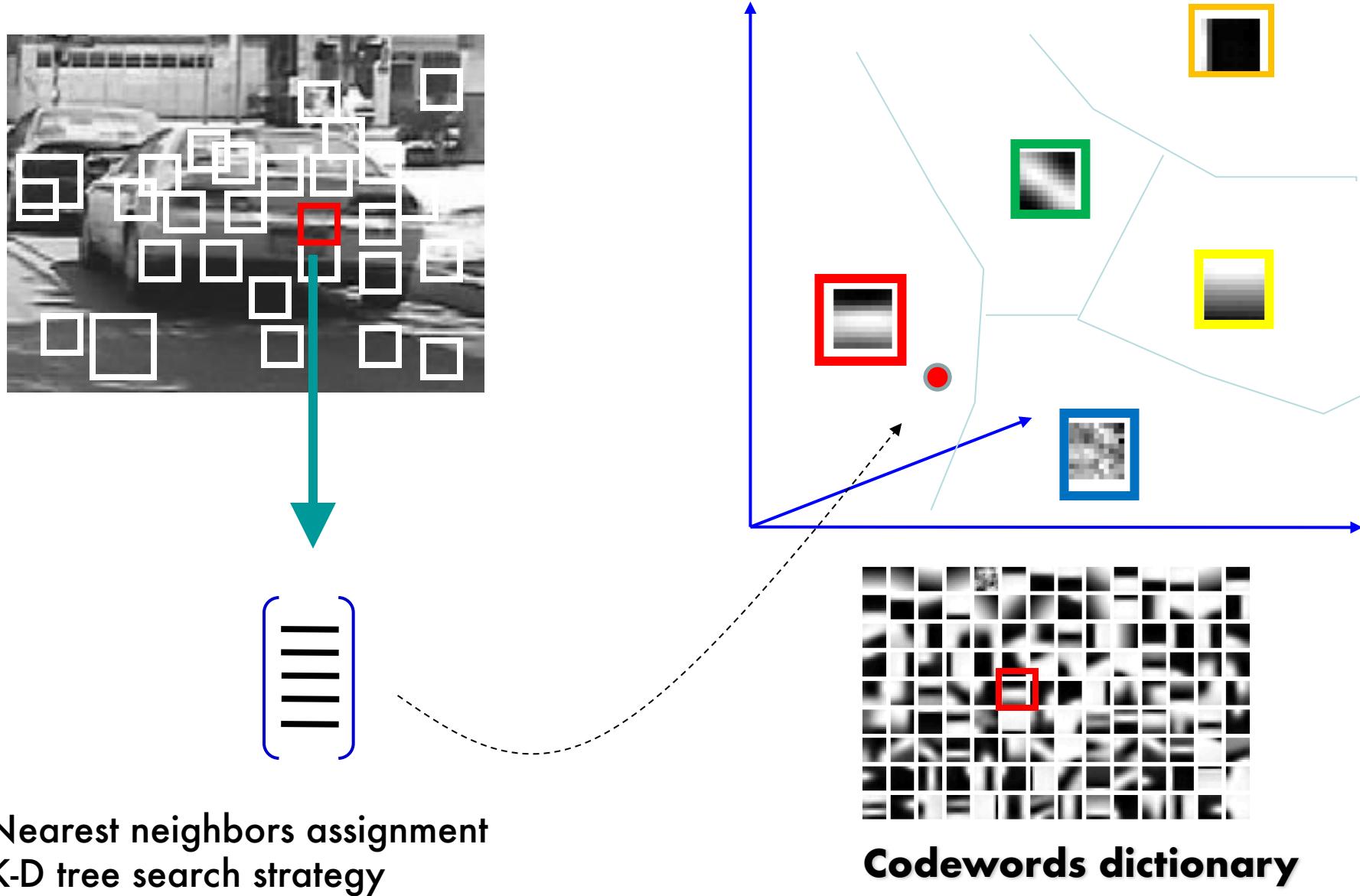


code word

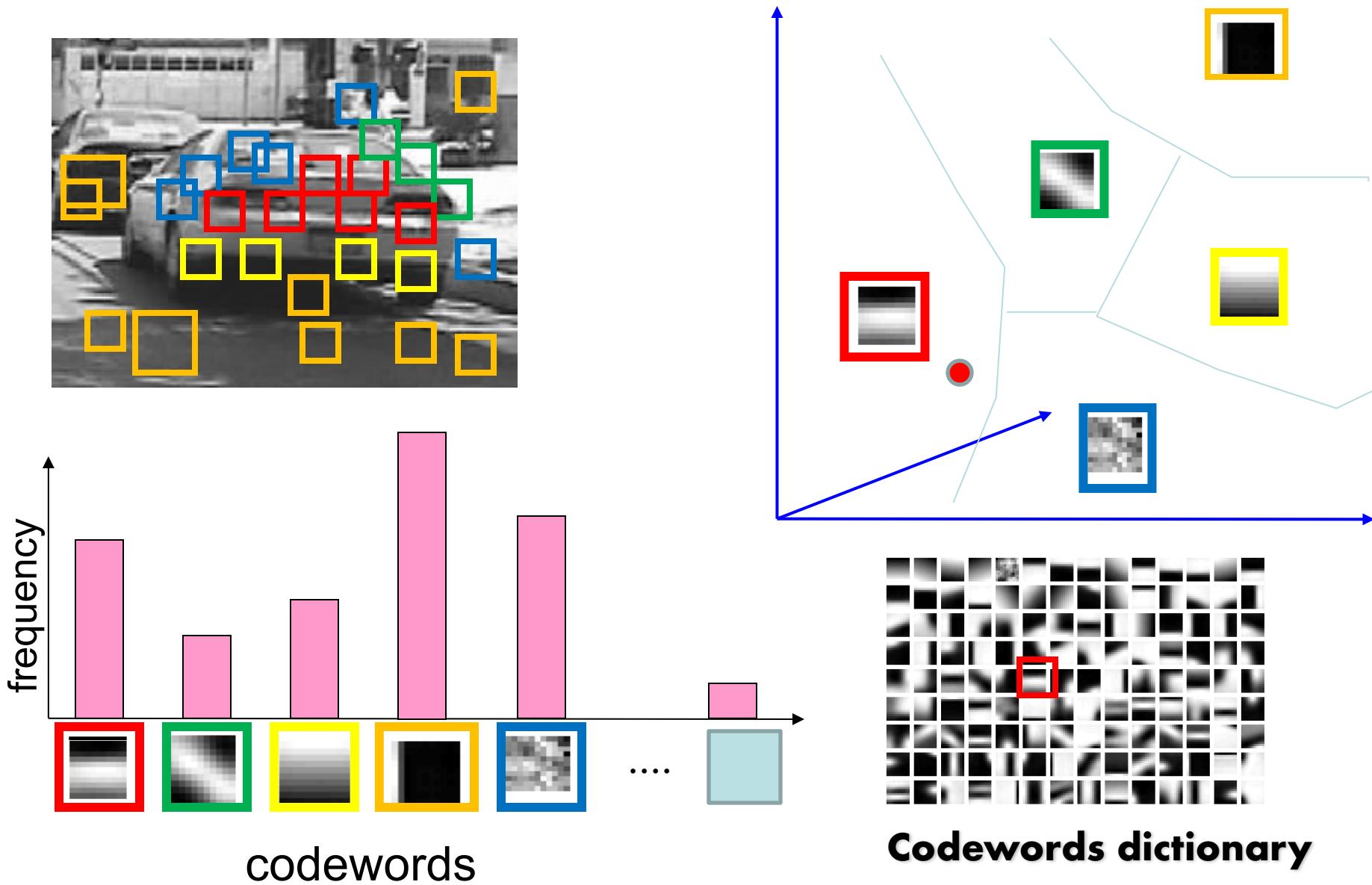
2. Codewords dictionary formation



3. Bag of word representation

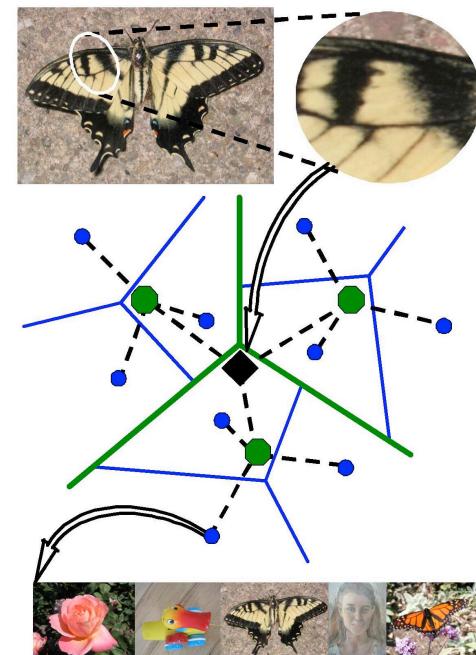


3. Bag of word representation



Visual vocabularies: Issues

- How to choose vocabulary size?
 - Too small: visual words not representative of the object appearance distribution
 - Too large: quantization artifacts, sparse histograms, overfitting
- Computational efficiency
 - Vocabulary trees
(Nister & Stewenius, 2006)



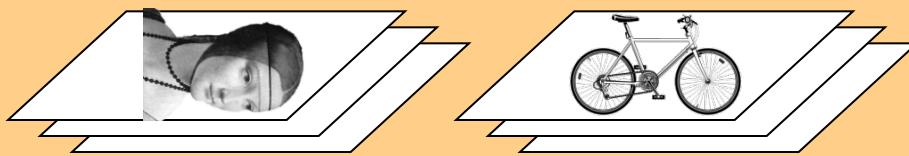
Invariance issues

- Scale? Rotation? View point? Occlusions?
 - Implicit
 - Depends on detectors and descriptors



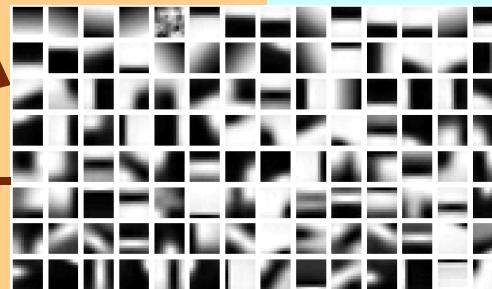
Kadir and Brady. 2003

Representation



1. feature detection
& representation

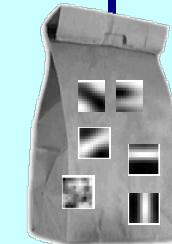
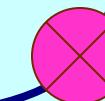
2. codewords dictionary



3. image representation



recognition



learning

**category models
(and/or) classifiers**

**category
decision**

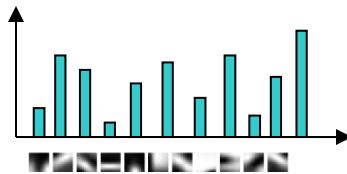
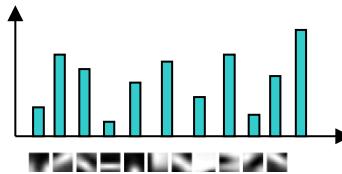
Category models



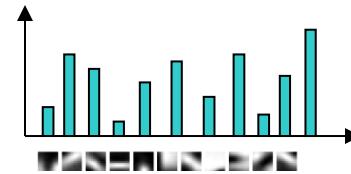
...



Class 1



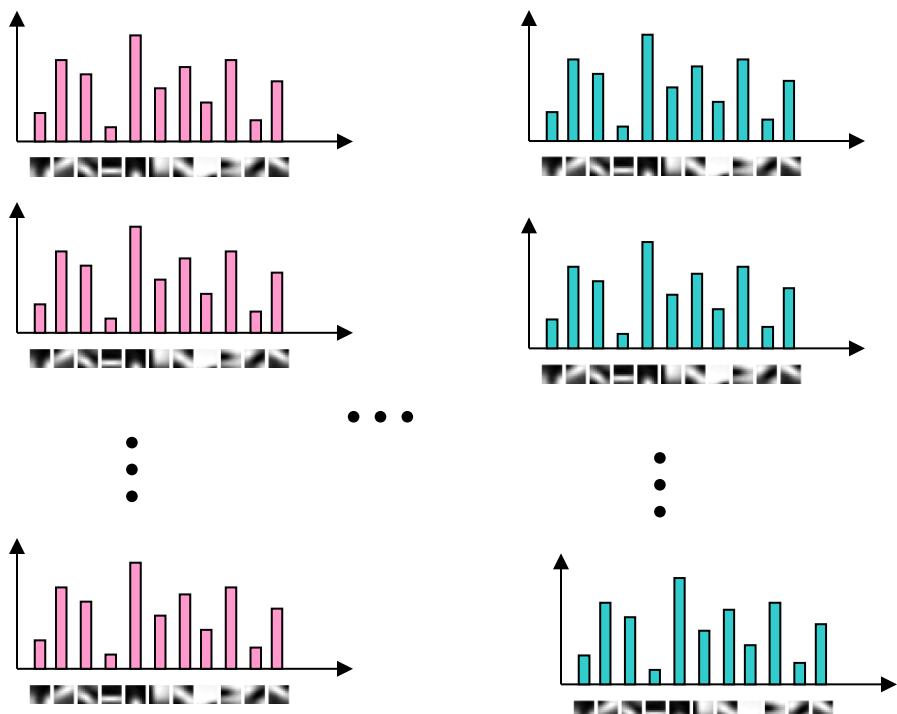
...



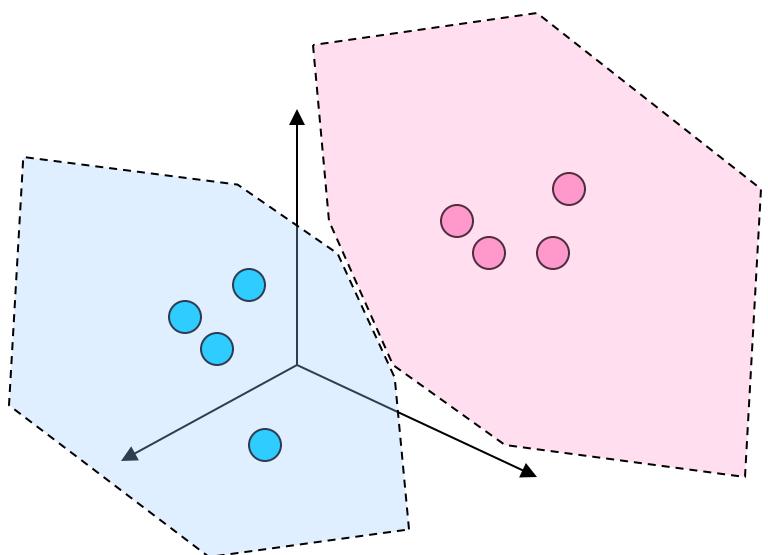
Class N

Discriminative classifiers

category models

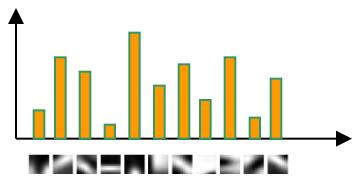


model space



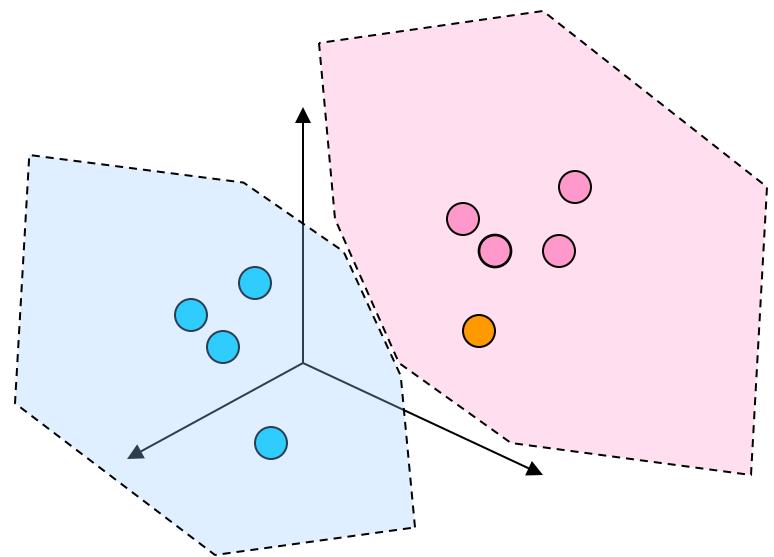
Discriminative classifiers

Query image



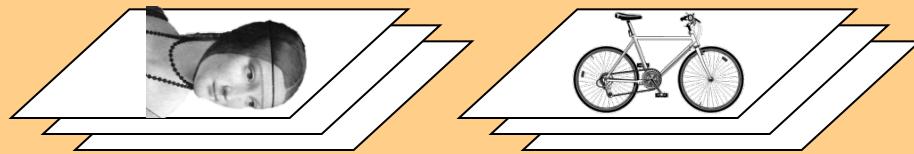
Winning class: pink

model space



- Nearest neighbors
- Linear classifier
- SVM

Representation



1. feature detection & representation

2.
codewords dictionary

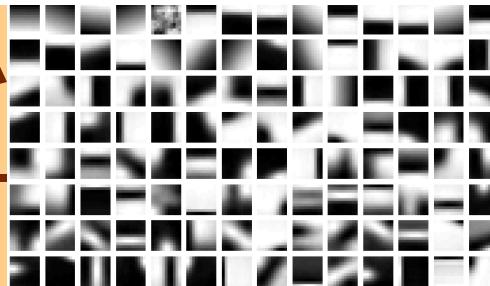
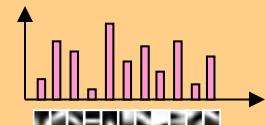
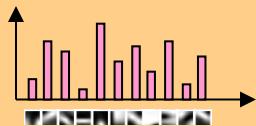


image representation

- 3.



category models

Next Lecture

- Object classification – BoW models part 2
- 2D object detection

Appendix

Object categorization: the statistical viewpoint



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

vs.

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

- **Bayes rule:**

$$p(A|B) = \frac{p(B|A) p(A)}{p(B)}$$

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

Object categorization: the statistical viewpoint



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

vs.

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

- **Bayes rule:**

$$p(A|B) = \frac{p(B|A) p(A)}{p(B)}$$

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

posterior ratio

likelihood ratio

prior ratio

Object categorization: the statistical viewpoint

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

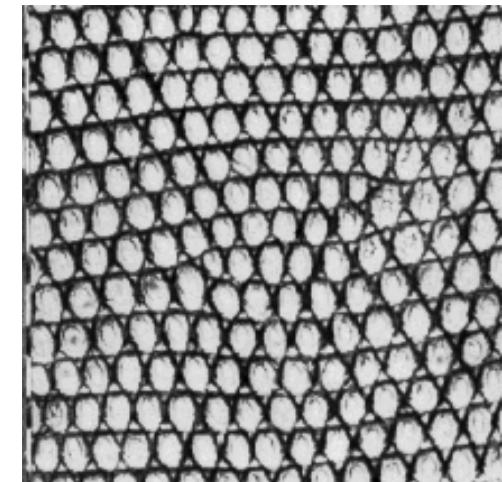
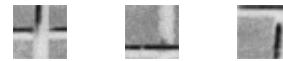
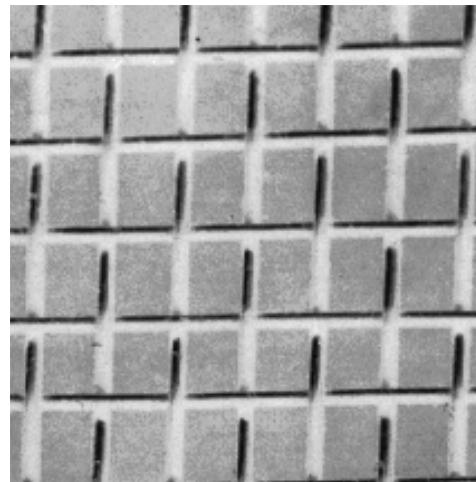
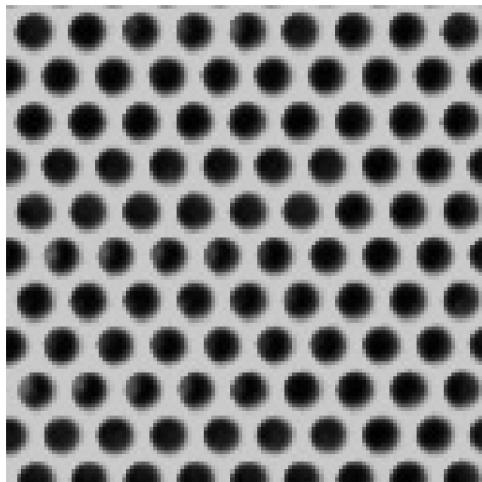
- **Bayes rule:**

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

posterior ratio likelihood ratio prior ratio

Representing textures

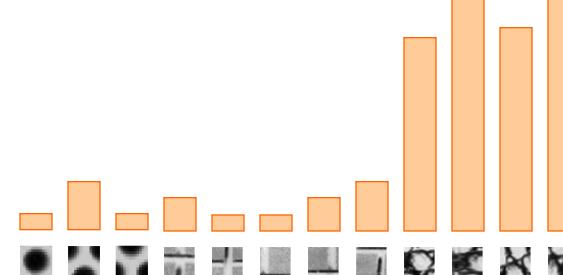
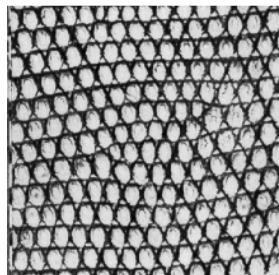
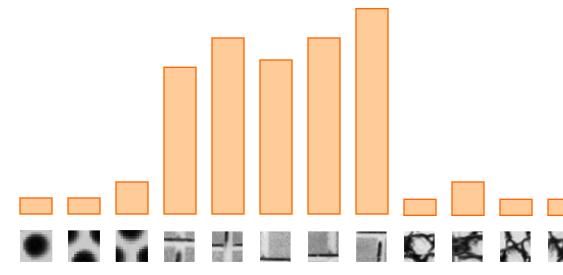
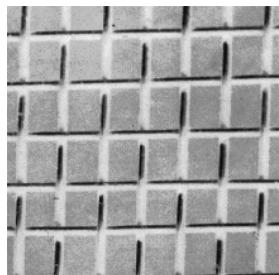
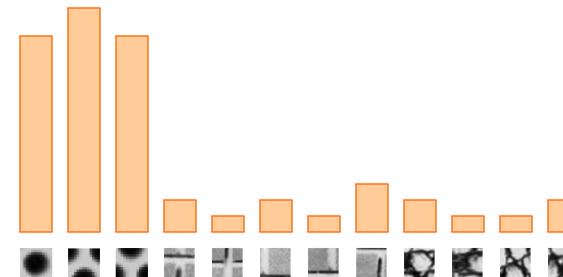
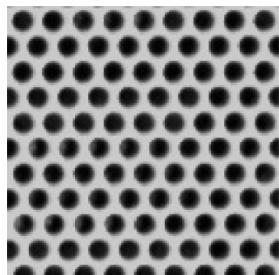
- Texture is characterized by the repetition of basic elements or *textons*
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Credit slide: S. Lazebnik

Representing textures



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Credit slide: S. Lazebnik