



Part 1: Bag-of-words models

by Li Fei-Fei (Princeton)

Object



Bag of 'words'

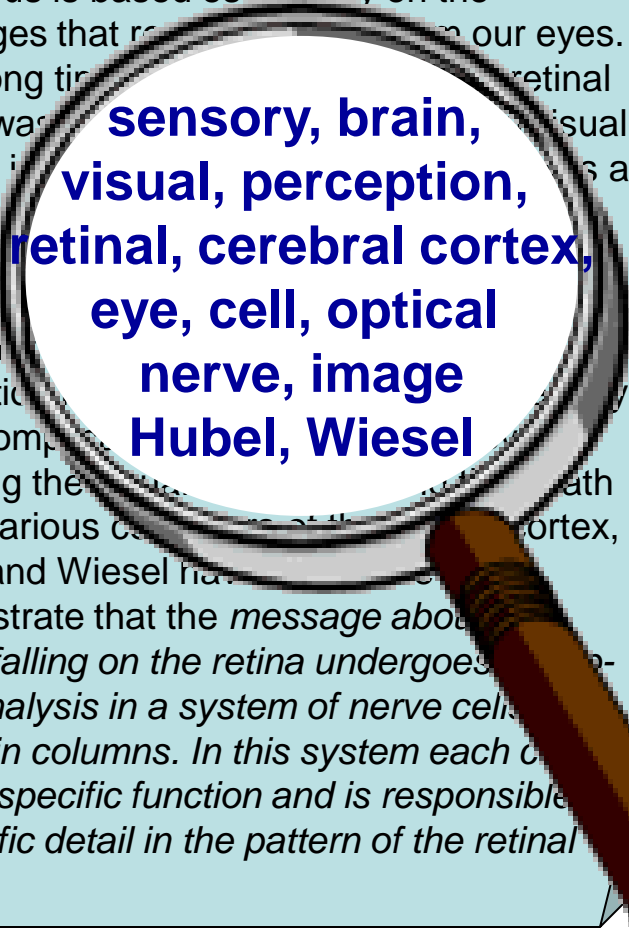


Analogy to documents

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 reach our eyes.

For a long time, the retinal image was considered as a movie screen. The image is discovered by the eye, and we know that perception is more complex than following the path to the various centers of the cortex, Hubel and Wiesel have demonstrated that the *message about the image falling on the retina undergoes a*

wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for 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% increase in exports to \$750bn, compared with \$560bn in 2004.

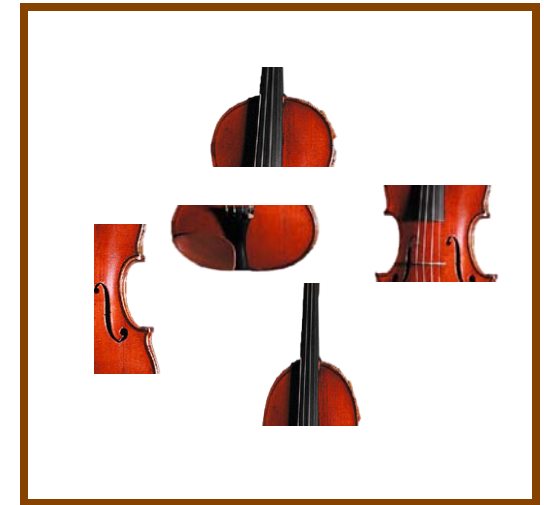
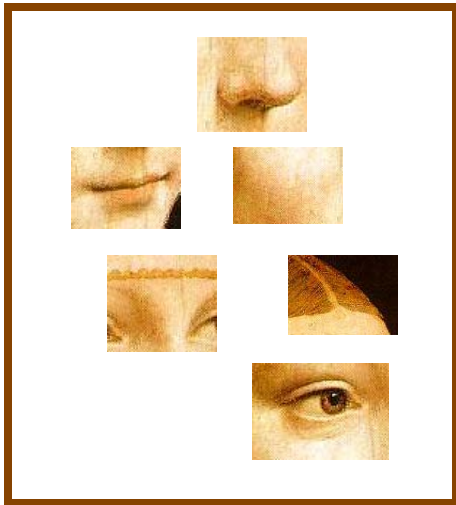
The increase will annoy the US because of China's deliberate policy to keep the yuan undervalued. The US government agrees that the yuan is undervalued and also needs to be allowed to rise. The demand so far has been for a free market country. China has not allowed the yuan against the dollar to rise and permitted it to trade within a narrow band but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.



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

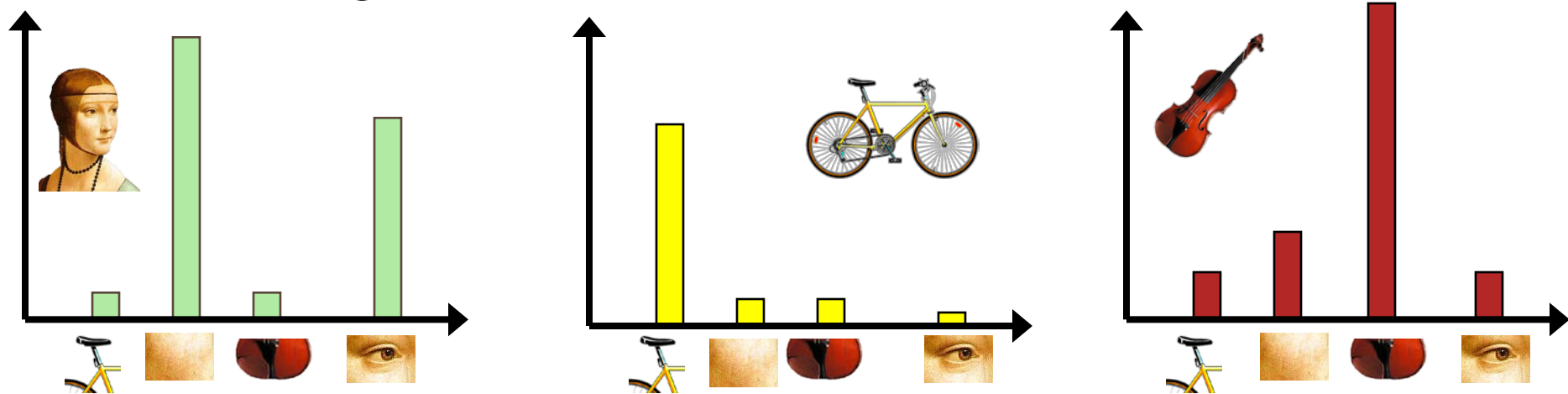
A clarification: definition of “BoW”

- Looser definition
 - Independent features

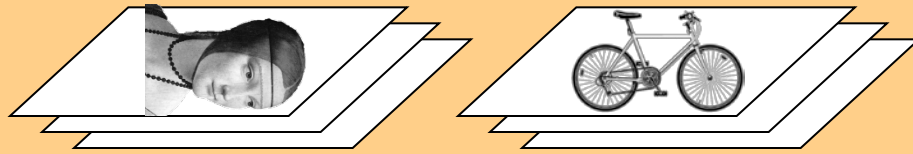


A clarification: definition of “BoW”

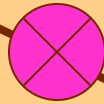
- Looser definition
 - Independent features
- Stricter definition
 - Independent features
 - histogram representation



learning



feature detection
& representation



codewords dictionary

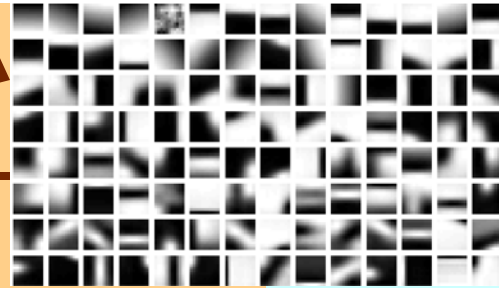
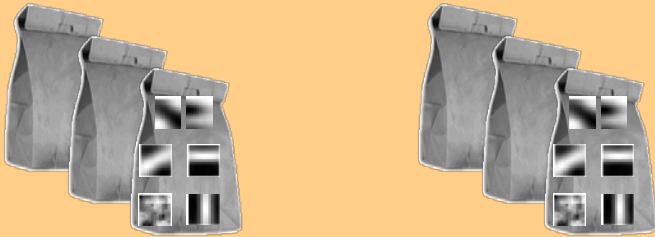
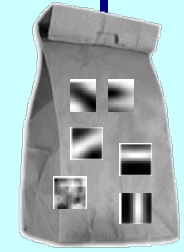
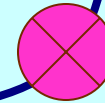


image representation



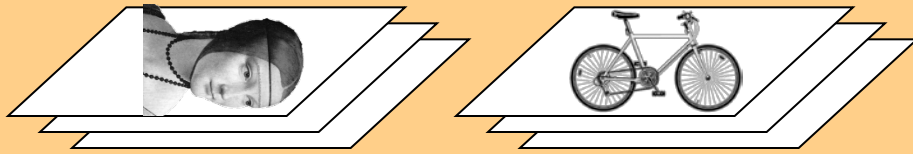
**category models
(and/or) classifiers**

recognition

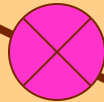


**category
decision**

Representation



1. feature detection
& representation



2. codewords dictionary

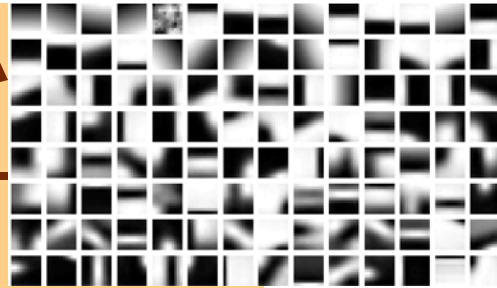
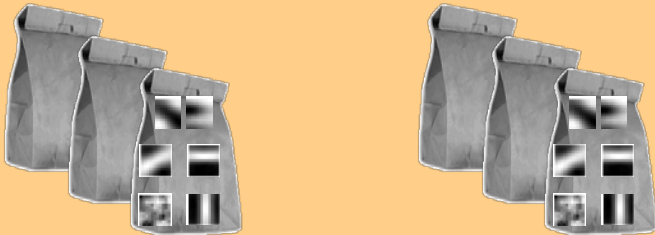
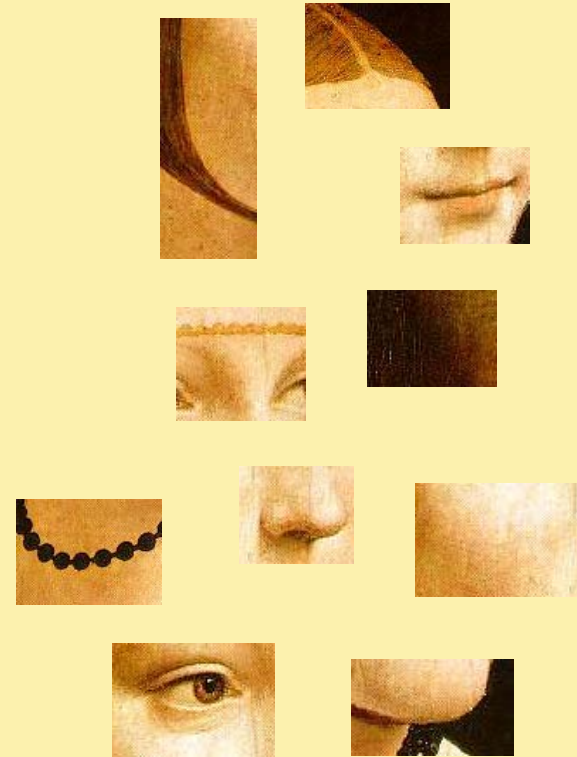


image representation

3.



1. Feature detection and representation



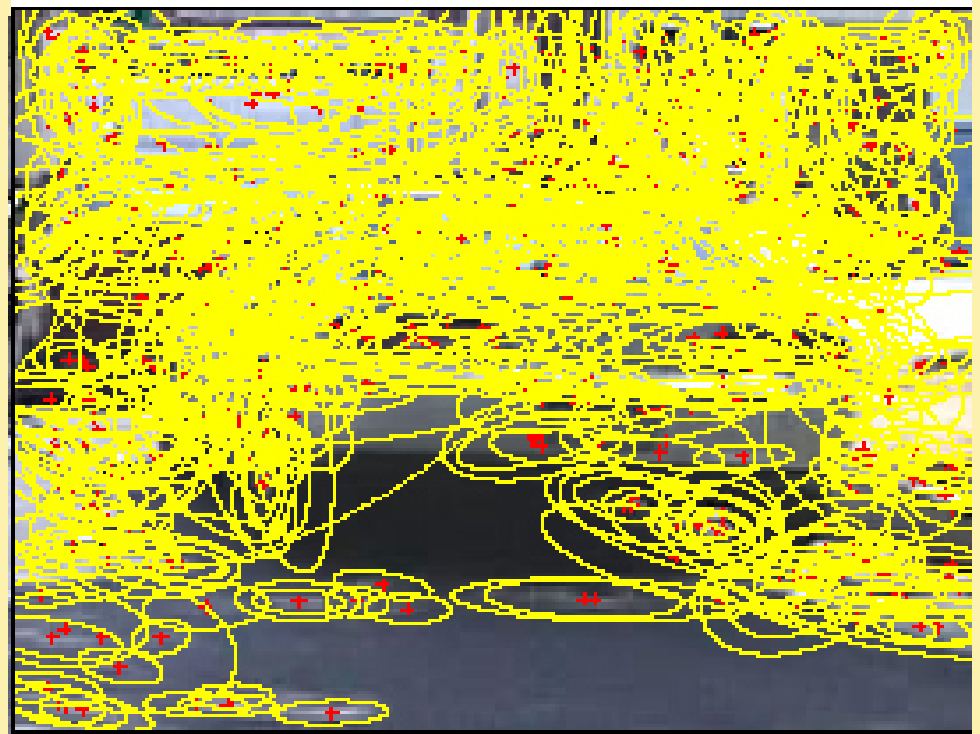
1. Feature detection and representation

- 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



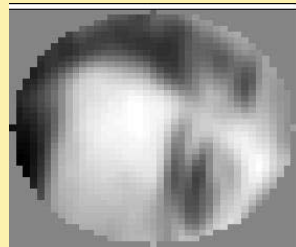
1. Feature detection and representation

- 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)

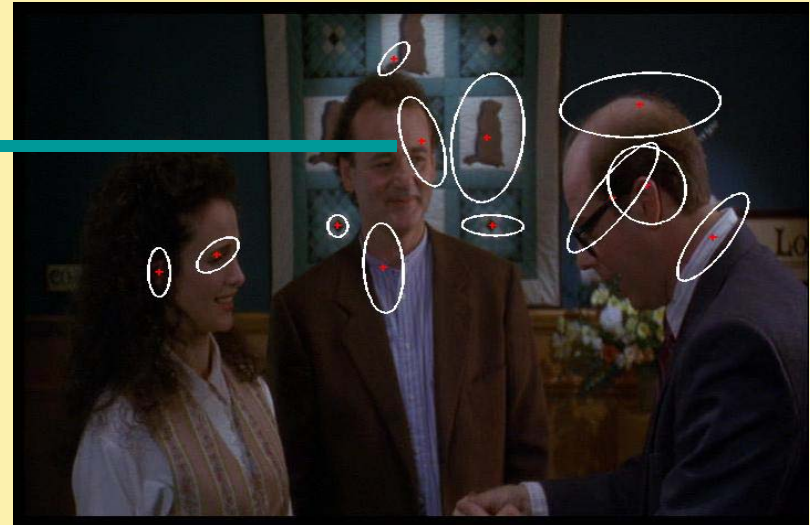
1. Feature detection and representation



[Lowe'99]



**Normalize
patch**



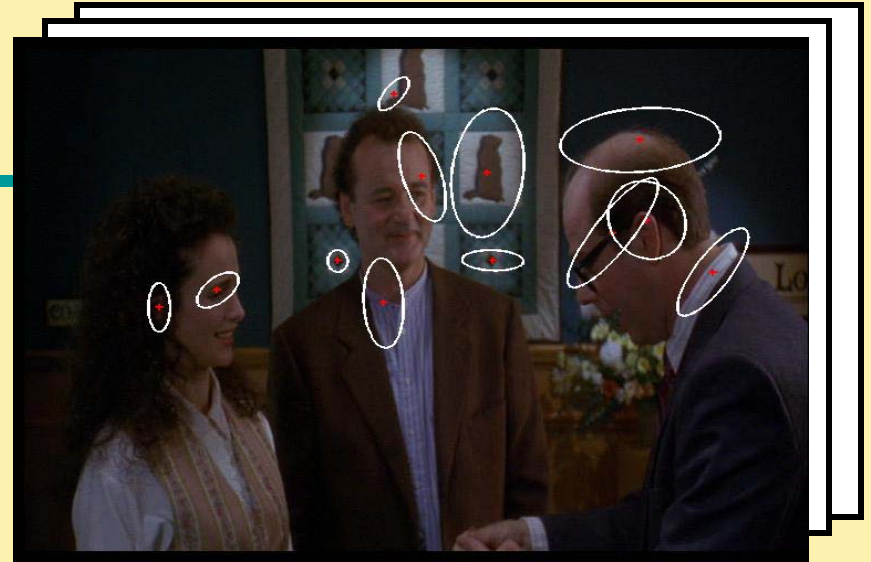
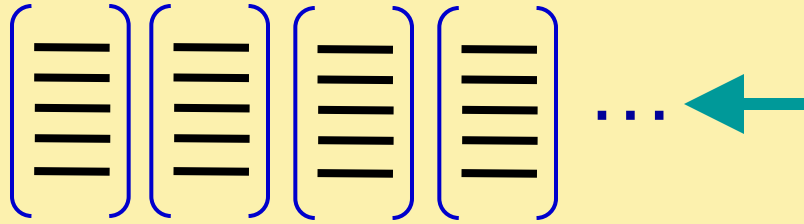
Detect patches

[Mikojczyk and Schmid '02]

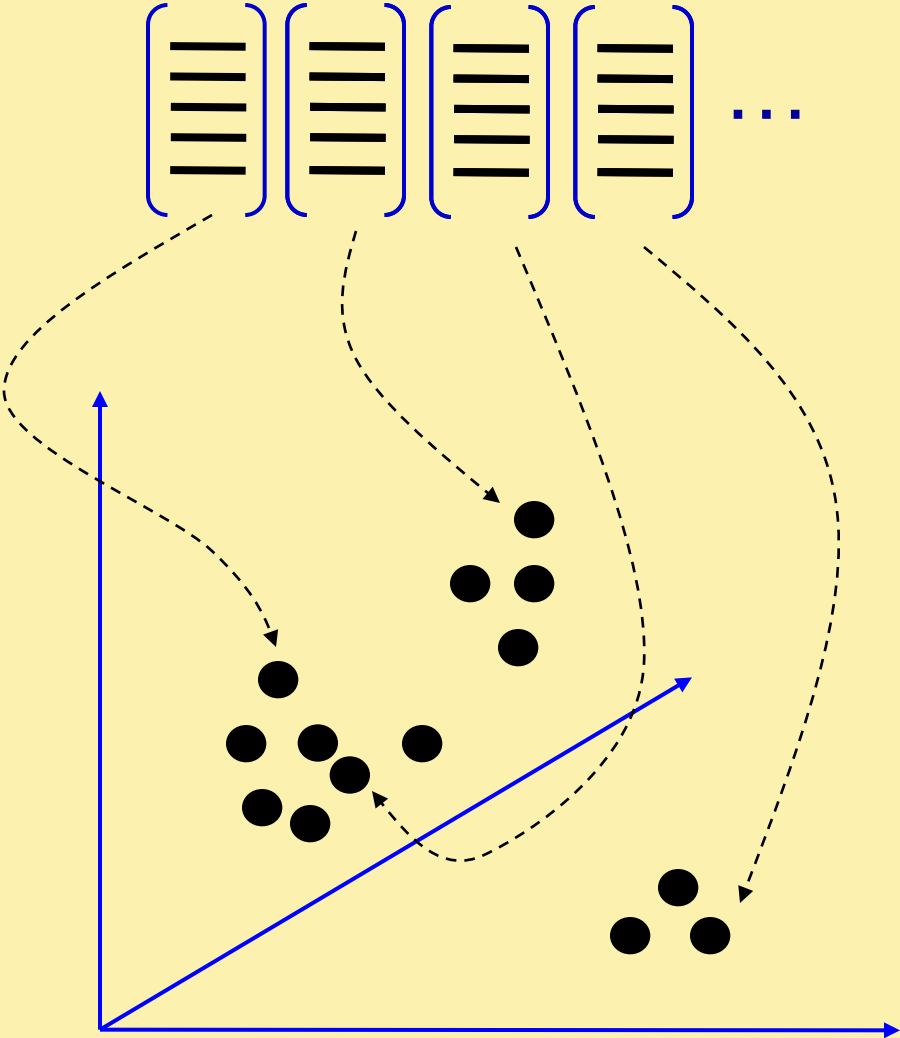
[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

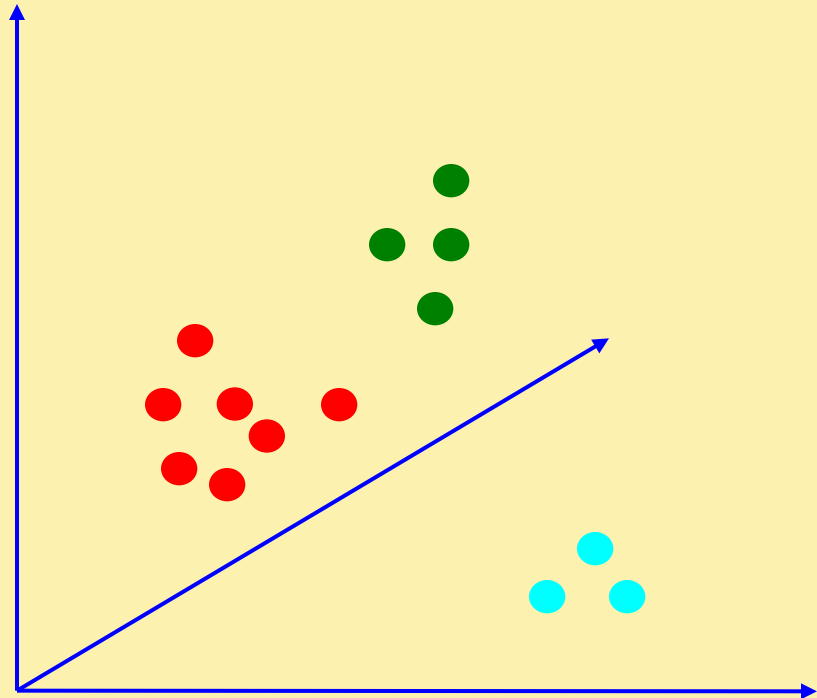
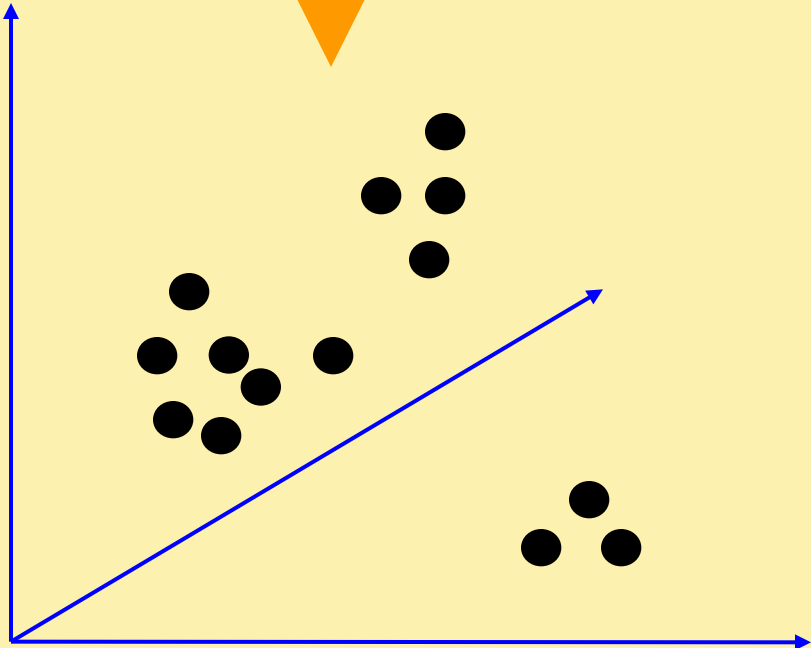
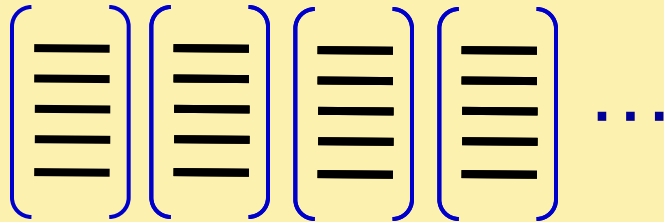
1. Feature detection and representation



2. Codewords dictionary formation



2. Codewords dictionary formation



Vector quantization

2. Codewords dictionary formation

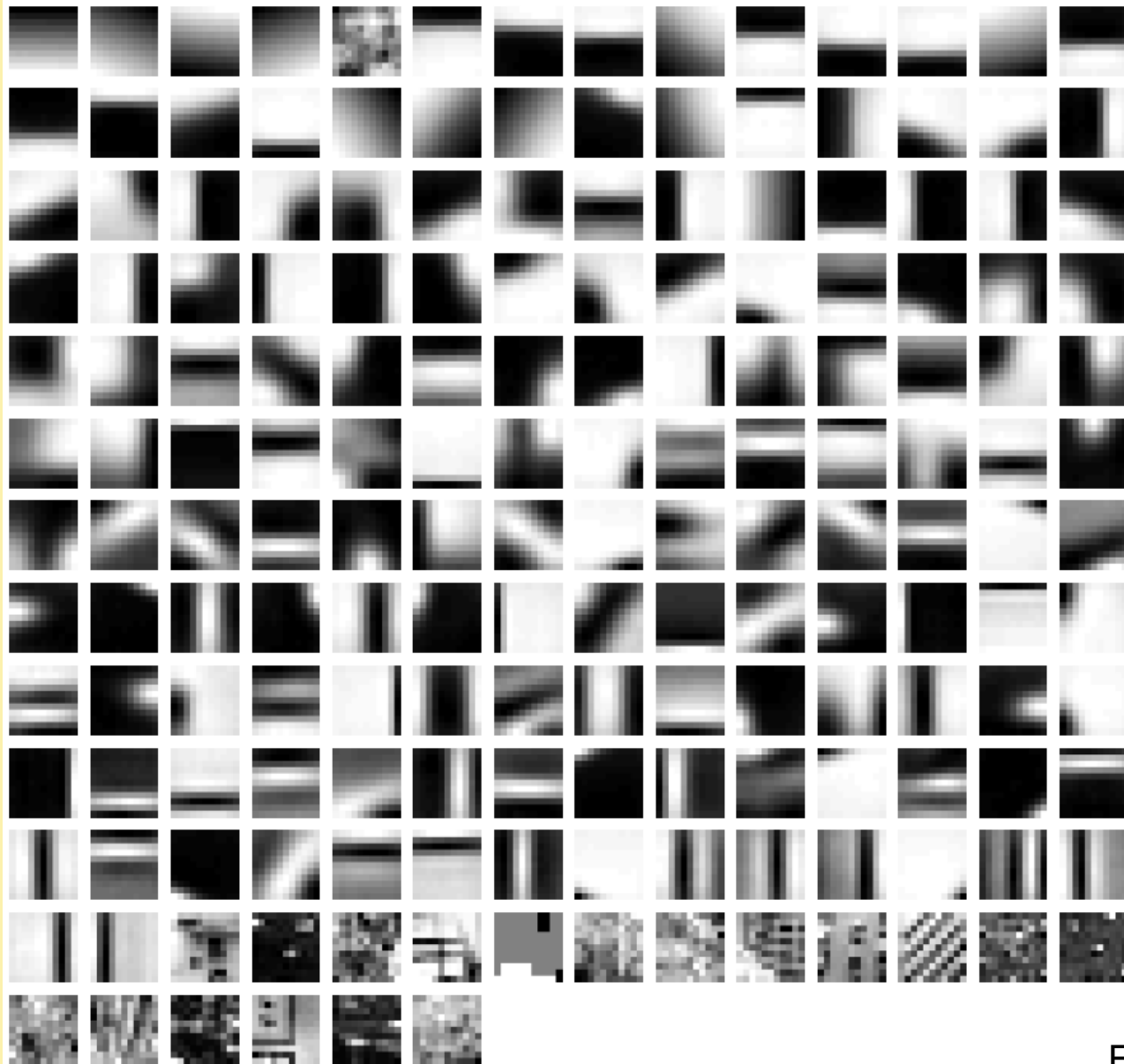
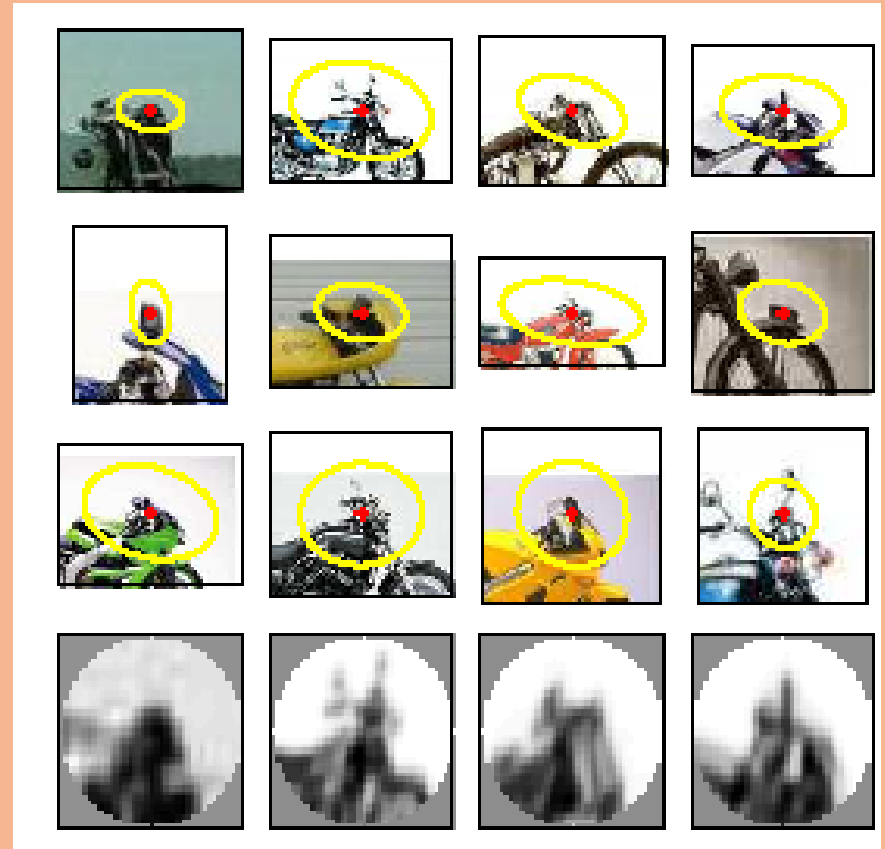
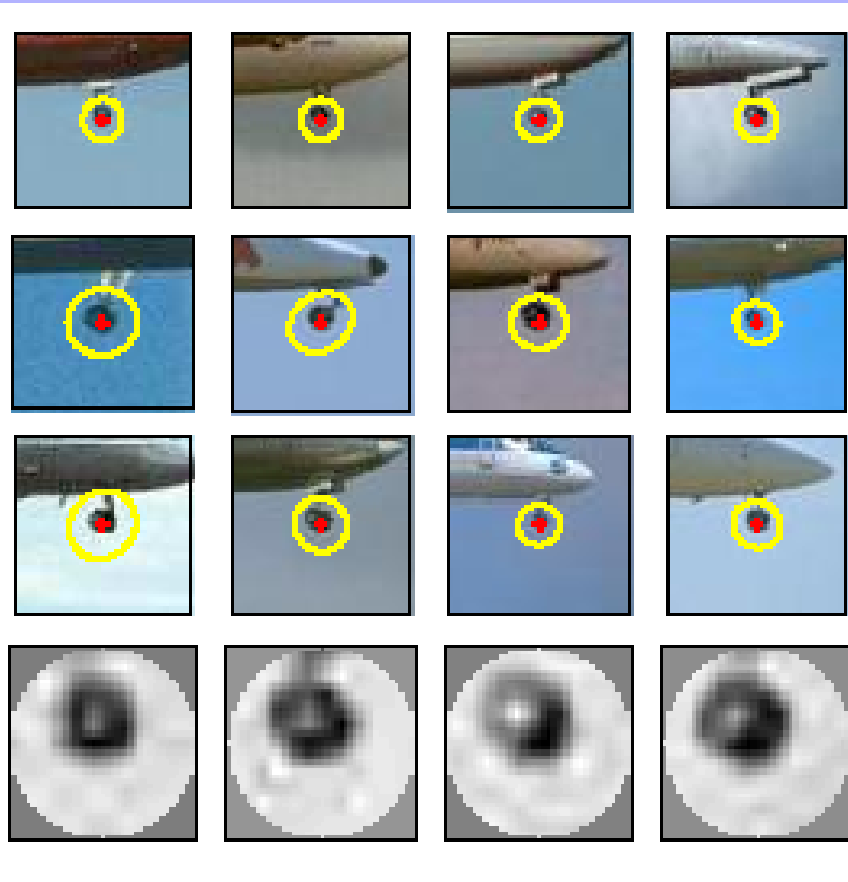


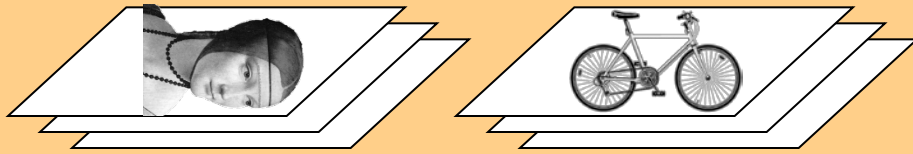
Image patch examples of codewords



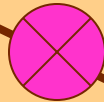
3. Image representation



Representation



1. feature detection
& representation



2. **codewords dictionary**

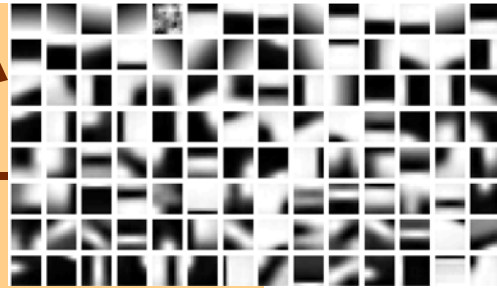
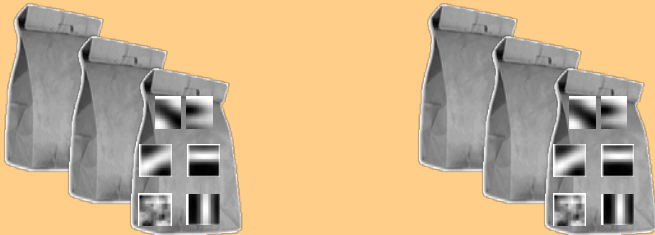
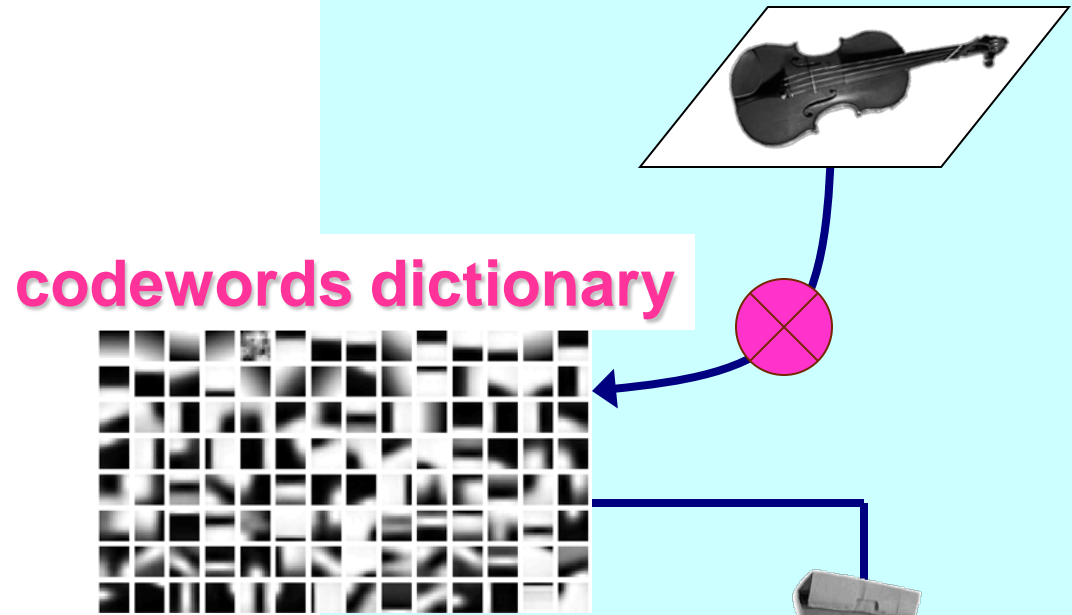


image representation

3.

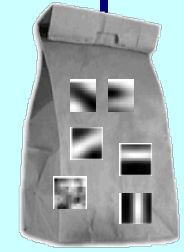


Learning and Recognition



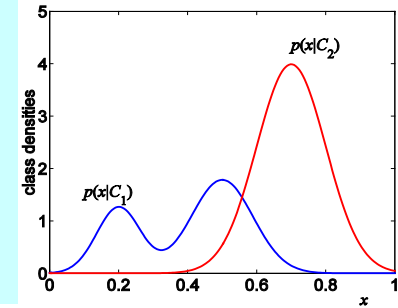
**category models
(and/or) classifiers**

**category
decision**

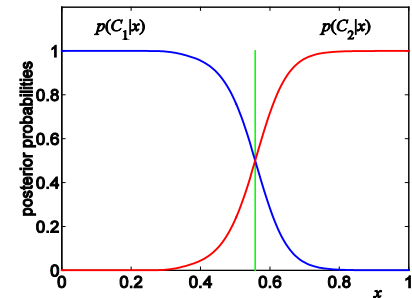


Learning and Recognition

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

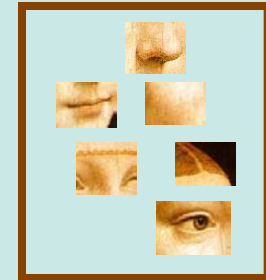
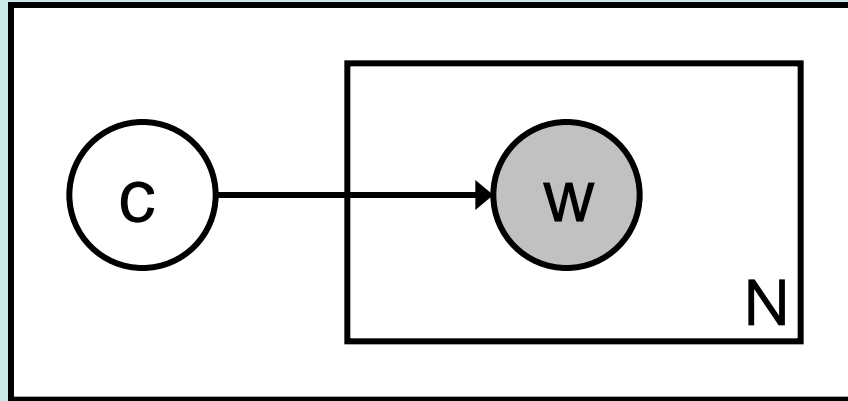
2. Hierarchical Bayesian text models (pLSA and LDA)

- 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
 - $w_n = [0, 0, \dots, 1, \dots, 0, 0]^T$
- \mathbf{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



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

Object class
decision

Prior prob. of
the object classes

Image likelihood
given the class

Our in-house database contains 1776 images in seven classes¹: faces, buildings, trees, cars, phones, bikes and books. Fig. 2 shows some examples from this dataset.

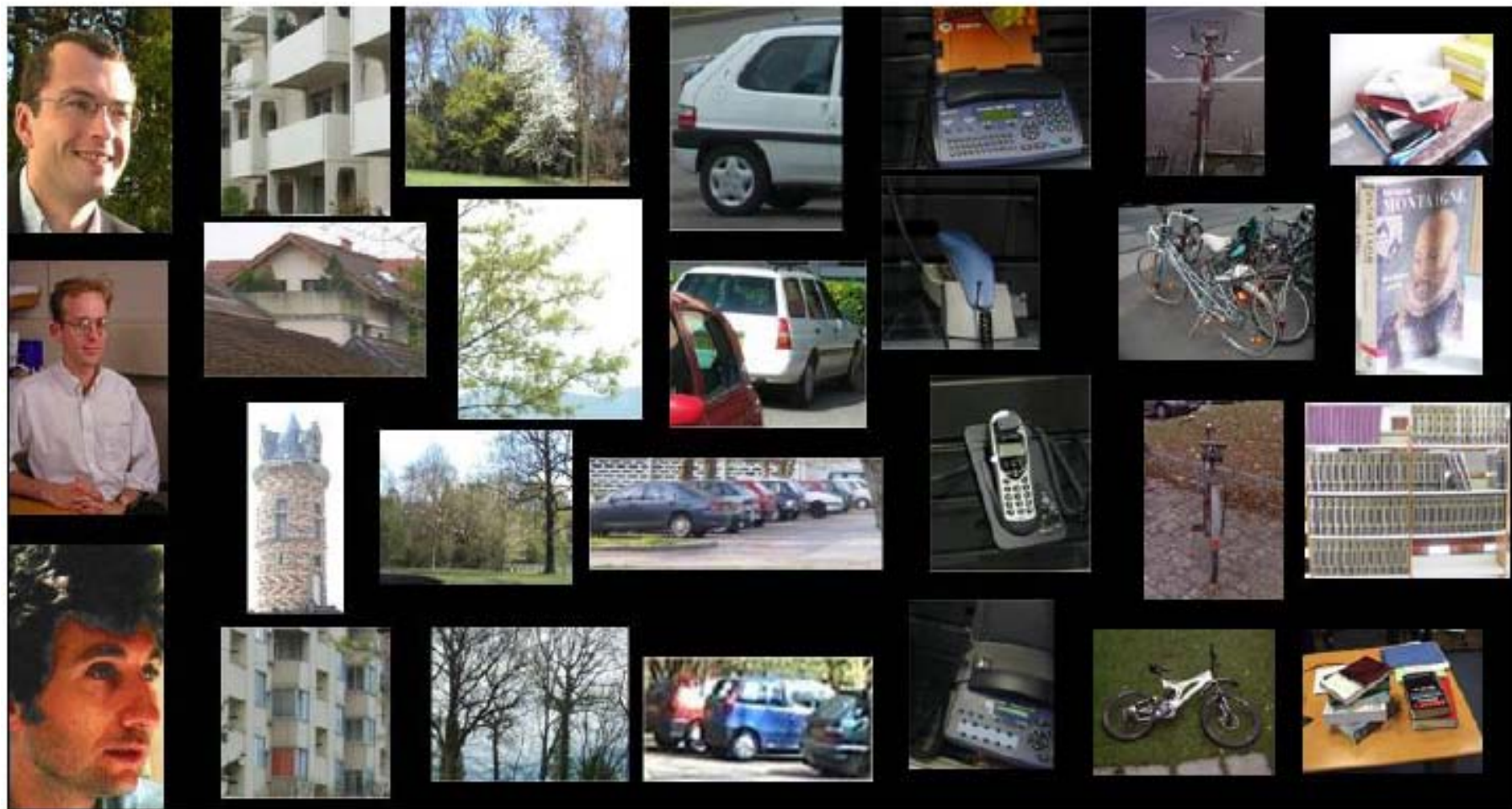
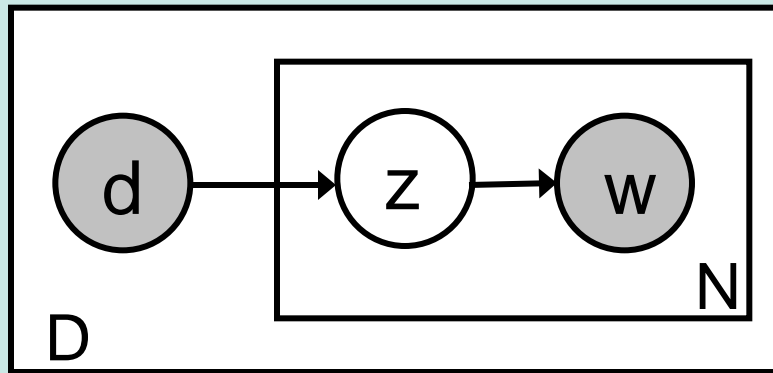


Table 1. Confusion matrix and the mean rank for the best vocabulary ($k=1000$).

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

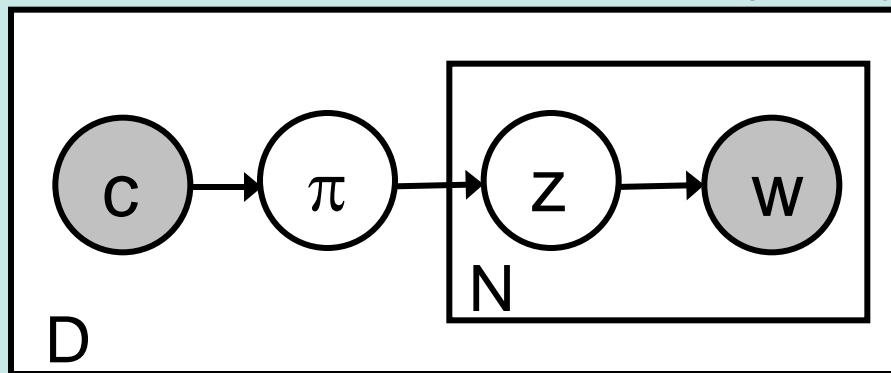
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 2001

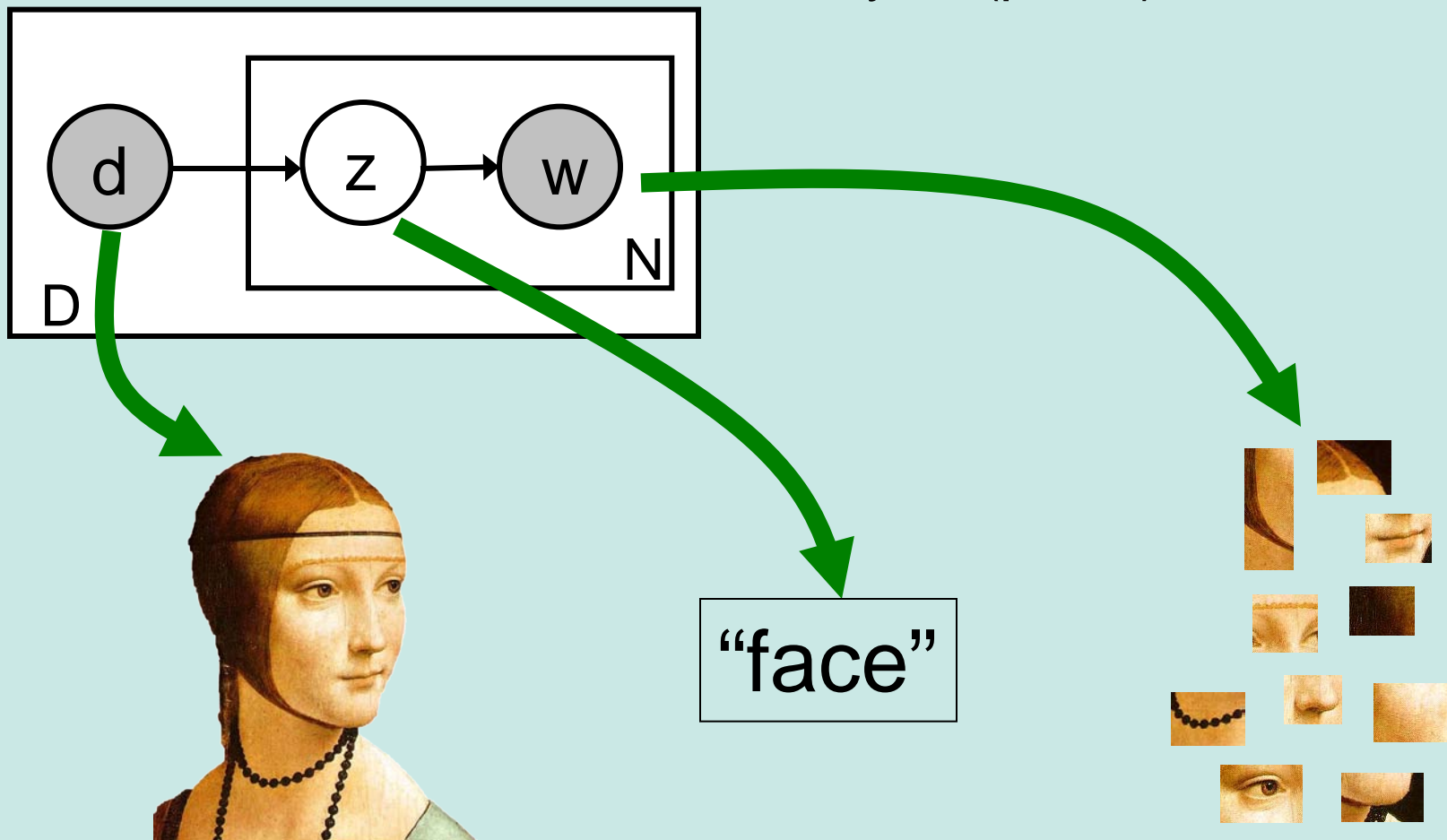
Latent Dirichlet Allocation (LDA)



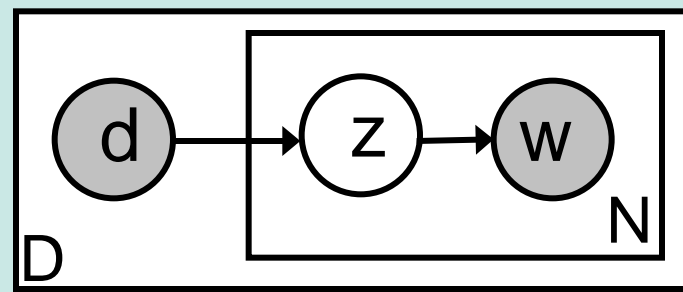
Blei et al., 2001

Case #2: Hierarchical Bayesian text models

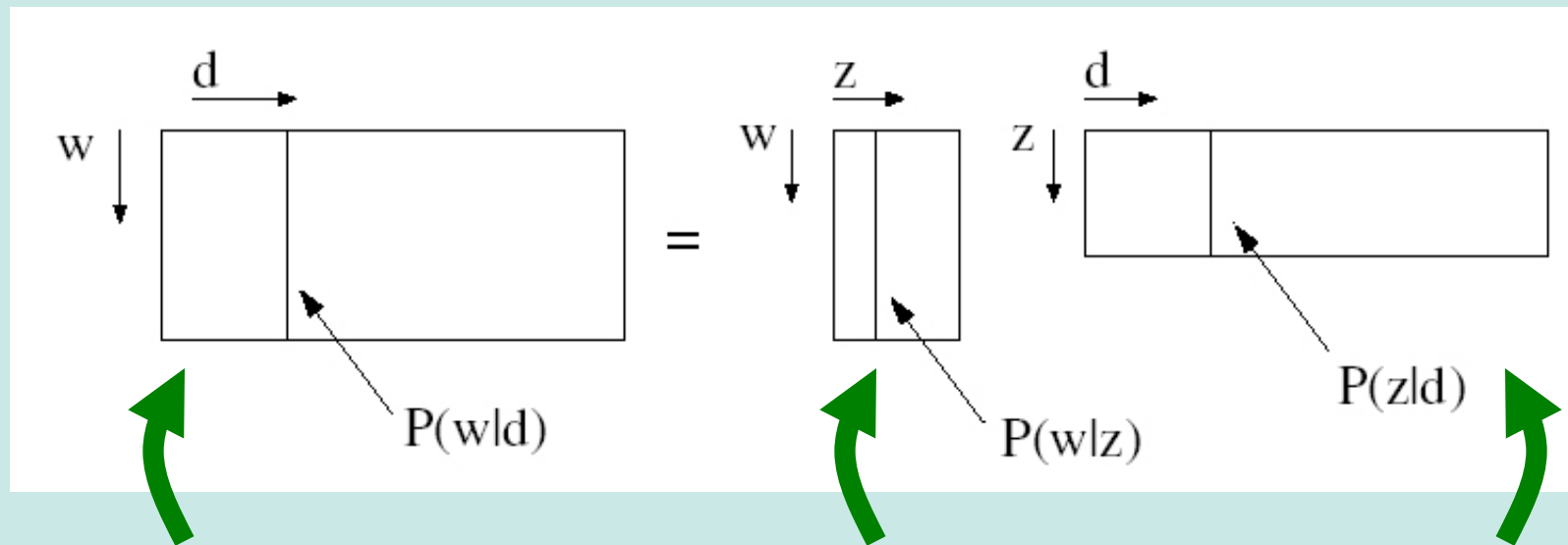
Probabilistic Latent Semantic Analysis (pLSA)



Case #2: the pLSA model



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

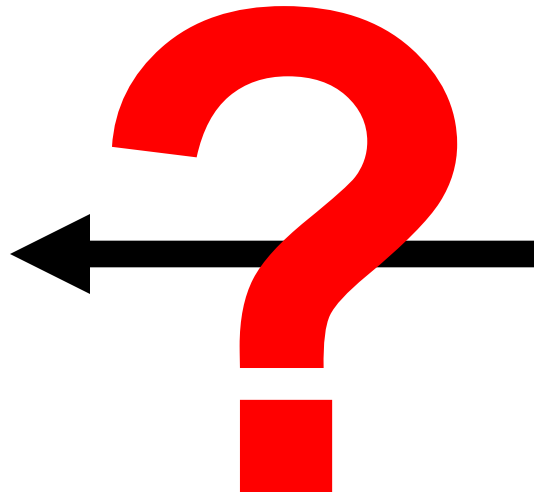


Observed codeword distributions

Codeword distributions per theme (topic)

Theme distributions per image

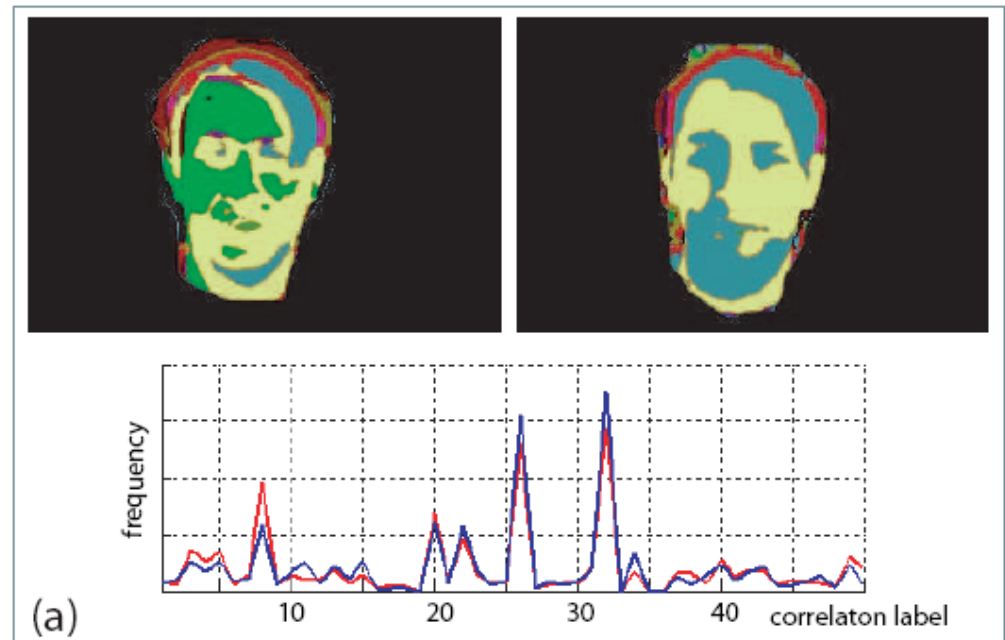
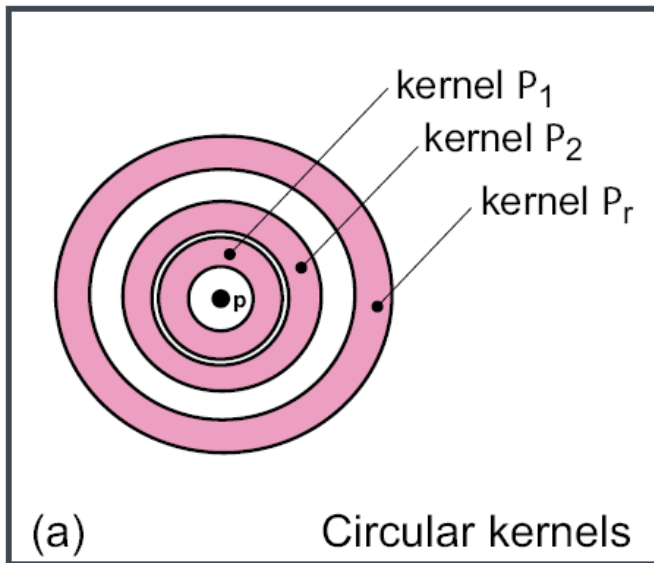
What about spatial info?



What about spatial info?



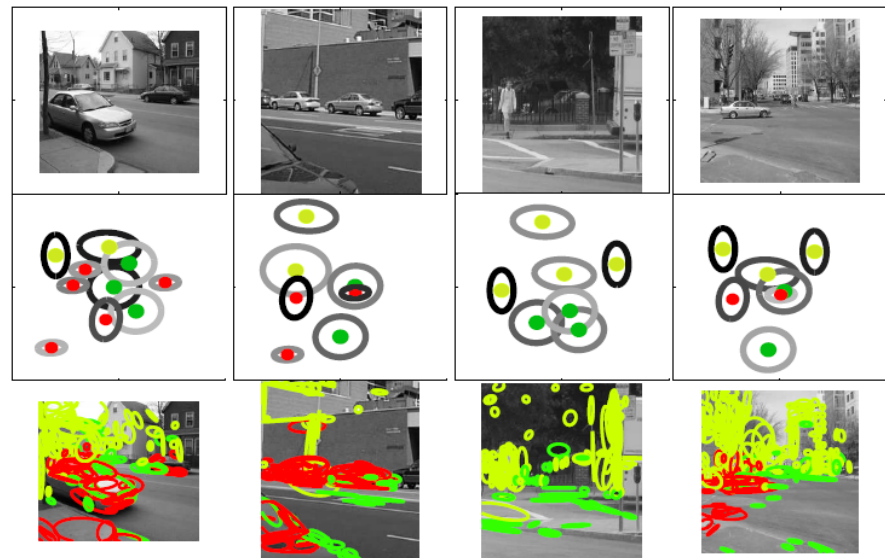
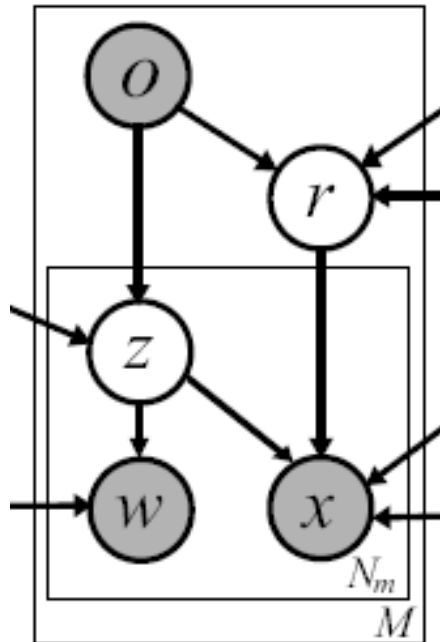
- Feature level
 - Spatial influence through correlogram features: Savarese, Winn and Criminisi, CVPR 2006



What about spatial info?



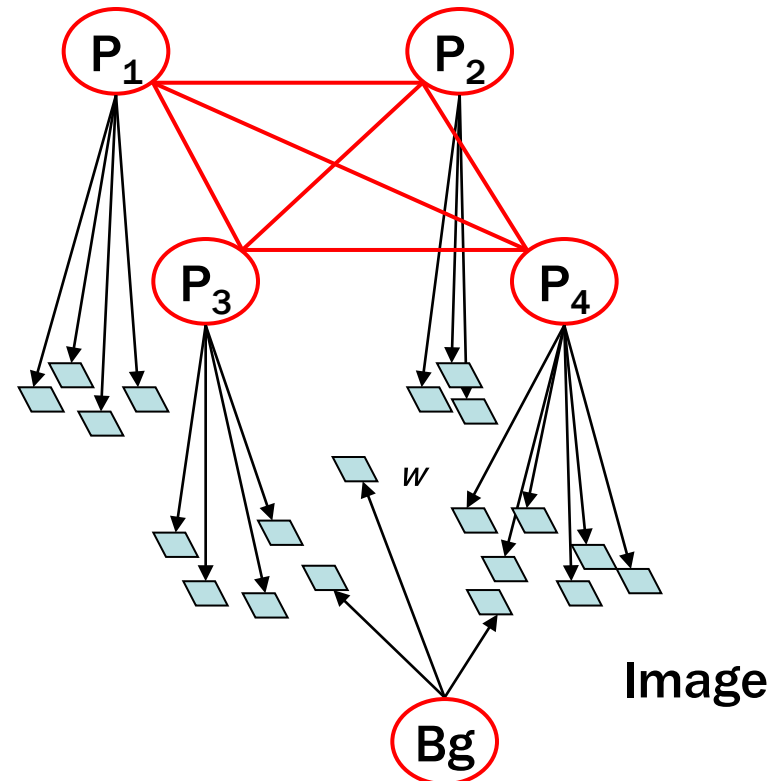
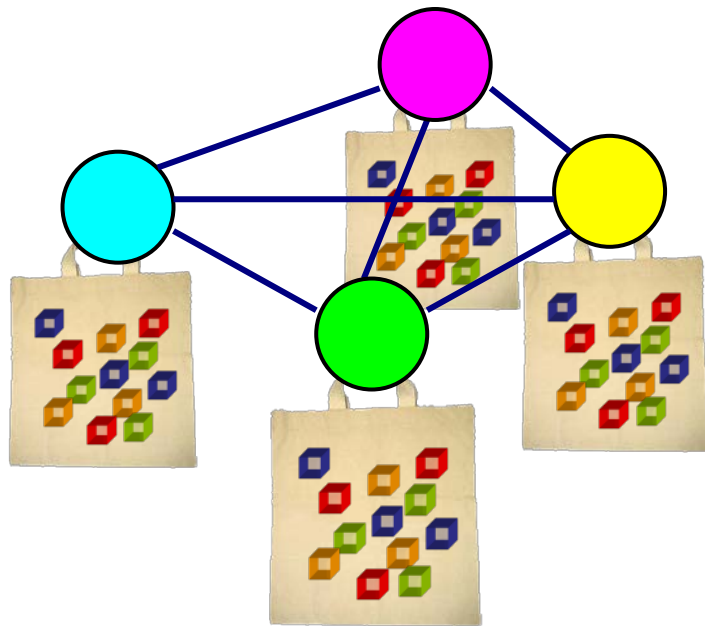
- Feature level
- Generative models
 - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
 - Niebles & Fei-Fei, CVPR 2007



What about spatial info?



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





Model properties

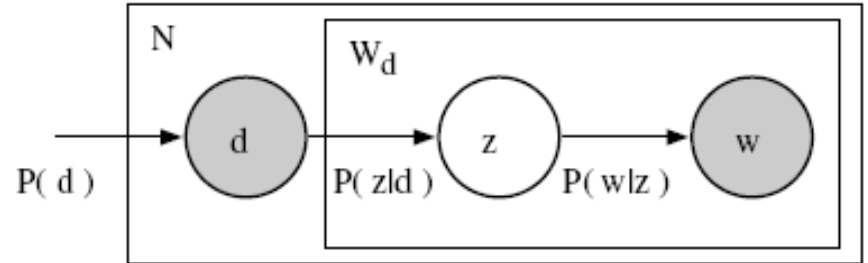
- Intuitive
 - Analogy to documents

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach us from our eyes. For a long time, the path of the retinal image was thought to be a direct path to the visual centers in the brain. As a movie camera records a scene, the retinal image is thought to be a direct record of the scene. However, Hubel and Wiesel discovered that the path of the retinal image is more complex than this. Following the path of the retinal image to the various centers of the visual cortex, Hubel and Wiesel have demonstrated that the message about the image falling on the retina undergoes a point-by-point analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

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

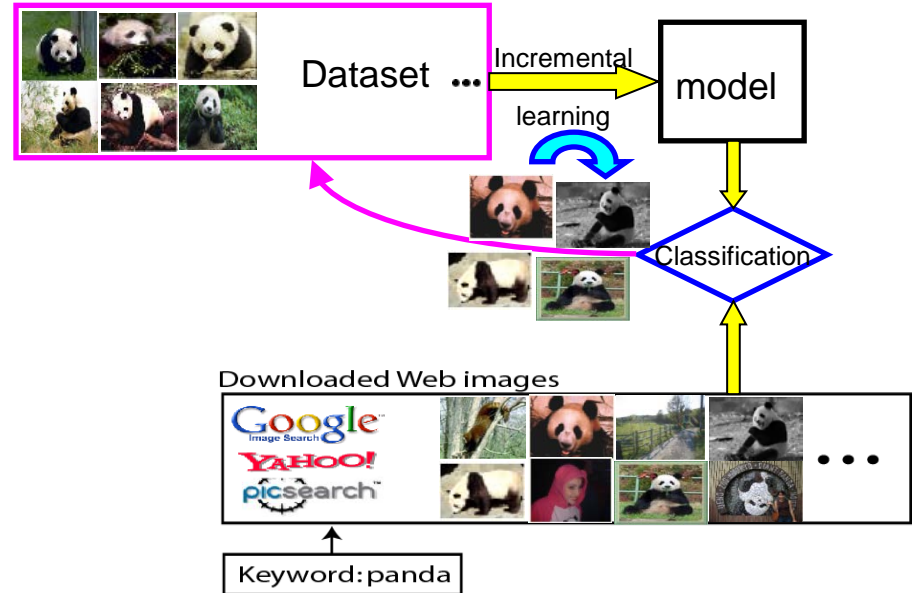


Model properties



Sivic, Russell, Efros, Freeman, Zisserman, 2005

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

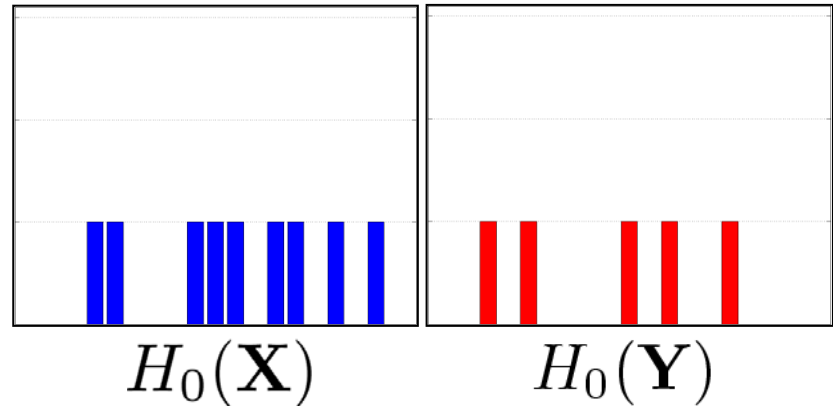
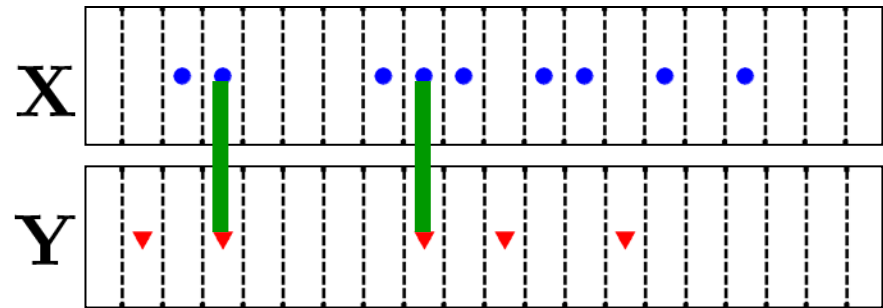


Li, Wang & Fei-Fei, CVPR 2007



Model properties

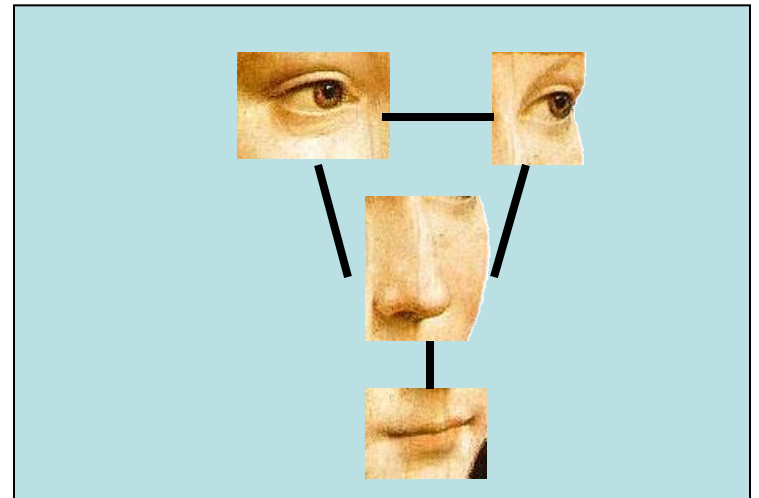
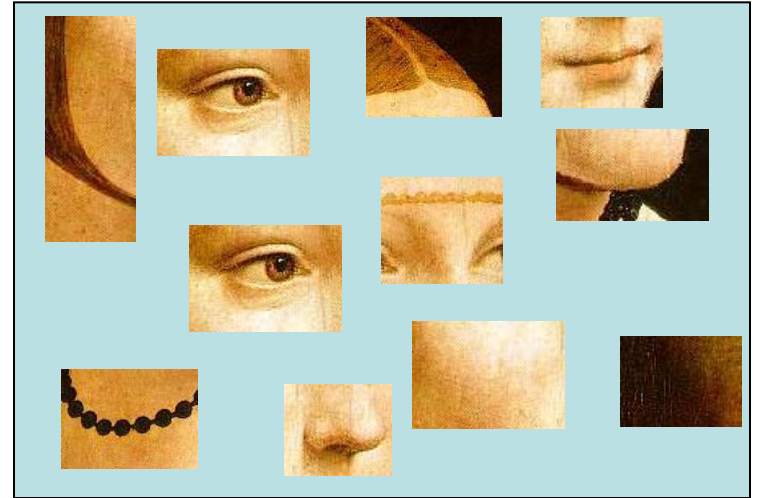
- Intuitive
- generative models
- Discriminative method
 - Computationally efficient



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