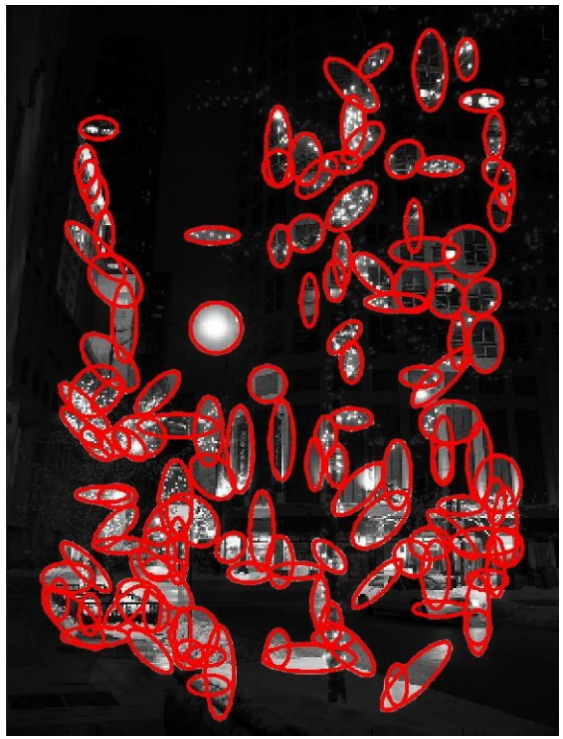
Lecture 11
Visual recognition

- 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



Feature Detection e.g. DoG

Feature Description e.g. SIFT

- Estimation
- Matching
- Indexing
- Recognition

What's visual recognition?

Classification:

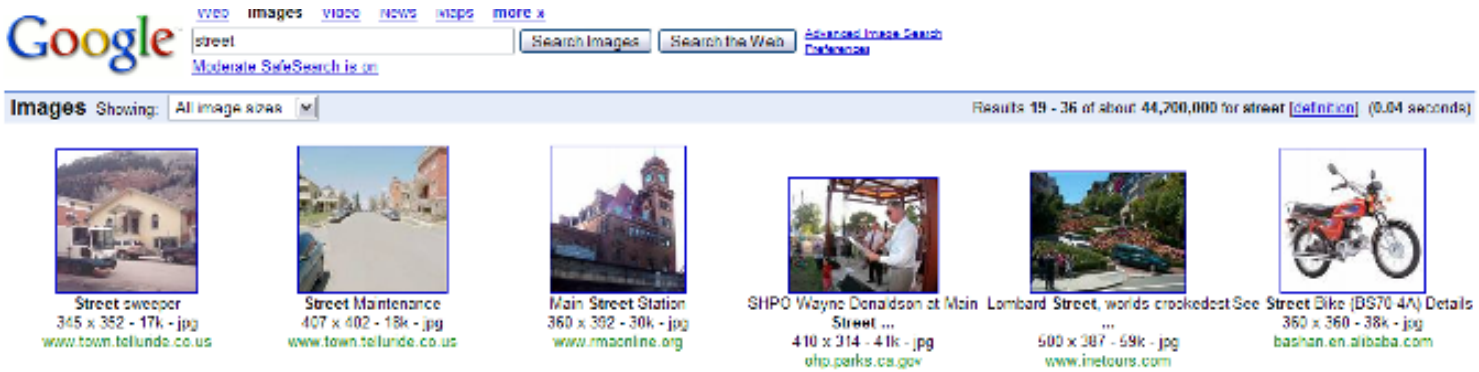
Does this image contain a building? [yes/no]

Classification:

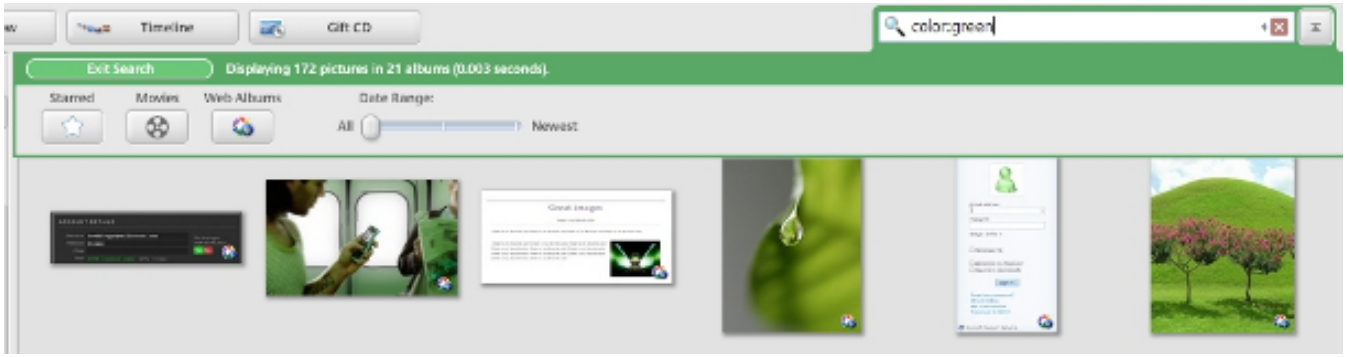
Is this an beach?

Image Search or Indexing

flickr webshots Ask Images Cydral picsearch Google Image Search altavista



Organizing photo collections

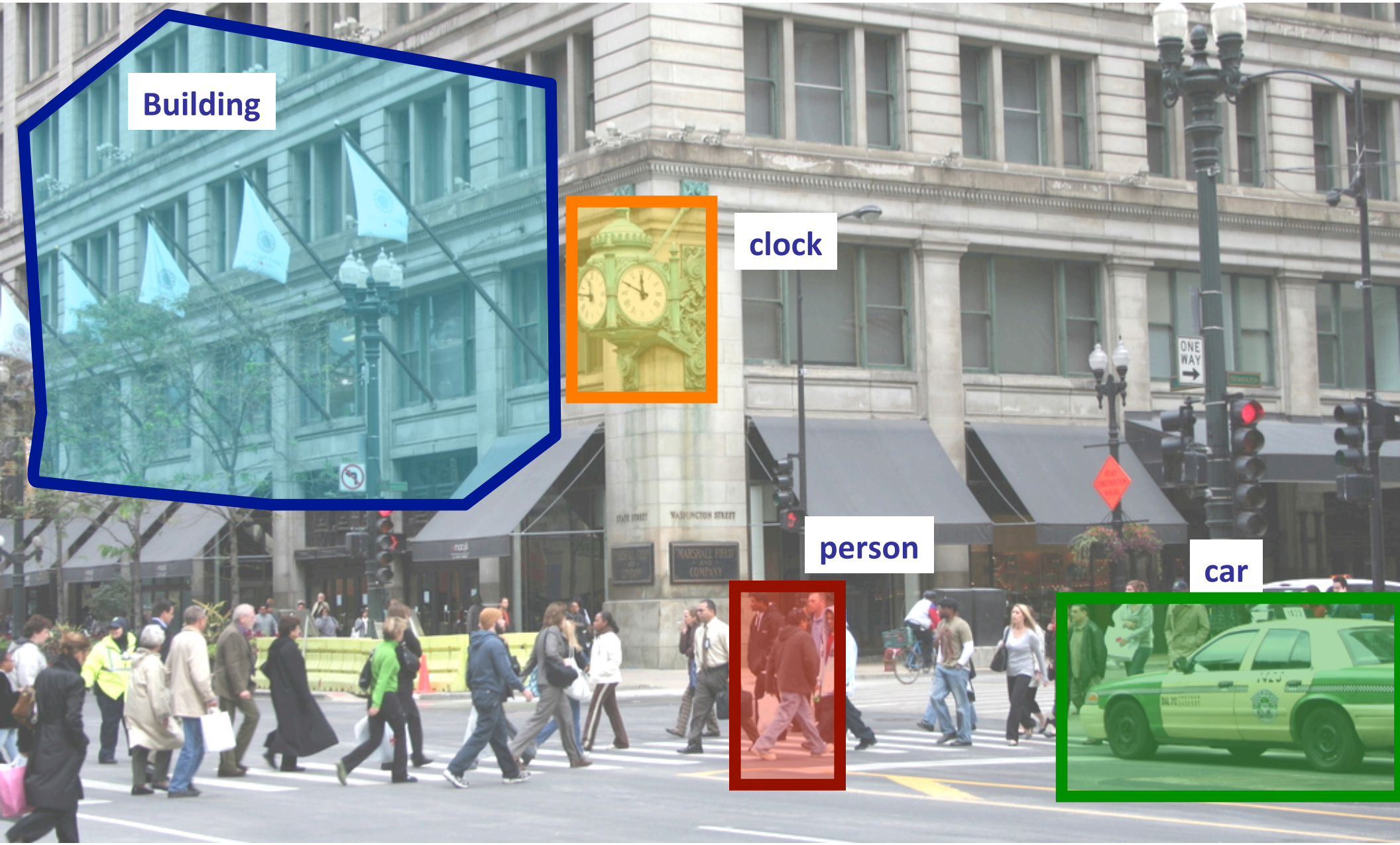


Detection:

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

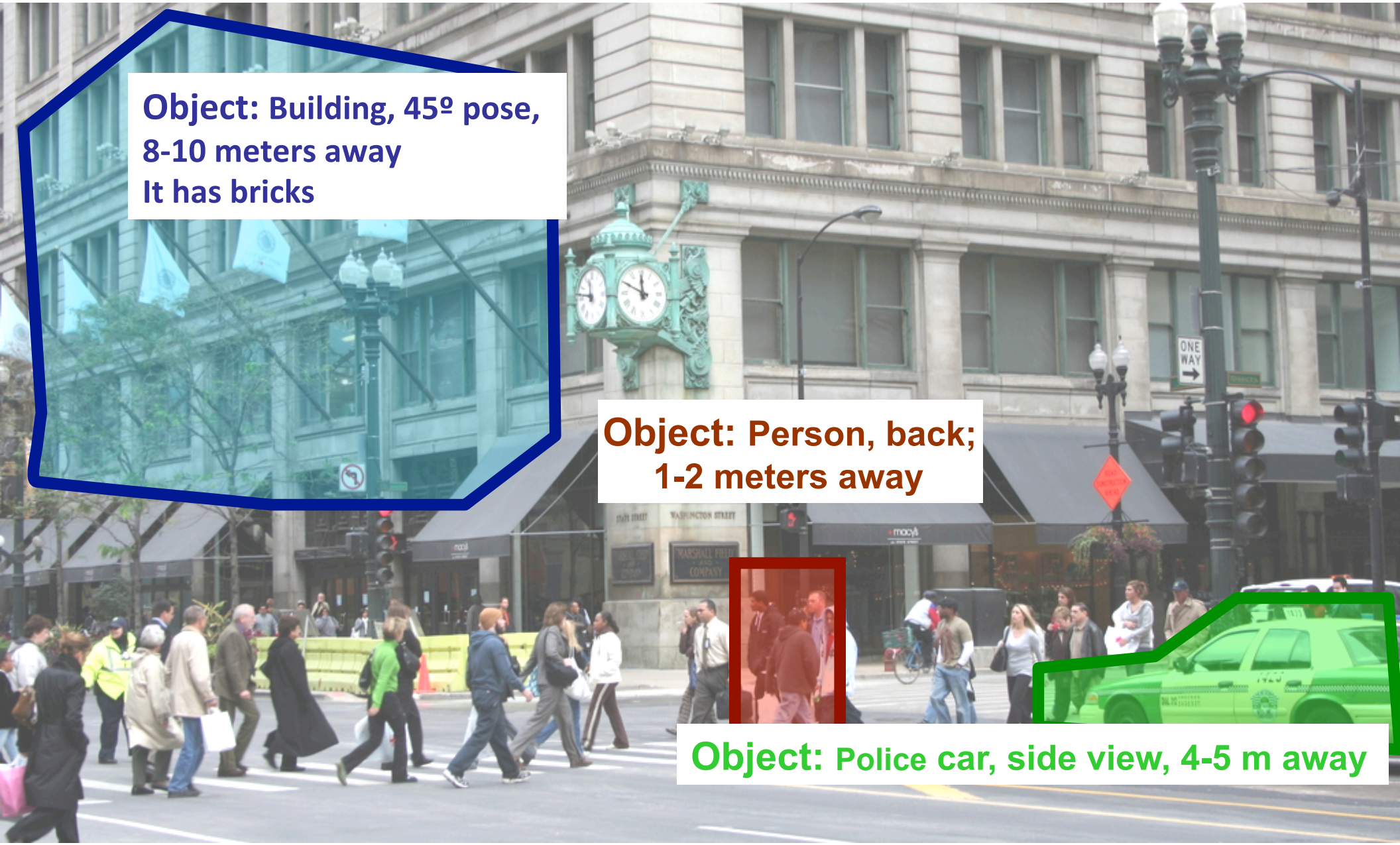
Which building is this? *Marshall Field* building in Chicago

Categorization vs Single instance recognition

Where is the crunchy nut?

Recognizing landmarks in mobile platforms

Detection: Estimating object semantic & geometric attributes



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?

Challenges: viewpoint variation

Michelangelo 1475-1564

slide credit: Fei-Fei, Fergus & Torralba

Challenges: illumination

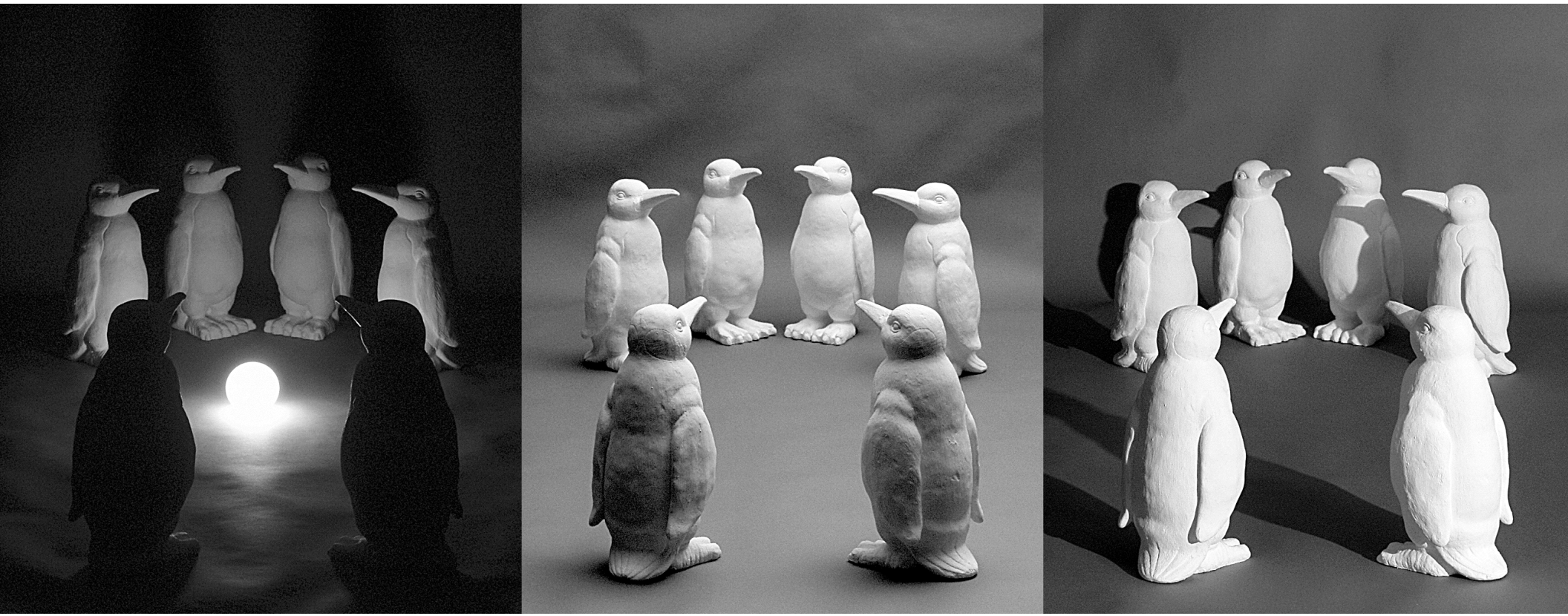


image credit: J. Koenderink

Challenges: scale

slide credit: Fei-Fei, Fergus & Torralba

Challenges: deformation

Challenges: occlusion

Magritte, 1957

slide credit: Fei-Fei, Fergus & Torralba

Challenges: background clutter

Kilmeny Niland. 1995

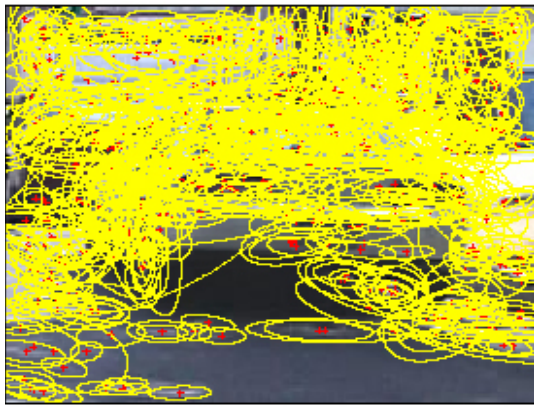
Challenges: intra-class variation

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

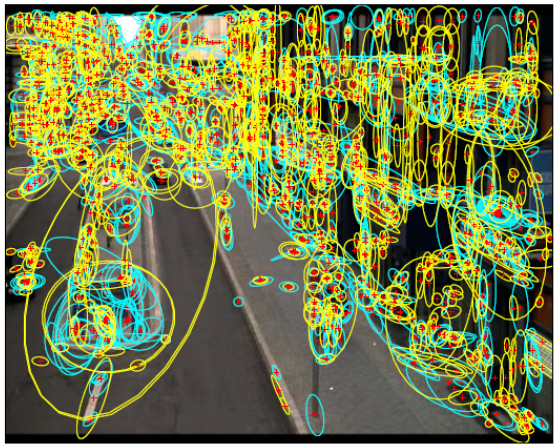
Representation

- Building blocks: Sampling strategies



Interest operators

Dense, uniformly



Multiple interest operators

Randomly

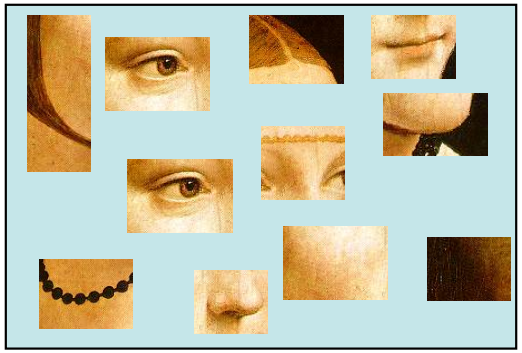
Image credits: F-F. Li, E. Nowak, J. Sivic

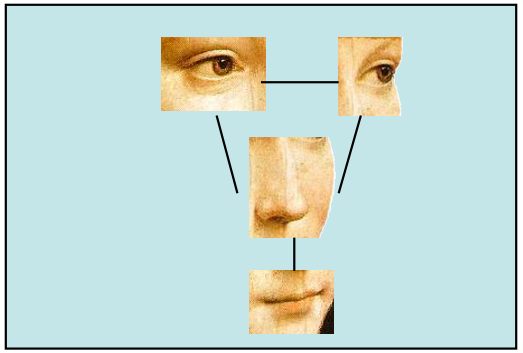
Representation

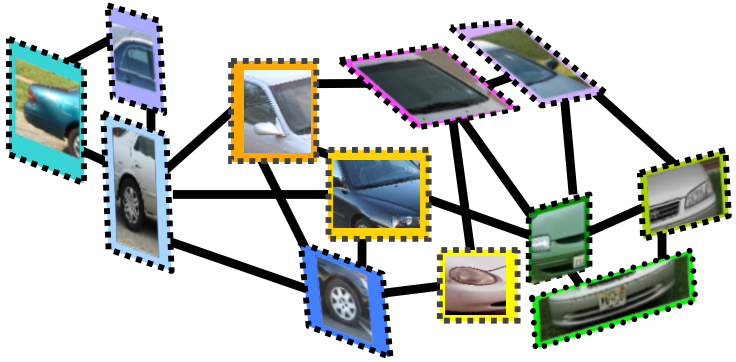
- Building blocks: Choice of descriptors
  [SIFT, HOG, codewords….]

Representation

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

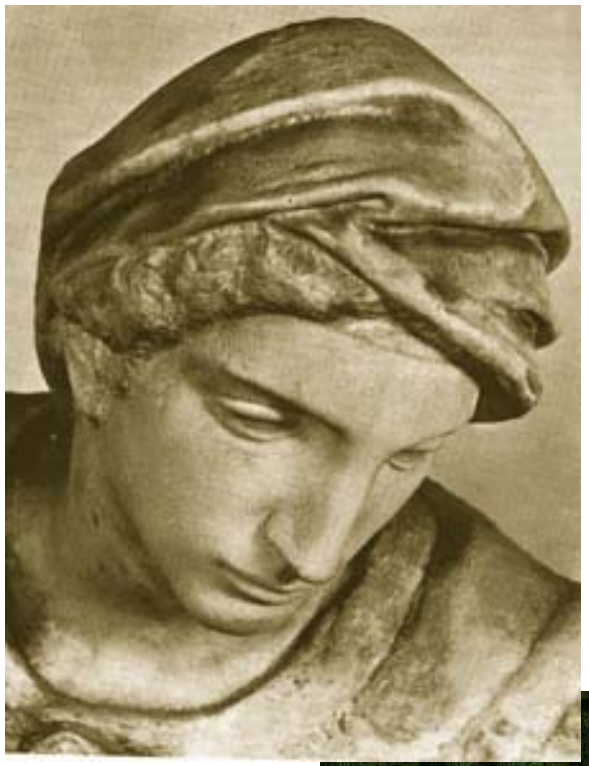


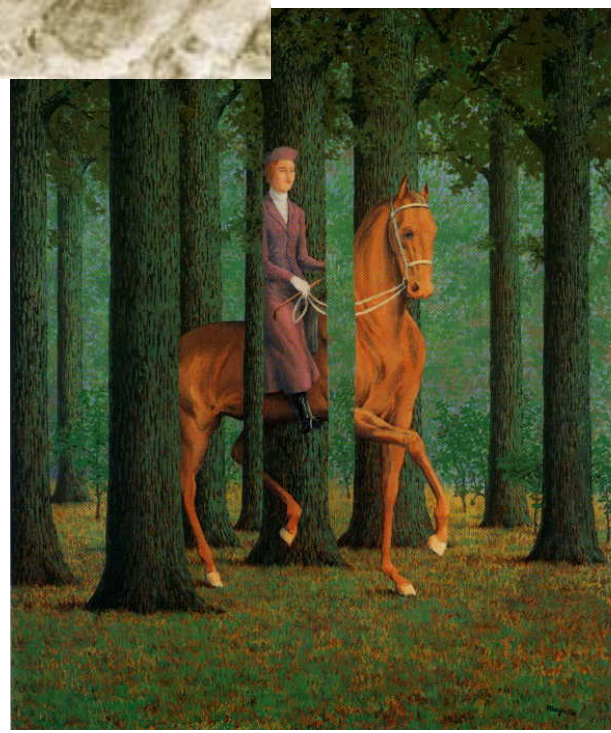




Representation

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



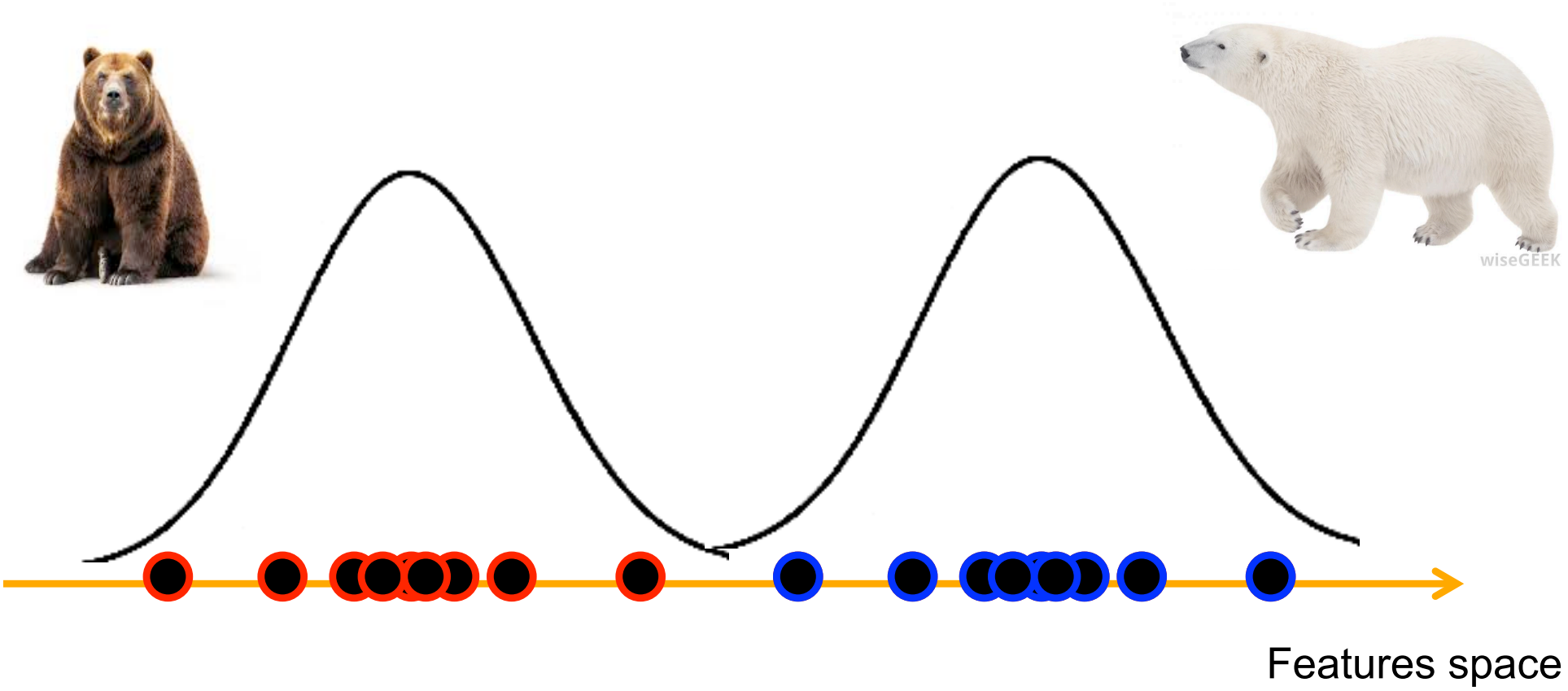


Representation

- How to handle intra-class variability?
  - → It is convenient to describe object categories using probabilistic models
  - → Generative – vs – discriminative

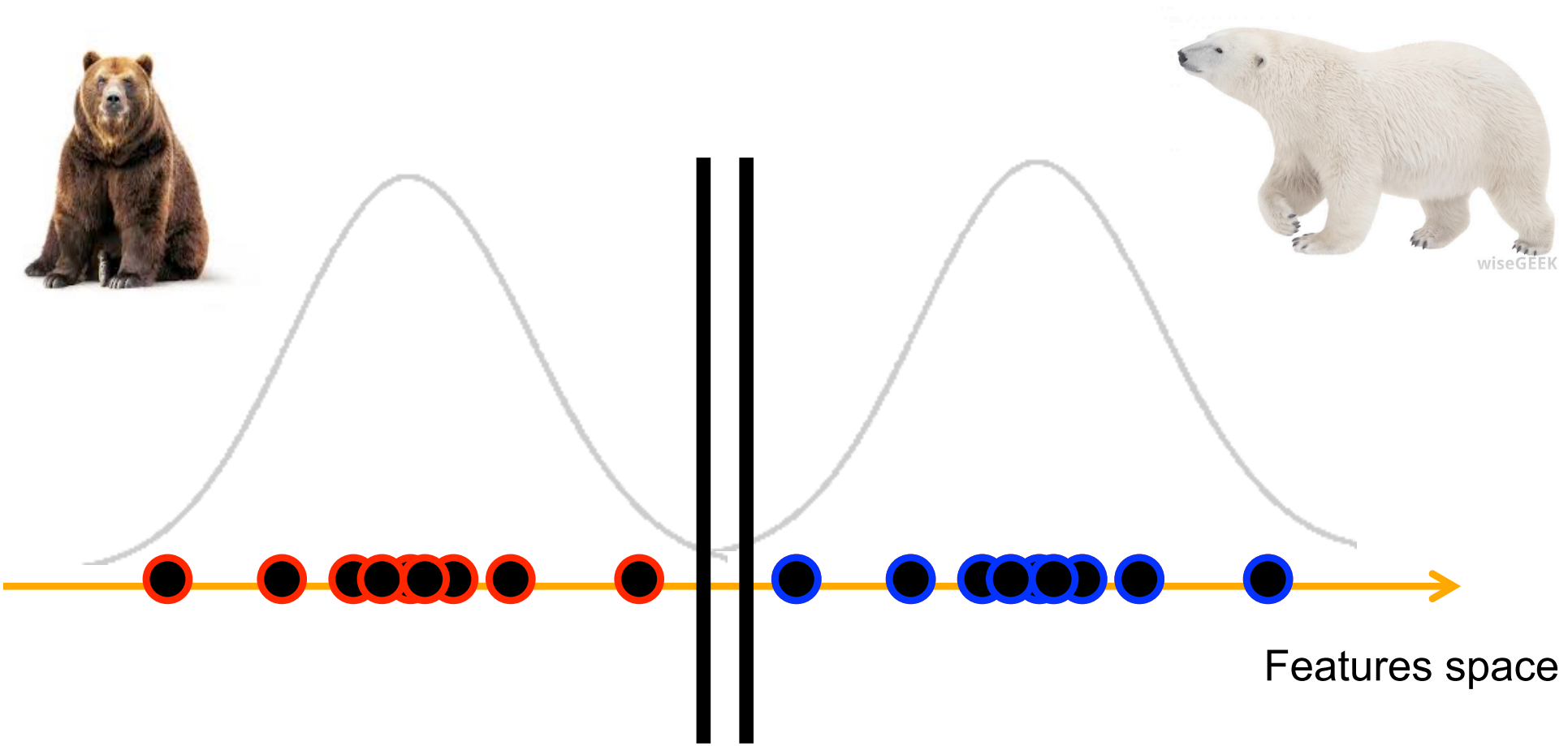
Generative – vs – discriminative

- 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 Part based models
  - Constellation models: Weber et al 2000; Fergus et al 200
  - Star models: ISM (Leibe et al 05)
- 3D part based models:
  - multi-aspects: Sun, et al, 2009

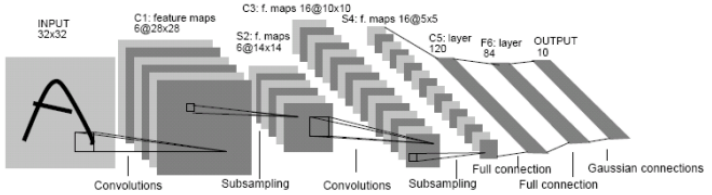
Discriminative models

Nearest neighbor

10^6 examples

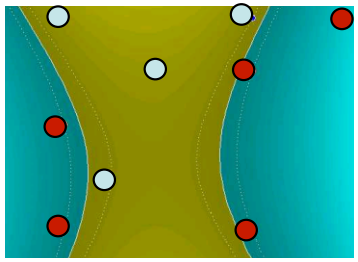
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

Neural networks



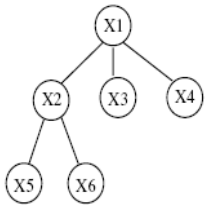
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines



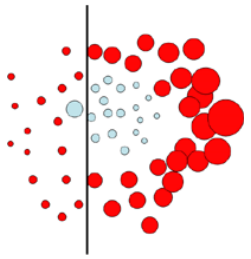
Guyon, Vapnik, Heisele, Serre, Poggio…

Latent SVM
Structural SVM



Felzenszwalb 00
Ramanan 03…

Boosting



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

Courtesy of Vittorio Ferrari
Slide credit: Kristen Grauman

Slide adapted from Antonio Torralba

Basic properties

- Representation
  - How to represent an object category; which classification scheme?

- Learning
  - How to learn the classifier, given training data

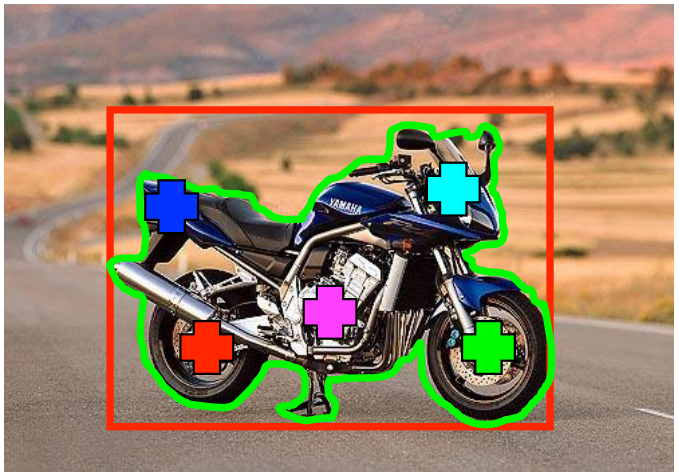
- Recognition
  - How the classifier is to be used on novel data

Learning

- Learning parameters
- Generative functions or separating functions?

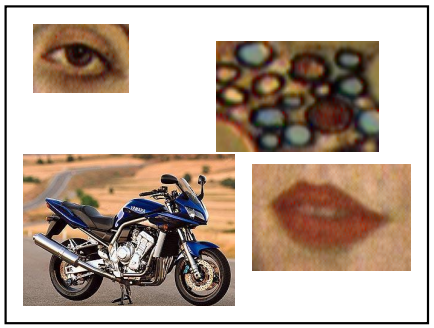
Learning

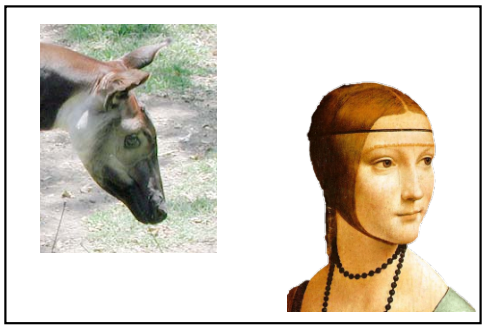
- Learning parameters
- Generative functions or separating functions?
- Level of supervision
  - Noisy labels; image labels; bounding box; manual segmentation; part annotations
- Batch/incremental
- Priors



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:
  - Issue of overfitting
  - Negative images for discriminative methods





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:

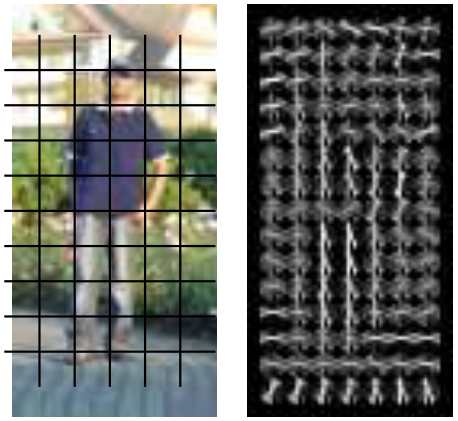
  – Sliding Windows

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

Dalal-Triggs pedestrian detector

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05

Section 17.1 [FP]



Represent an object as a collection of HoG templates

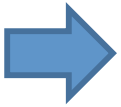
1. Extract fixed-sized window at each position and scale
2. Compute HOG (histogram of gradient) features within each window
3. Score the window with a linear SVM classifier
4. Perform non-maxima suppression to remove overlapping detections with lower scores

Non-max suppression

Section 17.1 [FP]

Non-max suppression

Also: Canny ’86, …. Desai et al , 2009



Recognition

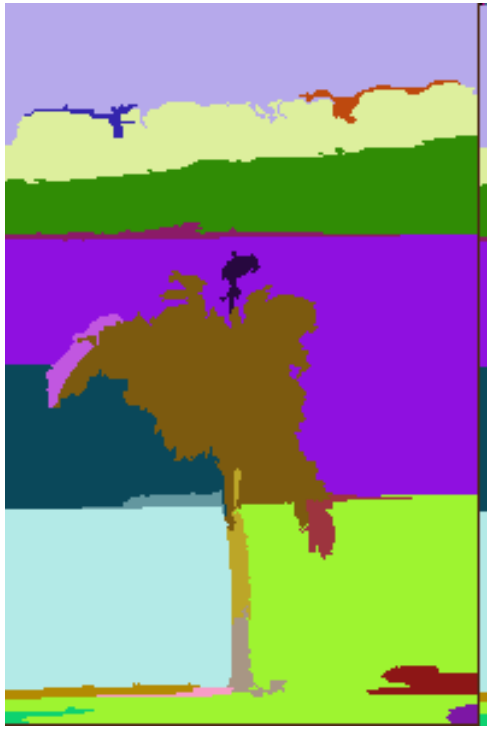
- Recognition task
- Search strategy:
  - Sliding Windows

Simple!

But computational expensive..

Recognition

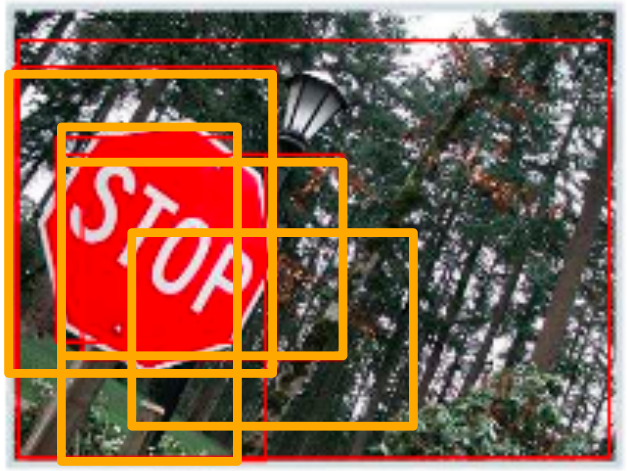
- Recognition task
- Search strategy:
  - Sliding Windows
  - Bottom-up cues (segmentation)

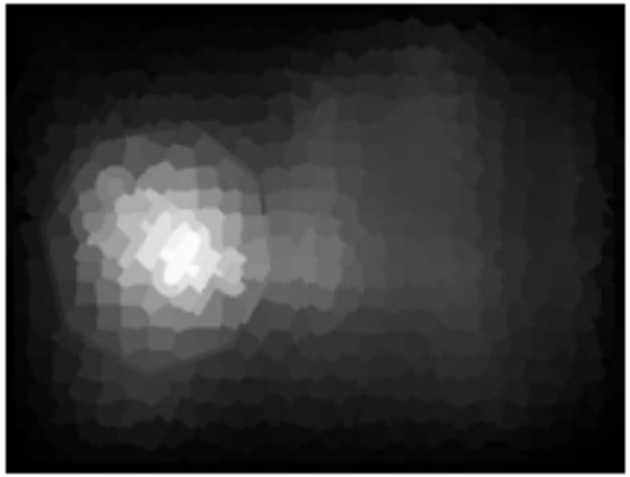


Felzenszwalb and Huttenlocher, 2004

Recognition

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





Jia & Han, 13
Alexe, et al 10
…

Recognition

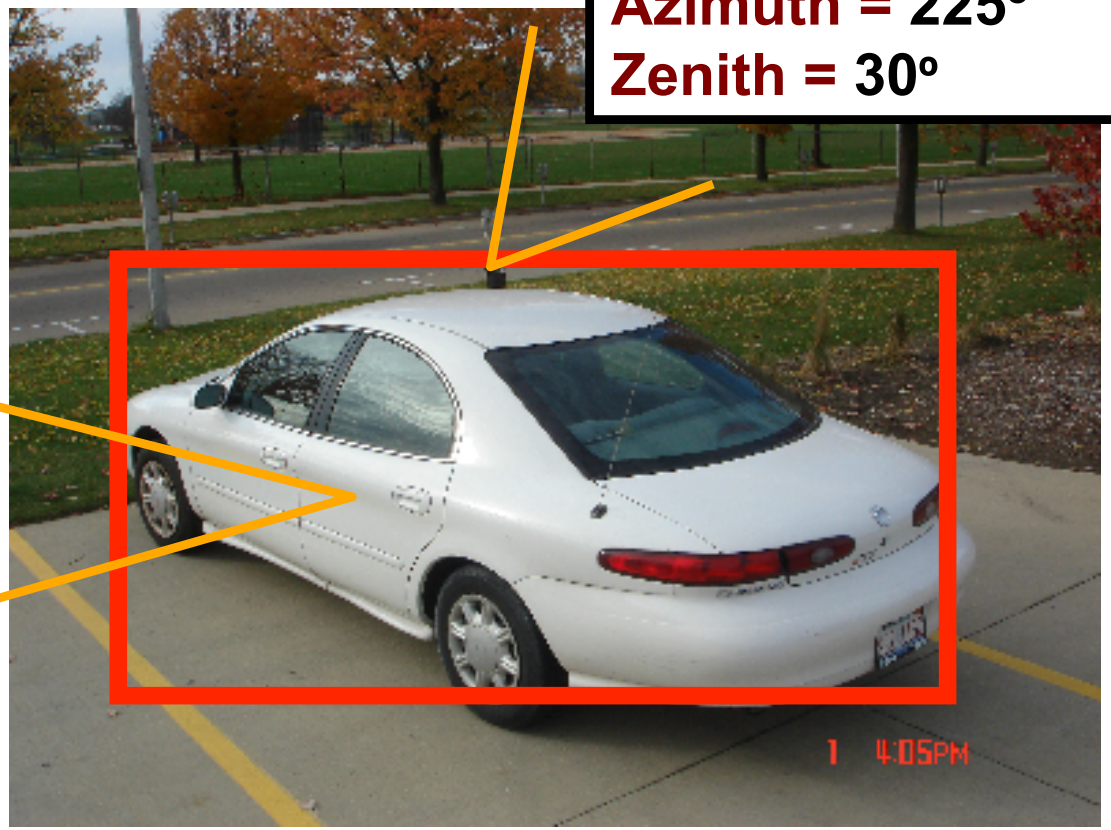
– Recognition task

– Search strategy

– Attributes

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

Category: car
Azimuth = 225º
Zenith = 30º



- 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, 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 r… our eyes. For a long ti… retinal image wa… isual centers i… s a movie s… image … discove… know th… perceptio… y more com… following the … ath to the various c… ortex, Hubel and Wiesel ha… demonstrate that the *message abou… image falling on the retina undergoes … p-wise analysis in a system of nerve cell… stored in columns. In this system each c… has its specific function and is responsible… a specific detail in the pattern of the retinal image.*

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

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% … $750bn, compared wi… $660bn. T… annoy th… China's… deliber… agrees … yuan is … governo… also need… demand so … country. China … yuan against the doll… permitted it to trade within a narrow… but the US wants the yuan to be allowed … freely. However, Beijing has made it cl… it will take its time and tread carefully be… allowing the yuan to rise further in value.

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

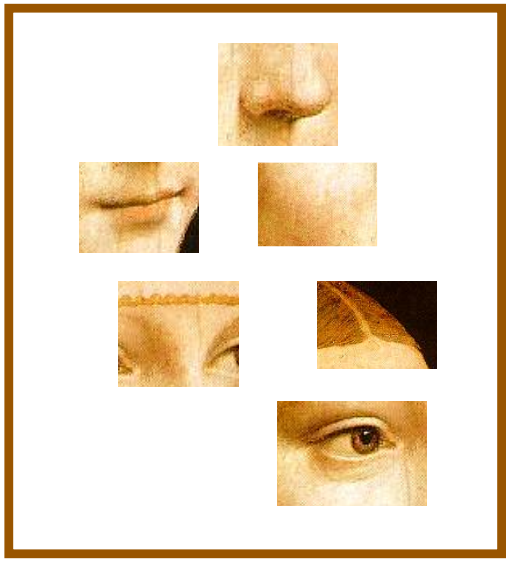
Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004

Object → Bag of 'words'

definition of “BoW”

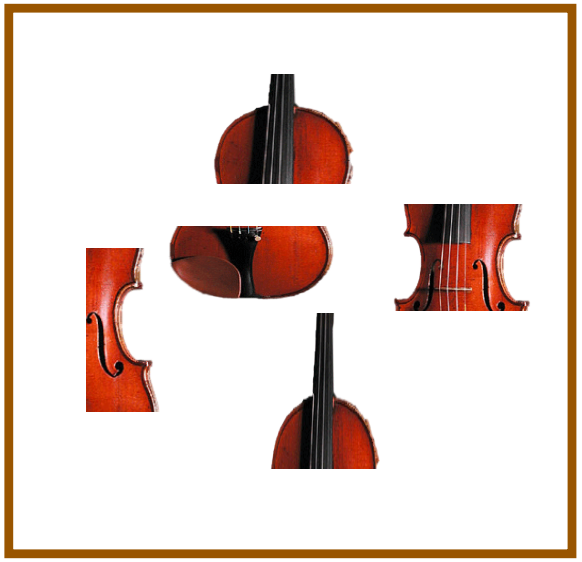
– Independent features

face



bike

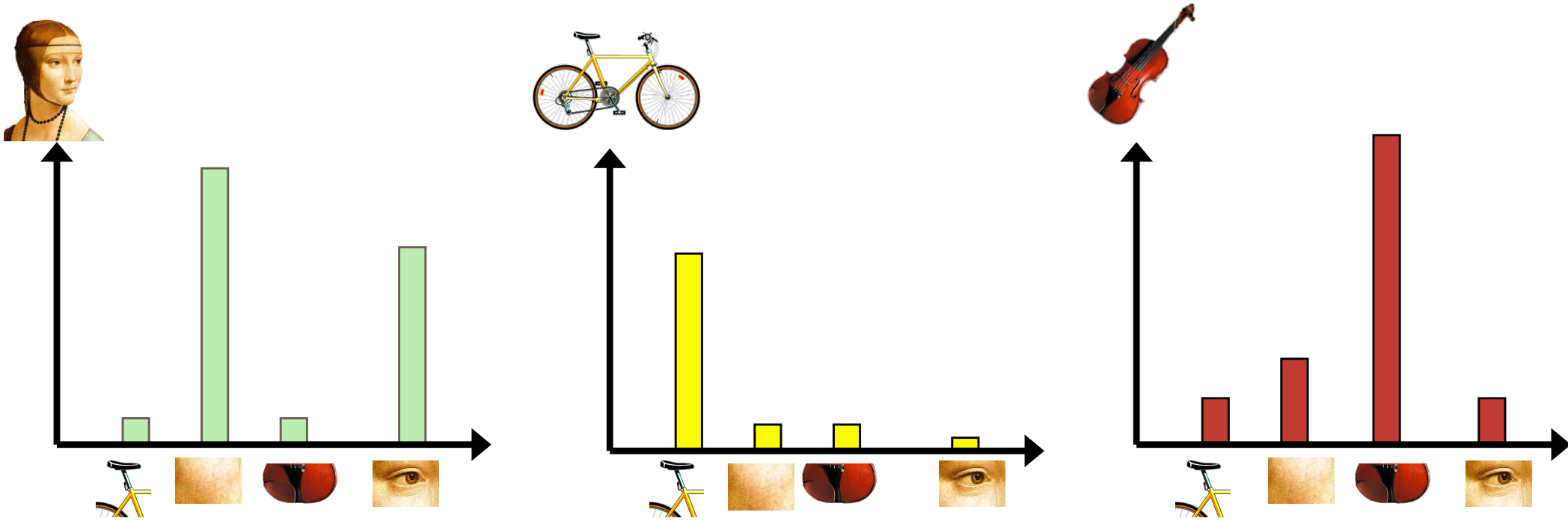
violin



definition of “BoW”

– Independent features

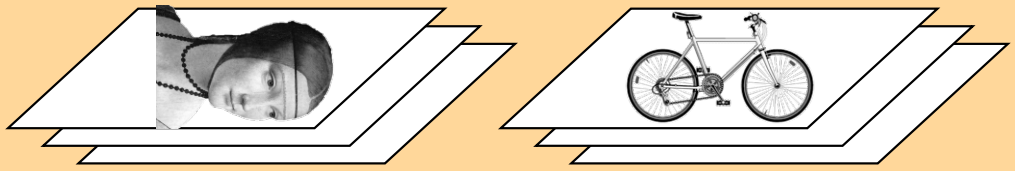
– histogram representation



codewords dictionary

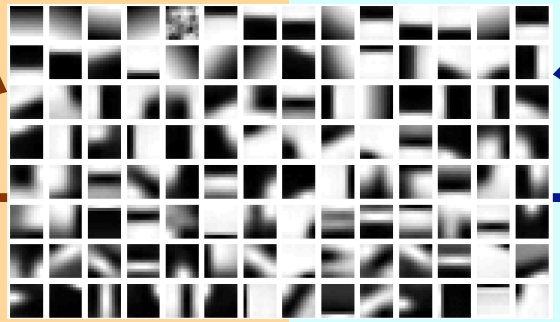
Representation

recognition

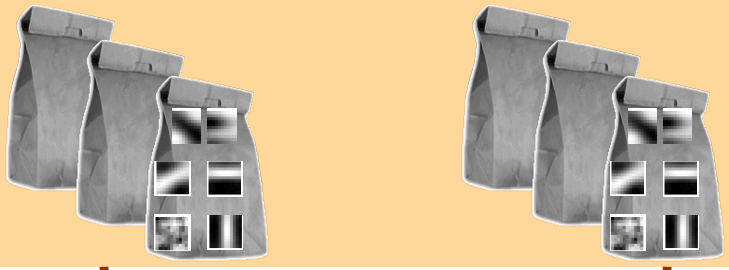


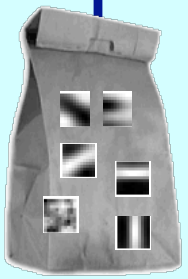
1. feature detection & description

2. codewords dictionary



3. BOW representation



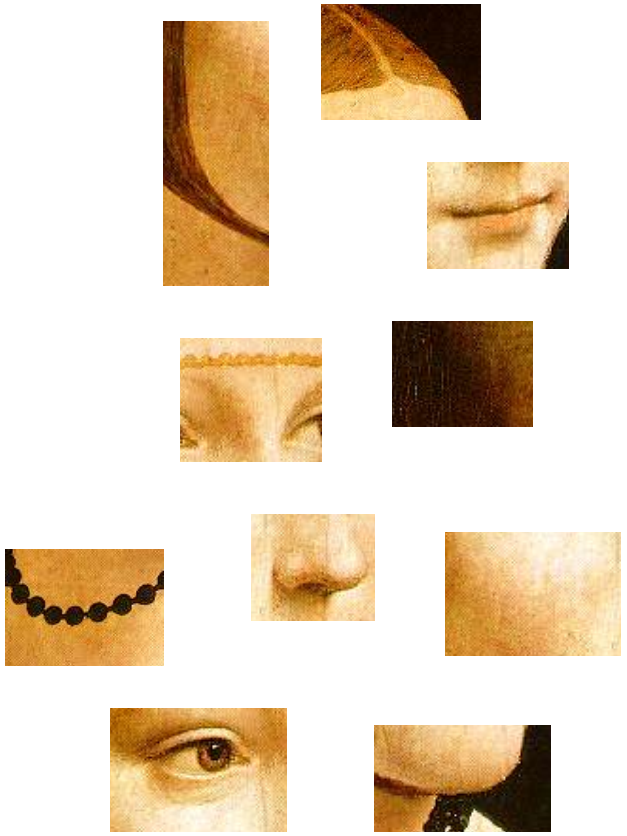


learning

category models (and/or) classifiers

category decision

1.Feature detection and description

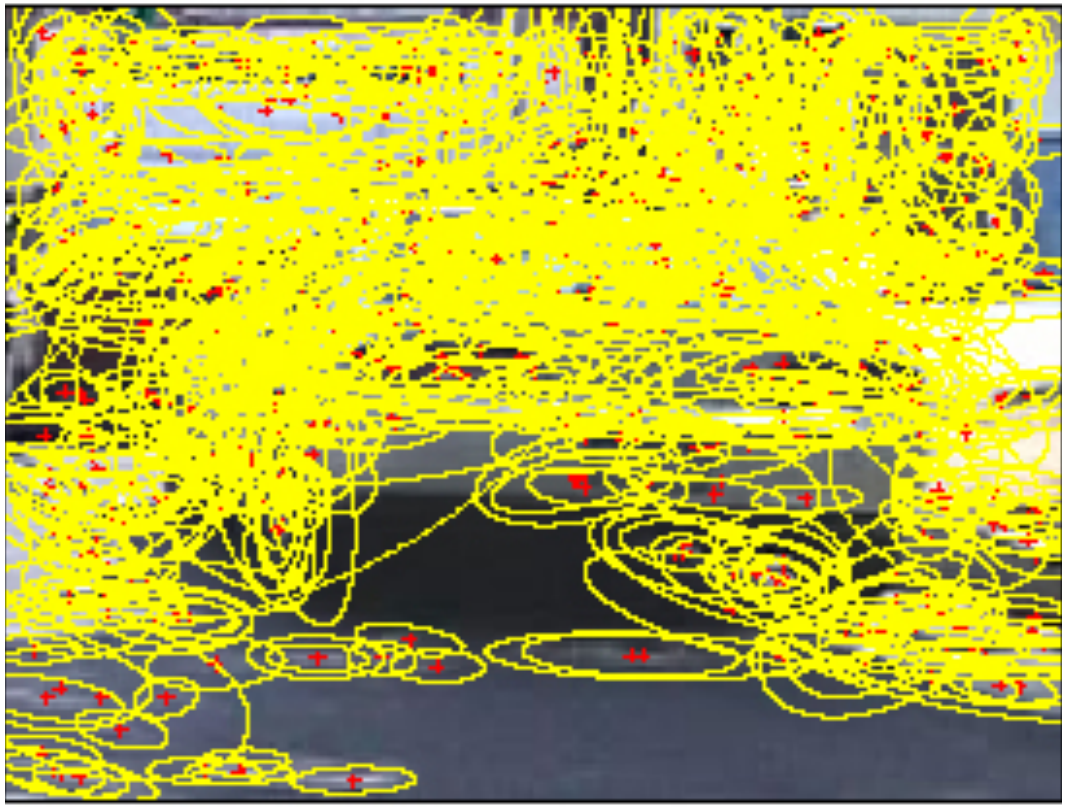


1.Feature detection and description

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

1.Feature detection and description

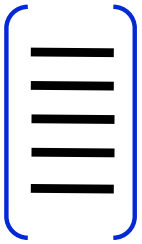
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic, et al. 2005



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



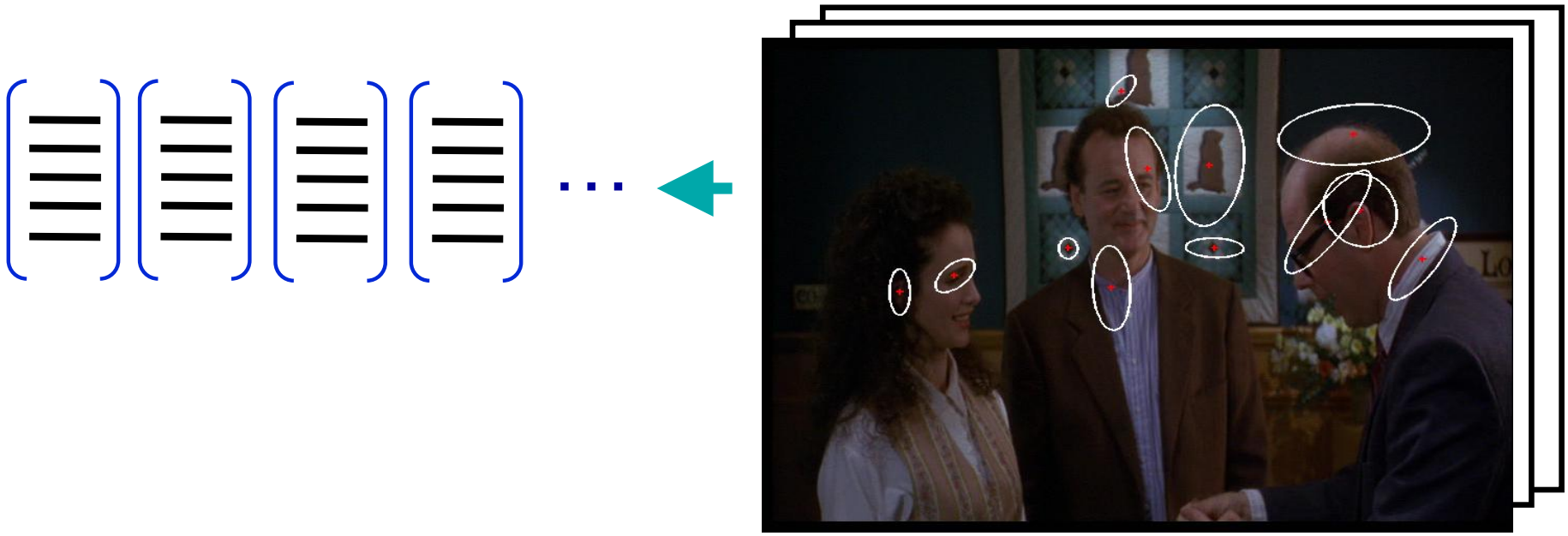
Compute descriptor

E.g. SIFT [Lowe'99]

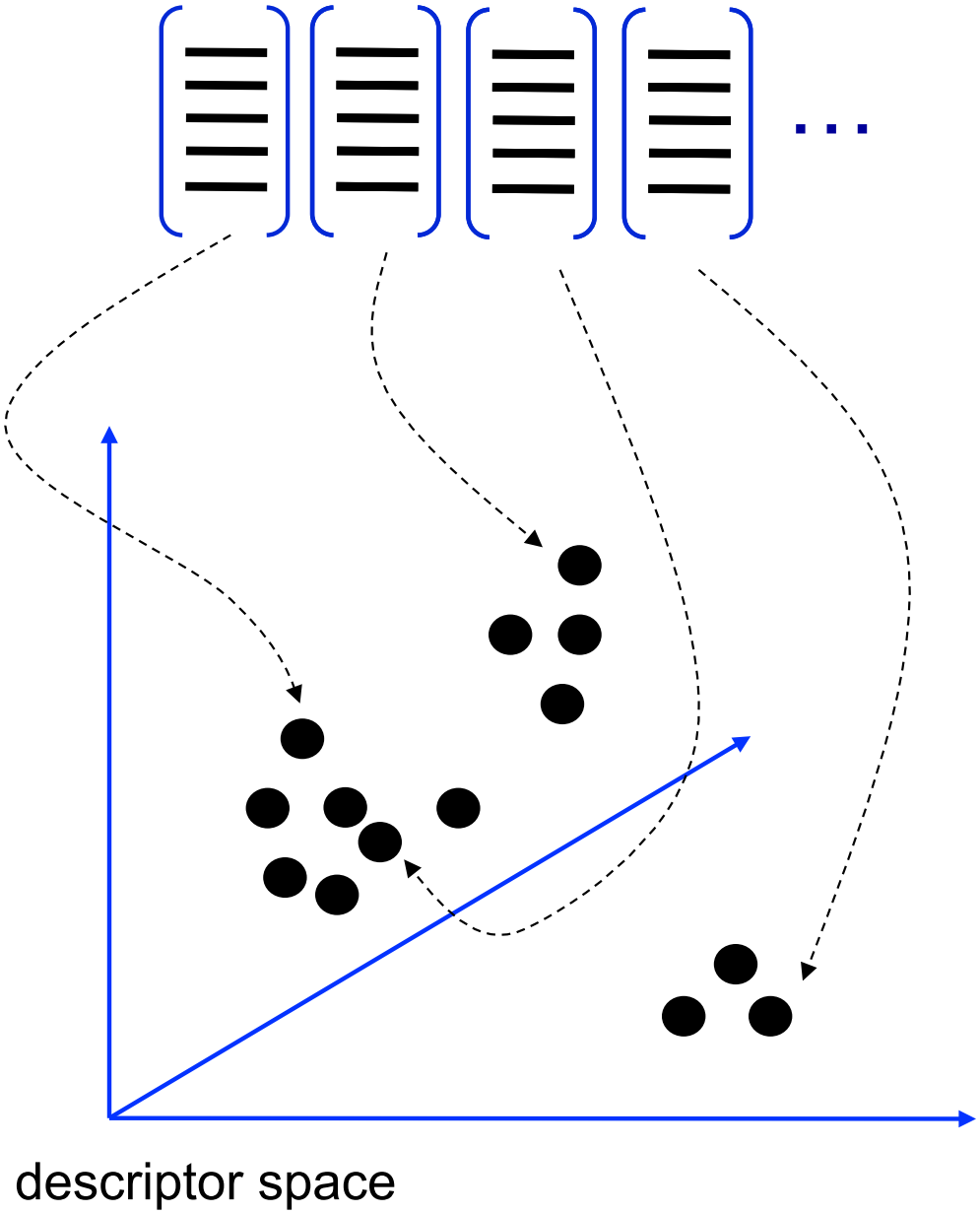
Detected features

Slide credit: Josef Sivic

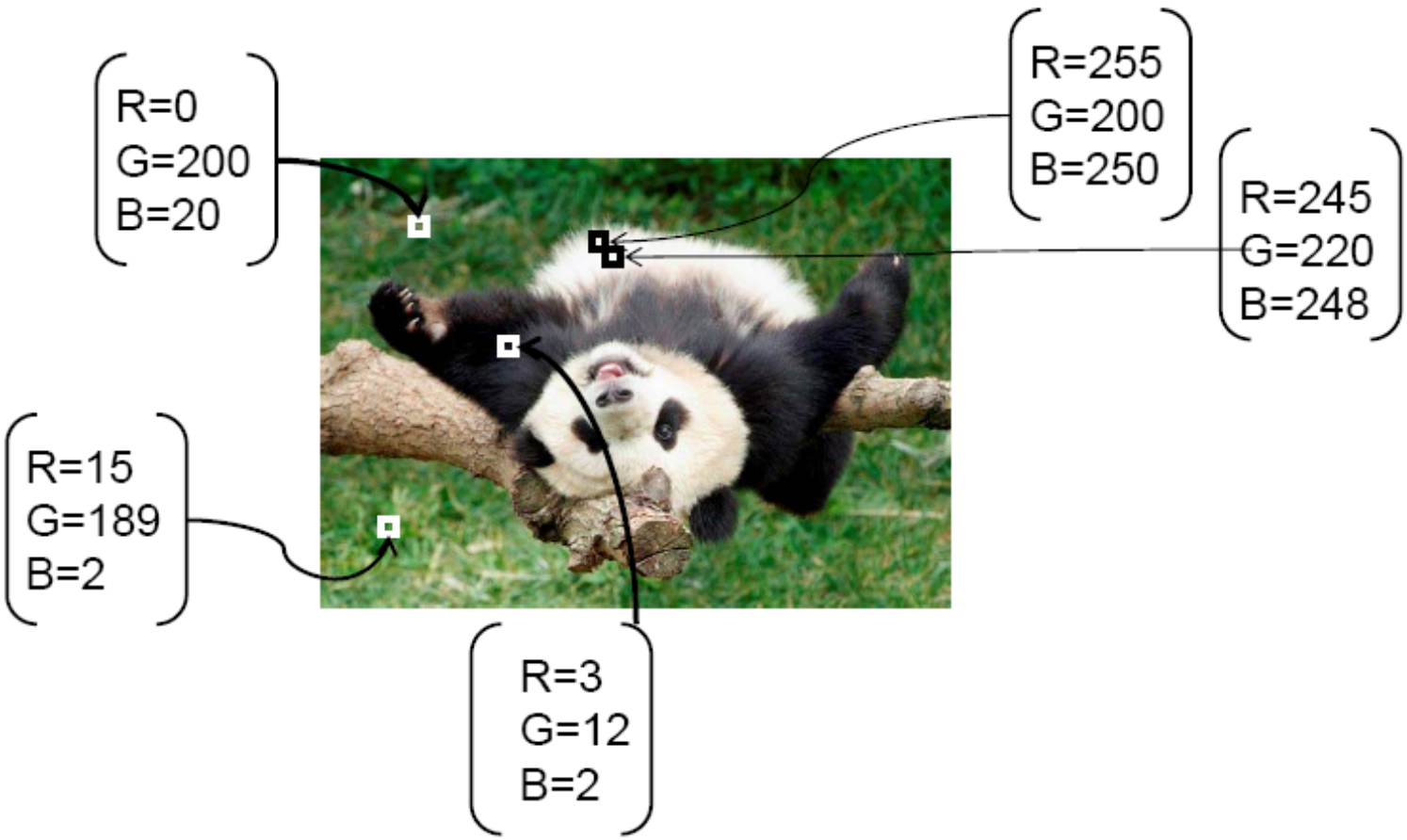
2. Codewords dictionary formation



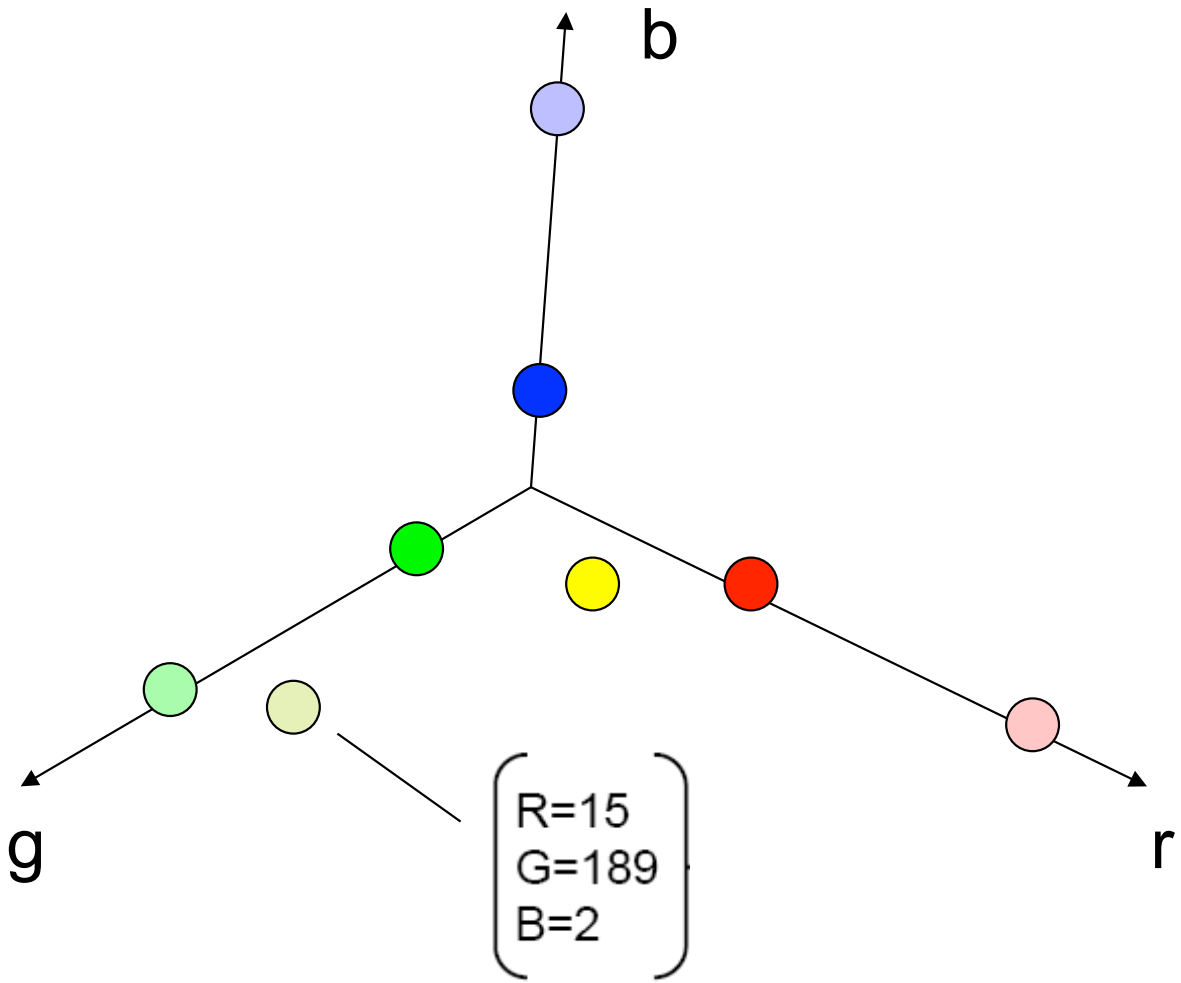
2. Codewords dictionary formation



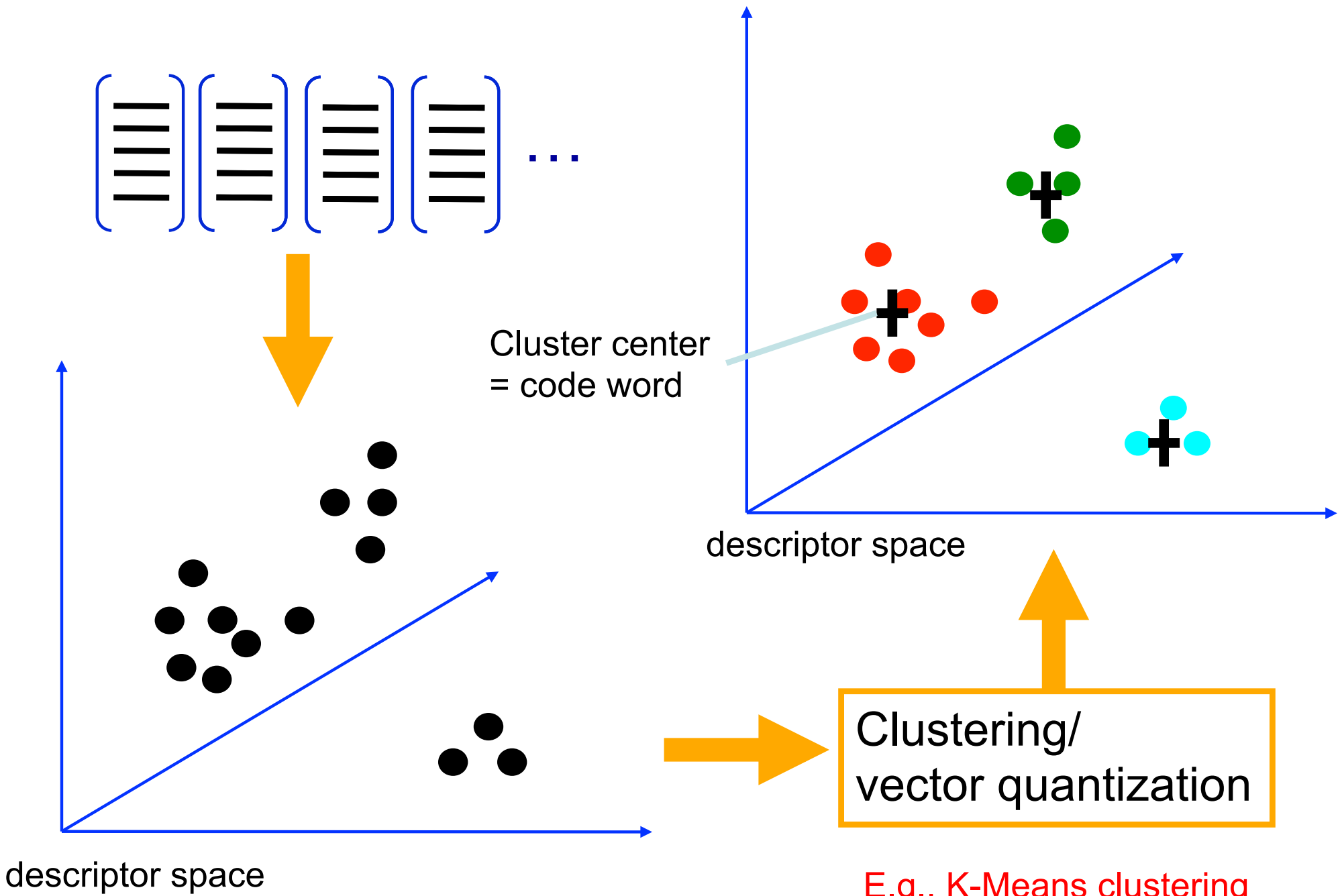
Example: color feature

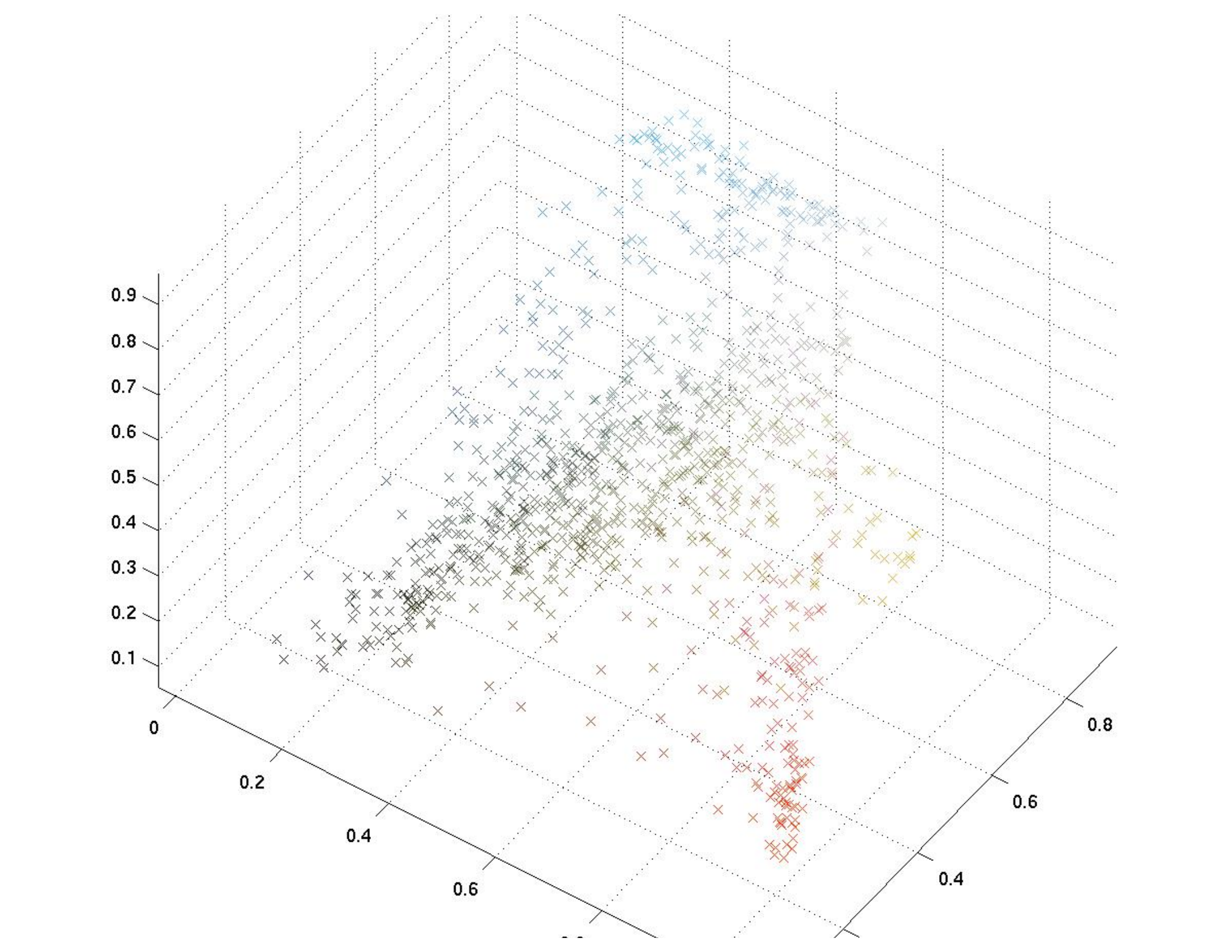


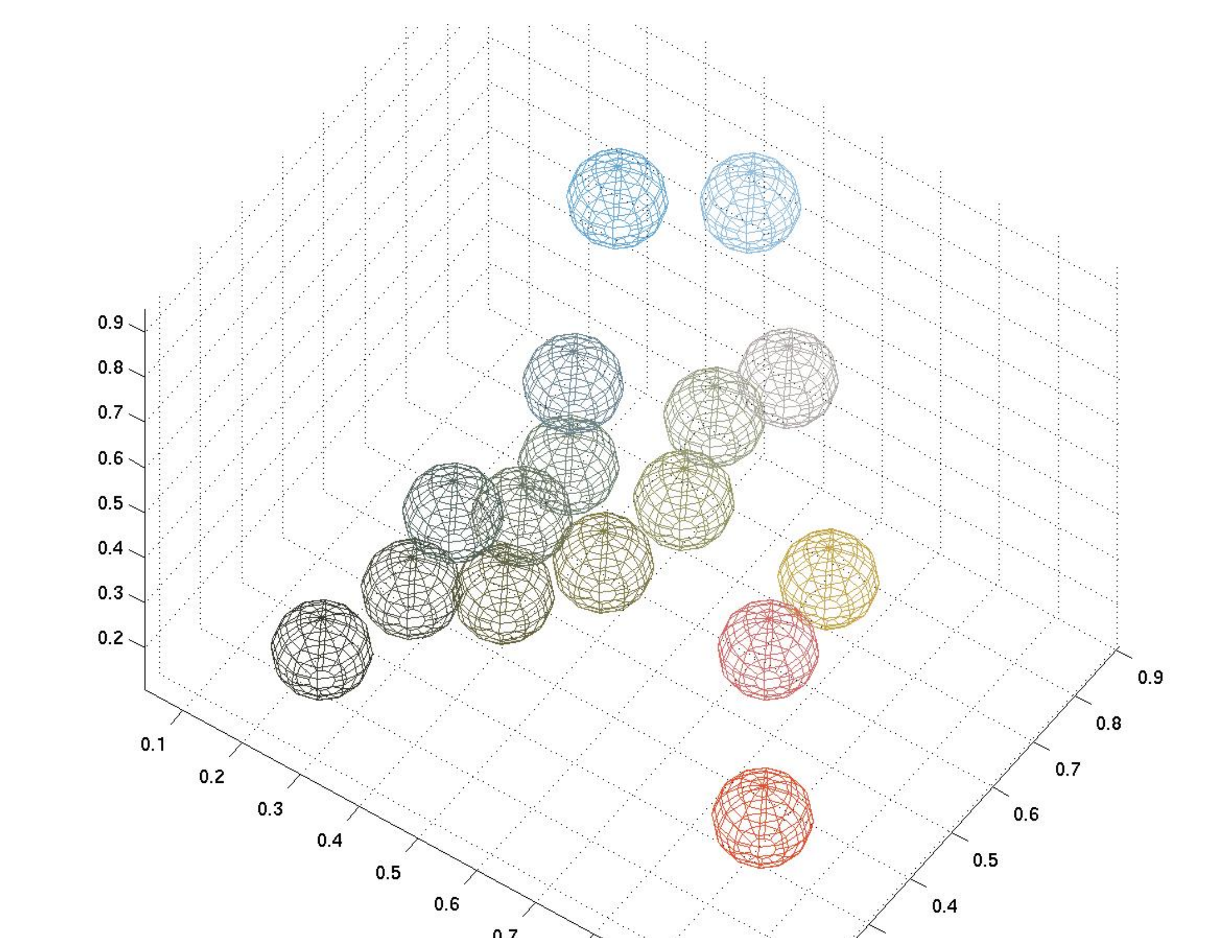
Example: color feature



2. Codewords dictionary formation

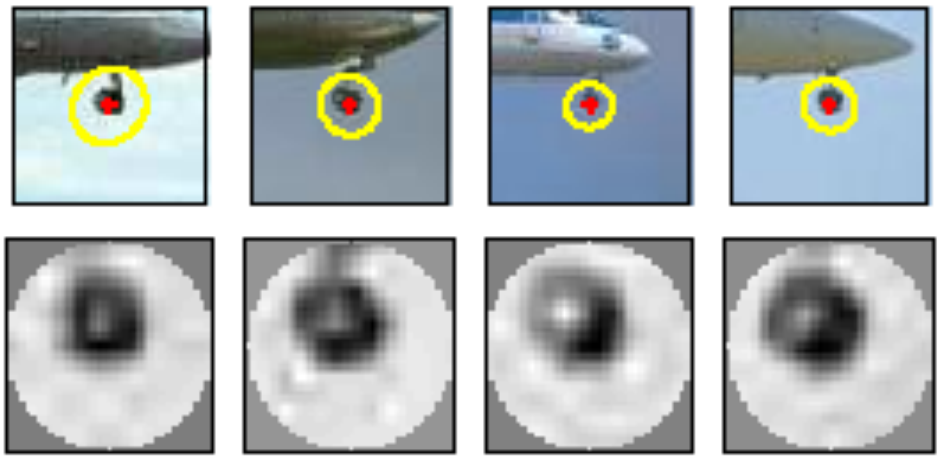



0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
0.6
0.4

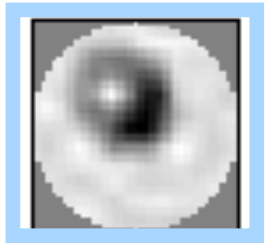

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.4
0.5
0.6
0.7
0.8
0.9

2. Codewords dictionary formation

- **Image patch examples of codewords**



Cluster



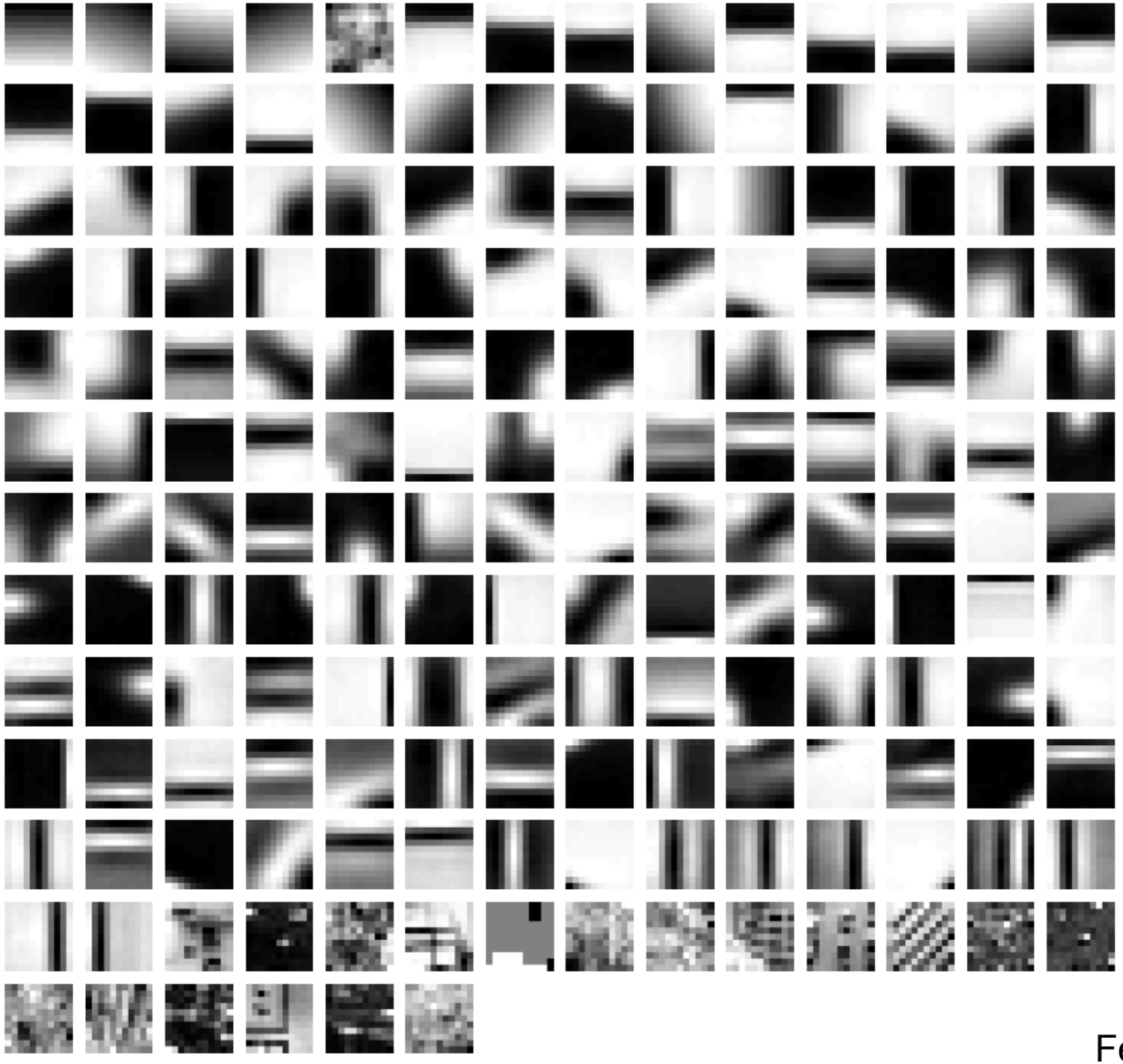
code word

Cluster

code word

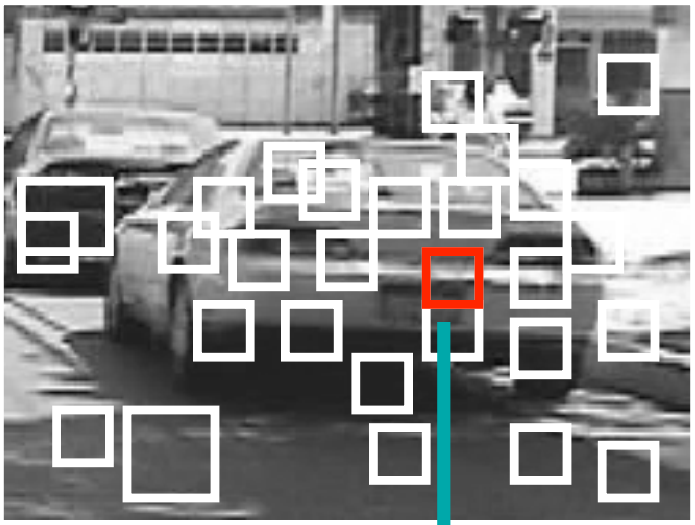
Sivic et al. 2005

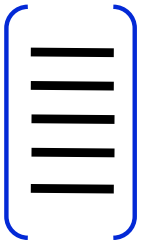
2. Codewords dictionary formation

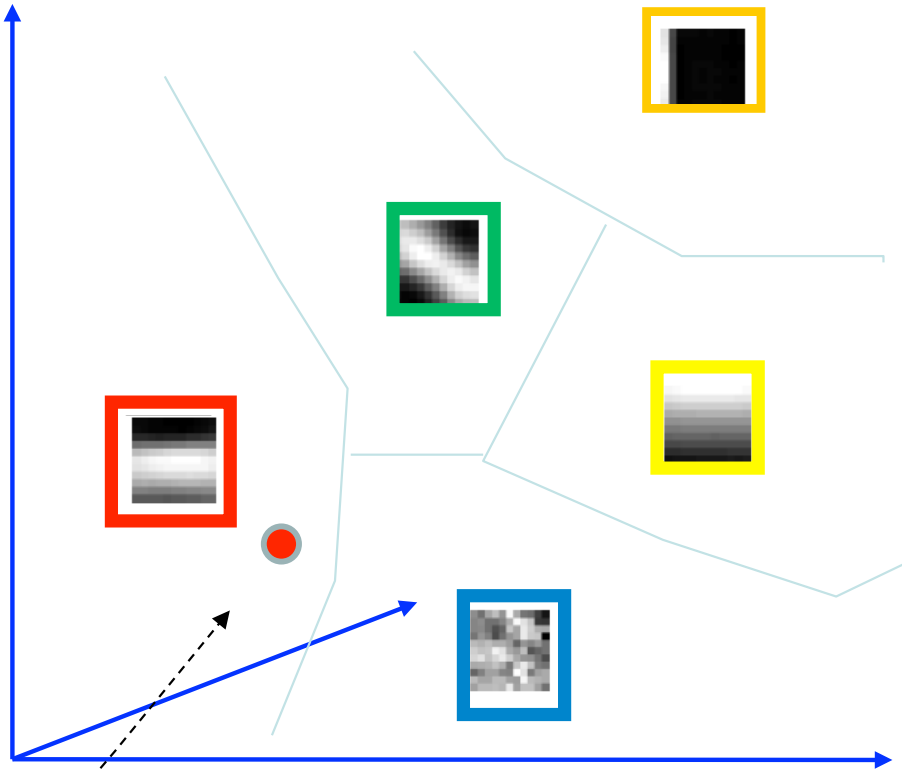


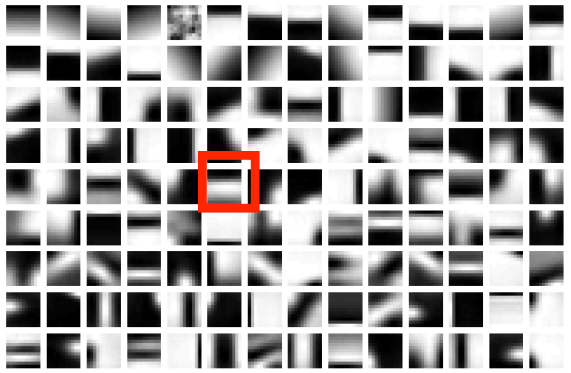
Fei-Fei et al. 2005

3. Bag of word representation





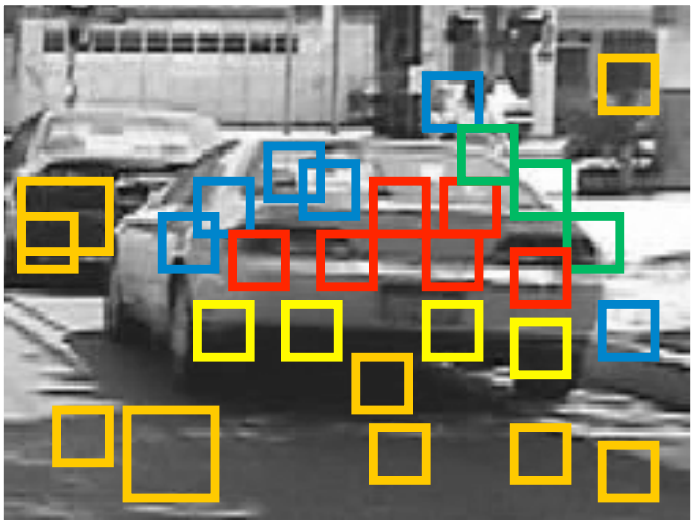


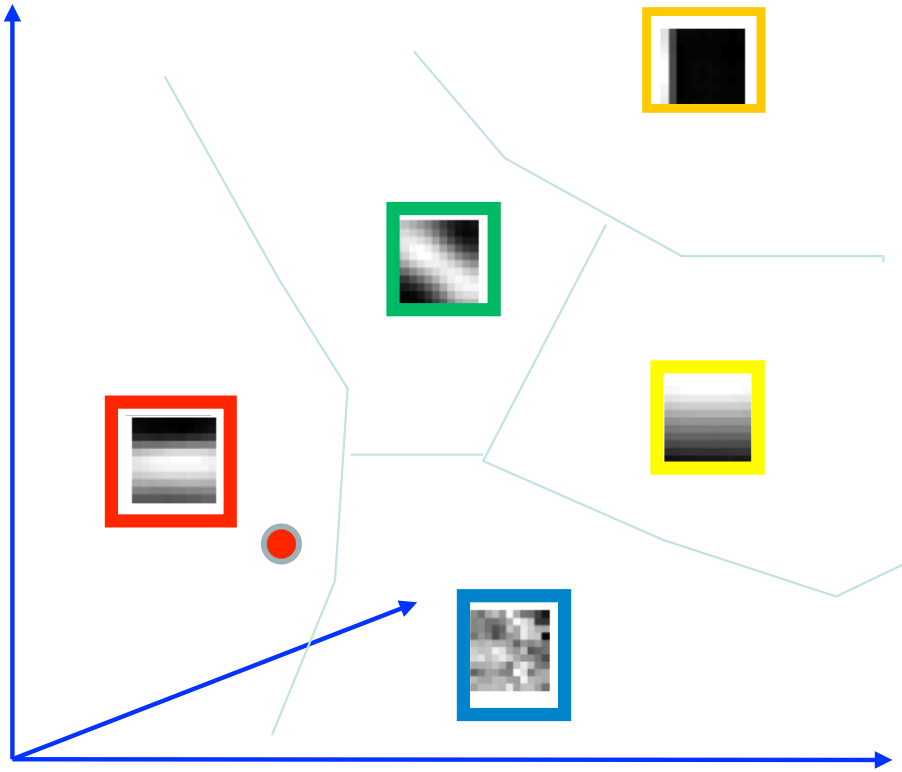


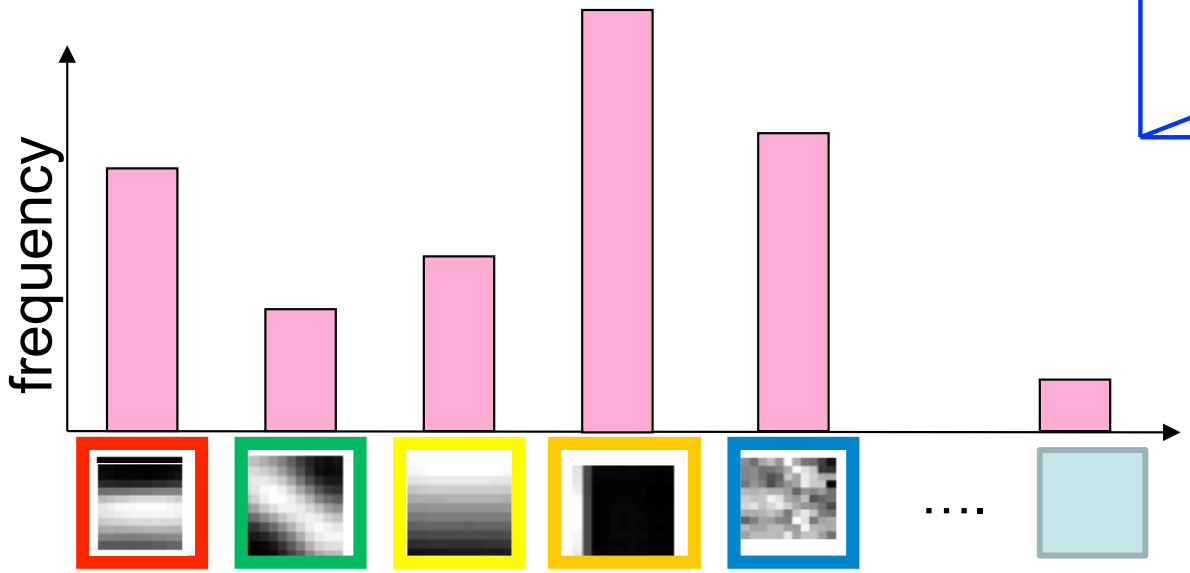
Codewords dictionary

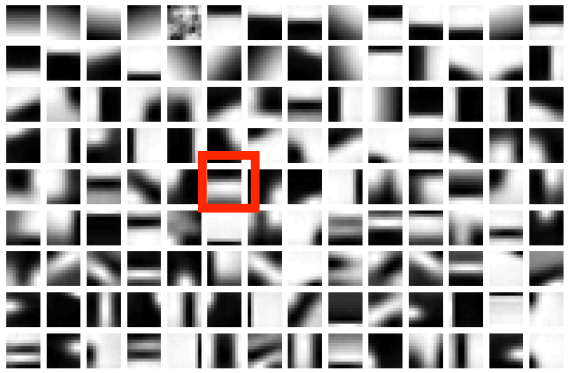
- Nearest neighbors assignment
- K-D tree search strategy

3. Bag of word representation







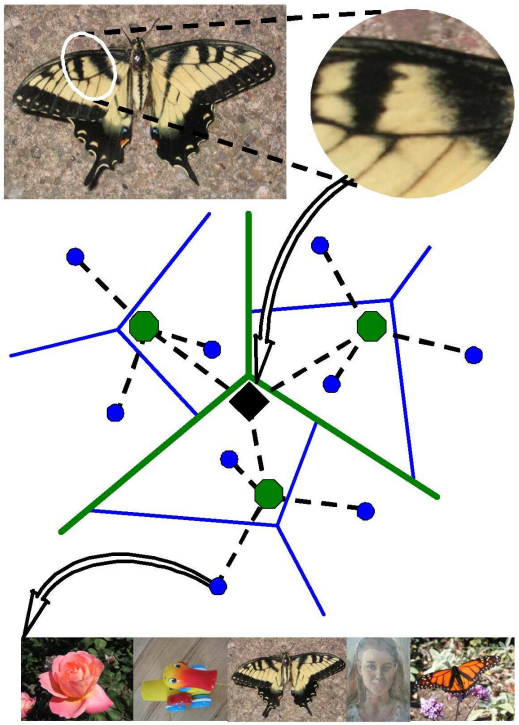


Codewords dictionary

Visual vocabularies: Issues

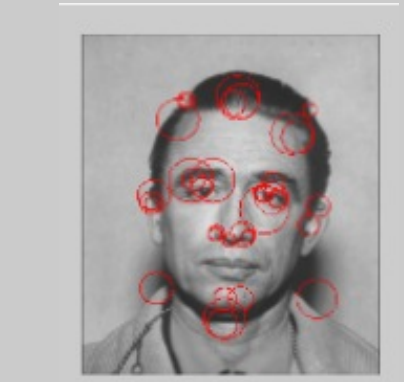
- How to choose vocabulary size?
  - Too small: visual words not representative of the object appearance distribution
  - Too large: quantization artifacts, sparse histograms, overfitting

- Computational efficiency
  - Vocabulary trees (Nister & Stewenius, 2006)



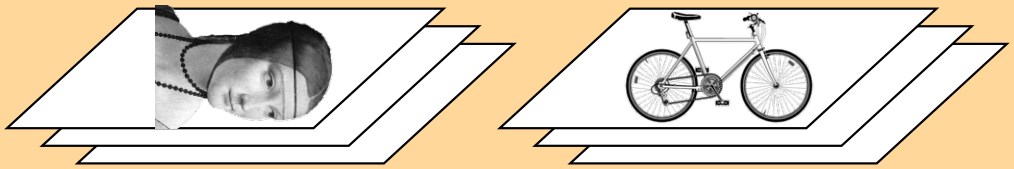
Invariance issues

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



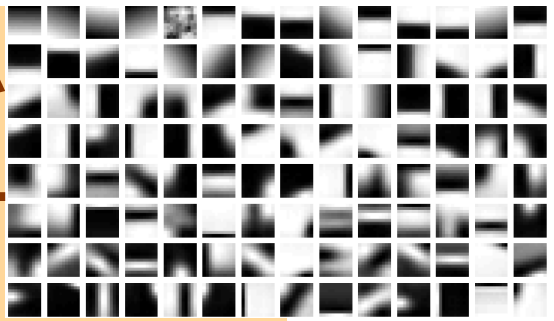
Kadir and Brady. 2003

Representation



2.

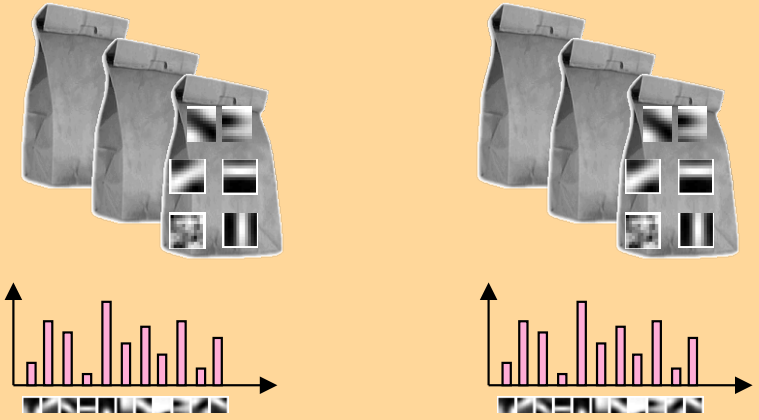
codewords dictionary



1. feature detection & representation

image representation

3.



category models

Next Lecture

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

Appendix

Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

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

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

Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule: $p(A|B) = \frac{p(B|A)\ p(A)}{p(B)}$

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

Object categorization: the statistical viewpoint

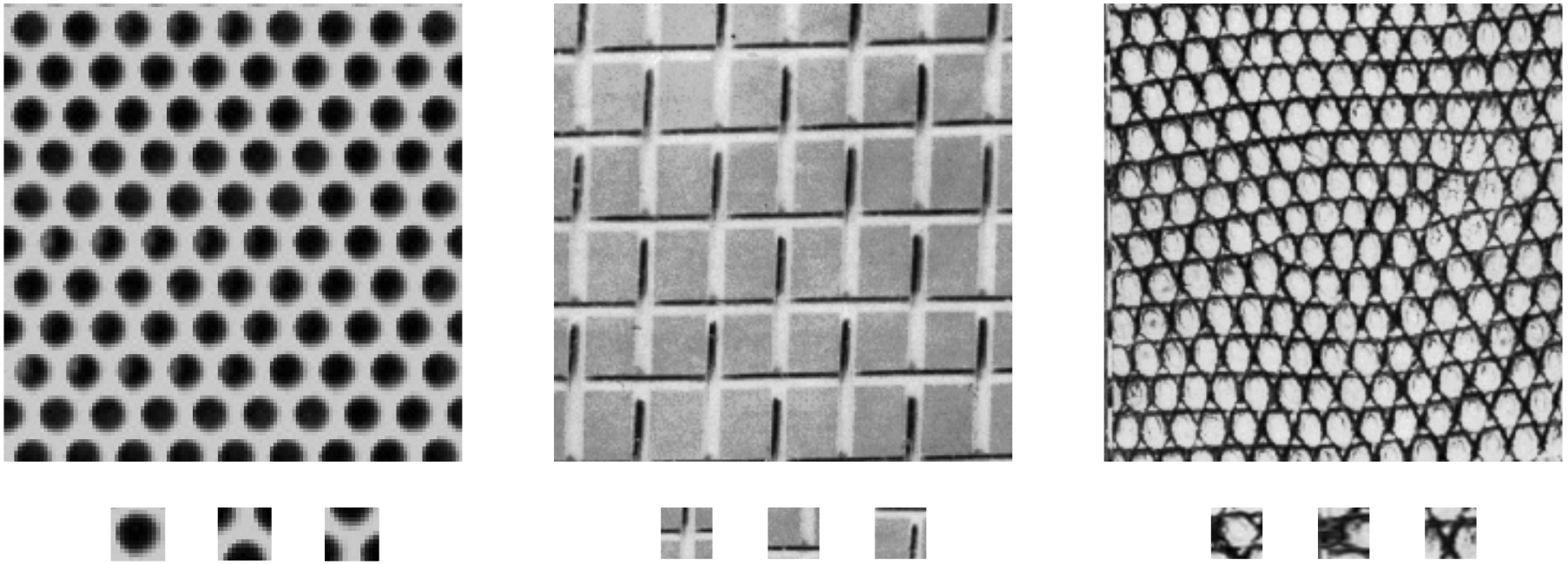
- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

- Bayes rule:

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

Representing textures

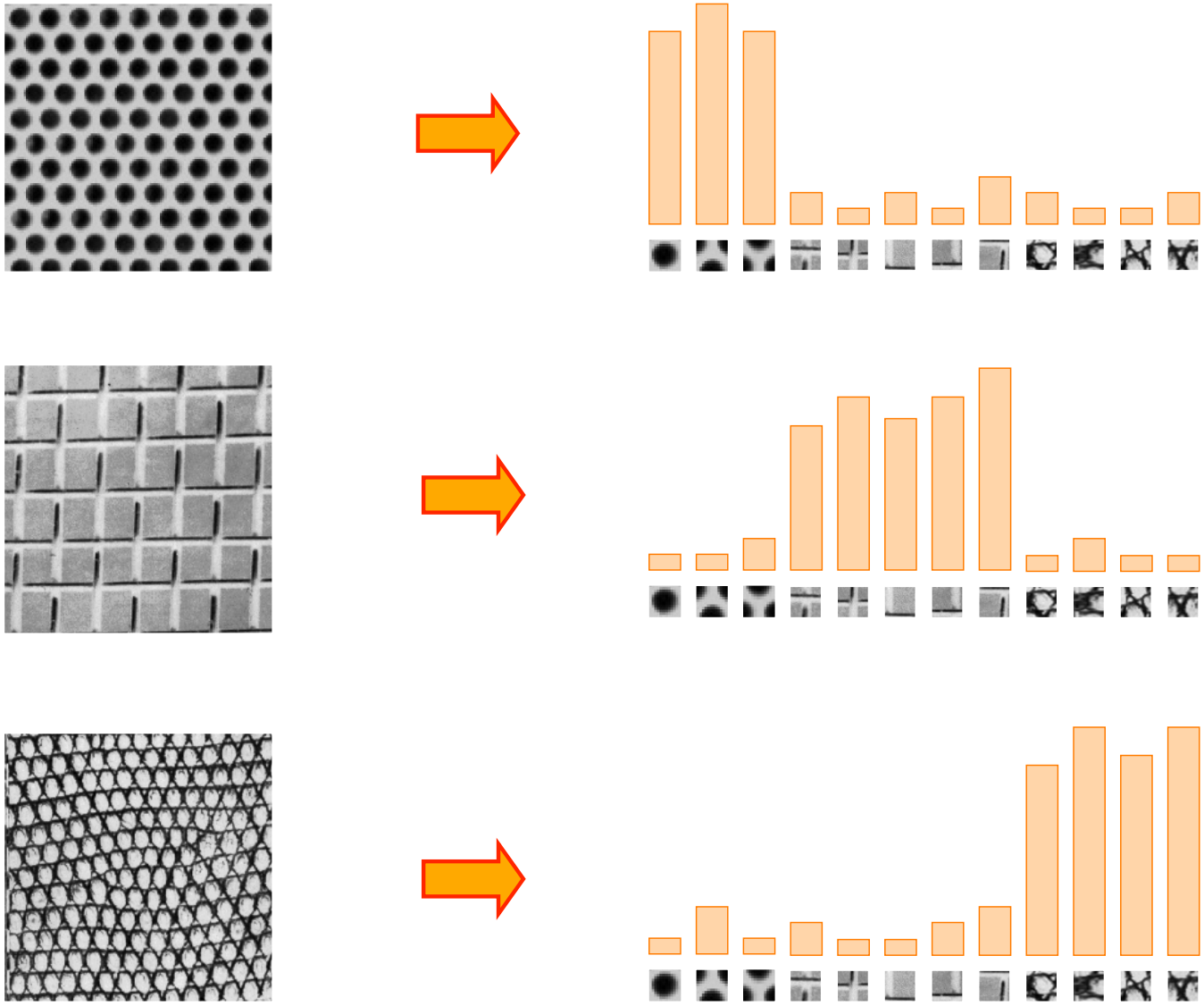
- Texture is characterized by the repetition of basic elements or *textons*
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



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