CEE598 - Visual Sensing for Civil Infrastructure Eng. & Mgmt.

Session 19 – Object Recognition II

Mani Golparvar-Fard
Department of Civil and Environmental Engineering
3129D, Newmark Civil Engineering Lab
e-mail: mgolpar@illinois.edu

Fun Time

Professor Hard working students Me?

Outline

- Object Recognition

- Introduction
- Recognition of single 3D objects
  - Bag of word models
  - Part based models
  - Models for 3D objects categorization

Challenges: object intra-class variation

Usual Challenges

- Variability due to:
  - View point
  - Illumination
  - Occlusions

Object categorization: the statistical viewpoint

$$p(excavator \mid image)$$

vs.

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

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

- Bayes rule:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Object categorization: the statistical viewpoint

$$p(excavator \mid image)$$

vs.

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

- Bayes rule:

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

Object categorization: the statistical viewpoint

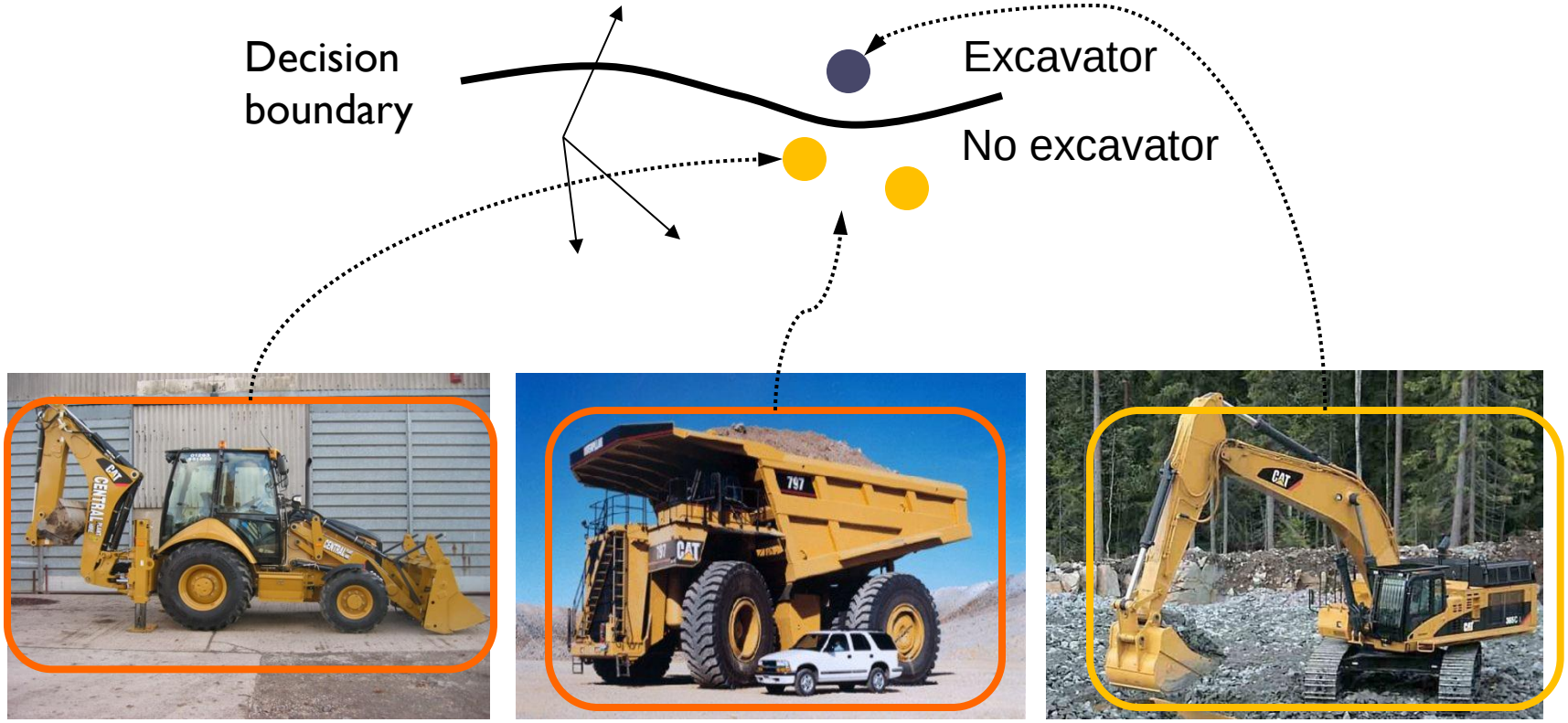
- **Discriminative methods** model posterior
- **Generative methods** model likelihood and prior

- Bayes rule:

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

Discriminative

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



Generative

$p(image \mid excavator)$

$p(image \mid no\ excavator)$

$p(image \mid excavator)$	$p(image \mid no\ excavator)$
Low	high
High	Low

Learning

- machine learning useful to model intraclass variability



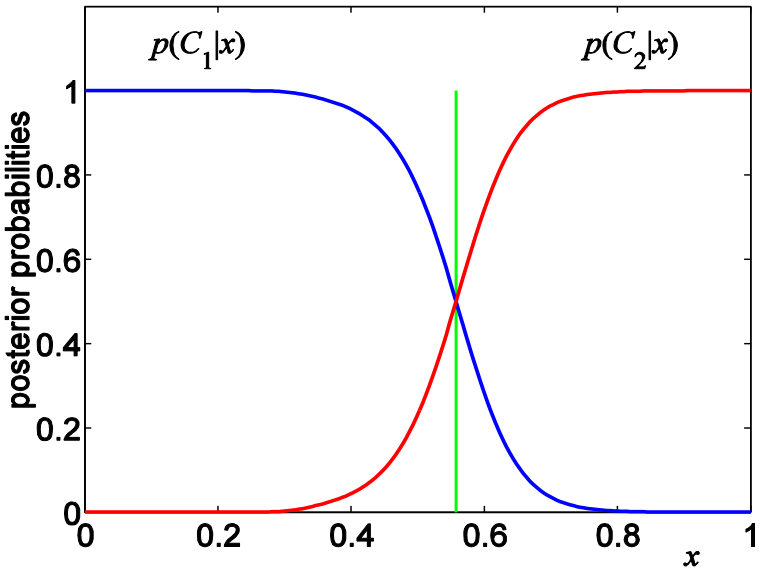

Learning

- Learning parameters: What are you maximizing?

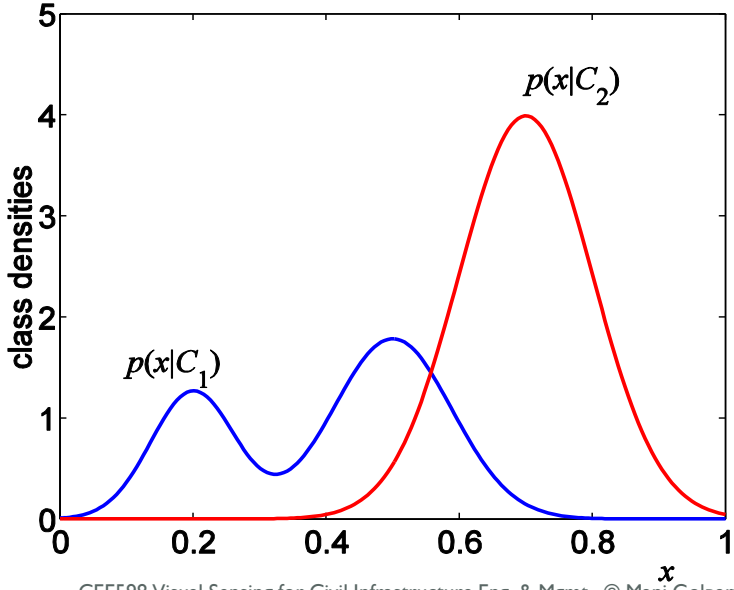
Likelihood (Gen.) or performances on train/validation set (Disc.)

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

discriminative



generative



Learning

- Learning parameters: What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

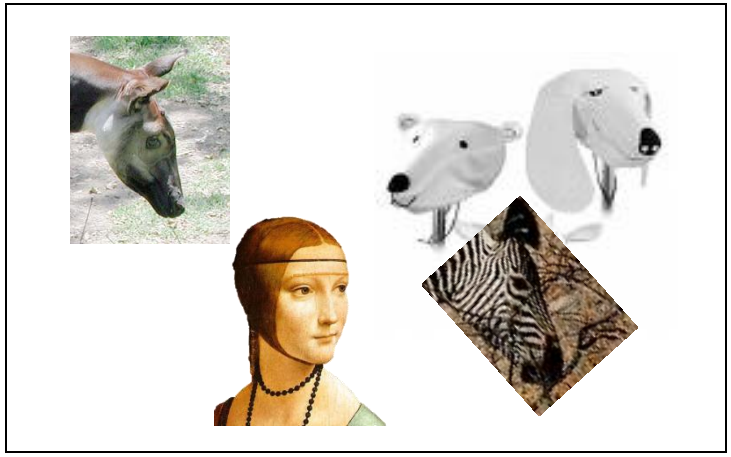
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

- Batch/incremental (on category and image level; user-feedback)

- Priors
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods

Contains a excavator

Choose Negative Images for Discriminative models



Bag of Words

Part of this segment is based on the tutorial *"Recognizing and Learning Object Categories: Year 2007"*

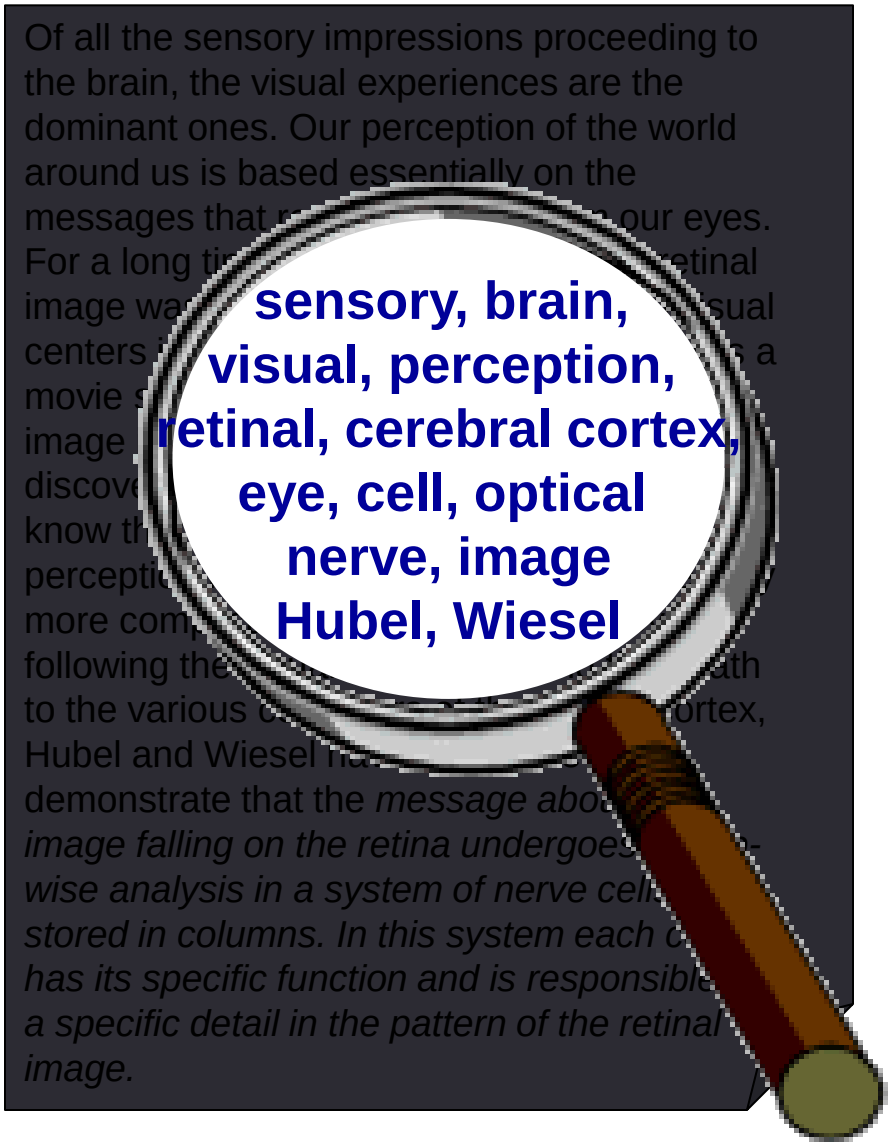
by Prof A. Torralba, R. Fergus and F. Li

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'

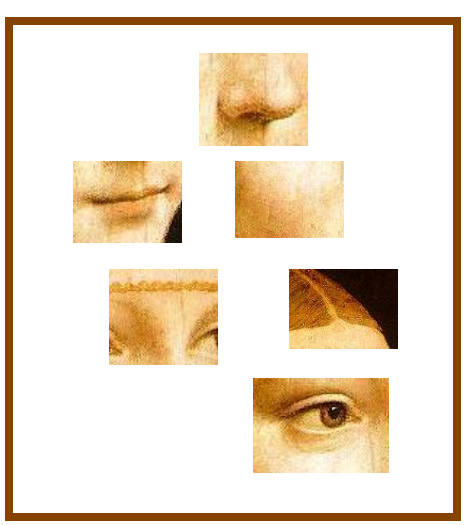
Analogy to documents



Definition of "BoW"

- Independent features

face

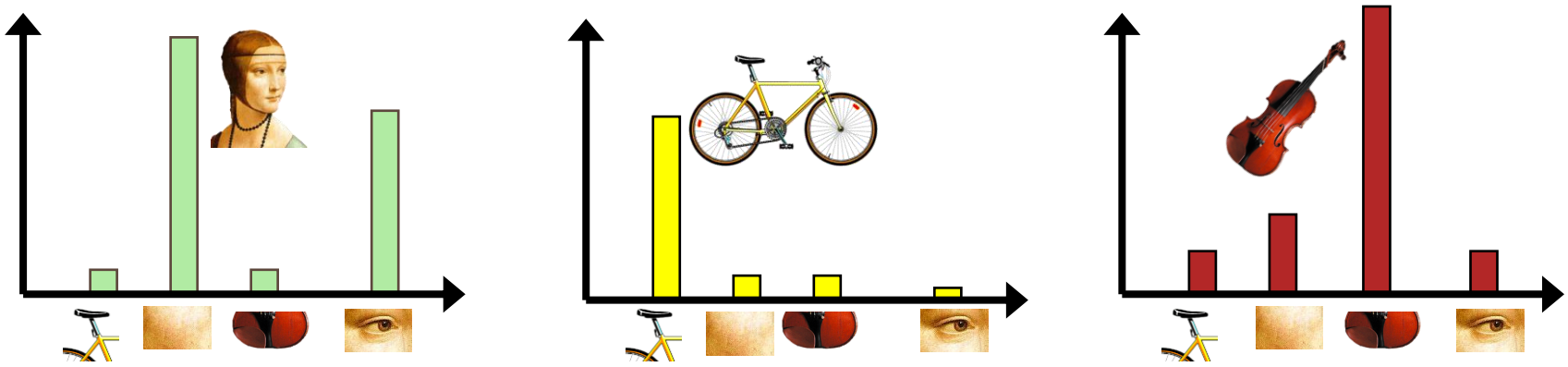


bike

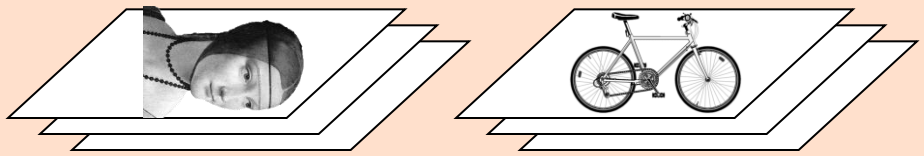
violin

definition of "BoW"

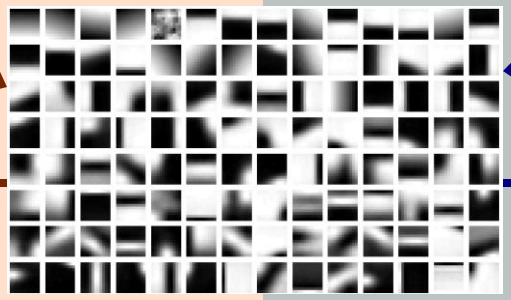
- Independent features
- histogram representation



Representation Learning and Recognition



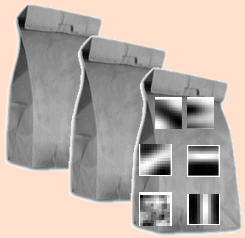
2. codewords dictionary

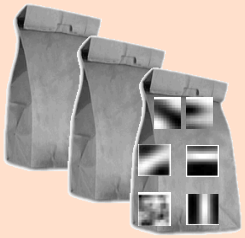


1. feature detection & representation

image representation

3.

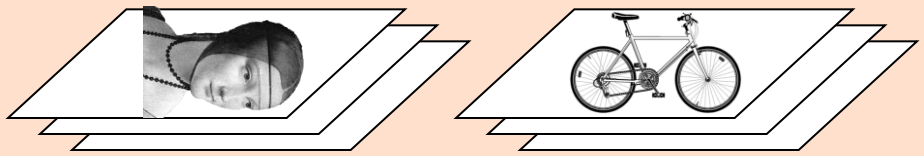




category models

category decision

Representation



1. feature detection & representation

2. codewords dictionary

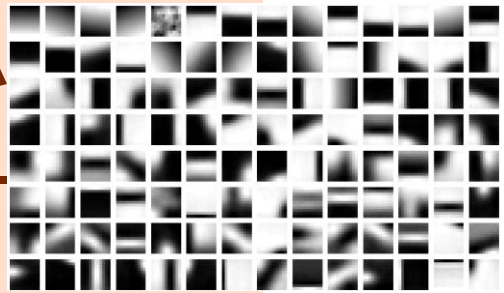
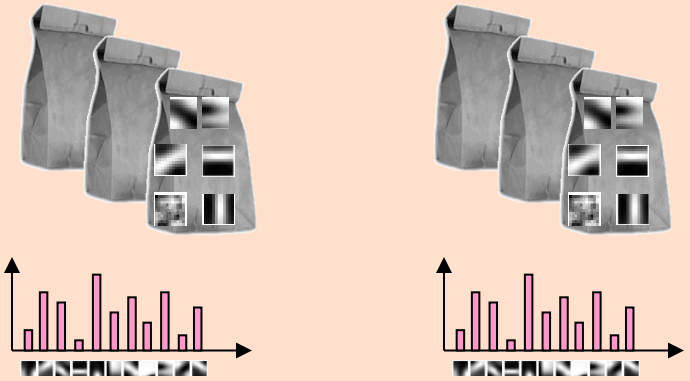


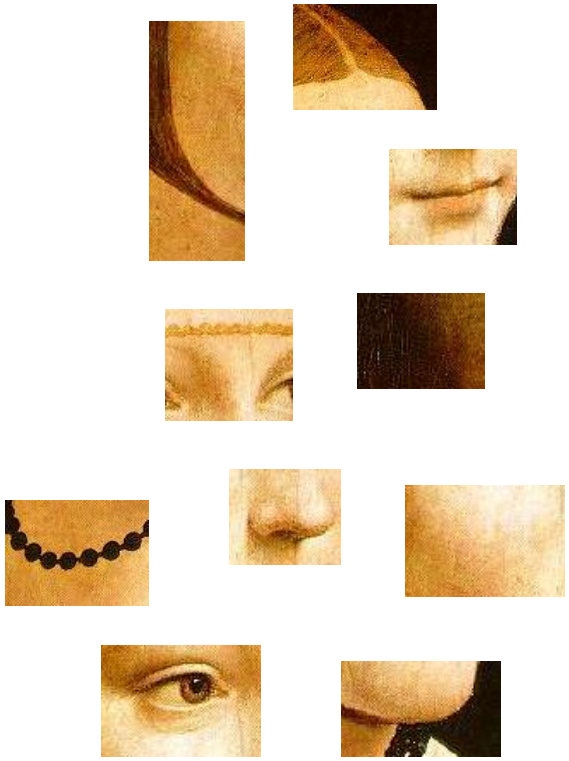
image representation

3.



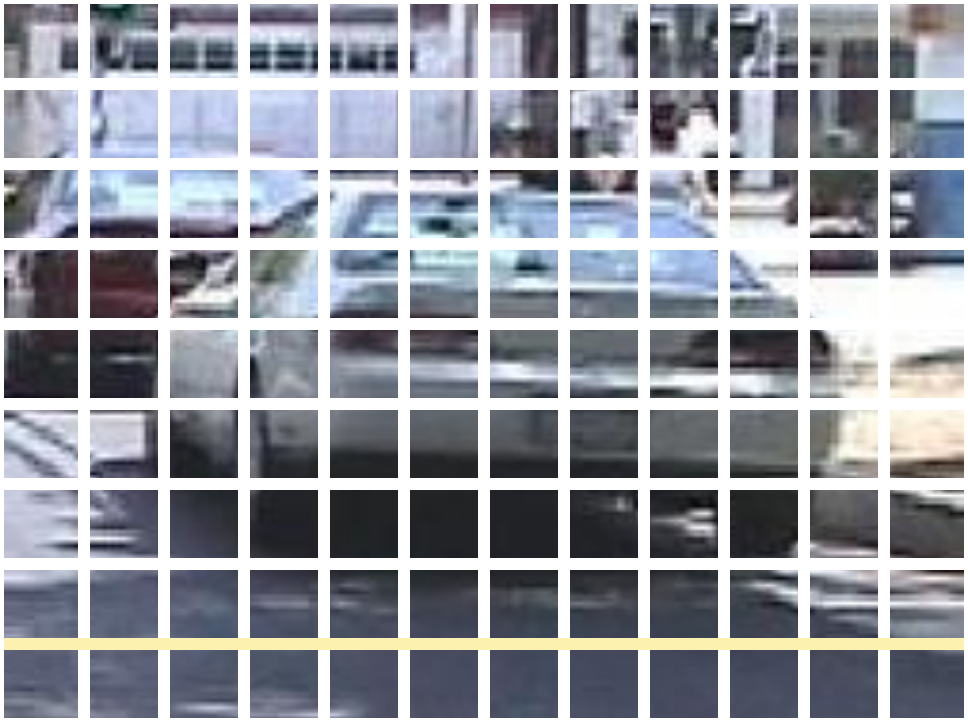
category models

1. Feature detection and representation



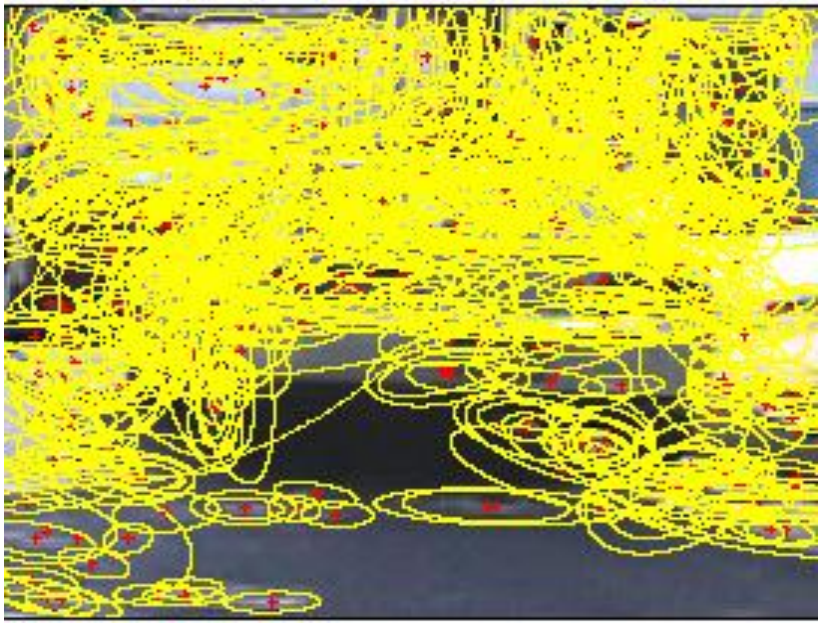
1. Feature detection and representation

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



1. Feature detection and representation

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



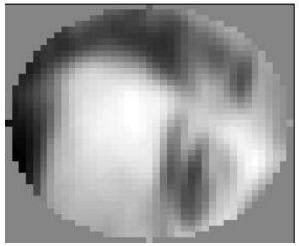
1. Feature detection and representation

- 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)

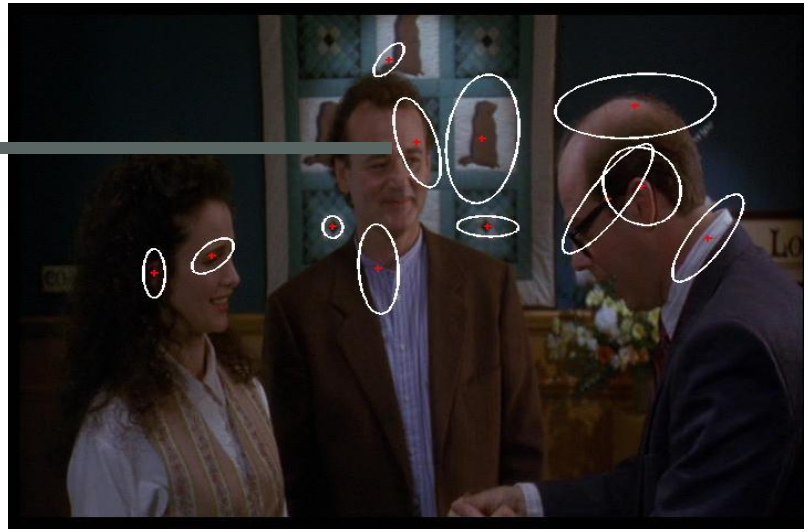
1. Feature detection and representation

Compute SIFT descriptor

[Lowe'99]



Normalize patch



Detect patches

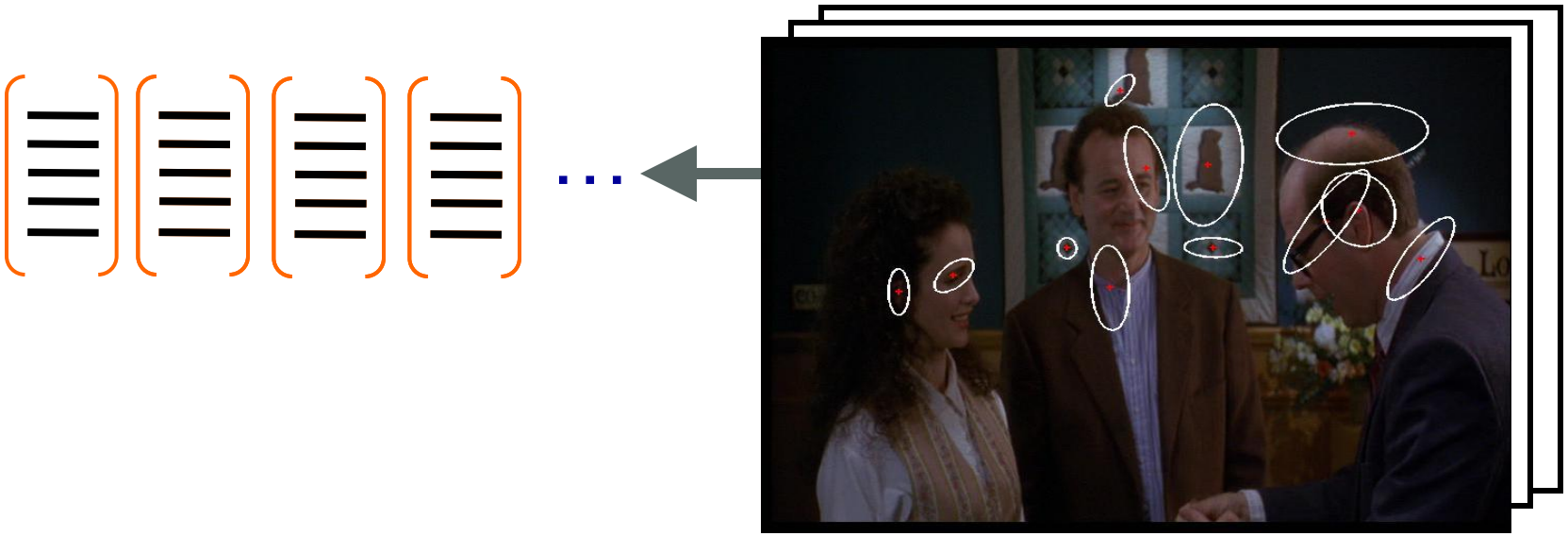
[Mikojaczyk and Schmid '02]

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

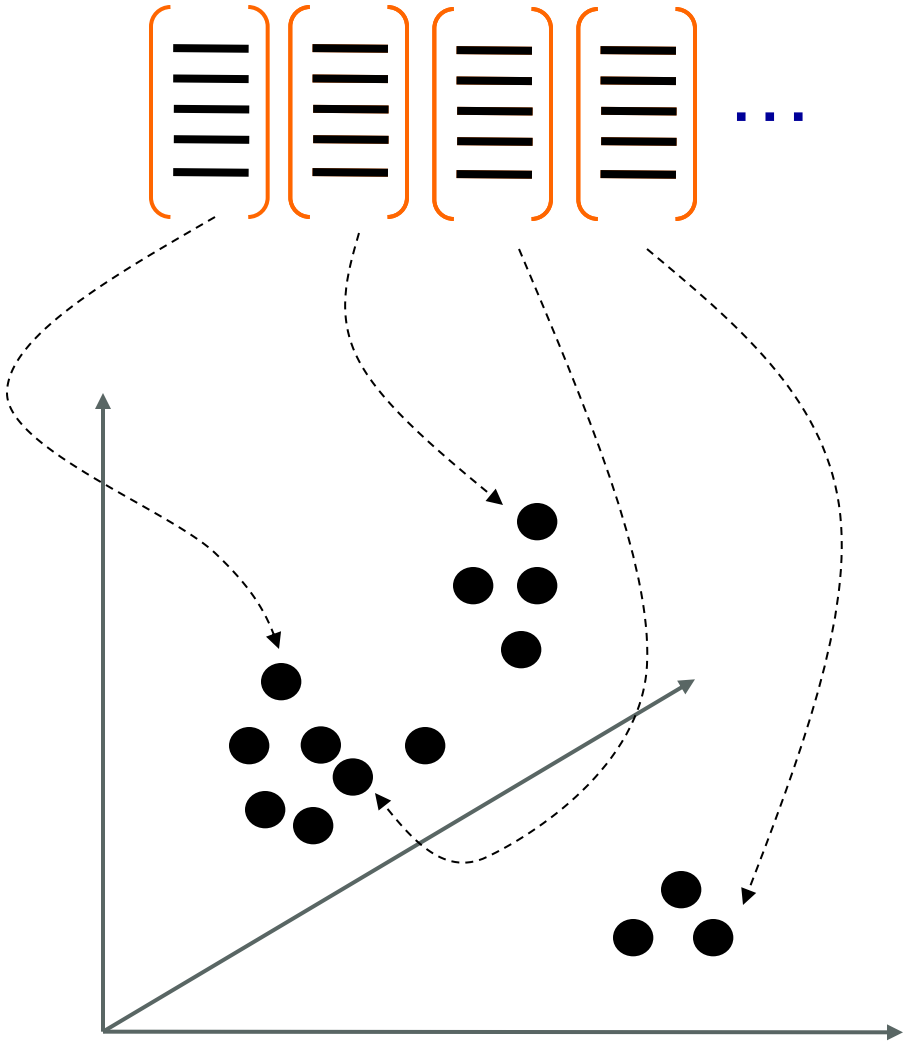
[Sivic & Zisserman, '03]

Slide credit: Josef Sivic

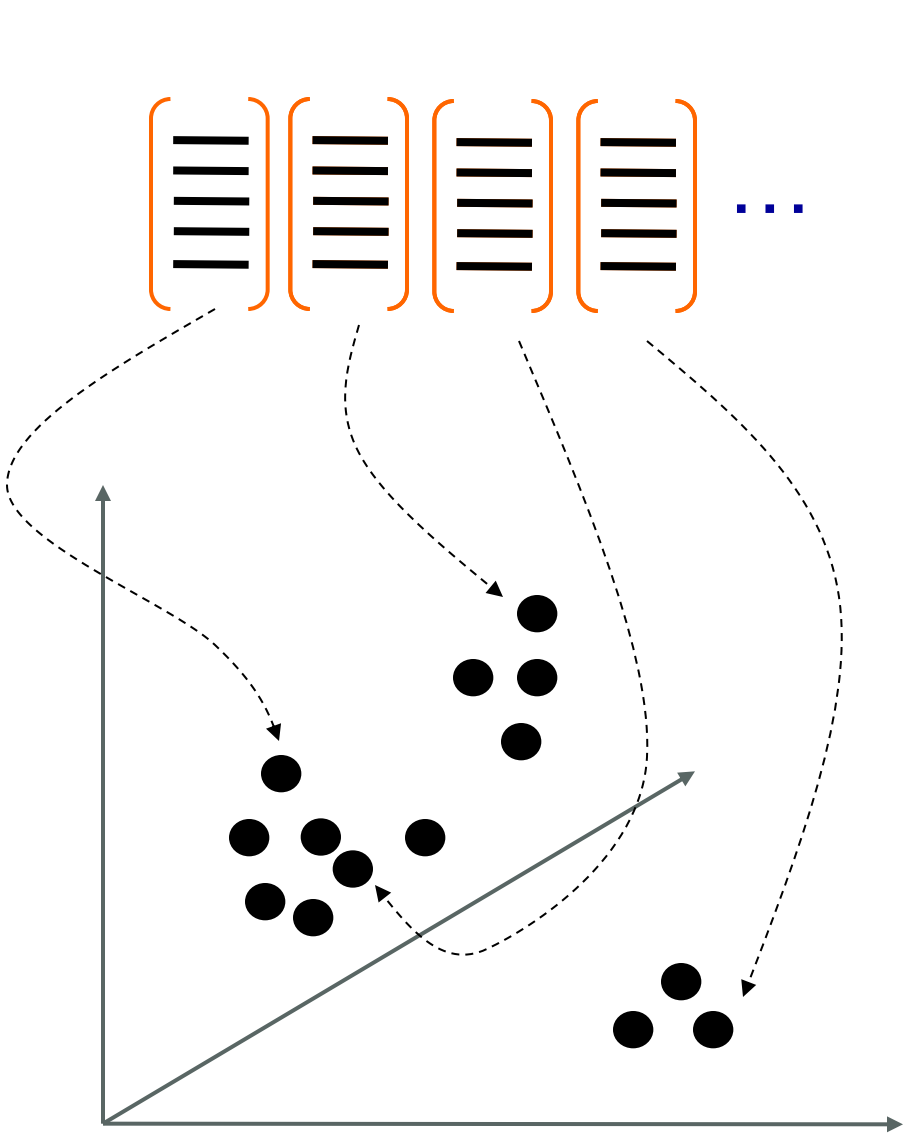
1. Feature detection and representation

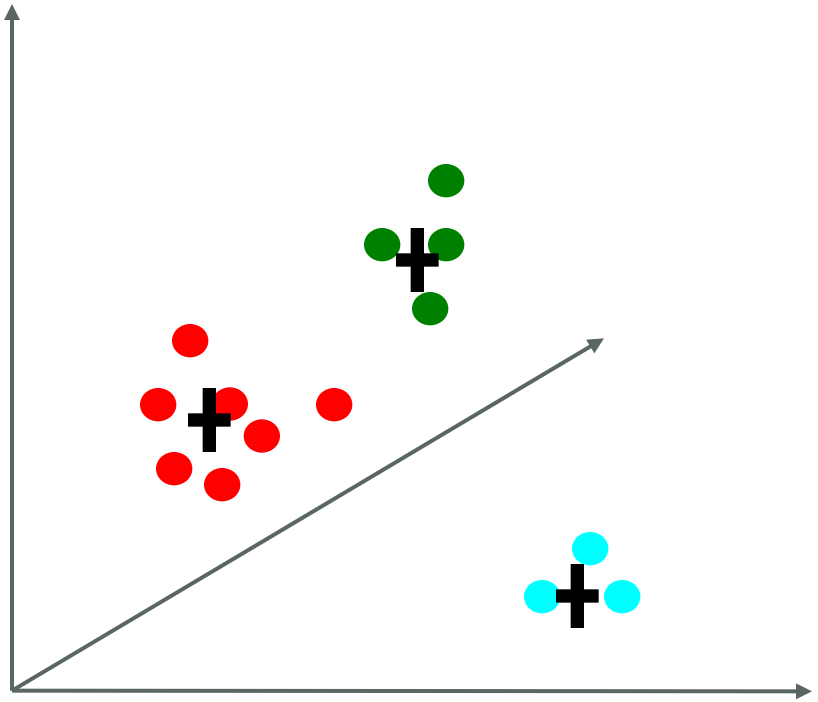


2. Codewords dictionary formation



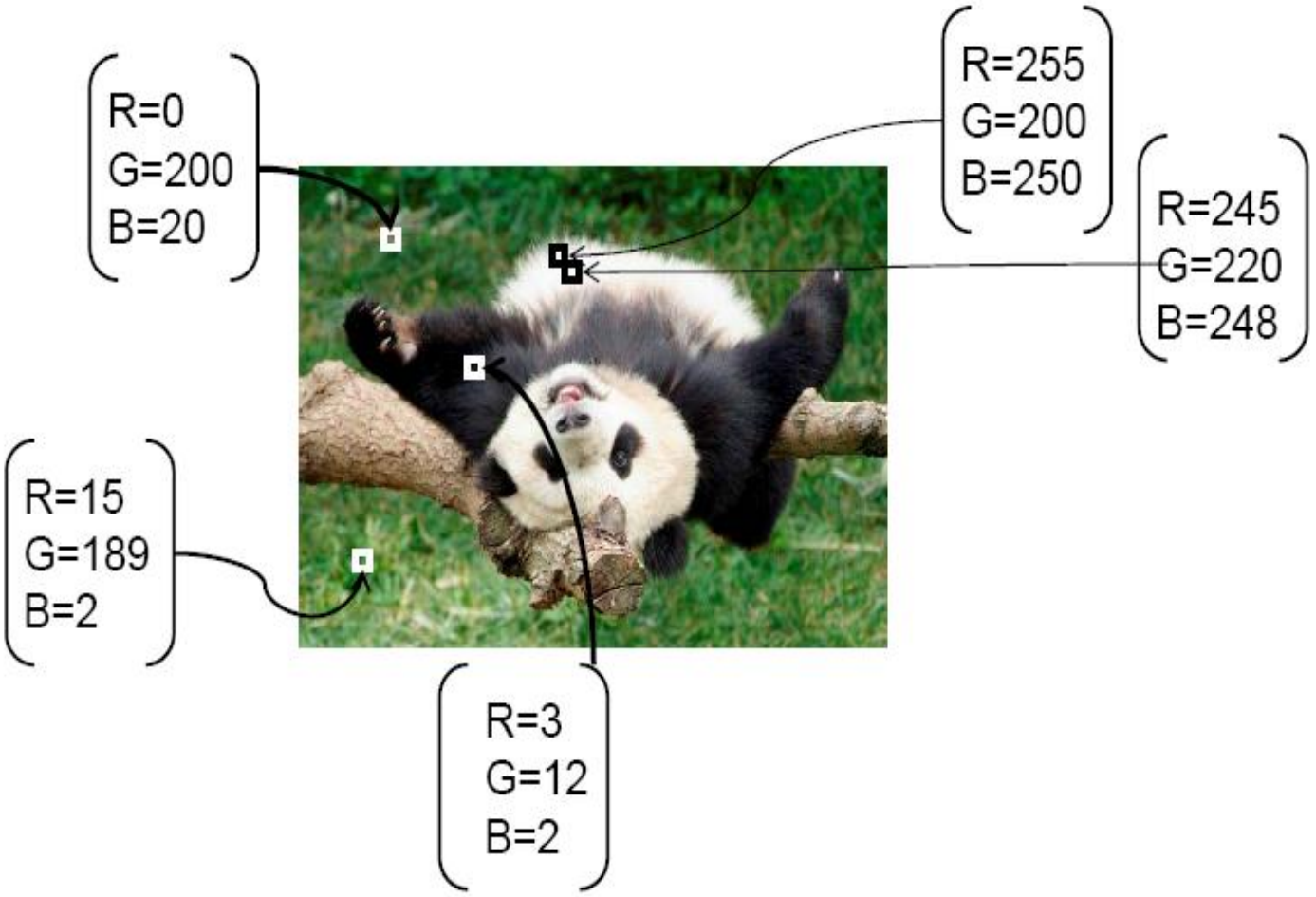
2. Codewords dictionary formation





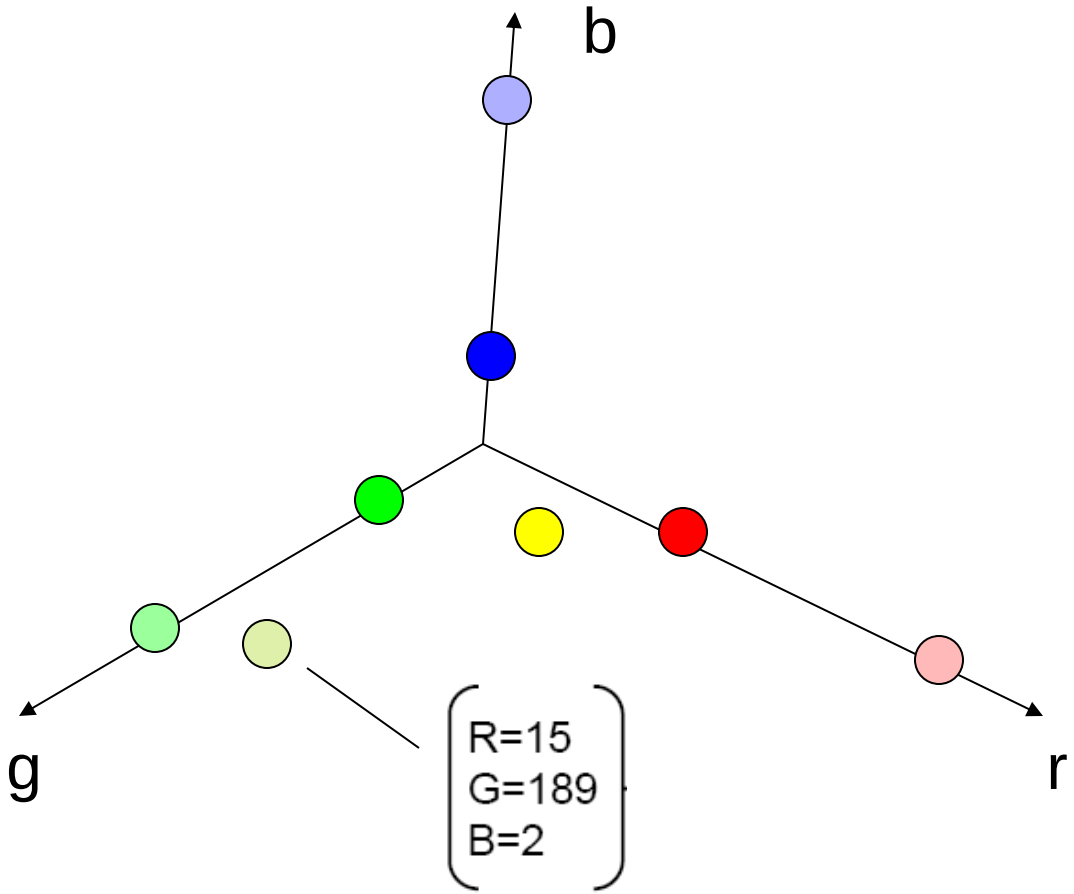
Clustering/
vector quantization

Example: color feature

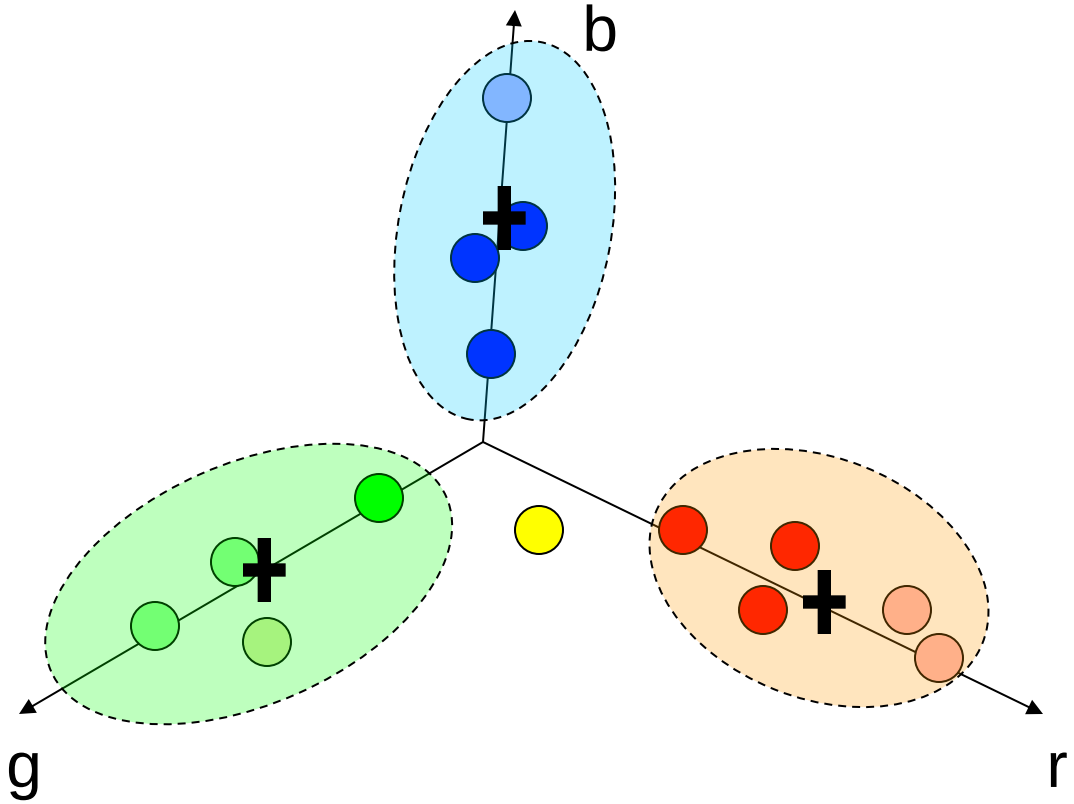


Source: K. Grauman

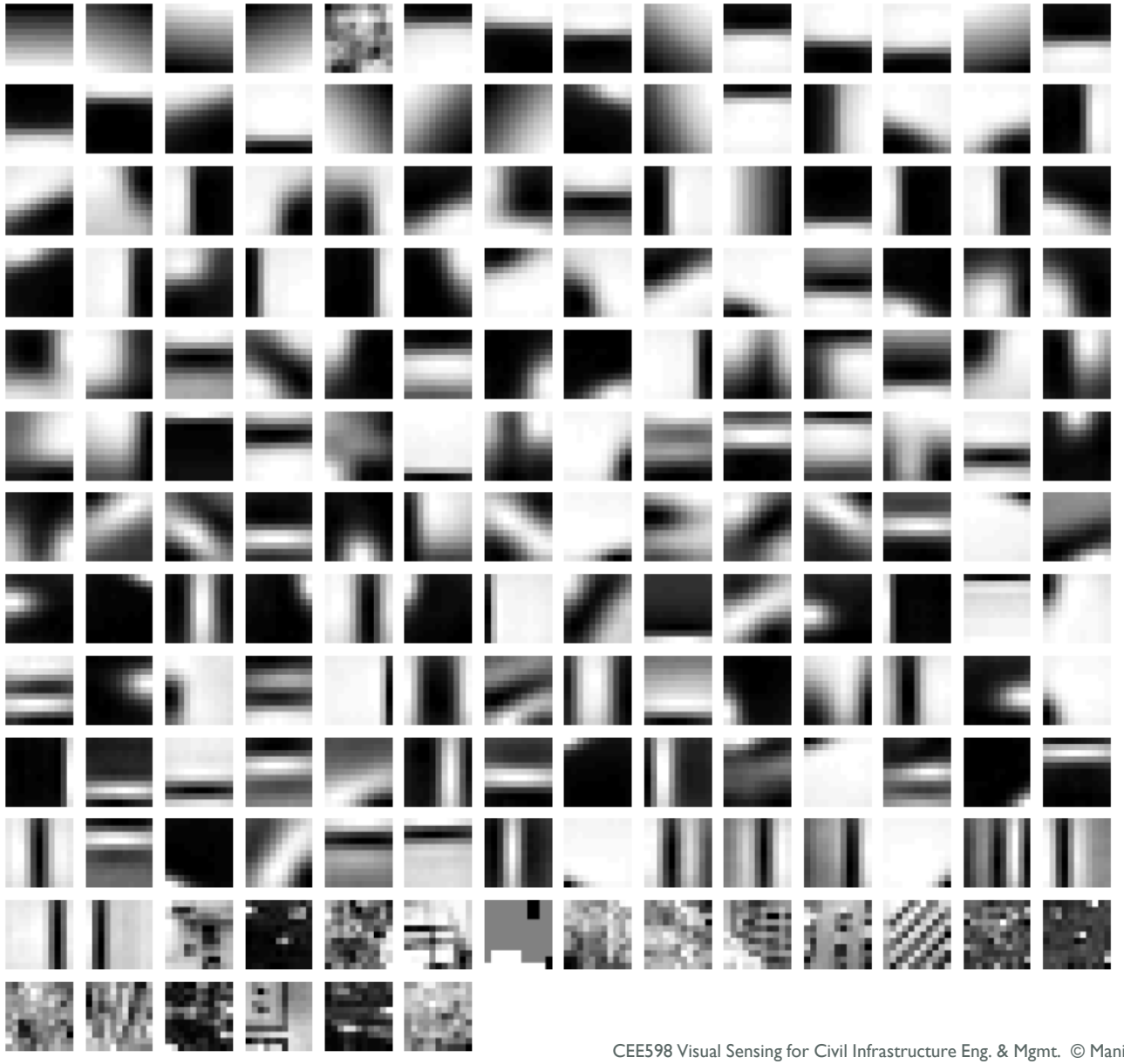
Example: color feature



Example: color feature

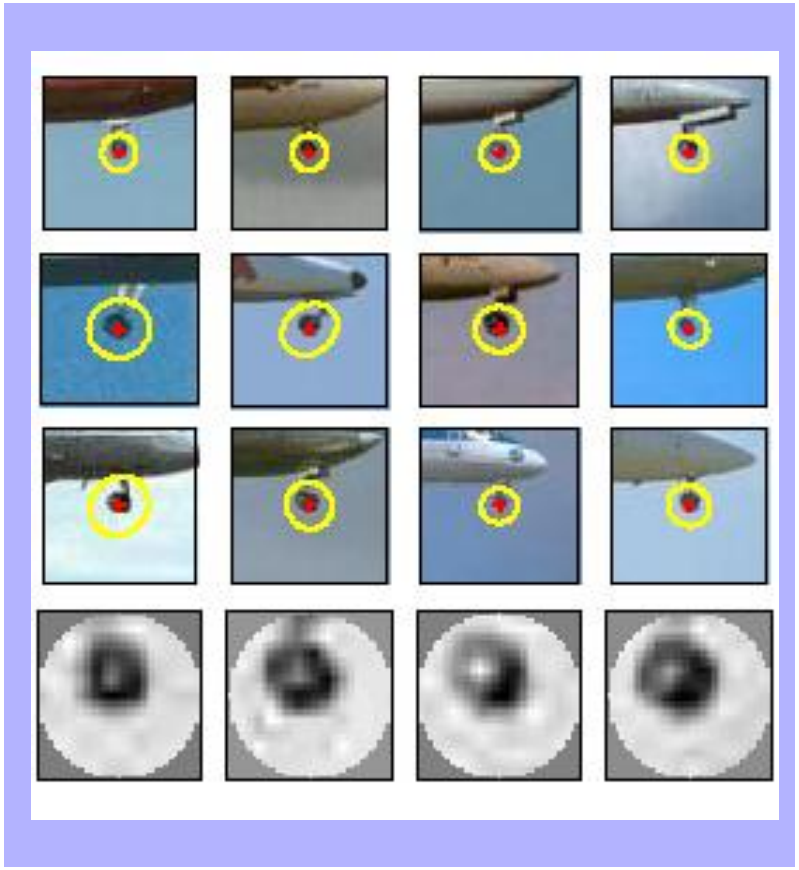


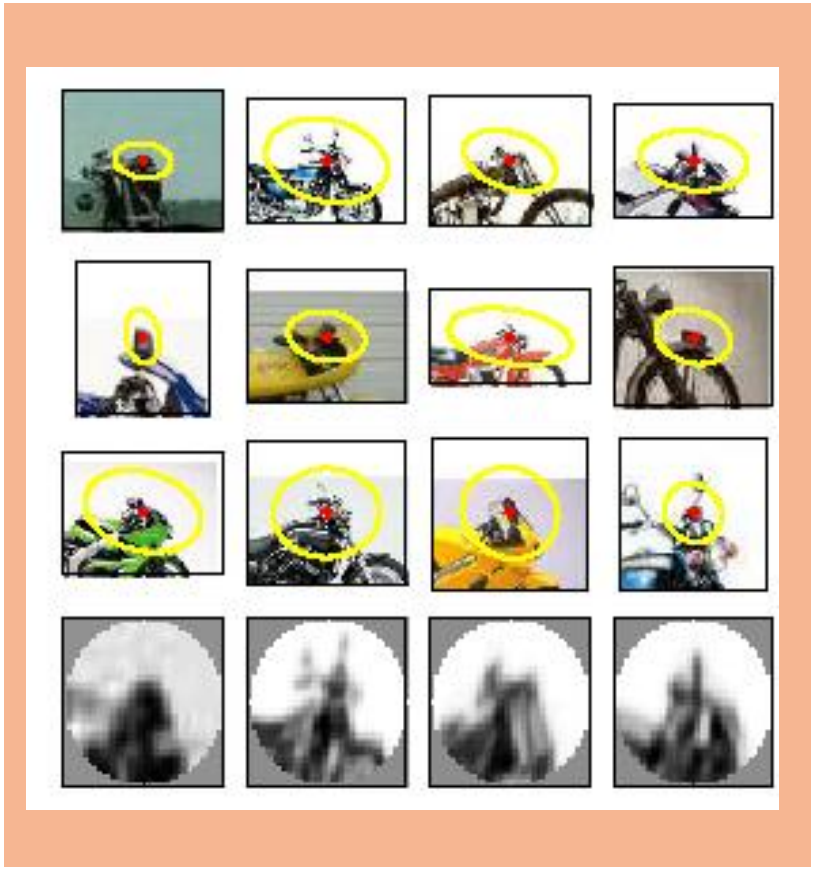
2. Codewords dictionary formation

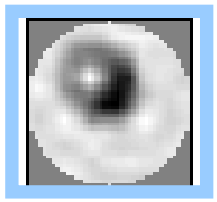


Fei-Fei et al. 2005

Image patch examples of codewords





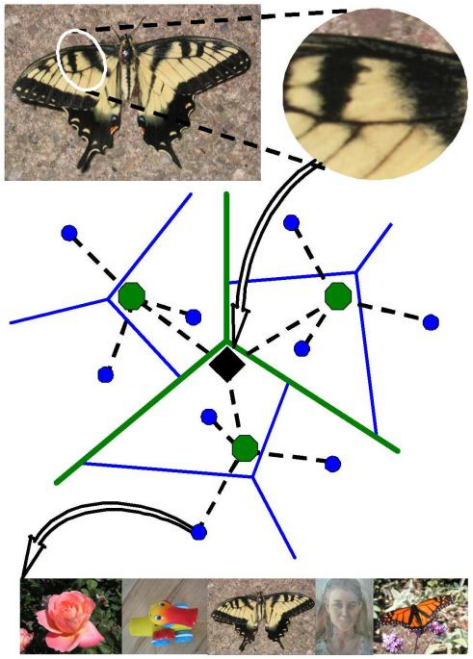


Sivic et al. 2005

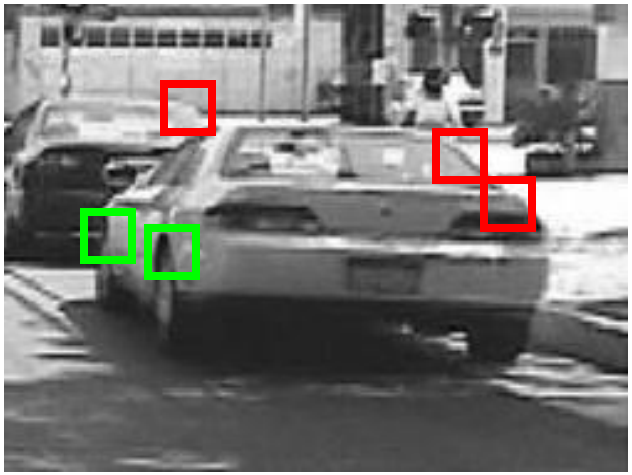
Visual vocabularies: Issues

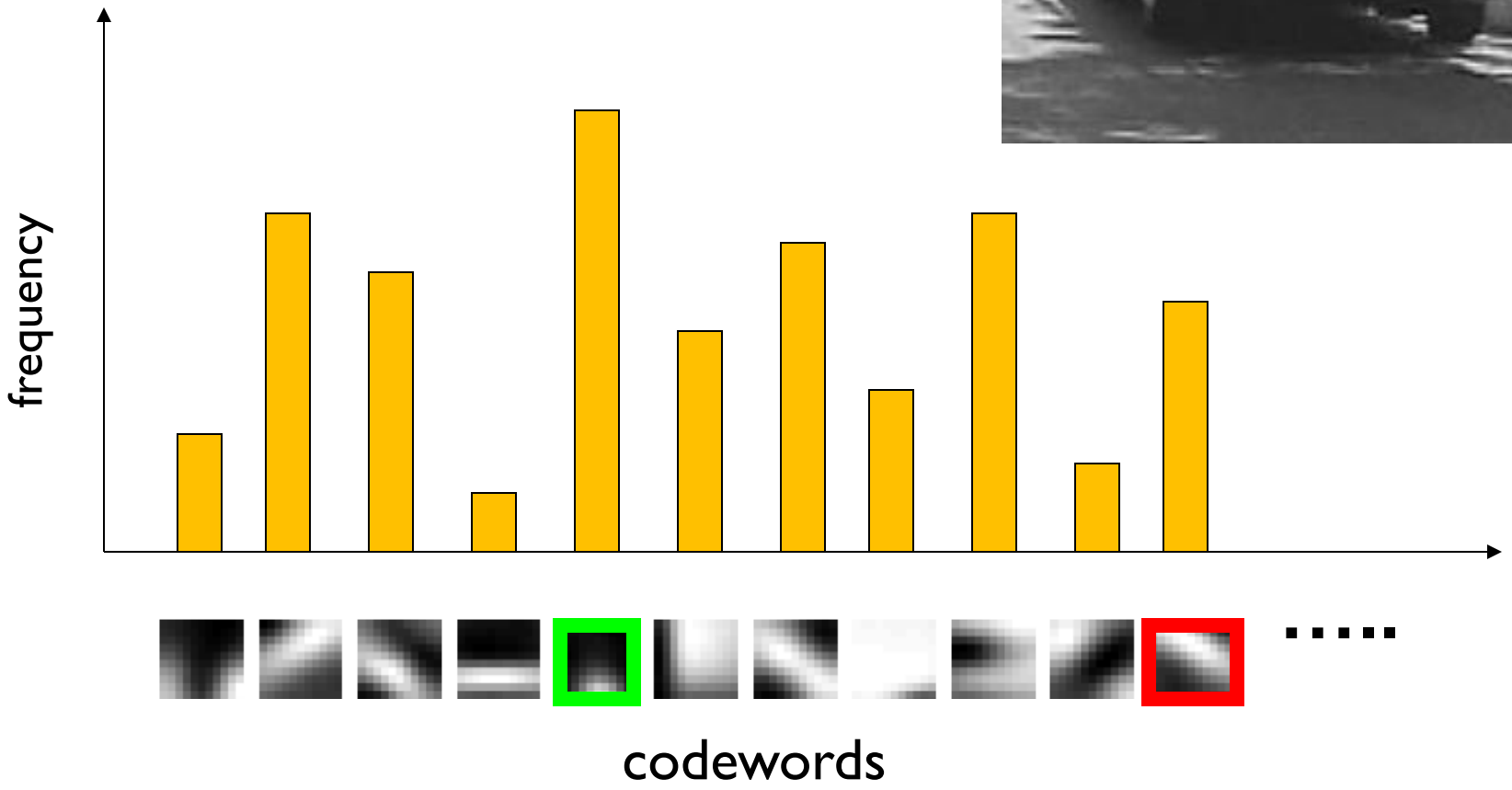
- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large: quantization artifacts, overfitting

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



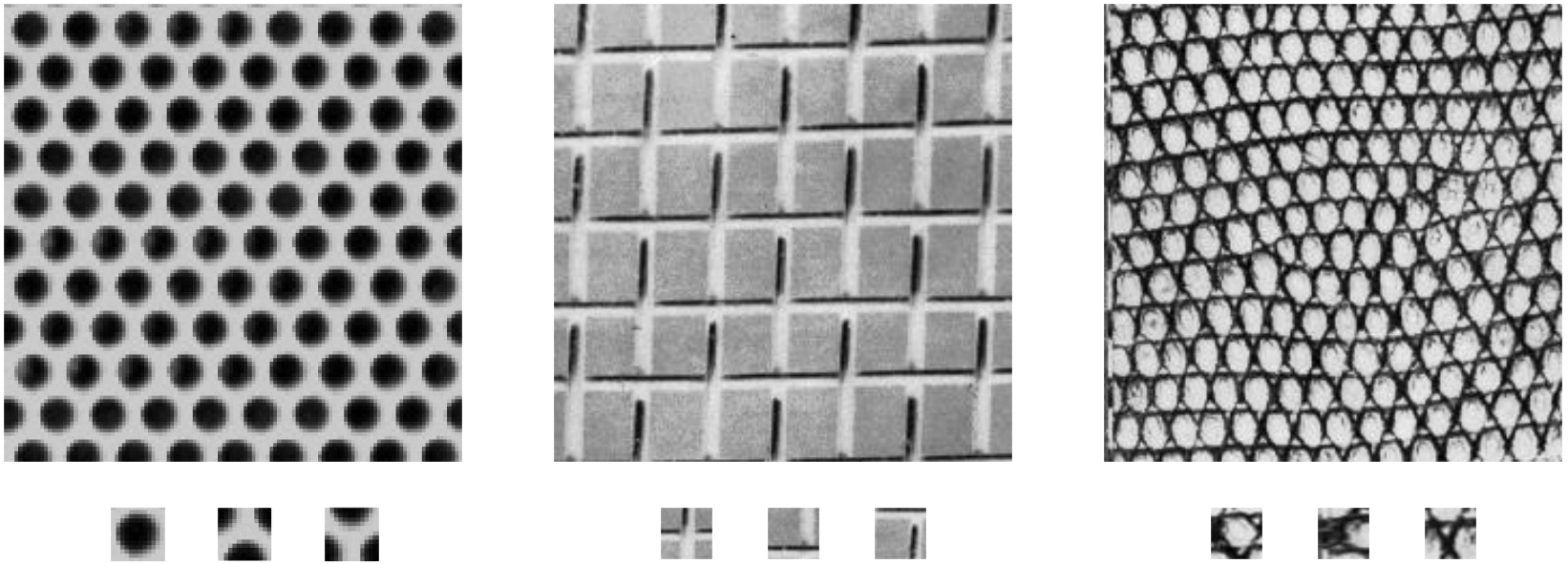
3. Image representation





Representing textures

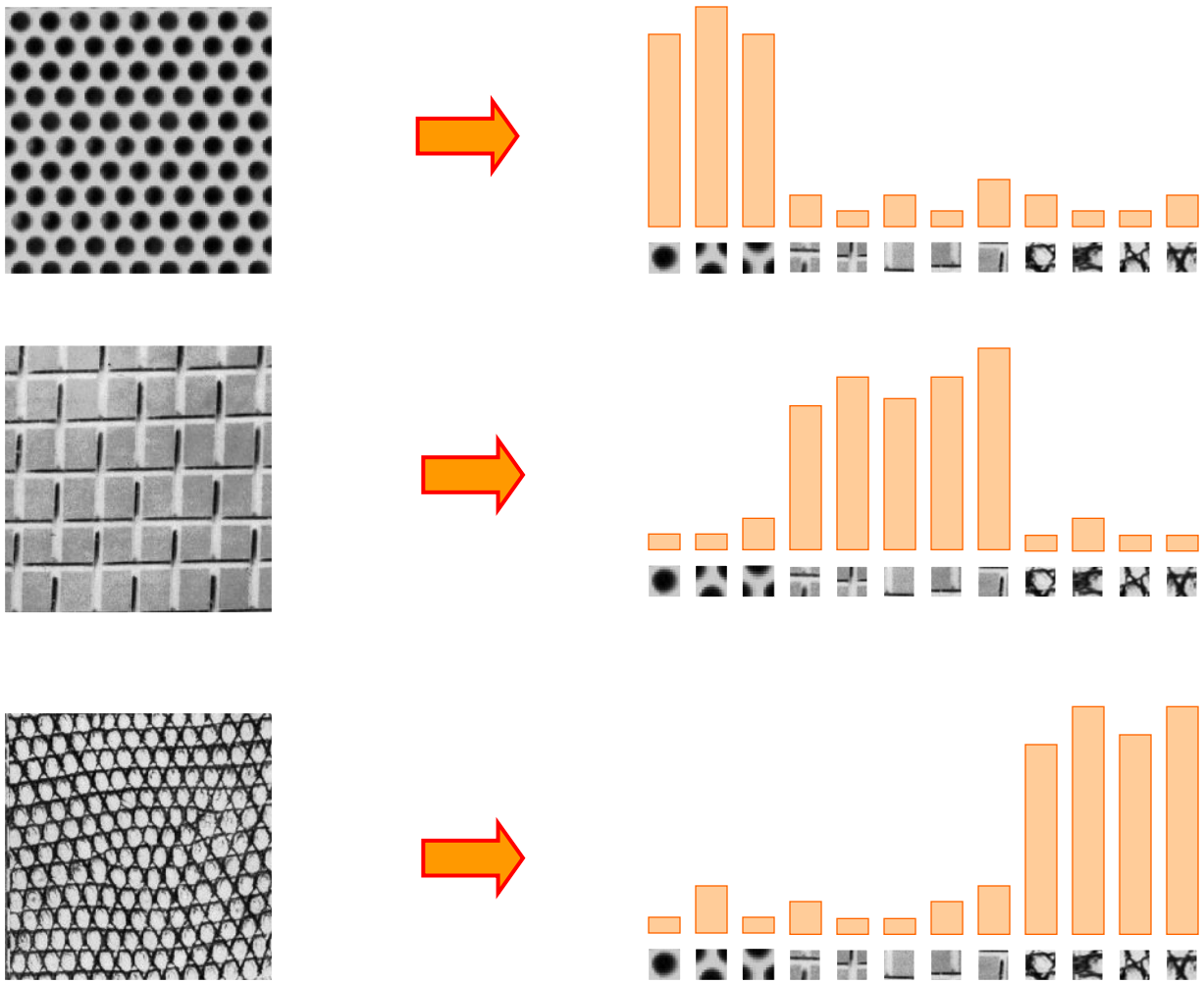
- Texture is characterized by the repetition of basic elements or *textons*
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



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

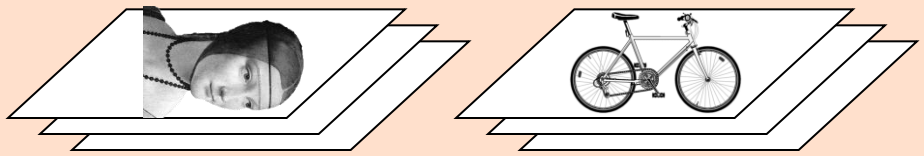
Representing textures



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

Representation



1. feature detection & representation

2. codewords dictionary

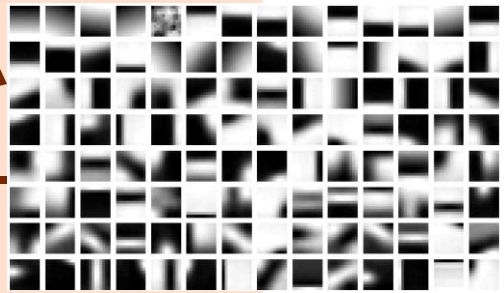
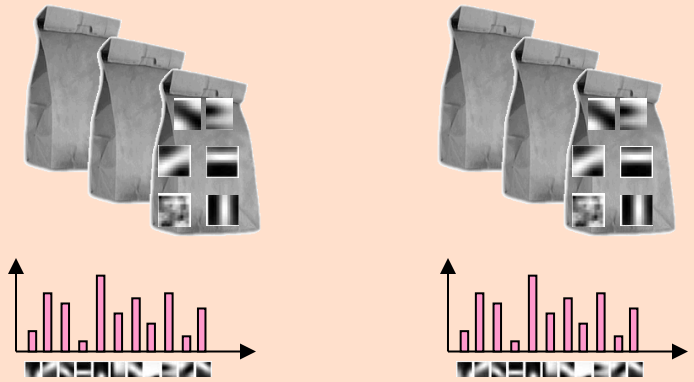


image representation

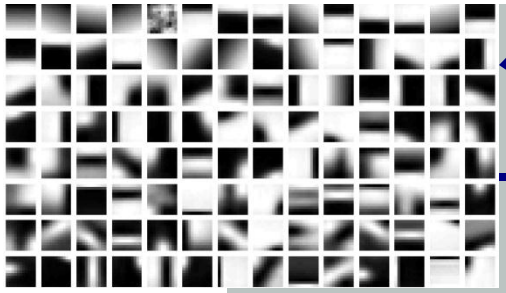
3.

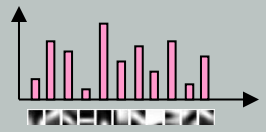


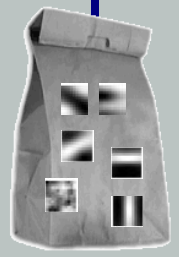
category models

Learning and Recognition

codewords dictionary





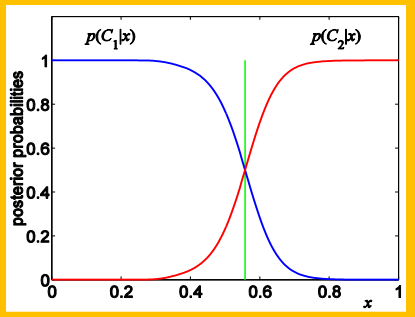


category models (and/or) classifiers

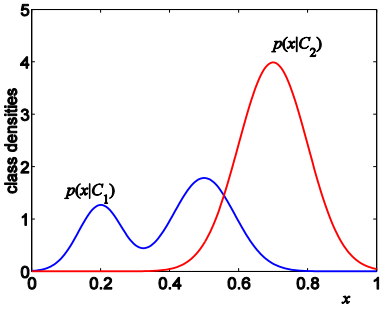
category decision

Learning and Recognition

1. Discriminative method:
   - NN (Nearest Neighbor)
   - SVM (Support Vector Machine)



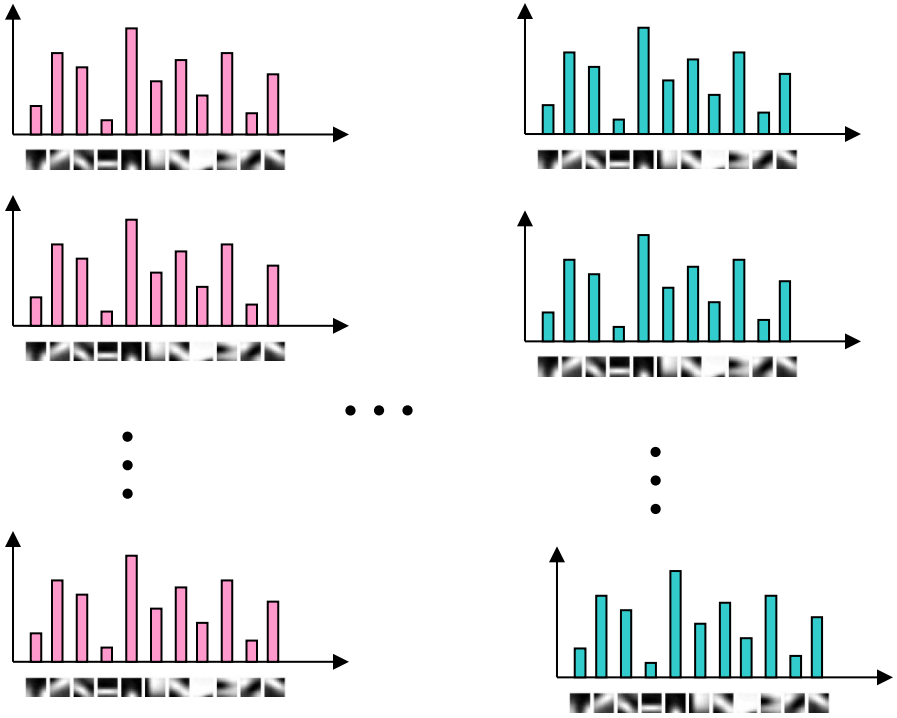
2. Generative method:
   - Graphical models



category models (and/or) classifiers

Discriminative classifiers

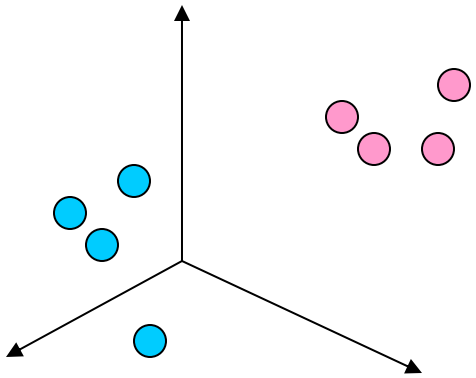
category models



Class 1

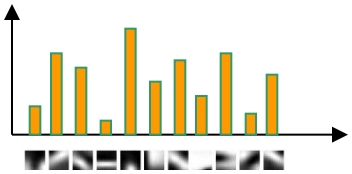
Class N

Model space

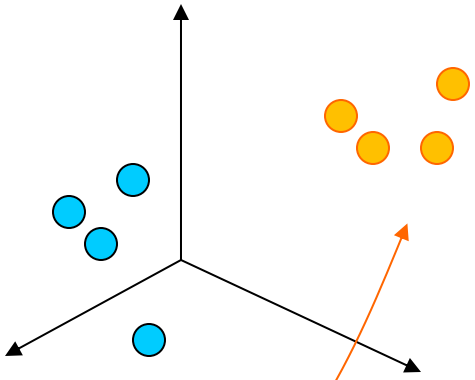


Discriminative classifiers

Query image



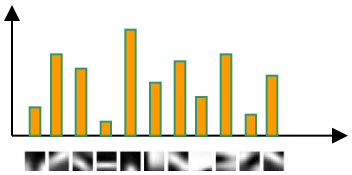
Model space



Winning class: Orange

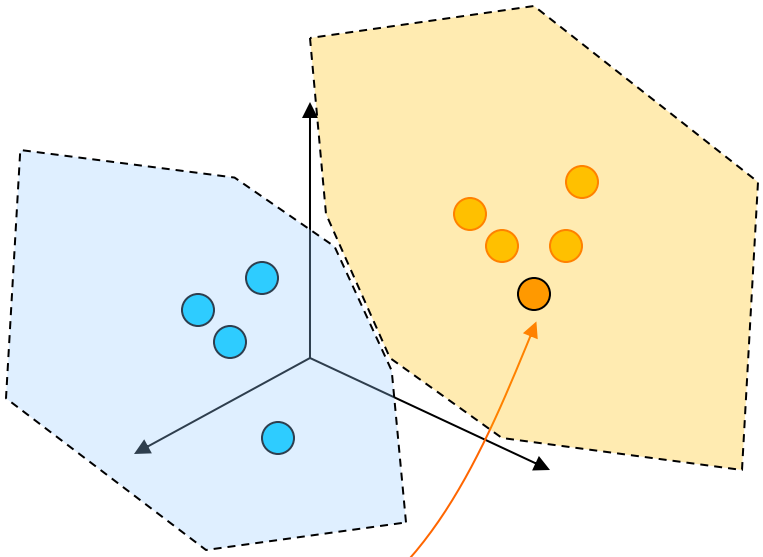
Discriminative classifiers

Query image



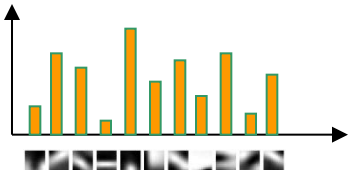
Winning class: Orange

Model space

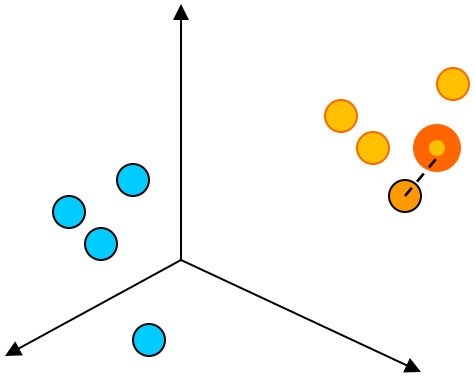


Discriminative classifiers

Query image



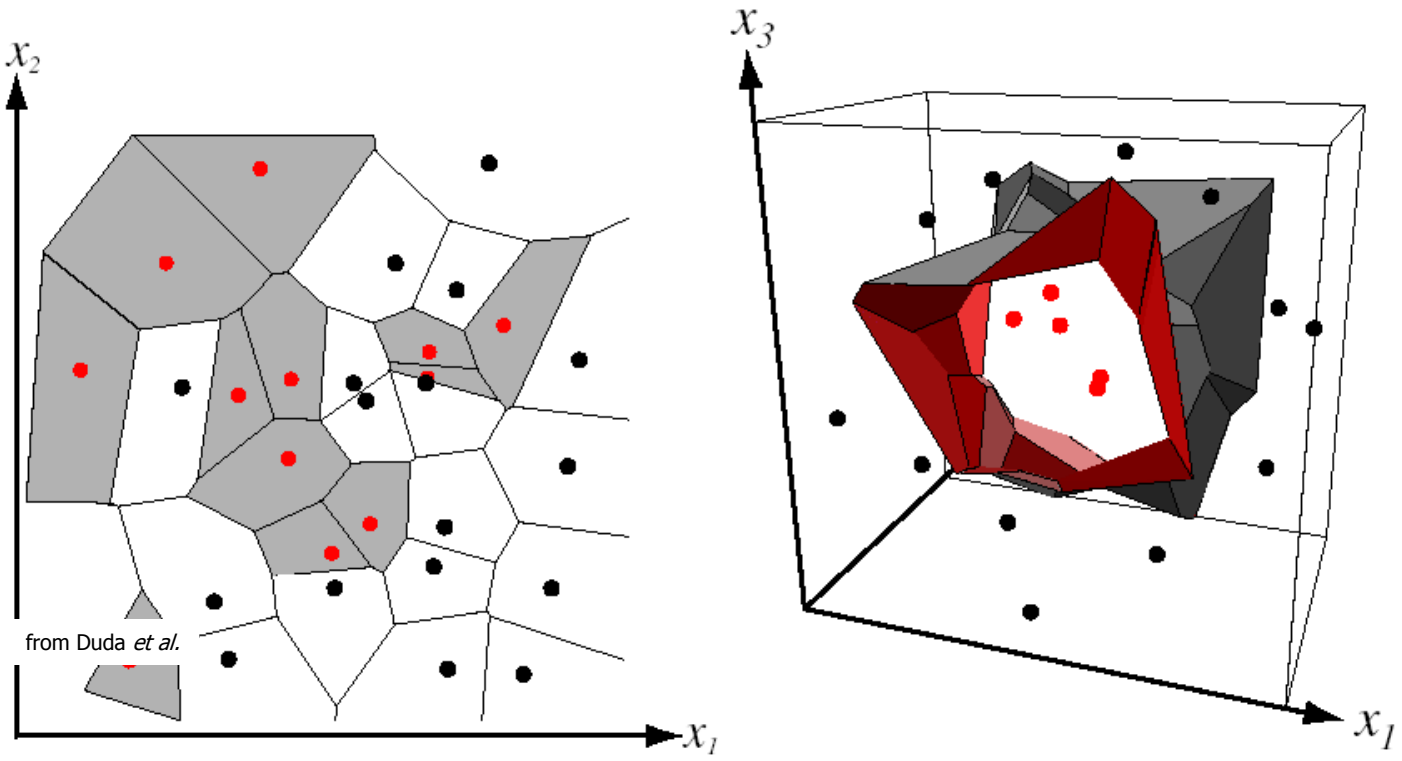
Model space



Winning class: Orange

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

Nearest Neighbors classifier

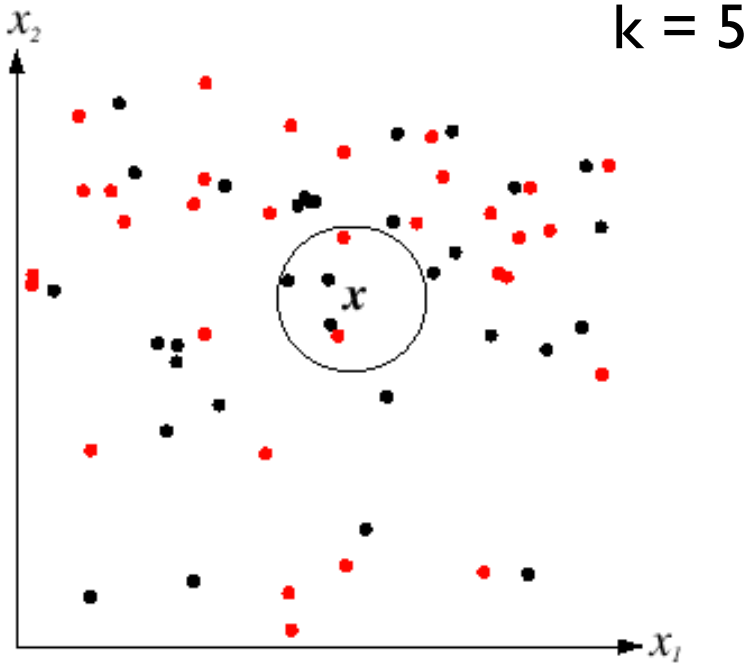


Voronoi partitioning of feature space for 2-category 2-D and 3-D data

Source: D. Lowe

K-Nearest Neighbors

- 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



Source: D. Lowe

Functions for comparing histograms

- L1 distance

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

- χ^2 distance

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

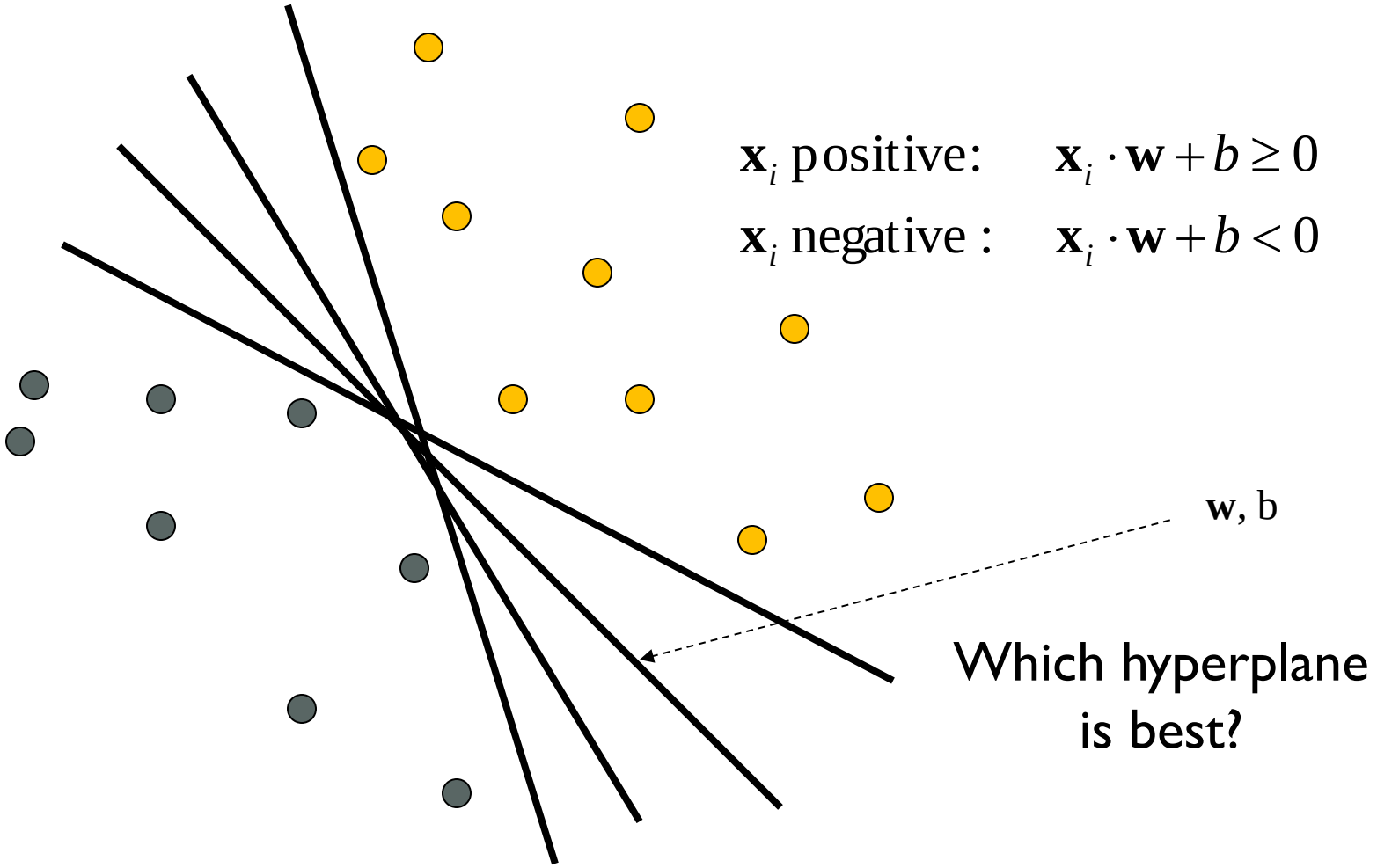
- Quadratic distance (*cross-bin*)

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

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

Linear classifiers

- Find linear function (*hyperplane*) to separate positive and negative examples

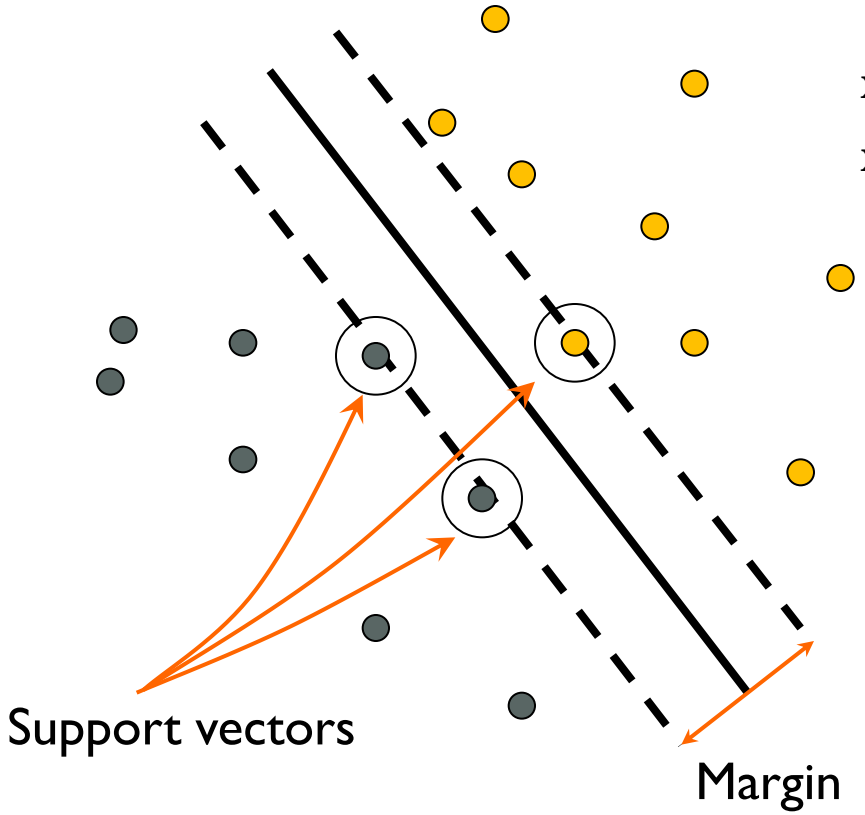


Support vector machines

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

Support vector machines

- Find **hyperplane** that maximizes the *margin* between the positive and negative examples



$\mathbf{x}_i$ positive $(y_i = 1)$: $\mathbf{x}_i \cdot \mathbf{w} + b \geq 1$

$\mathbf{x}_i$ negative $(y_i = -1)$: $\mathbf{x}_i \cdot \mathbf{w} + b \leq -1$

For support, vectors, $\mathbf{x}_i \cdot \mathbf{w} + b = \pm 1$

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

Therefore, the margin is $2 / \|\mathbf{w}\|$

Credit slide: S. Lazebnik

Finding the maximum margin hyperplane

1. Maximize margin $2/\|\mathbf{w}\|$
2. Correctly classify all training data:

$$\mathbf{x}_i \text{ positive}(y_i = 1): \quad \mathbf{x}_i \cdot \mathbf{w} + b \geq 1$$

$$\mathbf{x}_i \text{ negative}(y_i = -1): \quad \mathbf{x}_i \cdot \mathbf{w} + b \leq -1$$

- *Quadratic optimization problem*:

Minimize $\min_{\mathbf{w},b}\{\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=0}^{n}\alpha_i[c_i(\mathbf{w}\cdot\mathbf{x_i} - b) - 1]\}$

Subject to $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1$

Credit slide: S. Lazebnik

Finding the maximum margin hyperplane

- Solution:

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

learned weight

Support vector

- Classification function (decision boundary):

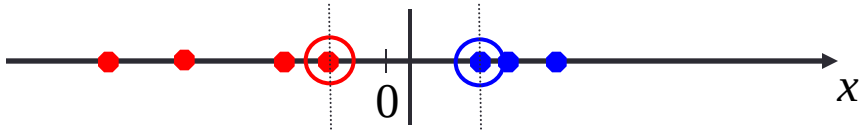
Test point

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

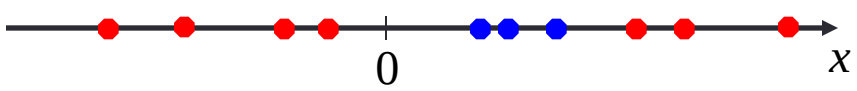
Credit slide: S. Lazebnik

Nonlinear SVMs

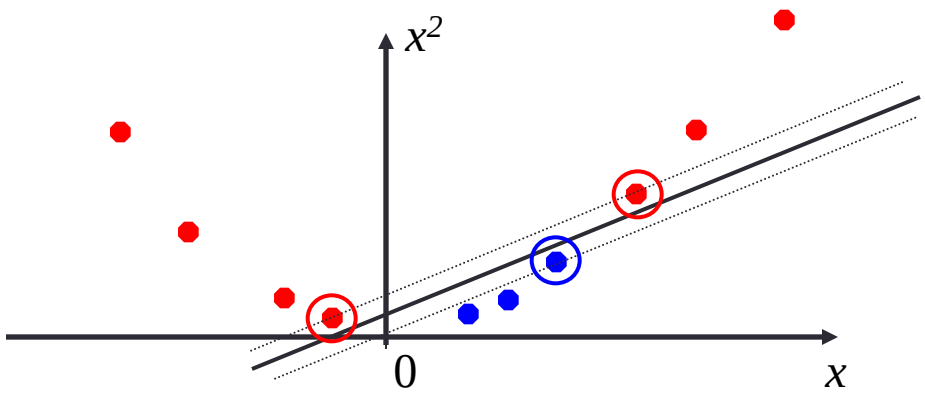
- Datasets that are linearly separable work out great:



- But what if the dataset is just too hard?



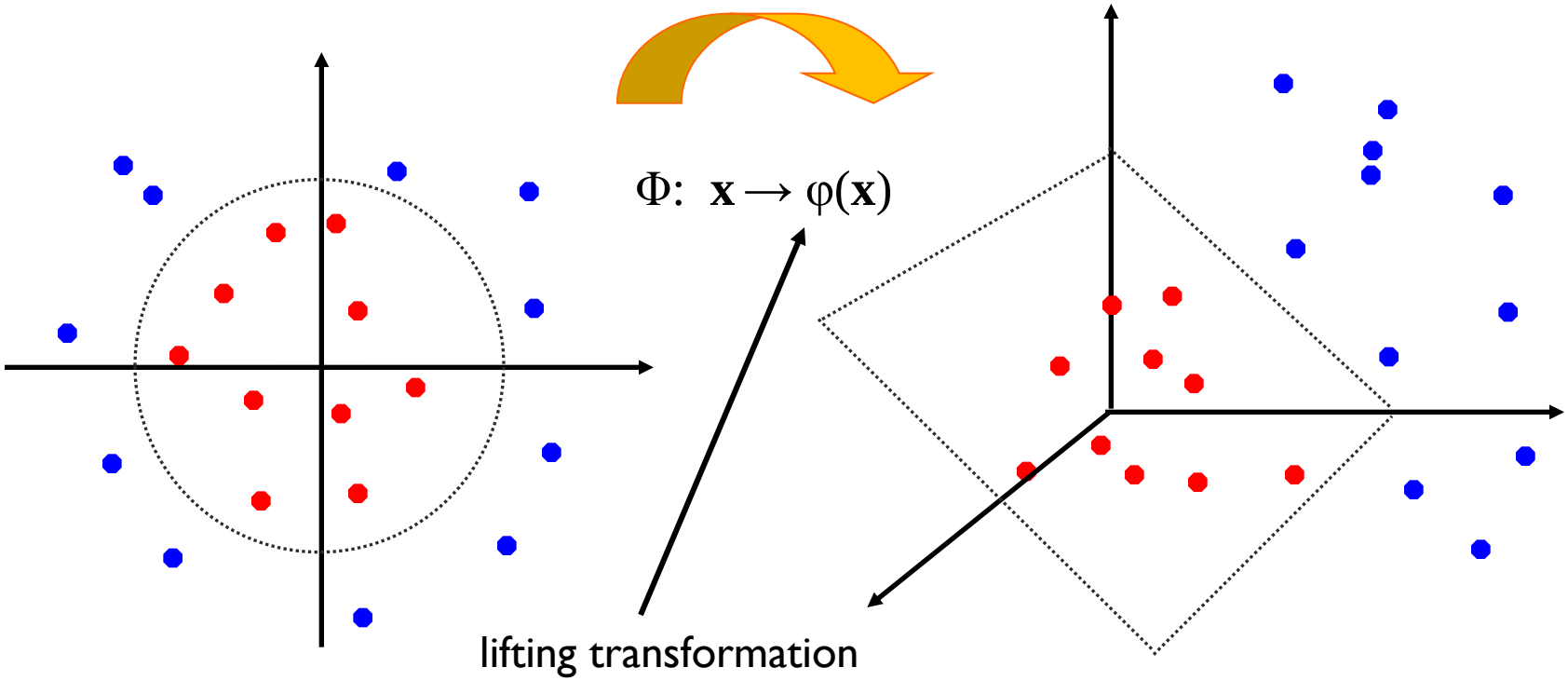
- We can map it to a higher-dimensional space:



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:

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

- *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:

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:

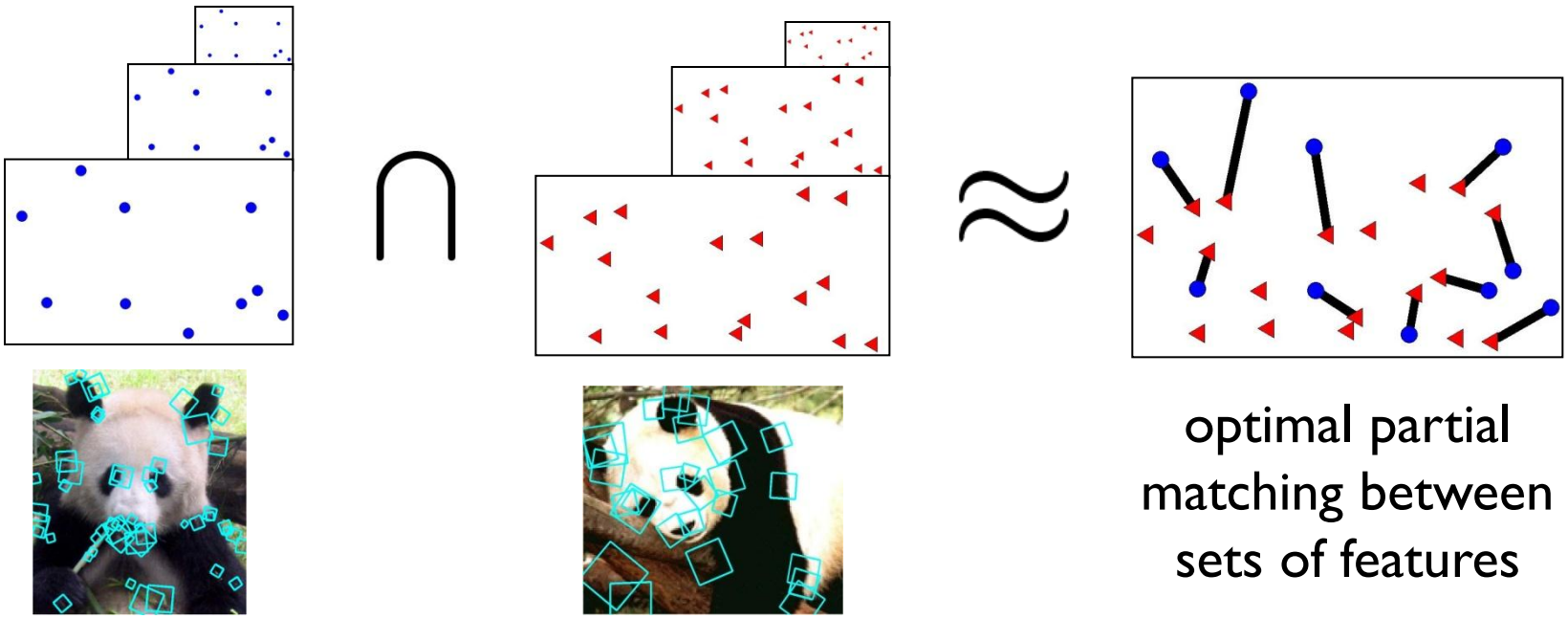
$$K(h_1, h_2) = \exp\left(-\frac{1}{A} D(h_1, h_2)^2 \right)$$

- *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

Pyramid match kernel

- Fast approximation of Earth Mover's Distance
- Weighted sum of histogram intersections at mutliple resolutions (linear in the number of features instead of cubic)



optimal partial matching between sets of features

K. Grauman and T. Darrell. The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features, ICCV 2005.

Summary: SVMs for image classification

1. Pick an image representation (in our case, bag of features)

2. Compute the matrix of kernel values between every pair of training examples

3. Feed the kernel matrix into your favorite SVM solver to obtain support vectors and weights

4. At test time: compute kernel values for your test example and each support vector, and combine them with the learned weights to get the value of the decision function

Credit slide: S. Lazebnik

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

Object recognition results

- ETH-80 database
  (Eichhorn and Chapelle 2004)
- Features:
  - Harris detector
  - PCA-SIFT descriptor, d=10

8 object classes

Kernel	Complexity	Recognition rate
Match *[Wallraven et al.]*	$O(dm^2)$	84%
Bhattacharyya affinity *[Kondor & Jebara]*	$O(dm^3)$	85%
Pyramid match	$O(dmL)$	84%

Slide credit: Kristen Grauman

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

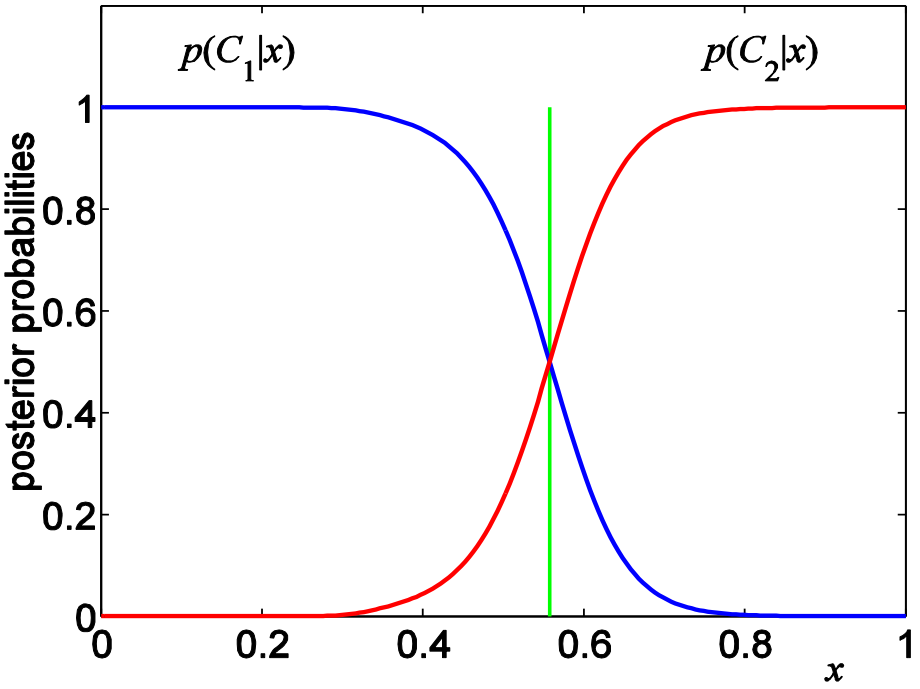
- Cons
  - No "direct" multi-class SVM, must combine two-class SVMs
  - Computation, memory
    - During training time, must compute matrix of kernel values for every pair of examples
    - Learning can take a very long time for large-scale problems

Of course, there are many other classifiers out there

- Neural networks
- Boosting
- Decision trees, …

Summary

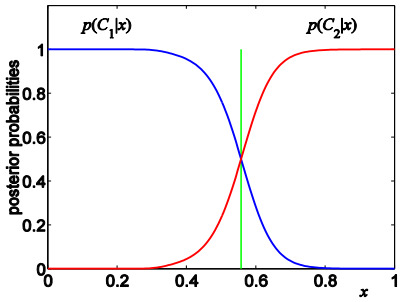
$$\frac{p(\text{class} \mid \text{image})}{p(\text{no class} \mid \text{image})} = \frac{p(\text{image} \mid \text{class})}{p(\text{image} \mid \text{no class})} \cdot \frac{p(\text{class})}{p(\text{no class})}$$



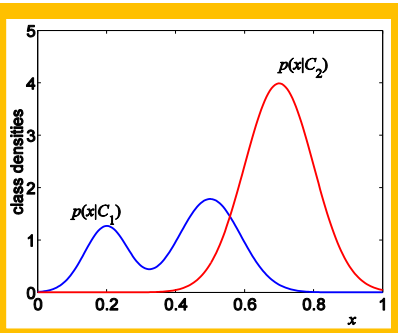
- Discriminative methods model posterior

Learning and Recognition

1. Discriminative method:
   - NN (Nearest Neighbor)
   - SVM (Support Vector Machine)



2. Generative method:
   - Graphical models



→ Model the probability distribution that produces a given bag of features

Generative Models

1. Naïve Bayes classifier
   - Csurka Bray, Dance & Fan, 2004

2. Hierarchical Bayesian text models (pLSA and LDA)
   - Background: Hoffman 2001, Blei, Ng & Jordan, 2004
   - Object categorization: Sivic et al. 2005, Sudderth et al. 2005
   - Natural scene categorization: Fei-Fei et al. 2005

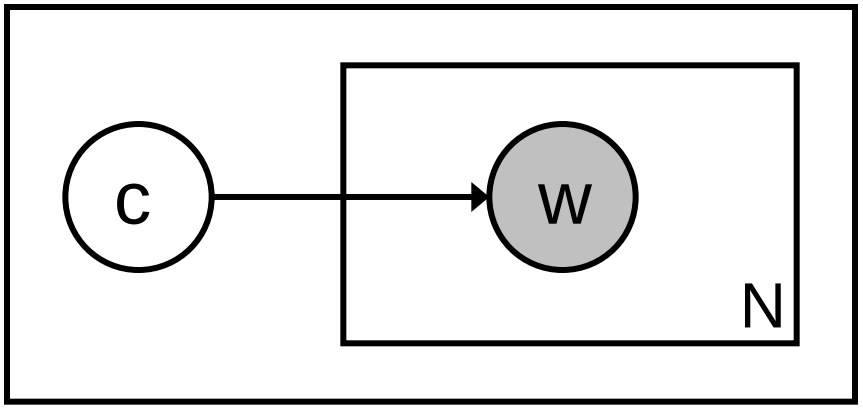
Some Notation

- **w:** a collection of all N codewords in the image

$\mathbf{w} = [w_1, w_2, \ldots, w_N]$

- **c:** category of the image

The Naïve Bayes model



Graphical model

$$p(c \mid w) \sim p(c)p(w \mid c)$$

Prior prob. of the object classes

Image likelihood given the class

Csurka et al. 2004

The Naïve Bayes model

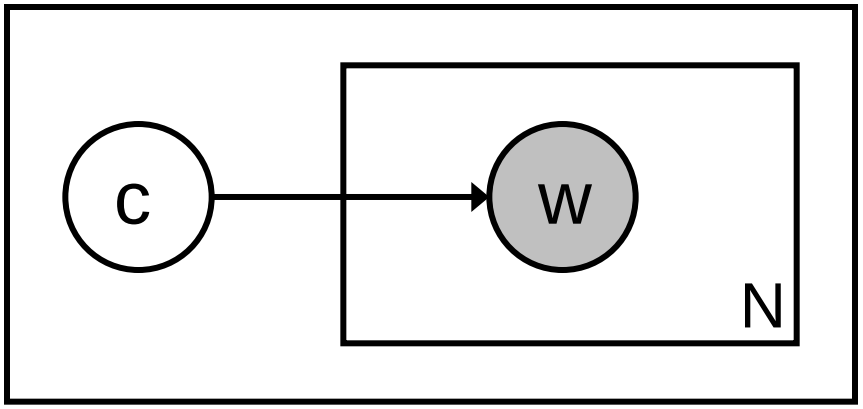
$$p(w_1, \ldots, w_N \mid c)$$

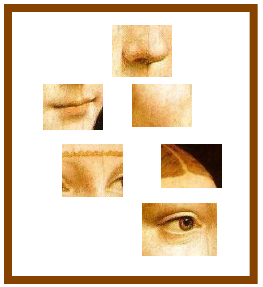
- Assume that each feature is conditionally independent *given the class*

$$p(w_1, \ldots, w_N \mid c) = \prod_{i=1}^{N} p(w_i \mid c)$$

Csurka et al. 2004

The Naïve Bayes model





$$c^* = \arg\max_c \; p(c \mid w) \propto p(c)\, p(w \mid c) = p(c) \prod_{n=1}^{N} p(w_n \mid c)$$

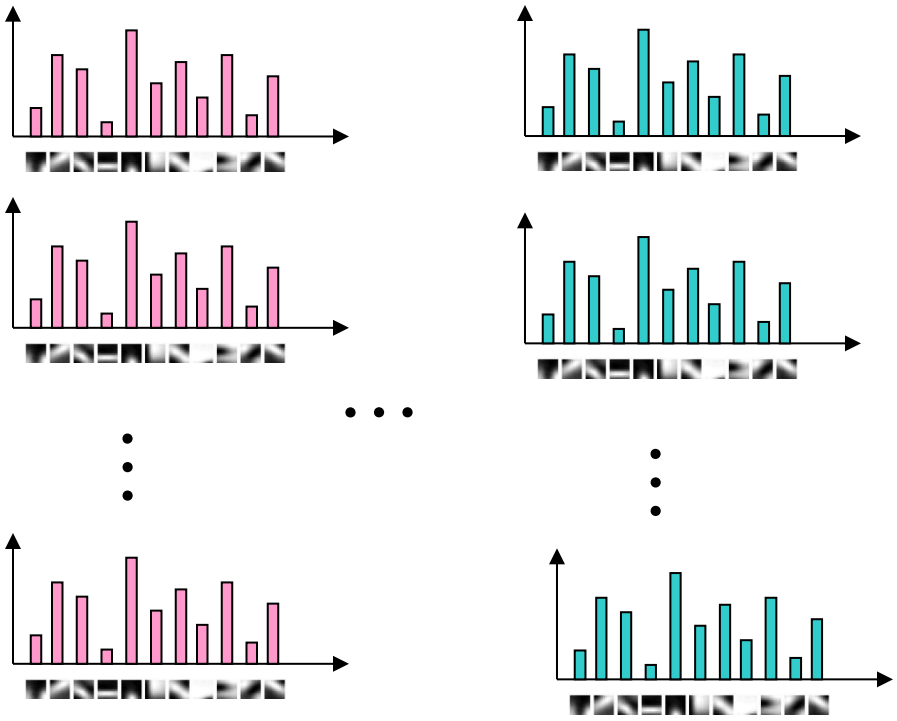
Object class decision

Likelihood of *nth* visual word given the class

Estimated by empirical frequencies of code words in images from a given class

Csurka et al. 2004

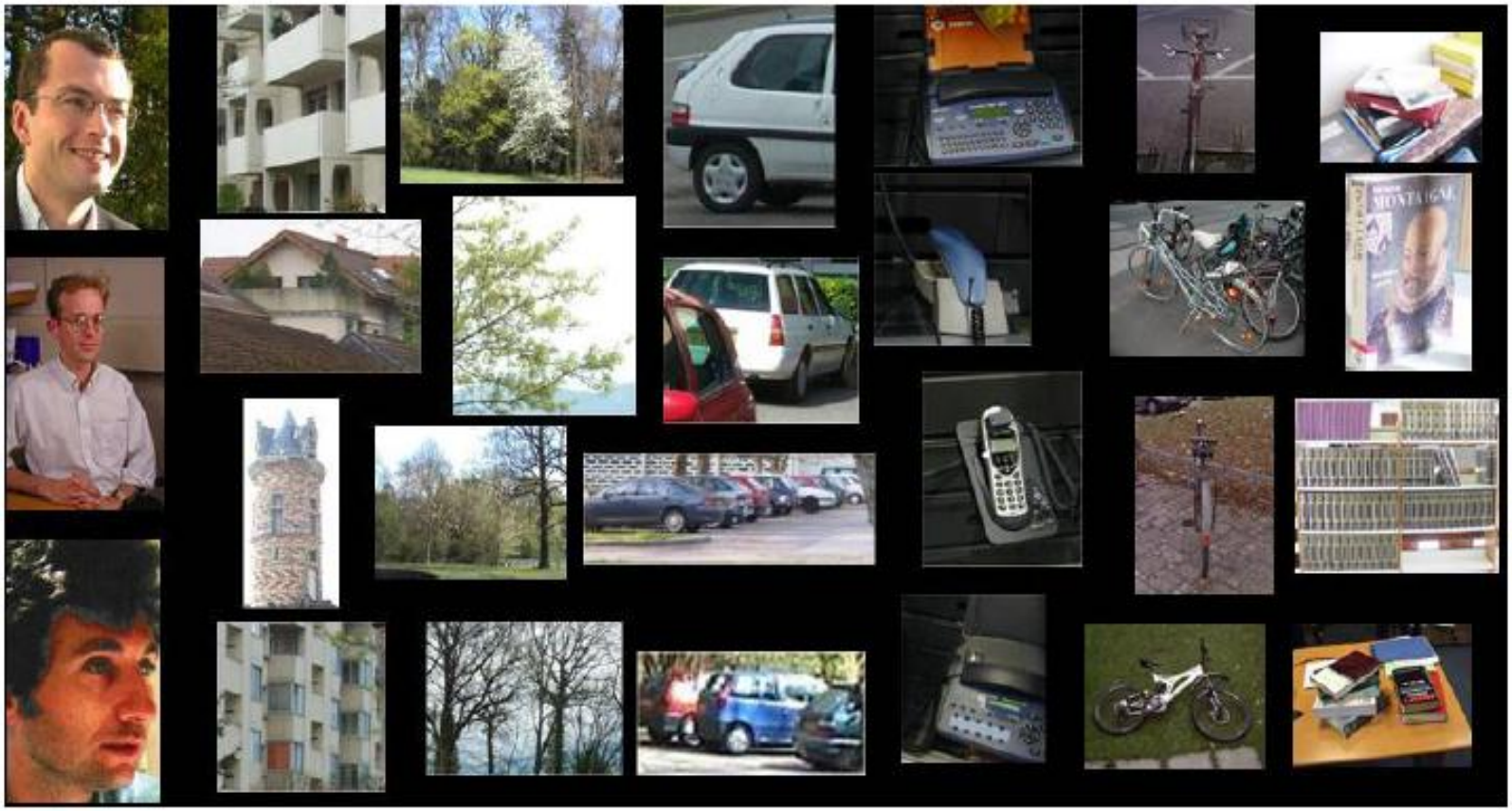
Category Models



Class 1

Class N

Our in-house database contains 1776 images in seven classes[1]: faces, buildings, trees, cars, phones, bikes and books. Fig. 2 shows some examples from this dataset.



Csurka et al. 2004

Table 1. Confusion matrix and the mean rank for the best vocabulary (*k=1000*).

True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**76**	4	2	3	4	4	13
buildings	2	**44**	5	0	5	1	3
trees	3	2	**80**	0	0	5	0
cars	4	1	0	**75**	3	1	4
phones	9	15	1	16	**70**	14	11
bikes	2	15	12	0	8	**73**	0
books	4	19	0	6	7	2	**69**
Mean ranks	1.49	1.88	1.33	1.33	1.63	1.57	1.57

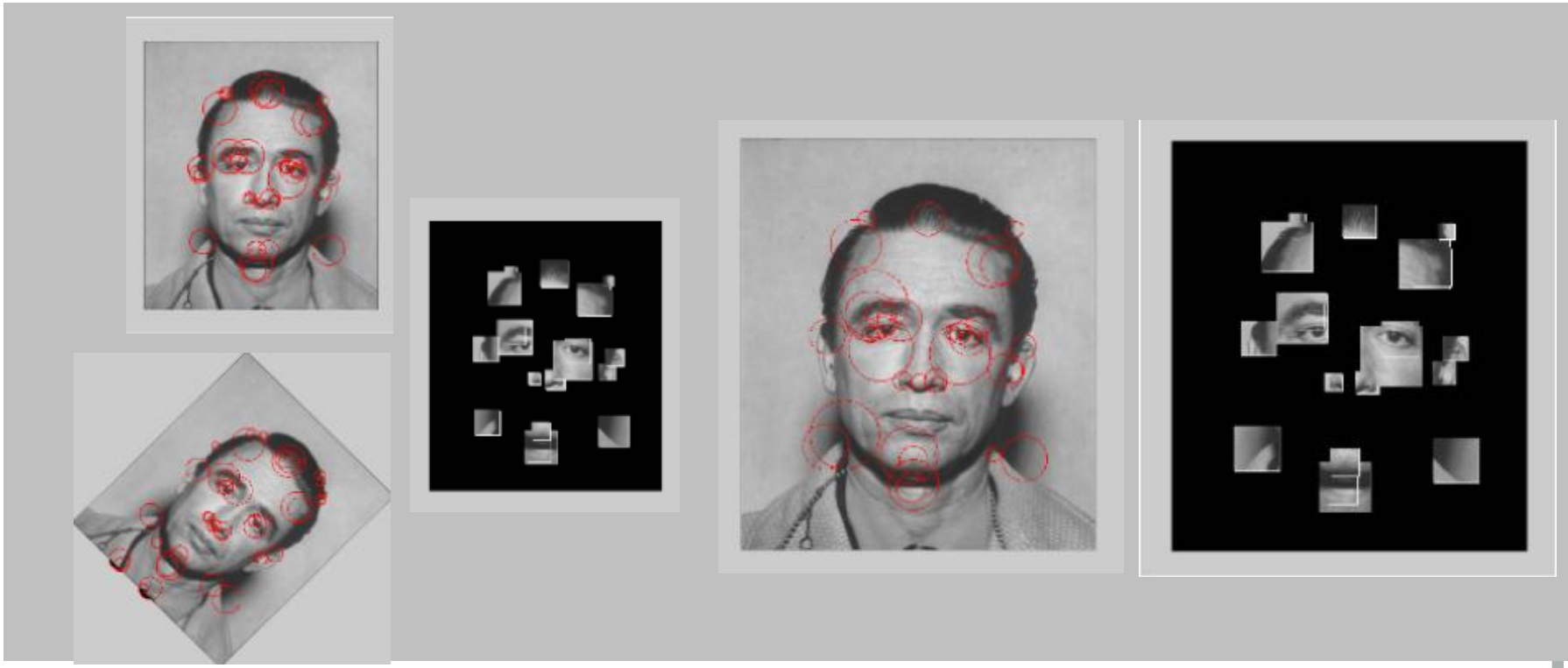
Csurka et al. 2004

Summary: Generative models

- Naïve Bayes
  - *Unigram models* in document analysis
  - Assumes **conditional independence** of words given class
  - Parameter estimation: frequency counting

Invariance issues

- Scale and rotation
  - Implicit
  - Detectors and descriptors



Kadir and Brady. 2003

Invariance issues

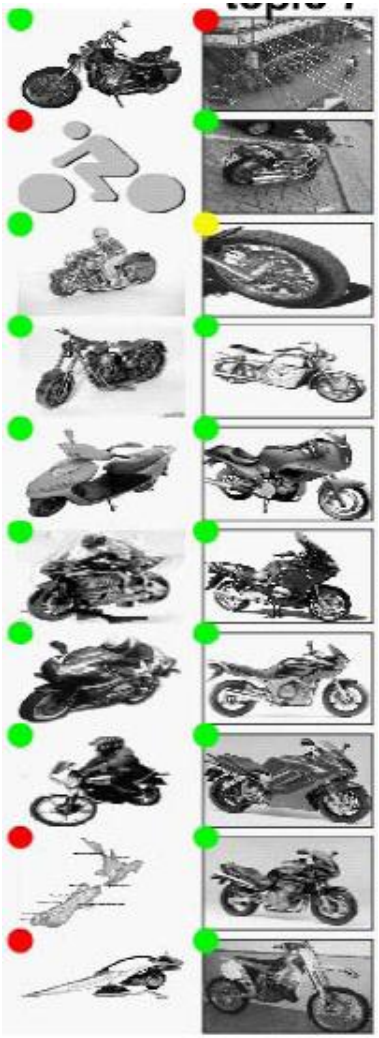
- Scale and rotation
- Occlusion
  - Implicit in the models
  - Codeword distribution: small variations
  - (In theory) Theme (z) distribution: different occlusion patterns

Invariance issues

- Scale and rotation
- Occlusion
- Translation
  - Encode (relative) location information
    - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
    - Niebles & Fei-Fei, 2007

Invariance issues

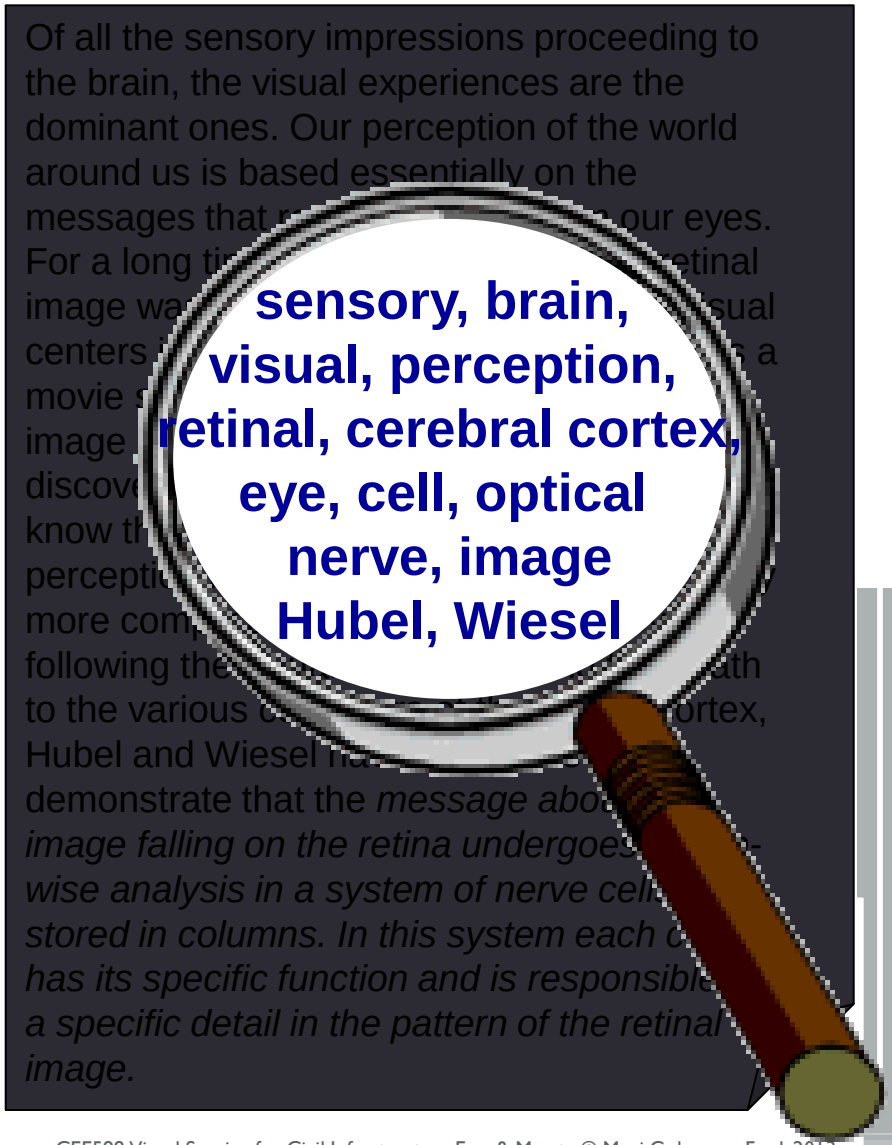
- Scale and rotation
- Occlusion
- Translation
- View point (in theory)
  - Codewords: detector and descriptor
  - Theme distributions: different view points



Fergus, Fei-Fei, Perona & Zisserman, 2005

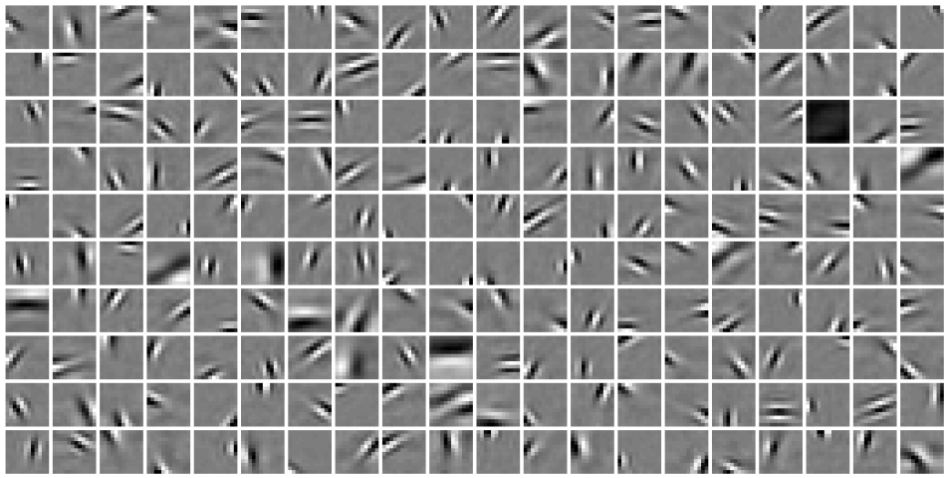
Model properties

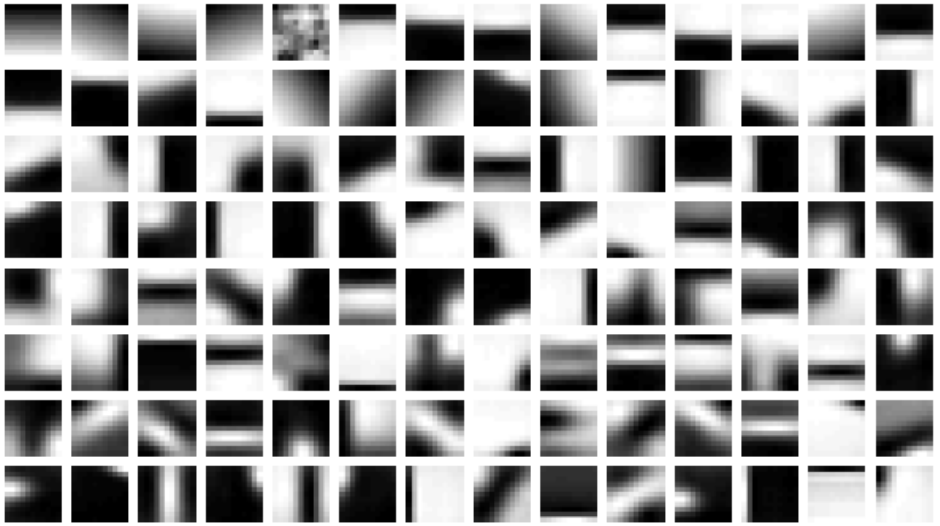
- Intuitive
  - Analogy to documents



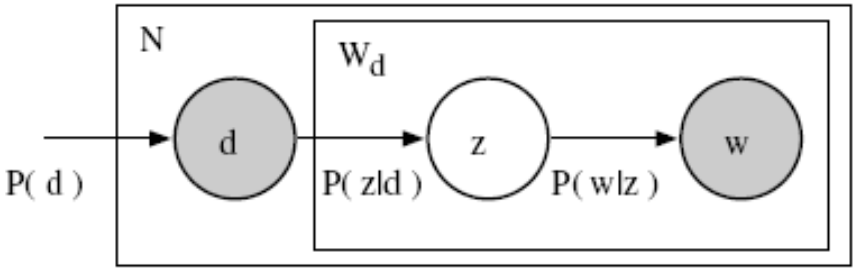
Model properties

- Intuitive
  - Analogy to documents
  - Analogy to human vision



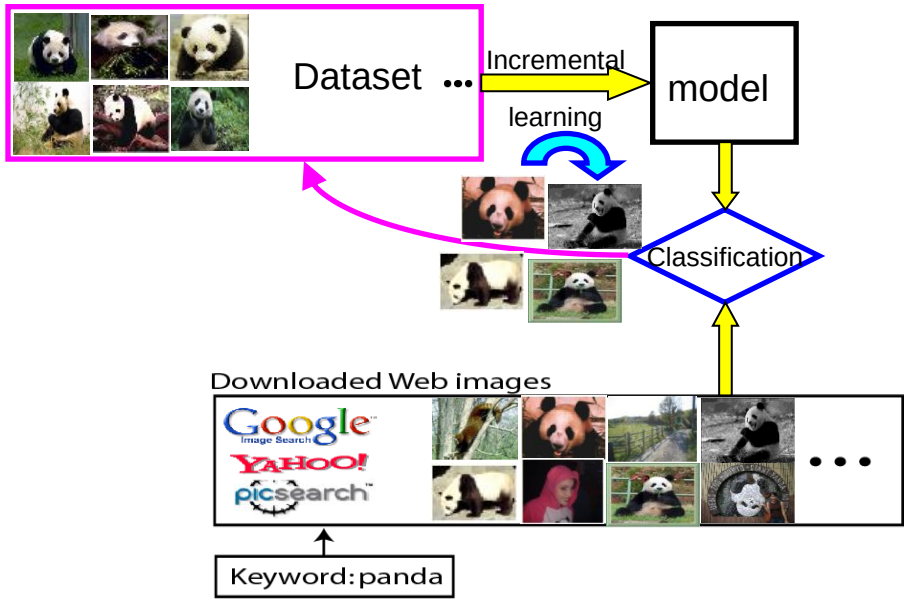


Model properties



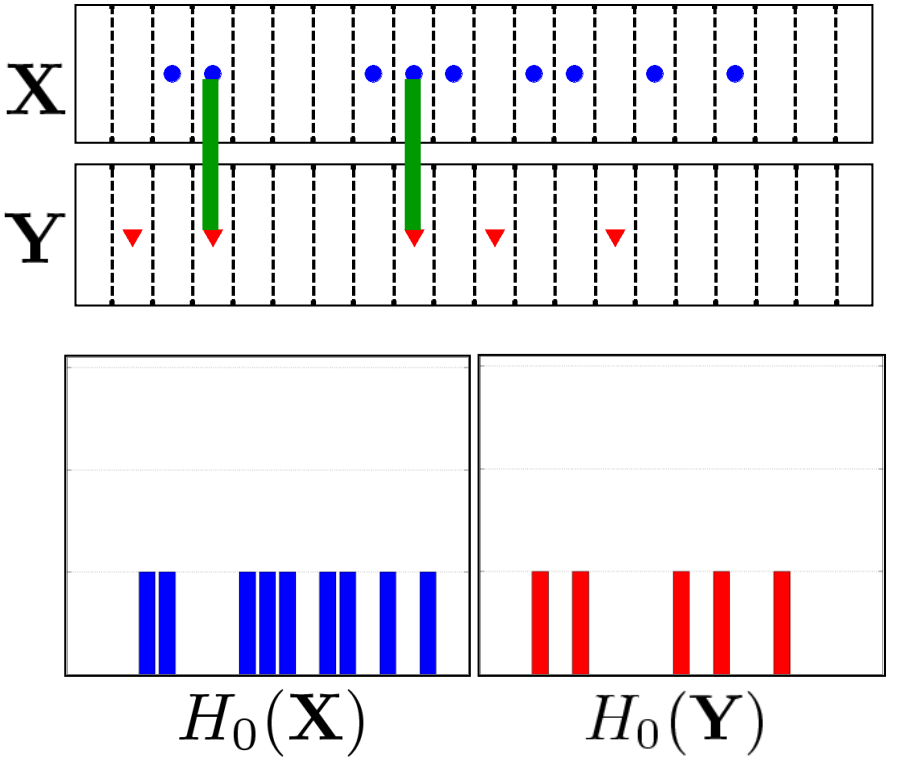
Sivic, Russell, Efros, Freeman, Zisserman, 2005

- Intuitive
- generative models
  - Convenient for weakly- or un-supervised, incremental training
  - Prior information
  - Flexibility (e.g. HDP)



Li, Wang & Fei-Fei, CVPR 2007

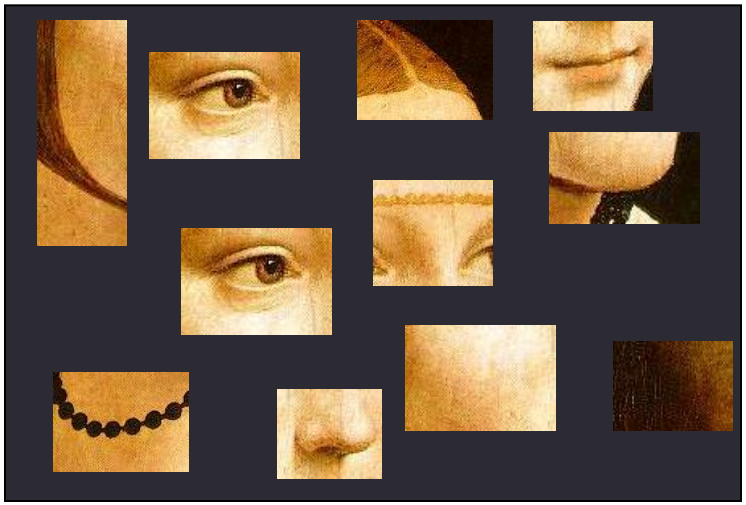
Model properties

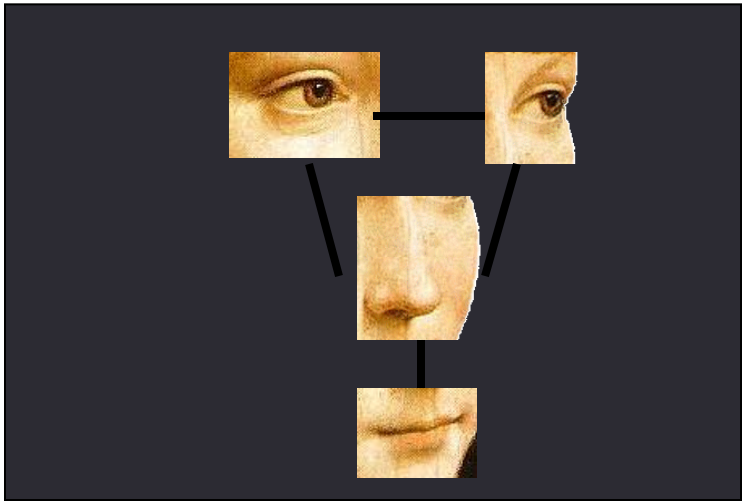


- Intuitive
- generative models
- Discriminative method
  - Computationally efficient
  - fast

Grauman et al. CVPR 2005

Model properties





- Intuitive
- generative models
- Discriminative method
- Learning and recognition relatively fast
  - Compare to other methods

Weakness of the model

- No rigorous geometric information of the object components
- It's intuitive to most of us that objects are made of parts – no such information
- Not extensively tested yet for
  - View point invariance
  - Scale invariance
- Segmentation and localization unclear

Outline

- Object Recognition

- Introduction
- Recognition of single 3D objects
  - Bag of world models
  - Part based models
  - Models for 3D objects categorization