



Part 1: Bag-of-words models

by Li Fei-Fei (Princeton)

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 reach our eyes. For a long time, the retinal image was considered as a movie screen. It is now discovered that the image is analyzed in a more complex way 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 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.*

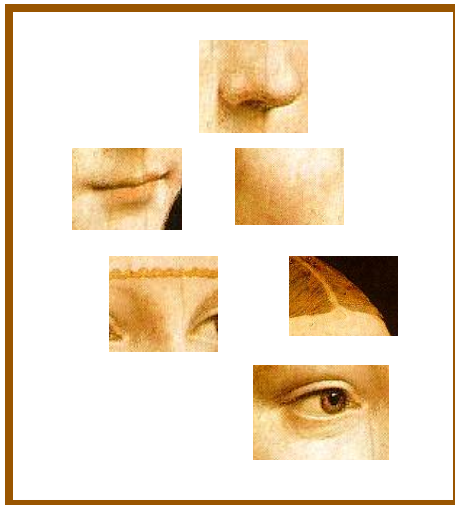
**sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel**

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 it will reduce the US's trade deficit. China's government has deliberately kept the yuan's value low to encourage exports. The government agrees that the yuan is undervalued. The government also needs to increase the demand for its goods in the rest of the world. China has been allowed to trade the yuan against the dollar since 2005 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**

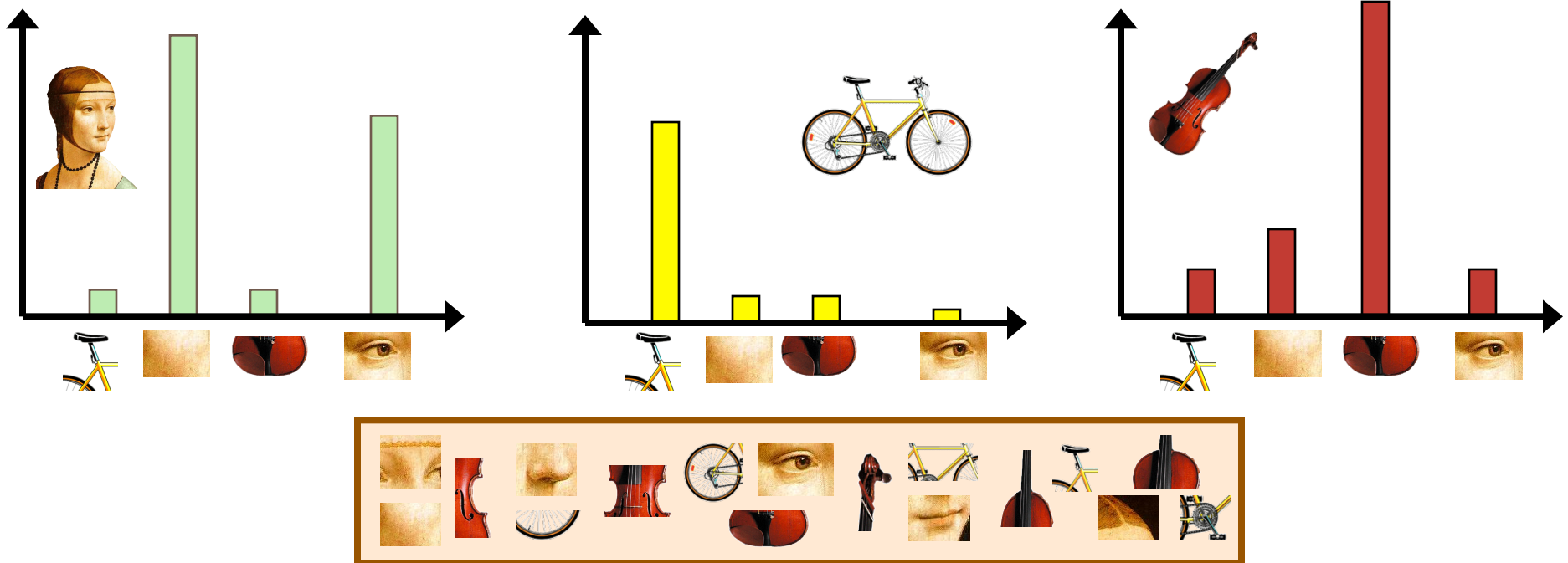
A clarification: definition of “BoW”

- Looser definition
 - Independent features

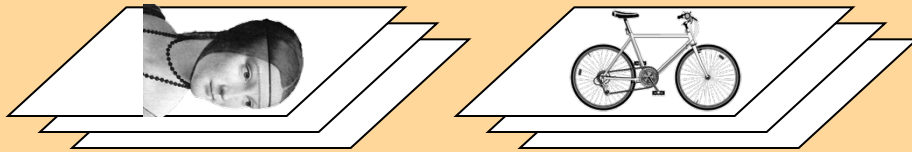


A clarification: definition of “BoW”

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



learning



feature detection
& representation

codewords dictionary

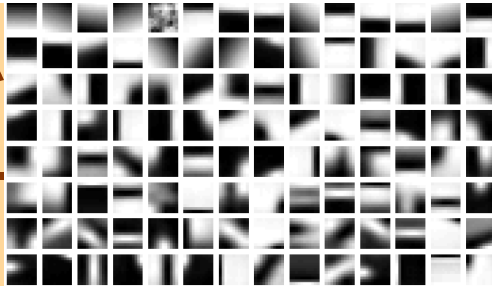
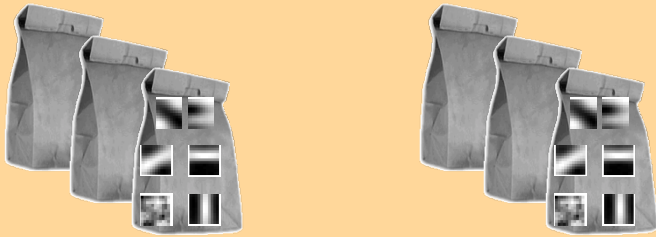
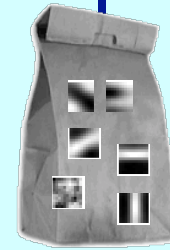


image representation



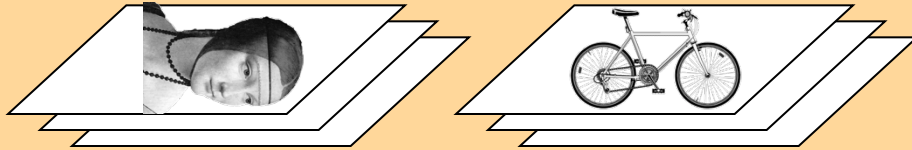
**category models
(and/or) classifiers**

recognition

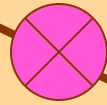


**category
decision**

Representation



1. feature detection & representation



2. codewords dictionary

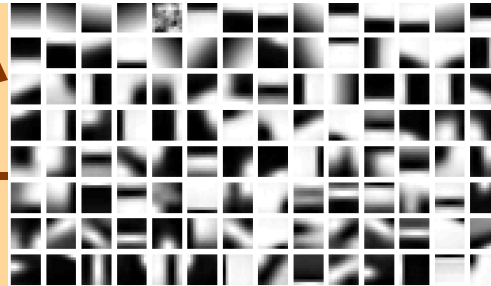
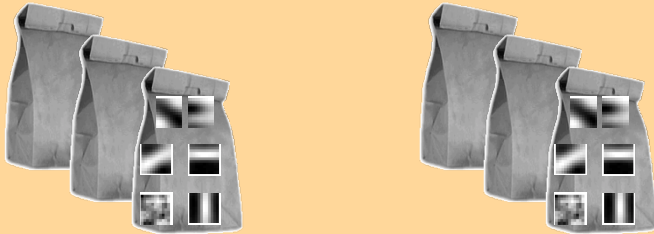
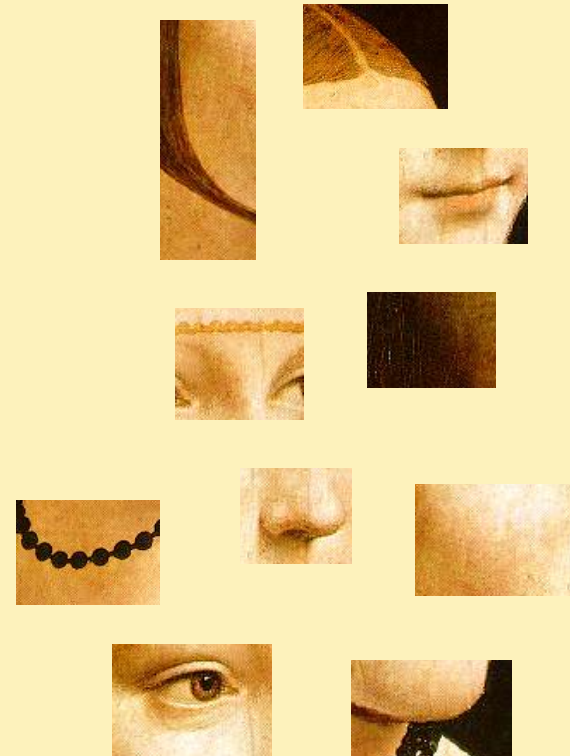


image representation

3.



1. Feature detection and representation



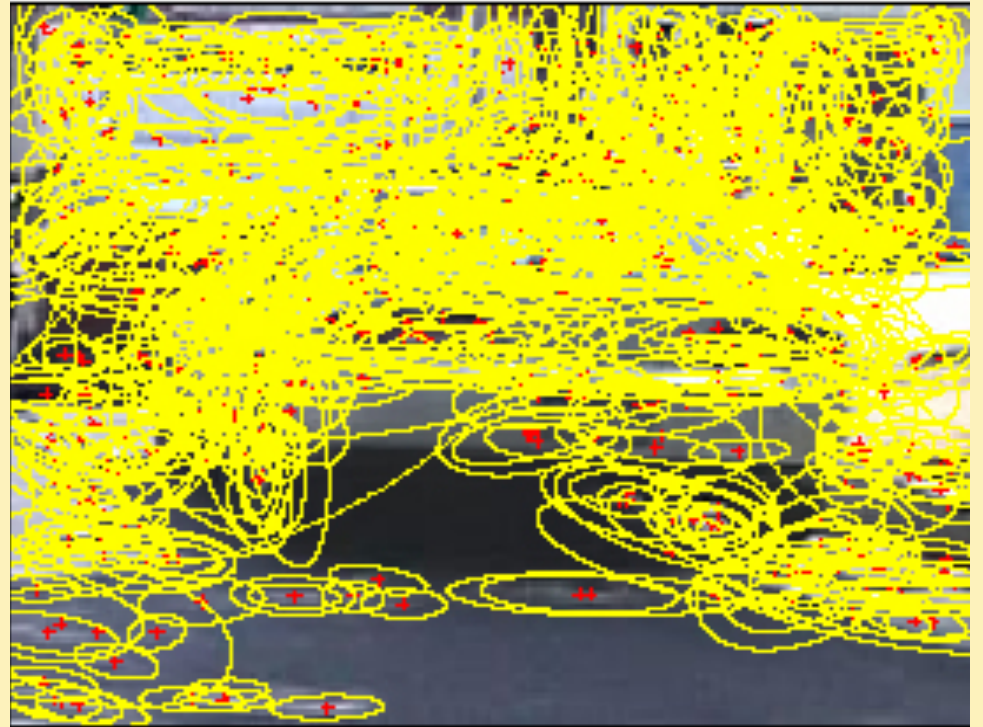
1. Feature detection and representation

- 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



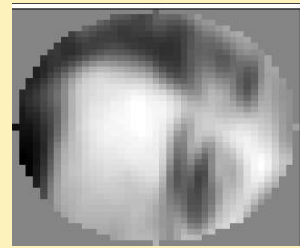
1. Feature detection and representation

- 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)

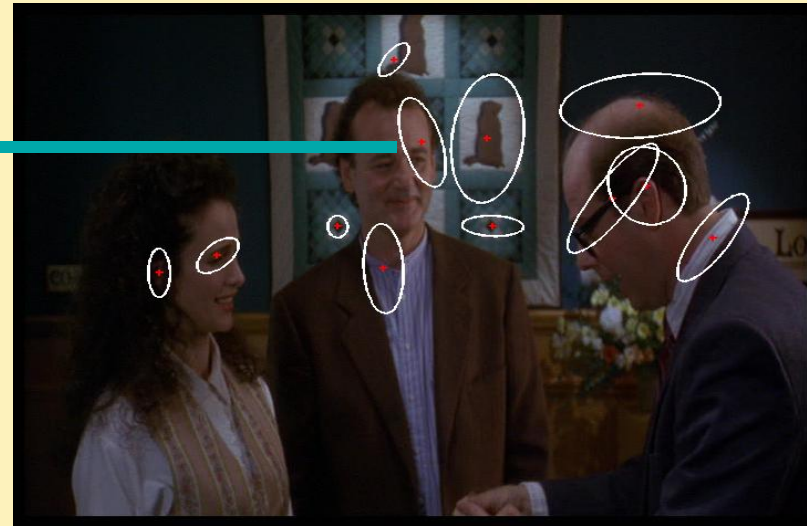
1. Feature detection and representation



**Compute
SIFT
descriptor**
[Lowe'99]



**Normalize
patch**



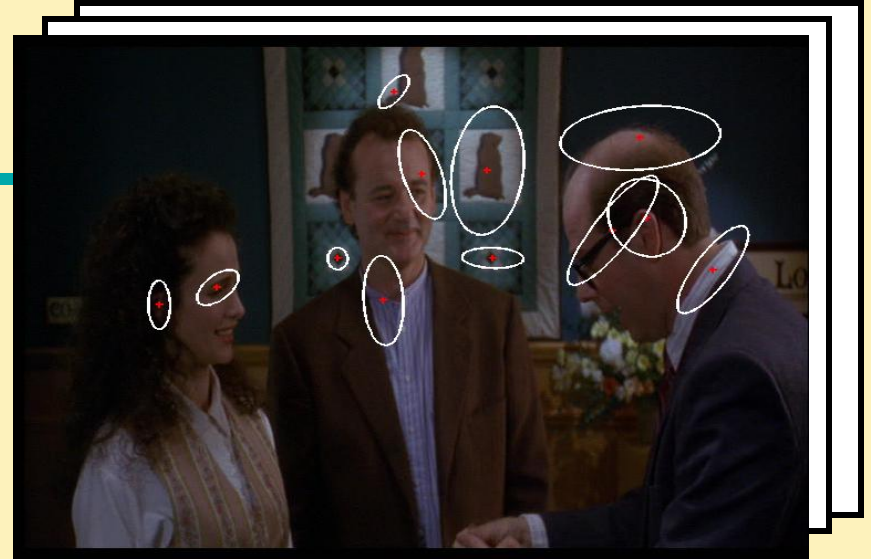
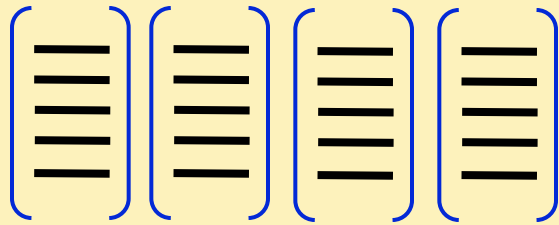
Detect patches

[Mikojczyk and Schmid '02]

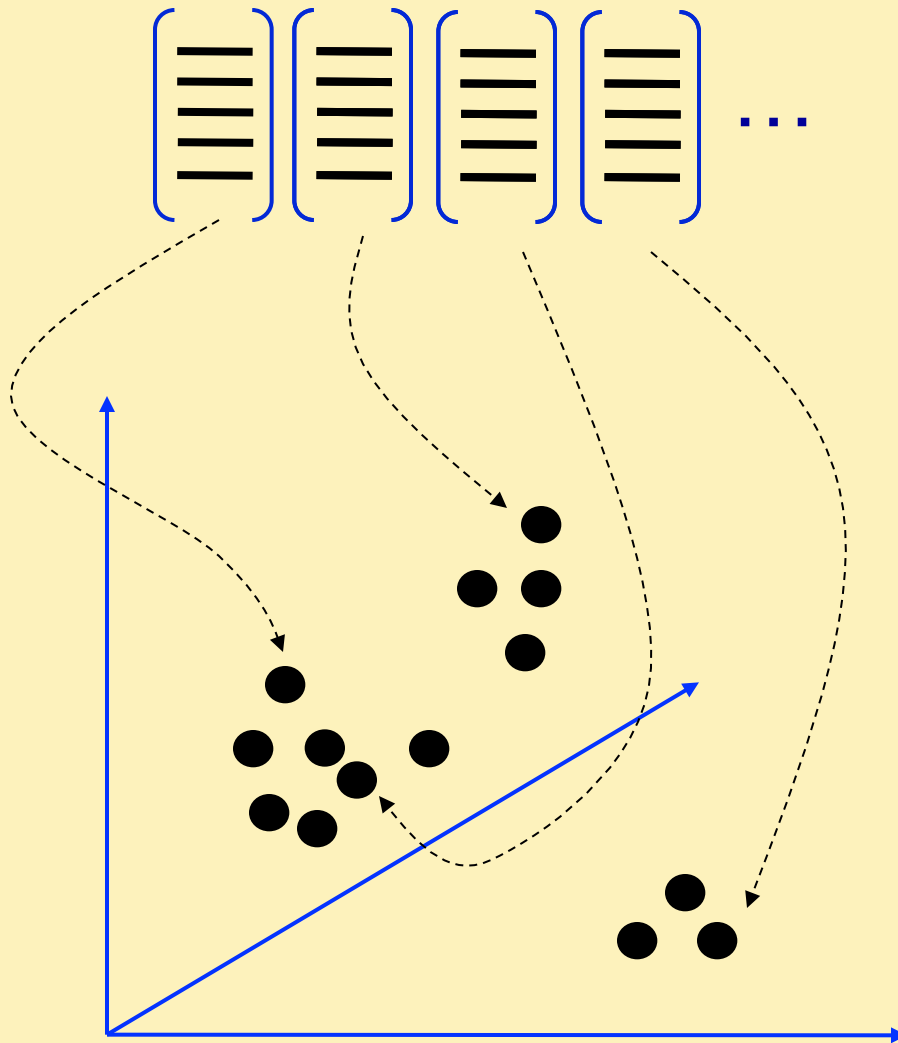
[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

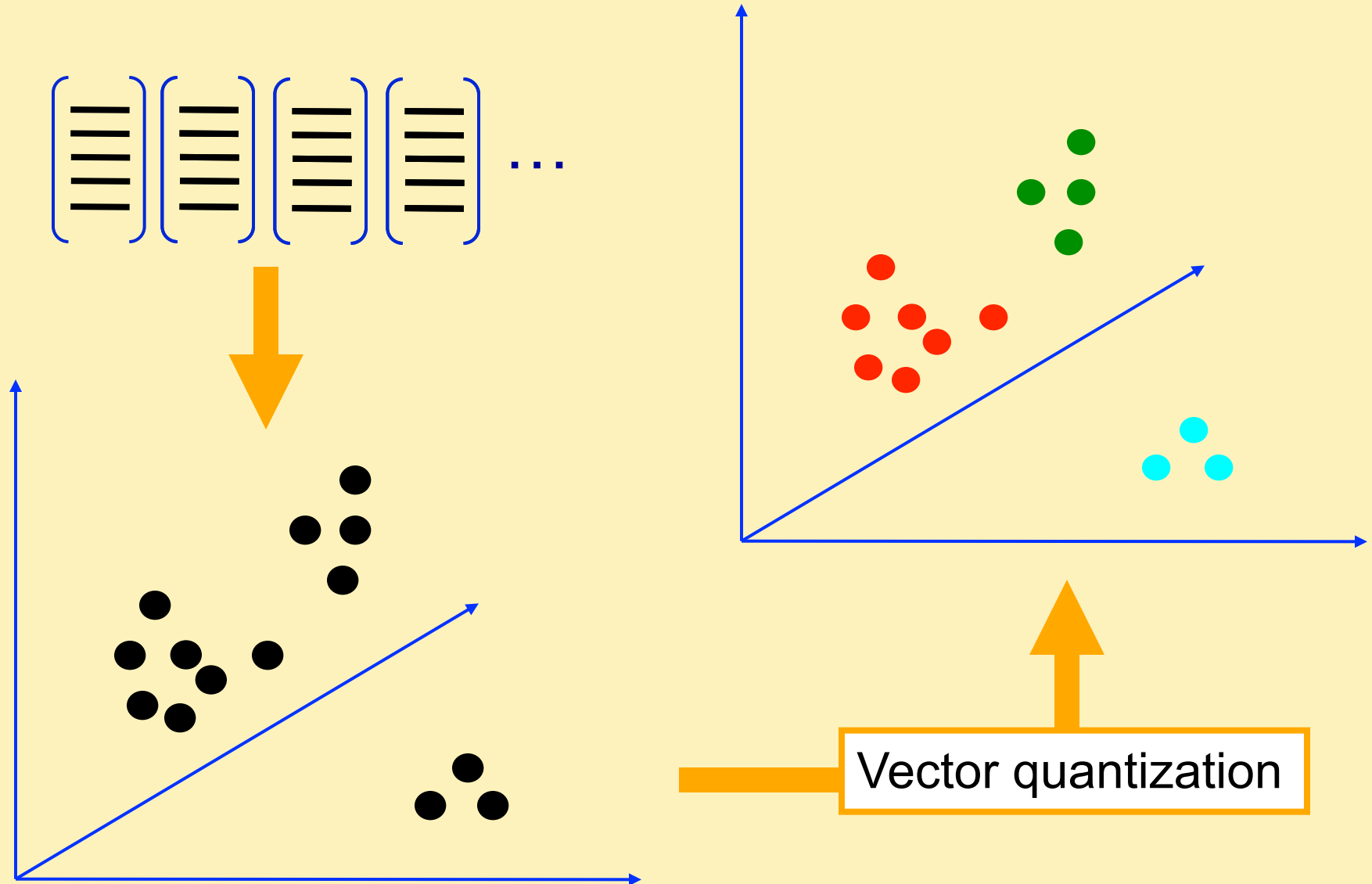
1. Feature detection and representation



2. Codewords dictionary formation



2. Codewords dictionary formation



2. Codewords dictionary formation

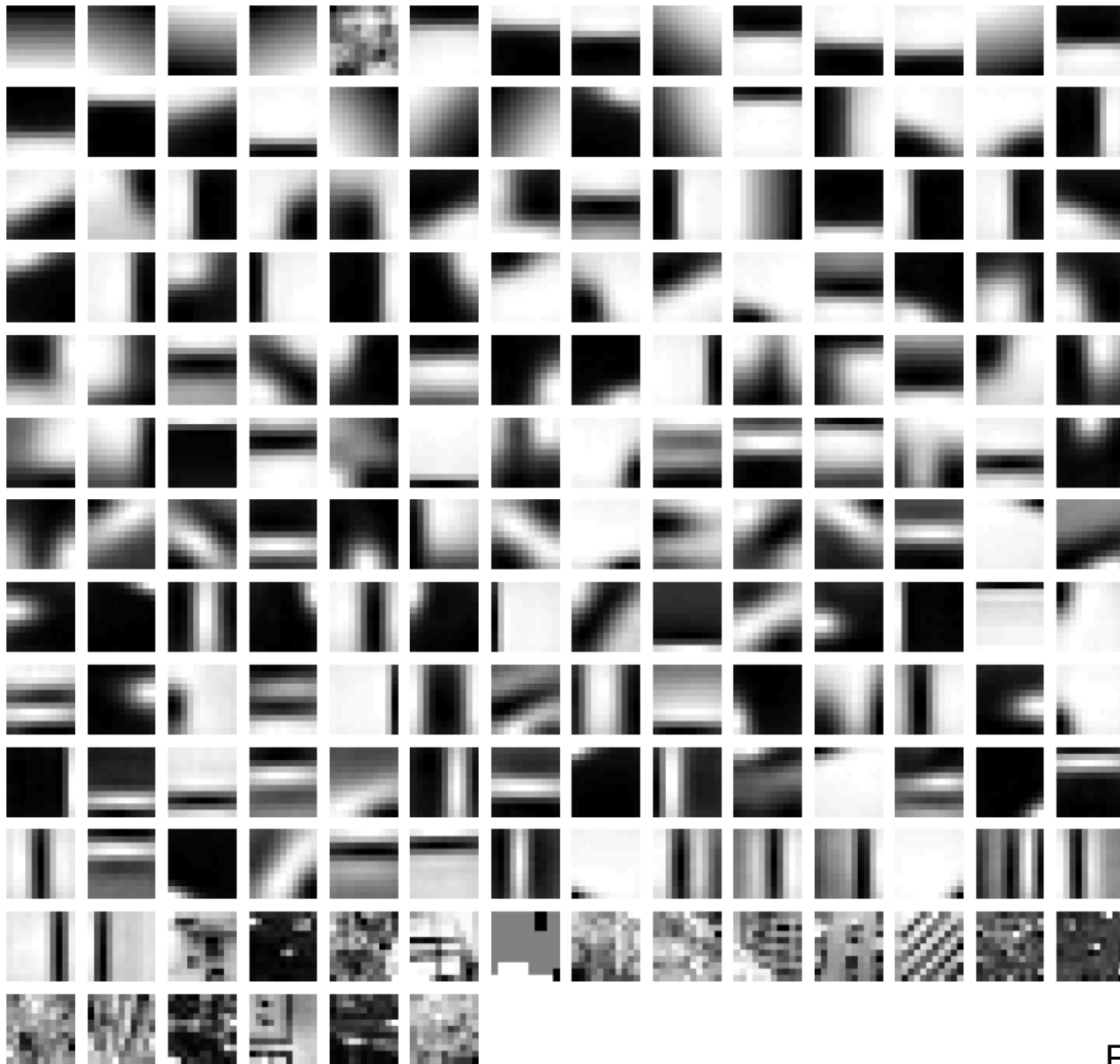
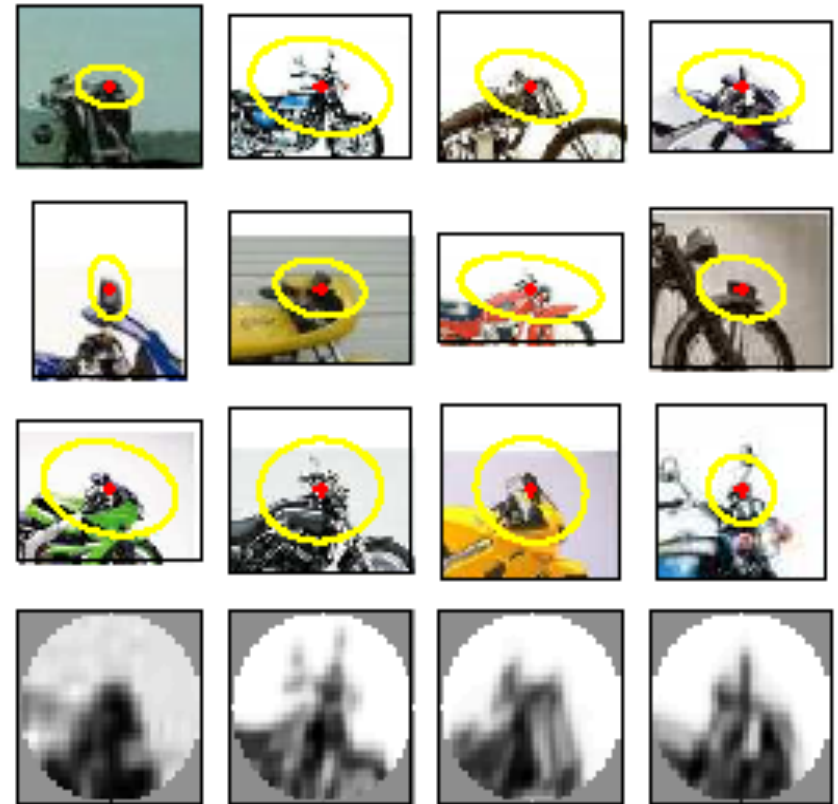
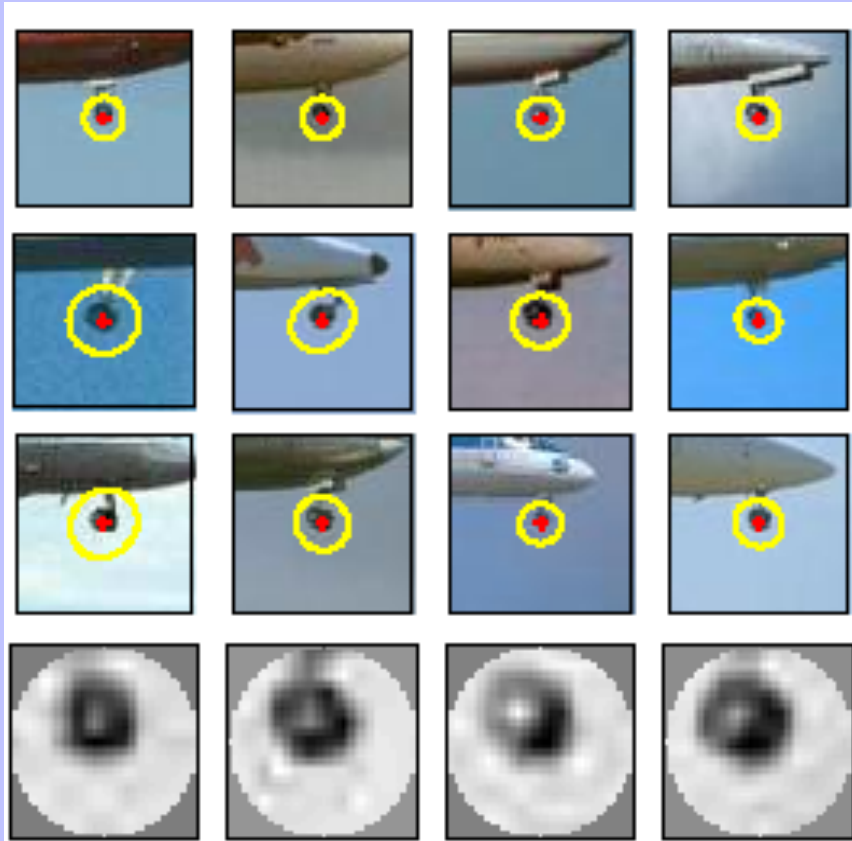


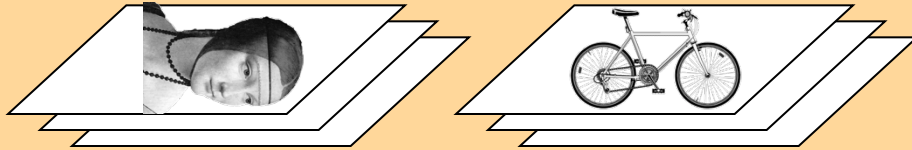
Image patch examples of codewords



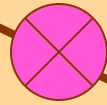
3. Image representation



Representation



1. feature detection & representation



2. codewords dictionary

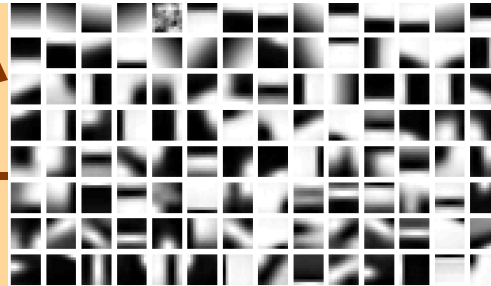
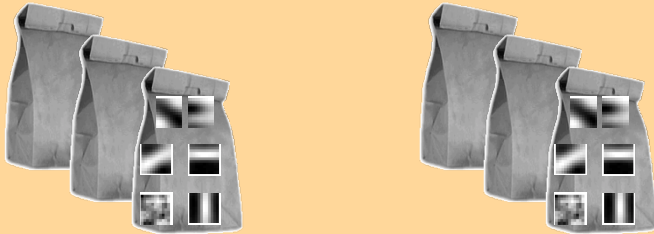


image representation

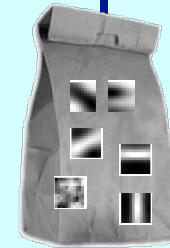
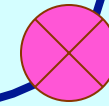
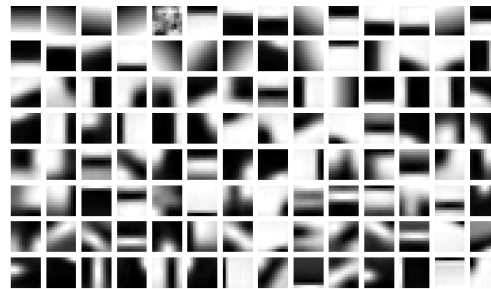
3.



Learning and Recognition

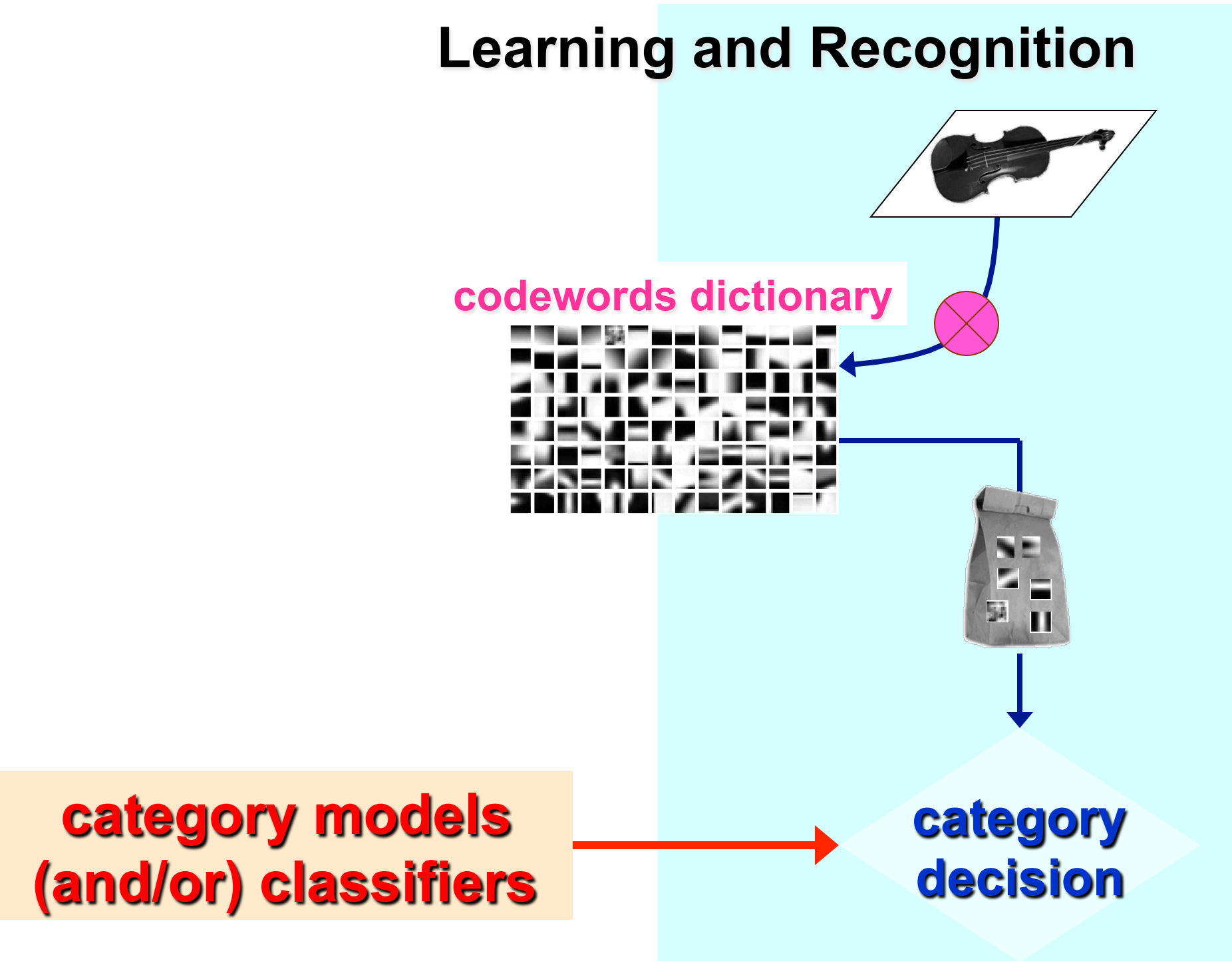


codewords dictionary



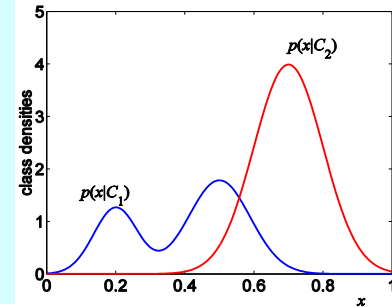
**category models
(and/or) classifiers**

**category
decision**

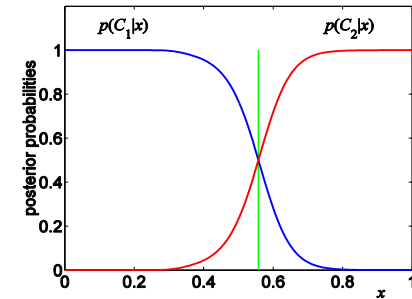


Learning and Recognition

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

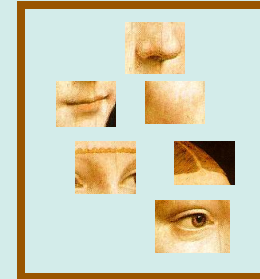
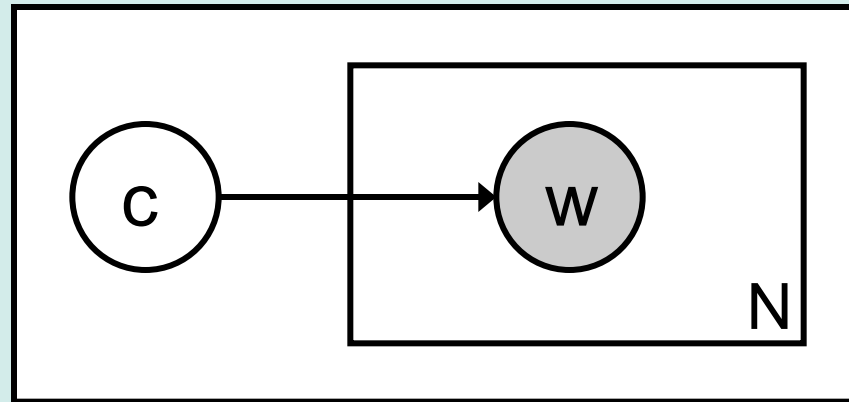
2. Hierarchical Bayesian text models (pLSA and LDA)

- 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.

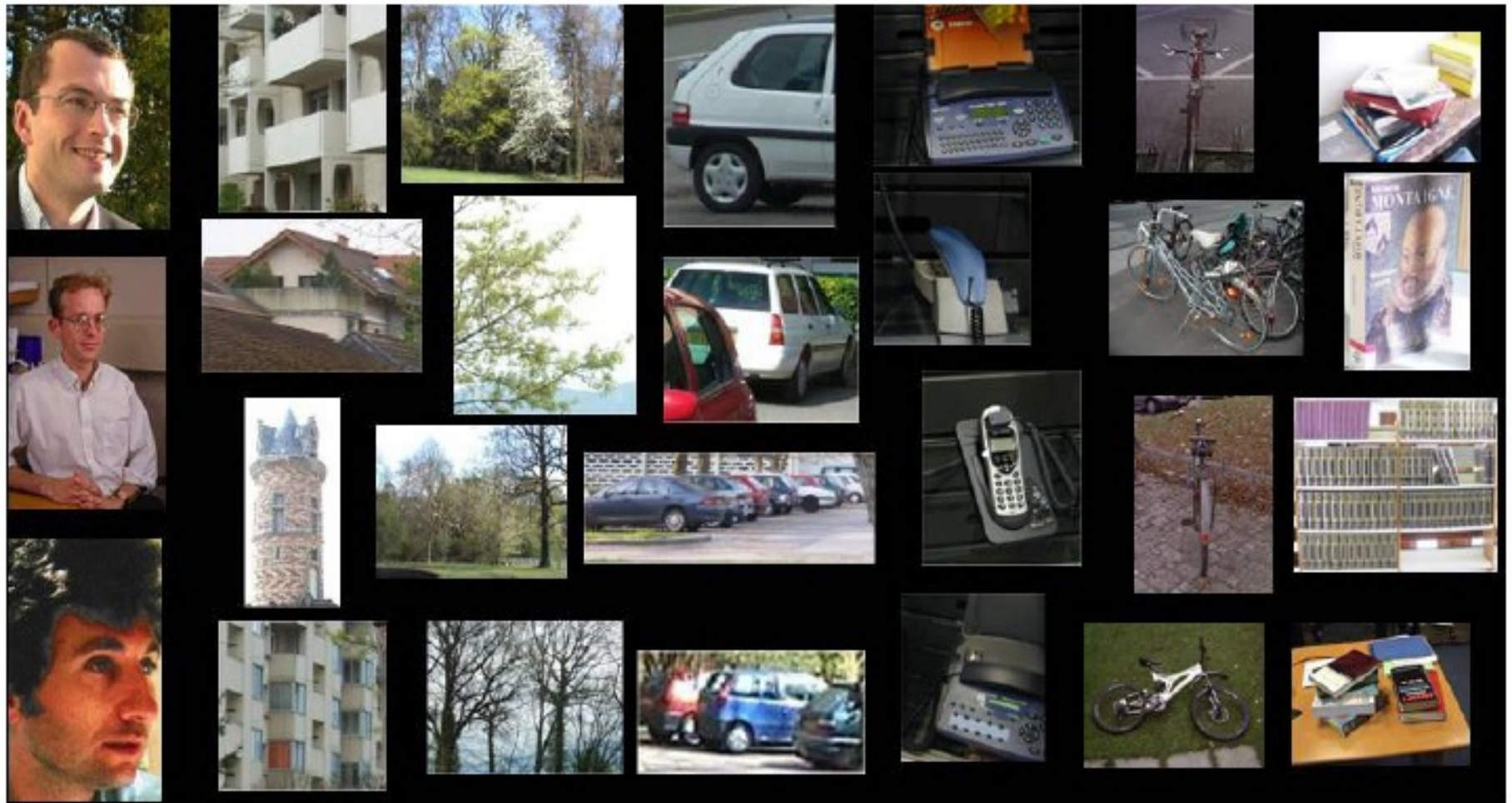
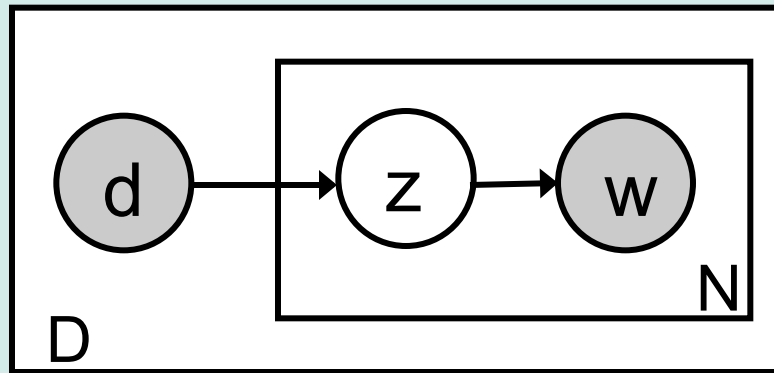


Table 1. Confusion matrix and the mean rank for the best vocabulary ($k=1000$).

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

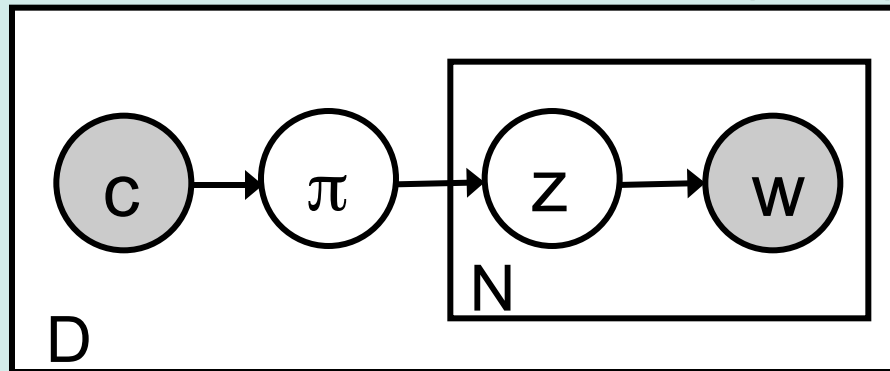
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 2001

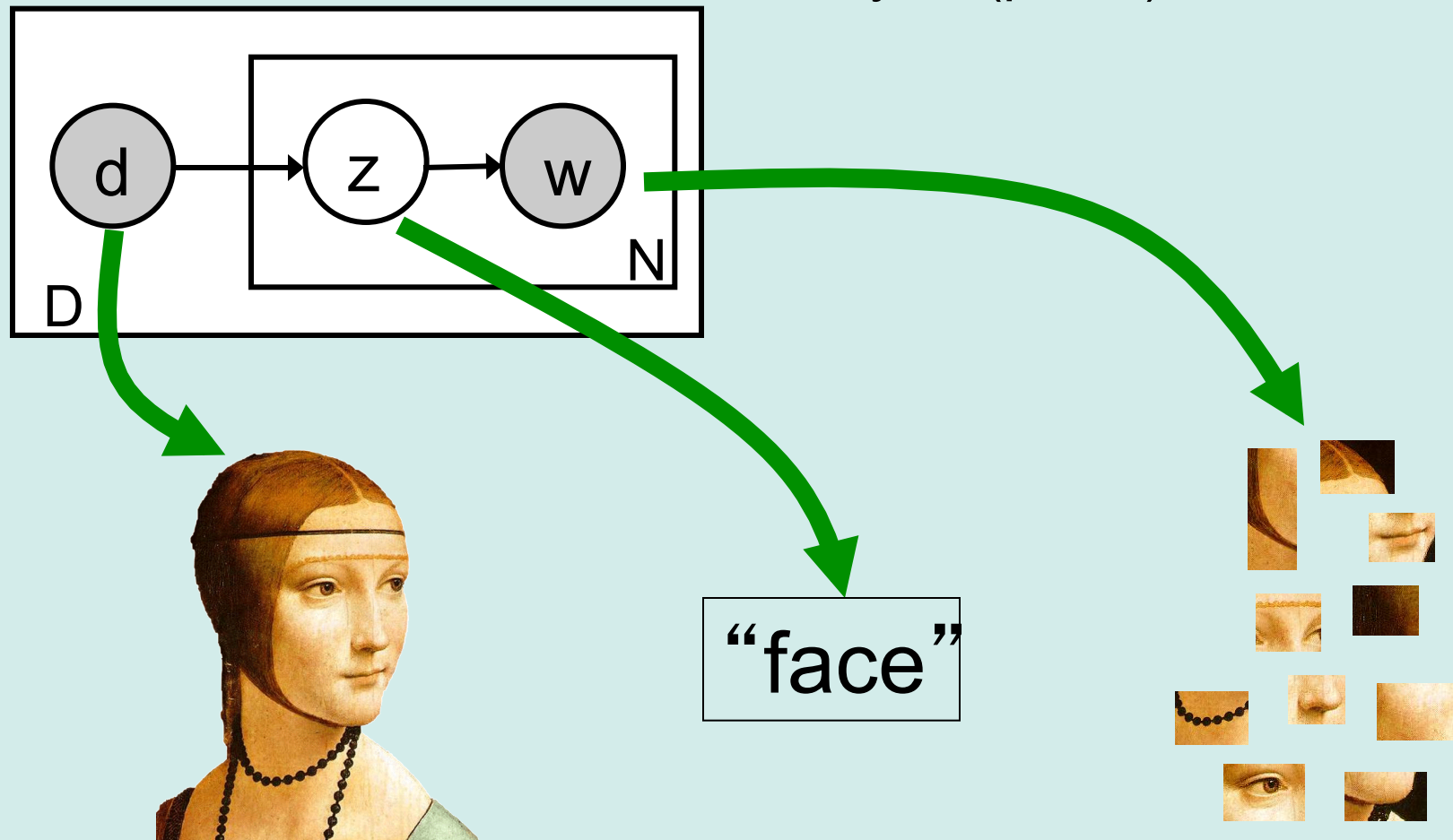
Latent Dirichlet Allocation (LDA)



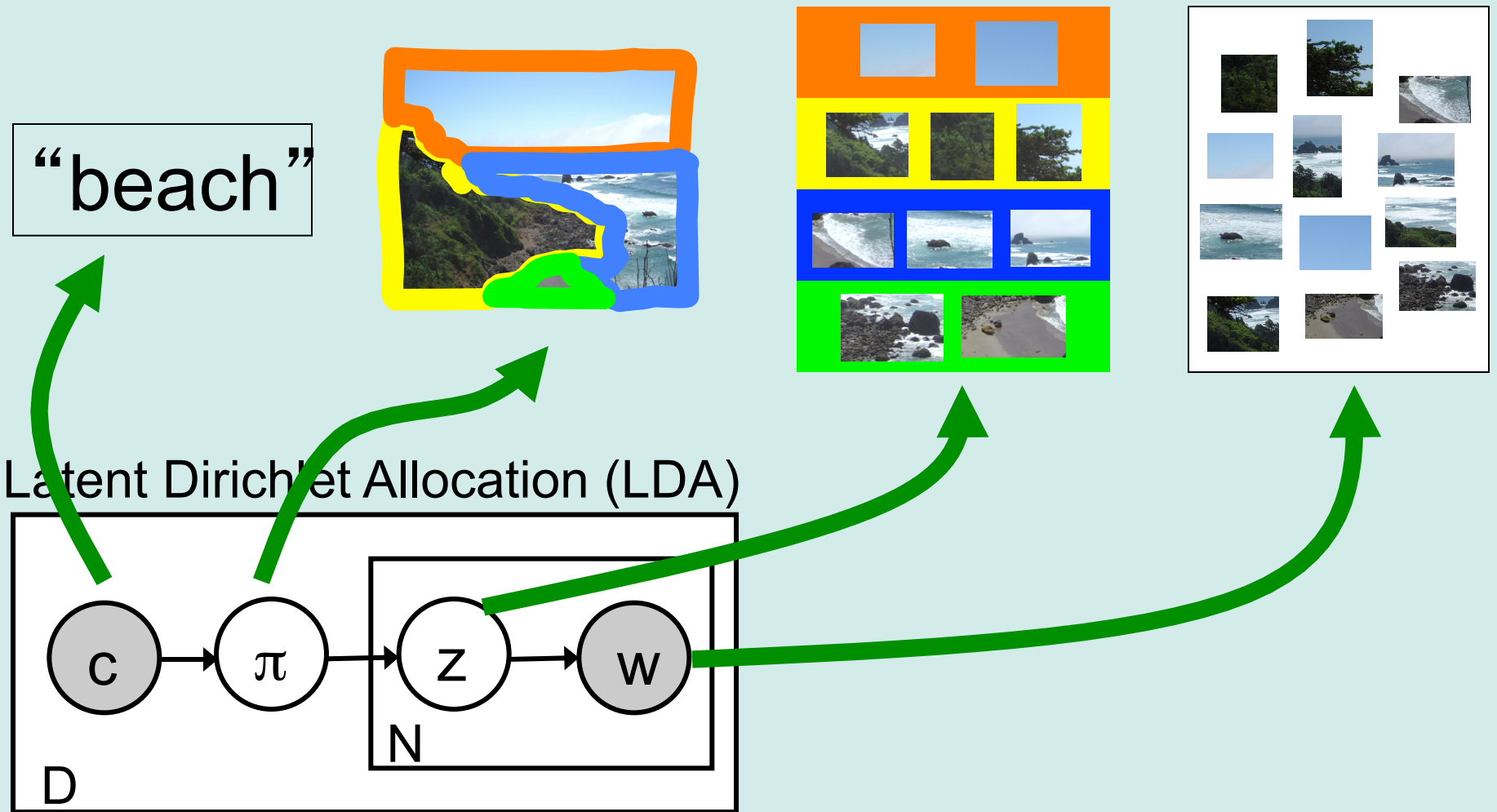
Blei et al., 2001

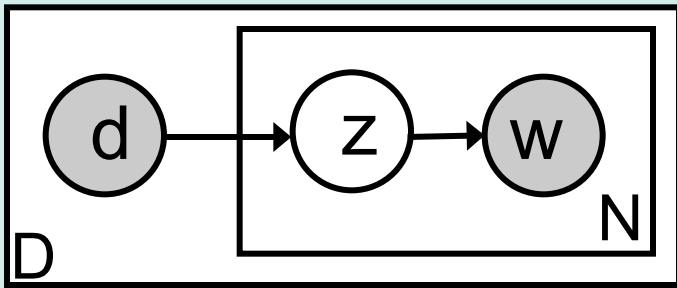
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



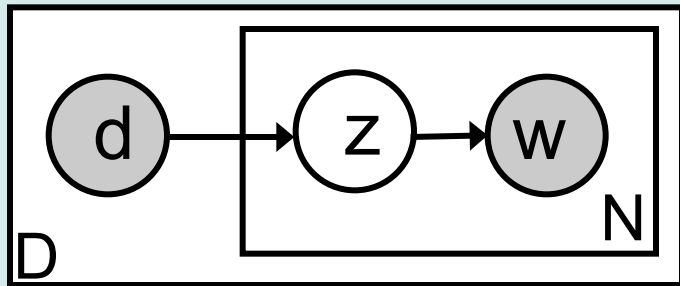
Case #2: Hierarchical Bayesian text models





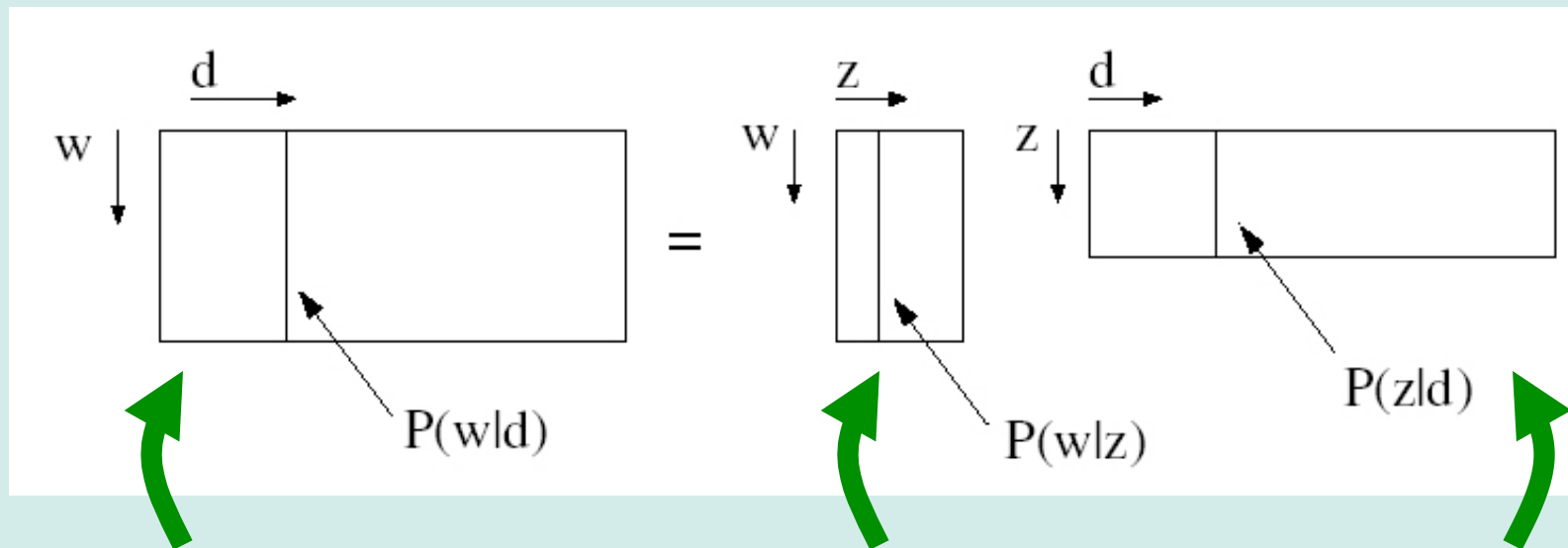
Case #2: the pLSA model





Case #2: the pLSA model

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



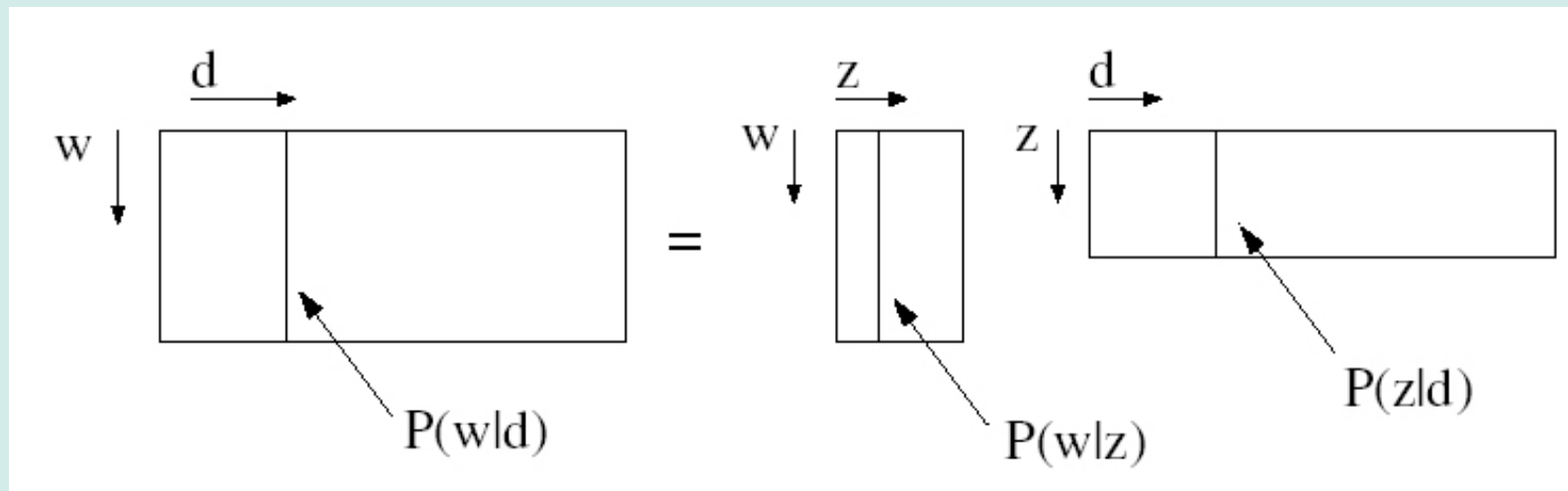
Observed codeword distributions

Codeword distributions per theme (topic)

Theme distributions per image

Case #2: Recognition using pLSA

$$z^* = \arg \max_z p(z | d)$$



Case #2: Learning the pLSA parameters

Observed counts of
word i in document j

$$L = \prod_{i=1}^M \prod_{j=1}^N P(w_i | d_j)^{n(w_i, d_j)}$$

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

Maximize likelihood of data using EM

M ... number of codewords

N ... number of images

Demo

- Course website


A demonstration of bag-of-words classifiers - Microsoft Internet Explorer provided by Insight Broadband

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Recycle Bin Mail Print Fax AutoLink AutoF

Address <http://people.csail.mit.edu/fergus/iccv2005/bagwords.html>

Google Search 100 blocked Check AutoLink AutoF



Two bag-of-words classifiers

**ICCV 2005 short courses on
Recognizing and Learning Object Categories**

A simple approach to classifying images is to treat them as a collection of regions, describing only their appearance and ignoring their location. This approach has been successfully used in the text community for analyzing documents and are known as "bag-of-words" models, since each document is represented by its distribution over fixed vocabulary(s). Using such a representation, methods such as probabilistic latent semantic analysis (pLSA) [1] (LDA) [2] are able to extract coherent topics within document collections in an unsupervised manner.

Recently, Fei-Fei et al. [3] and Sivic et al. [4] have applied such methods to the visual domain. The demo code implements pLSA, including a Naïve Bayes classifier for comparison. For comparison, a Naïve Bayes classifier is also provided which requires labelled training data, unlike pLSA.

The code consists of Matlab scripts (which should run under both Windows and Linux) and a couple of 32-bit Linux binaries for doing image classification. Hence the whole system will need to be run on Linux. The code is for teaching/research purposes only. If you find a bug, please email me at fergus@csail.mit.edu where csail point mit point edu.

Download

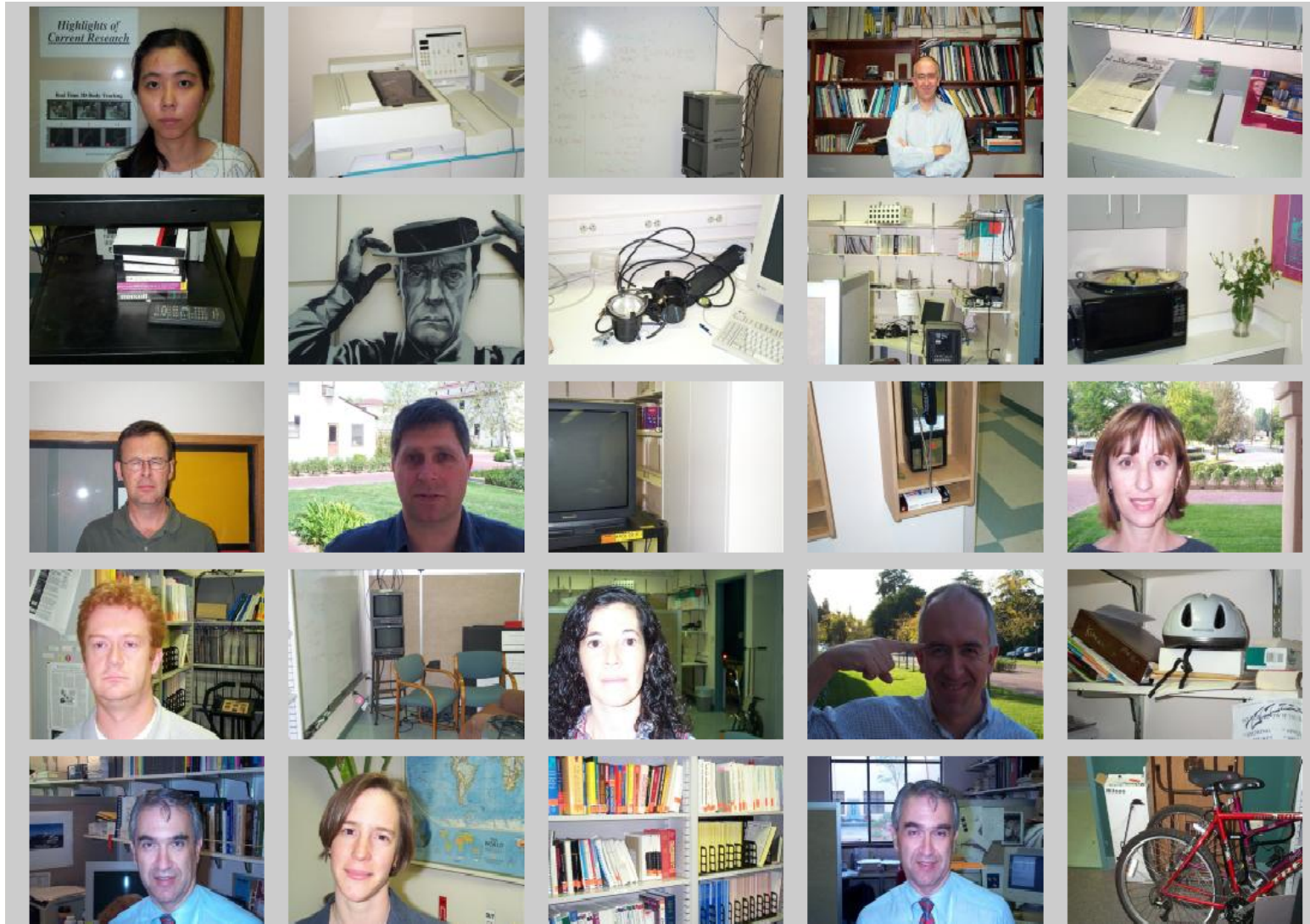
[Download](#) the code and datasets (32 Mbytes)

Operation of code

To run the demos:

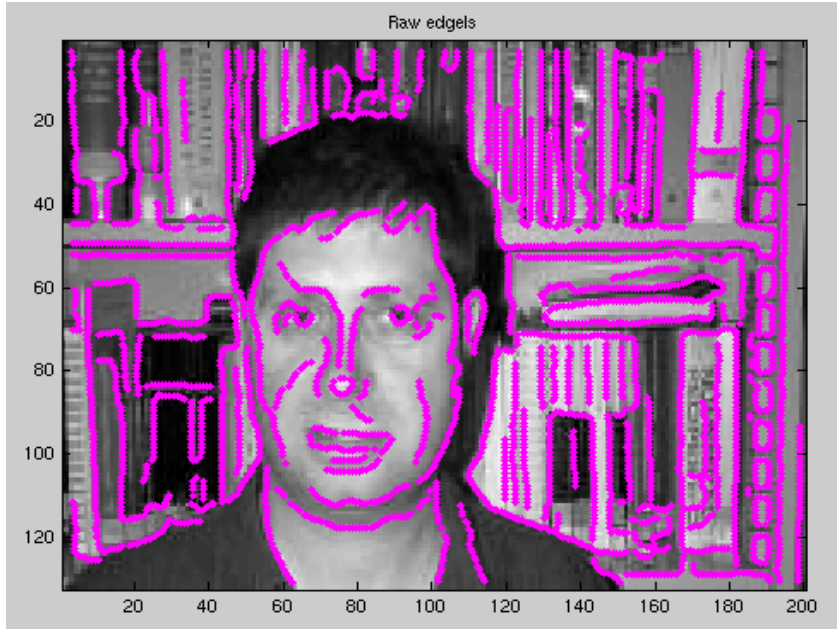
start Microsoft Outlook We... 未名空间(mitbbs.co... A demonstration of b... ICCV200

task: face detection – no labeling



Demo: feature detection

- Output of crude feature detector
 - Find edges
 - Draw points randomly from edge set
 - Draw from uniform distribution to get scale

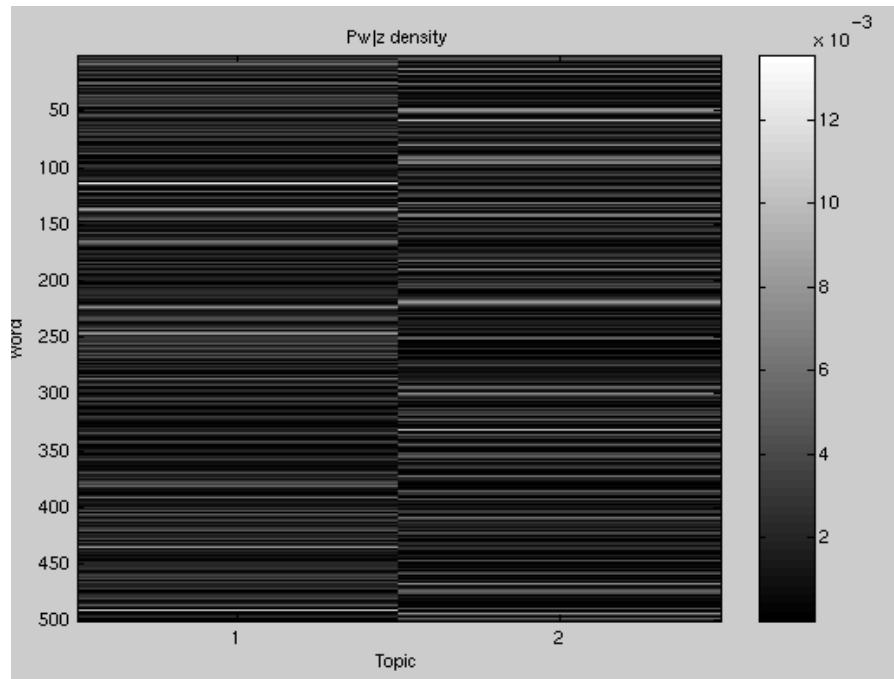


Demo: learnt parameters

- Learning the model: `do_plsa('config_file_1')`
- Evaluate and visualize the model: `do_plsa_evaluation('config_file_1')`

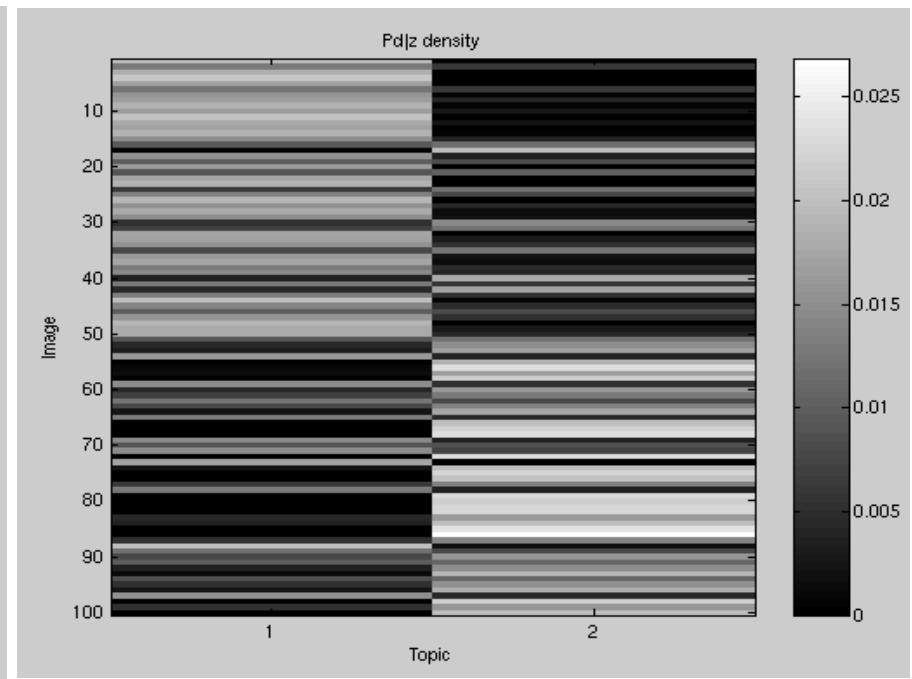
Codeword distributions
per theme (topic)

$$p(w | z)$$

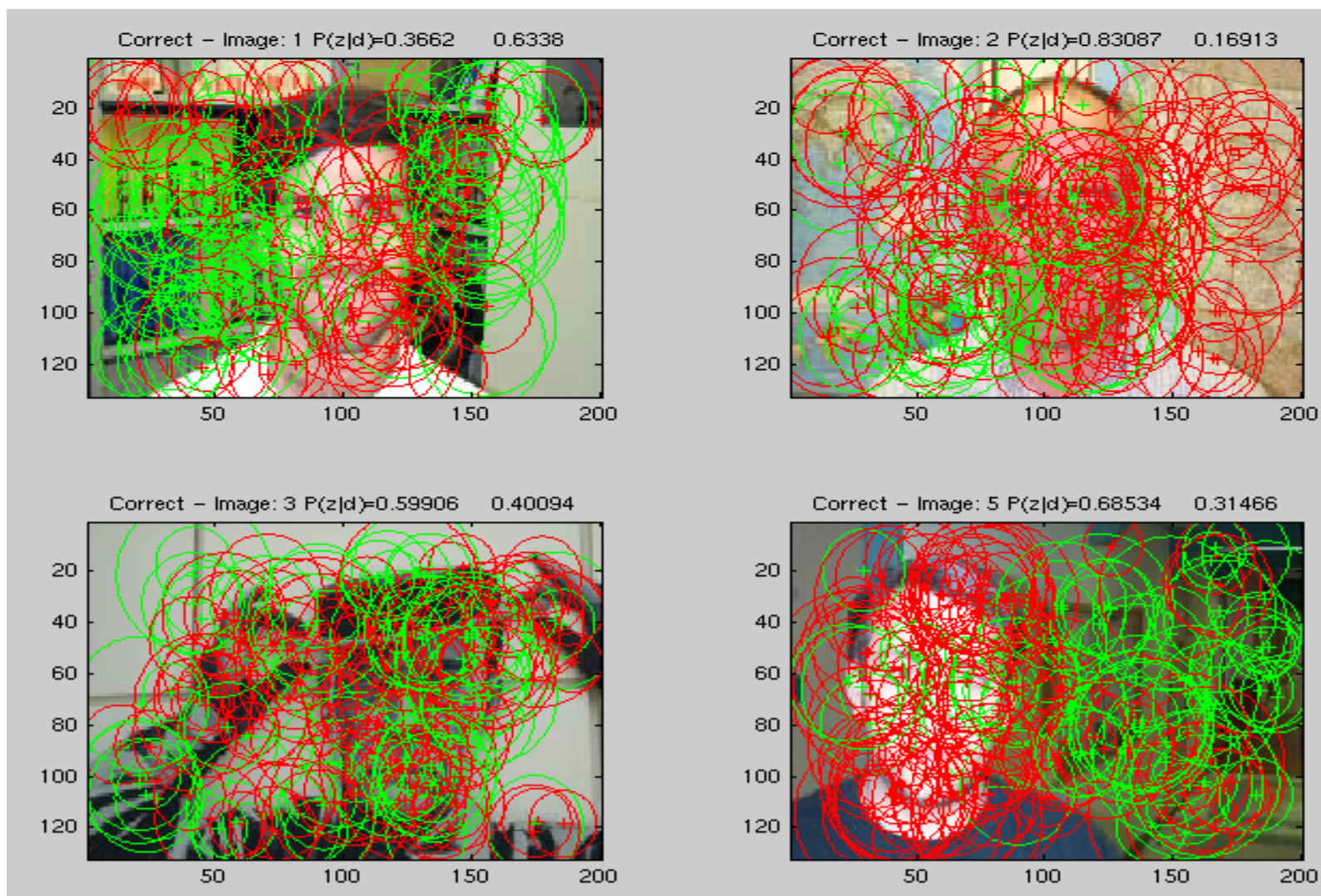


Theme distributions
per image

$$p(z | d)$$

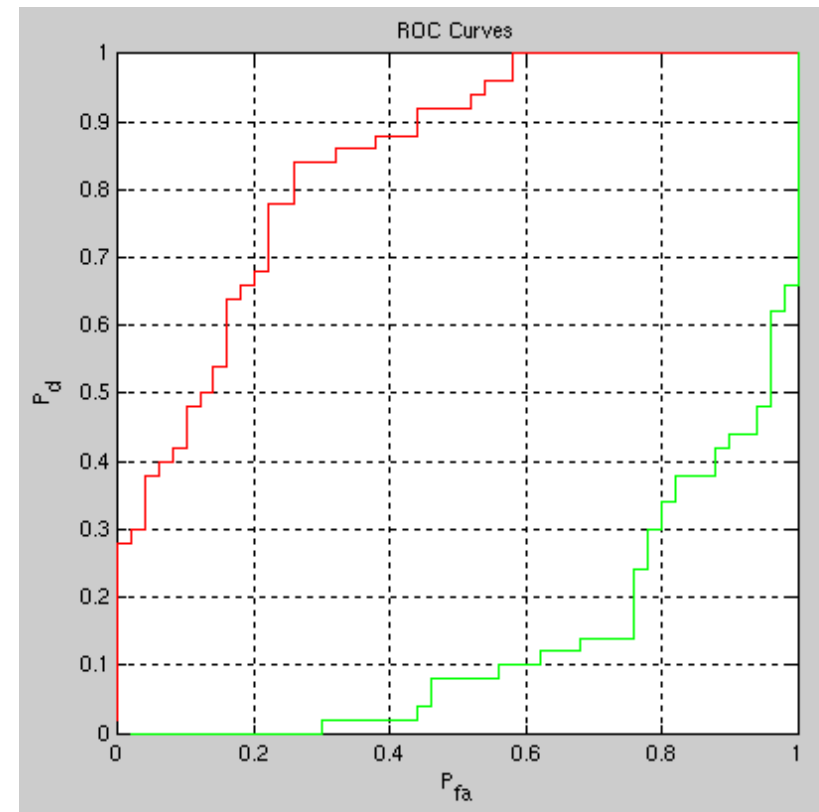
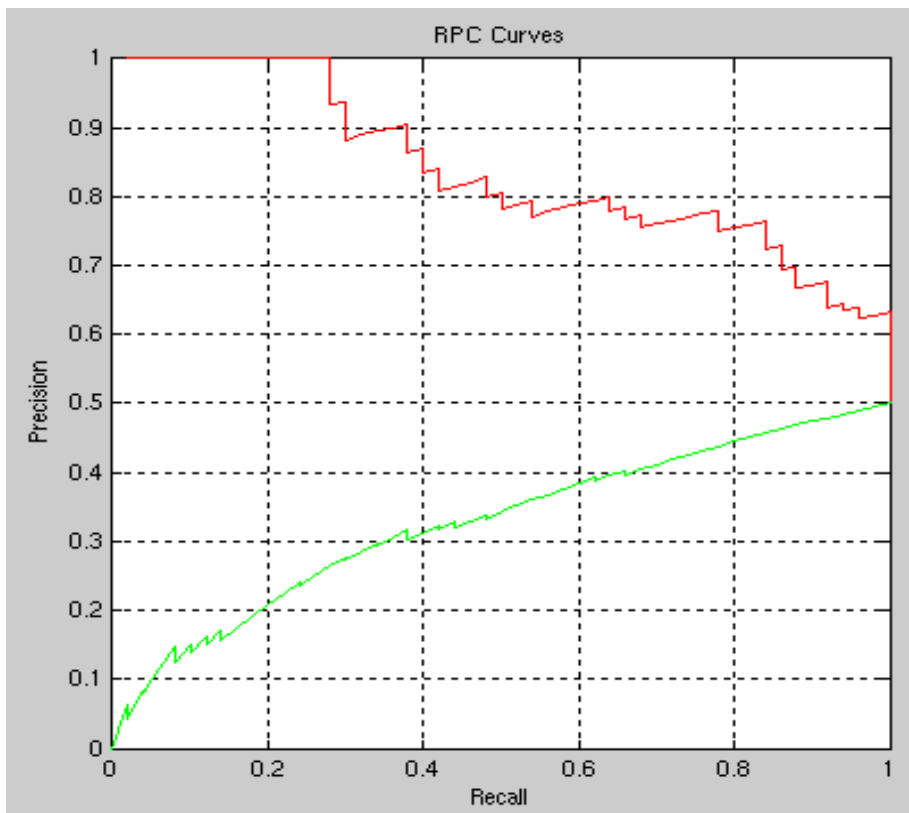


Demo: recognition examples



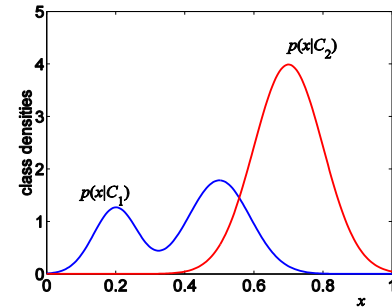
Demo: categorization results

- Performance of each theme

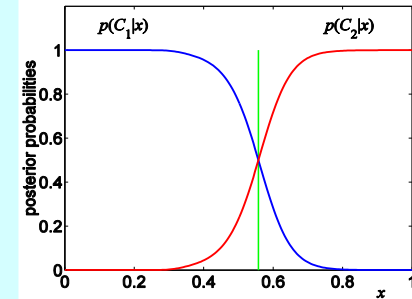


Learning and Recognition

1. Generative method:
 - graphical models

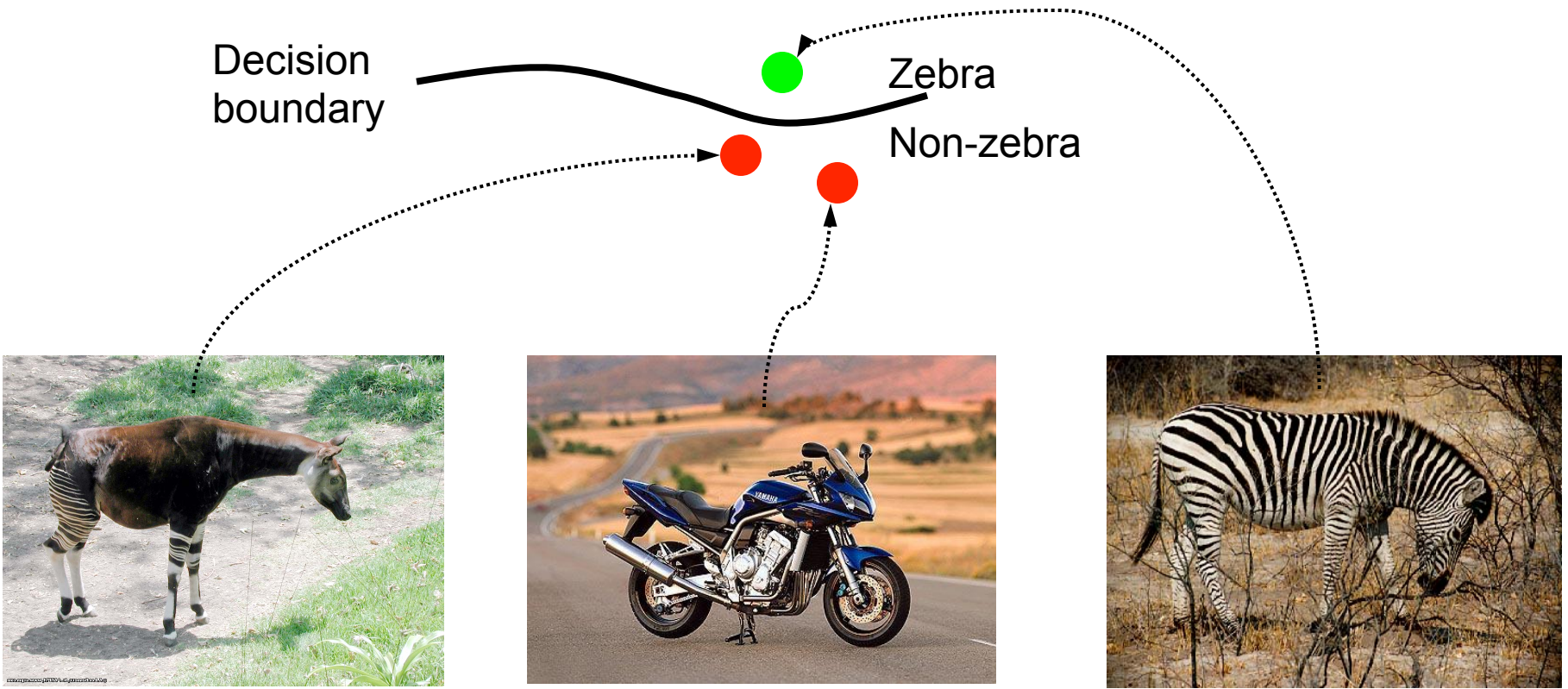


2. Discriminative method:
 - SVM



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

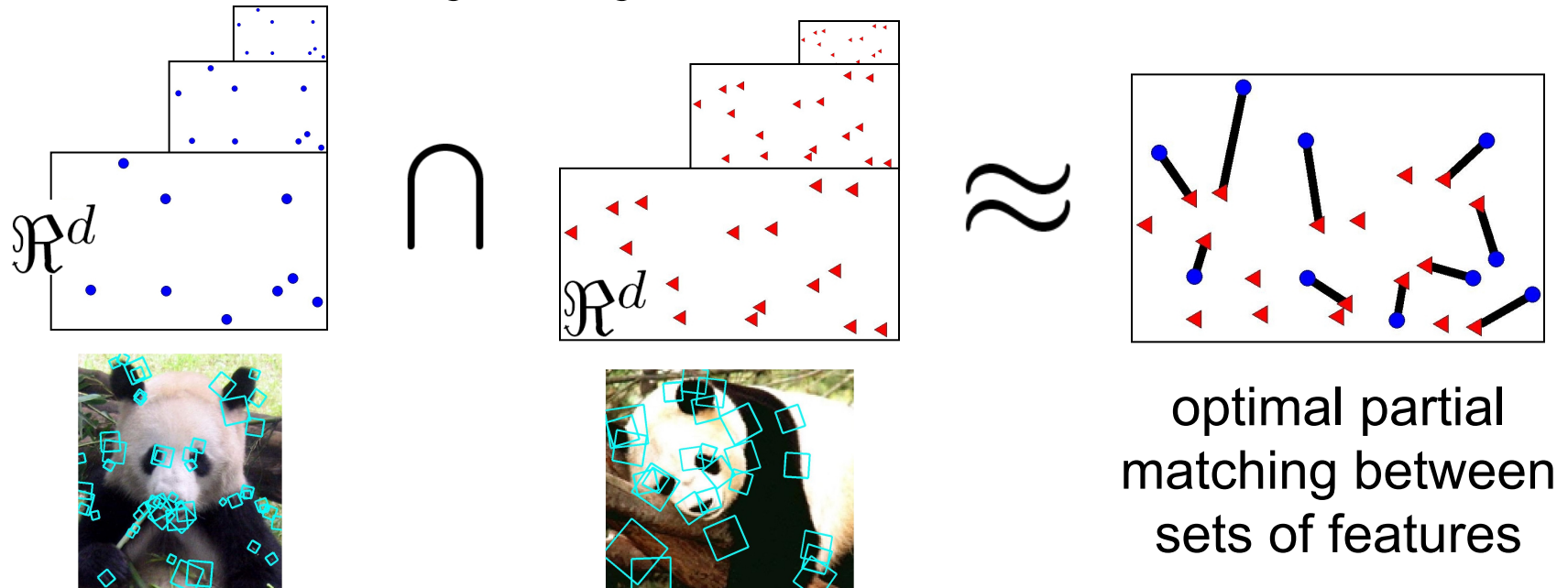
Discriminative methods based on 'bag of words' representation



Discriminative methods based on 'bag of words' representation

- Grauman & Darrell, 2005, 2006:
 - SVM w/ Pyramid Match kernels
- Others
 - Csurka, Bray, Dance & Fan, 2004
 - Serre & Poggio, 2005

Summary: Pyramid match kernel

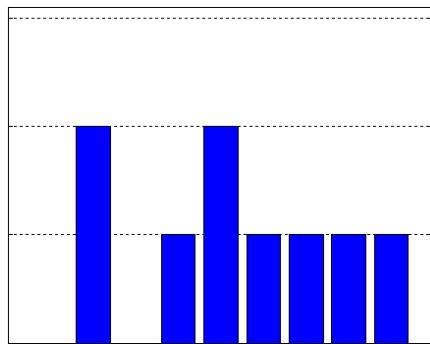
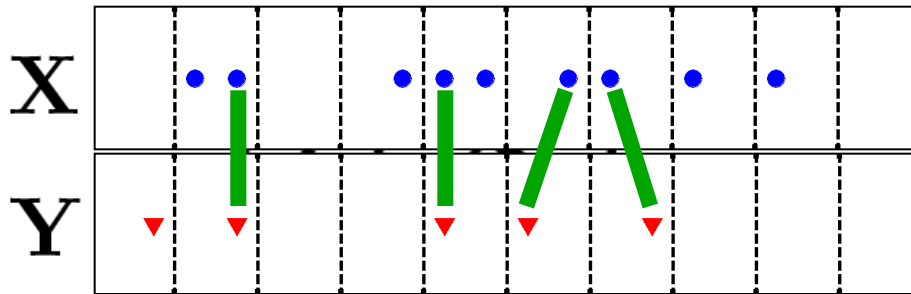


$$K_{\Delta} (\Psi(\mathbf{X}), \Psi(\mathbf{Y}))$$

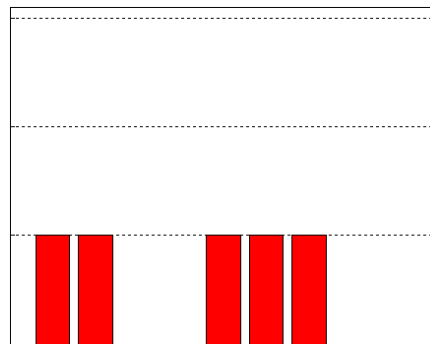
Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

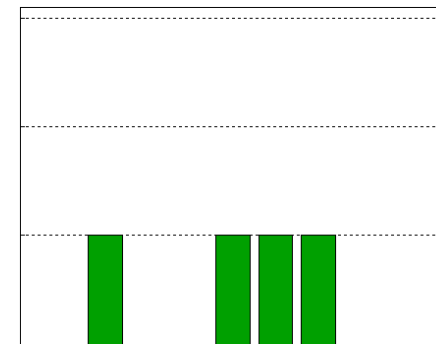
$$\mathcal{I}(H(\mathbf{X}), H(\mathbf{Y})) = \sum_{j=1}^r \min(H(\mathbf{X})_j, H(\mathbf{Y})_j)$$



$H(\mathbf{X})$



$H(\mathbf{Y})$



$$\mathcal{I}(H(\mathbf{X}), H(\mathbf{Y})) = 4$$

Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

$$\mathcal{I}(H(\mathbf{X}), H(\mathbf{Y})) = \sum_{j=1}^r \min(H(\mathbf{X})_j, H(\mathbf{Y})_j)$$

$$N_i = \overbrace{\mathcal{I}(H_i(\mathbf{X}), H_i(\mathbf{Y}))}^{\text{matches at this level}} - \overbrace{\mathcal{I}(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y}))}^{\text{matches at previous level}}$$

Difference in histogram intersections across levels counts *number of new pairs* matched

Pyramid match kernel

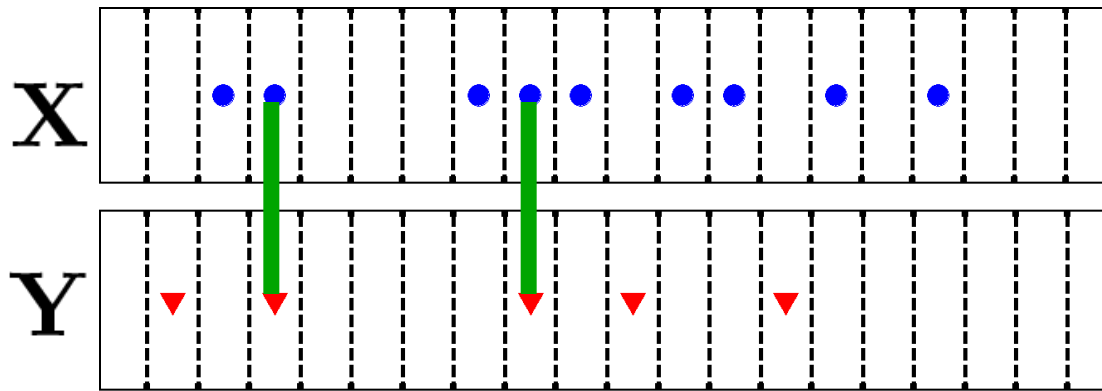
$$K_{\Delta} \left(\overbrace{\Psi(\mathbf{X}), \Psi(\mathbf{Y})}^{\text{histogram pyramids}} \right) = \sum_{i=0}^L \frac{1}{2^i} \left(\underbrace{\mathcal{I}(H_i(\mathbf{X}), H_i(\mathbf{Y})) - \mathcal{I}(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y}))}_{\text{number of newly matched pairs at level } i} \right)$$

↑
measure of difficulty of a match at level i

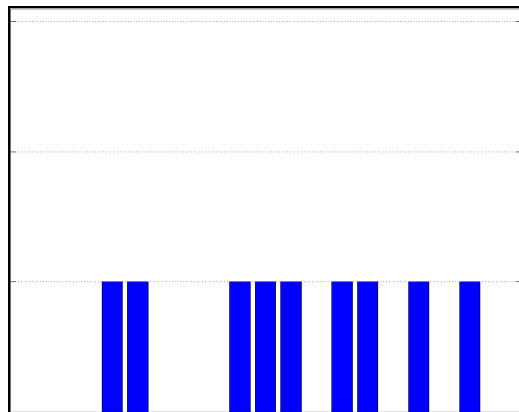
- Weights inversely proportional to bin size
- Normalize kernel values to avoid favoring large sets

Example pyramid match

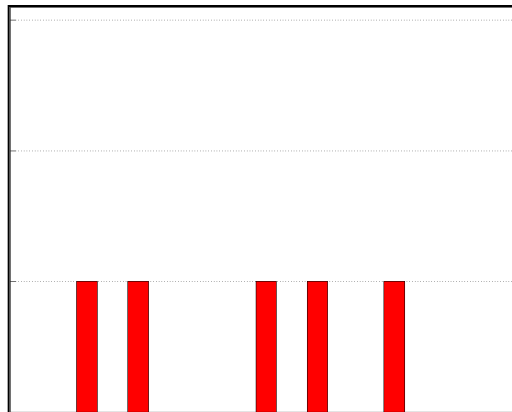
Level 0



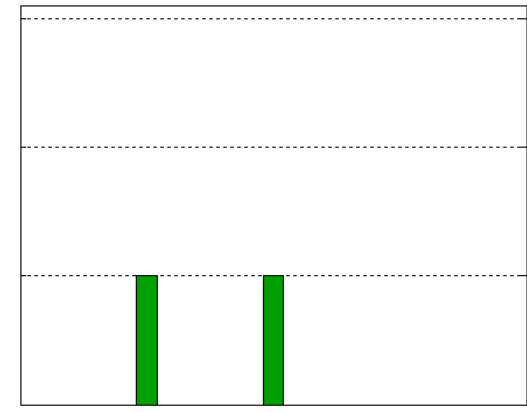
$$\begin{aligned} N_0 &= 2 \\ w_0 &= 1 \end{aligned}$$



$H_0(\mathbf{X})$



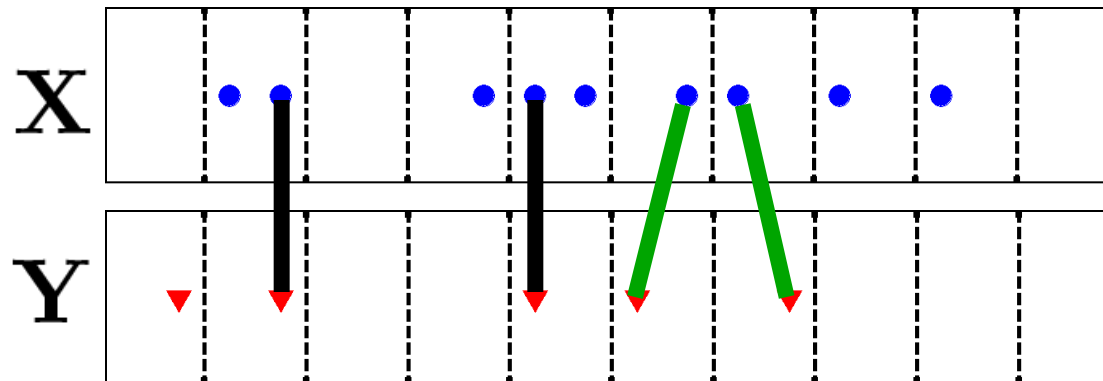
$H_0(\mathbf{Y})$



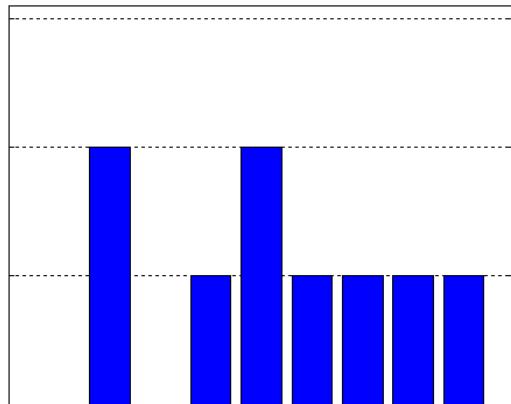
$\mathcal{I}_0 = 2$

Example pyramid match

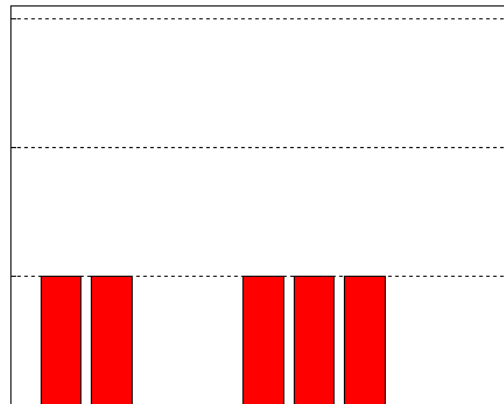
Level 1



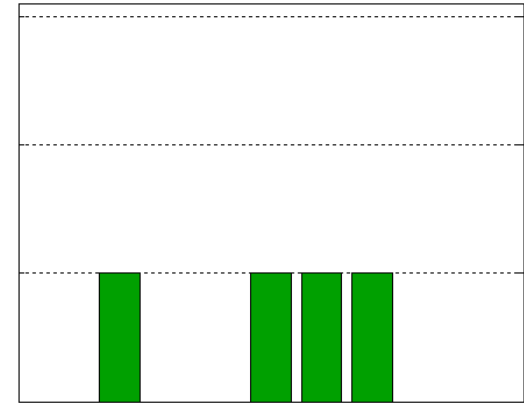
$$\begin{aligned} N_1 &= 4 - 2 = 2 \\ w_1 &= \frac{1}{2} \end{aligned}$$



$H_1(\mathbf{X})$



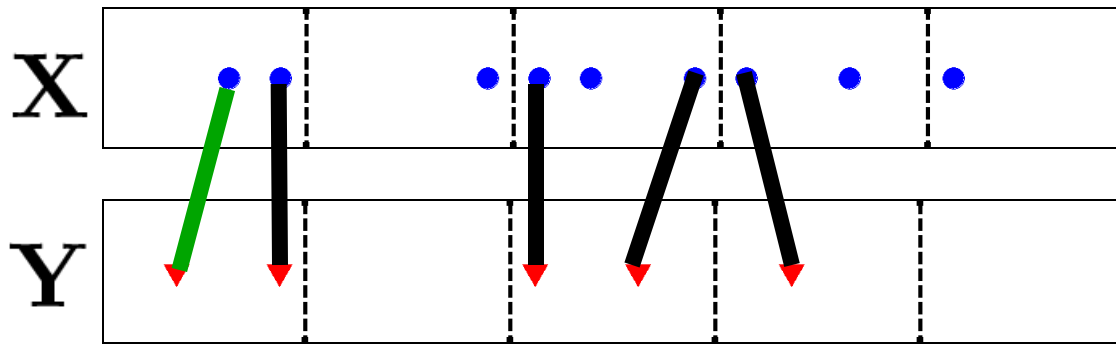
$H_1(\mathbf{Y})$



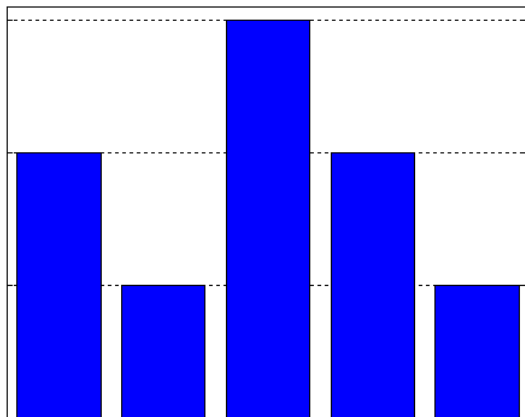
$\mathcal{I}_1 = 4$

Example pyramid match

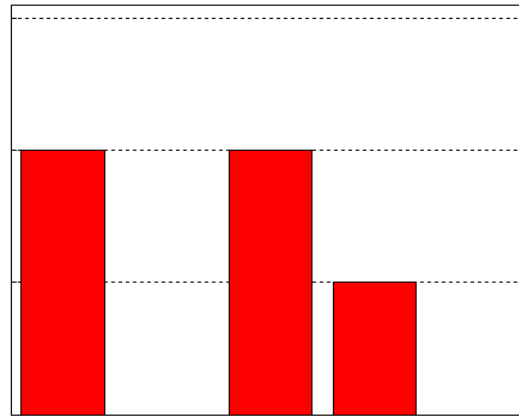
Level 2



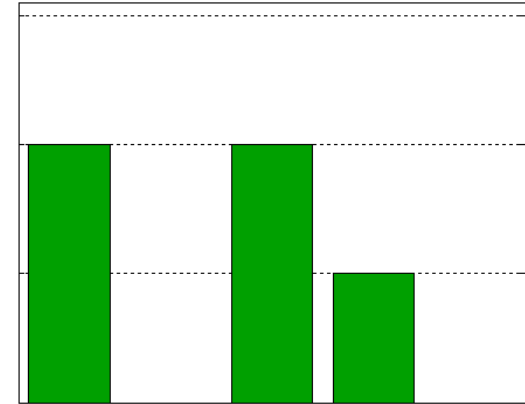
$$\begin{aligned} N_2 &= 5 - 4 = 1 \\ \rightarrow w_2 &= \frac{1}{4} \end{aligned}$$



$H_2(\mathbf{X})$



$H_2(\mathbf{Y})$

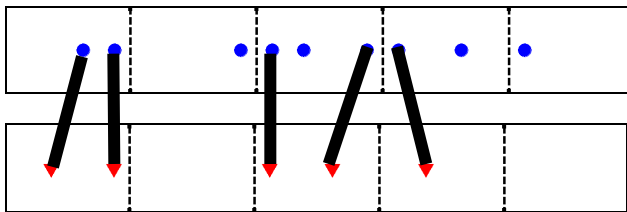


$\mathcal{I}_2 = 5$

Slide credit: Kristen Grauman

Example pyramid match

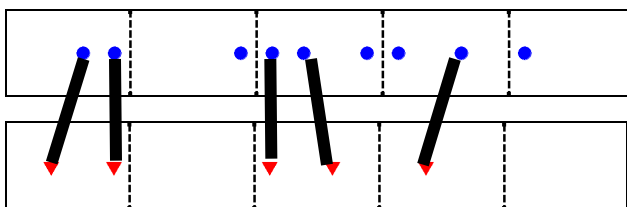
pyramid match



$$K_{\Delta} = \sum_{i=0}^L w_i N_i$$

$$= 1(2) + \frac{1}{2}(2) + \frac{1}{4}(1) = 3.25$$

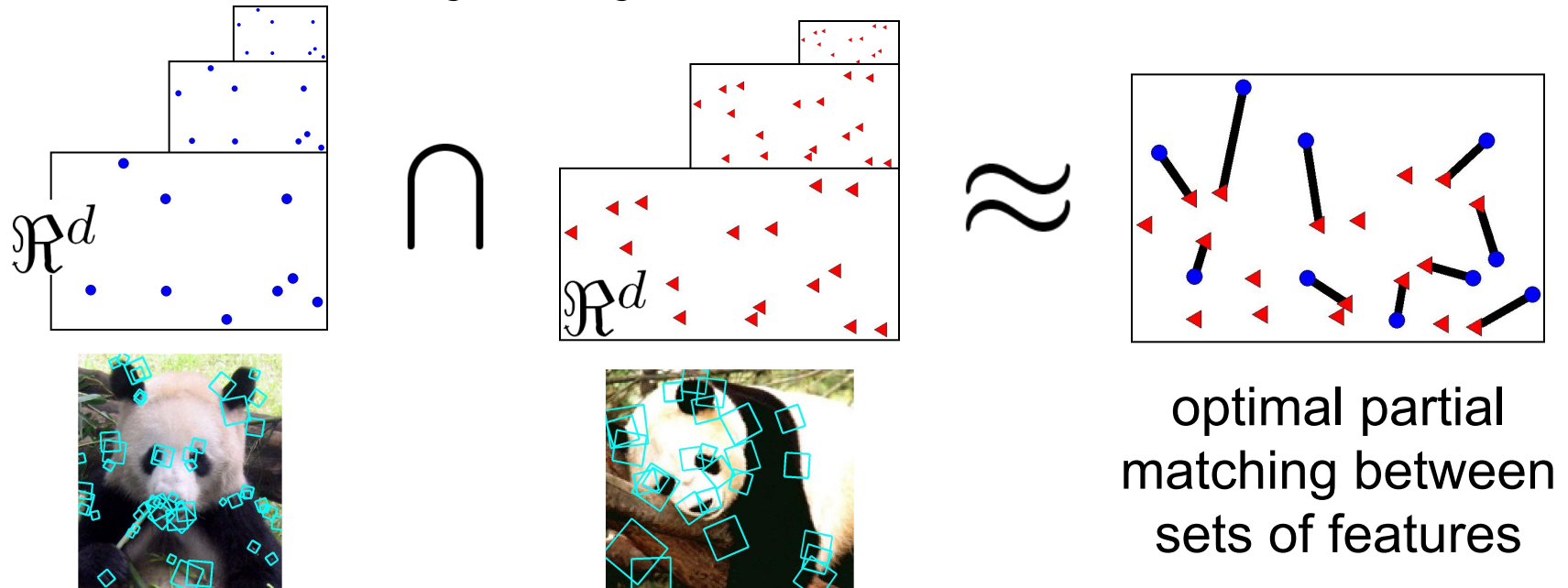
optimal match



$$K = \max_{\pi: \mathbf{X} \rightarrow \mathbf{Y}} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathcal{S}(\mathbf{x}_i, \pi(\mathbf{x}_i))$$

$$= 1(2) + \frac{1}{2}(3) = 3.5$$

Summary: Pyramid match kernel



$$K_{\Delta} (\Psi(\mathbf{X}), \Psi(\mathbf{Y})) = \sum_{i=0}^L \frac{1}{2^i} \left(\mathcal{I}(H_i(\mathbf{X}), H_i(\mathbf{Y})) - \mathcal{I}(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})) \right)$$

difficulty of a match at level i

number of new matches at level i

Object recognition results

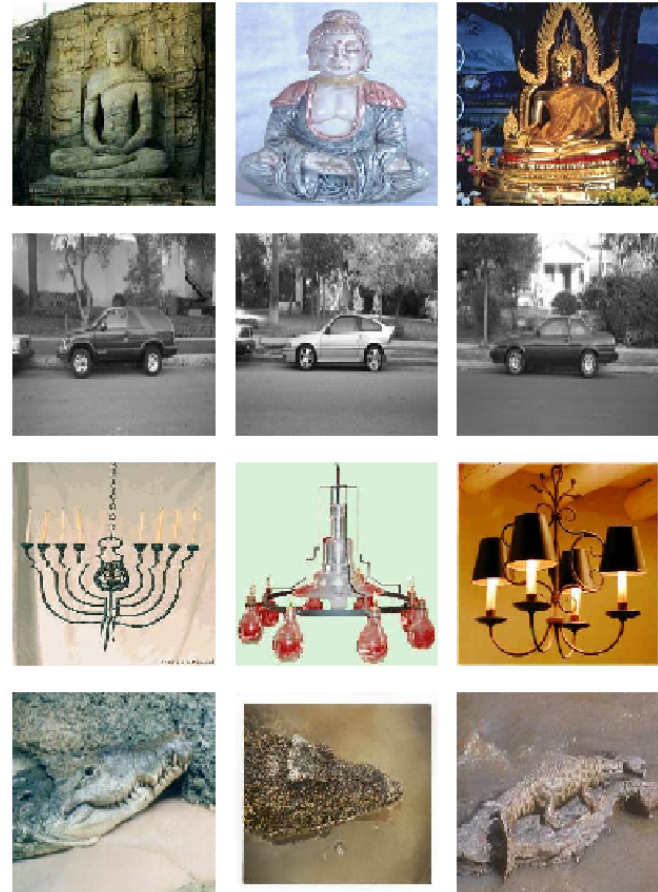
- ETH-80 database
8 object classes
(Eichhorn and Chapelle 2004)
- Features:
 - Harris detector
 - PCA-SIFT descriptor, $d=10$



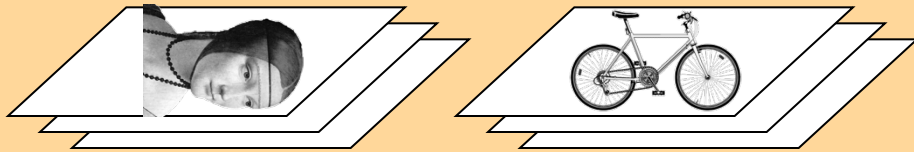
Kernel	Complexity	Recognition rate
Match [Wallraven et al.]	$O(dm^2)$	84%
Bhattacharyya affinity [Kondor & Jebara]	$O(dm^3)$	85%
Pyramid match	$O(dmL)$	84%

Object recognition results

- Caltech objects database
101 object classes
- Features:
 - SIFT detector
 - PCA-SIFT descriptor, $d=10$
- 30 training images / class
- **43% recognition rate**
(1% chance performance)
- 0.002 seconds per match



learning



feature detection
& representation

codewords dictionary

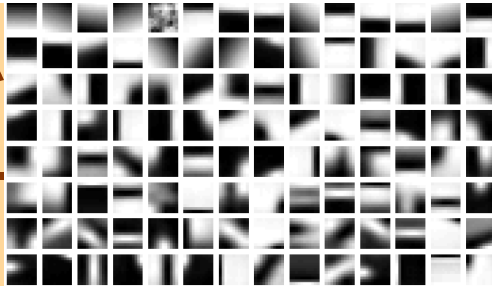
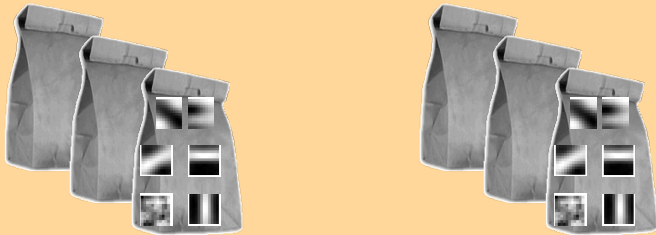


image representation



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

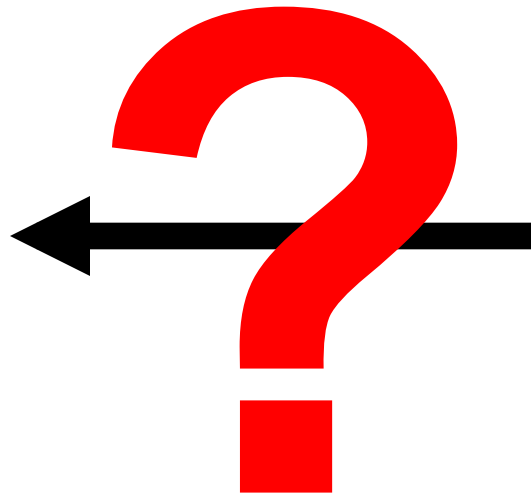
recognition



**category
decision**



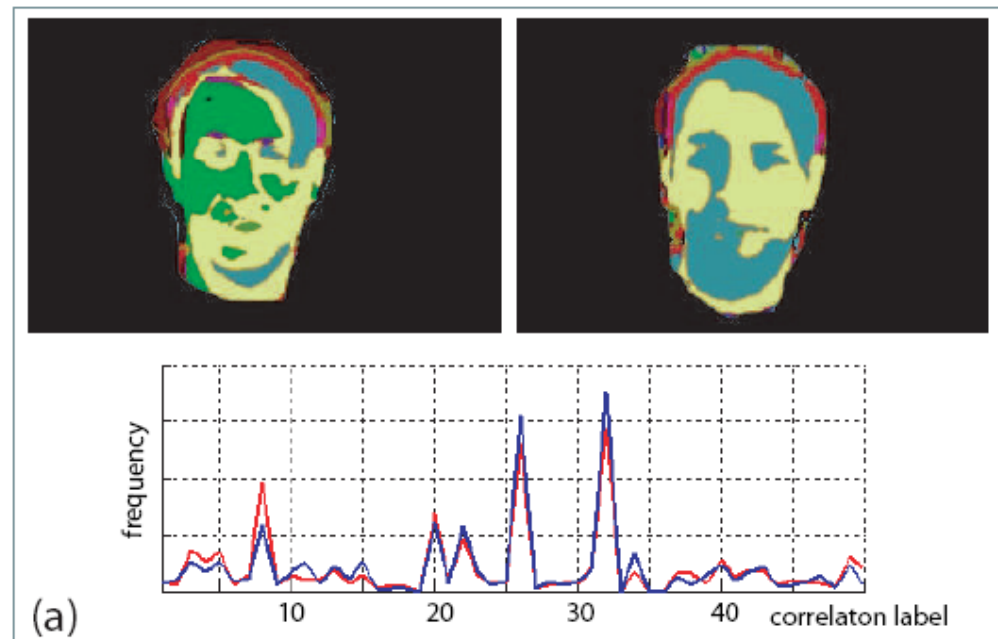
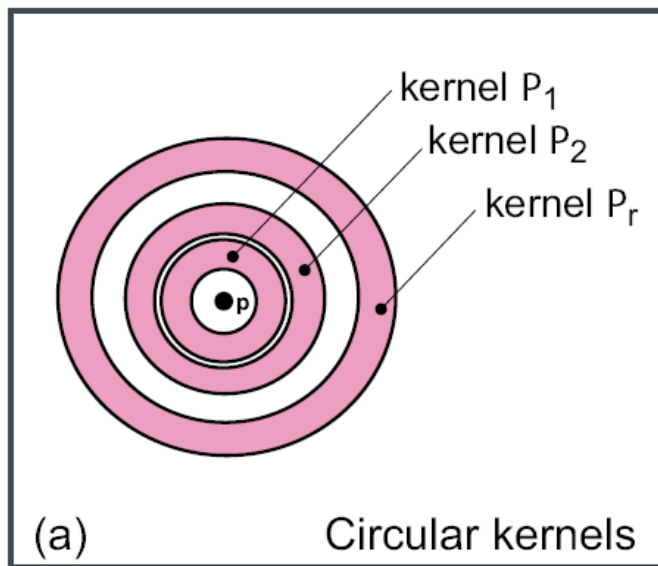
What about spatial info?



What about spatial info?



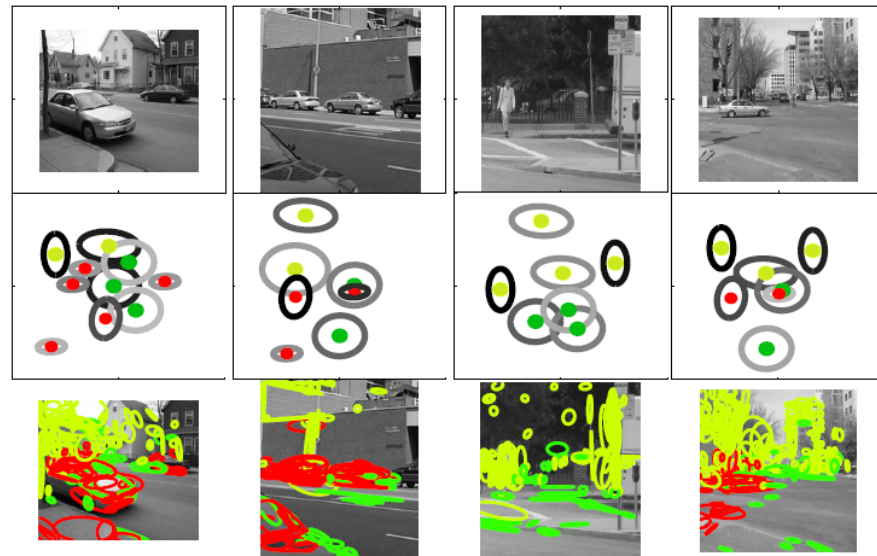
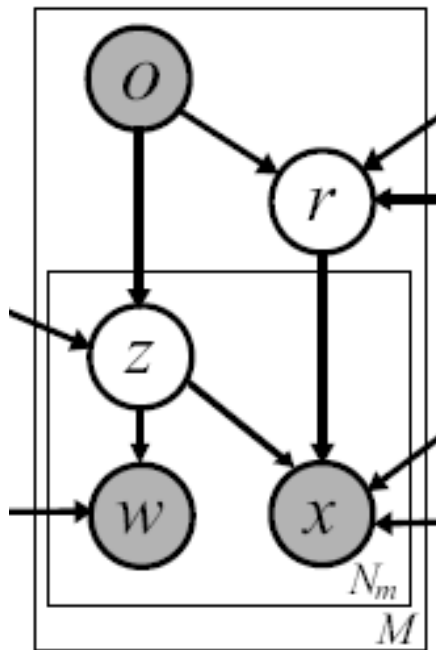
- Feature level
 - Spatial influence through correlogram features:
Savarese, Winn and Criminisi, CVPR 2006



What about spatial info?



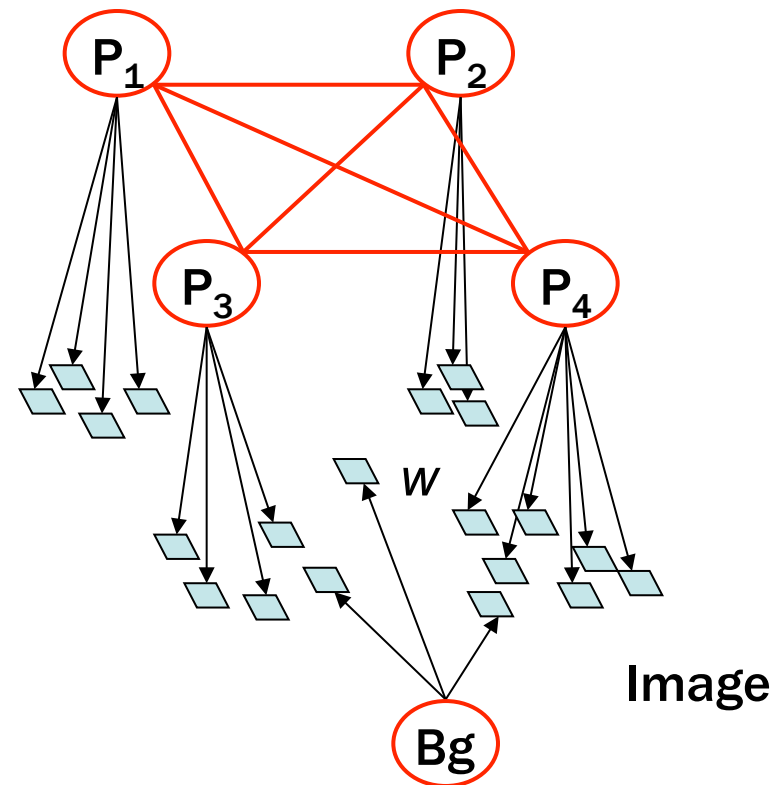
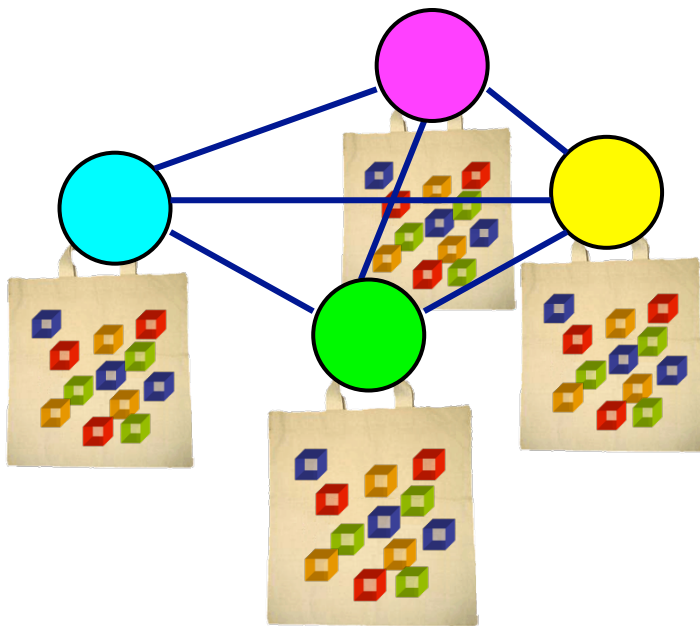
- Feature level
- Generative models
 - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
 - Niebles & Fei-Fei, CVPR 2007



What about spatial info?



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



What about spatial info?



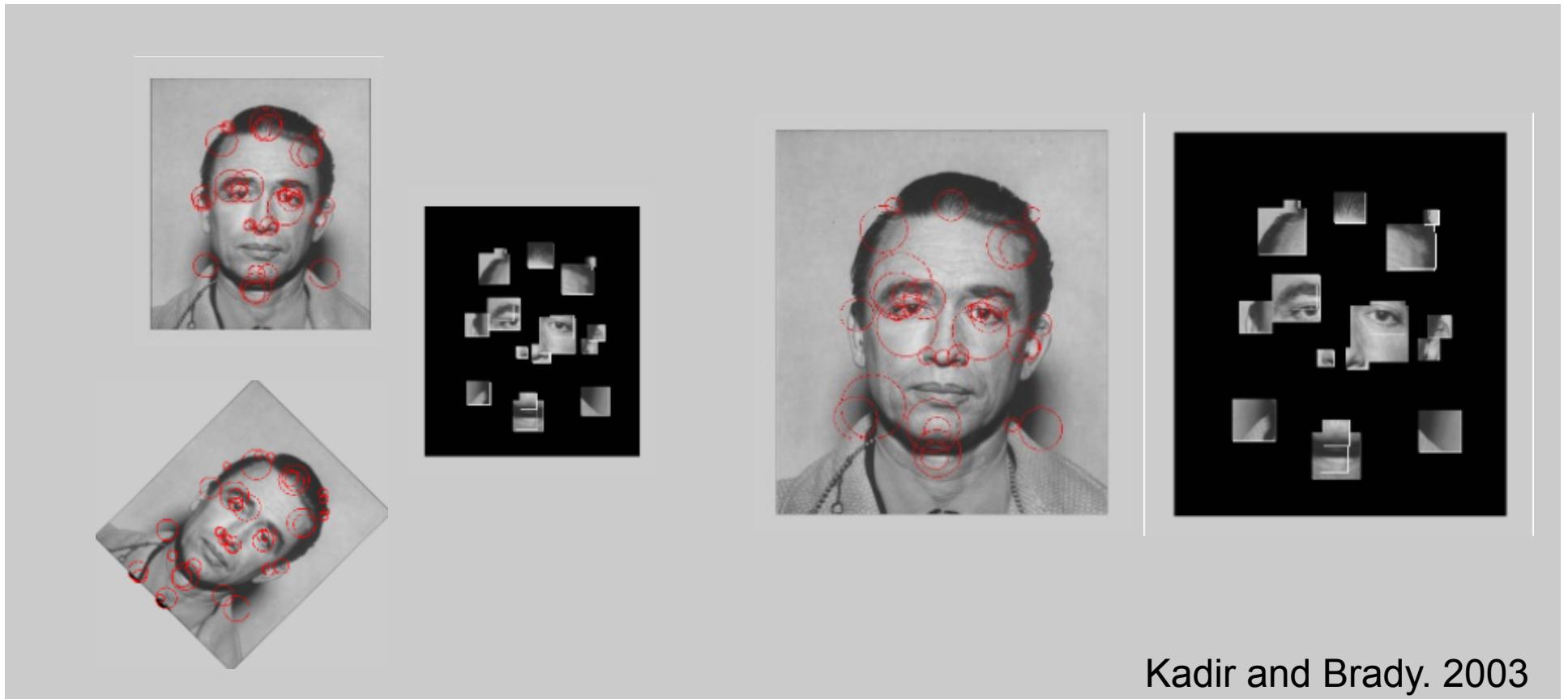
- Feature level
- Generative models
- Discriminative methods
 - Lazebnik, Schmid & Ponce, 2006



Invariance issues



- Scale and rotation
 - Implicit
 - Detectors and descriptors



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

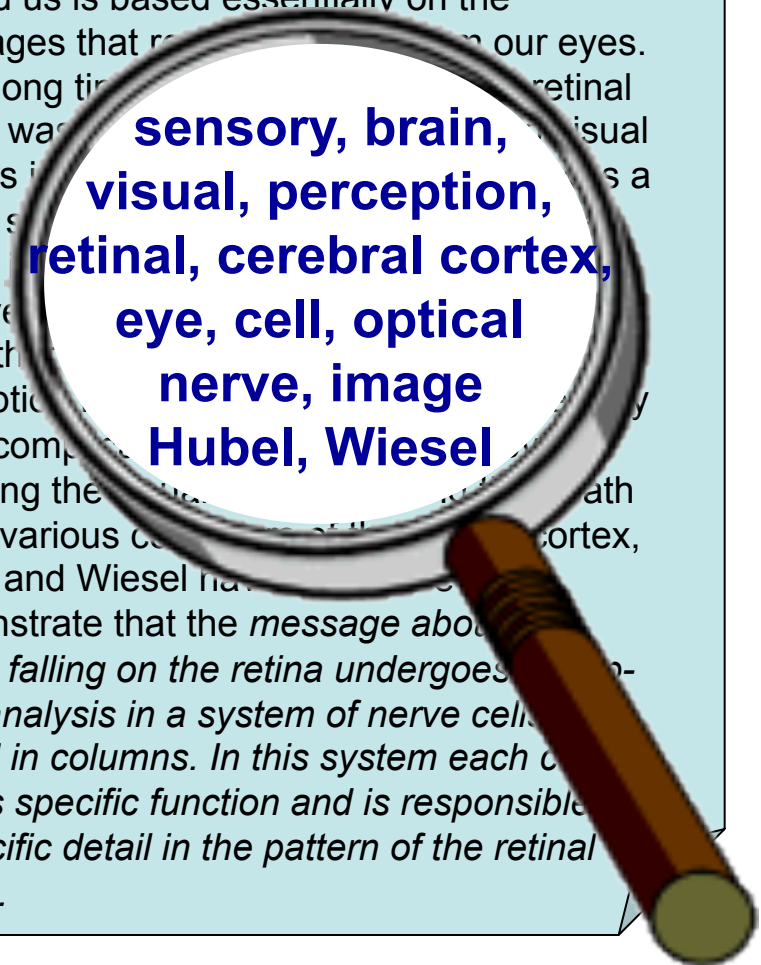




Model properties

- 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 our eyes. For a long time, the retinal image was considered as a simple image of the world. It was a movie screen. The visual centers in the brain were discovered only after the discovery of the retinal image. We now know that the visual perception is a more complex process. Following the path to the various cells of the 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.

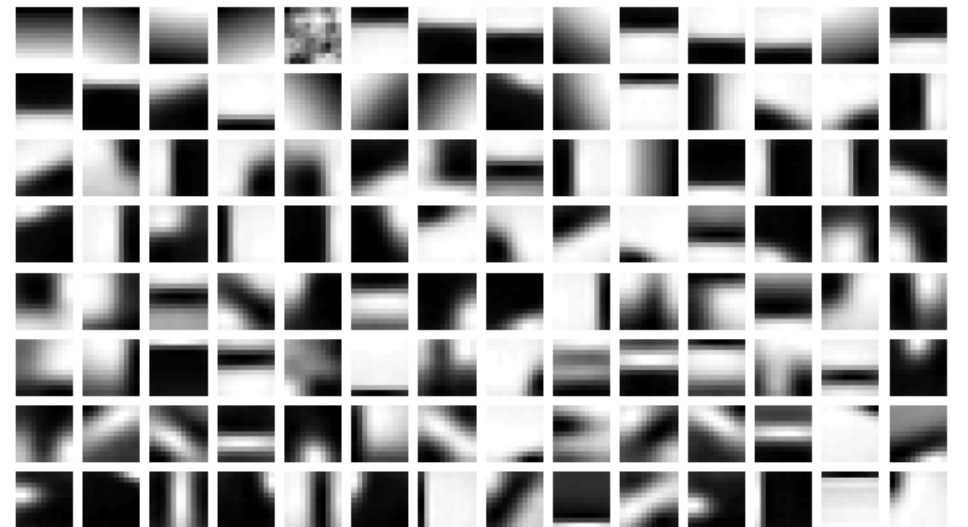
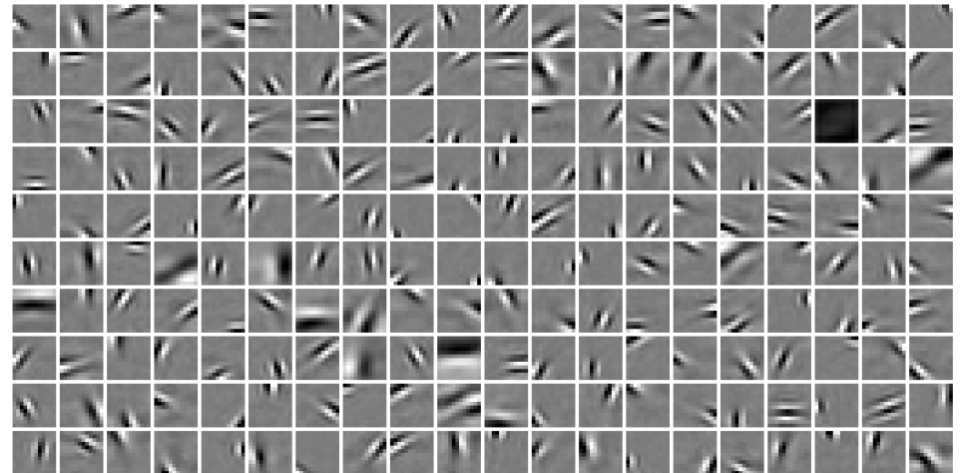


**sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel**

Model properties

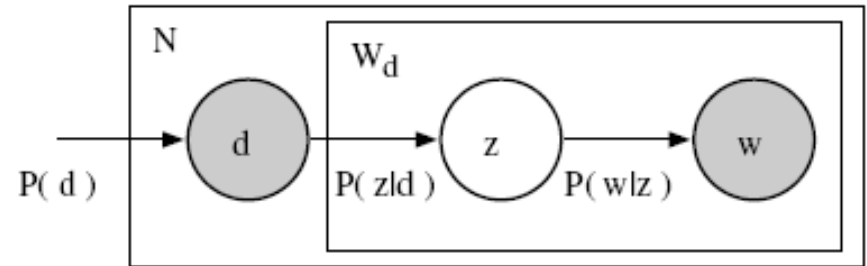


- Intuitive
 - Analogy to documents
 - Analogy to human vision



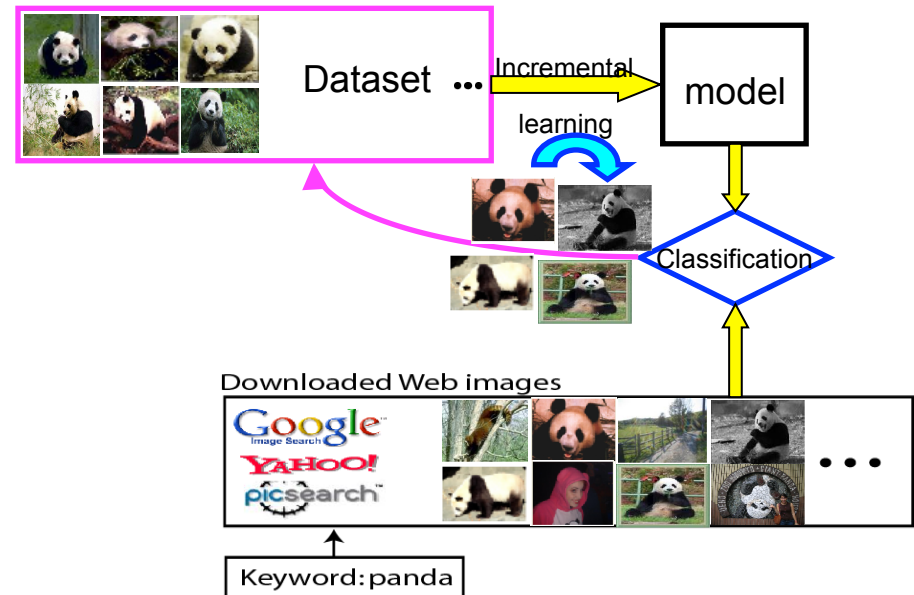


Model properties



Sivic, Russell, Efros, Freeman, Zisserman, 2005

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

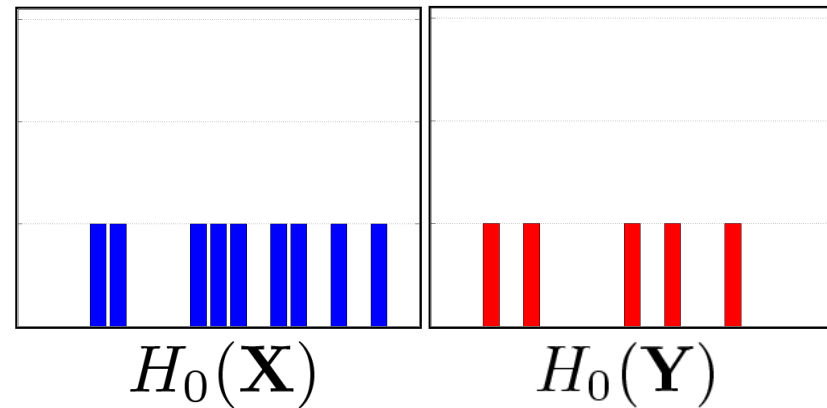
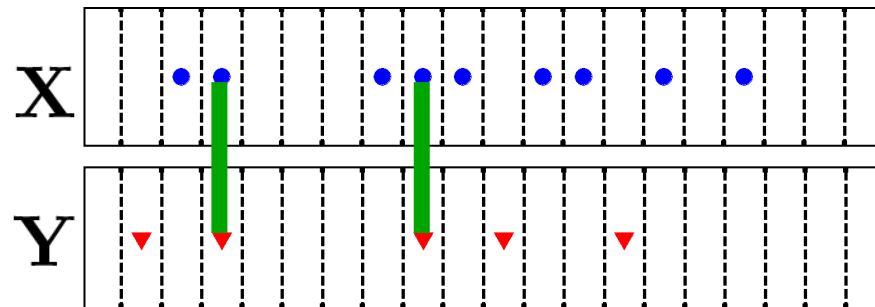


Li, Wang & Fei-Fei, CVPR 2007



Model properties

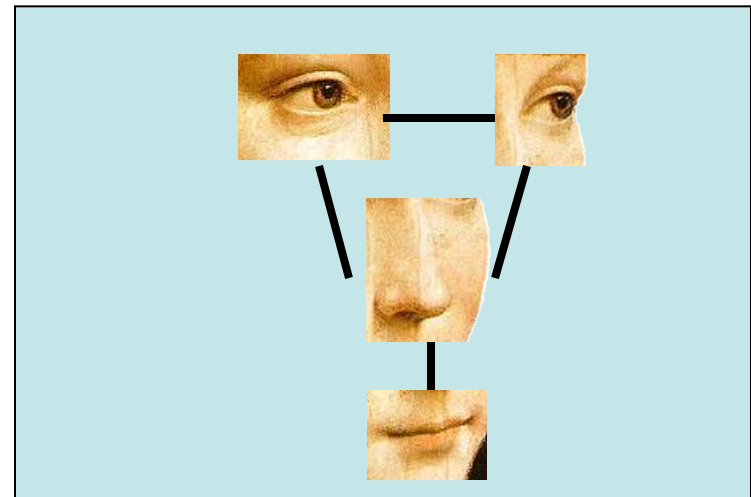
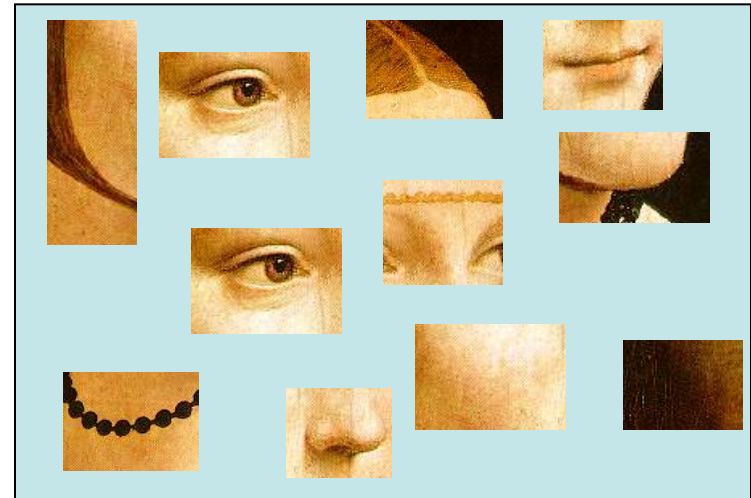
- Intuitive
- generative models
- Discriminative method
 - Computationally efficient



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