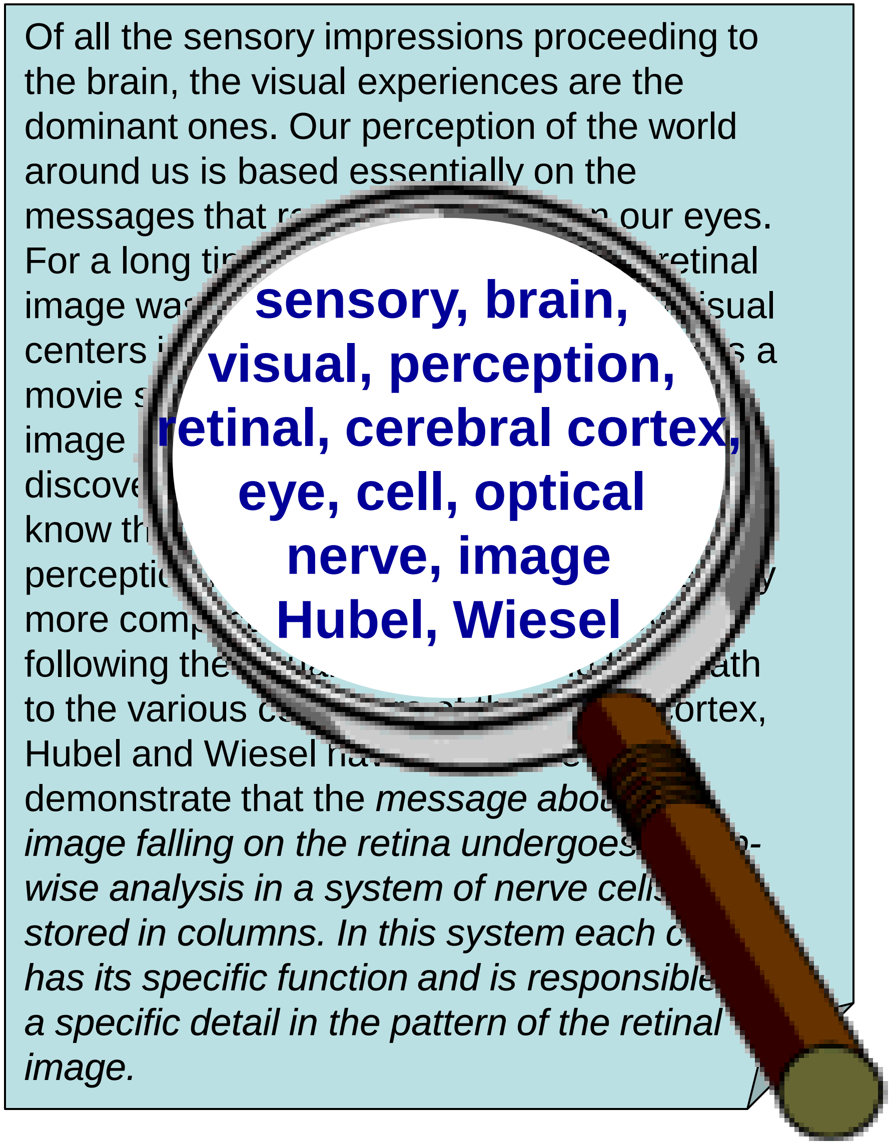
Part 1: Bag-of-words models

by Li Fei-Fei (Princeton)

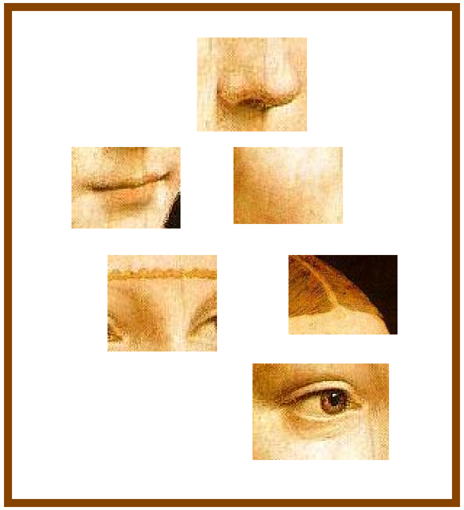
Object → **Bag of ‘words’**

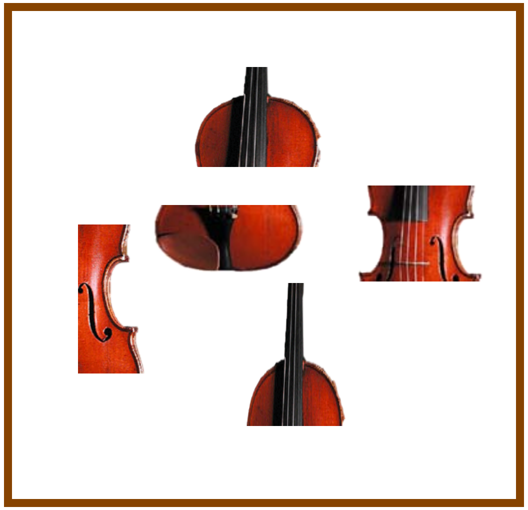
Analogy to documents



A clarification: definition of “BoW”

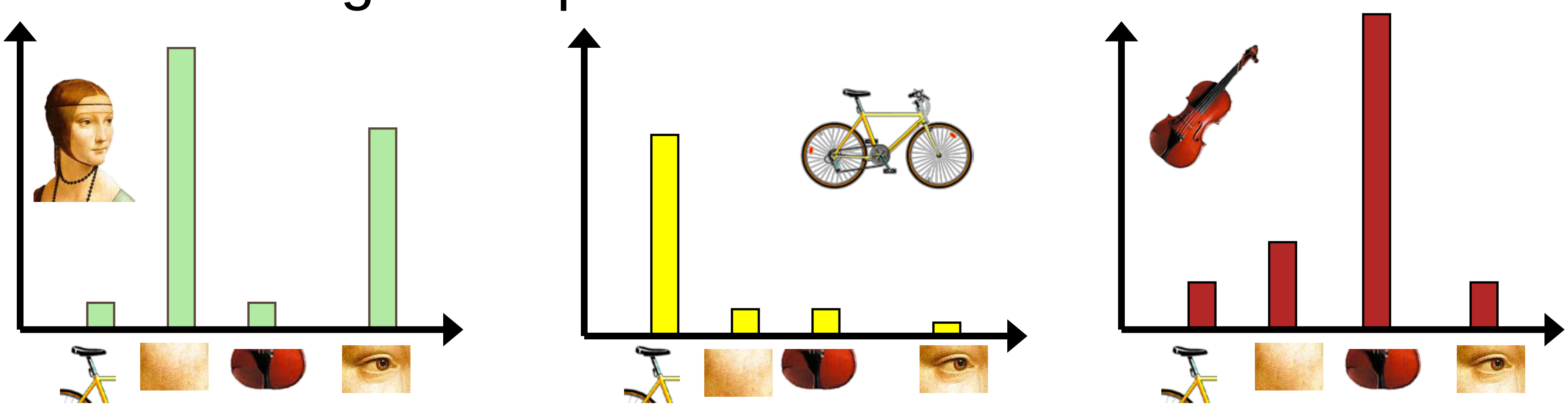
- Looser definition
  - Independent features

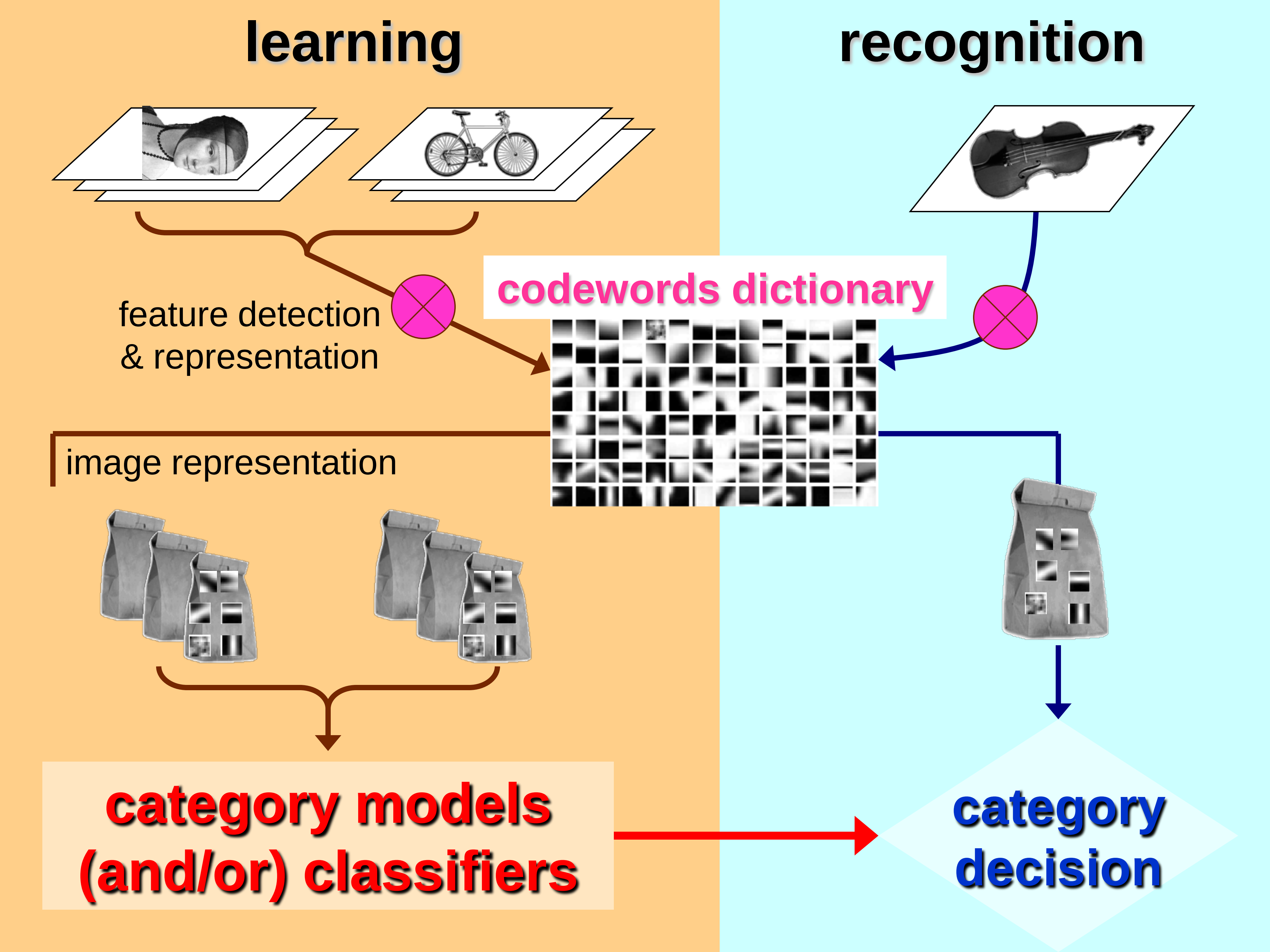




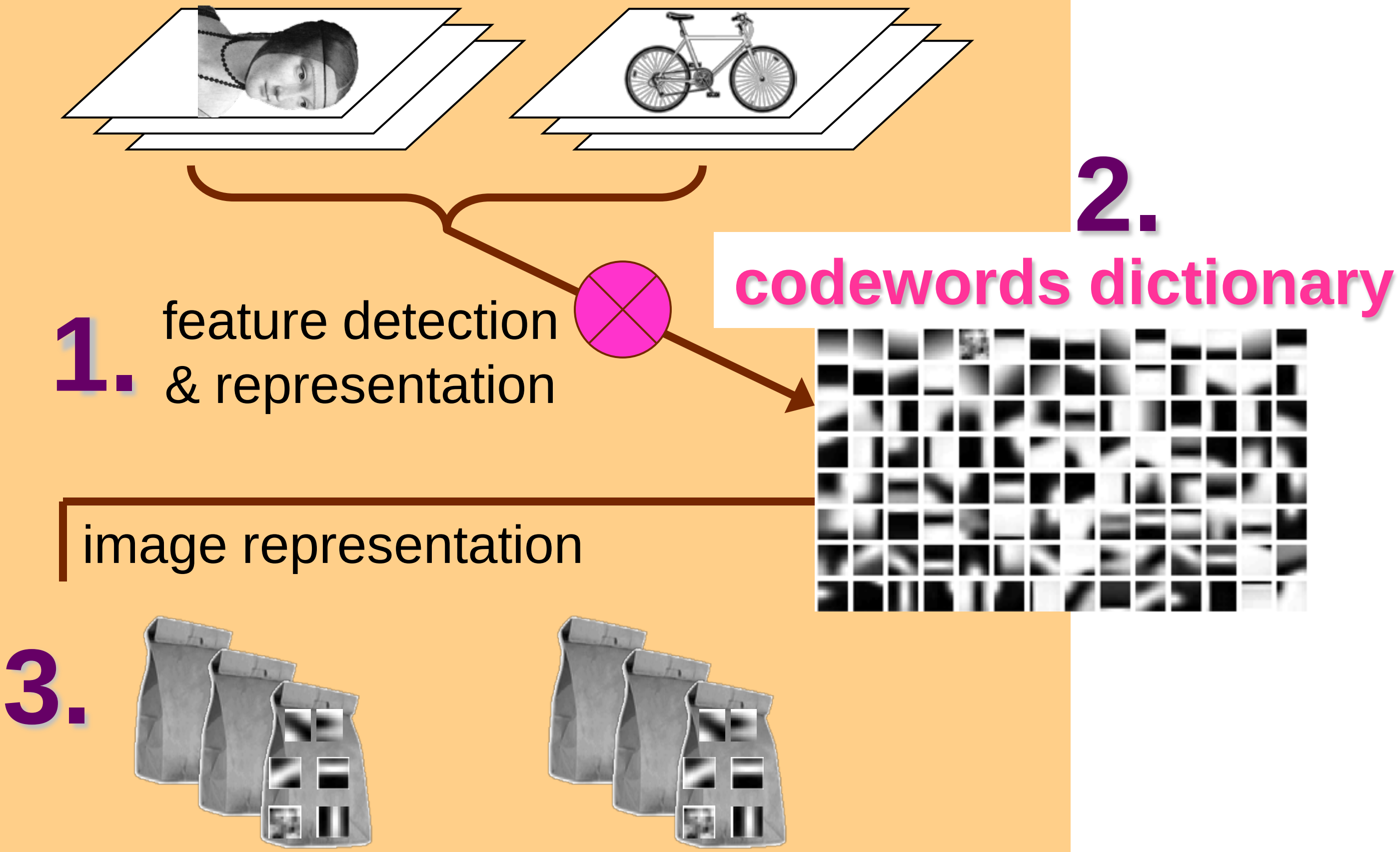
A clarification: definition of “BoW”

- Looser definition
  - Independent features
- Stricter definition
  - Independent features
  - histogram representation

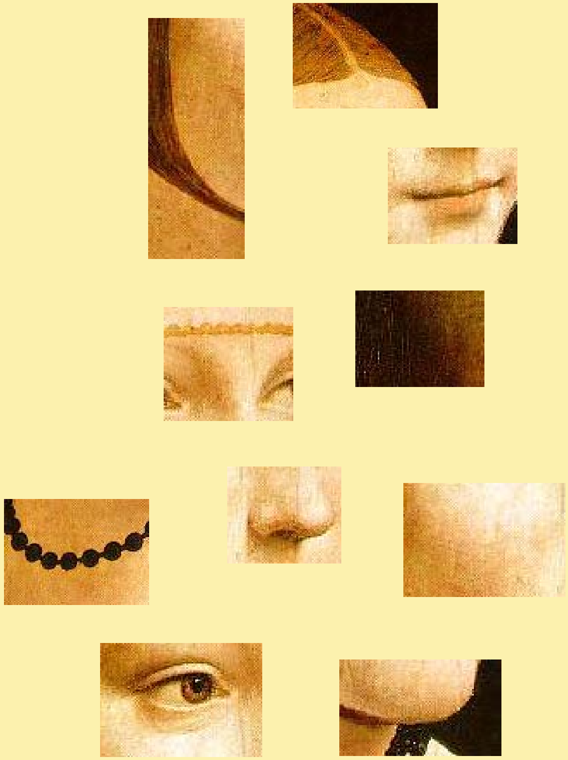



learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

Representation

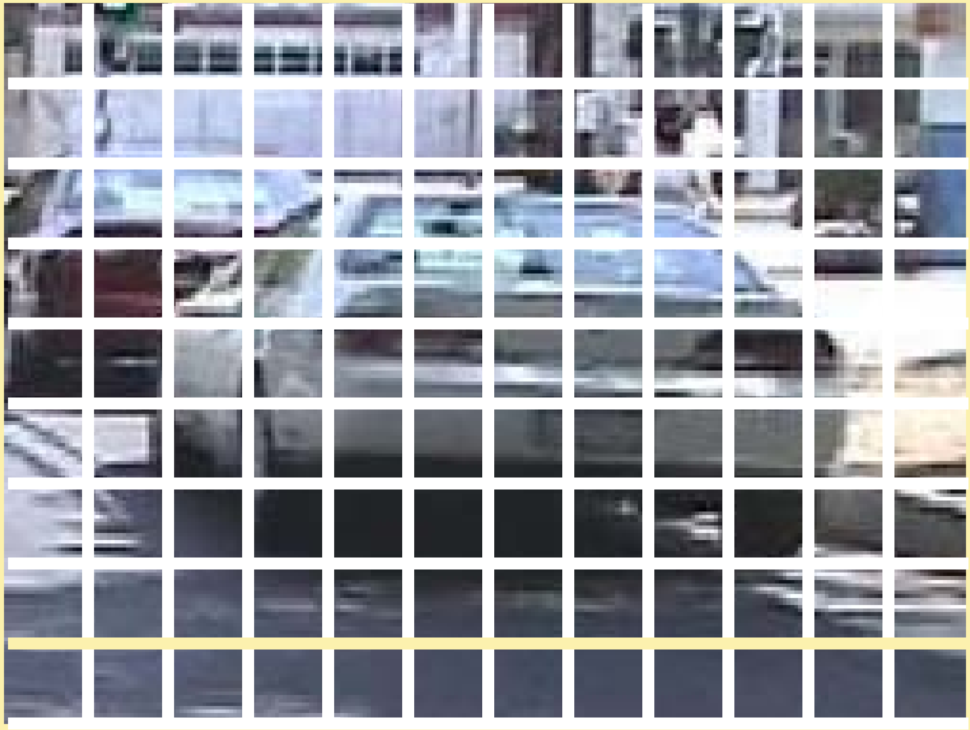


1.Feature detection and representation



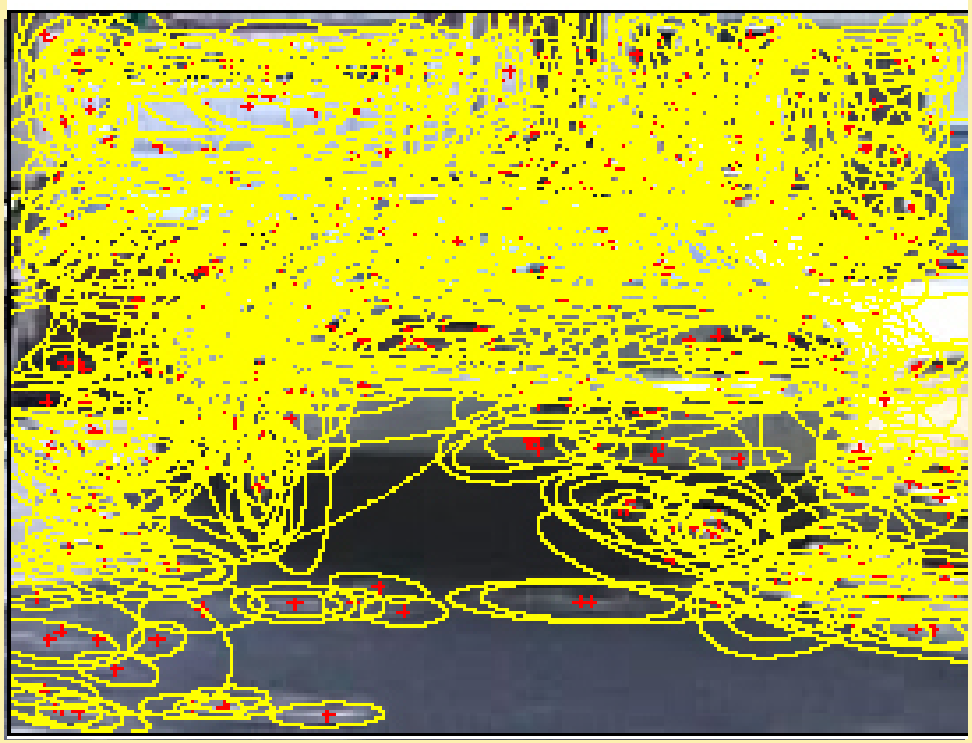
1.Feature detection and representation

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



1.Feature detection and representation

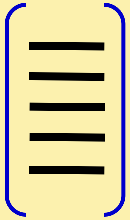
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic, et al. 2005



1.Feature detection and representation

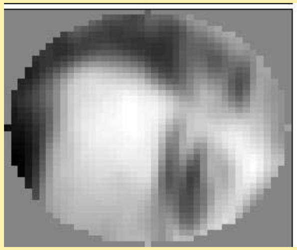
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, Bray, Dance & Fan, 2004
  - Fei-Fei & Perona, 2005
  - Sivic, Russell, Efros, Freeman & Zisserman, 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation based patches (Barnard, Duygulu, Forsyth, de Freitas, Blei, Jordan, 2003)

1.Feature detection and representation

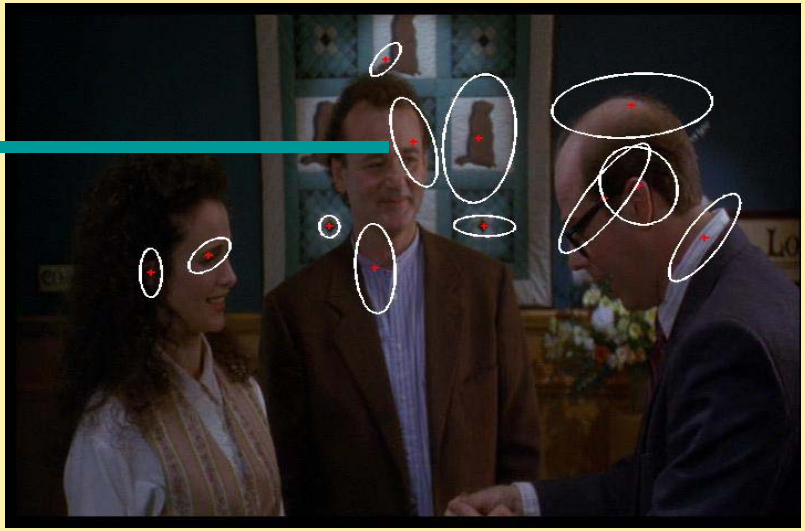


Compute SIFT descriptor

[Lowe’99]



Normalize patch



Detect patches

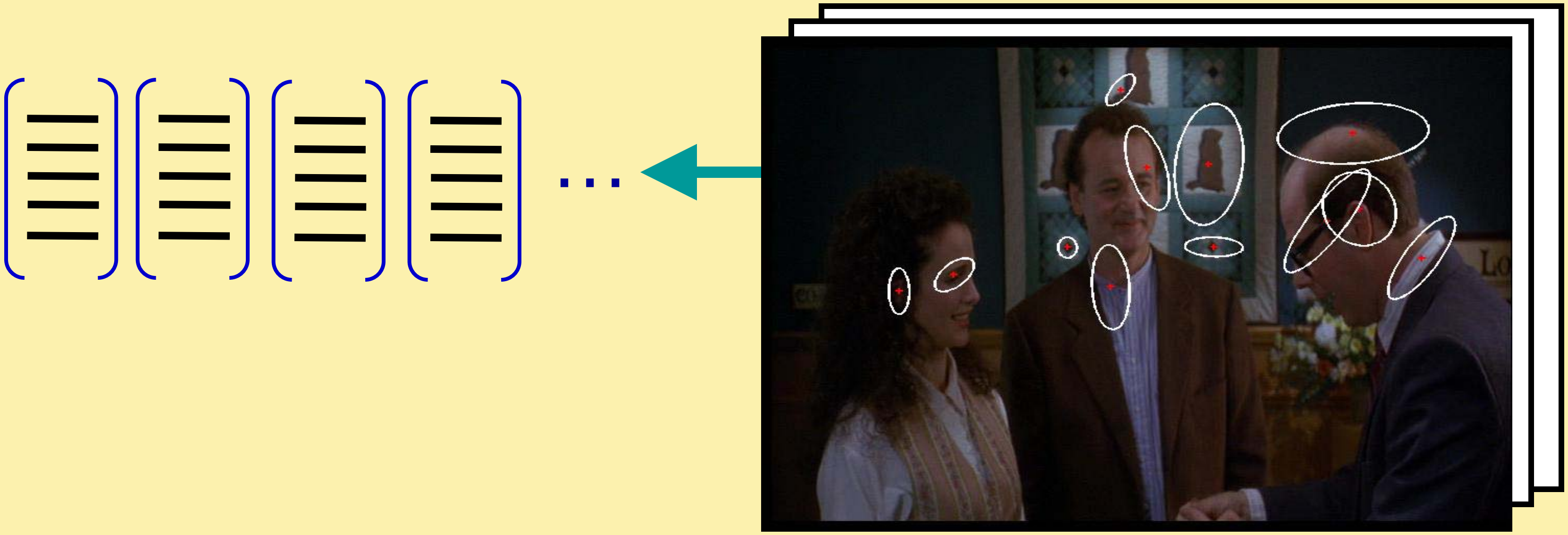
[Mikojaczyk and Schmid ’02]

[Mata, Chum, Urban & Pajdla, ’02]

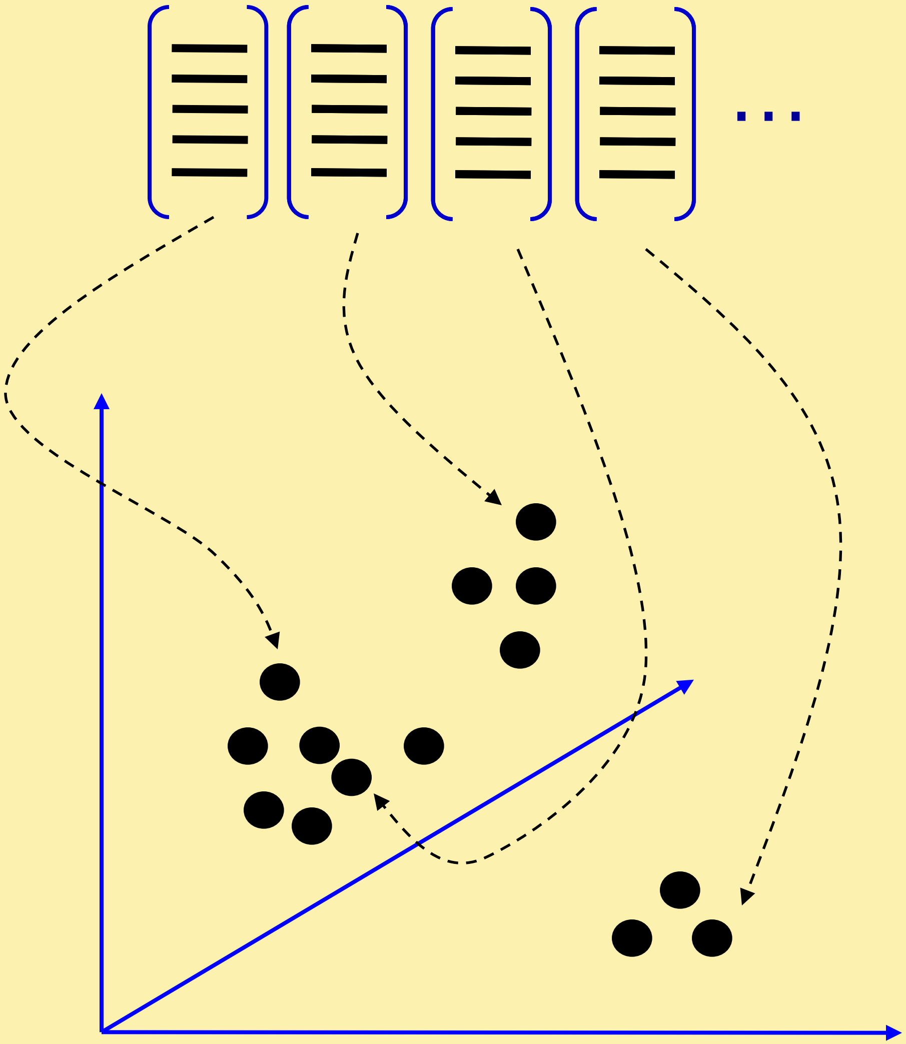
[Sivic & Zisserman, ’03]

Slide credit: Josef Sivic

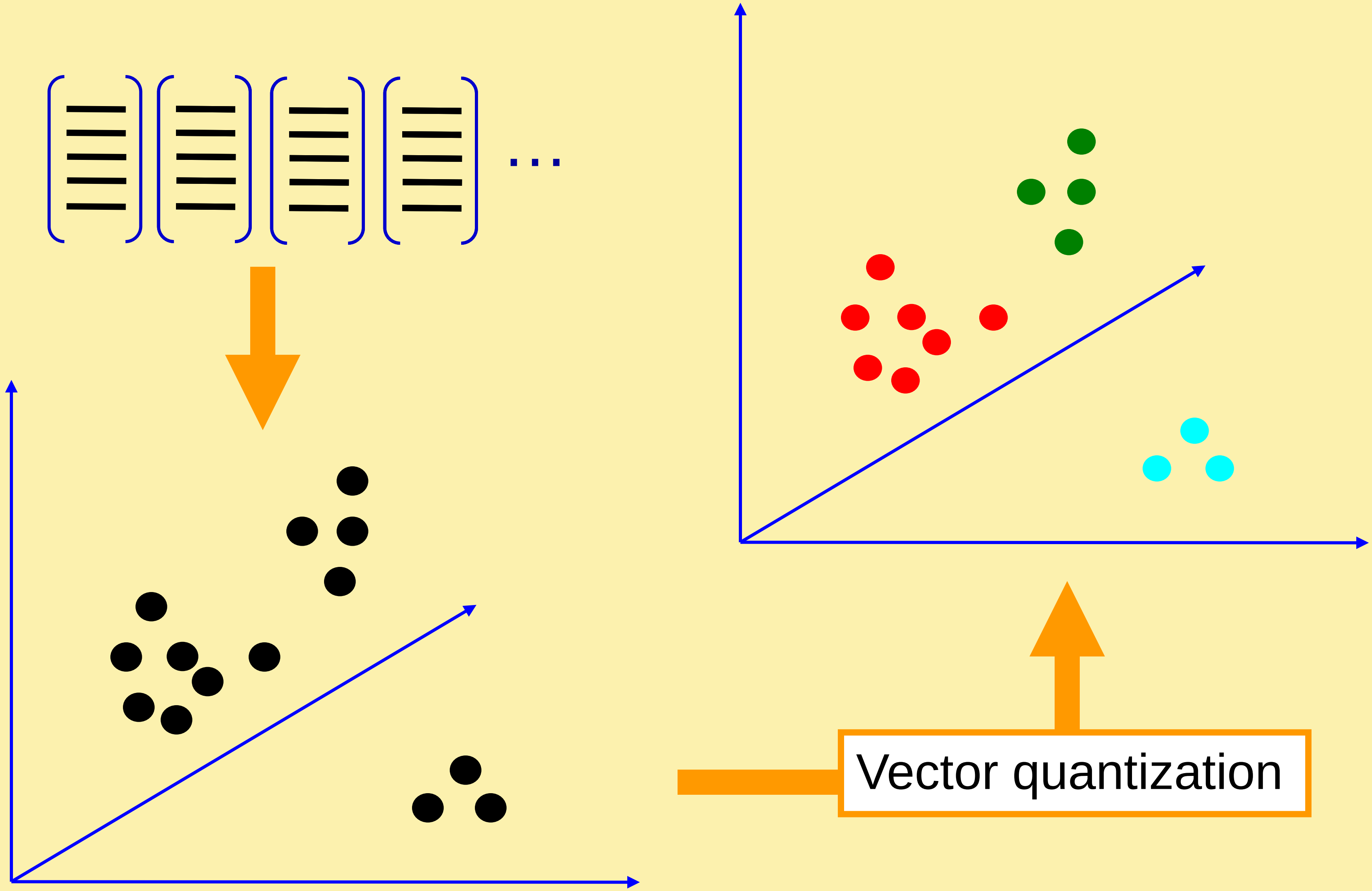
1.Feature detection and representation



2. Codewords dictionary formation

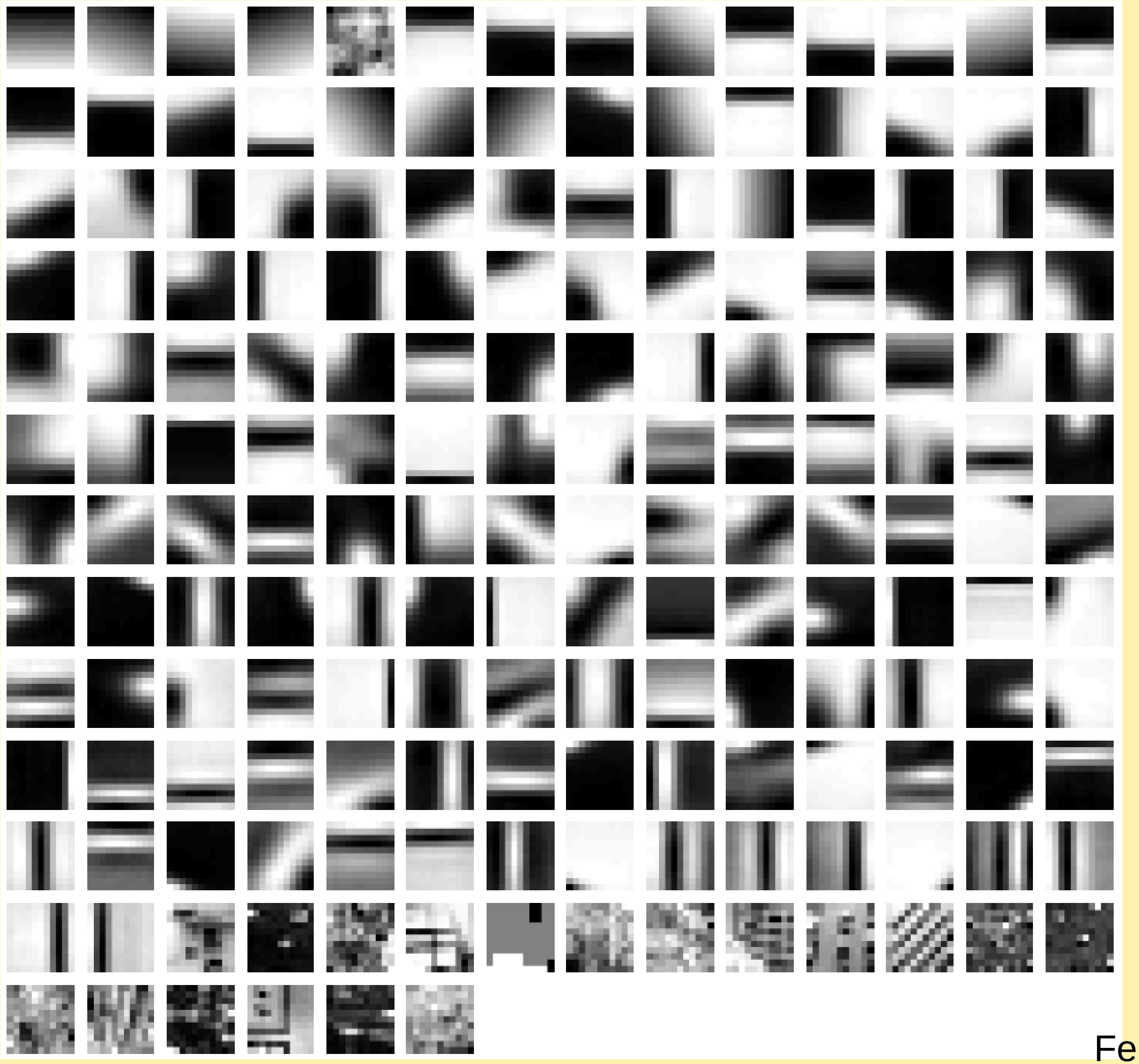


2. Codewords dictionary formation



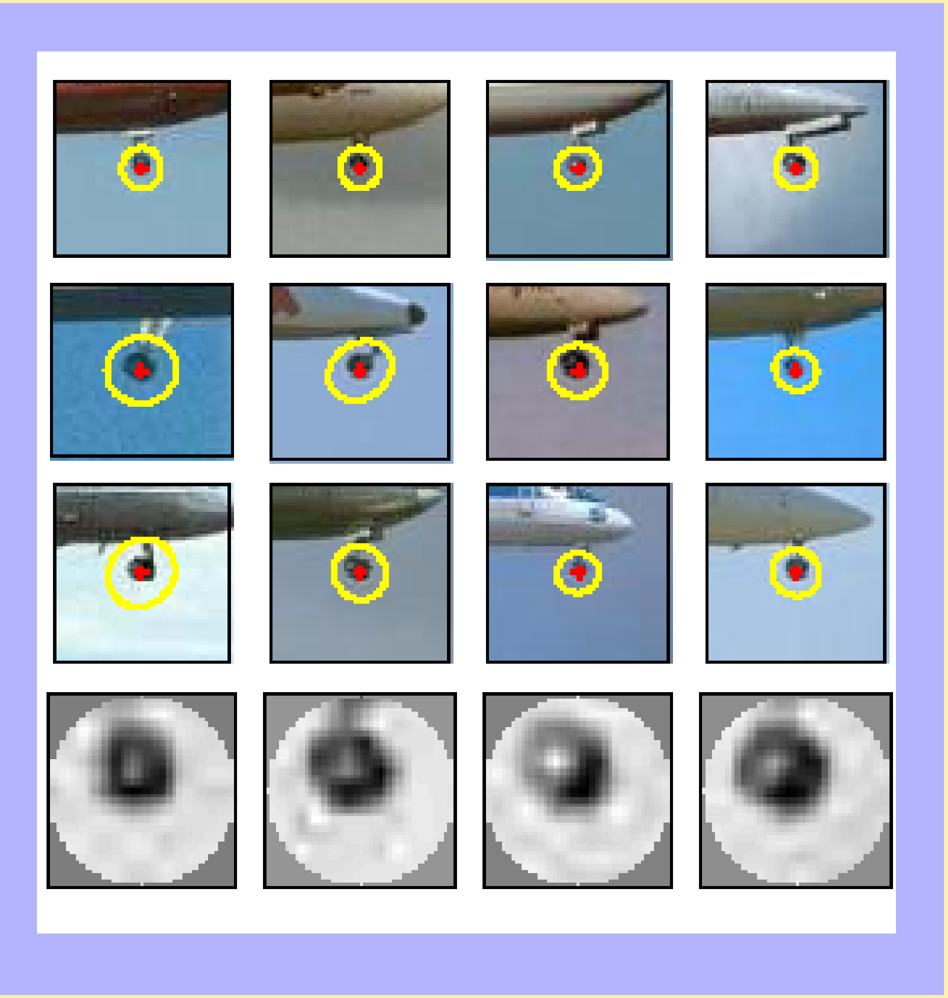
Slide credit: Josef Sivic

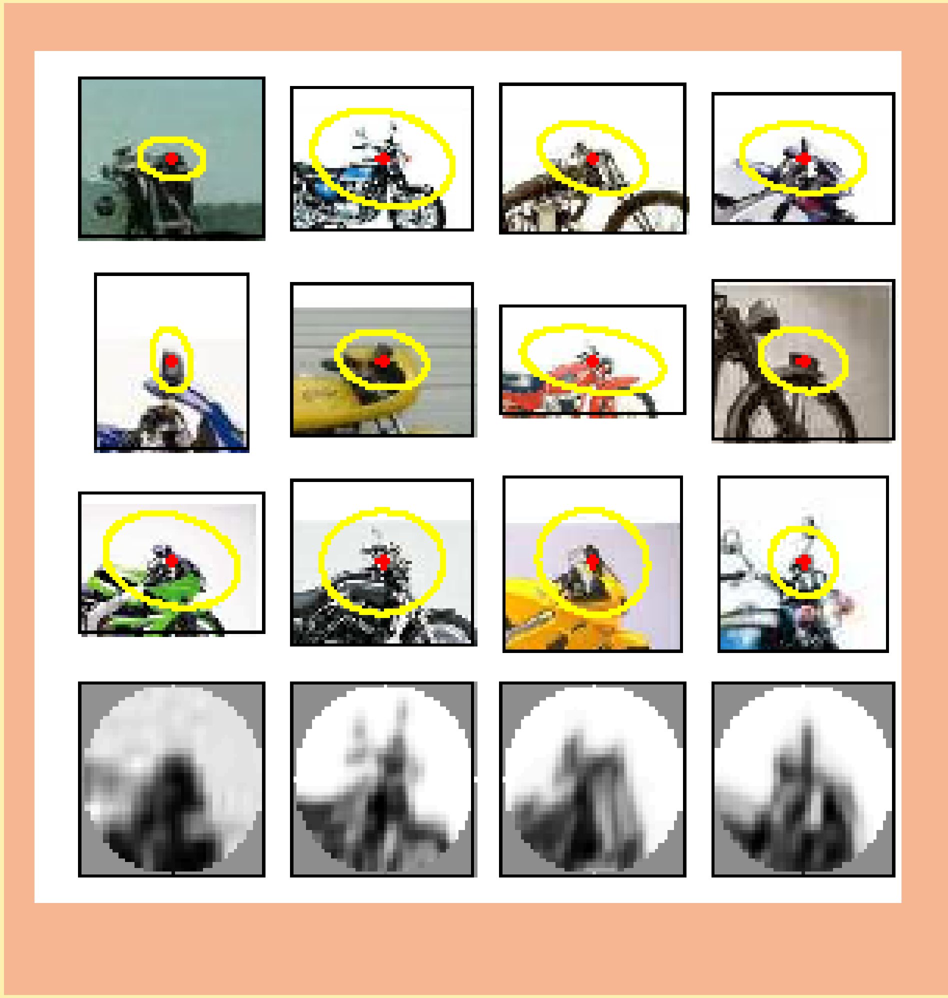
2. Codewords dictionary formation



Fei-Fei et al. 2005

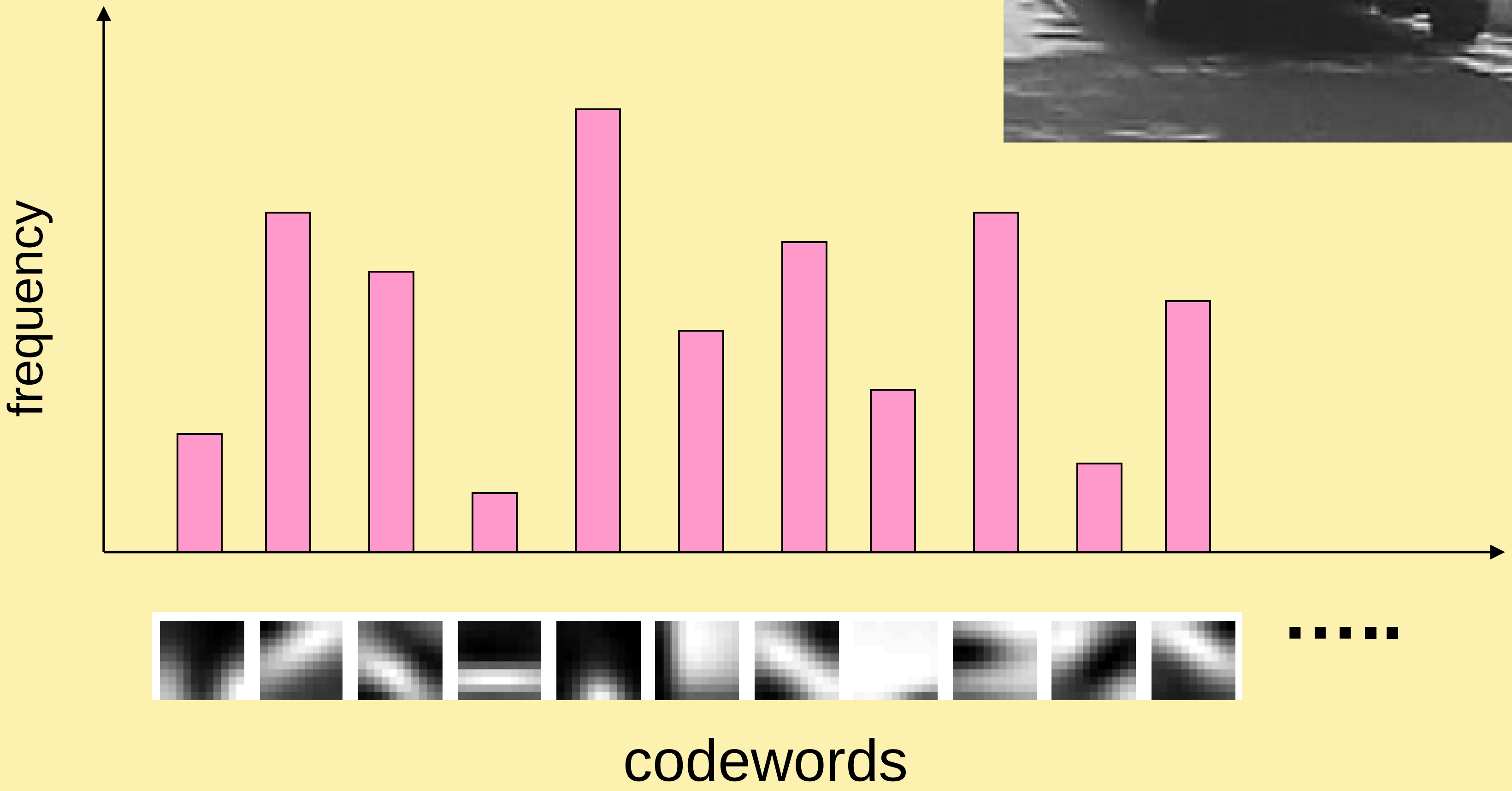
Image patch examples of codewords



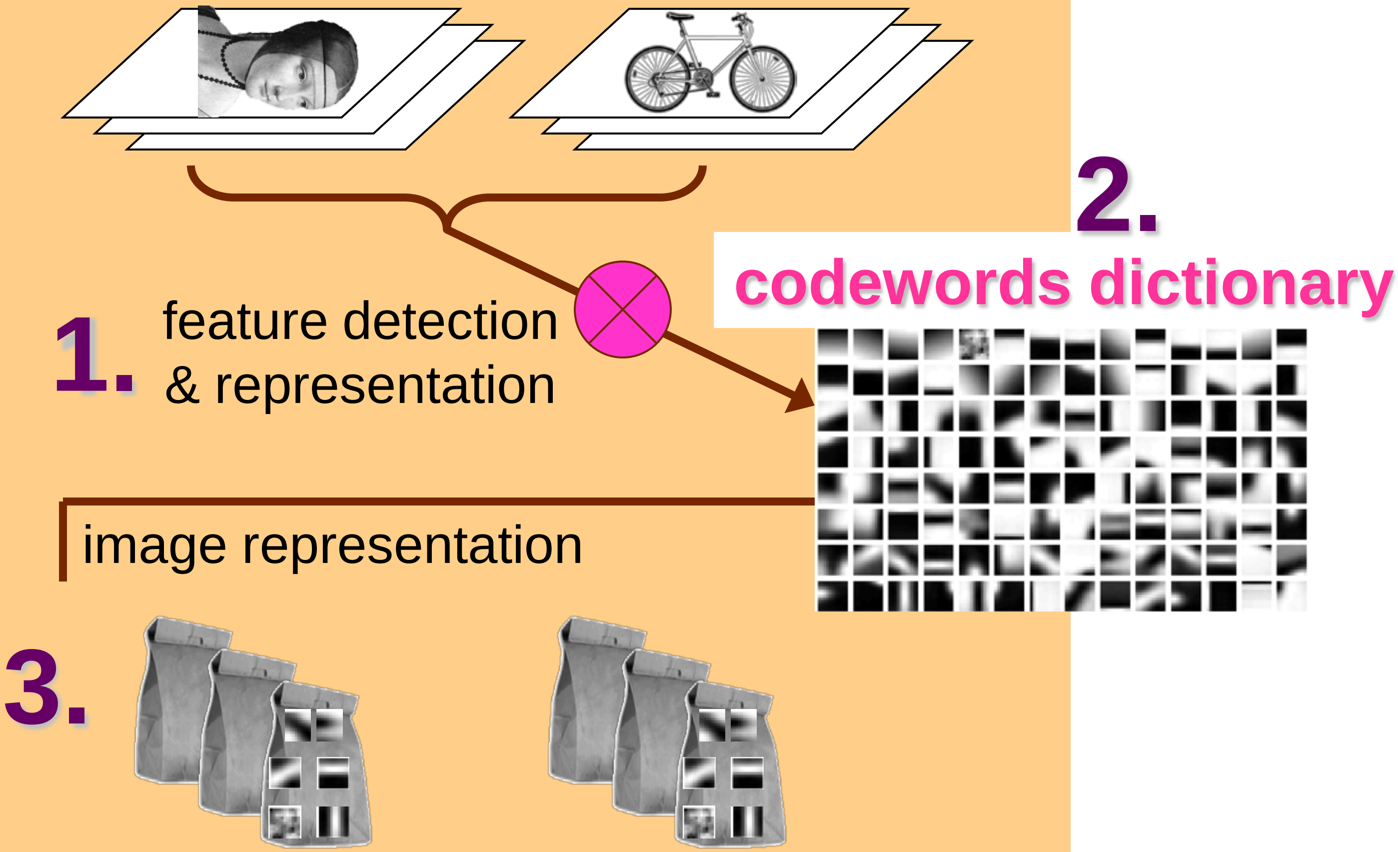


Sivic et al. 2005

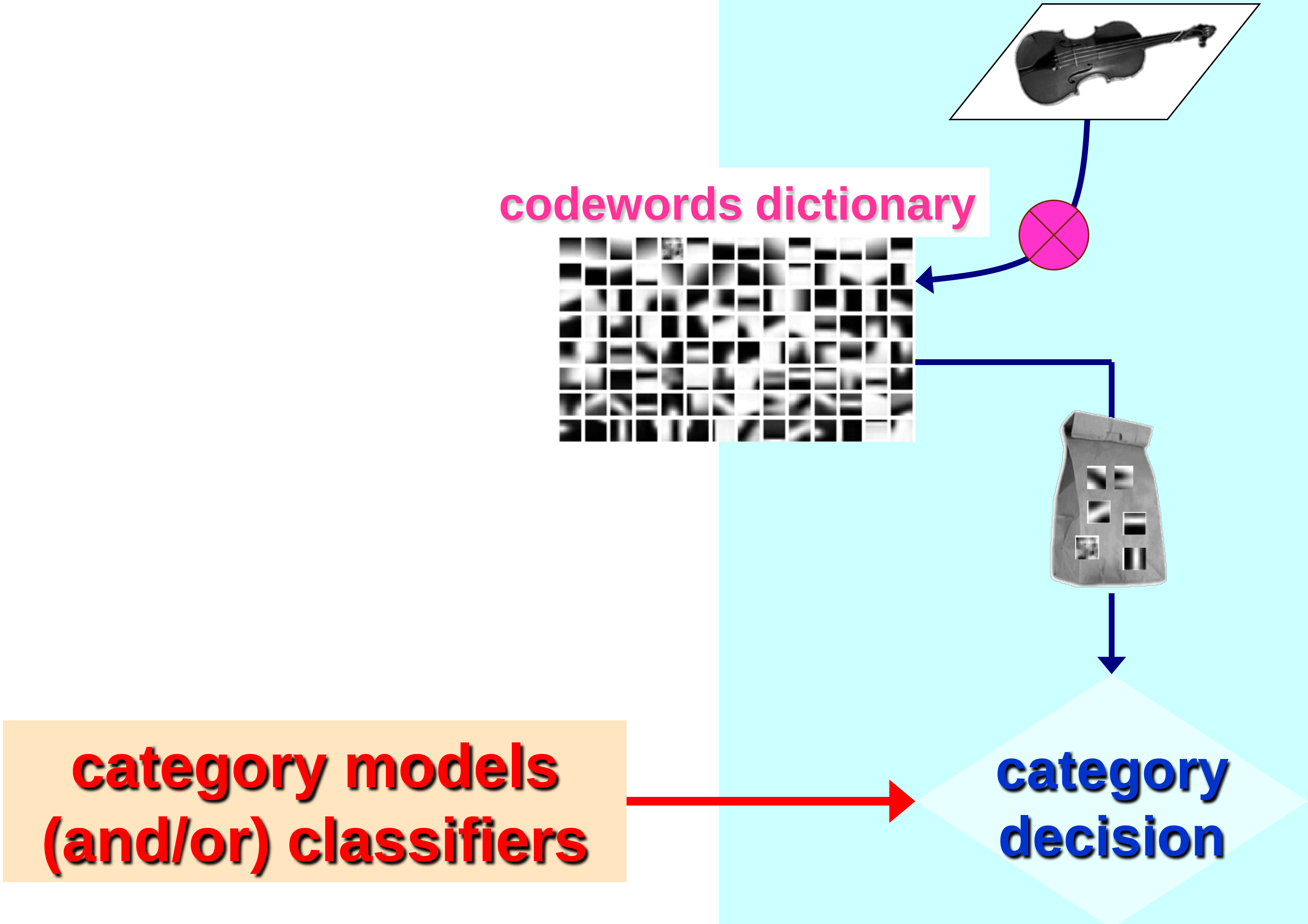
3. Image representation



Representation

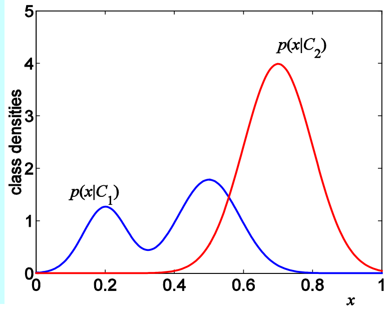


Learning and Recognition

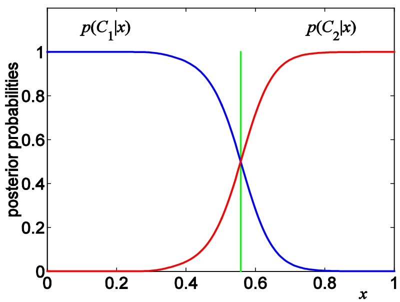


Learning and Recognition

1. Generative method:
   - graphical models



2. Discriminative method:
   - SVM



category models (and/or) classifiers

2 generative models

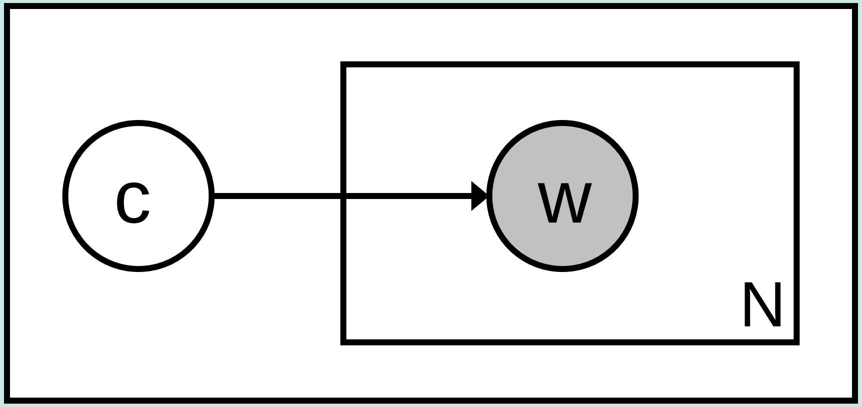
1. Naïve Bayes classifier
   – Csurka Bray, Dance & Fan, 2004

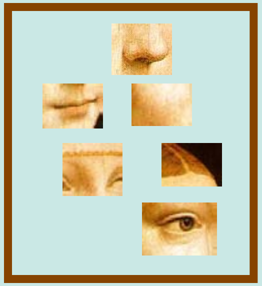
2. Hierarchical Bayesian text models (pLSA and LDA)
   – Background: Hoffman 2001, Blei, Ng & Jordan, 2004
   – Object categorization: Sivic et al. 2005, Sudderth et al. 2005
   – Natural scene categorization: Fei-Fei et al. 2005

First, some notations

- w_n: each patch in an image
  - $w_n = [0,0,\dots 1,\dots,0,0]^T$
- **w**: a collection of all N patches in an image
  - $\mathbf{w} = [w_1,w_2,\dots,w_N]$
- d_j: the j^{th} image in an image collection
- c: category of the image
- z: theme or topic of the patch

Case #1: the Naïve Bayes model





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

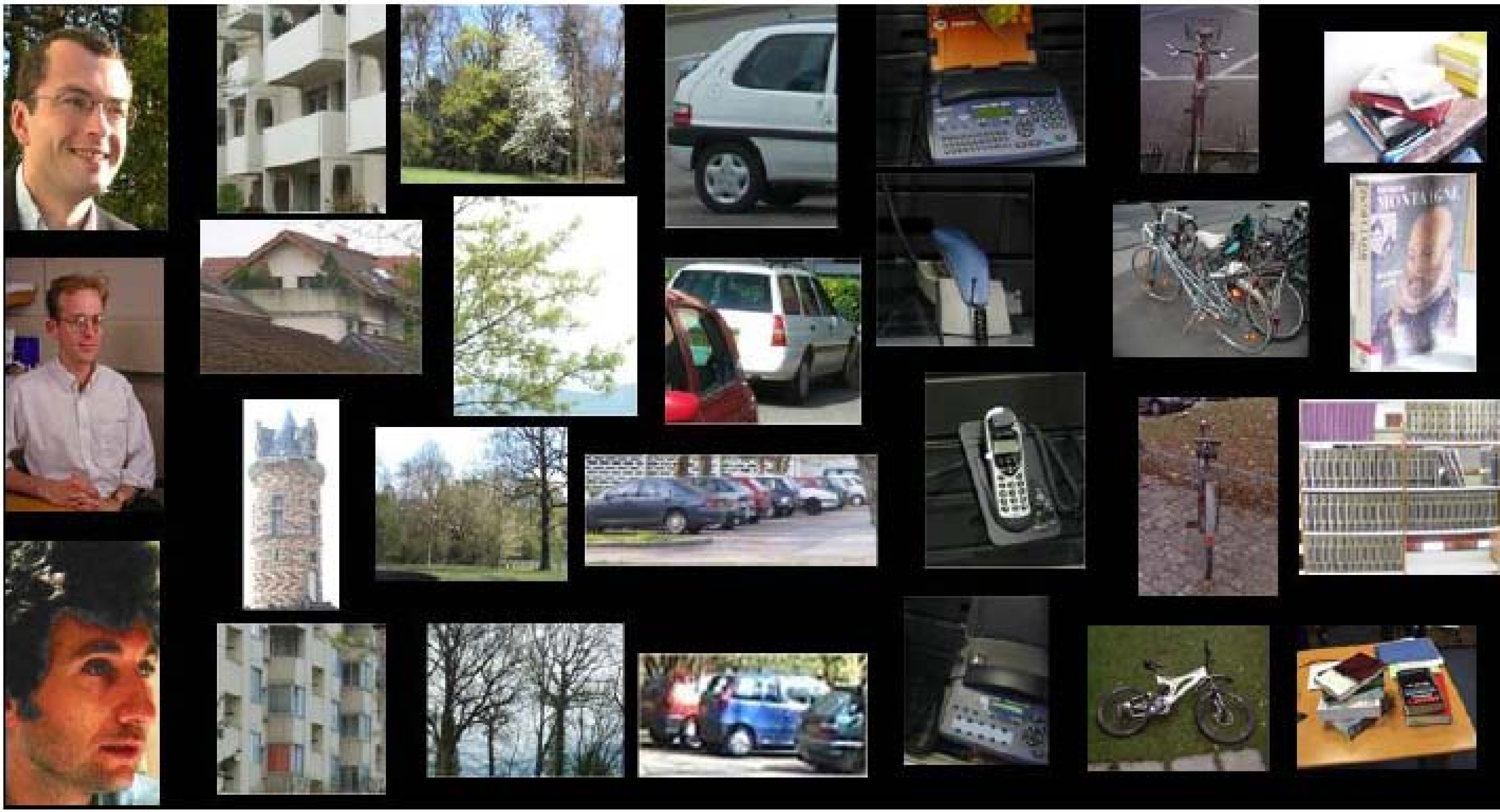
Object class decision

Prior prob. of the object classes

Image likelihood given the class

Csurka et al. 2004

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

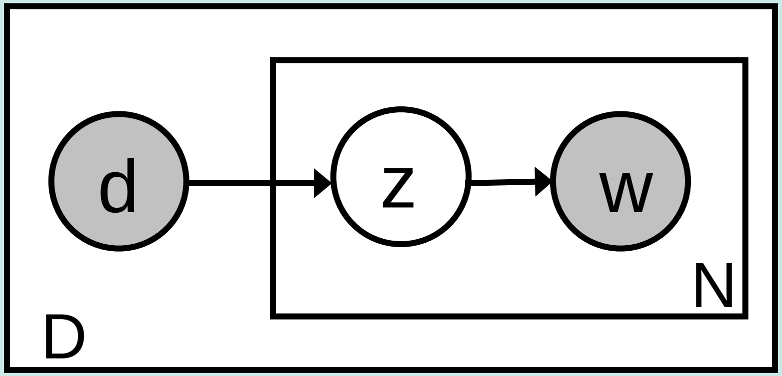
Table 1. Confusion matrix and the mean rank for the best vocabulary (*k=1000*).

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

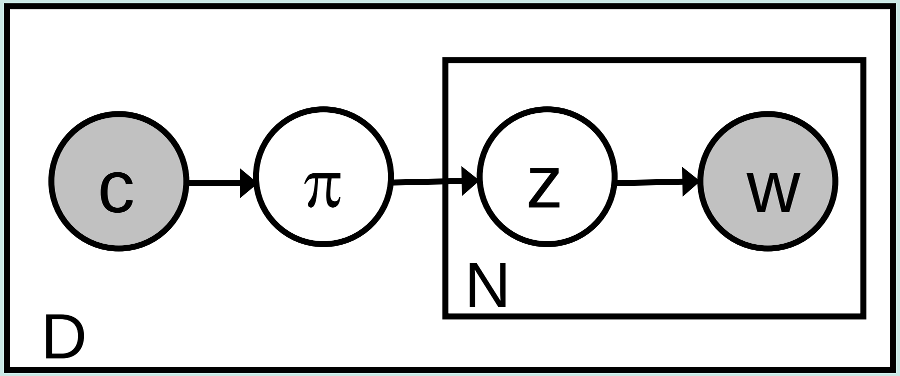
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 2001

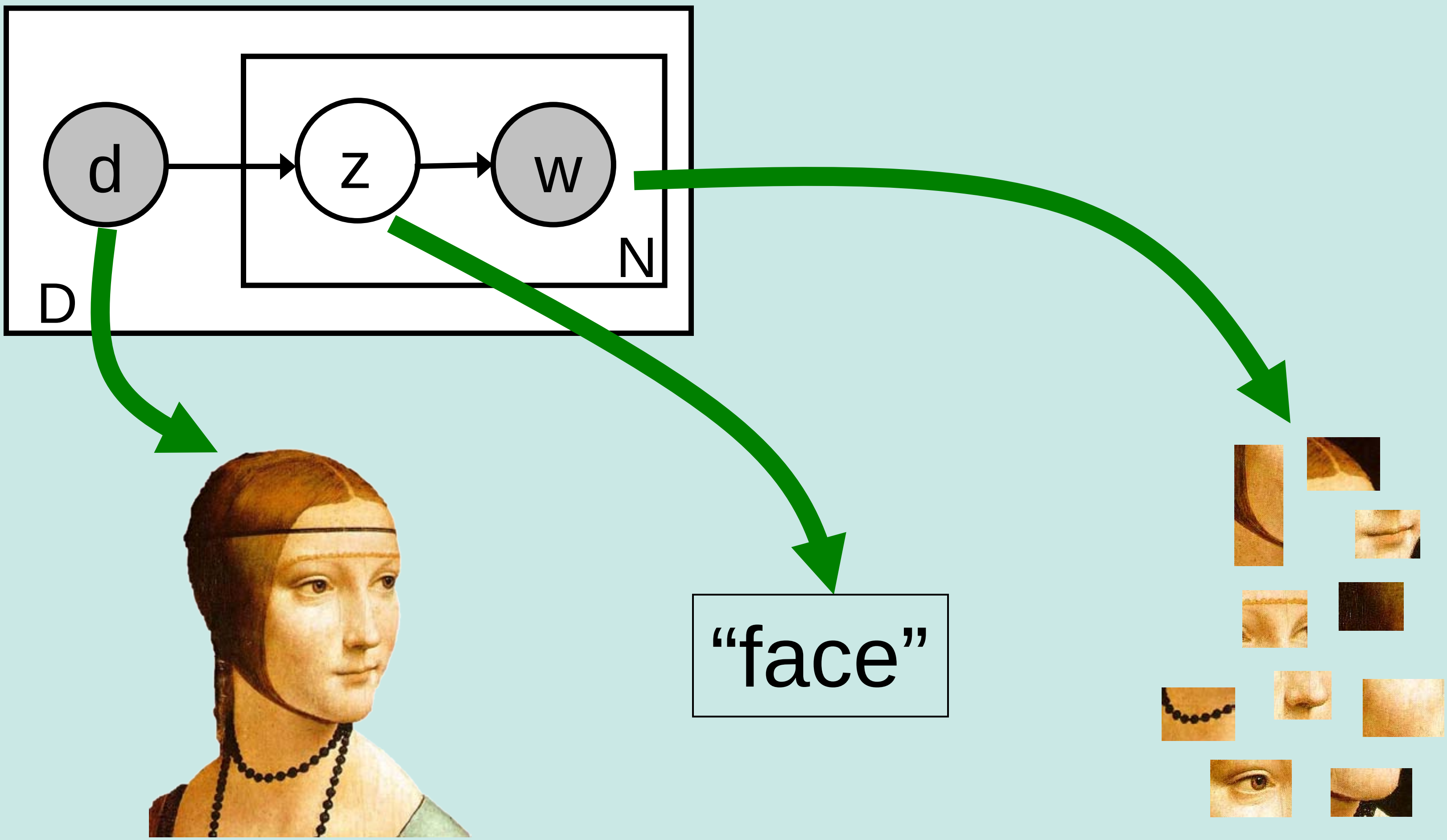
Latent Dirichlet Allocation (LDA)



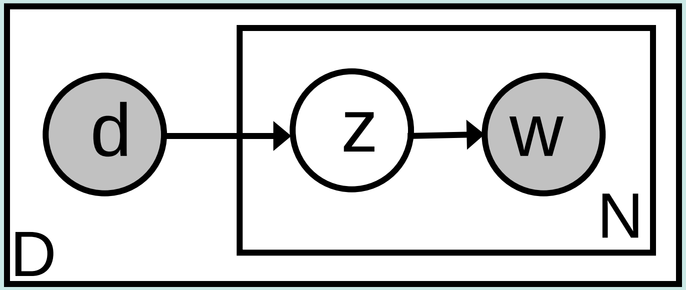
Blei et al., 2001

Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)

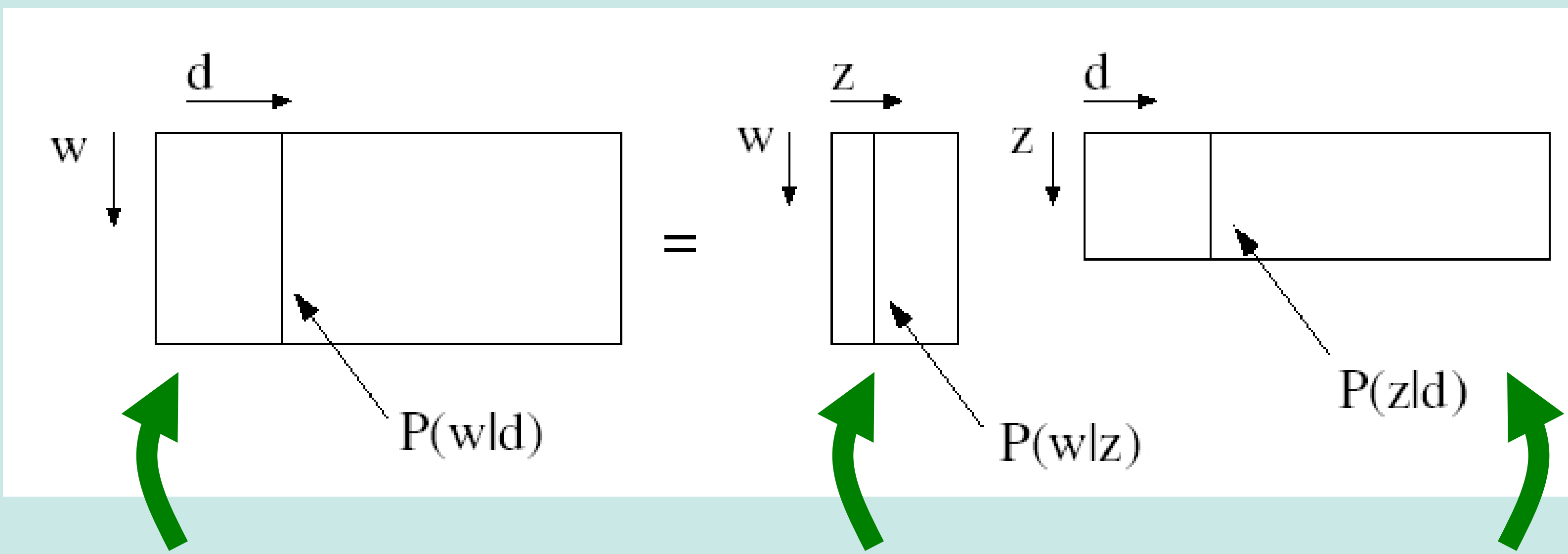


Sivic et al. ICCV 2005



Case #2: the pLSA model

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$



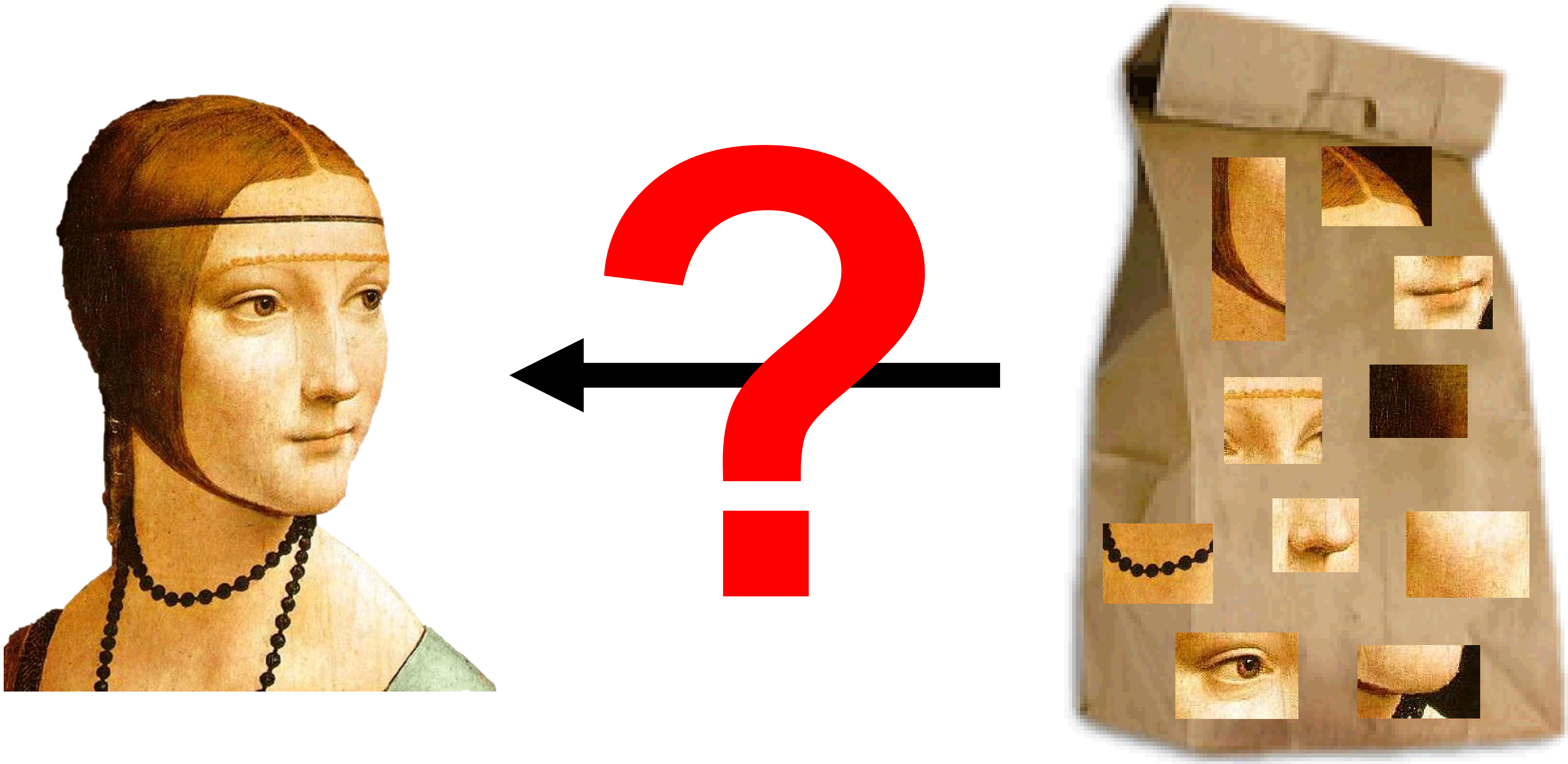
Observed codeword distributions

Codeword distributions per theme (topic)

Theme distributions per image

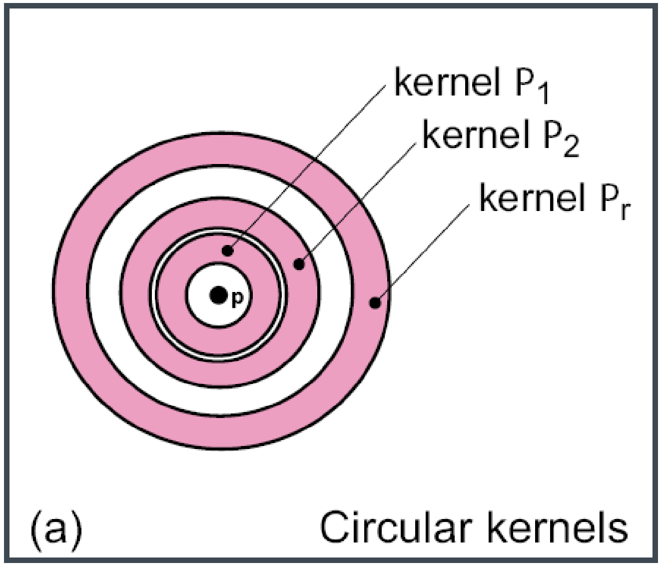
Slide credit: Josef Sivic

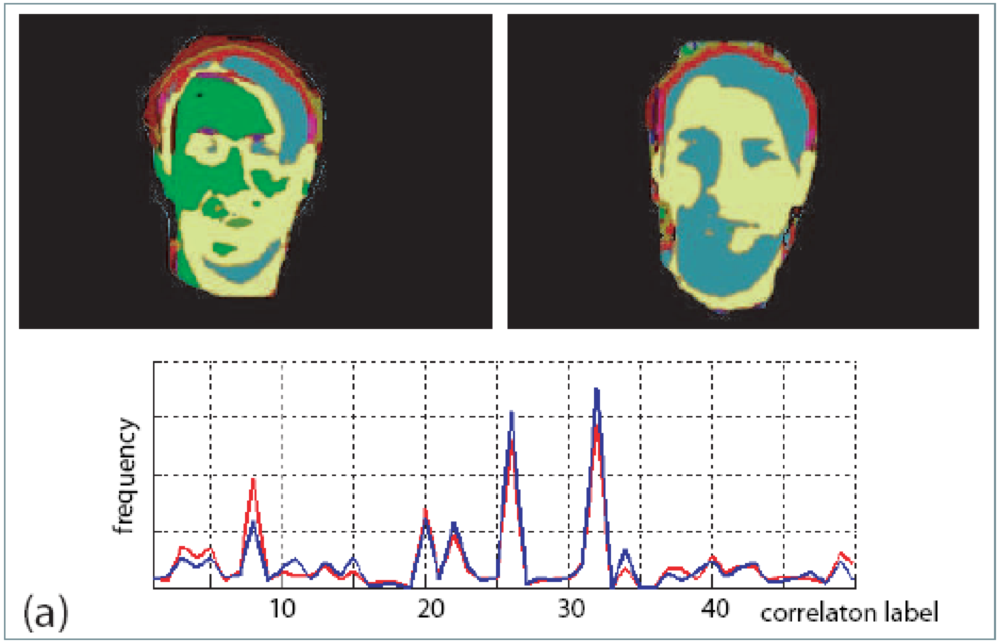
What about spatial info?



What about spatial info?

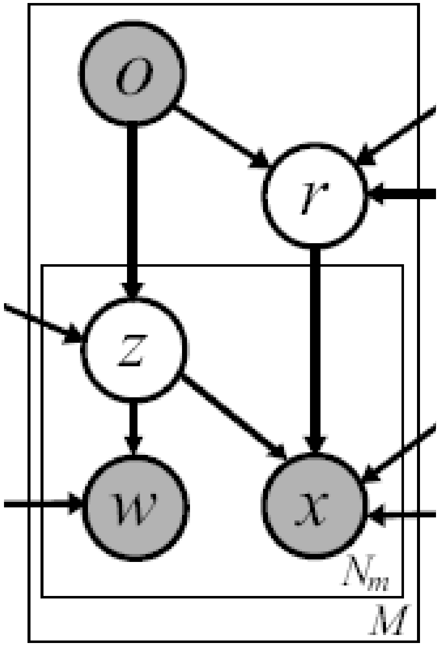
- Feature level
  - Spatial influence through correlogram features: Savarese, Winn and Criminisi, CVPR 2006

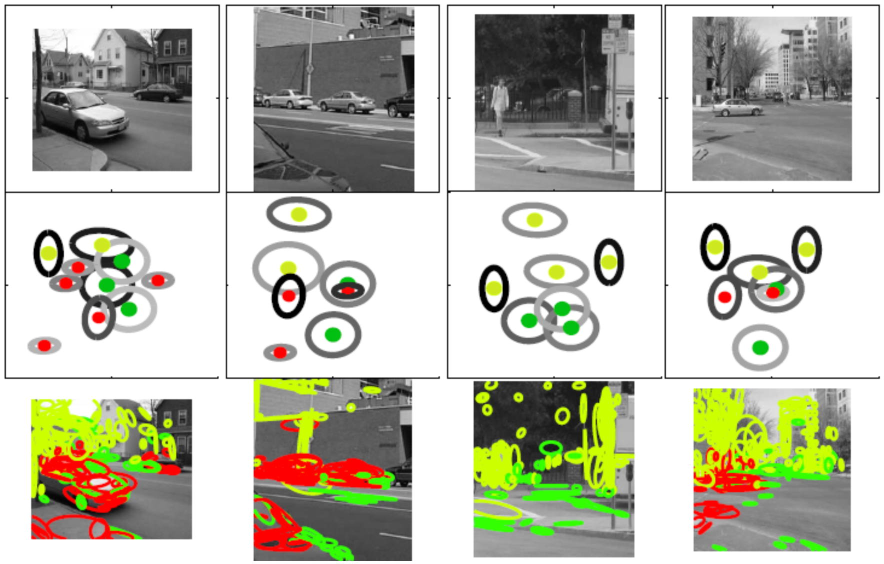




What about spatial info?

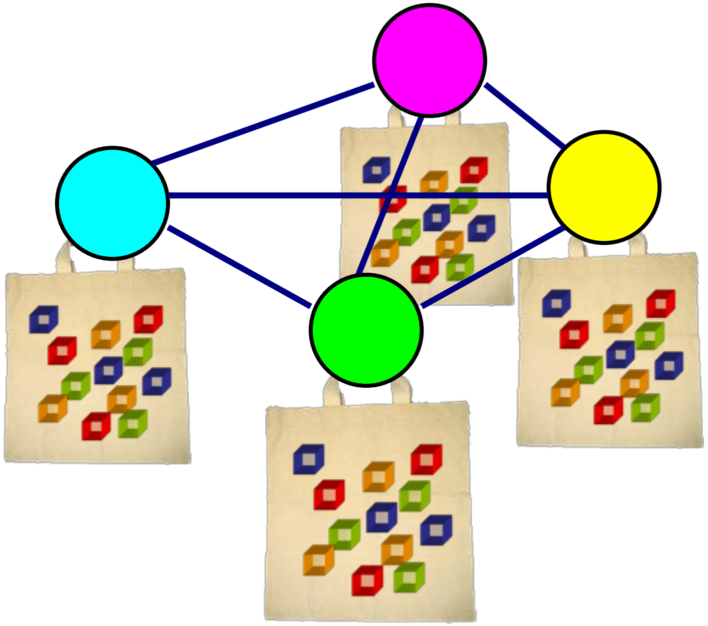
- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007

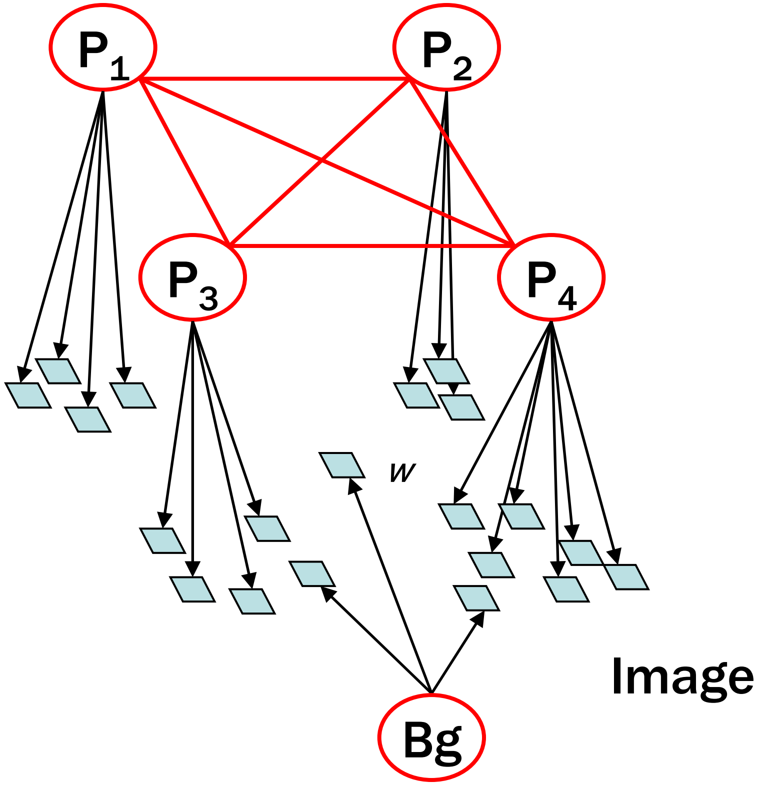




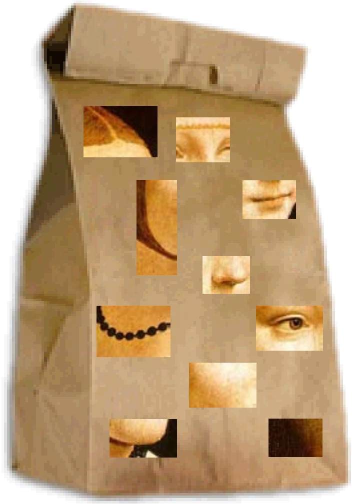
What about spatial info?

- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007

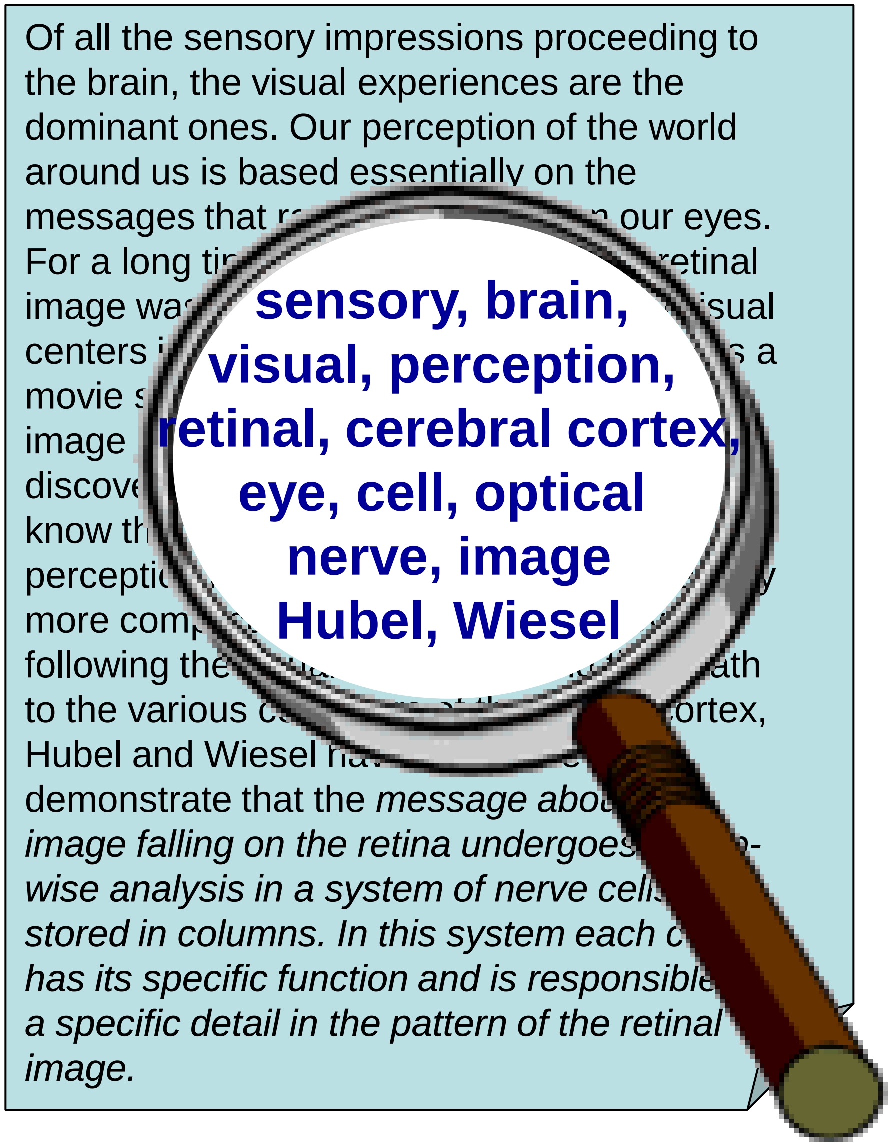




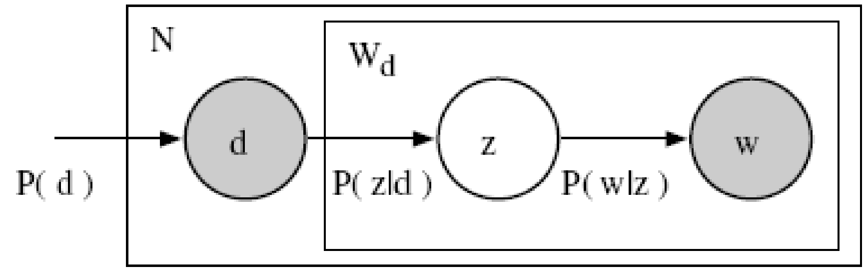
Model properties



- Intuitive
  - Analogy to documents

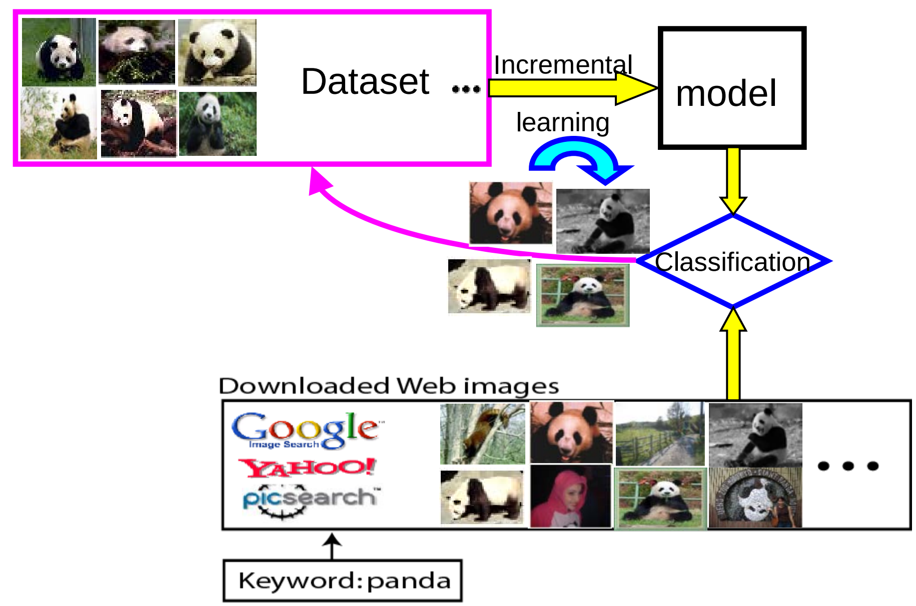


Model properties



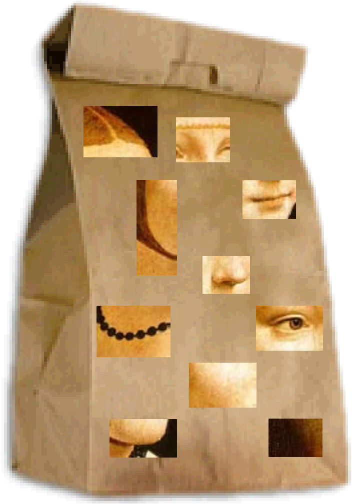
Sivic, Russell, Efros, Freeman, Zisserman, 2005

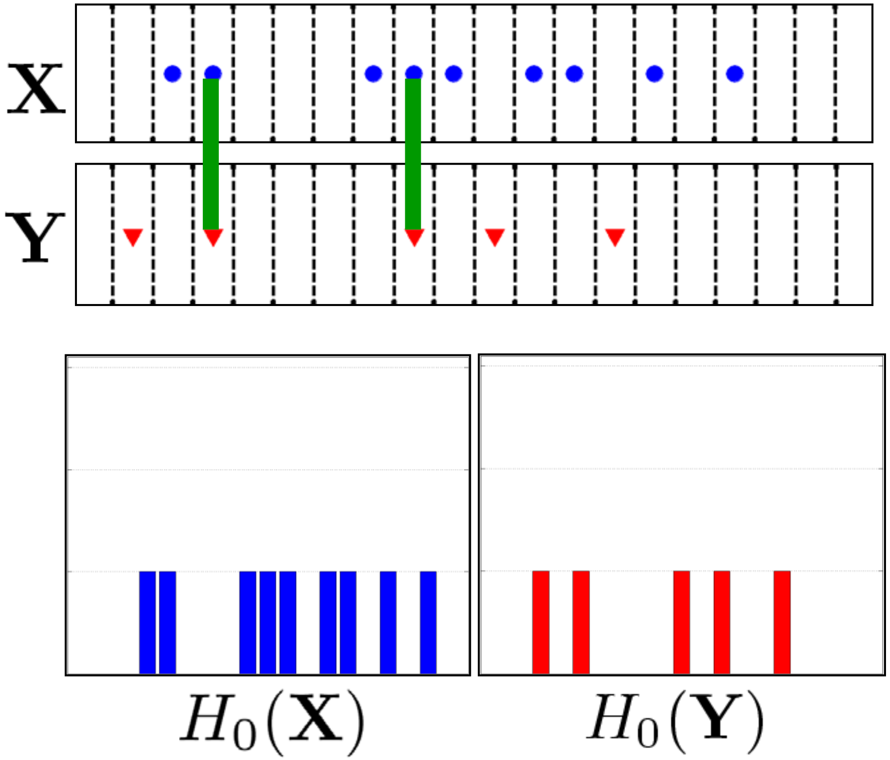
- Intuitive
- generative models
  - Convenient for weakly- or un-supervised, incremental training
  - Prior information
  - Flexibility (e.g. HDP)



Li, Wang & Fei-Fei, CVPR 2007

Model properties

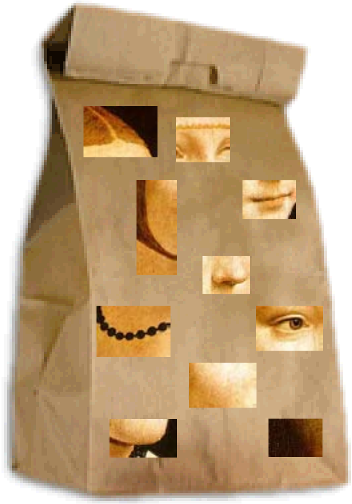




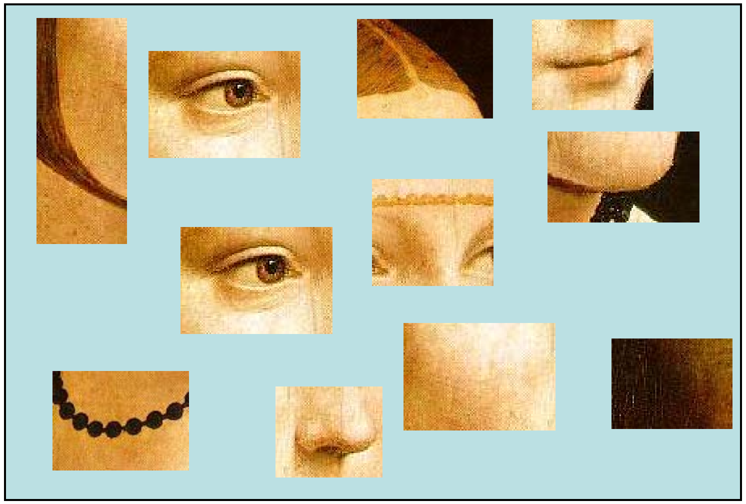
- Intuitive
- generative models
- Discriminative method
  - Computationally efficient

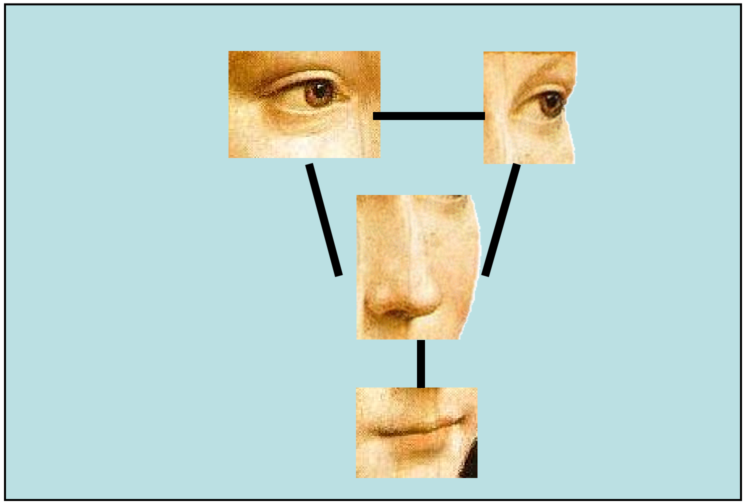
Grauman et al. CVPR 2005

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