Lecture 12
Visual recognition

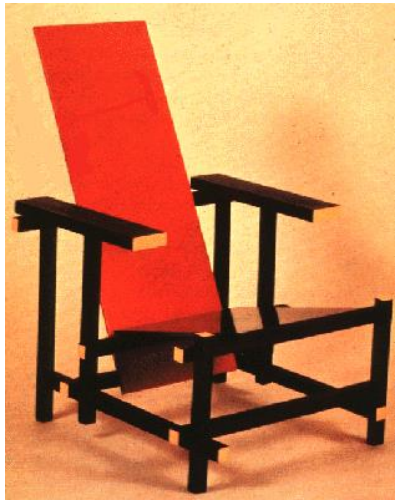
- Bag of words models for object recognition and classification
  - Discriminative methods
  - Generative methods

Challenges

Variability due to:

- View point
- Illumination
- Occlusions
- Intra-class variability

Challenges: intra-class variation



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

Part 1: Bag-of-words models

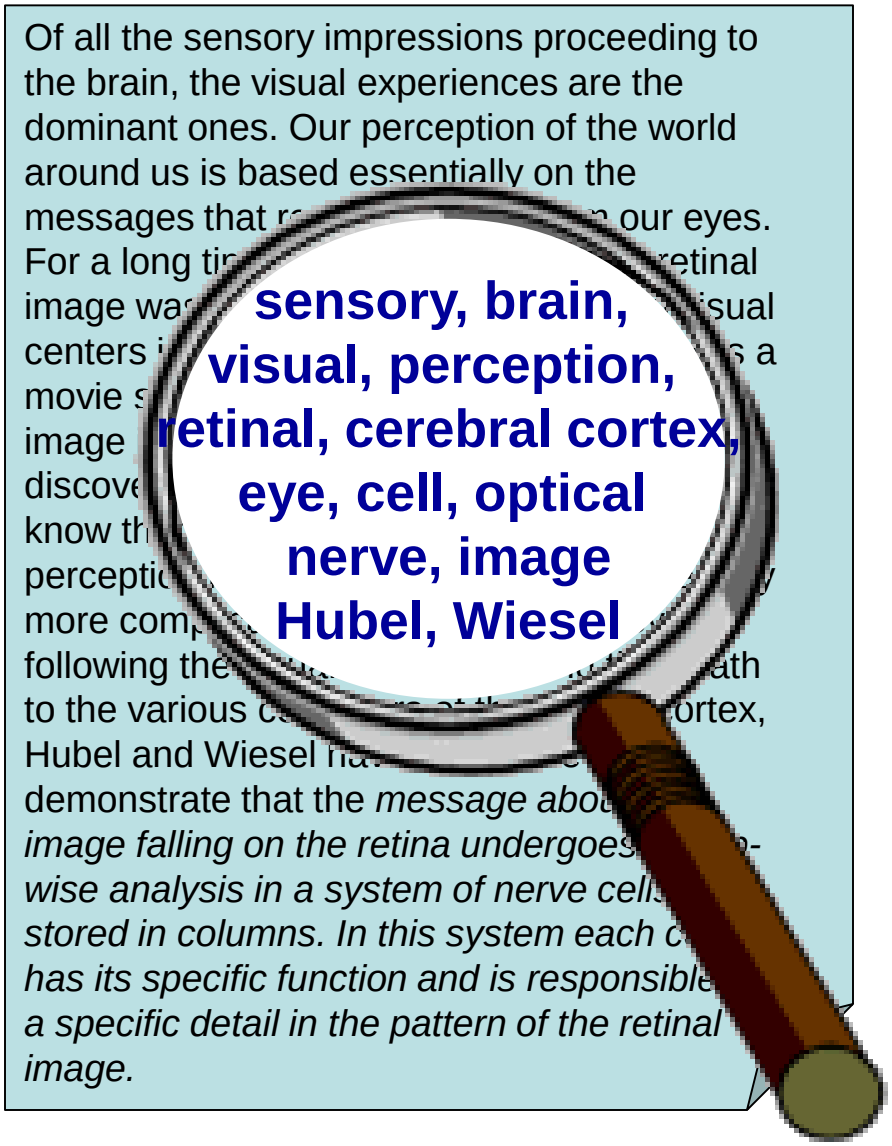
This segment is based on the tutorial *"**Recognizing and Learning Object Categories: Year 2007**",* by Prof A. Torralba, R. Fergus and F. Li

Related works

- 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
- Object categorization
  - Csurka, Bray, Dance & Fan, 2004; Sivic, Russell, Efros, Freeman & Zisserman, 2005; Sudderth, Torralba, Freeman & Willsky, 2005;
- Natural scene categorization
  - Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman & Munoz, 2006

Object → Bag of ‘words’

Analogy to documents



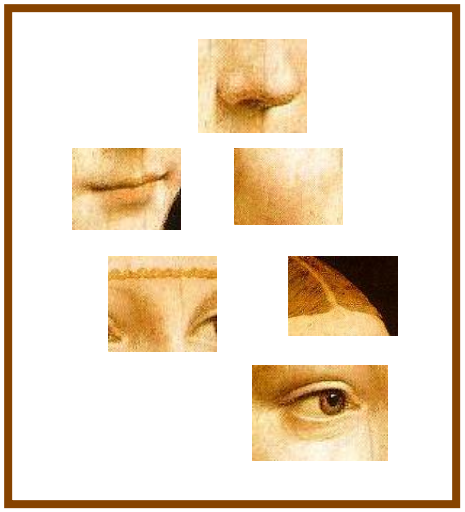
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that r… our eyes. For a long ti… retinal image wa… isual centers i… s a movie s… image … discove… know th… perceptio… more com… following the… ath to the various c… ortex, Hubel and Wiesel ha… demonstrate that the *message abou… image falling on the retina undergoes… -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.*

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 … e yuan against the dol… nd permitted it to trade within a narrow … but the US wants the yuan to be allowed … de freely. However, Beijing has made it cl… t it will take its time and tread carefully be… allowing the yuan to rise further in value.

definition of “BoW”

– Independent features

face

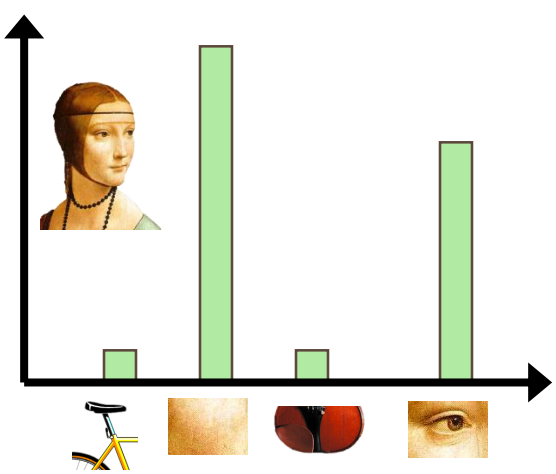


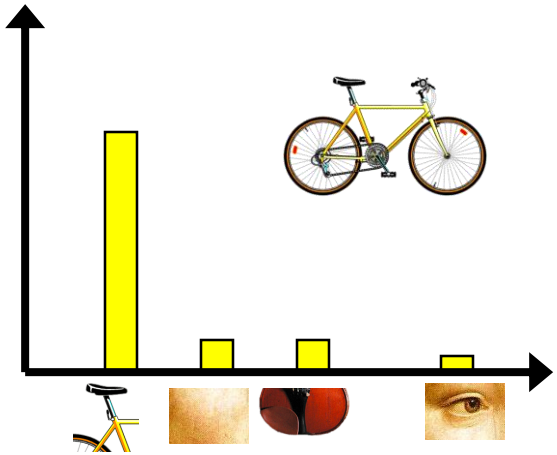
bike

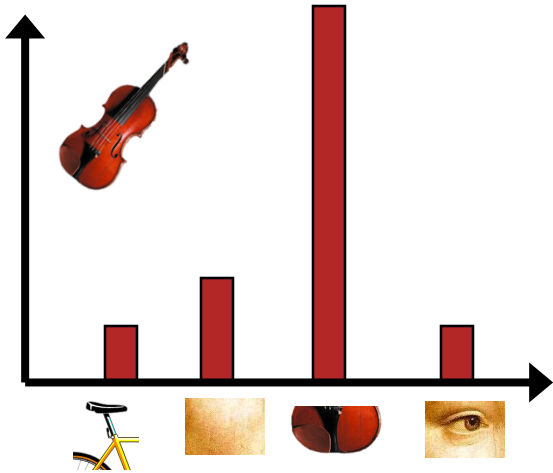
violin

definition of “BoW”

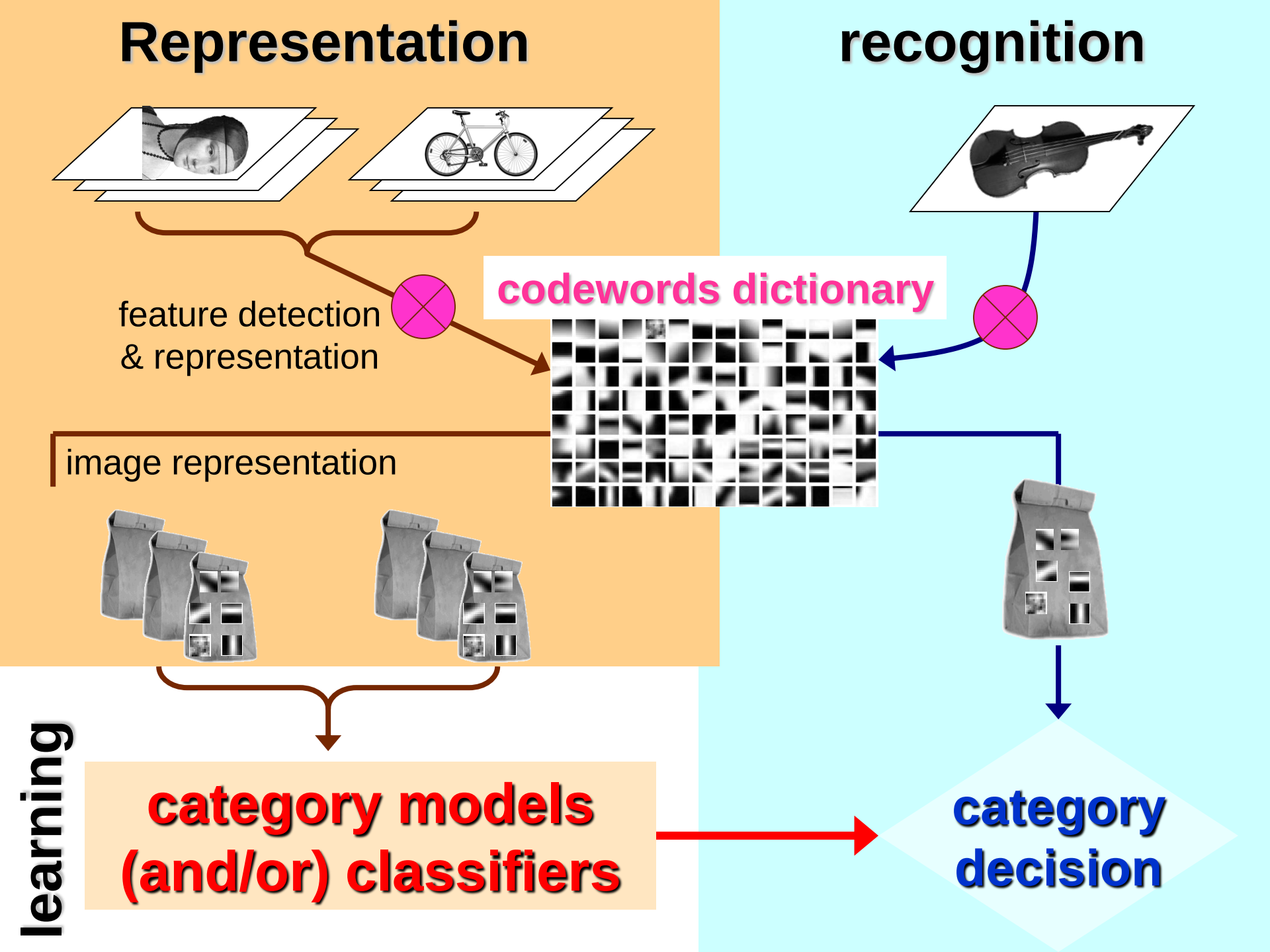
- Independent features
- histogram representation



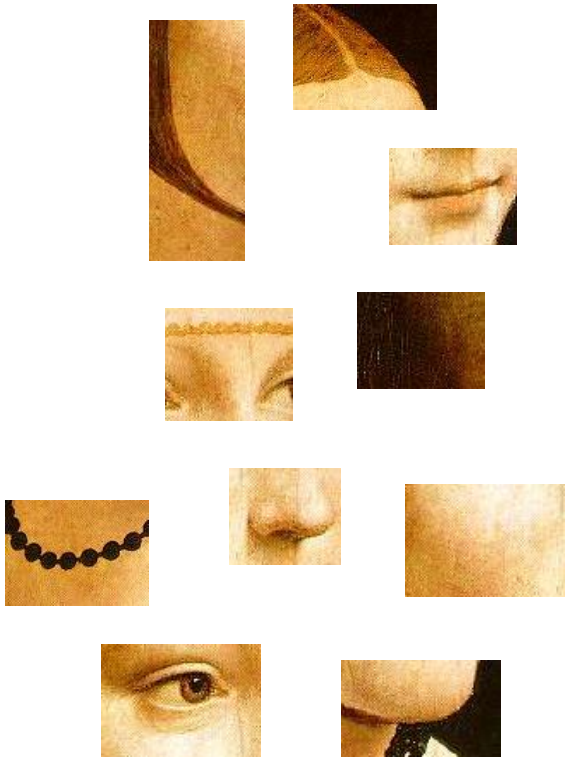




codewords dictionary

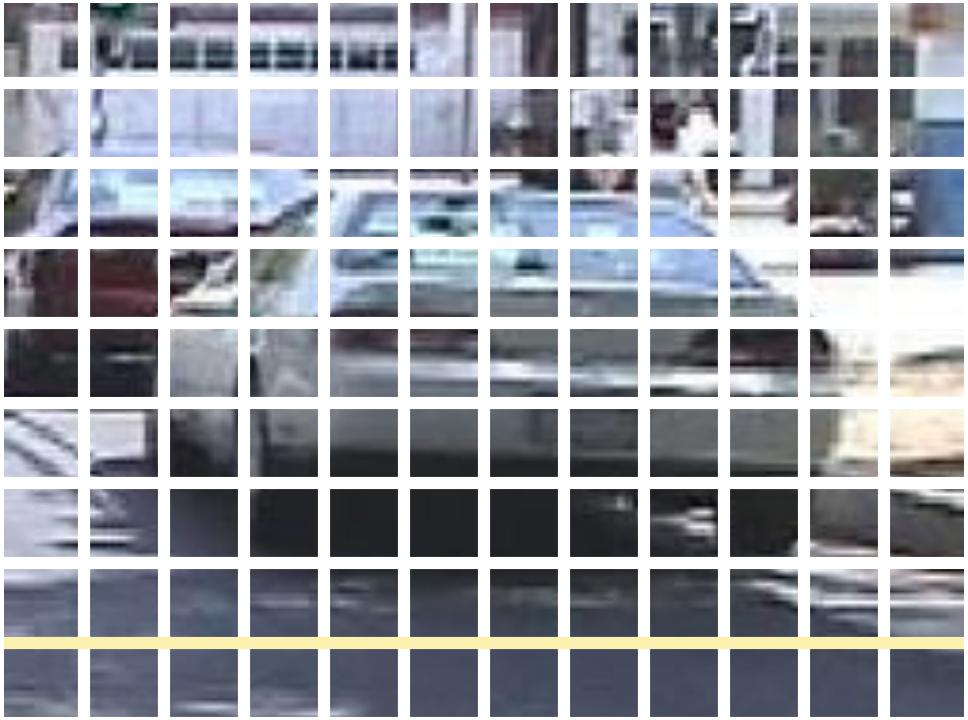

Representation
recognition
feature detection
& representation
codewords dictionary
image representation
learning
category models
(and/or) classifiers
category
decision

1.Feature detection and description



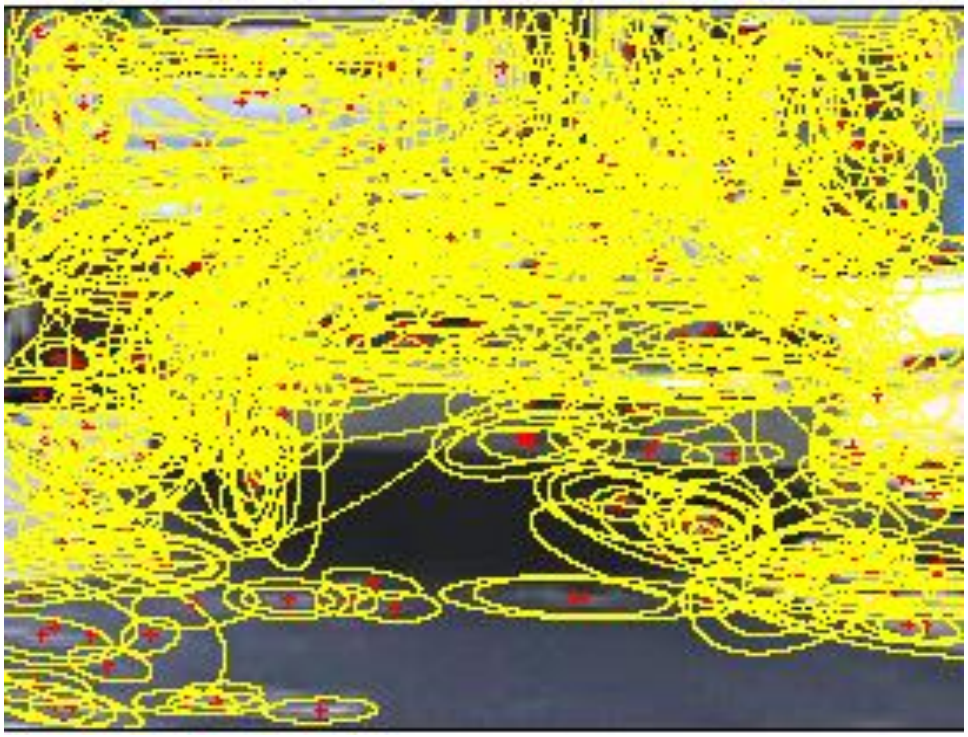
1.Feature detection and description

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



1.Feature detection and description

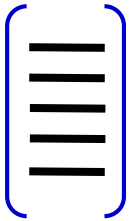
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic, et al. 2005

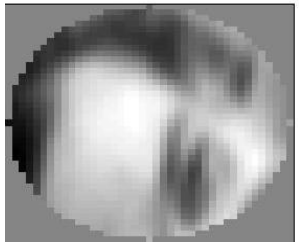


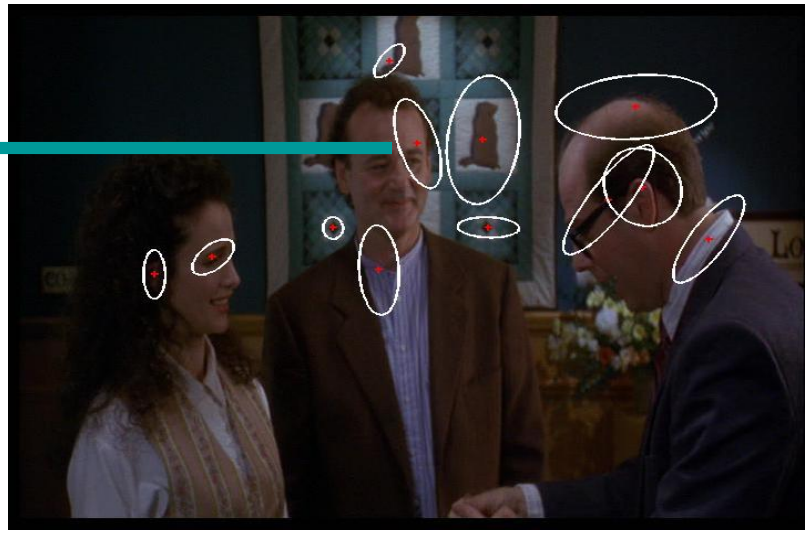
1.Feature detection and description

- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, Bray, Dance & Fan, 2004
  - Fei-Fei & Perona, 2005
  - Sivic, Russell, Efros, Freeman & Zisserman, 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 SIFT descriptor

[Lowe'99]

Normalize patch

Detect patches

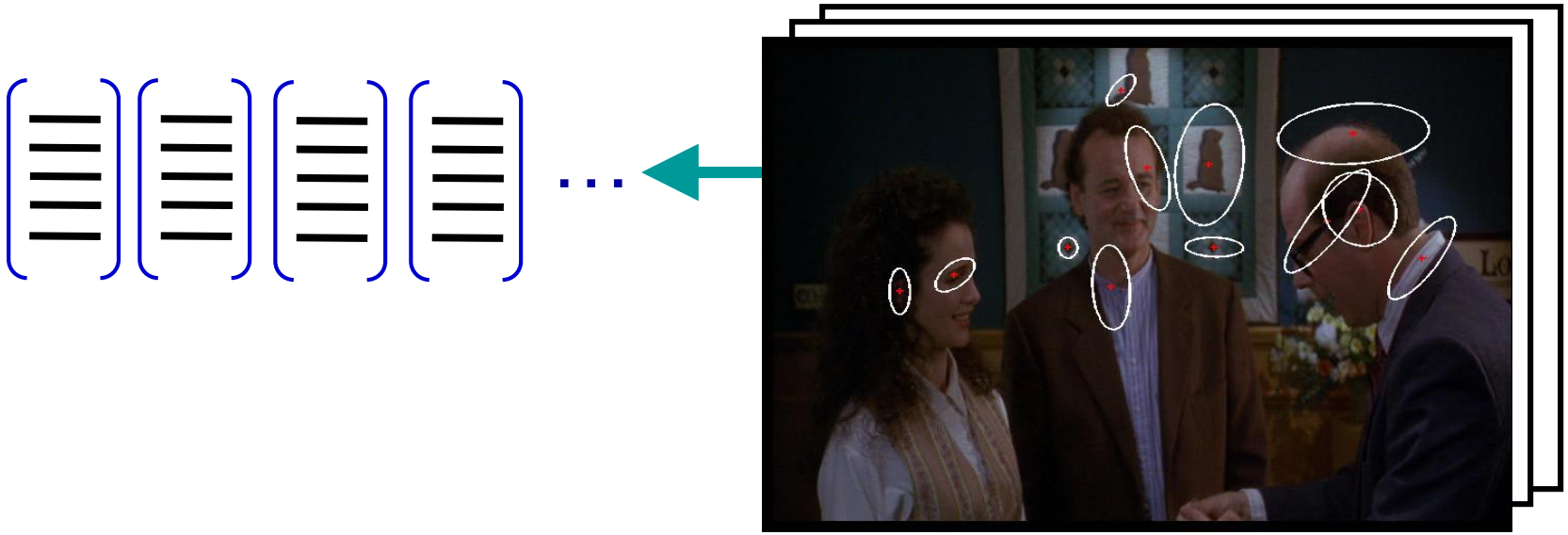
[Mikojaczyk and Schmid '02]

[Mata, Chum, Urban & Pajdla, '02]

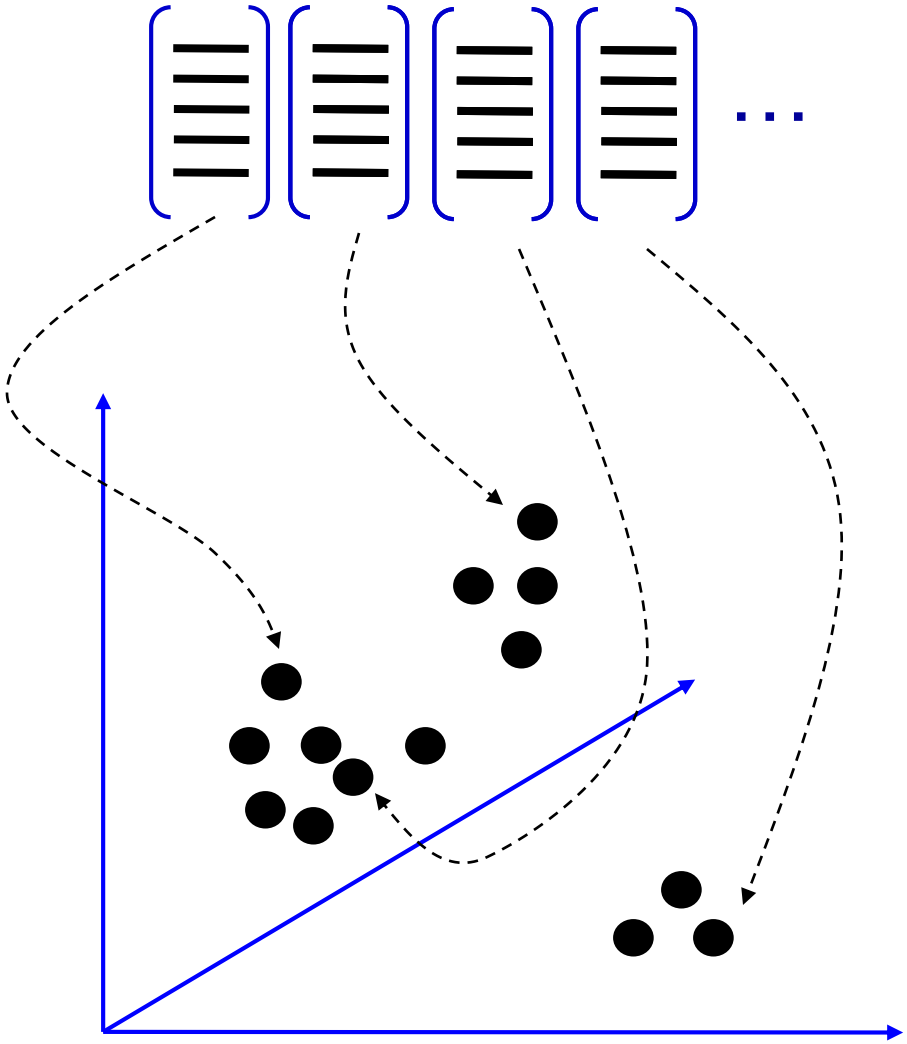
[Sivic & Zisserman, '03]

Slide credit: Josef Sivic

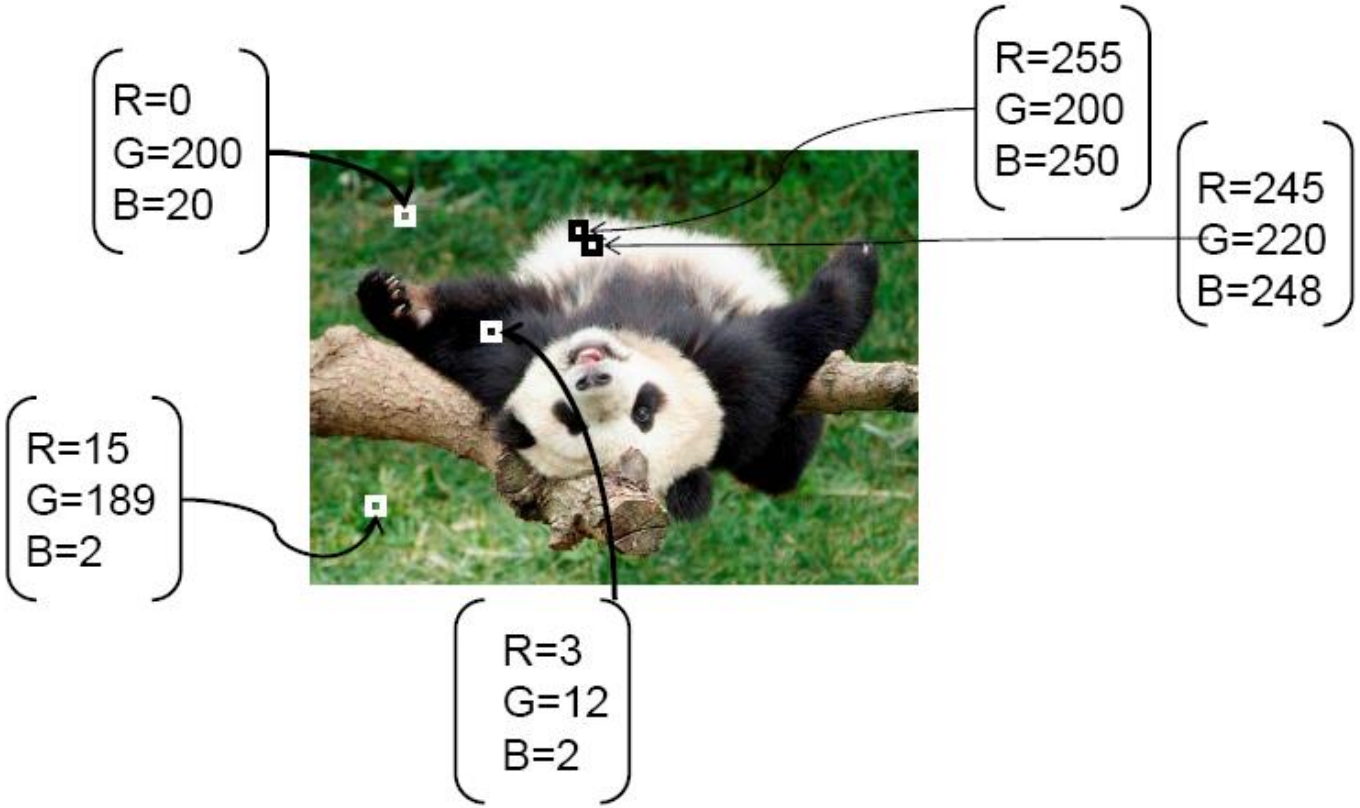
2. Codewords dictionary formation



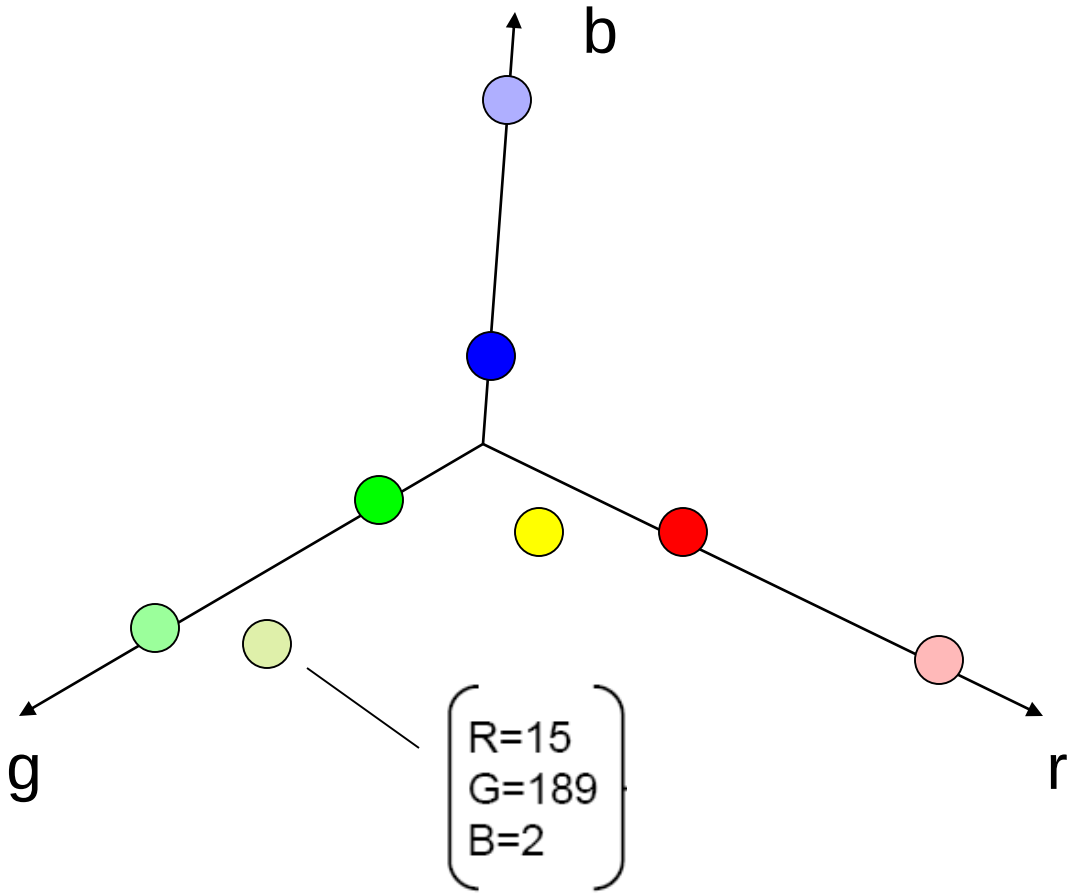
2. Codewords dictionary formation



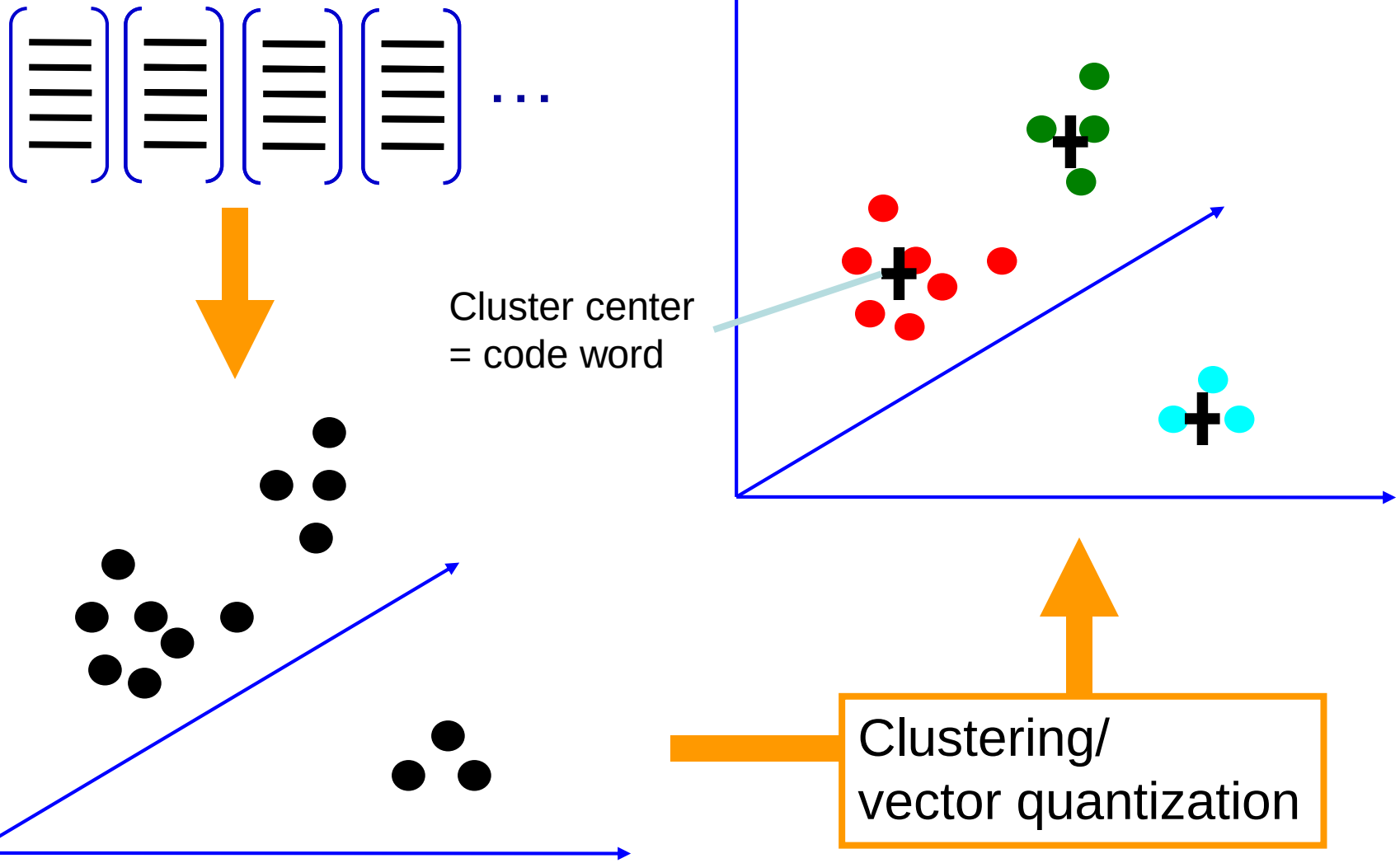
Example: color feature



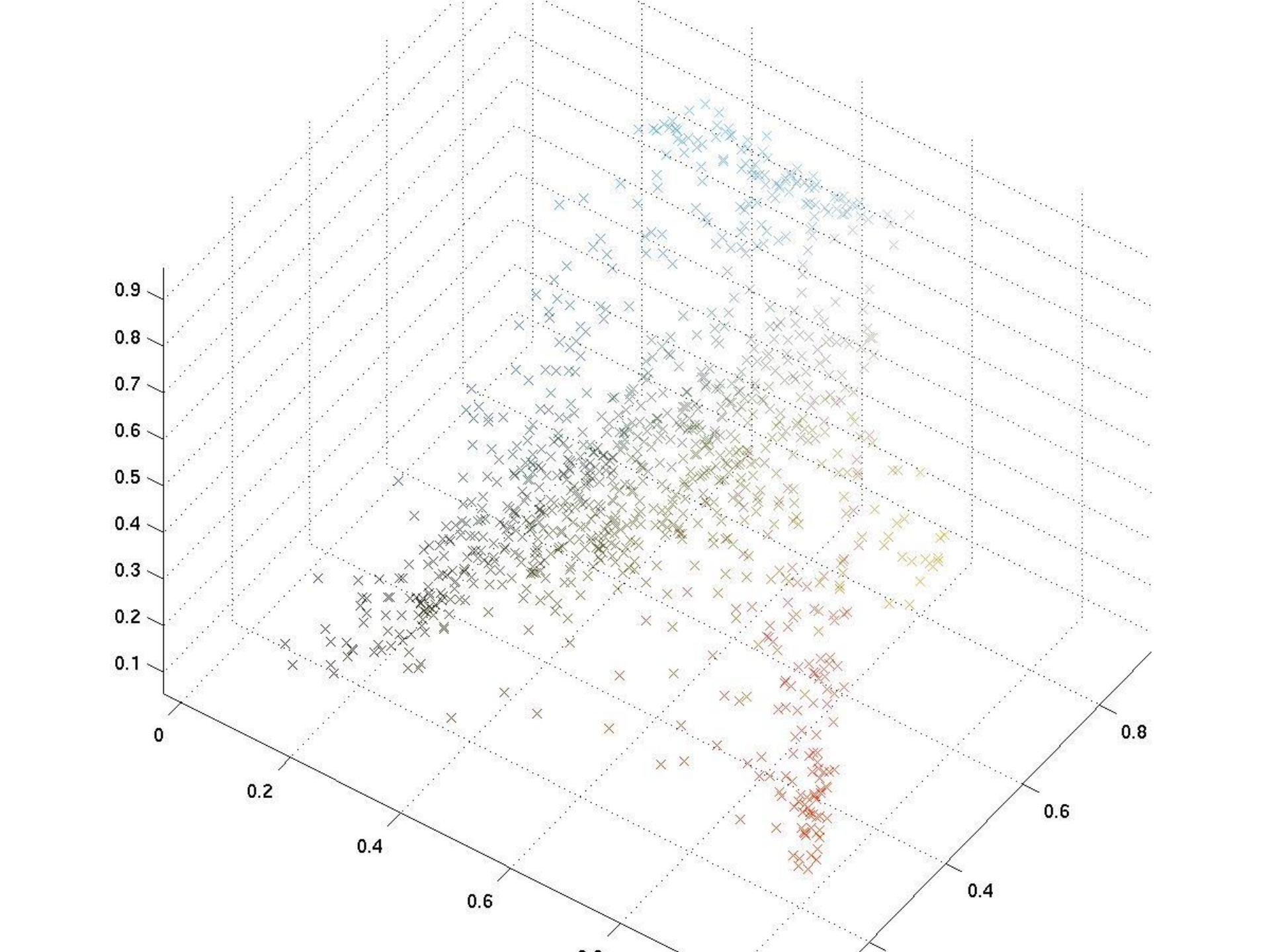
Example: color feature

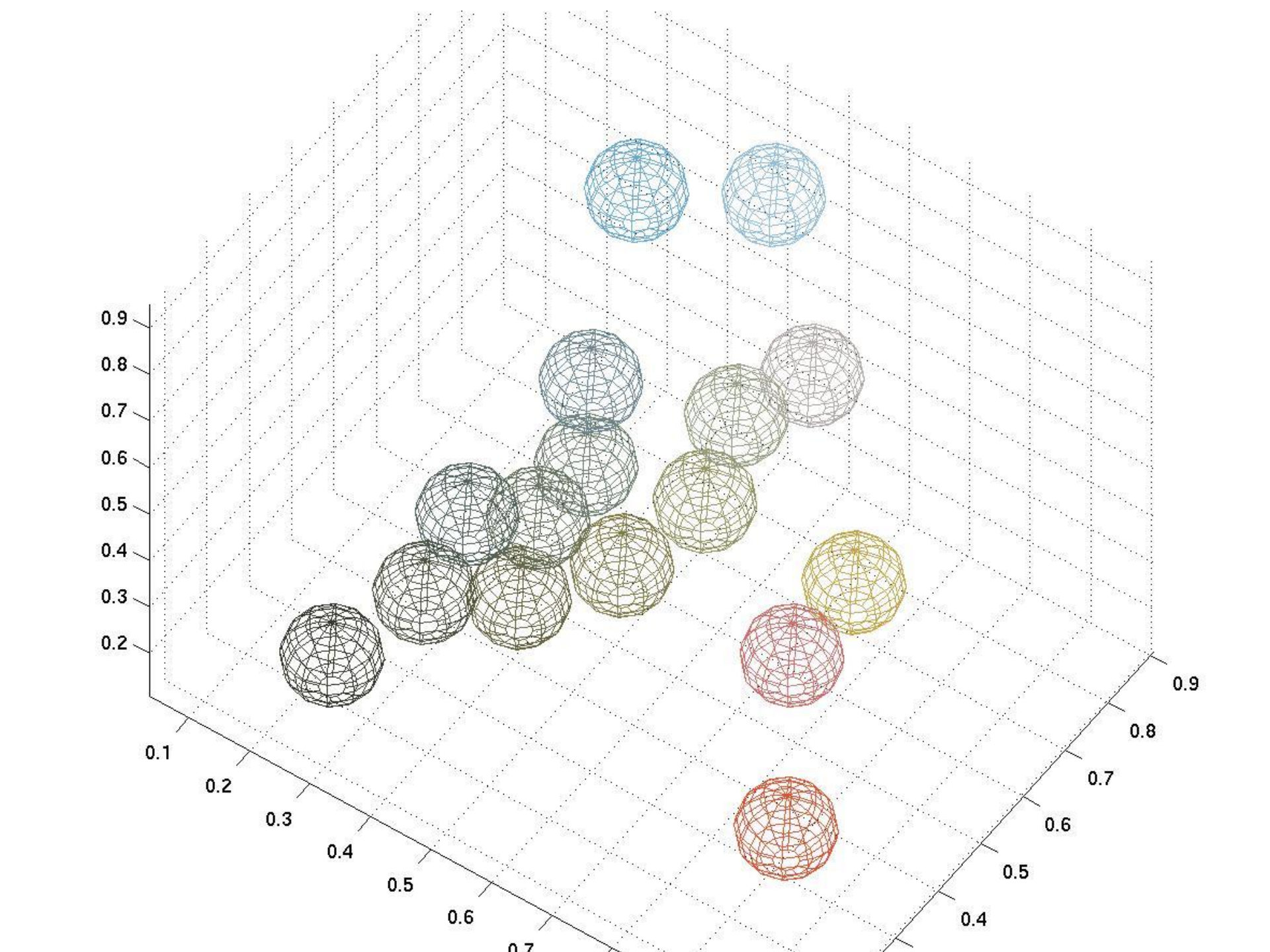


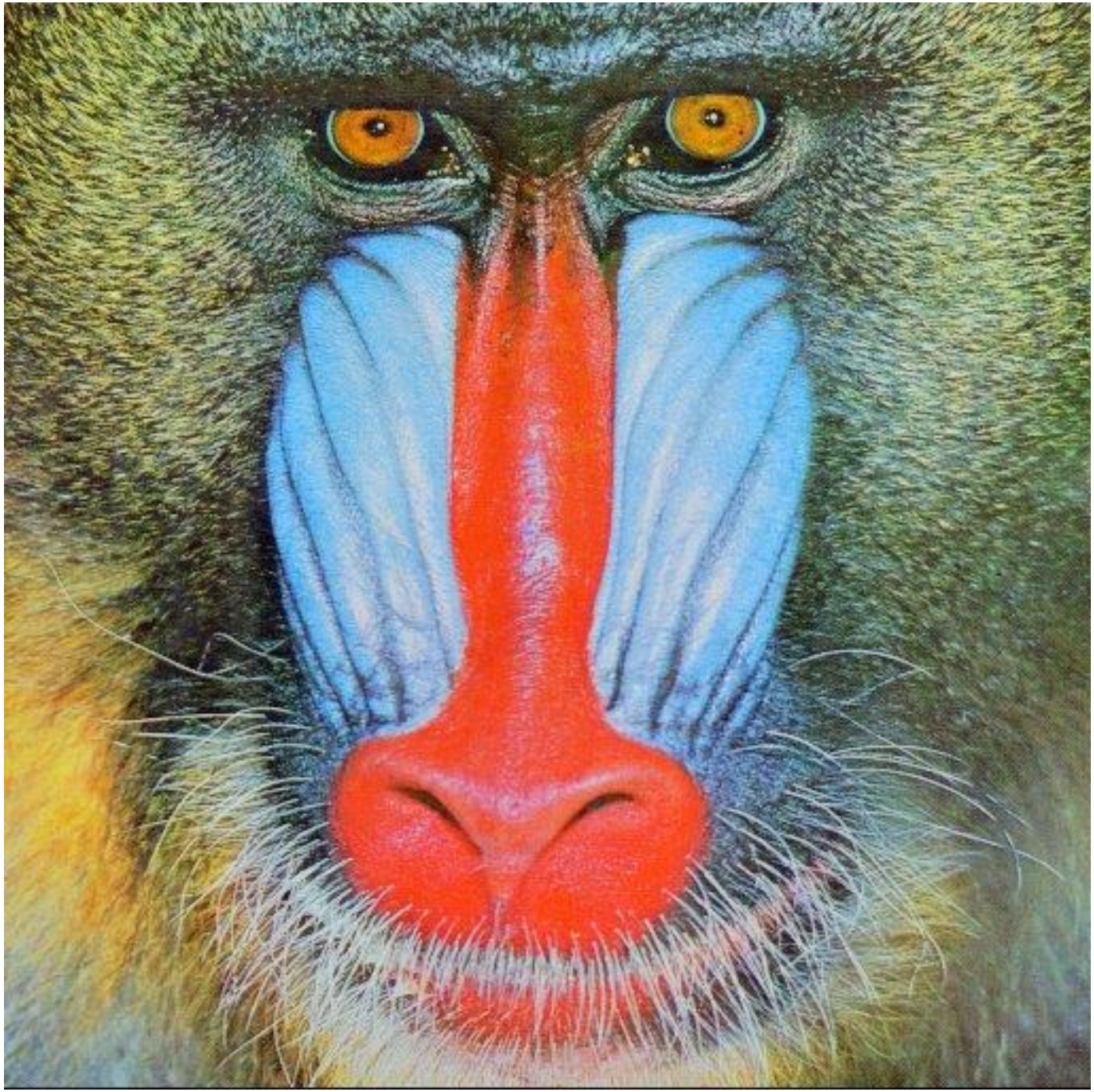
2. Codewords dictionary formation



E.g., Kmeans, see CS131A

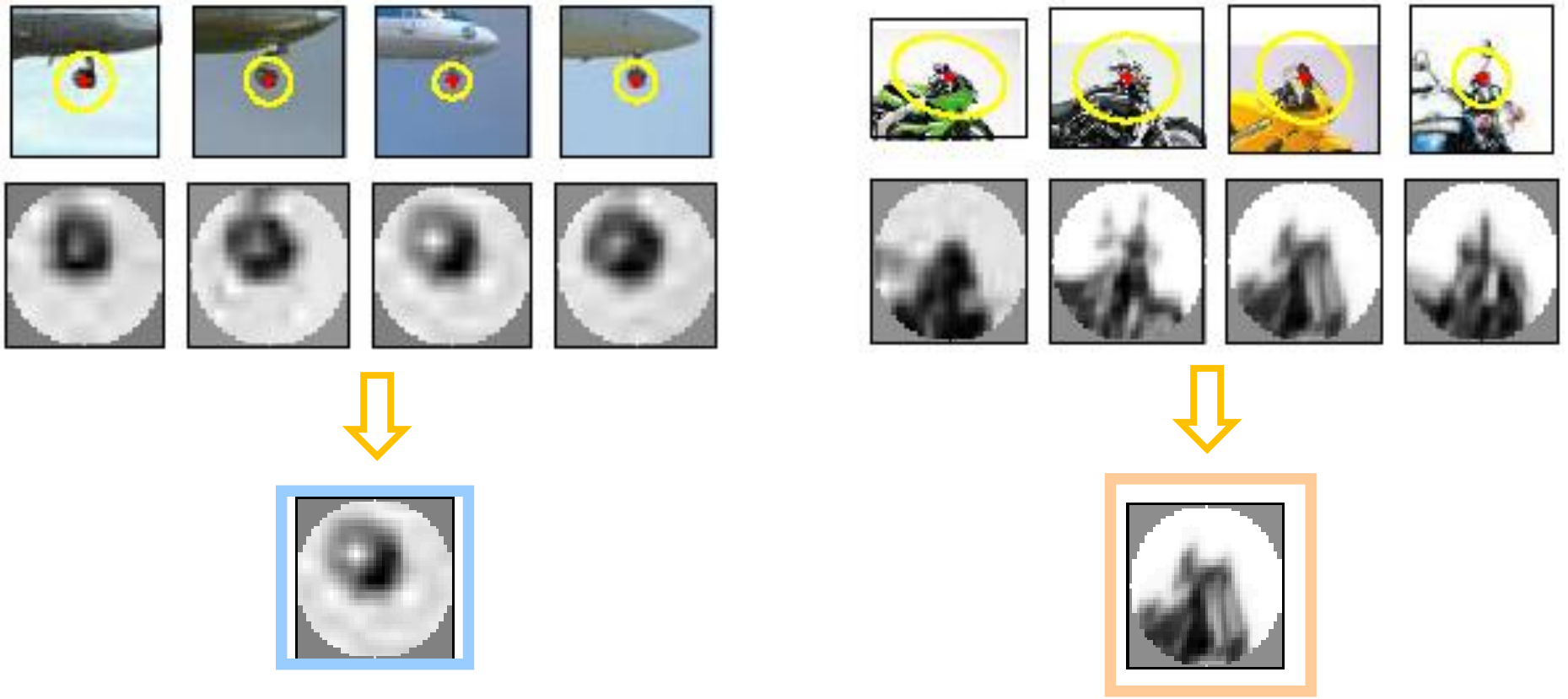

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
0.6
0.4


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



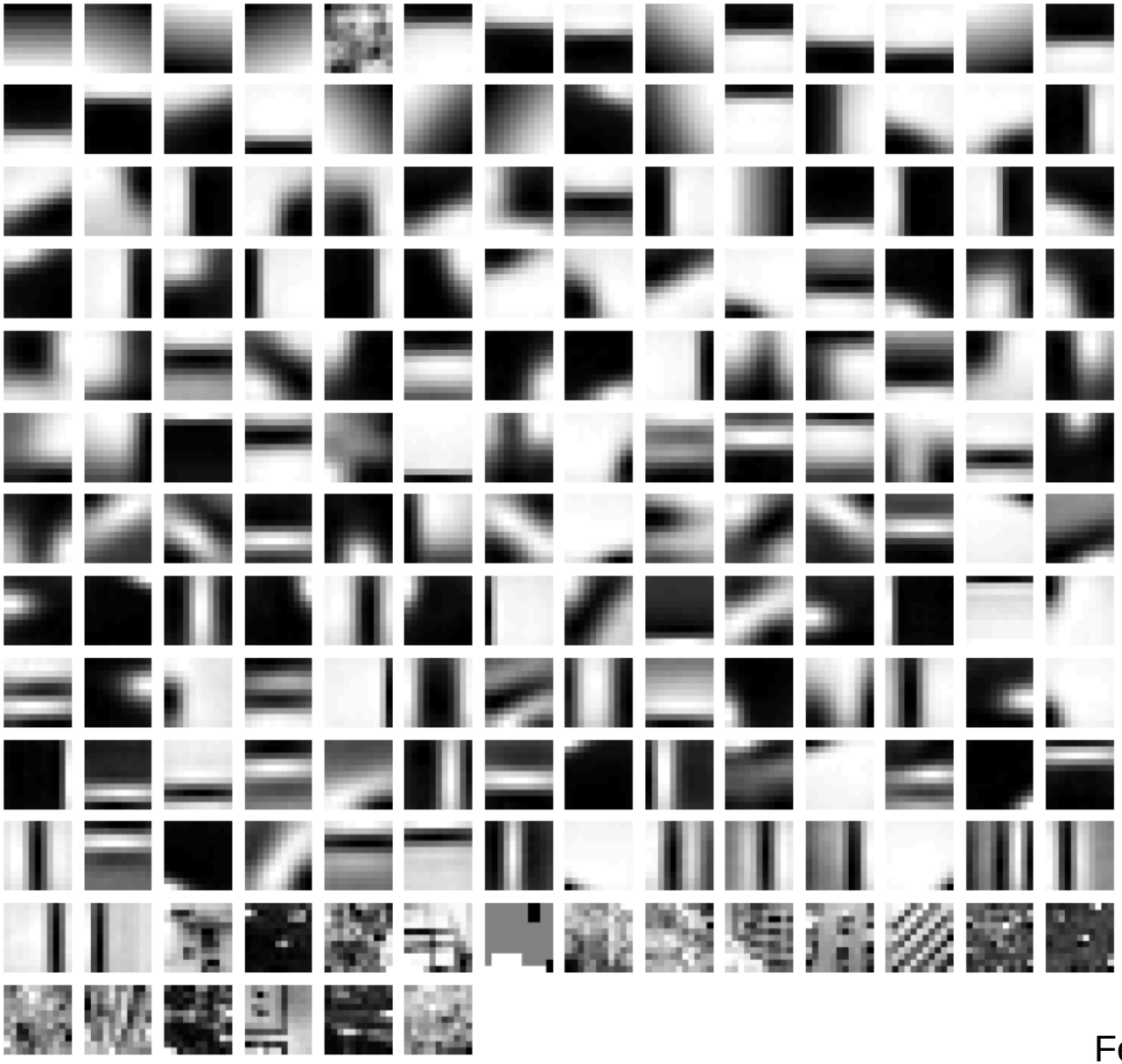
2. Codewords dictionary formation

- **Image patch examples of codewords**



Sivic et al. 2005

2. Codewords dictionary formation



Fei-Fei et al. 2005

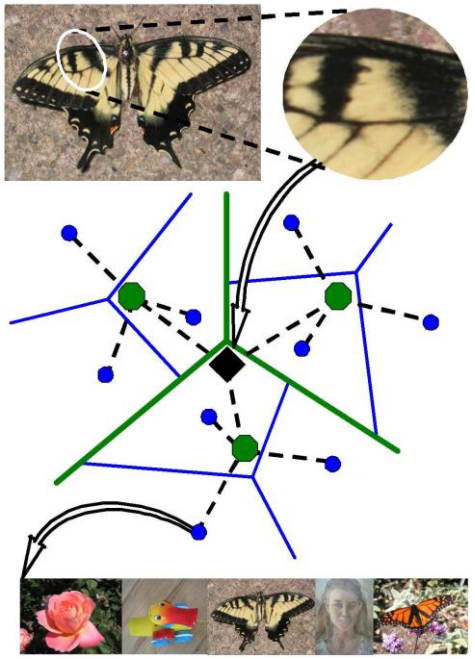
2. Codewords dictionary formation

- **Typically a codeword dictionary is obtained from a training set comprising all the object classes of interests**

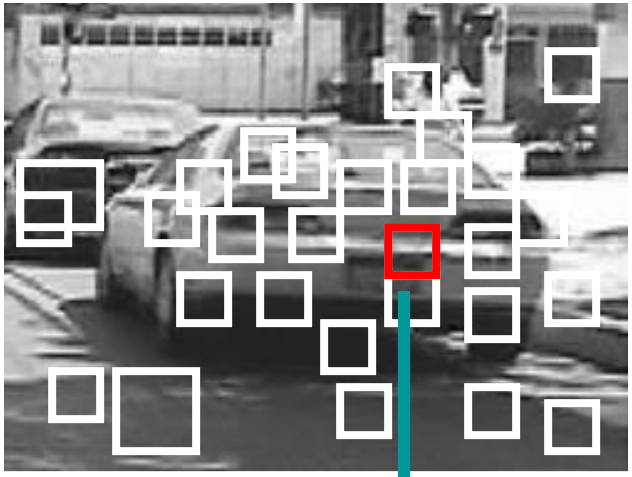
Visual vocabularies: Issues

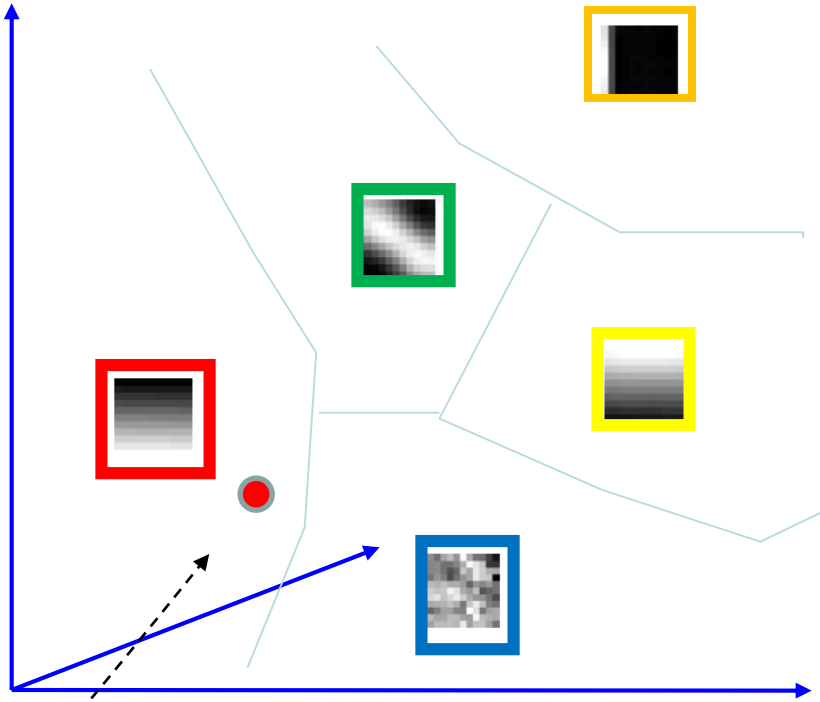
- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large: quantization artifacts, overfitting

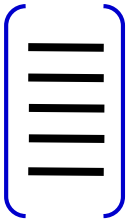
- Computational efficiency
  - Vocabulary trees
    (Nister & Stewenius, 2006)

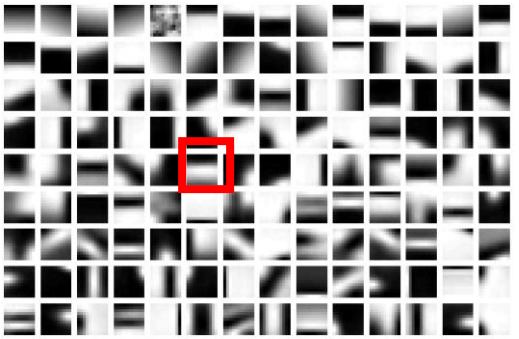


3. Bag of word representation





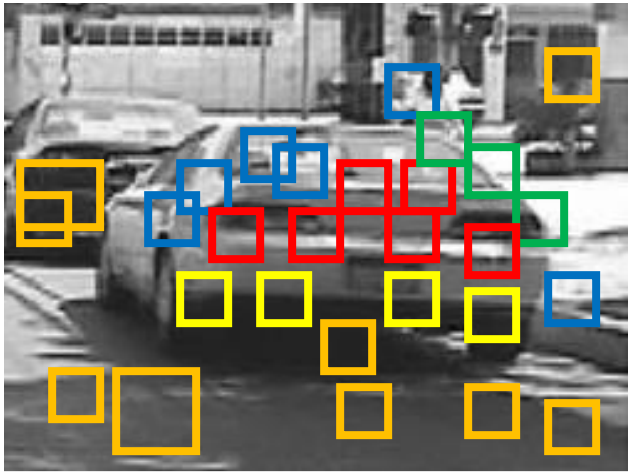


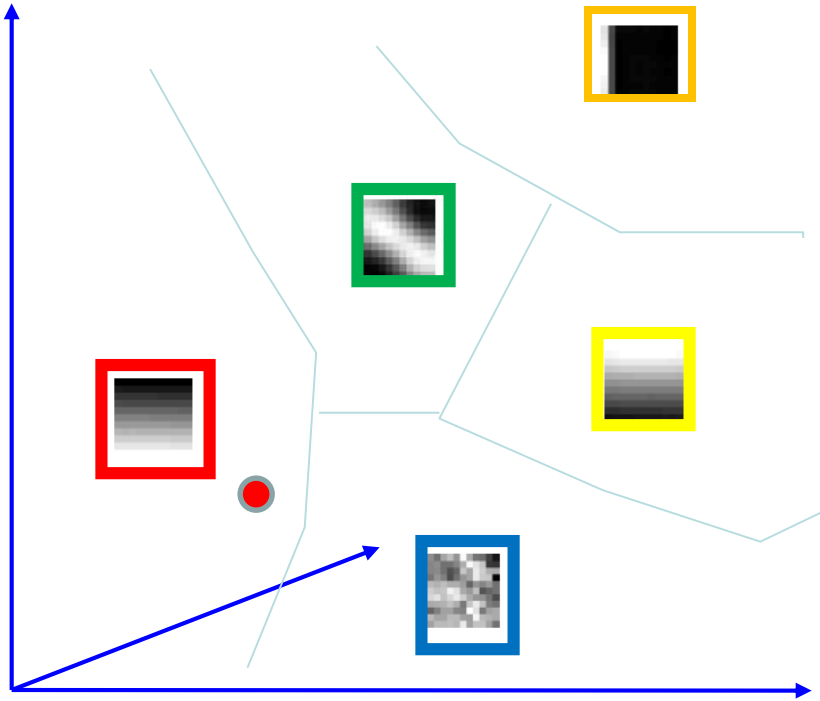


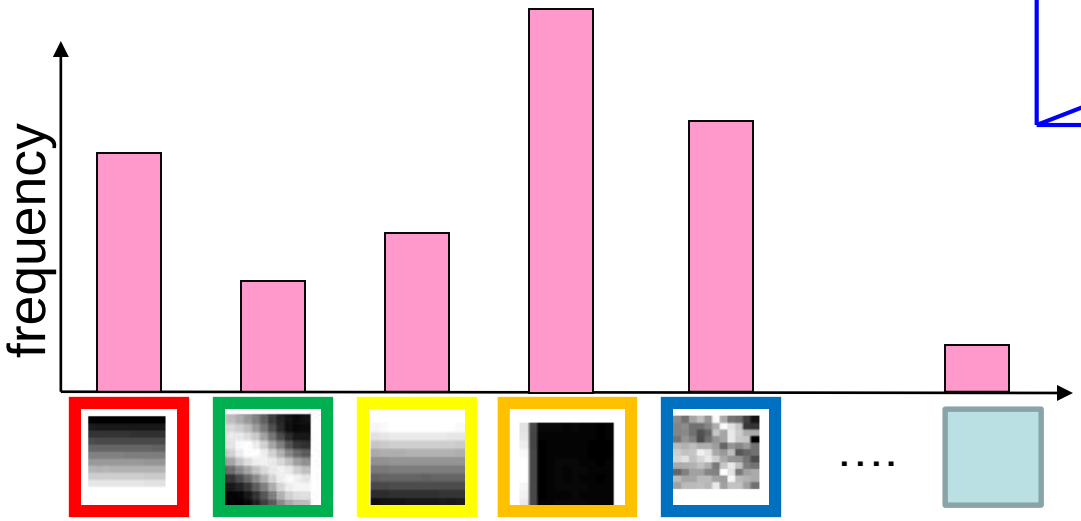
Codewords dictionary

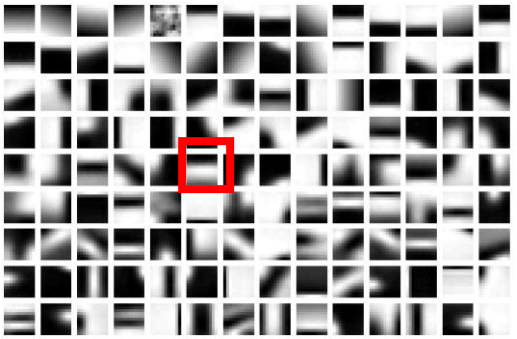
- Nearest neighbors assignment
- K-D tree search strategy

3. Bag of word representation





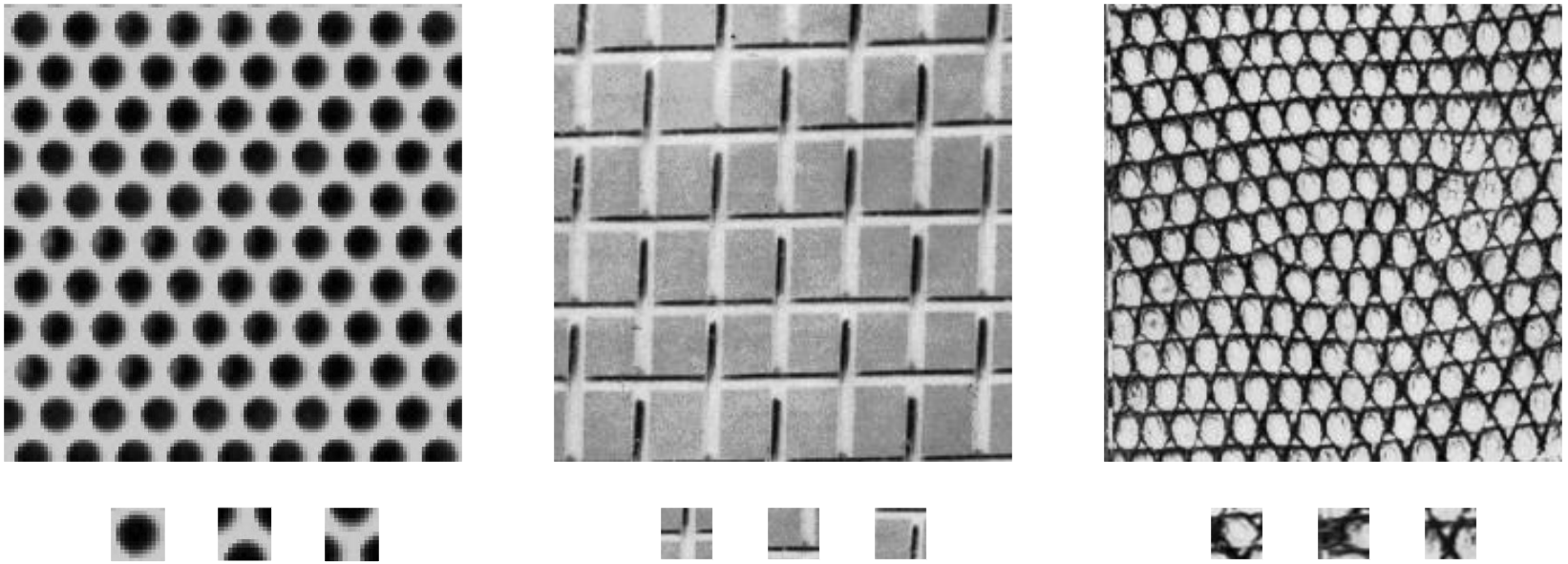




Codewords dictionary

Representing textures

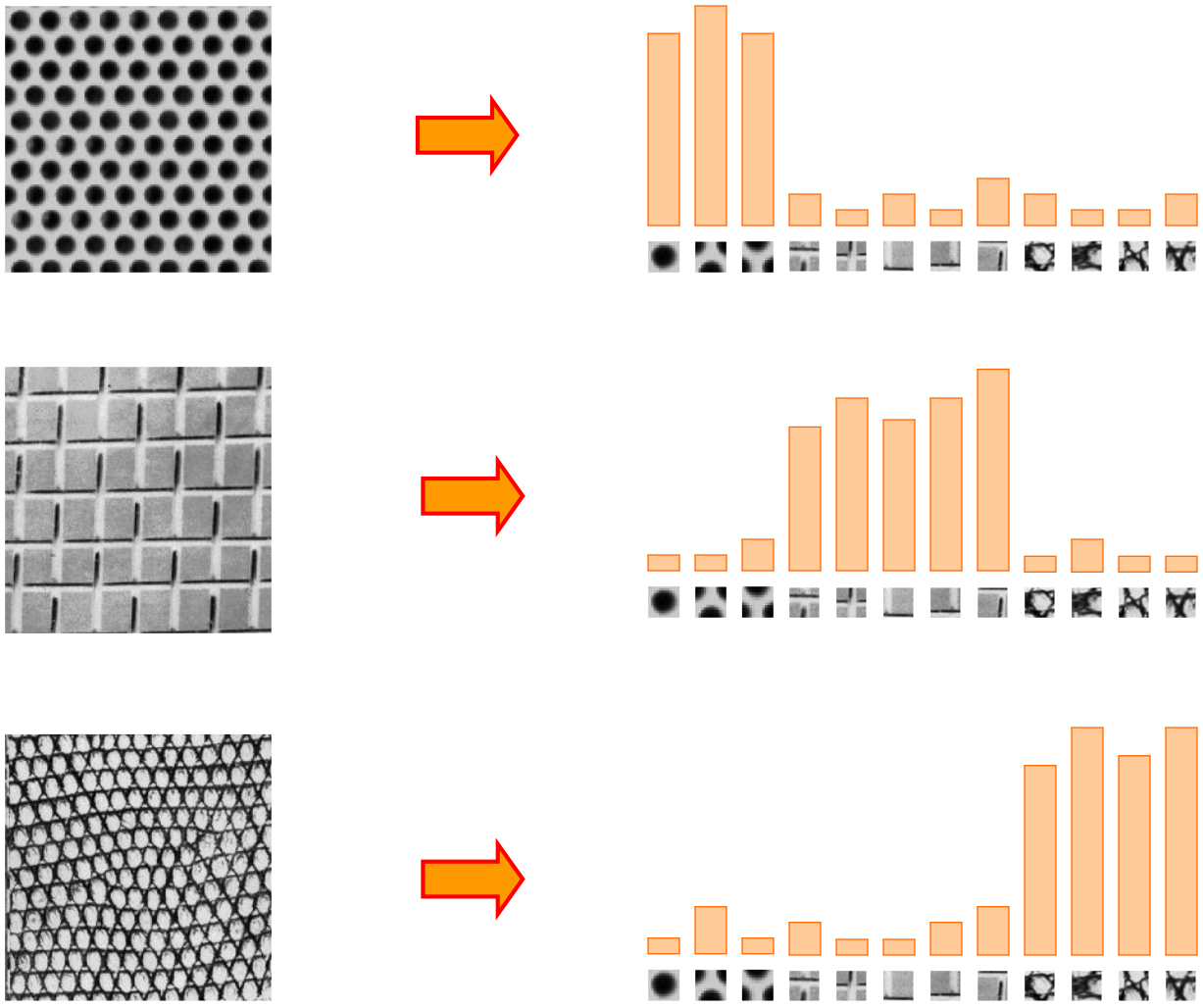
- Texture is characterized by the repetition of basic elements or *textons*
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



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

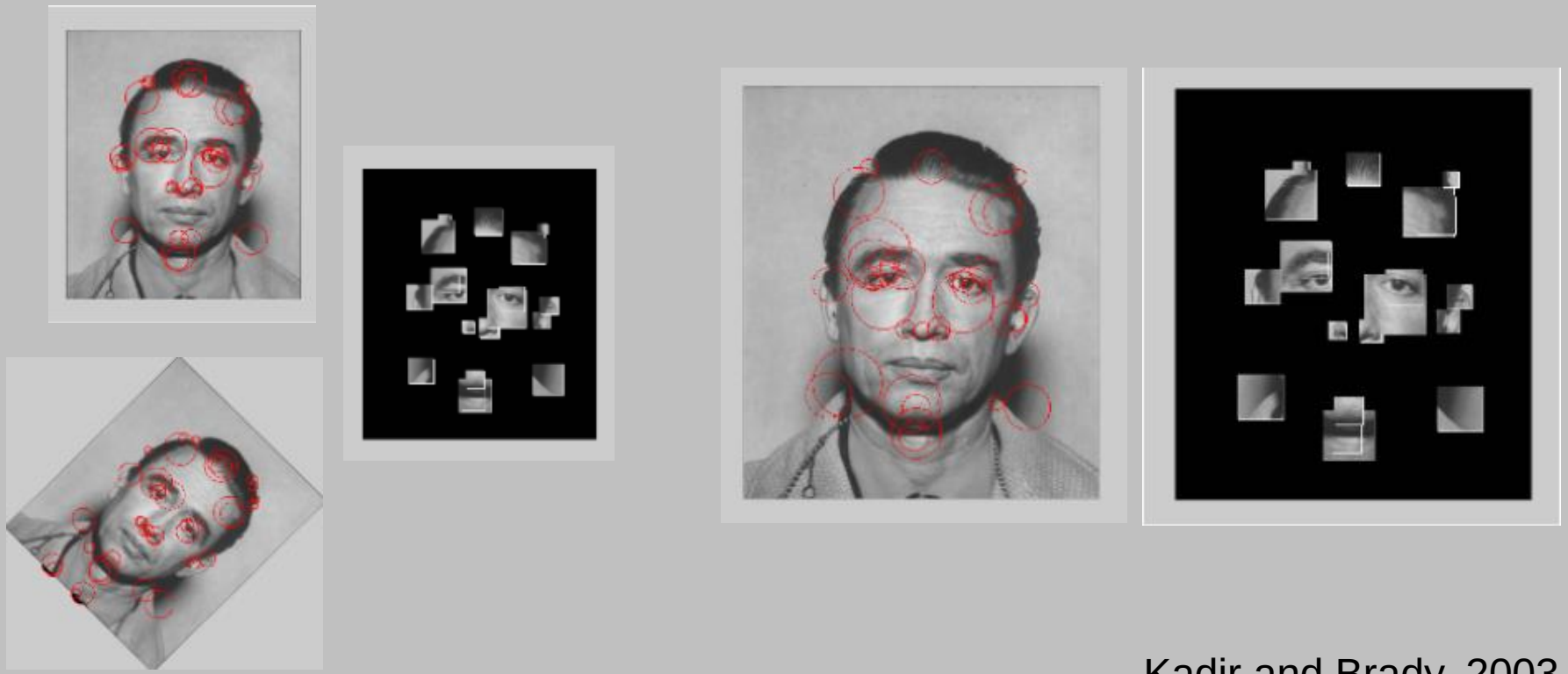


Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Credit slide: S. Lazebnik

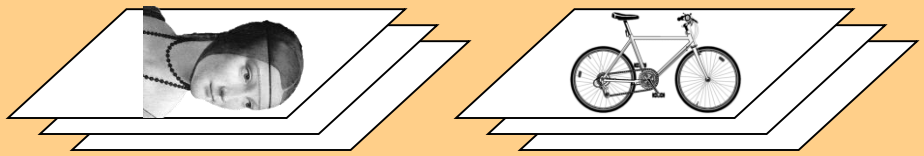
Invariance issues

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



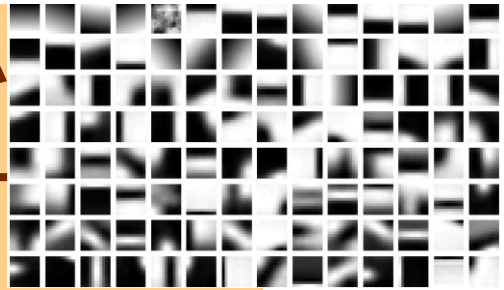
Kadir and Brady. 2003

Representation



2.

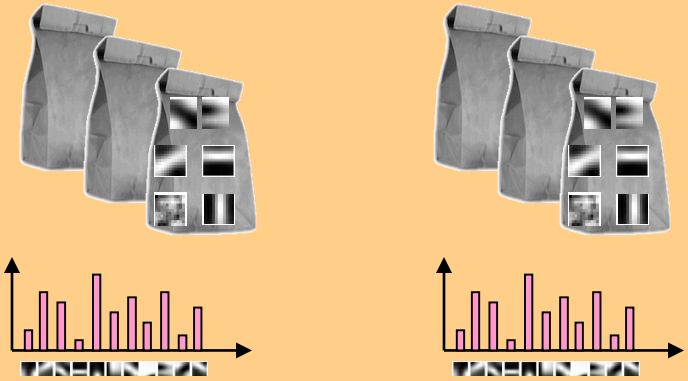
codewords dictionary



1. feature detection & representation

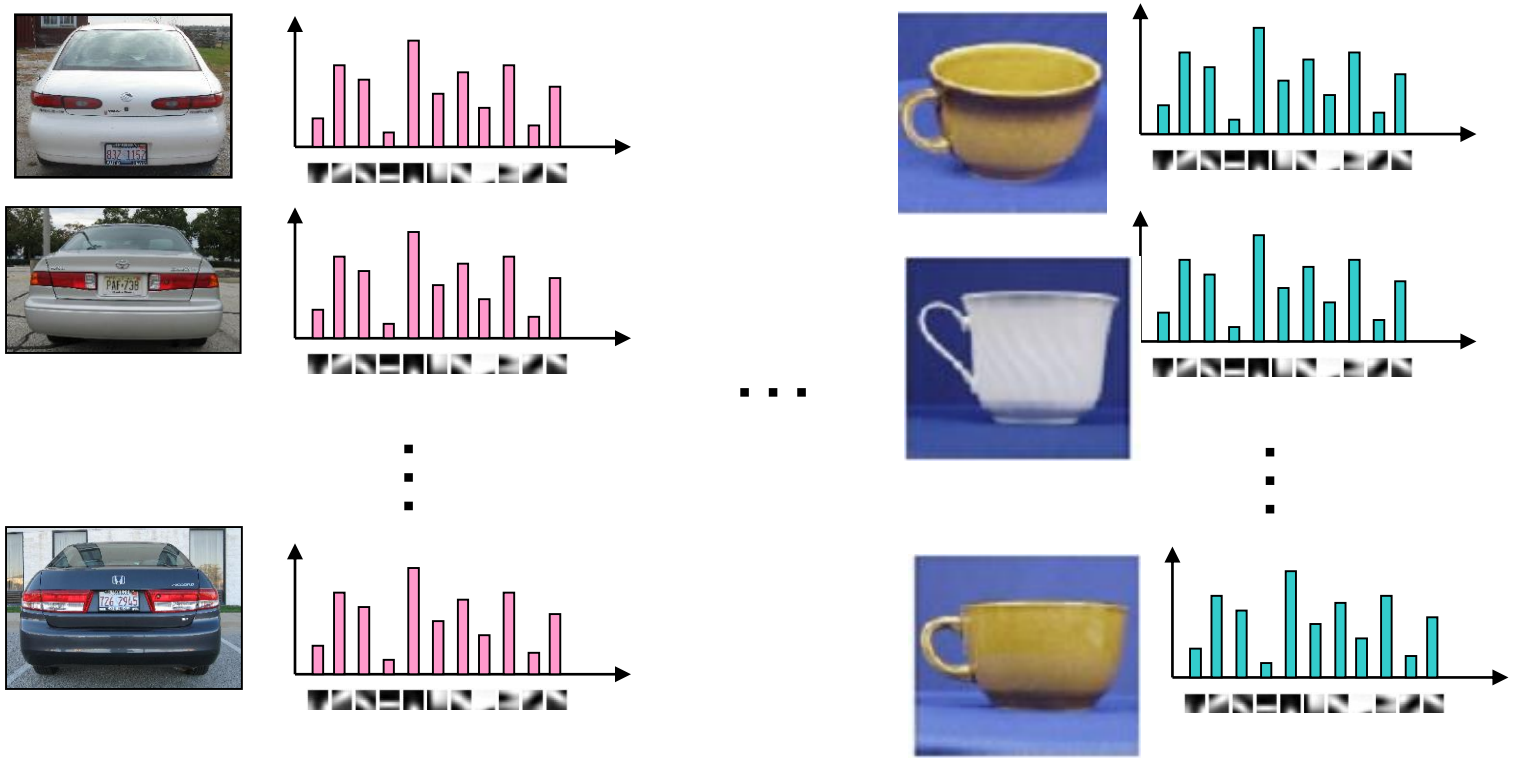
image representation

3.



category models

Category models

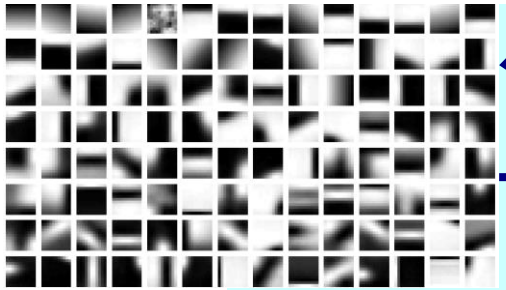


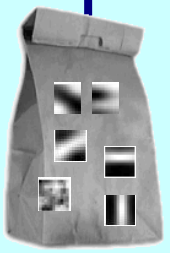
Class 1

Class N

Recognition

codewords dictionary





category models (and/or) classifiers

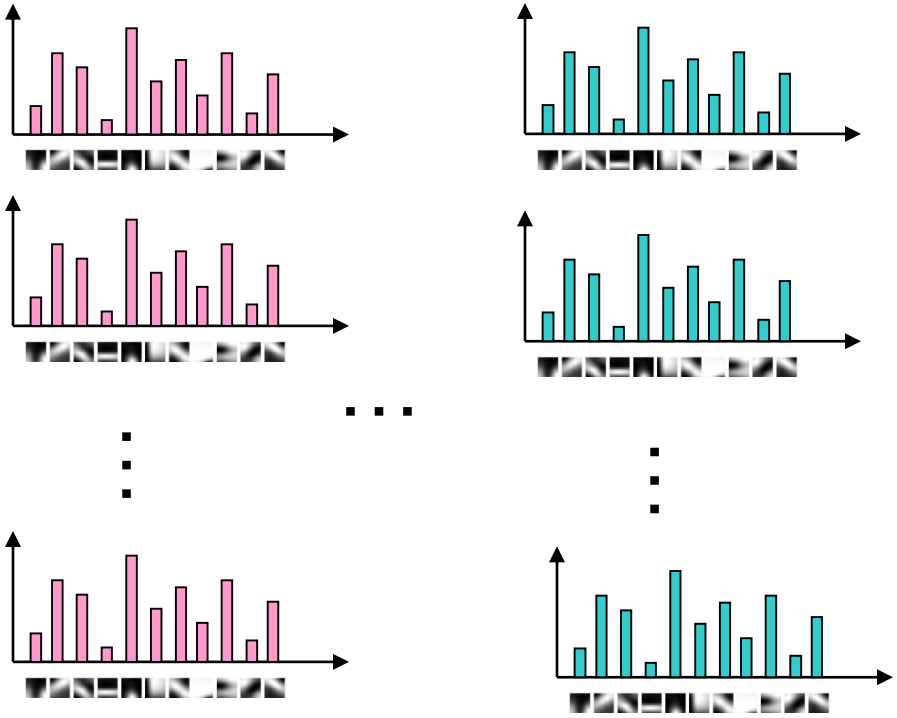
category decision

Lecture 12
Visual recognition

- Bag of words models for object recognition and classification
  - Discriminative methods
    - Nearest neighborhood
    - Linear classifier
    - SVM
  - Generative methods

Discriminative classifiers

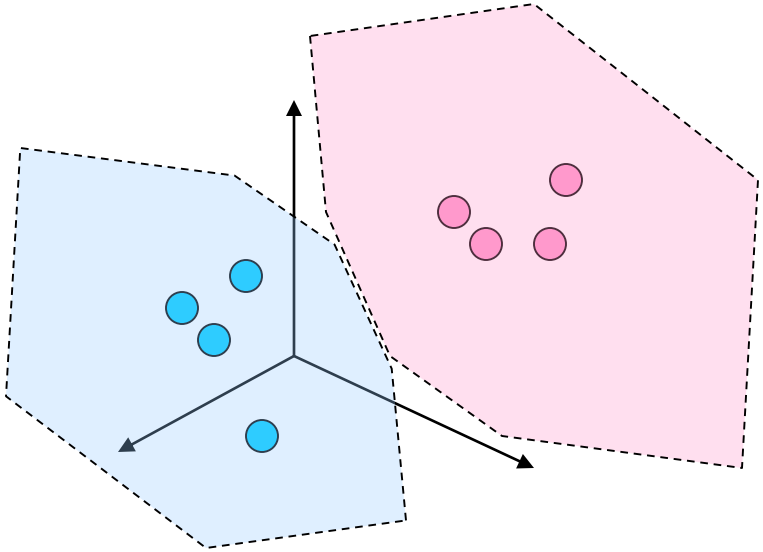
category models



Class 1

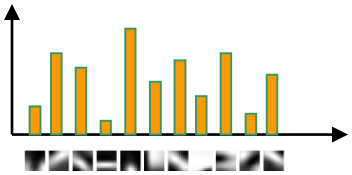
Class N

Model space



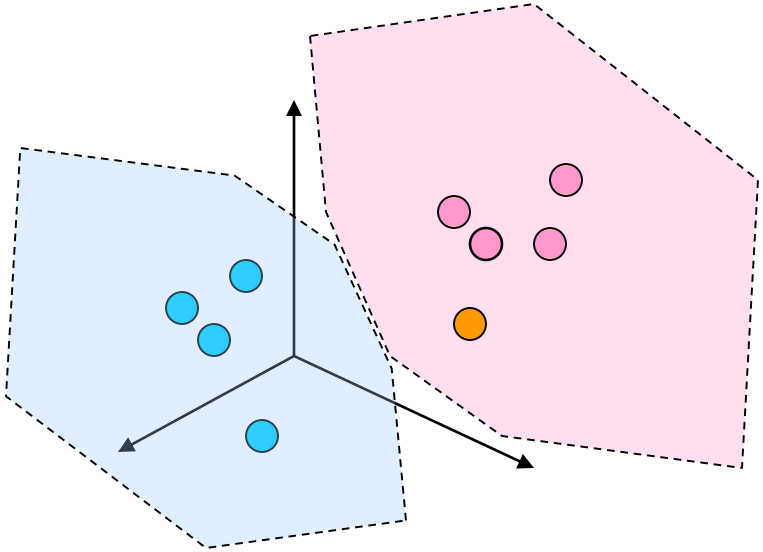
Discriminative classifiers

Query image



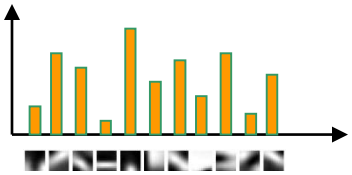
Winning class: pink

Model space



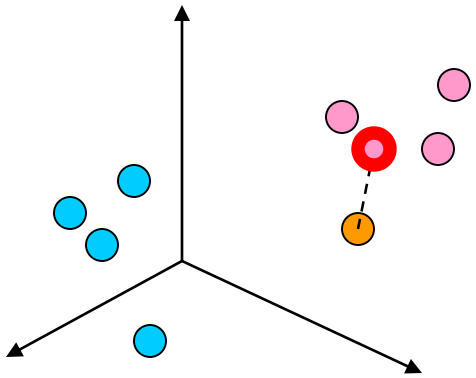
Nearest Neighbors classifier

Query image



Winning class: pink

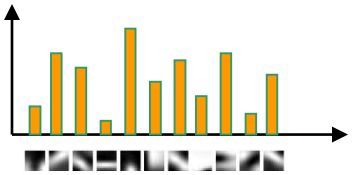
Model space



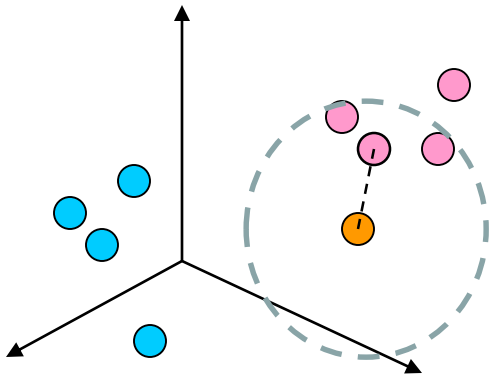
- Assign label of nearest training data point to each test data point

K- Nearest Neighbors classifier

Query image



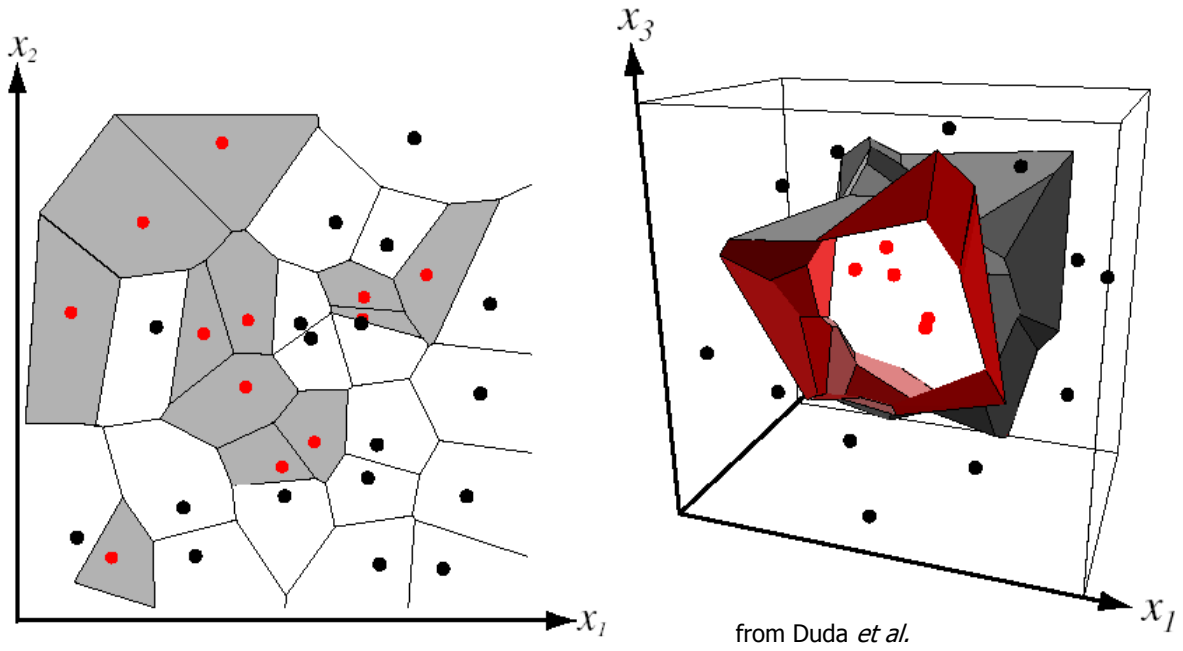
Model space



Winning class: pink

- For a new point, find the k closest points from training data
- Labels of the k points “vote” to classify
- Works well provided there is lots of data and the distance function is good

K- Nearest Neighbors classifier



from Duda *et al.*

- Voronoi partitioning of feature space for 2-category 2-D and 3-D data

- For k dimensions: k-D tree = space-partitioning data structure for organizing points in a *k*-dimensional space
- Enable efficient search
- Nice tutorial: http://www.cs.umd.edu/class/spring2002/cmsc420-0401/pbasic.pdf

Functions for comparing histograms

Jan Puzicha, Yossi Rubner, Carlo Tomasi, Joachim M. Buhmann: Empirical Evaluation of Dissimilarity Measures for Color and Texture. ICCV 1999

- L1 distance

$$D(h_1, h_2) = \sum_{i=1}^{N} | h_1(i) - h_2(i) |$$

- χ^2 distance

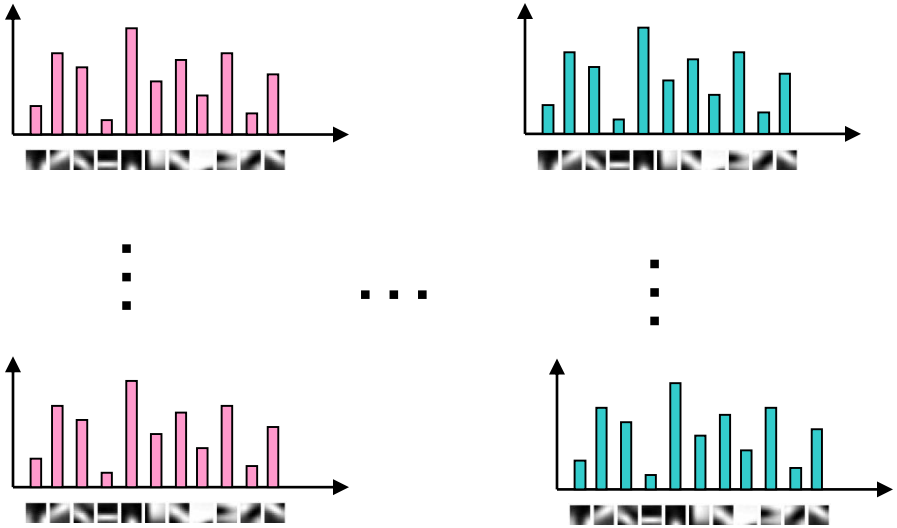
$$D(h_1, h_2) = \sum_{i=1}^{N} \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)}$$

- Quadratic distance (*cross-bin*)

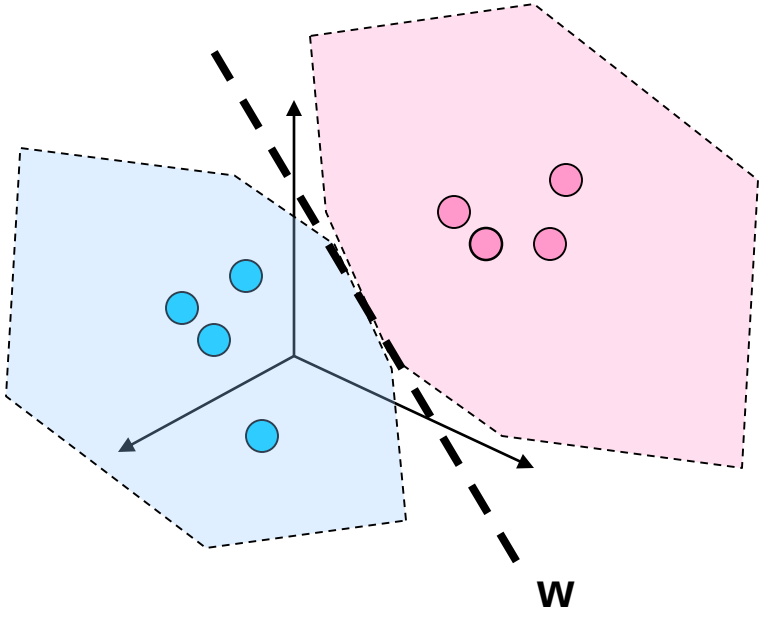
$$D(h_1, h_2) = \sum_{i,j} A_{ij} (h_1(i) - h_2(j))^2$$

Discriminative classifiers (linear classifier)

category models



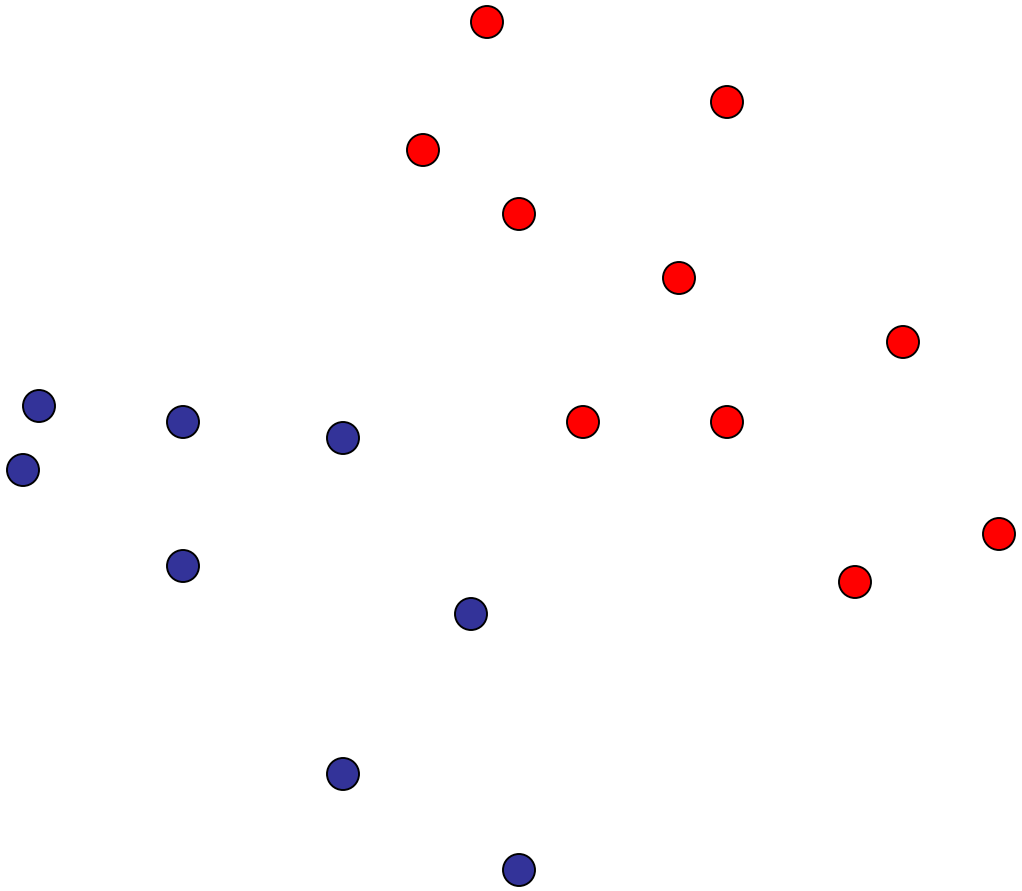
Model space



- For a linear classifier, the training data is used to learn **w** and then discarded
- Only **w** is needed for classifying new data

Linear classifiers

- We want to classify two classes of points
- Each point x_i can have two labels y_i: {-1, +1}

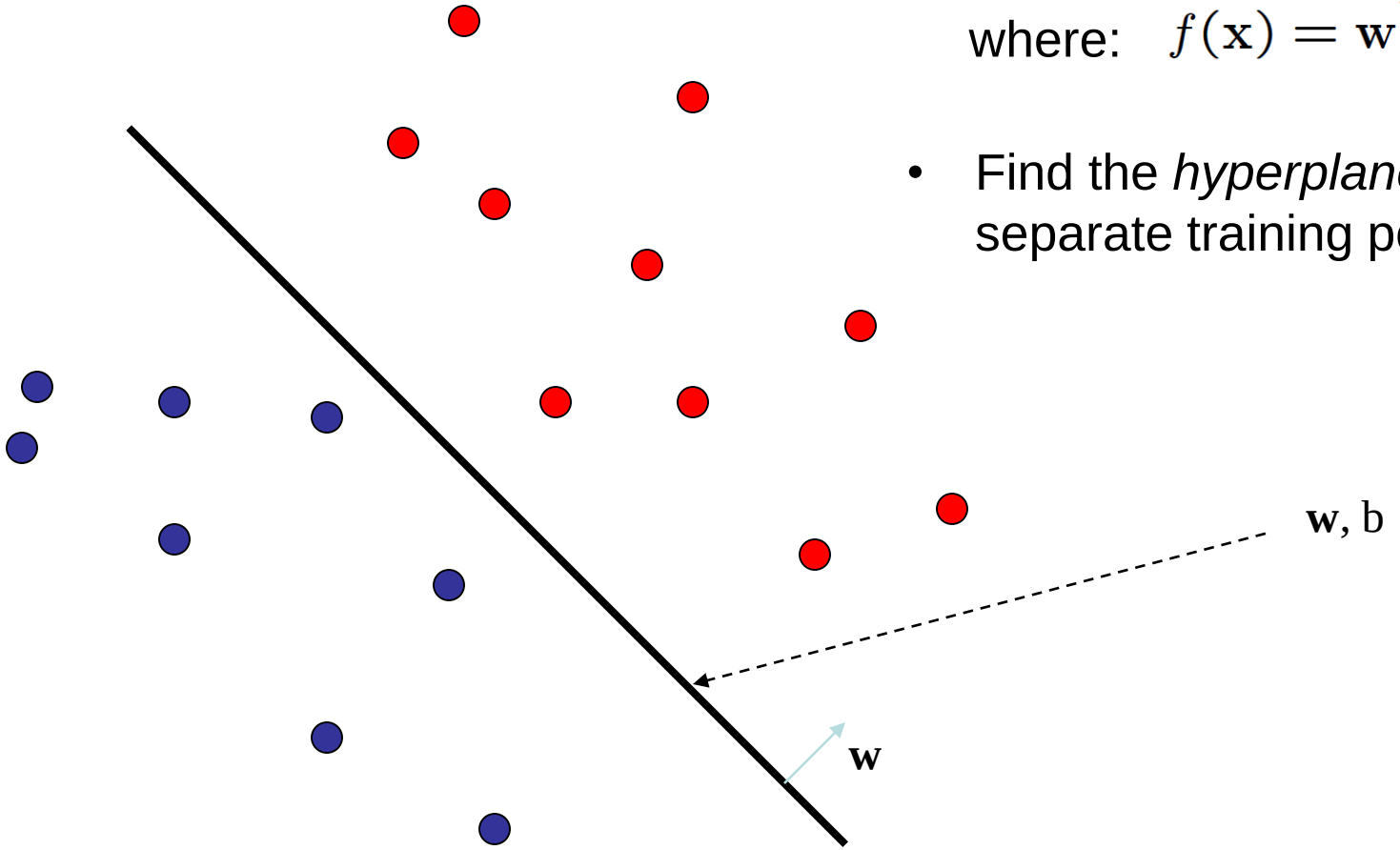


Linear classifiers

- GOAL: learn a classifier f(X) such that: $f(\mathbf{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$

where: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$

- Find the *hyperplane* (w,b) to separate training points

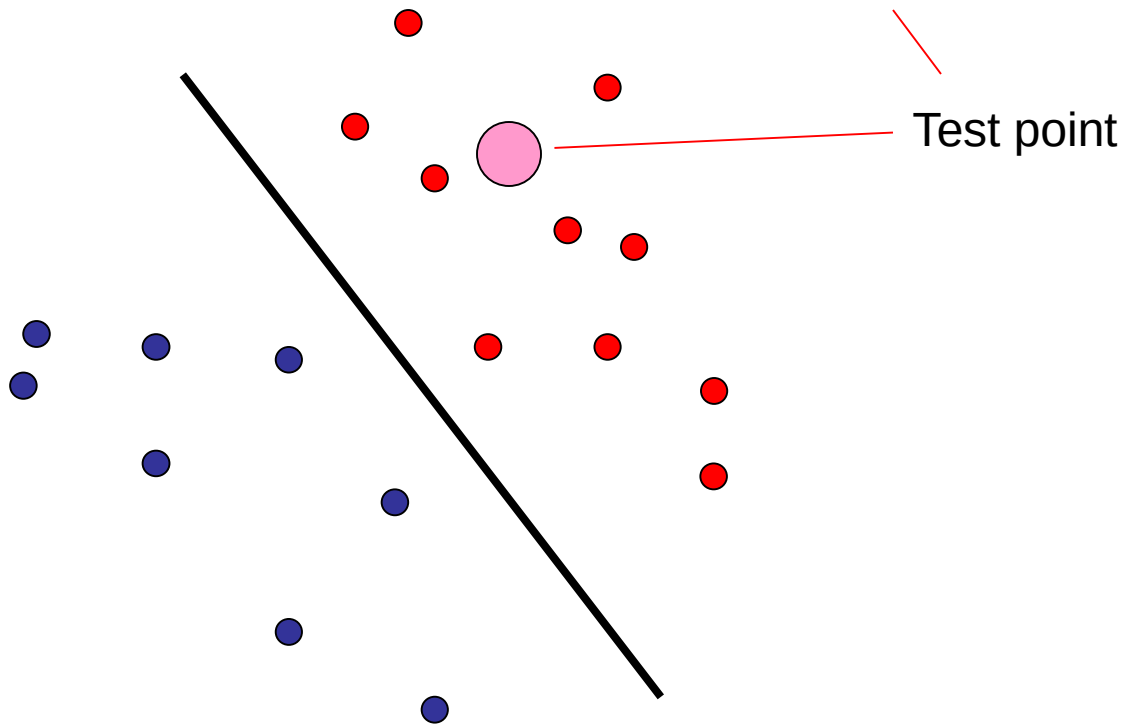


Linear classifiers

- Once w,b are learnt, we can do classification:

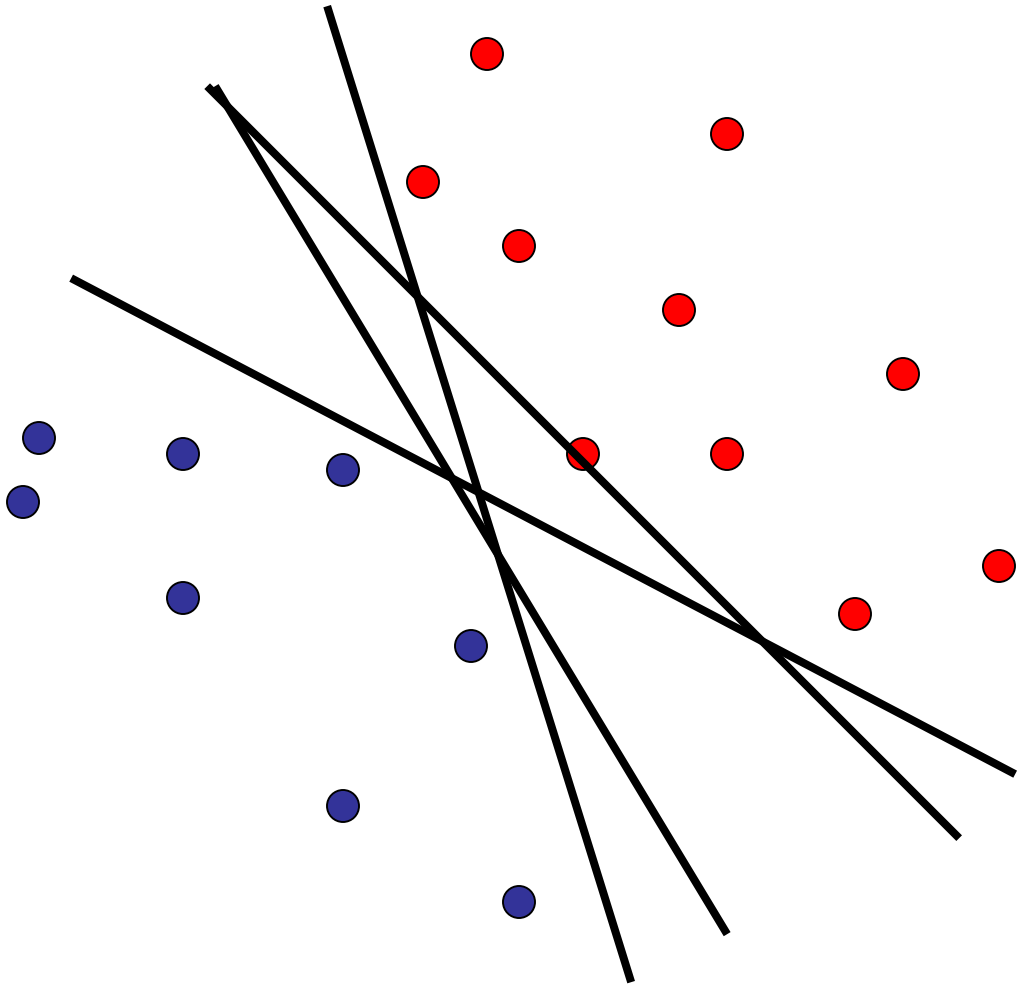
$$if \quad \mathbf{x}\cdot\mathbf{w}+b \geq 0 \rightarrow \quad class\ 1$$

$$if \quad \mathbf{x}\cdot\mathbf{w}+b < 0 \rightarrow \quad class\ 2$$



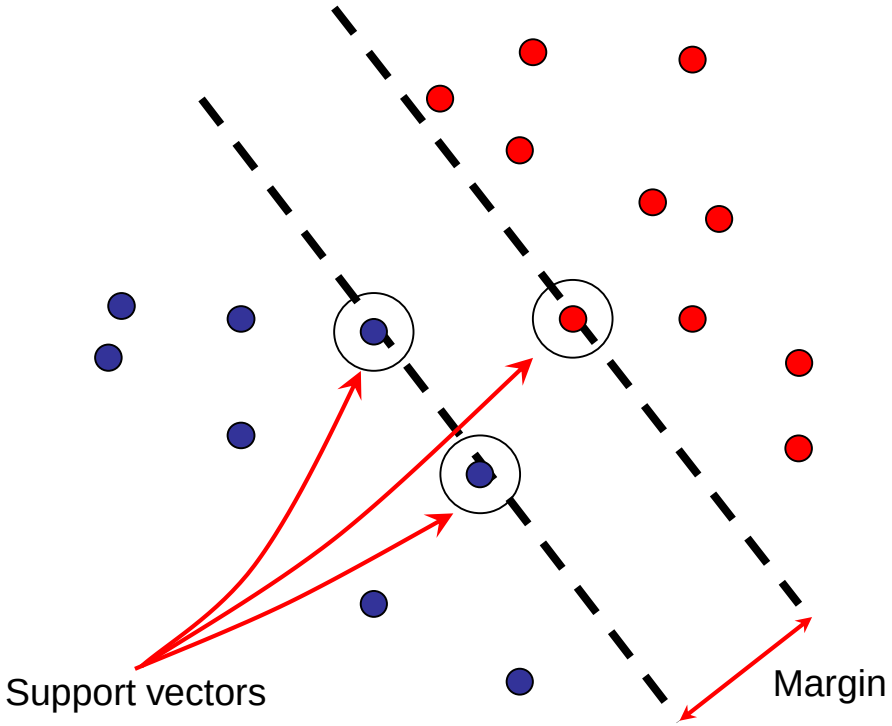
Linear classifiers

Which hyperplane is best?



Support vector machines

- Select two hyperplanes such:
  - They separate the training points
  - There are no points between them
  - Their distance is maximized



Support vectors: $\mathbf{x}_i \cdot \mathbf{w} + b = \pm 1$

Distance between point and hyperplane: $\frac{|\mathbf{x}_i \cdot \mathbf{w} + b|}{\|\mathbf{w}\|}$

Margin = $2 / \|\mathbf{w}\|$

Solution:

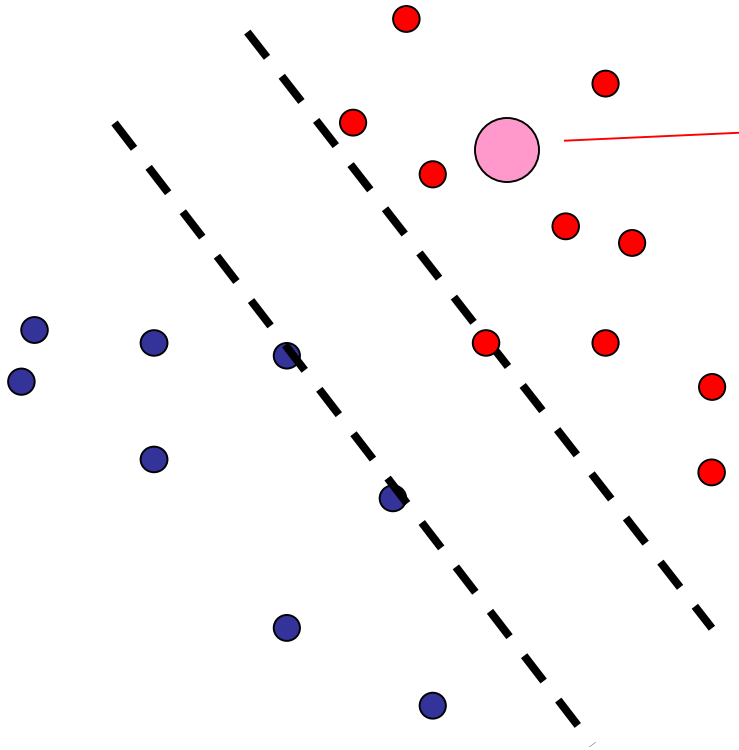
$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$b = \frac{1}{N} \sum_{i-1}^{N} (\mathbf{x}_i \cdot \mathbf{w} - y_i)$$

- The region bounded by them is called "the margin".
- Maximum margin solution: most stable under perturbations of the inputs

Support vector machines

- Classification: $f(\mathbf{x}) = \sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b$



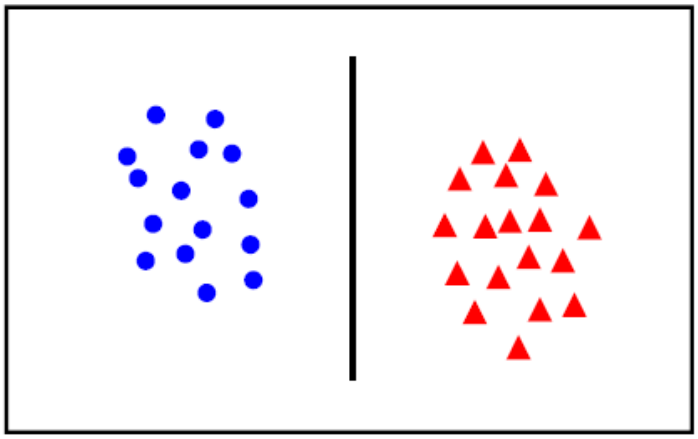
$$if \quad f(\mathbf{x}) \geq 0 \rightarrow \quad \mathbf{x} \in class\ 1$$

$$if \quad f(\mathbf{x}) < 0 \rightarrow \quad \mathbf{x} \in class\ 2$$

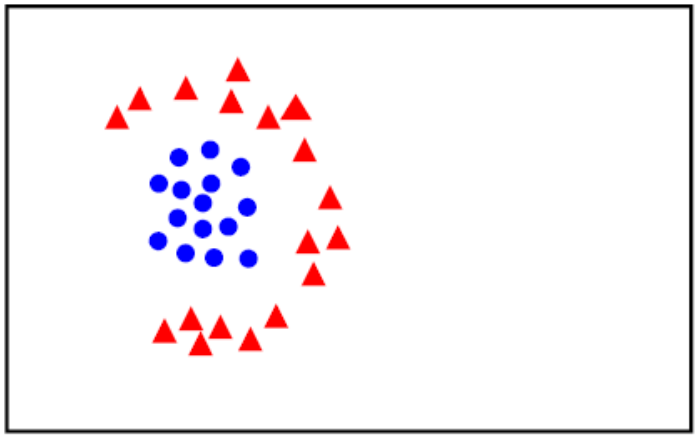
C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998

Linear separability

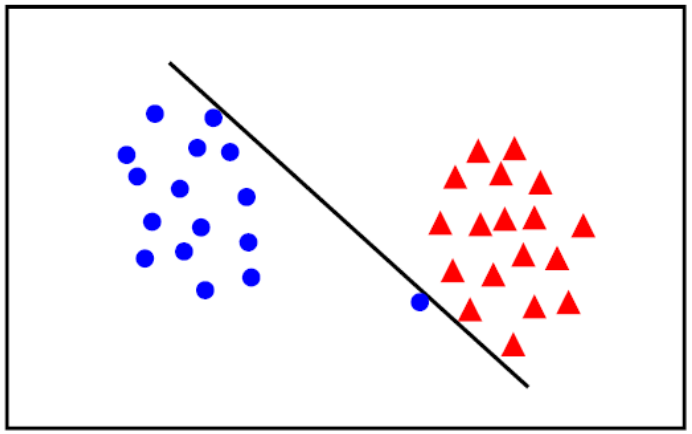
Linearly separable



Linearly **not** separable



Linearly separable with small margins



Courtesy of A Zisserman

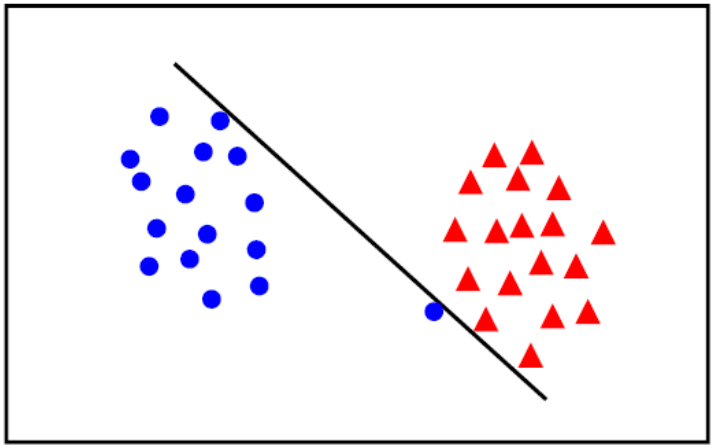
Linear separability

Two possible solutions:

- Introduce soft variables (through slack variables)

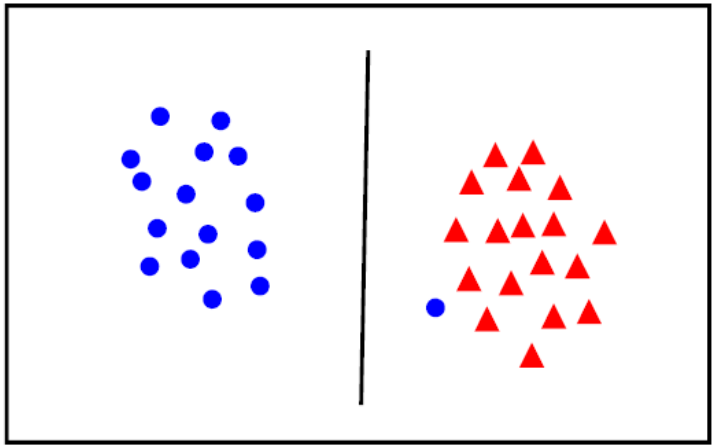
- Non linear separation function (e.g., non linear SVM)

Soft margins



The points can be linearly separated but there is a very narrow margin

Or they cannot be separated at all



IDEA: still seek at large margin solution, even though one constraint is violated

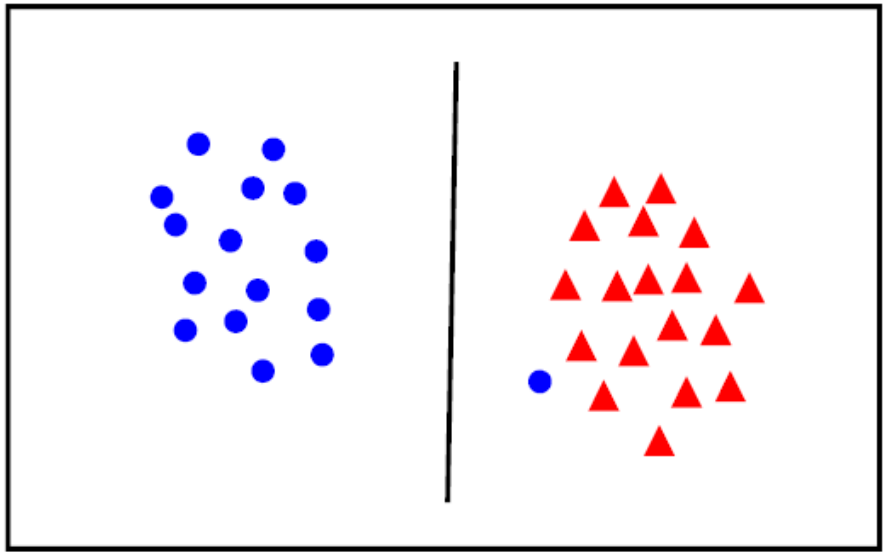
In general there is a trade off between the margin and the number of mistakes on the training data

Courtesy of A Zisserman

Soft margins

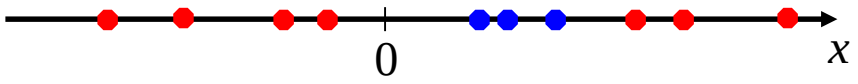
By Corinna Cortes and Vladimir N. Vapnik, 1995

- Use soft margin violations instead of the hard one: Find hyperplanes that split the examples as “cleanly” as possible, while still maximizing the distance to the nearest cleanly split examples.
- Introduce slack variables to measure the degree of misclassification of the data

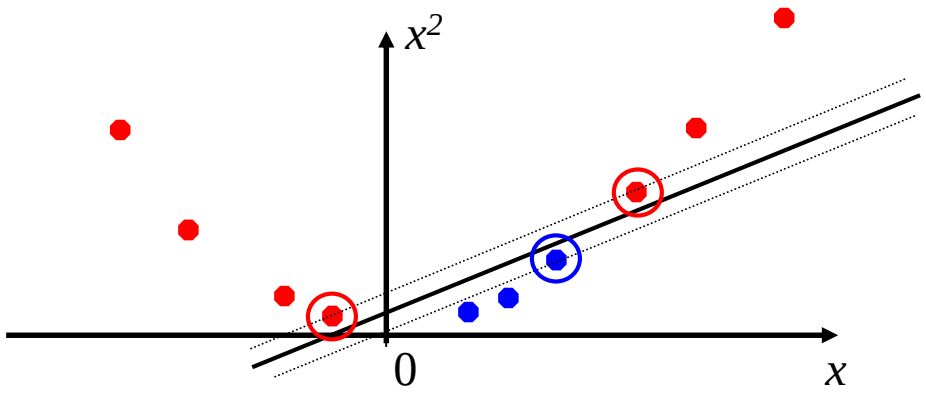


Nonlinear SVMs

- Given a non-linearly separable dataset:



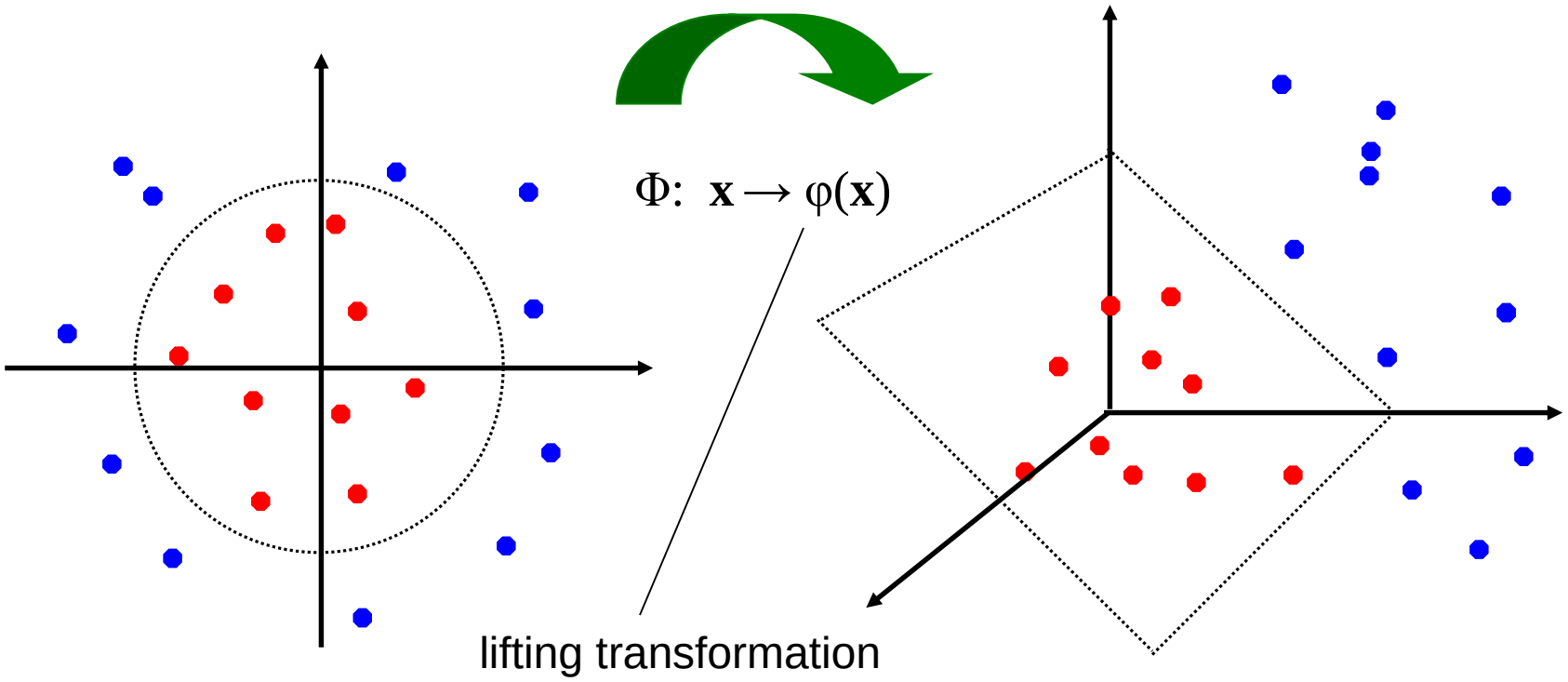
- Map it to a higher-dimensional space:



Slide credit: Andrew Moore

Nonlinear SVMs

- General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable:



Slide credit: Andrew Moore

Nonlinear SVMs

- Nonlinear decision boundary in the original feature space:

$$\sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b \longrightarrow \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

- *The kernel K* = product of the lifting transformation $\boldsymbol{\varphi}(\mathbf{x})$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x}_i) \cdot \boldsymbol{\varphi}(\mathbf{x}_j)$$

NOTE:
- It is not required to compute $\boldsymbol{\varphi}(\mathbf{x})$ explicitly:
- The kernel must satisfy the “Mercer inequality”

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998

Kernels for bags of features

- Histogram intersection kernel:

$$I(h_1, h_2) = \sum_{i=1}^{N} \min(h_1(i), h_2(i))$$

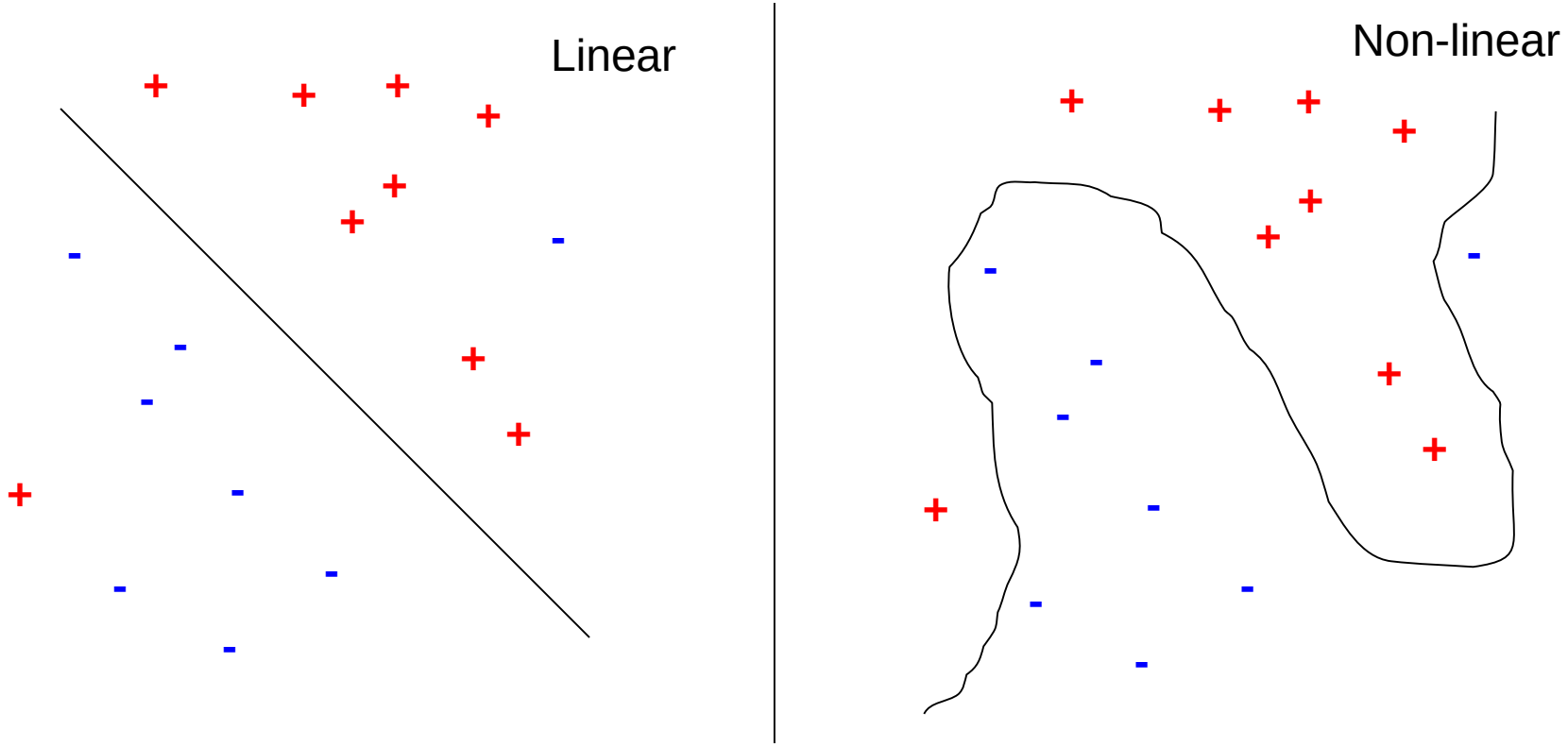
- Generalized Gaussian kernel:

$$K(h_1, h_2) = \exp\left(-\frac{1}{A} D(h_1, h_2)^2 \right)$$

- *D* can be Euclidean distance, χ^2 distance etc…

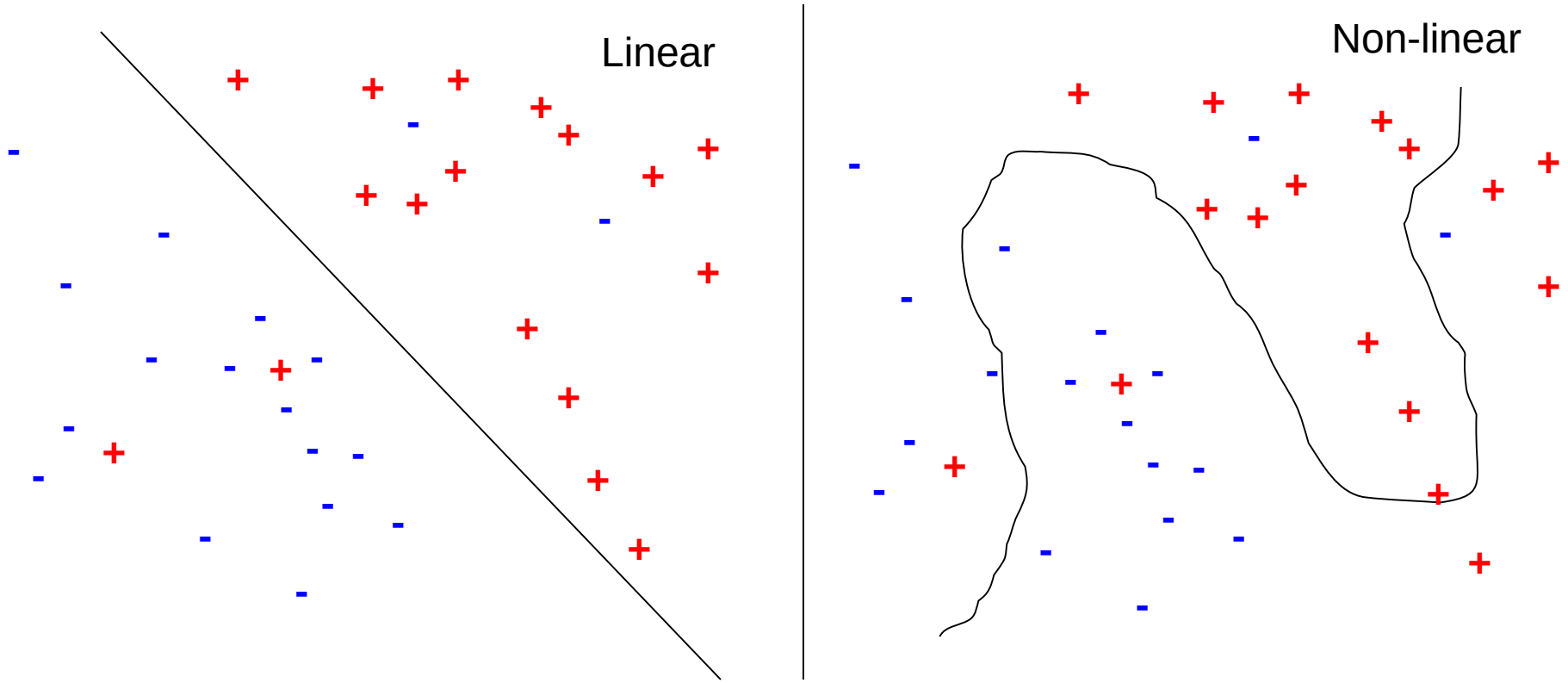
J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, Local Features and Kernels for Classifcation of Texture and Object Categories: A Comprehensive Study, IJCV 2007

Which classifier to use?



Which classifier to use?

Let’s add more training data!



A more complex model can over fit the data if these are not enough!

What about multi-class SVMs?

- No “definitive” multi-class SVM formulation
- In practice, we have to obtain a multi-class SVM by combining multiple two-class SVMs
- One vs. others
  - Traning: learn an SVM for each class vs. the others
  - Testing: apply each SVM to test example and assign to it the class of the SVM that returns the highest decision value
- One vs. one
  - Training: learn an SVM for each pair of classes
  - Testing: each learned SVM “votes” for a class to assign to the test example

Credit slide: S. Lazebnik

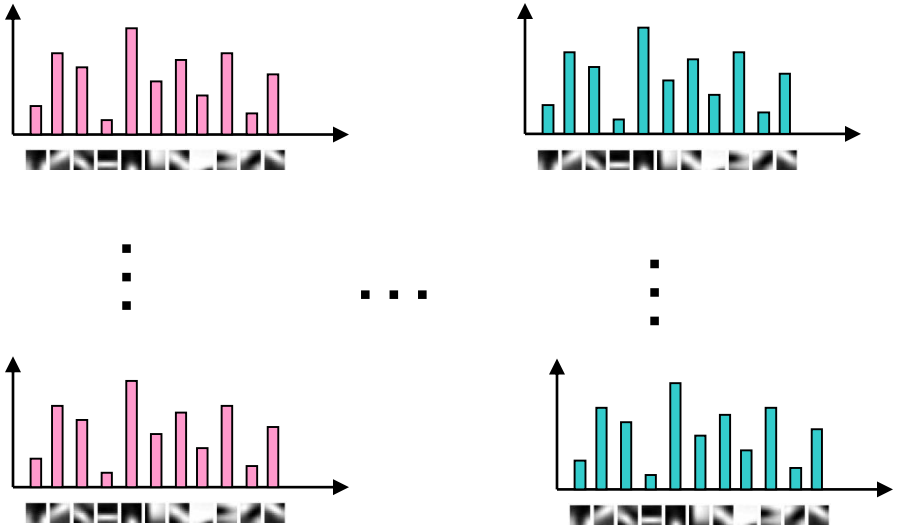
SVMs: Pros and cons

- Pros
  - Many publicly available SVM packages: http://www.kernel-machines.org/software
  - Kernel-based framework is very powerful, flexible
  - SVMs work very well in practice, even with very small training sample sizes

- Cons
  - No “direct” multi-class SVM, must combine two-class SVMs
  - Computation, memory
    - During training time, must compute matrix of kernel values for every pair of examples
    - Learning can take a very long time for large-scale problems

Discriminative classifiers (linear classifier)

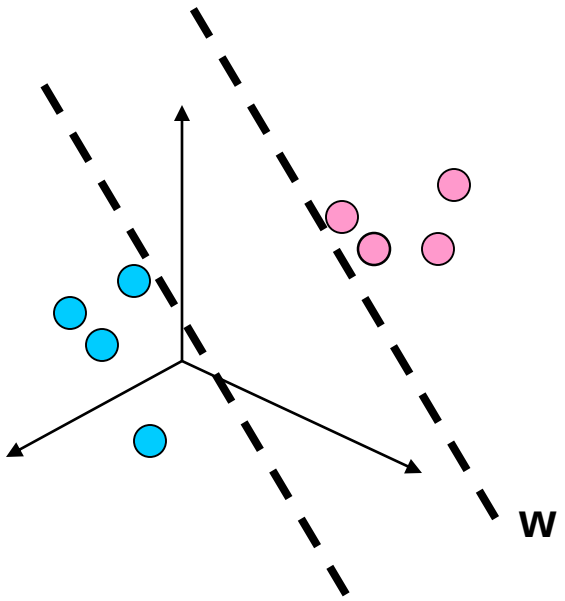
category models



Class 1

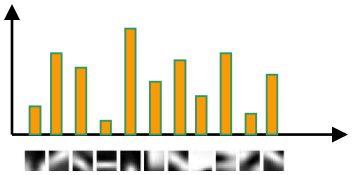
Class N

Model space



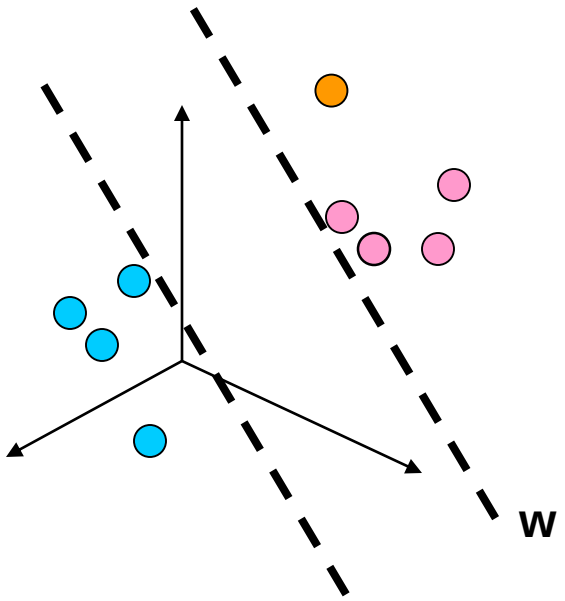
Discriminative classifiers (linear classifier)

Query image



Winning class: pink

Model space

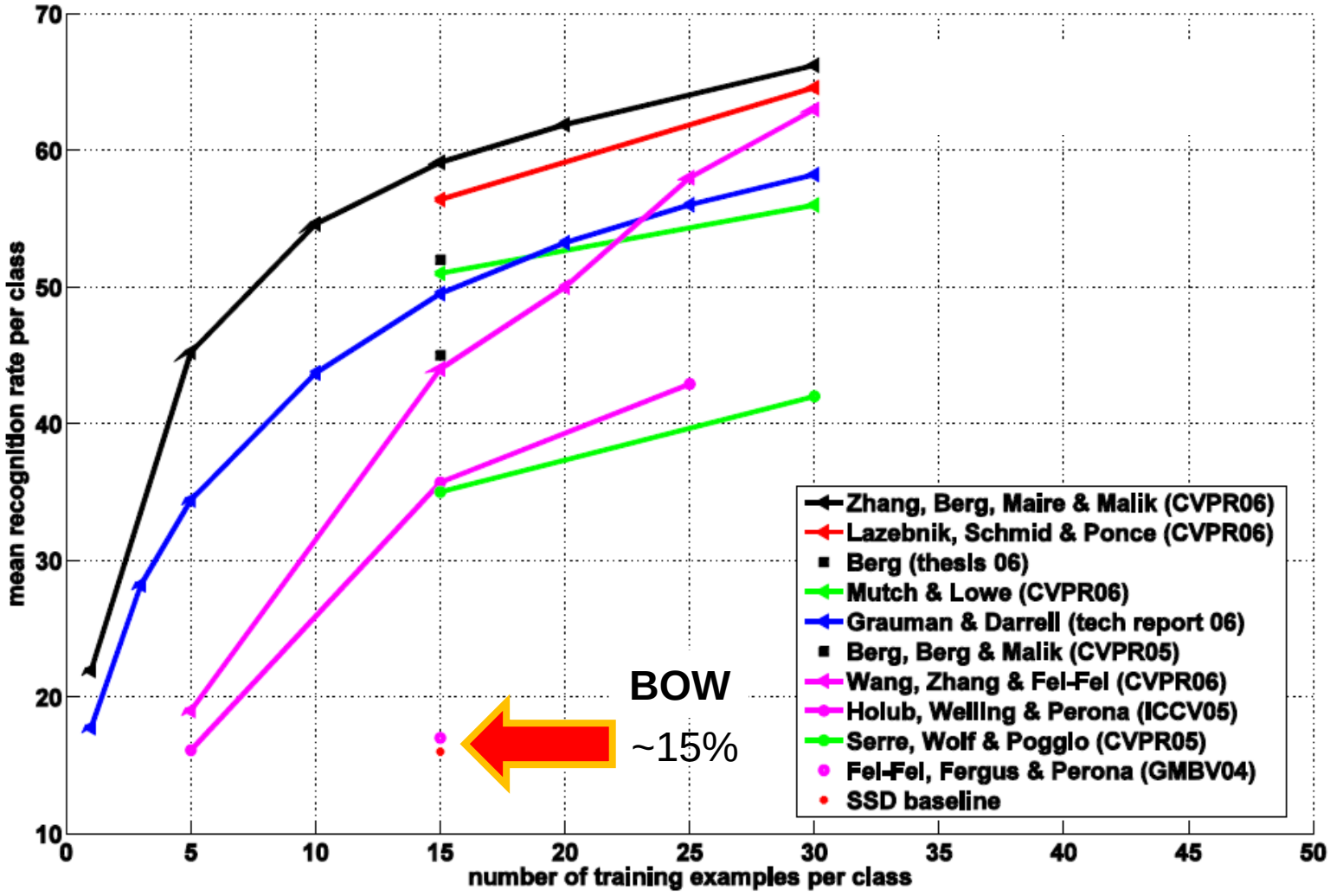


Caltech 101

Fei-Fei et al. (2004)

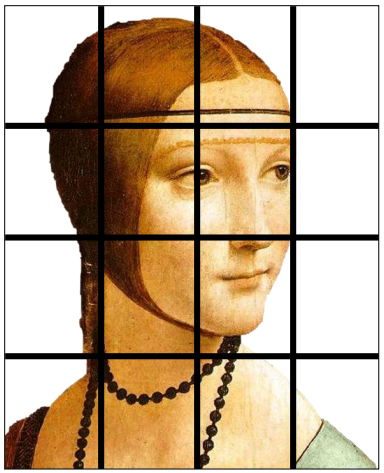
http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

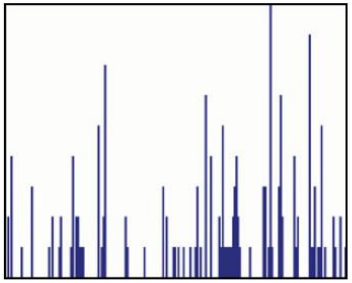
Caltech 101


mean recognition rate per class
number of training examples per class
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
40
45
50
Zhang, Berg, Maire & Malik (CVPR06)
Lazebnik, Schmid & Ponce (CVPR06)
Berg (thesis 06)
Mutch & Lowe (CVPR06)
Grauman & Darrell (tech report 06)
Berg, Berg & Malik (CVPR05)
Wang, Zhang & Fei-Fei (CVPR06)
Holub, Welling & Perona (ICCV05)
Serre, Wolf & Poggio (CVPR05)
Fei-Fei, Fergus & Perona (GMBV04)
SSD baseline
BOW
~15%

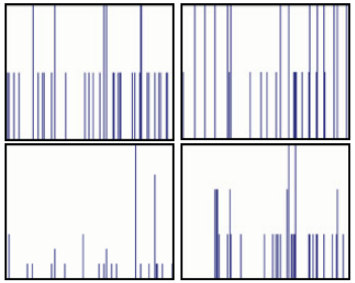
Spatial Pyramid Matching

Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. S. Lazebnik, C. Schmid, and J. Ponce.. 2006

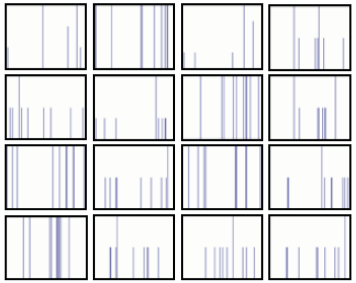




level 0



level 1



level 2

$$SPM(x_i, x) = \frac{1}{2^L} HIK_0(x_i, x) + \ldots + \frac{1}{L-l+1} HIK_l(x_i, x) + \ldots + HIK_L(x_i, x)$$

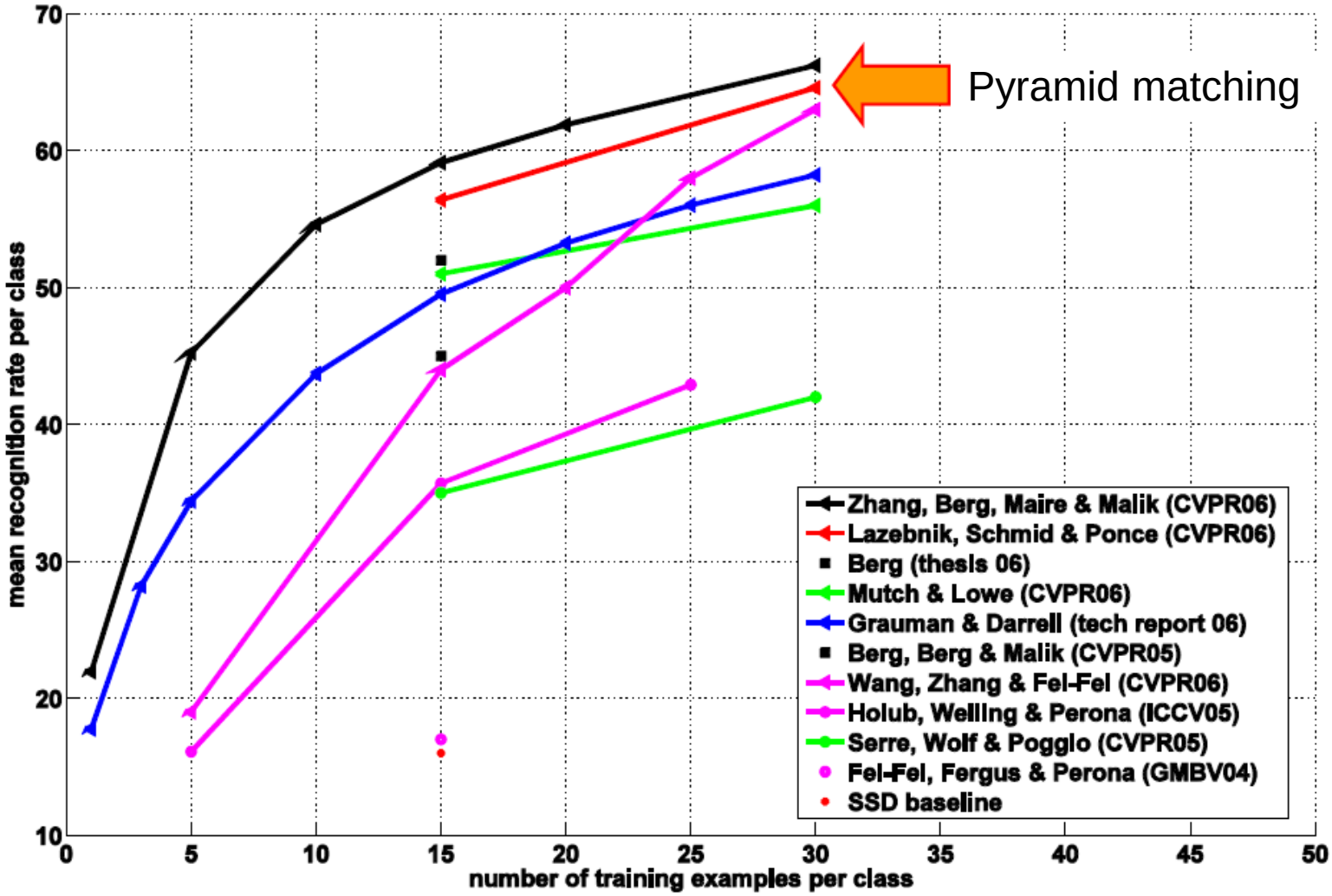
$$I(h_1, h_2) = \sum_{i=1}^{N} \min(h_1(i), h_2(i))$$

Caltech 101

Multi-class classification results (30 training images per class)

	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

Caltech 101


Pyramid matching
mean recognition rate per class
number of training examples per class
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
40
45
50
Zhang, Berg, Maire & Malik (CVPR06)
Lazebnik, Schmid & Ponce (CVPR06)
Berg (thesis 06)
Mutch & Lowe (CVPR06)
Grauman & Darrell (tech report 06)
Berg, Berg & Malik (CVPR05)
Wang, Zhang & Fei-Fei (CVPR06)
Holub, Welling & Perona (ICCV05)
Serre, Wolf & Poggio (CVPR05)
Fei-Fei, Fergus & Perona (GMBV04)
SSD baseline

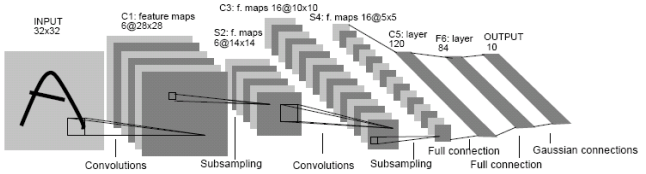
Discriminative models

Nearest neighbor

10^6 examples

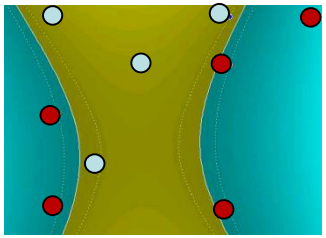
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

Neural networks



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

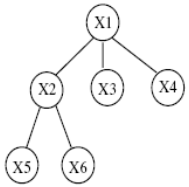
Support Vector Machines



Guyon, Vapnik, Heisele, Serre, Poggio…

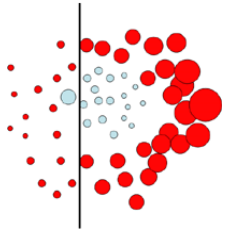
Latent SVM

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

Next Lecture

Bag of words models for object recognition and classification

- Generative methods