Part 1: Bag-of-words models

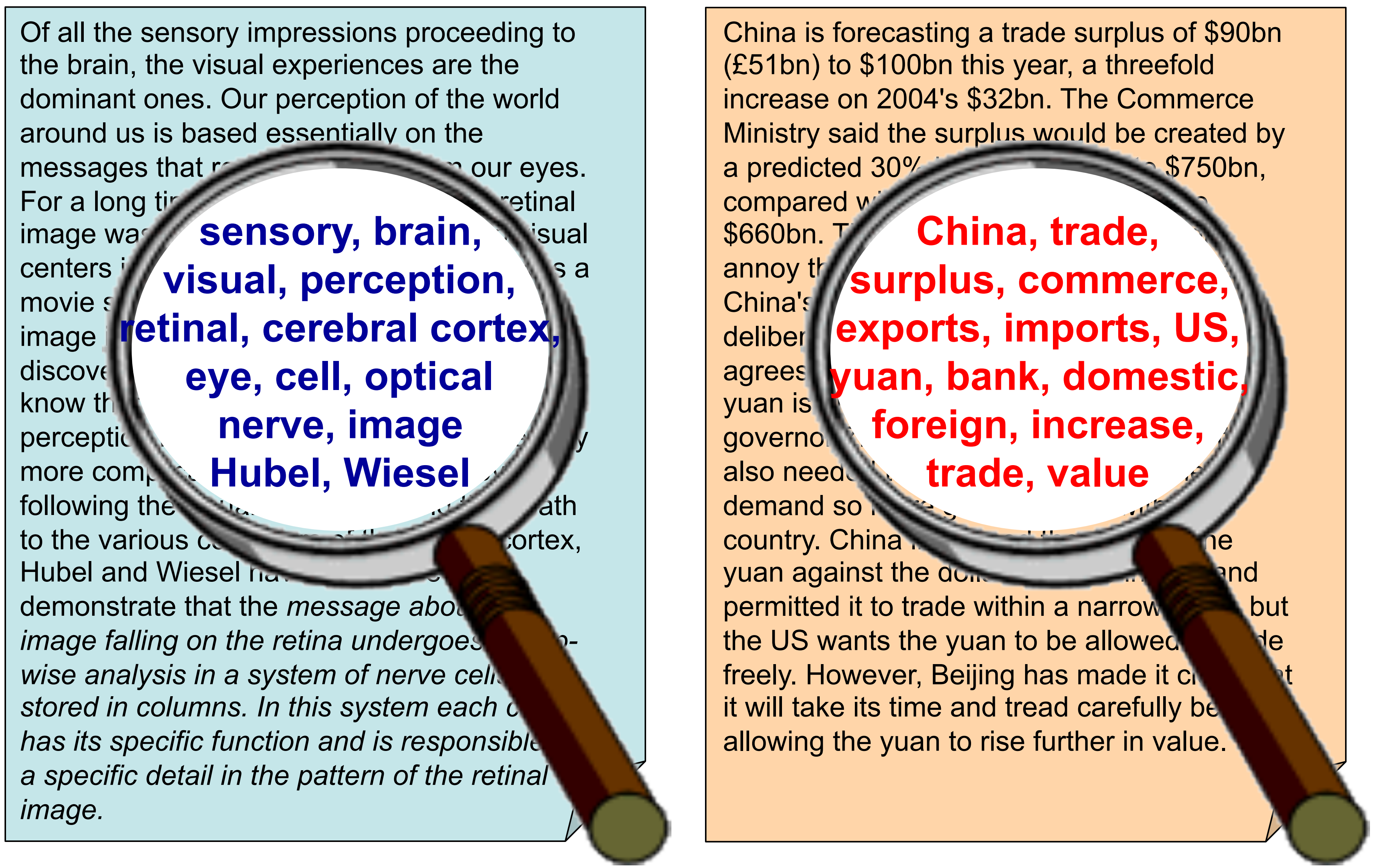
by Li Fei-Fei (Princeton)

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

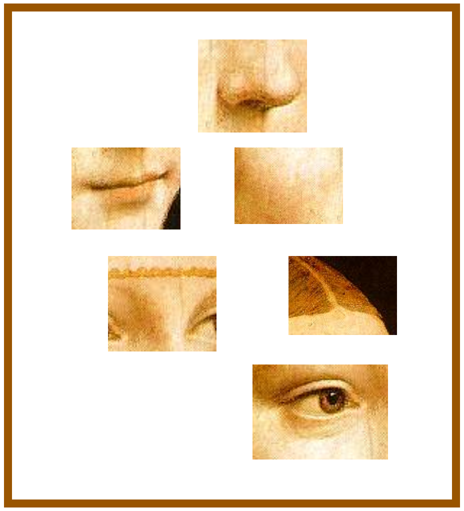
Object → Bag of ‘words’

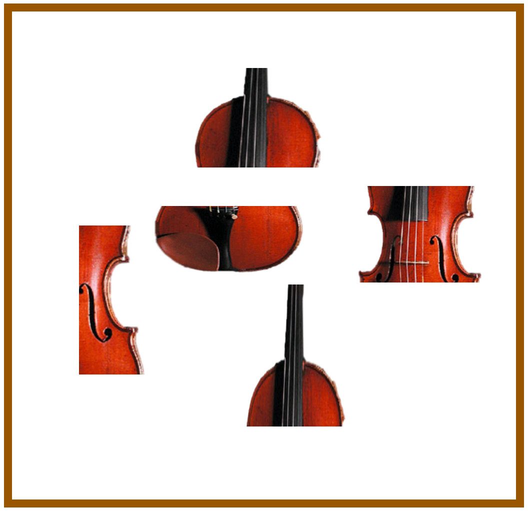
Analogy to documents



A clarification: definition of "BoW"

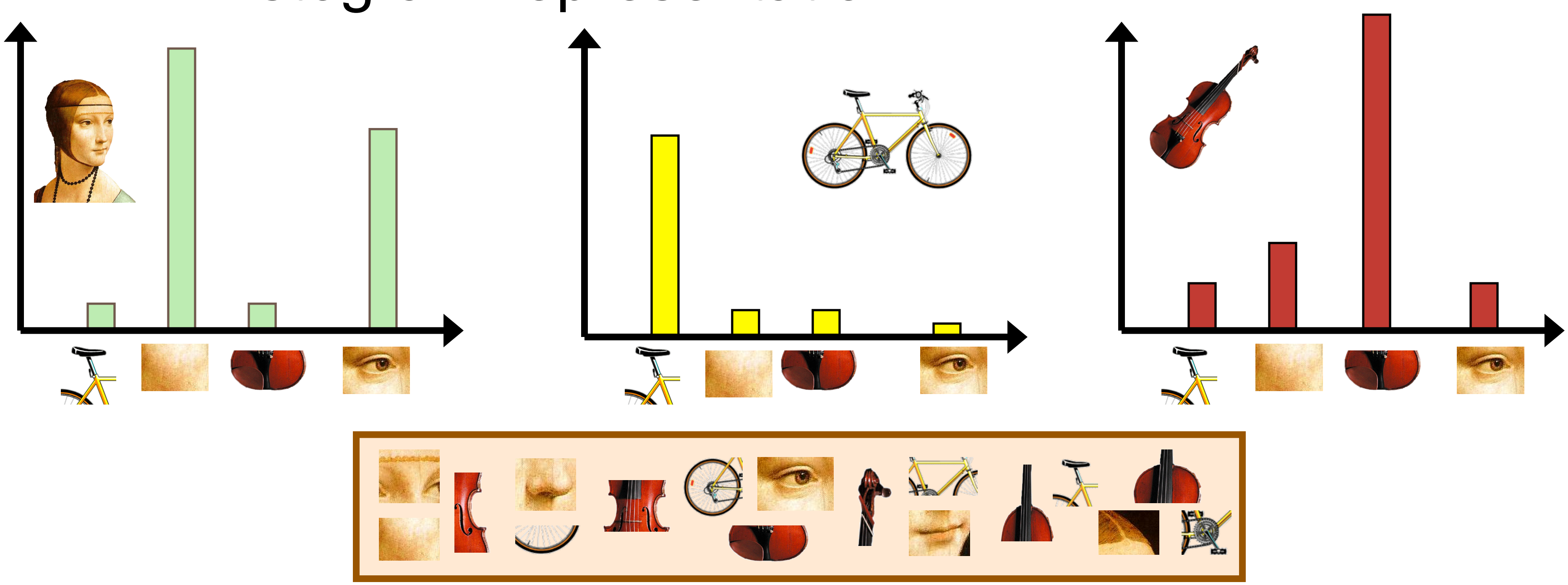
- Looser definition
  - Independent features

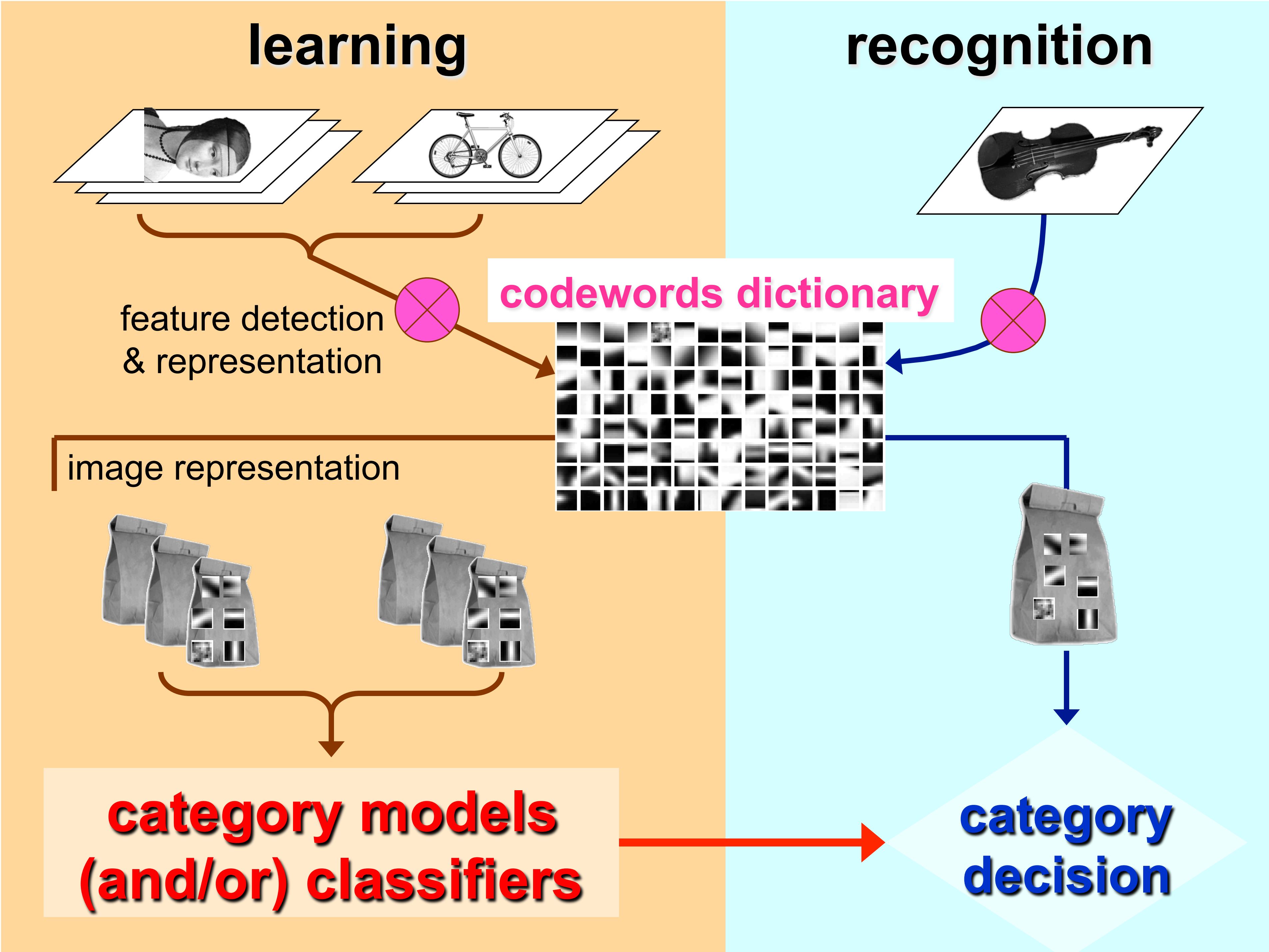




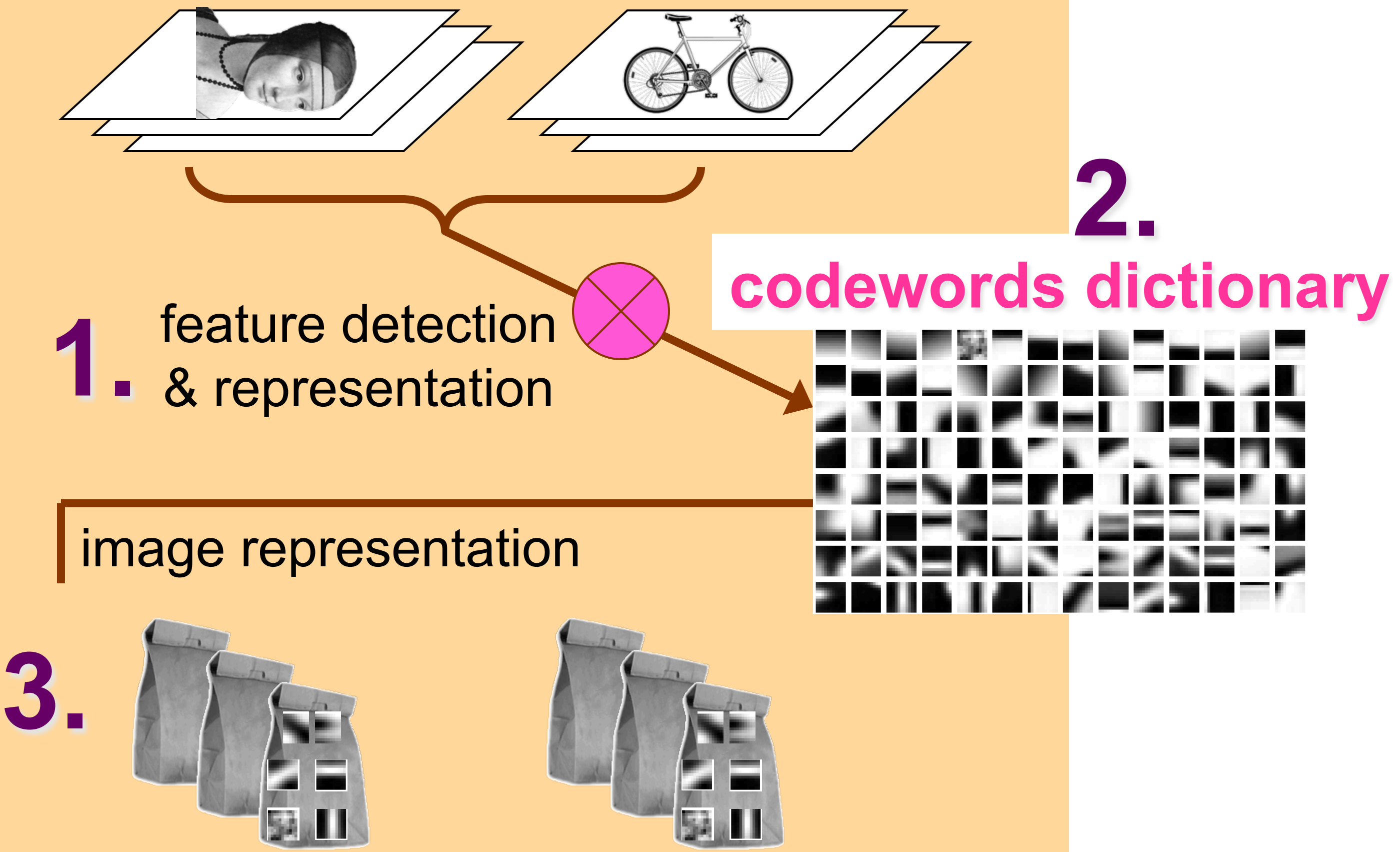
A clarification: definition of "BoW"

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

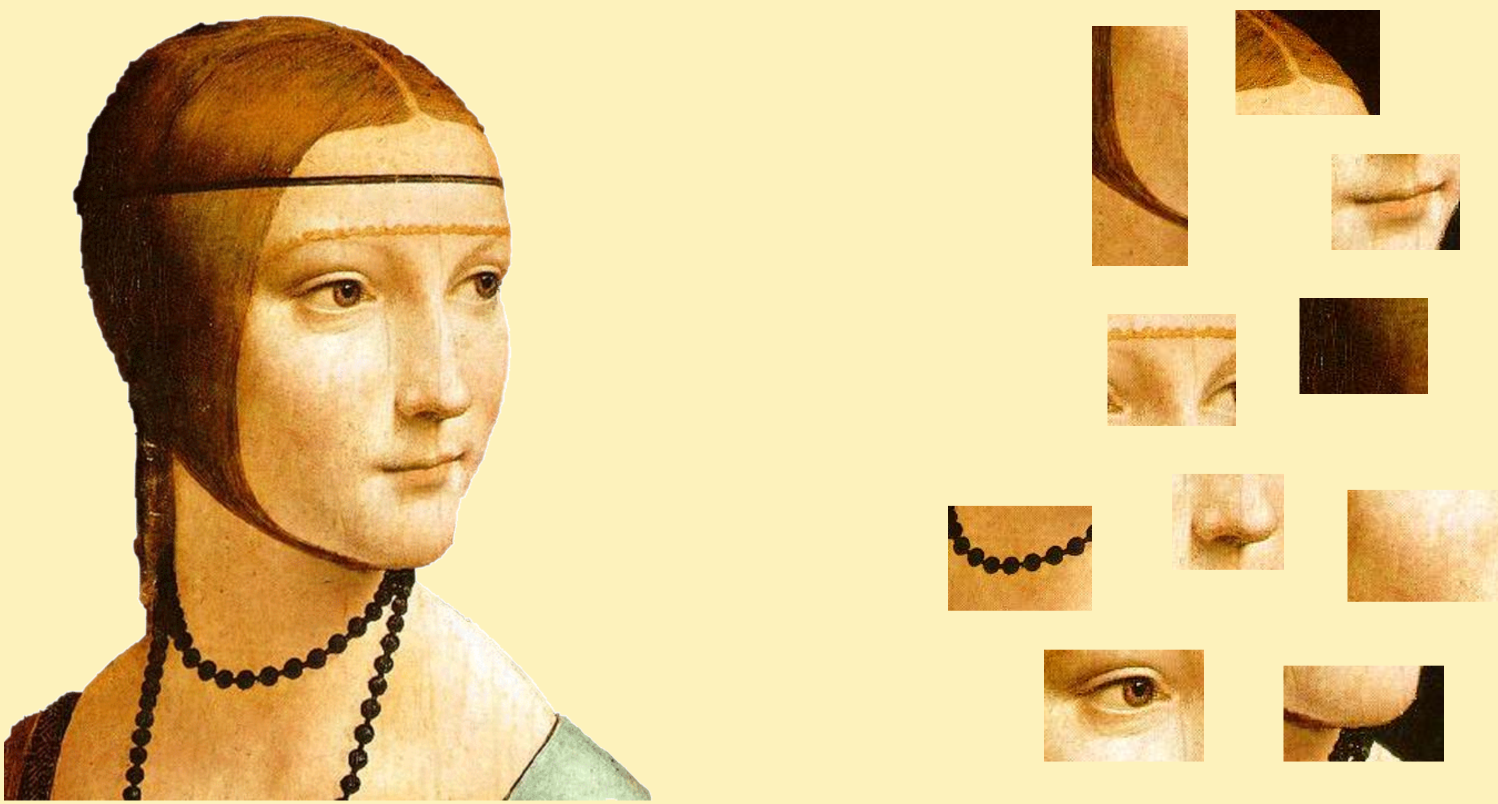



learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

Representation

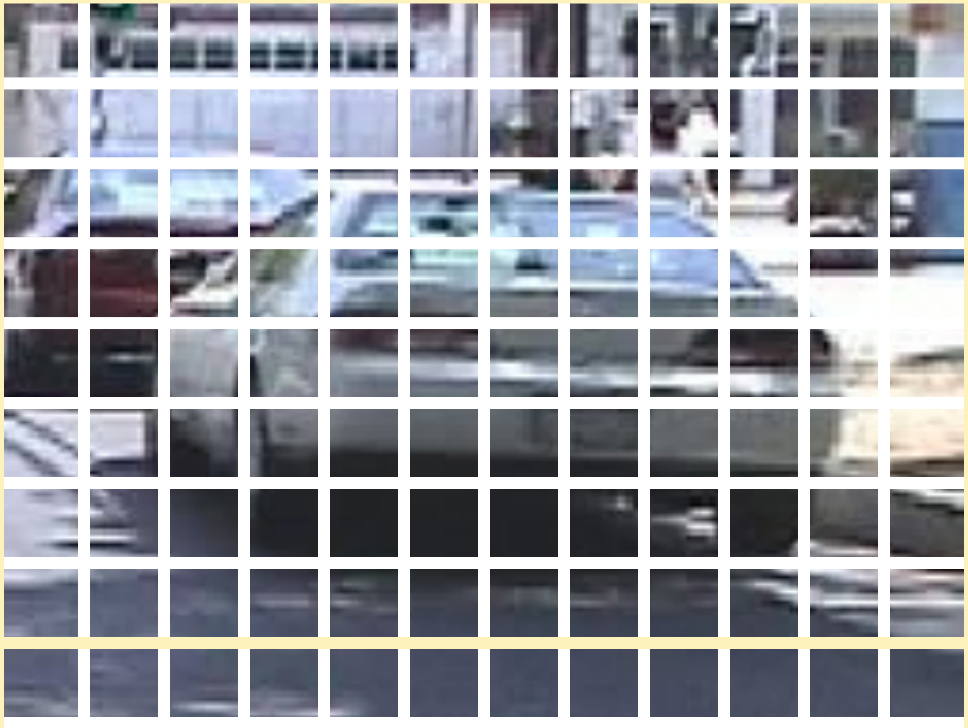


1.Feature detection and representation



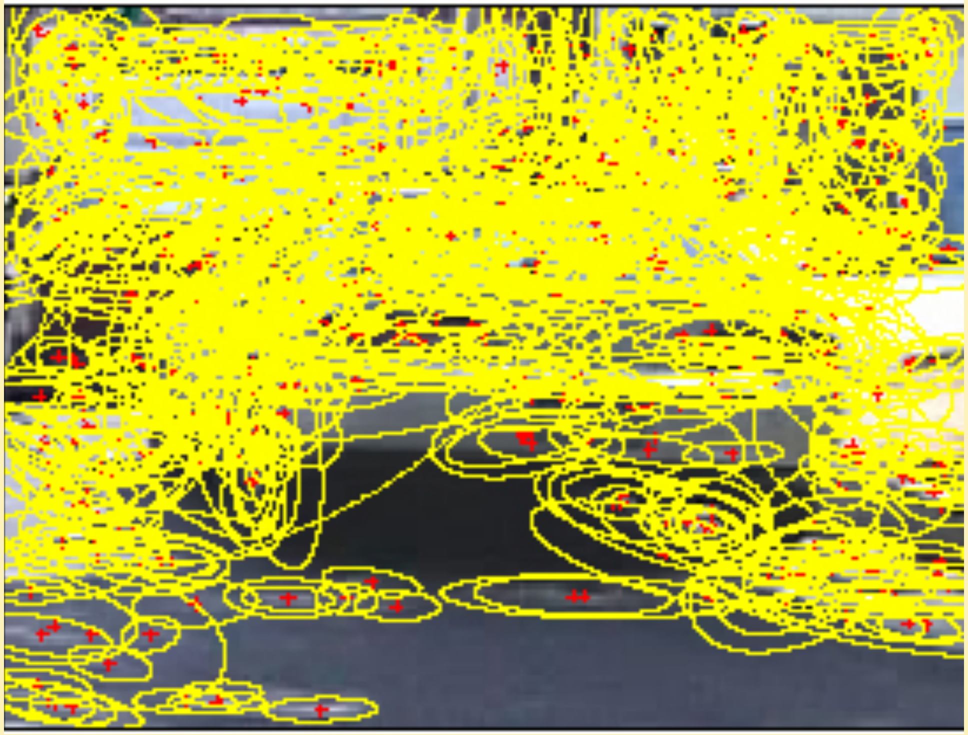
1.Feature detection and representation

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



1.Feature detection and representation

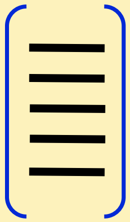
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic, et al. 2005



1.Feature detection and representation

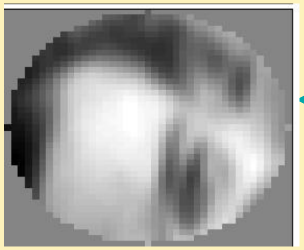
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, Bray, Dance & Fan, 2004
  - Fei-Fei & Perona, 2005
  - Sivic, Russell, Efros, Freeman & Zisserman, 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation based patches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

1.Feature detection and representation

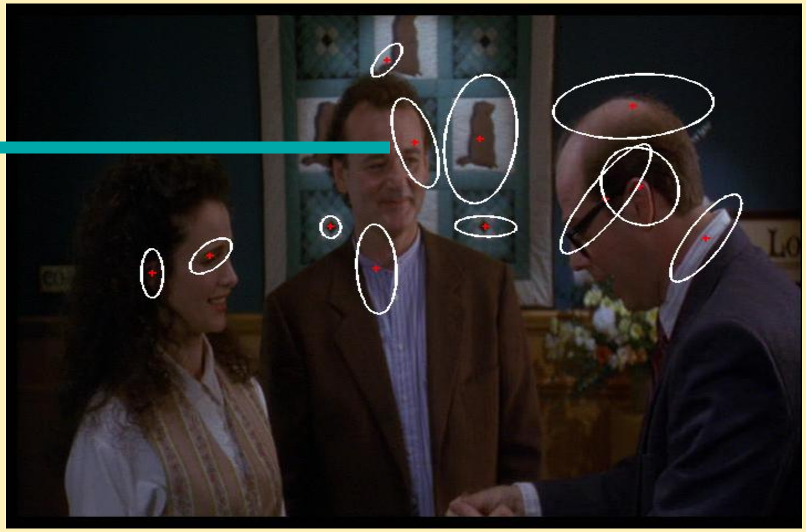


Compute SIFT descriptor

[Lowe’99]



Normalize patch



Detect patches

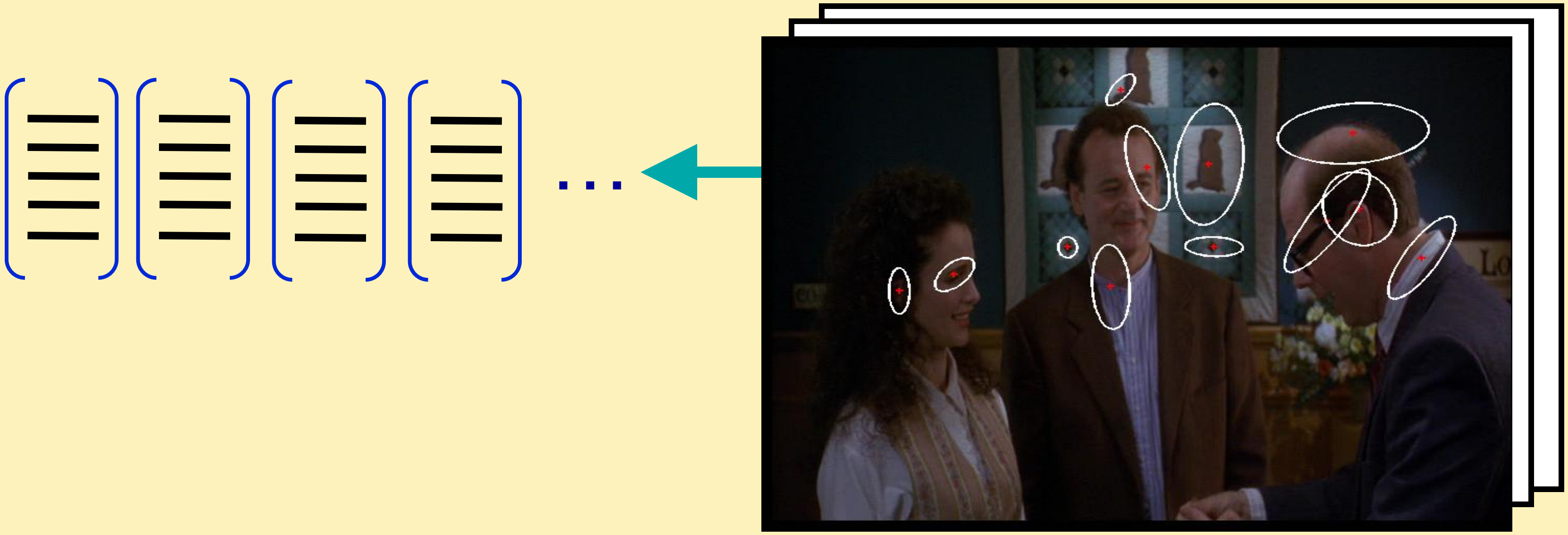
[Mikojaczyk and Schmid ’02]

[Mata, Chum, Urban & Pajdla, ’02]

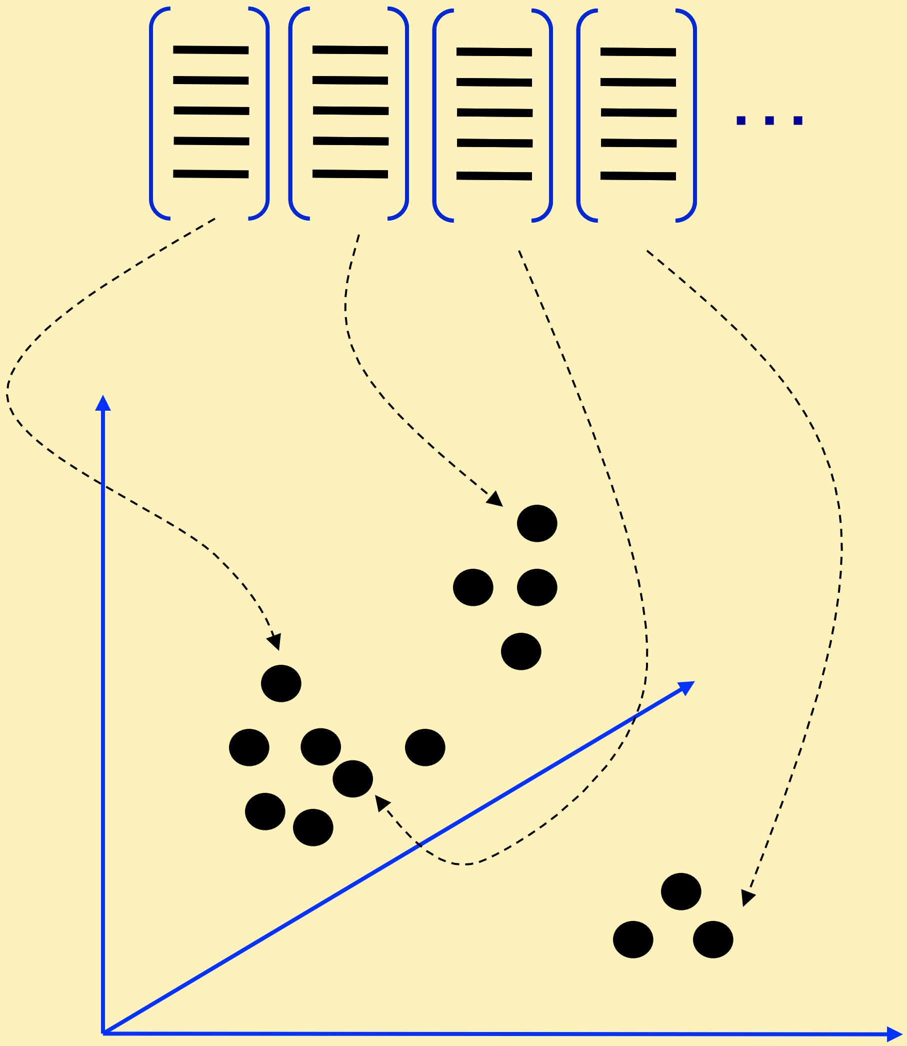
[Sivic & Zisserman, ’03]

Slide credit: Josef Sivic

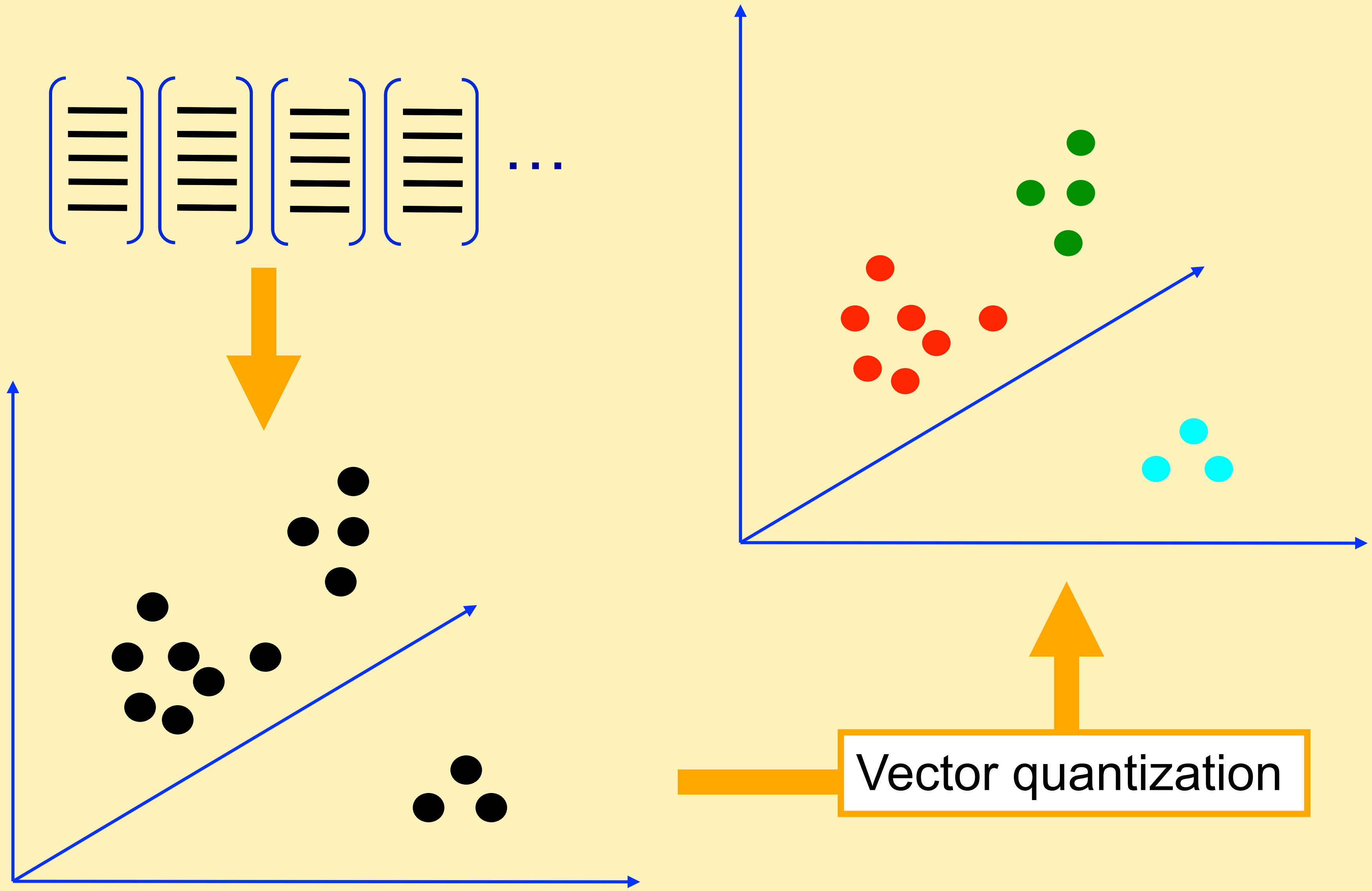
1.Feature detection and representation



2. Codewords dictionary formation

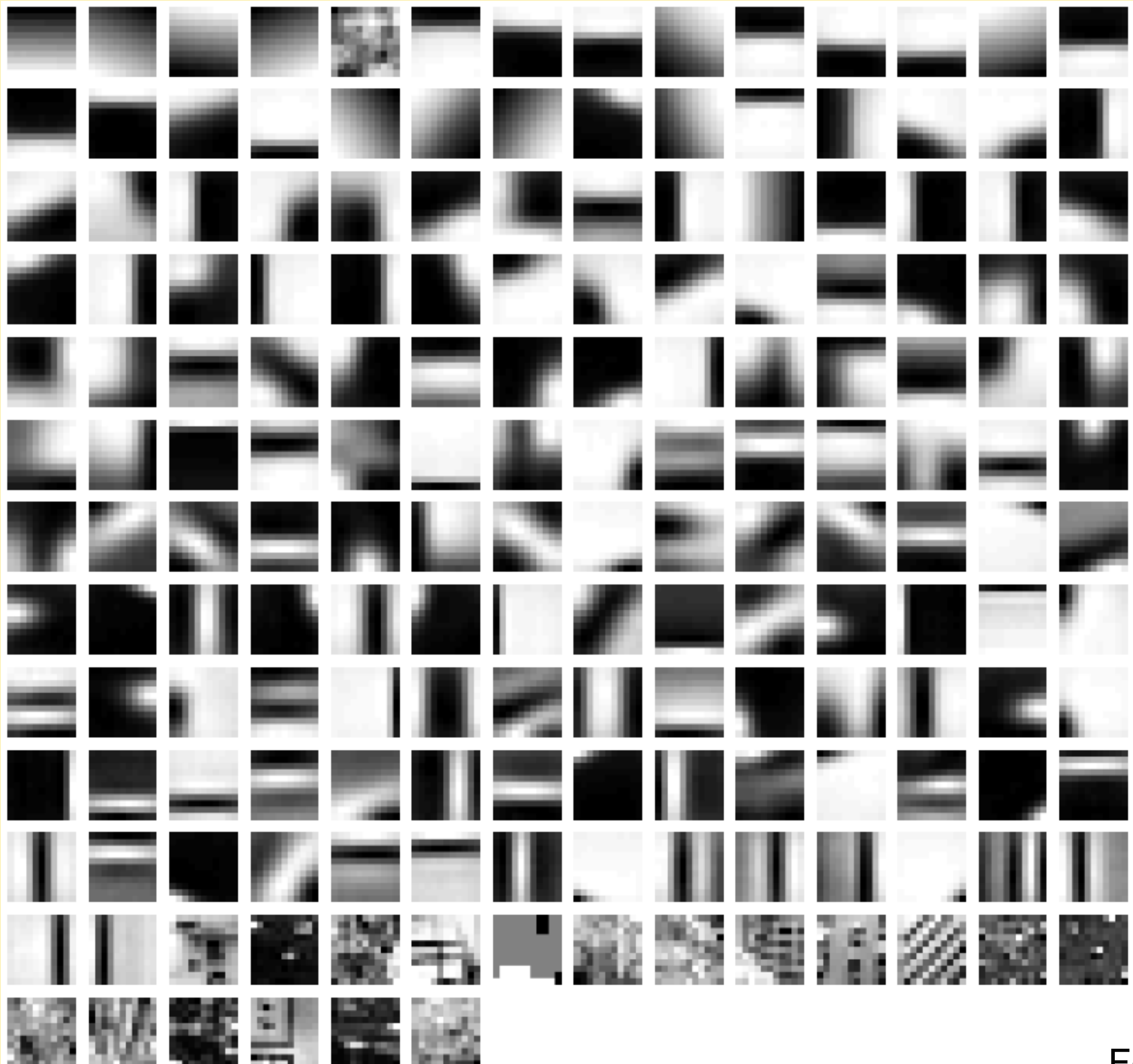


2. Codewords dictionary formation



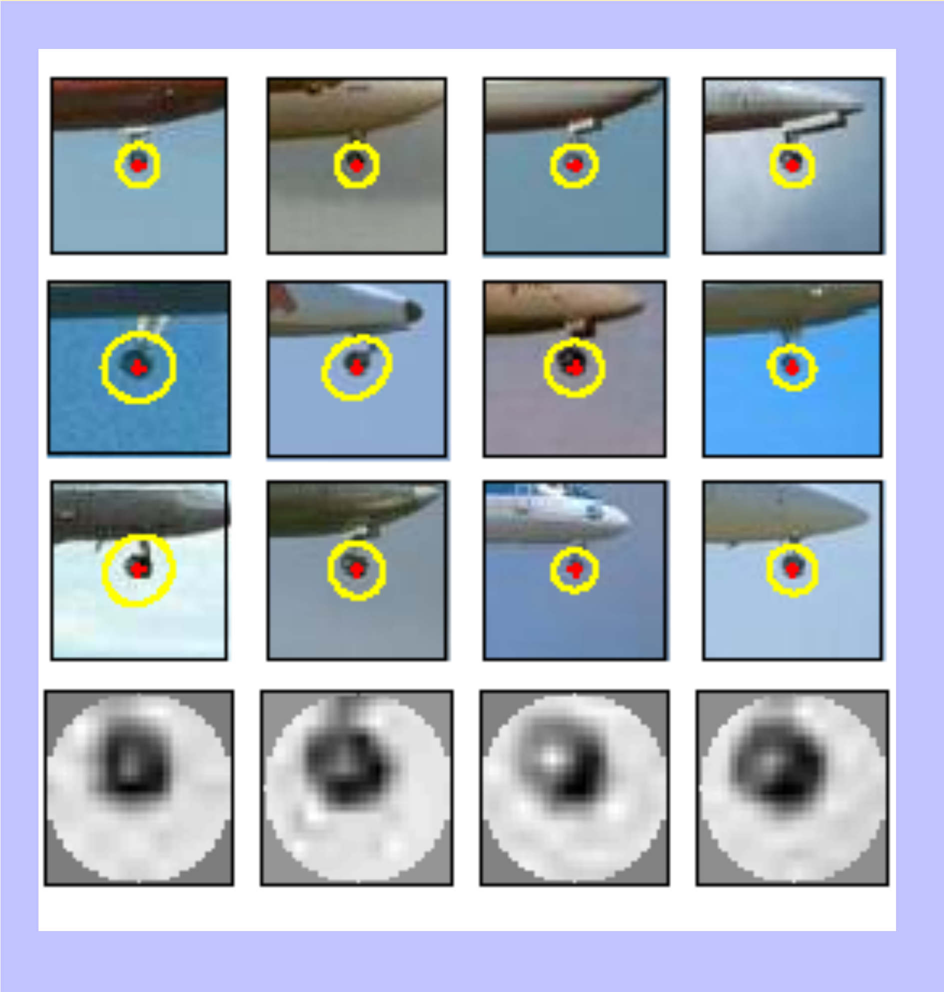
Slide credit: Josef Sivic

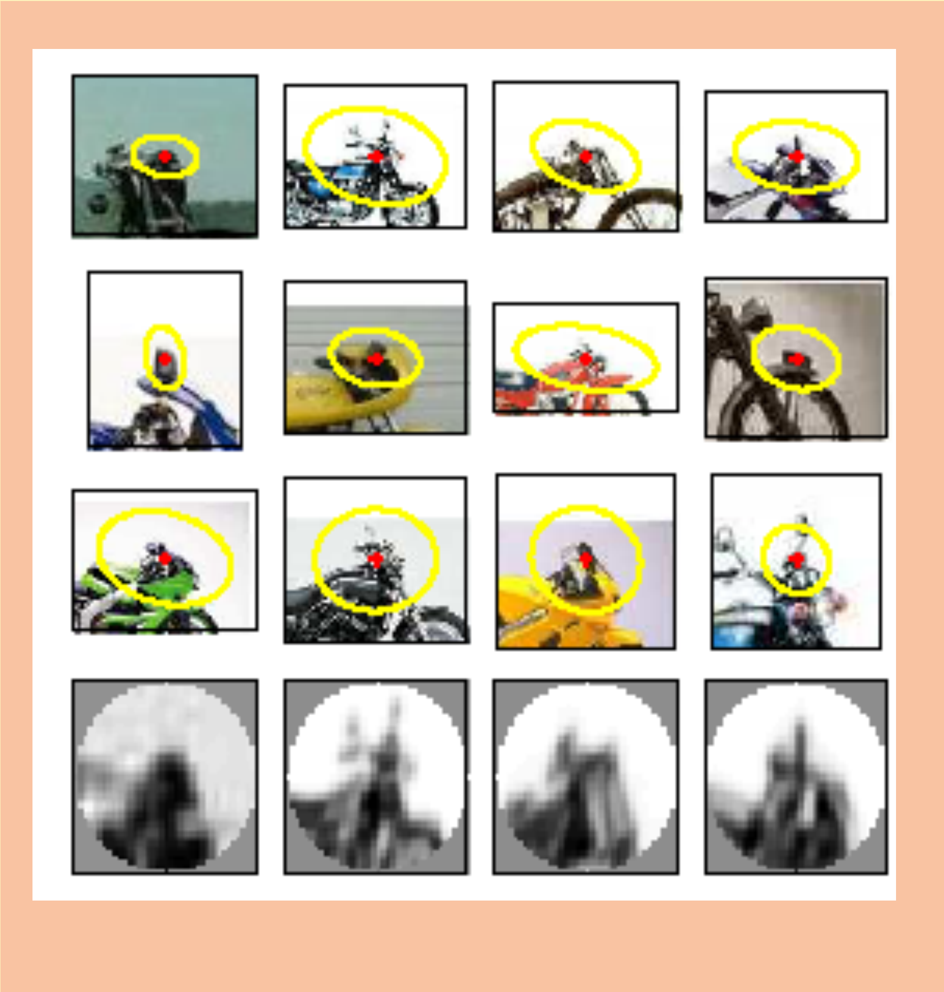
2. Codewords dictionary formation



Fei-Fei et al. 2005

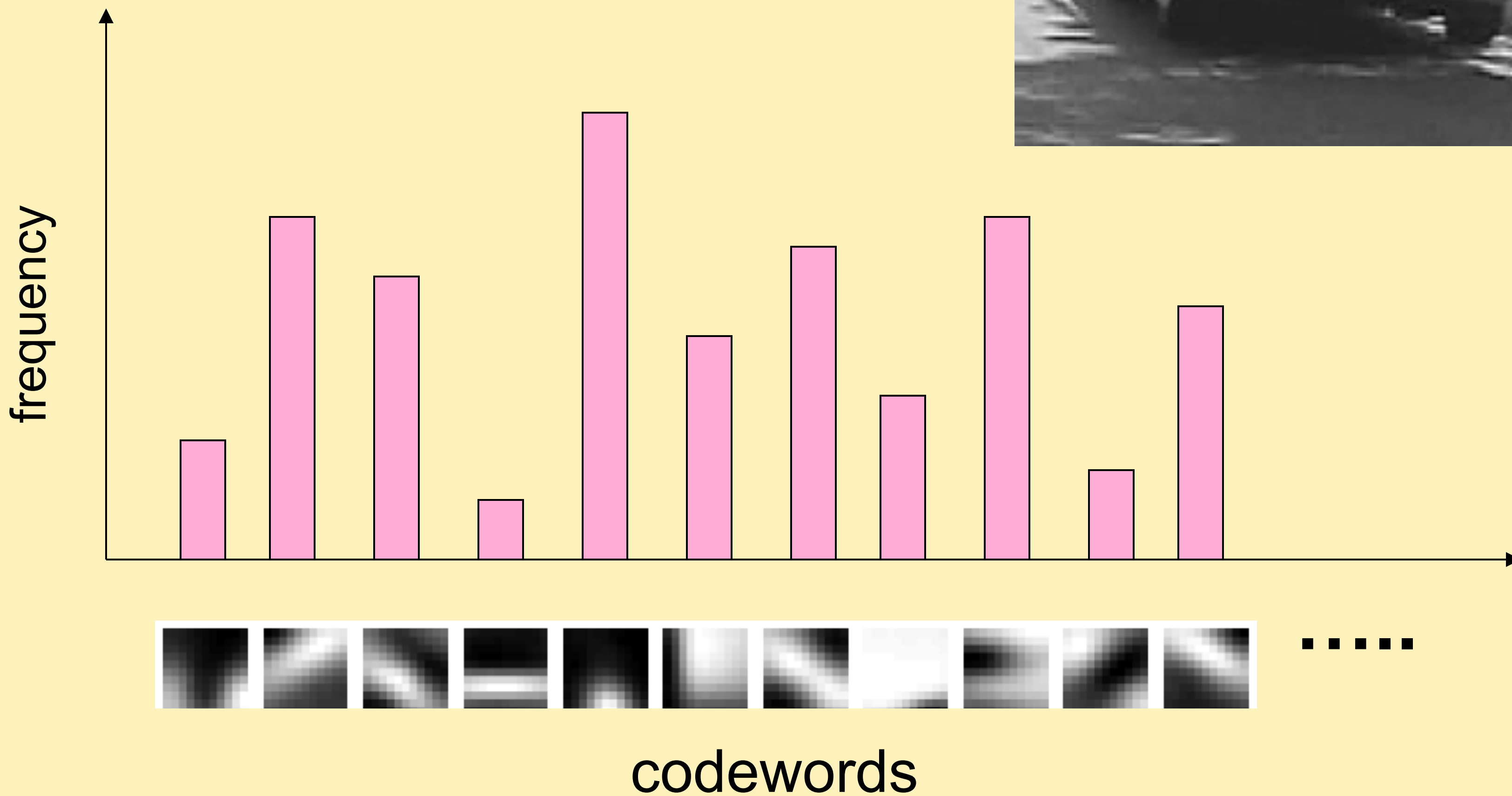
Image patch examples of codewords



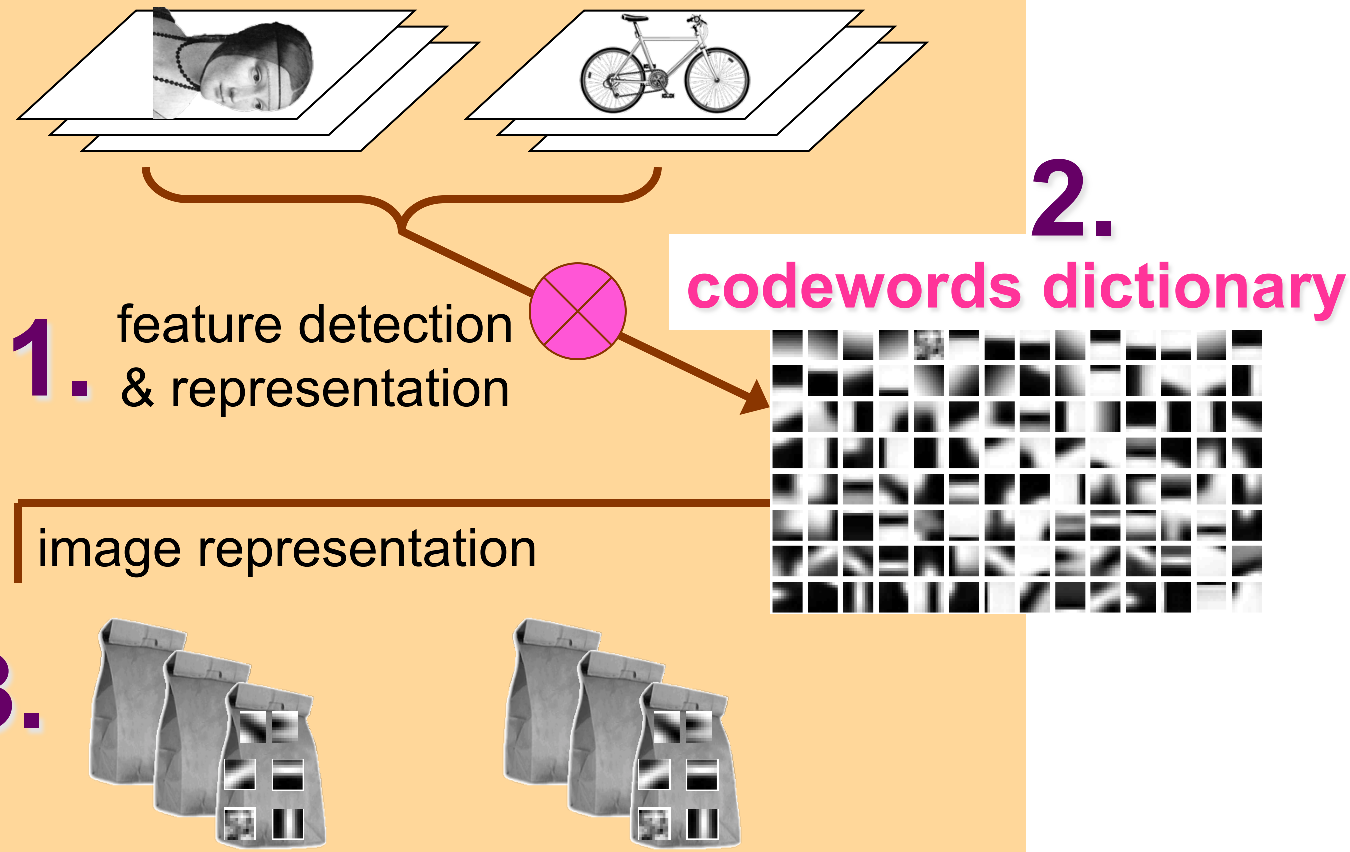


Sivic et al. 2005

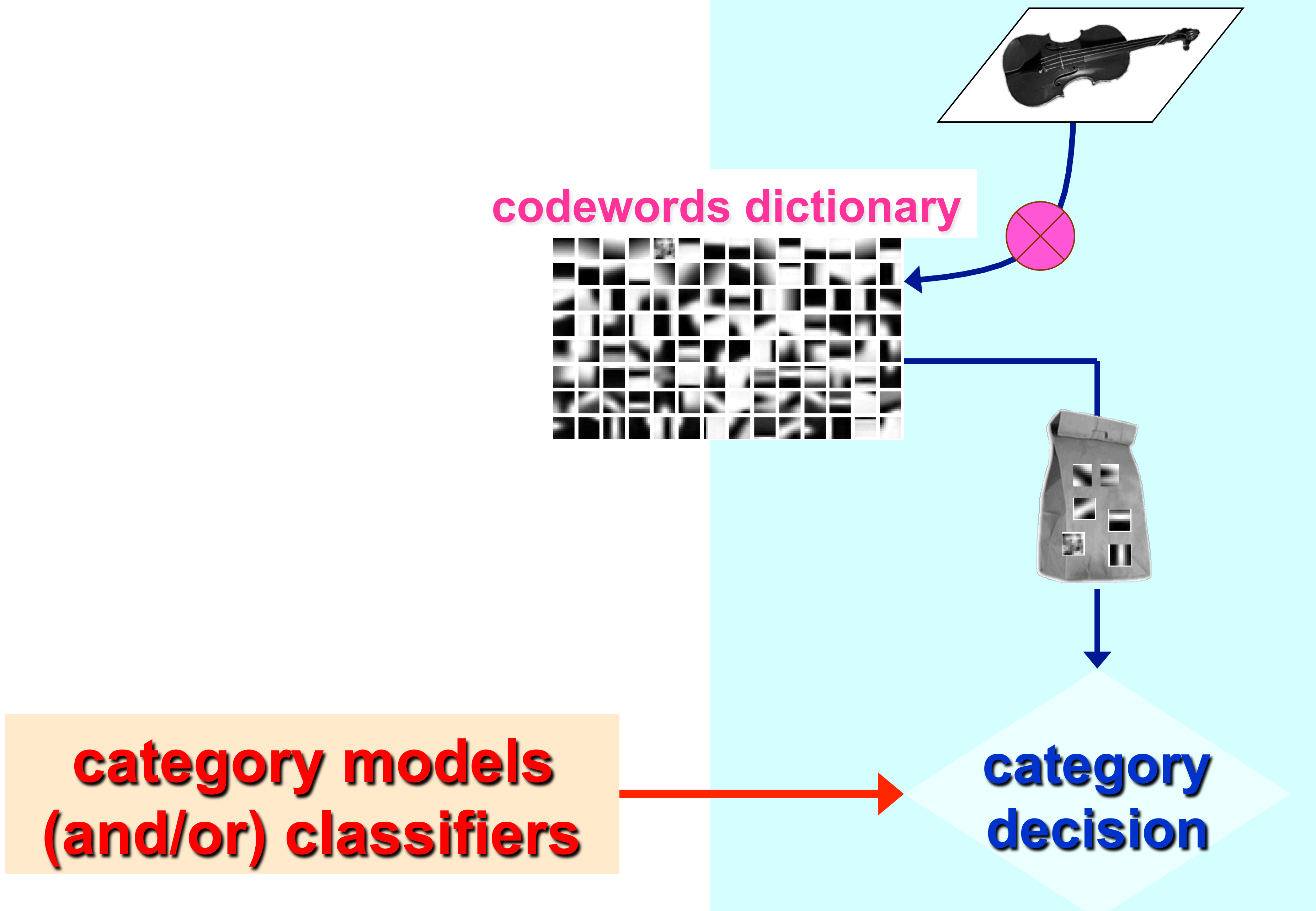
3. Image representation



Representation

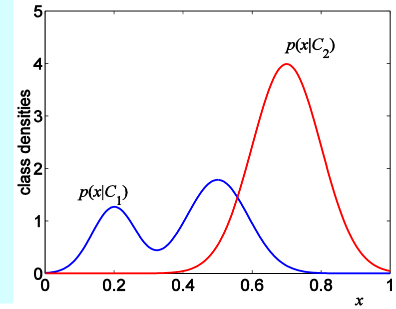


Learning and Recognition

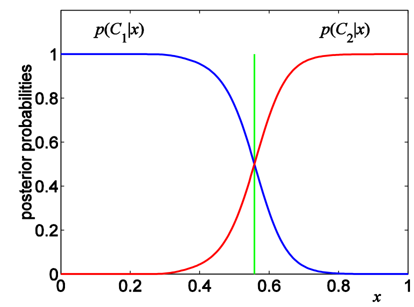


Learning and Recognition

1. Generative method:
 - graphical models



2. Discriminative method:
 - SVM



category models (and/or) classifiers

2 generative models

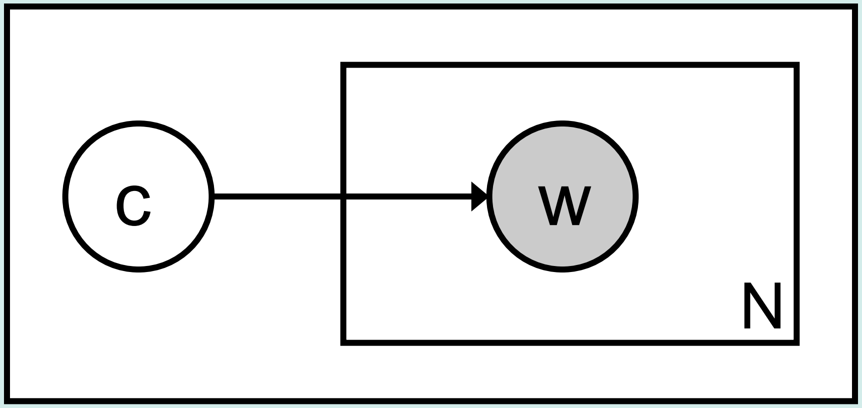
1. Naïve Bayes classifier
   – Csurka Bray, Dance & Fan, 2004

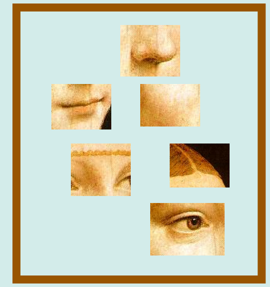
2. Hierarchical Bayesian text models (pLSA and LDA)
   – Background: Hoffman 2001, Blei, Ng & Jordan, 2004
   – Object categorization: Sivic et al. 2005, Sudderth et al. 2005
   – Natural scene categorization: Fei-Fei et al. 2005

First, some notations

- w_n: each patch in an image
  - $w_n = [0,0,\ldots1,\ldots,0,0]^T$
- **w**: a collection of all N patches in an image
  - $\mathbf{w} = [w_1,w_2,\ldots,w_N]$
- 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





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

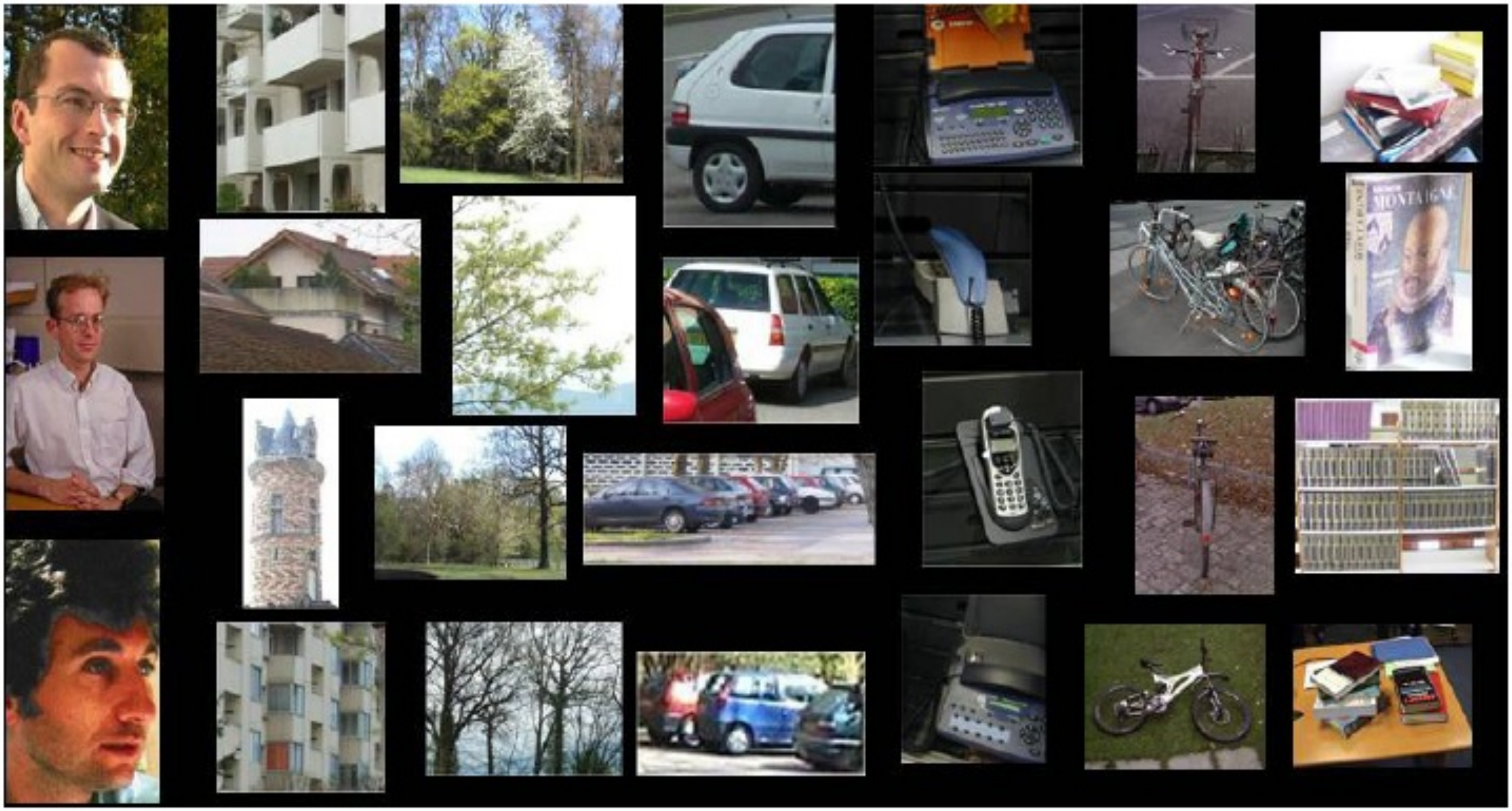
Object class decision

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

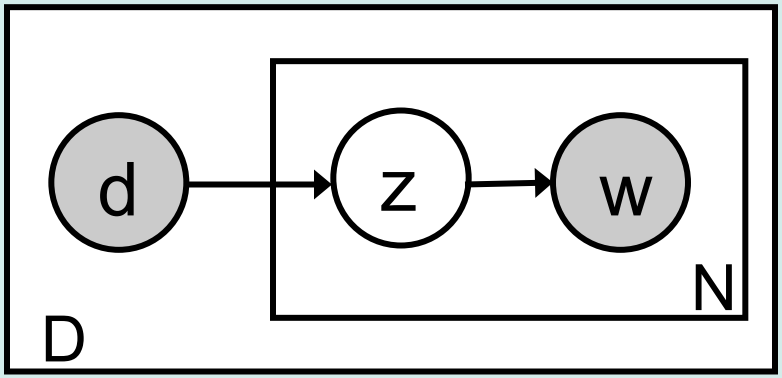
Table 1. Confusion matrix and the mean rank for the best vocabulary (*k=1000*).

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

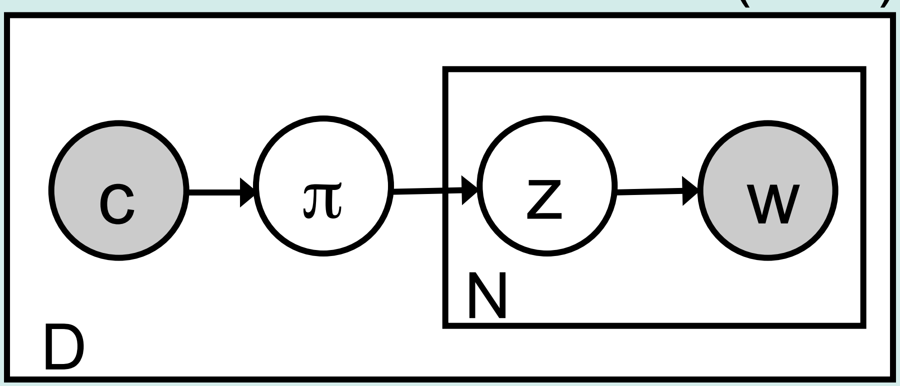
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 2001

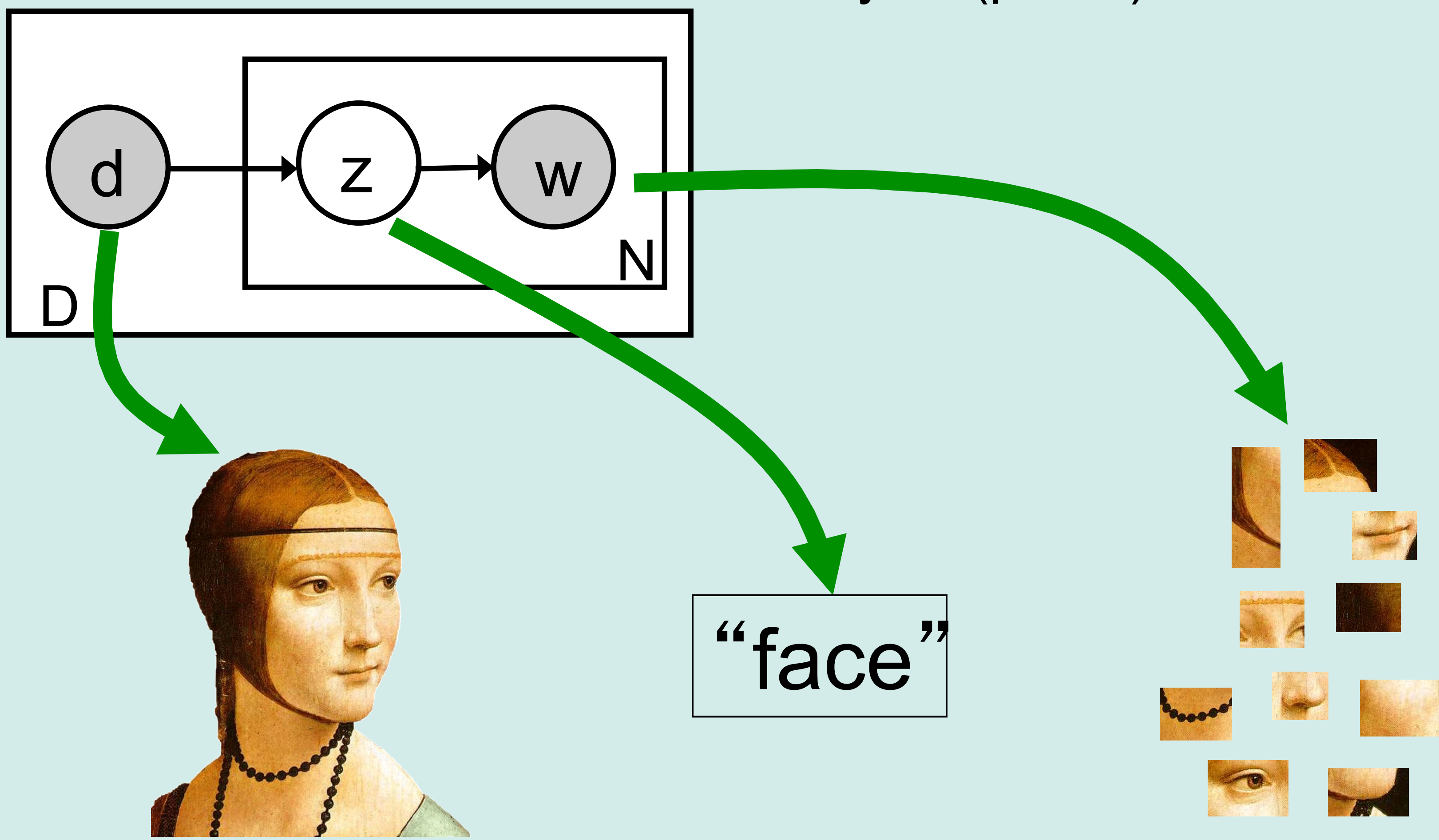
Latent Dirichlet Allocation (LDA)



Blei et al., 2001

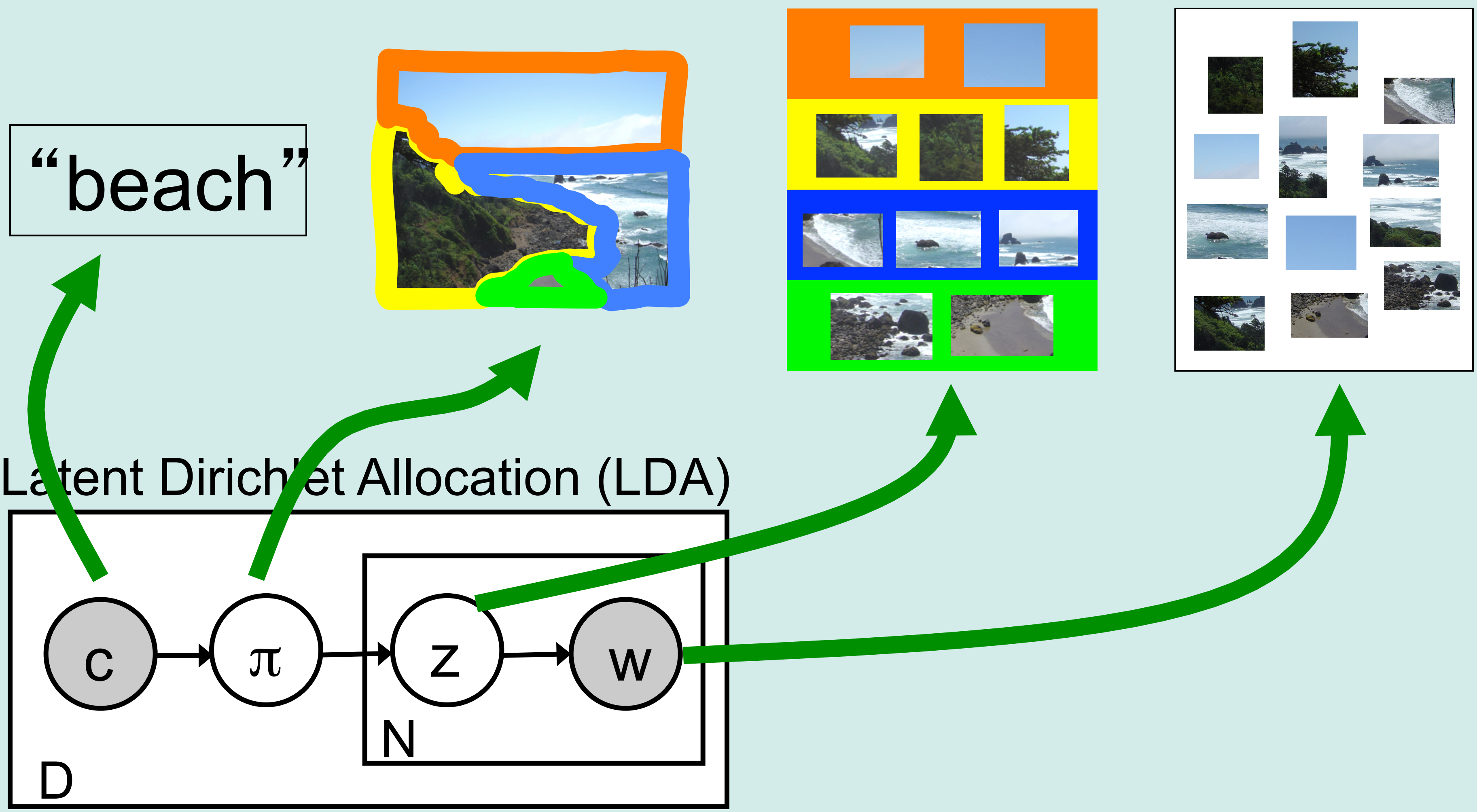
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)

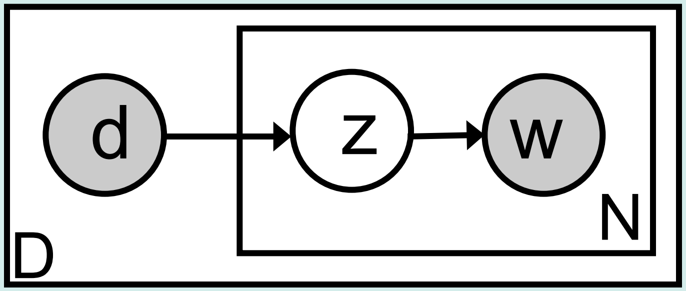


Sivic et al. ICCV 2005

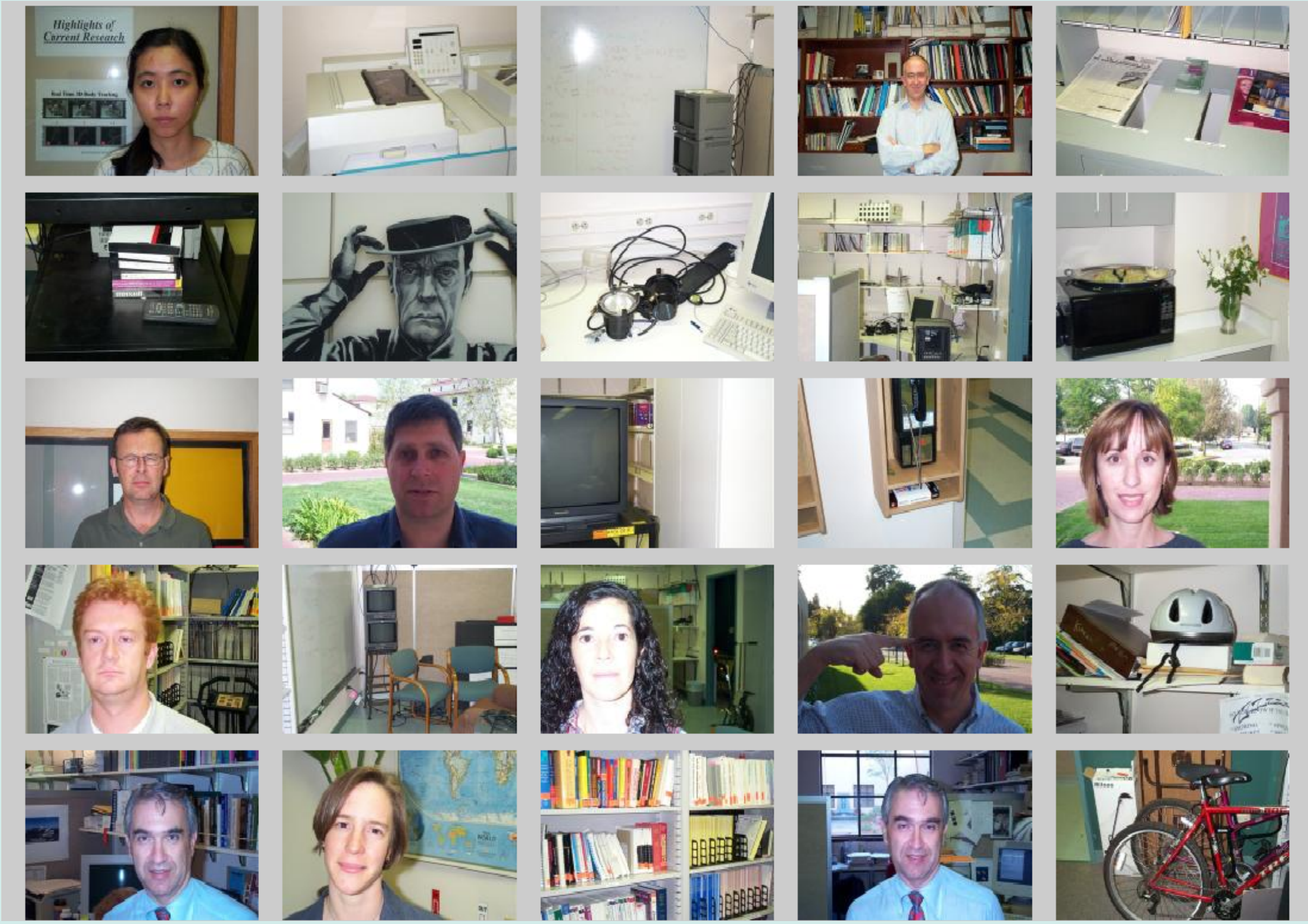
Case #2: Hierarchical Bayesian text models

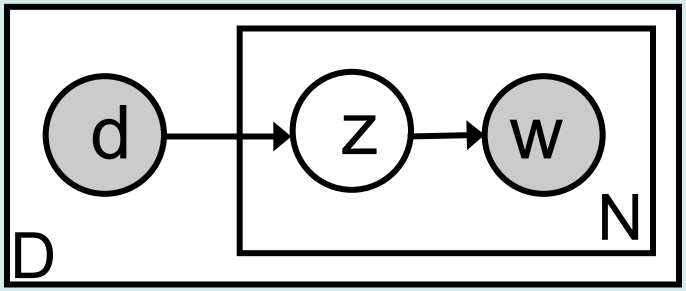


Fei-Fei et al. ICCV 2005



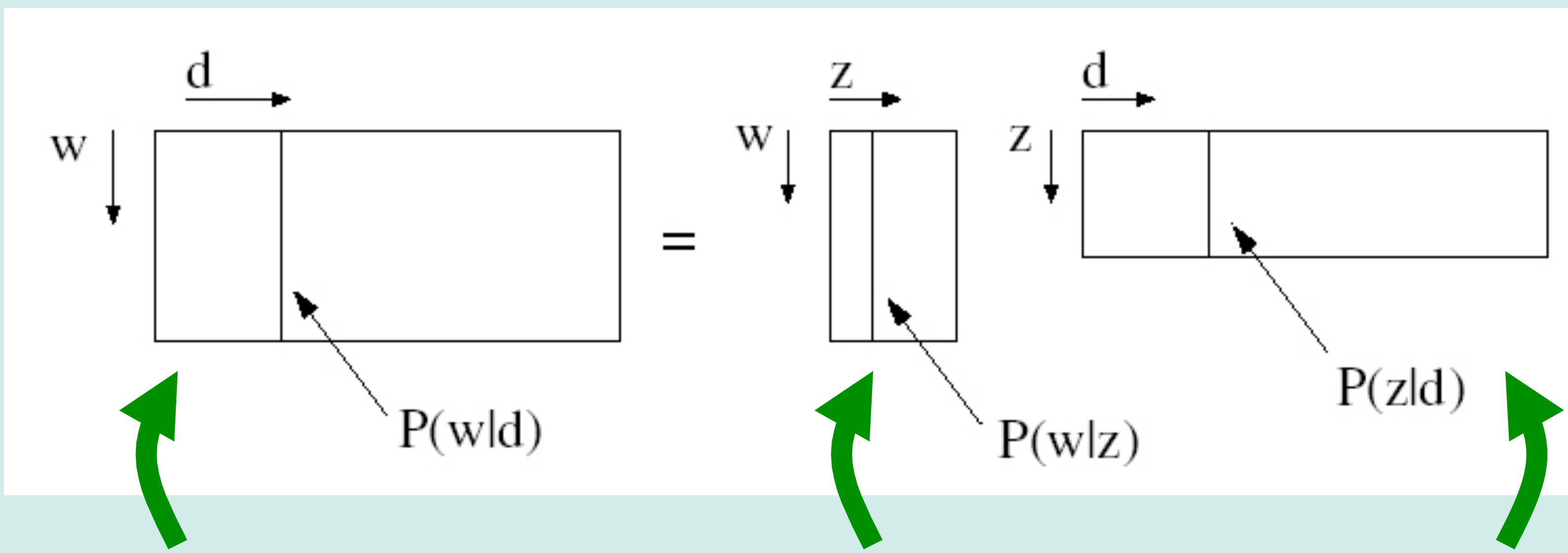
Case #2: the pLSA model





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

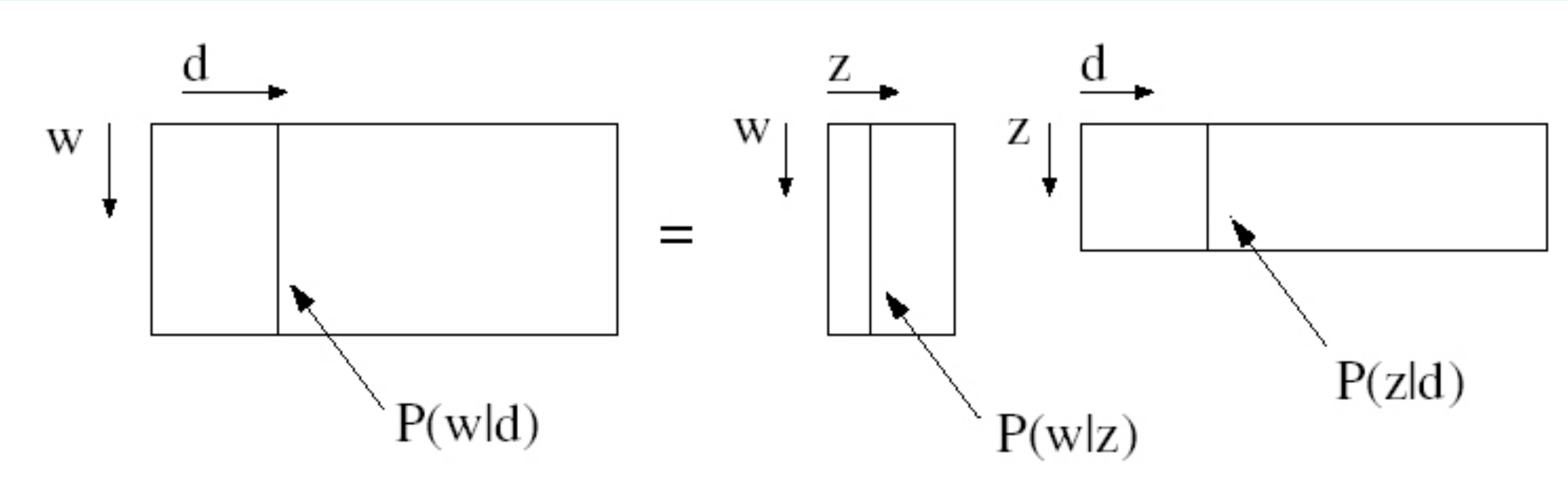
Codeword distributions per theme (topic)

Theme distributions per image

Slide credit: Josef Sivic

Case #2: Recognition using pLSA

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



Slide credit: Josef Sivic

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

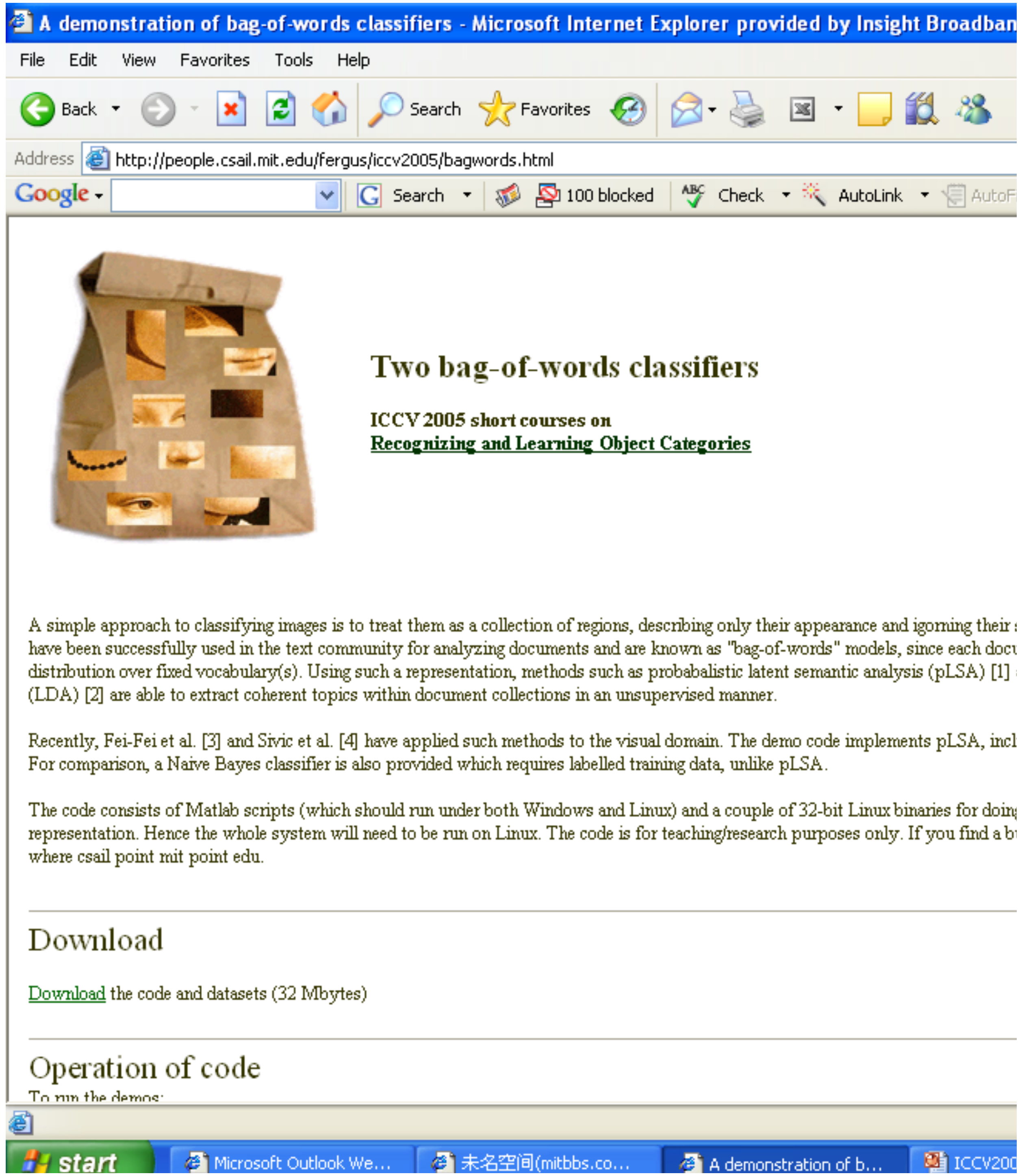
M … number of codewords

N … number of images

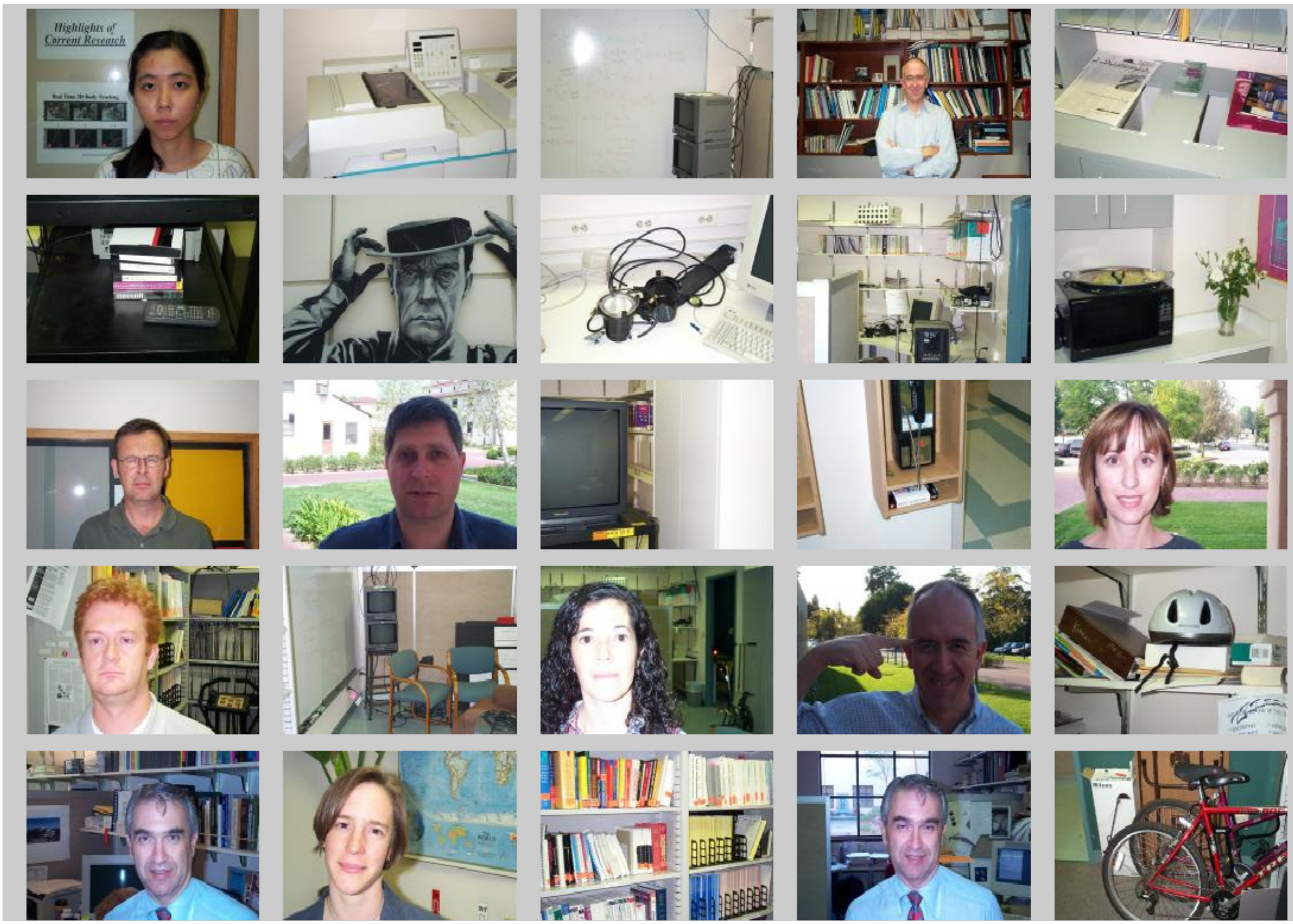
Slide credit: Josef Sivic

Demo

- Course website

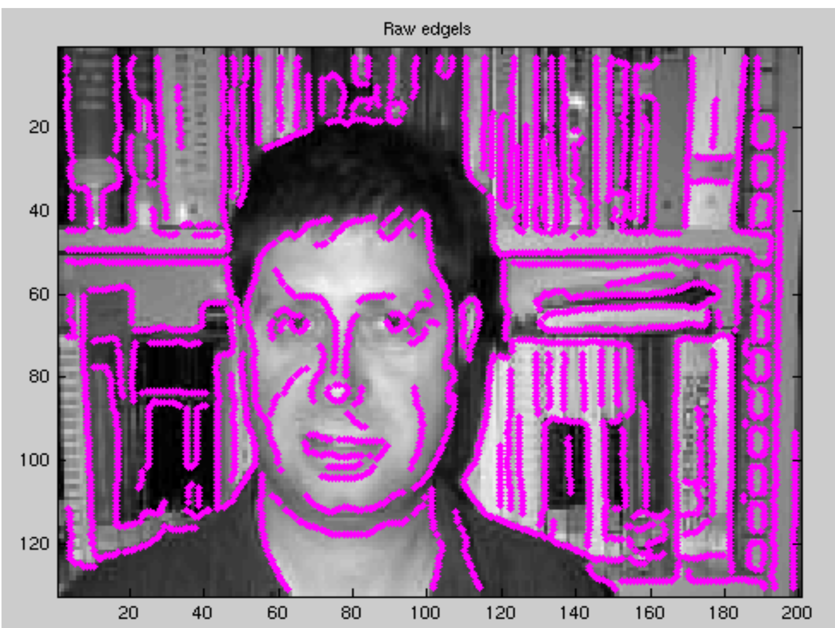


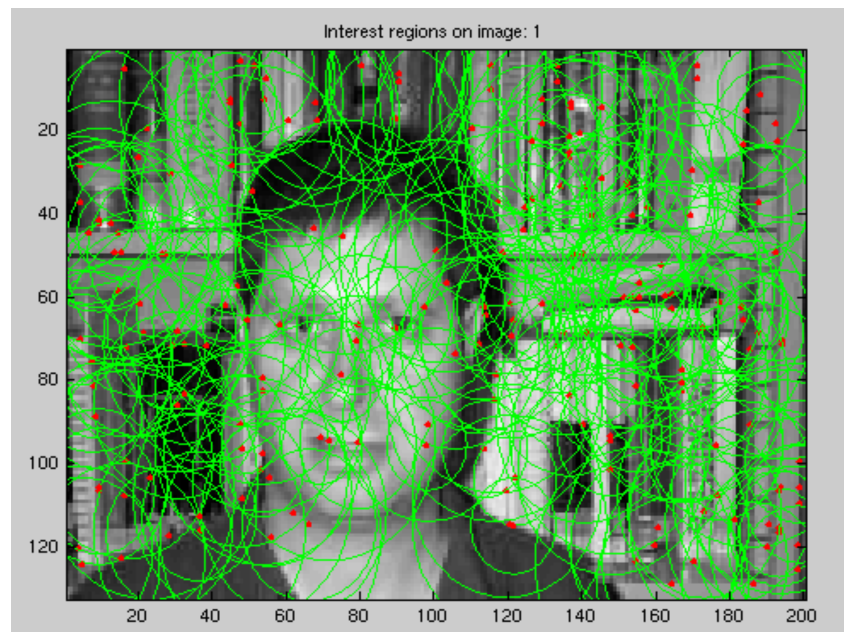
task: face detection – no labeling



Demo: feature detection

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



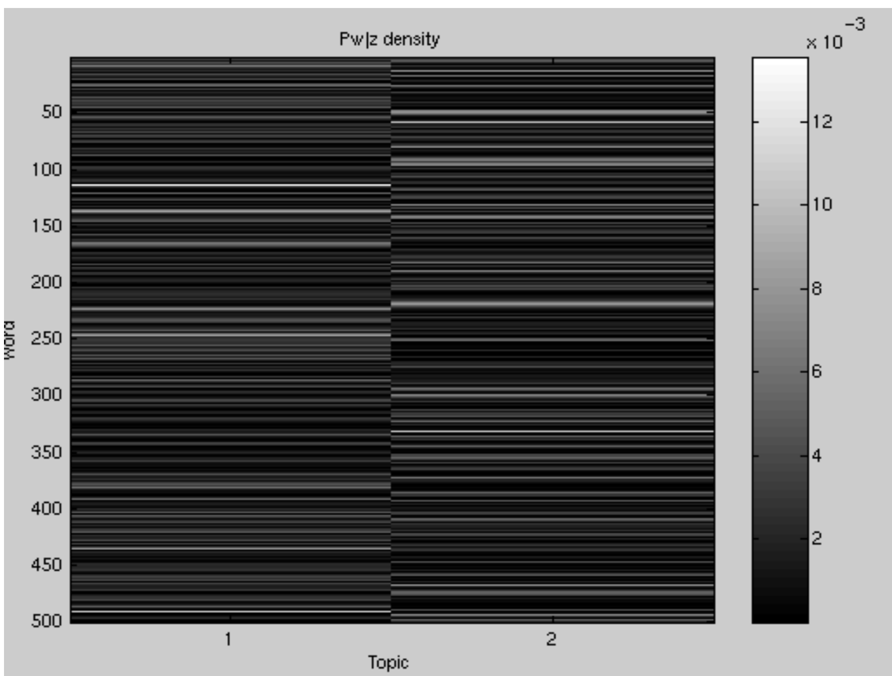


Demo: learnt parameters

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

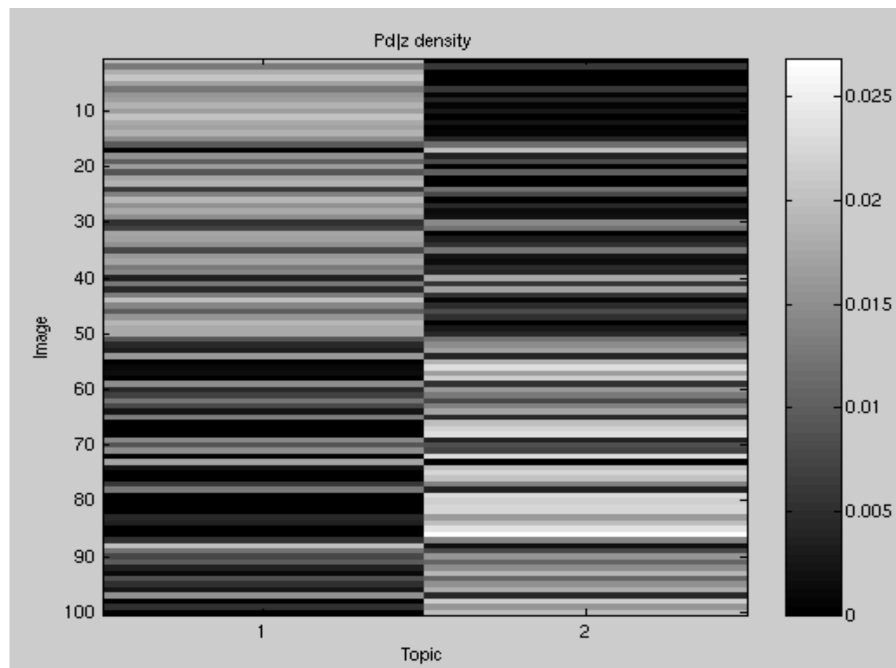
Codeword distributions per theme (topic)

$$p(w\,|\,z)$$

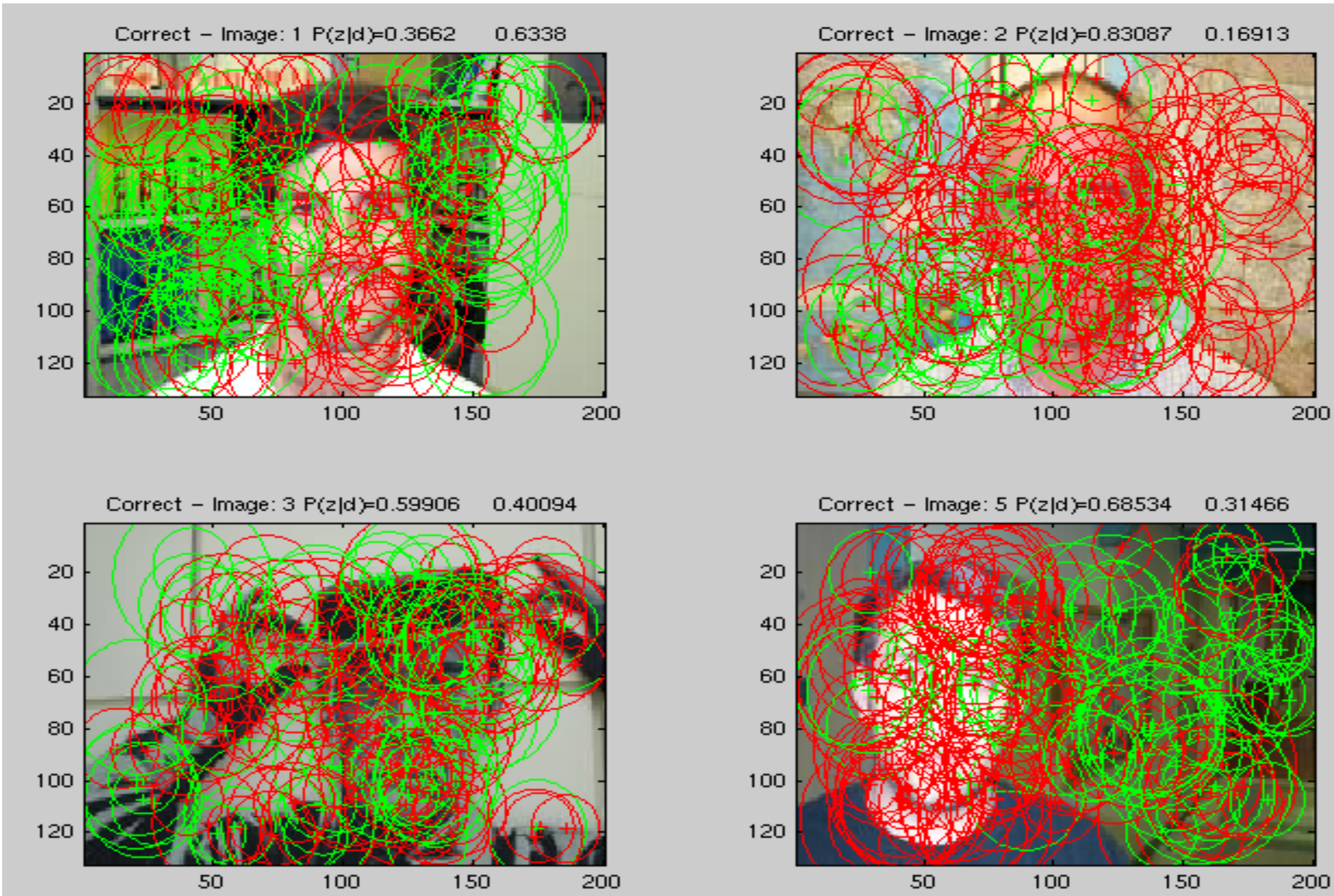


Theme distributions per image

$$p(z\,|\,d)$$

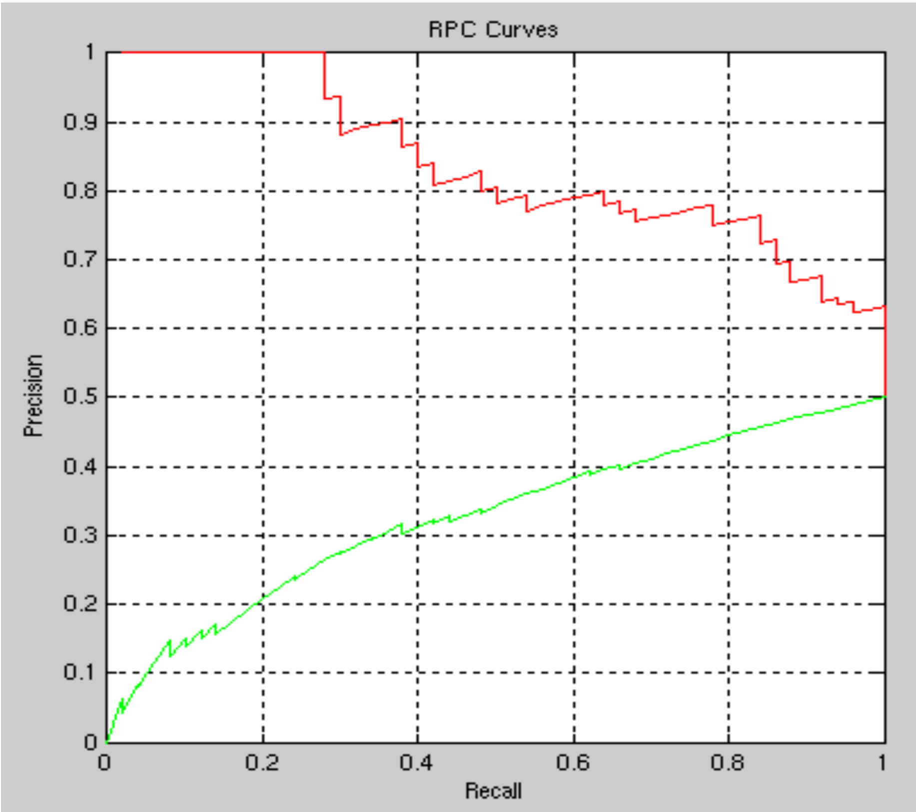


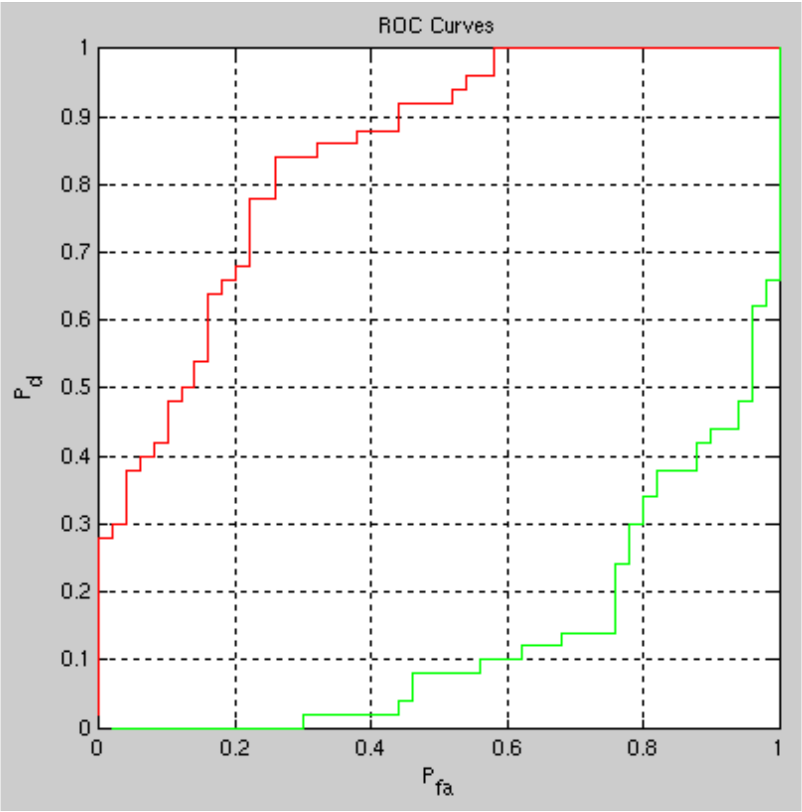
Demo: recognition examples



Demo: categorization results

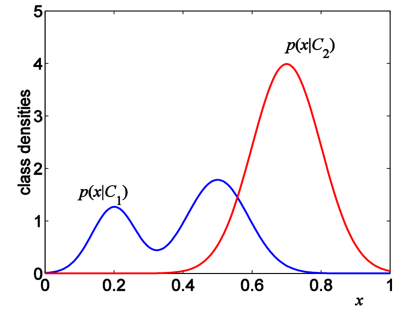
- Performance of each theme



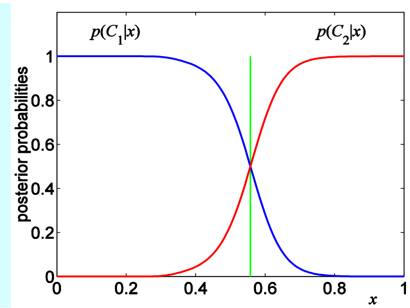


Learning and Recognition

1. Generative method:
   - graphical models

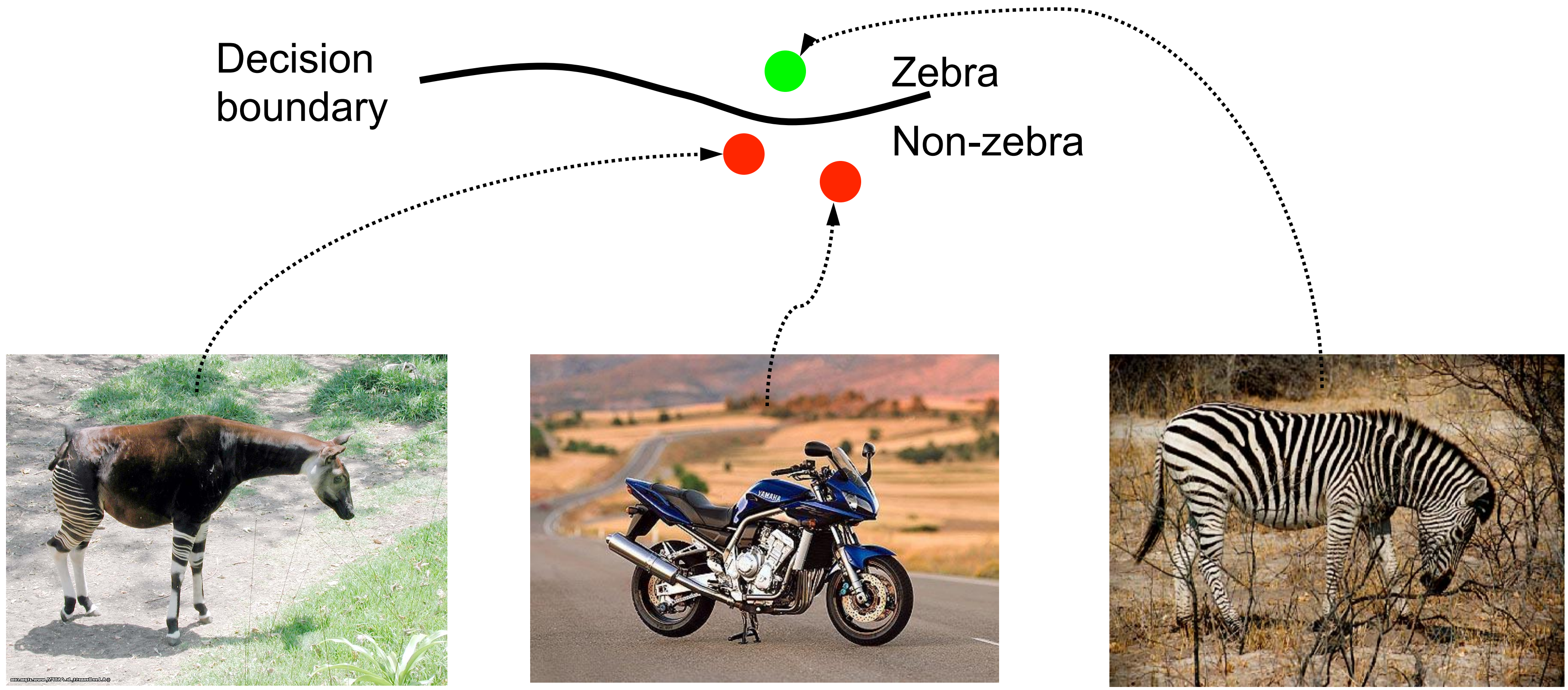


2. Discriminative method:
   - SVM



category models (and/or) classifiers

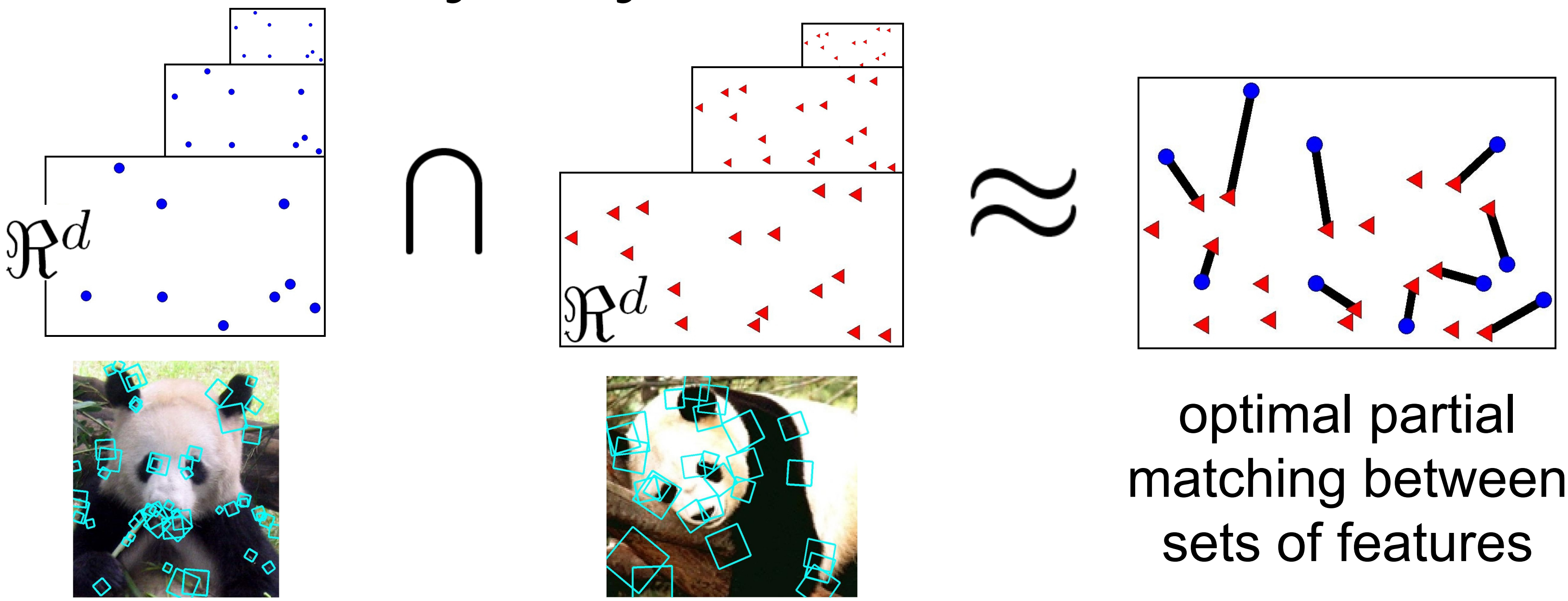
Discriminative methods based on ‘bag of words’ representation



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



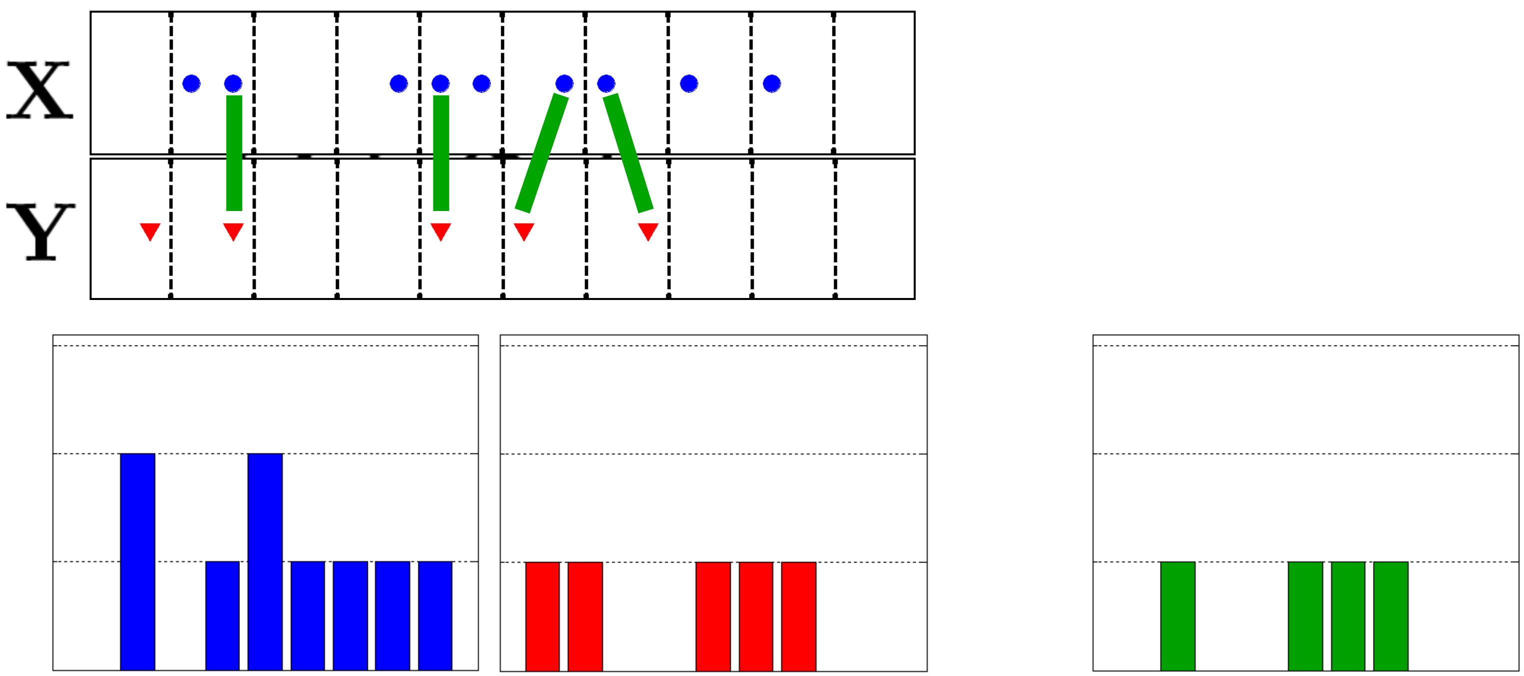
$$K_{\Delta}\left(\Psi(\mathbf{X}), \Psi(\mathbf{Y})\right)$$

Grauman & Darrell, 2005, Slide credit: Kristen Grauman

Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

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



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

Slide credit: Kristen Grauman

Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

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

$$\boxed{N_i = \underbrace{\mathcal{I}\left(H_i(\mathbf{X}), H_i(\mathbf{Y})\right)}_{\text{matches at this level}} - \underbrace{\mathcal{I}\left(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})\right)}_{\text{matches at previous level}}}$$

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

Slide credit: Kristen Grauman

Pyramid match kernel

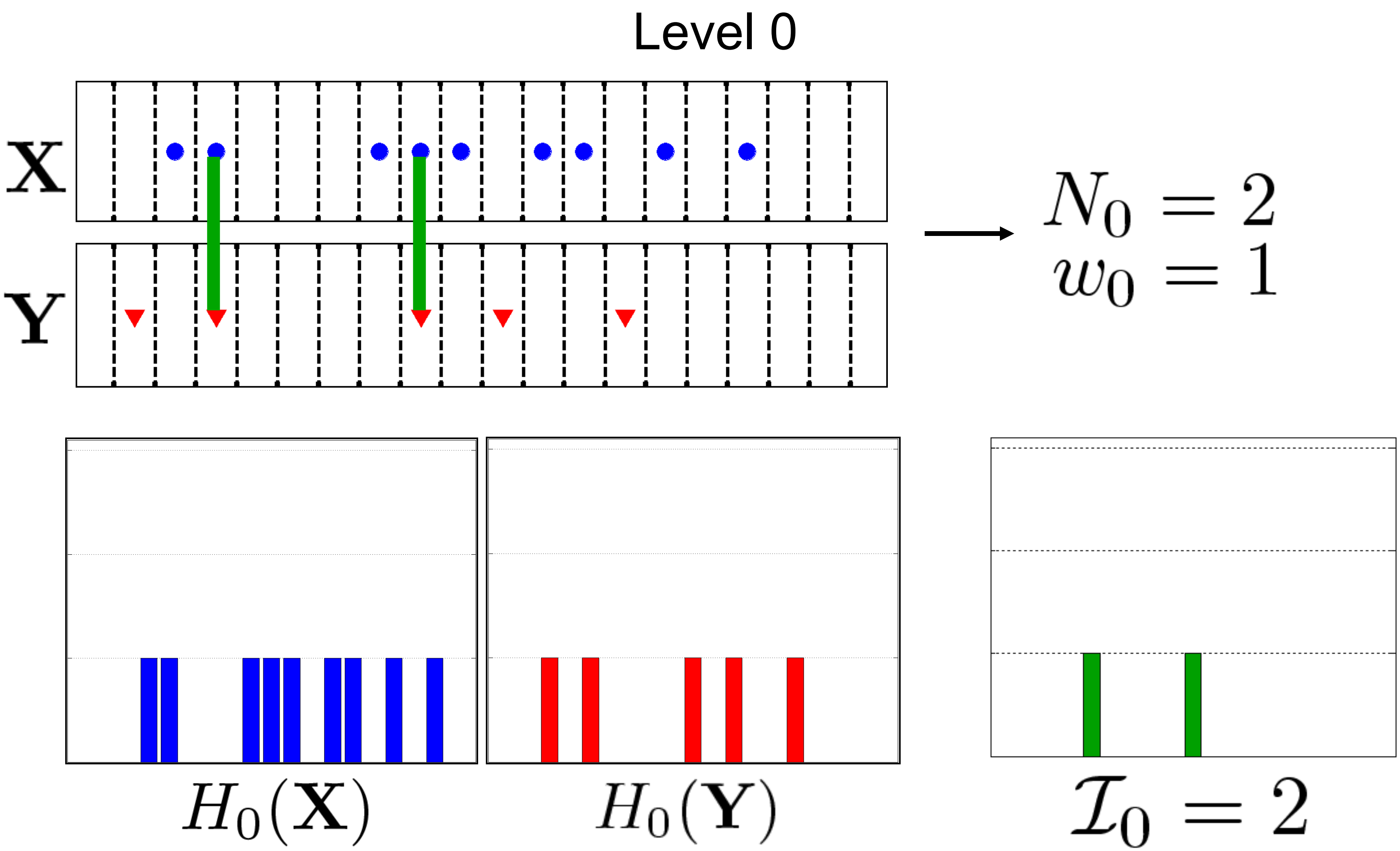
$$K_\Delta\left(\overbrace{\Psi(\mathbf{X}),\Psi(\mathbf{Y})}^{\text{histogram pyramids}}\right) =$$

$$\sum_{i=0}^{L} \underbrace{\frac{1}{2^i}}_{\substack{\text{measure of difficulty of}\\ \text{a match at level } i}} \Big(\underbrace{\mathcal{I}\left(H_i(\mathbf{X}), H_i(\mathbf{Y})\right) - \mathcal{I}\left(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})\right)}_{\text{number of newly matched pairs at level } i} \Big)$$

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

Slide credit: Kristen Grauman

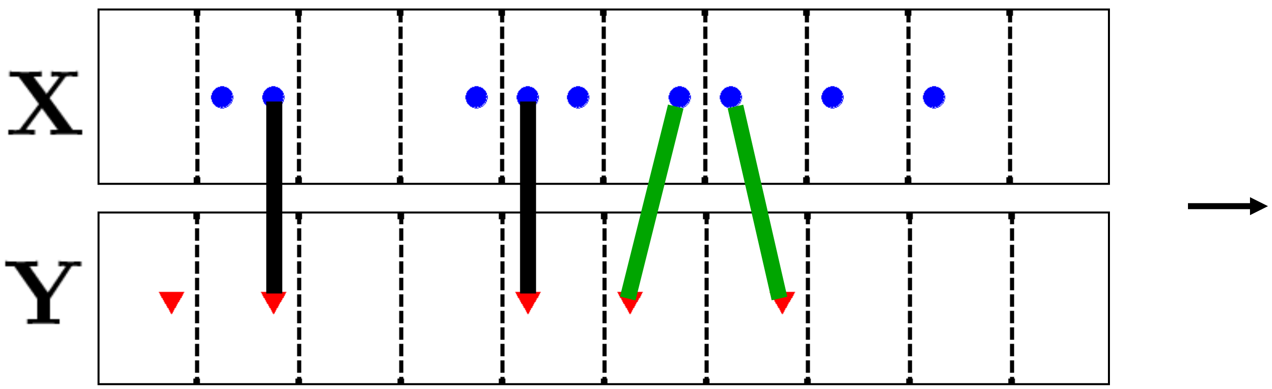
Example pyramid match



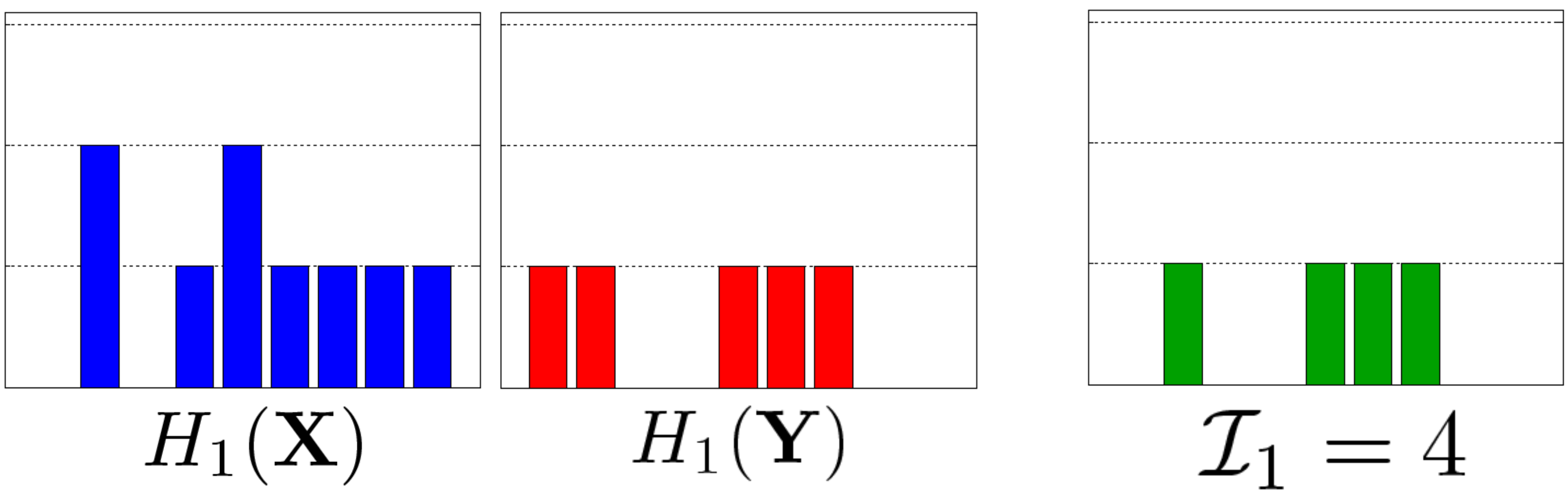
Slide credit: Kristen Grauman

Example pyramid match

Level 1



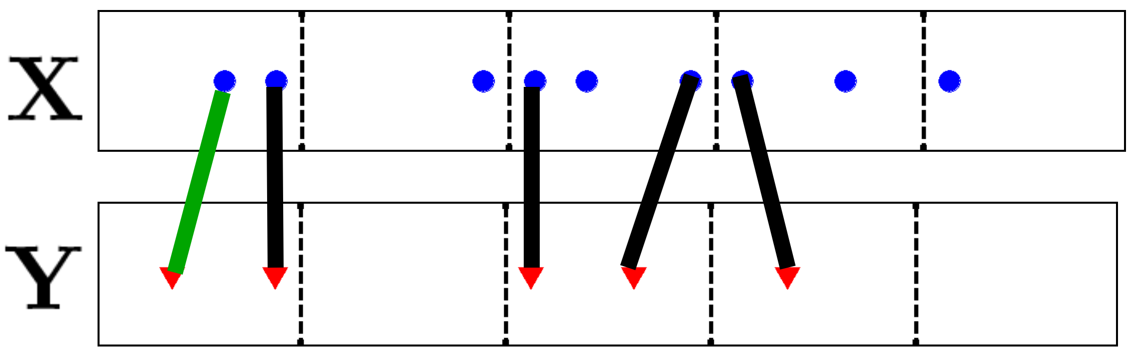
$$N_1 = 4 - 2 = 2$$
$$w_1 = \frac{1}{2}$$



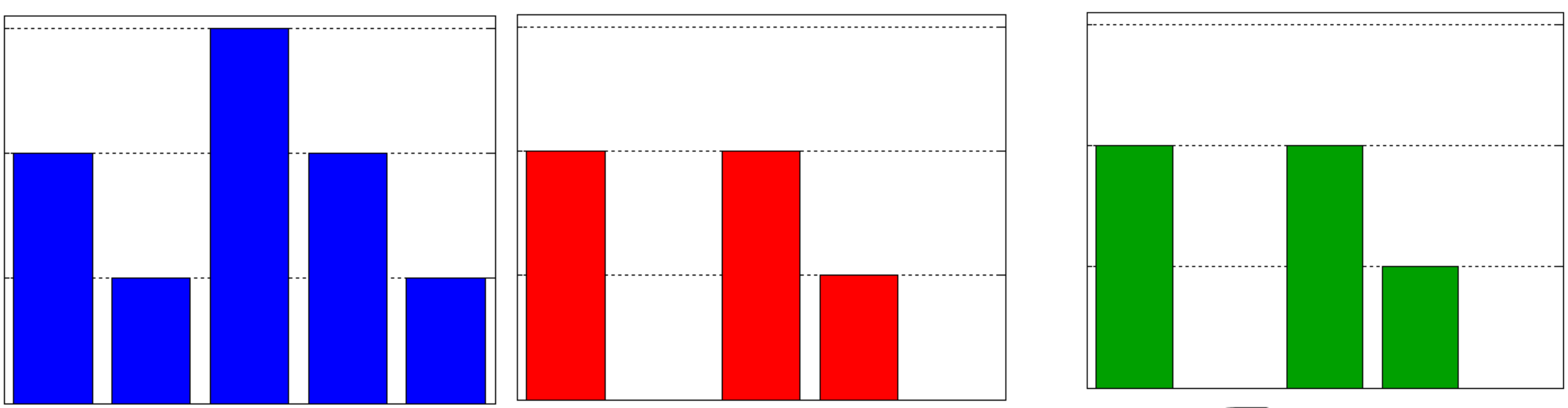
Slide credit: Kristen Grauman

Example pyramid match

Level 2



$$\rightarrow \begin{array}{l} N_2 = 5 - 4 = 1 \\ w_2 = \frac{1}{4} \end{array}$$

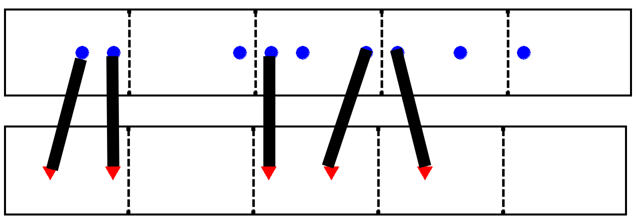


$H_2(\mathbf{X})$ $H_2(\mathbf{Y})$ $\mathcal{I}_2 = 5$

Slide credit: Kristen Grauman

Example pyramid match

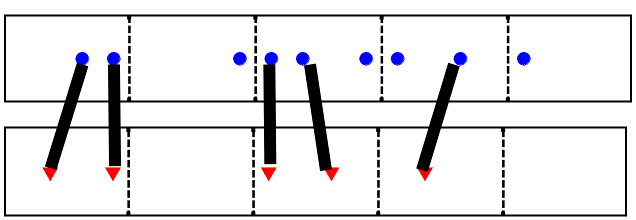
pyramid match



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

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

optimal match

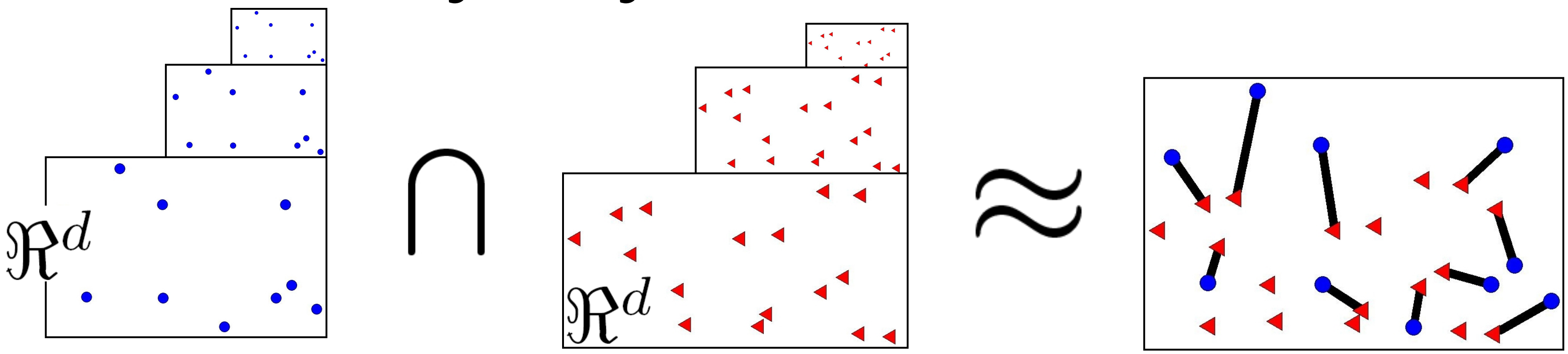


$$K = \max_{\pi : \mathbf{X} \to \mathbf{Y}} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathcal{S}(\mathbf{x}_i, \pi(\mathbf{x}_i))$$

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

Slide credit: Kristen Grauman

Summary: Pyramid match kernel



optimal partial matching between sets of features

$$K_\Delta\left(\Psi(\mathbf{X}),\Psi(\mathbf{Y})\right) = \sum_{i=0}^{L} \underbrace{\frac{1}{2^i}}_{\text{difficulty of a match at level i}} \Big(\underbrace{\mathcal{I}\left(H_i(\mathbf{X}),H_i(\mathbf{Y})\right) - \mathcal{I}\left(H_{i-1}(\mathbf{X}),H_{i-1}(\mathbf{Y})\right)}_{\text{number of new matches at level i}} \Big)$$

Slide credit: Kristen Grauman

Object recognition results

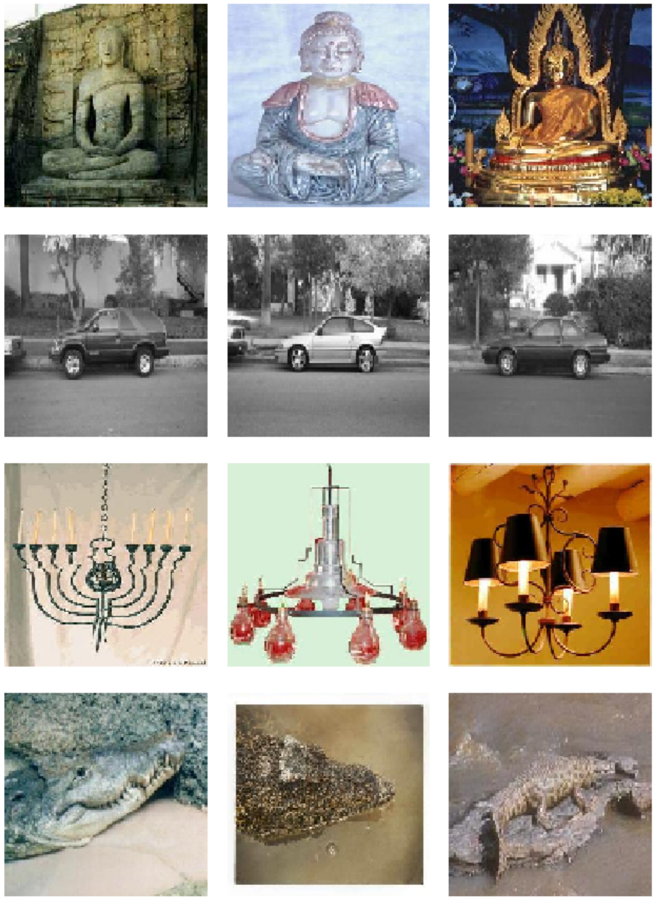
- ETH-80 database
  8 object classes
  (*Eichhorn and Chapelle 2004*)
- Features:
  - Harris detector
  - PCA-SIFT descriptor, *d*=10

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%

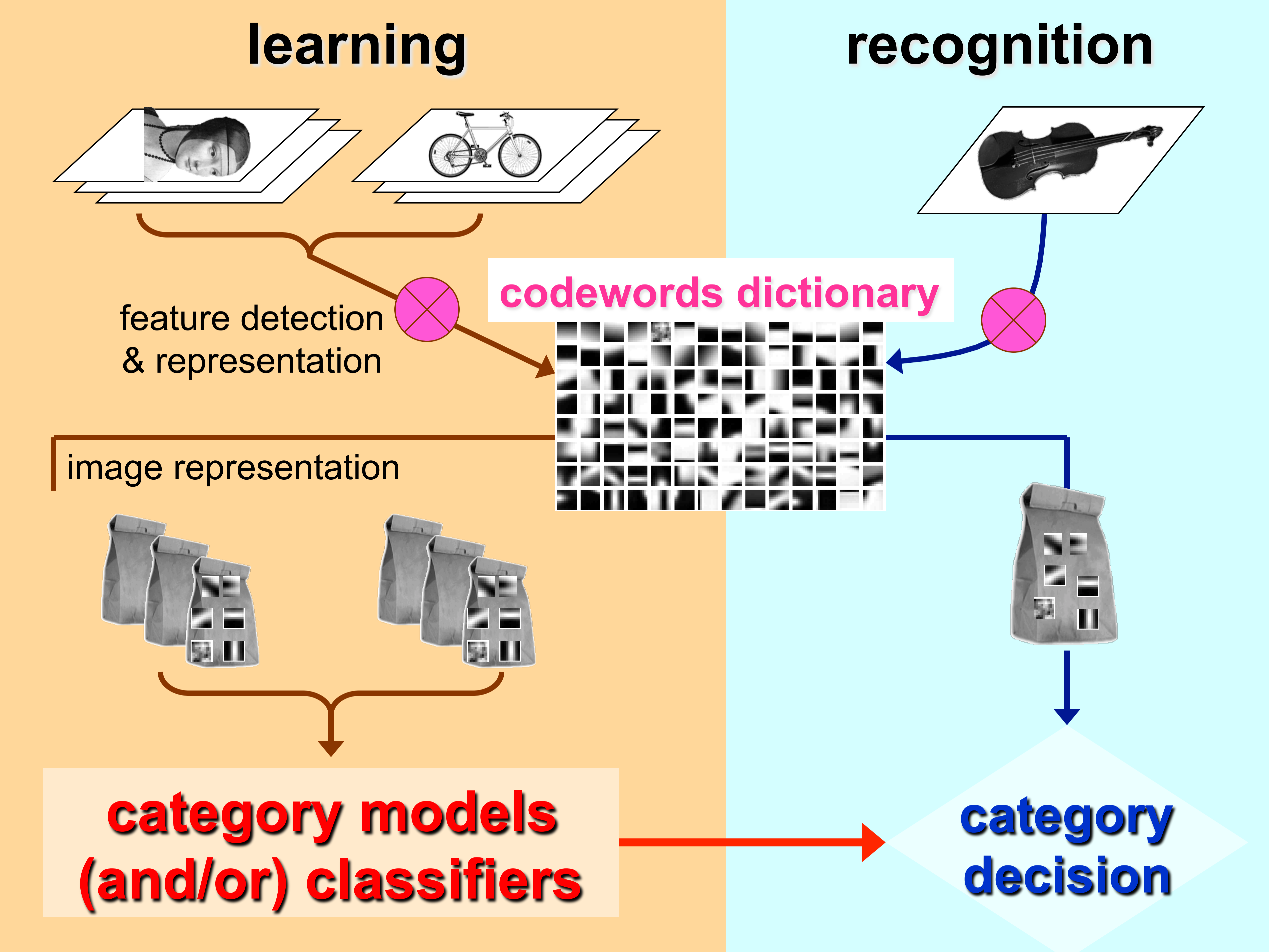
Slide credit: Kristen Grauman

Object recognition results

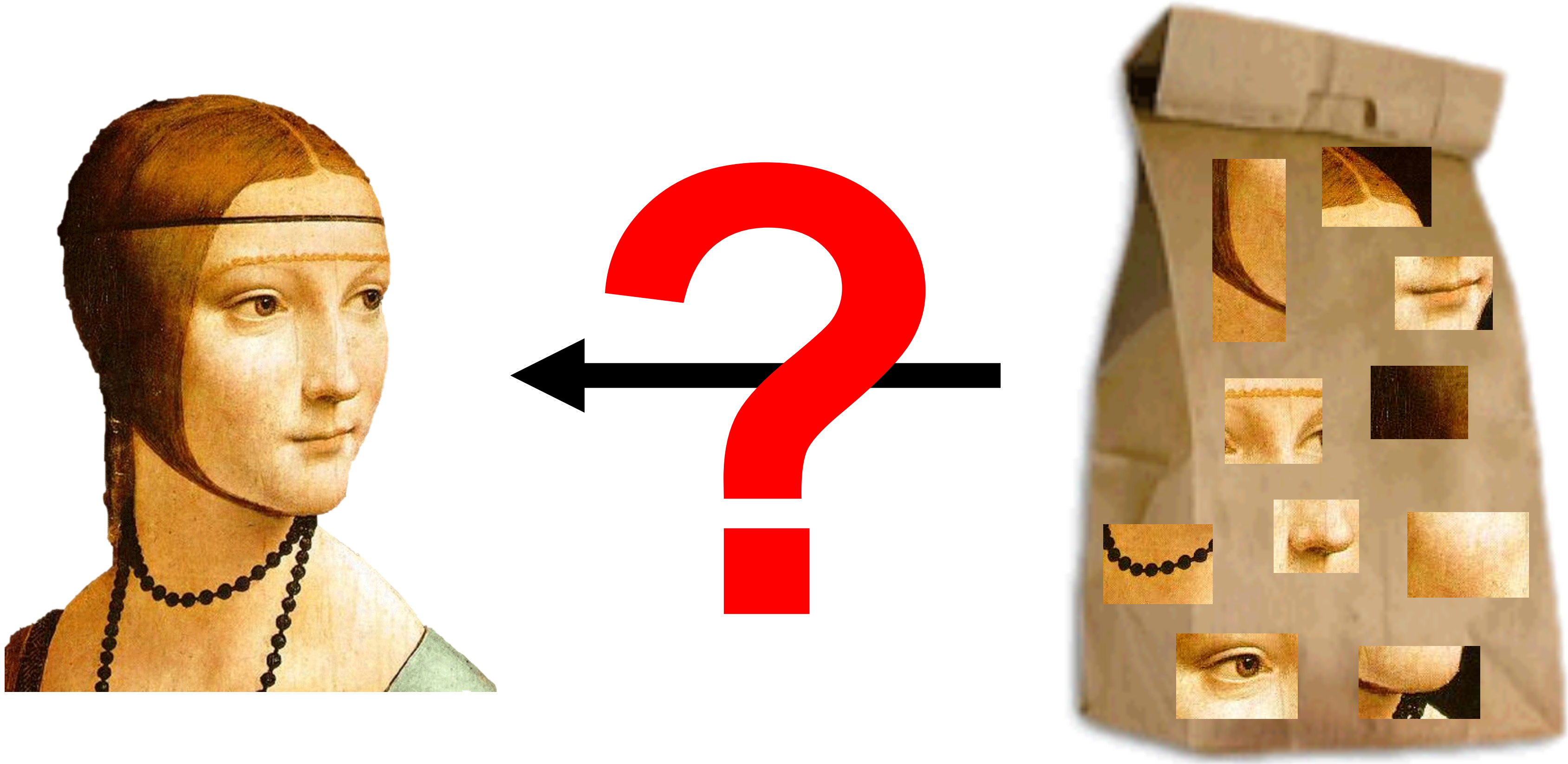
- Caltech objects database 101 object classes
- Features:
  - SIFT detector
  - PCA-SIFT descriptor, d=10
- 30 training images / class
- 43% recognition rate
  (1% chance performance)
- 0.002 seconds per match



Slide credit: Kristen Grauman

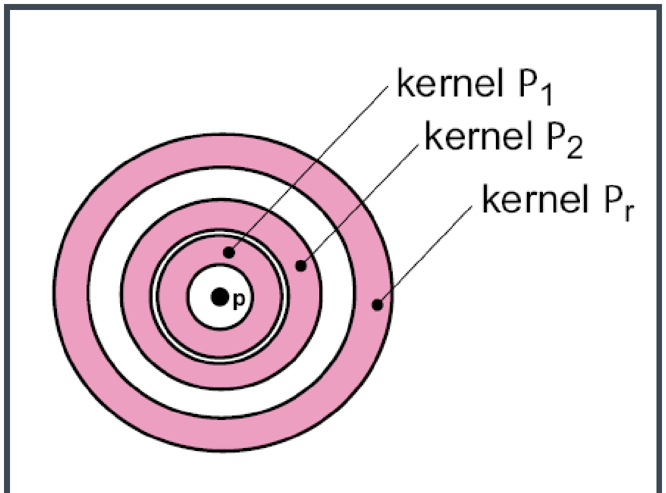

learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

What about spatial info?

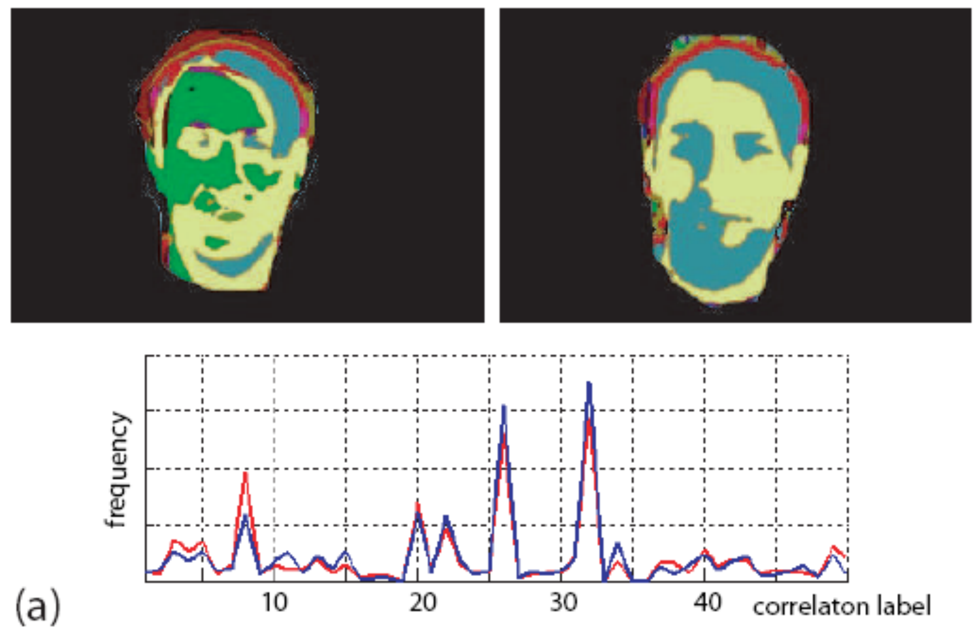


What about spatial info?

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



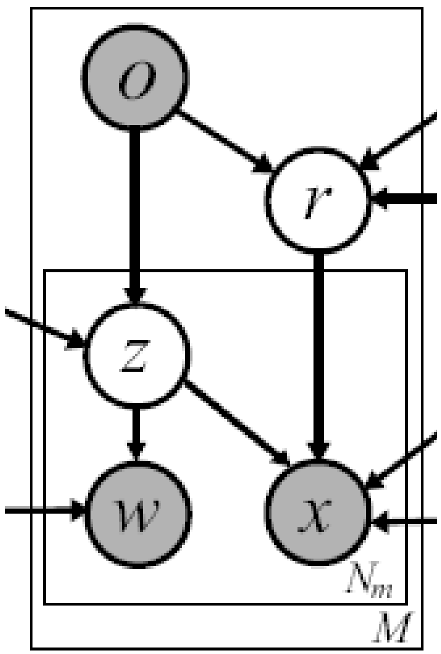
(a) Circular kernels

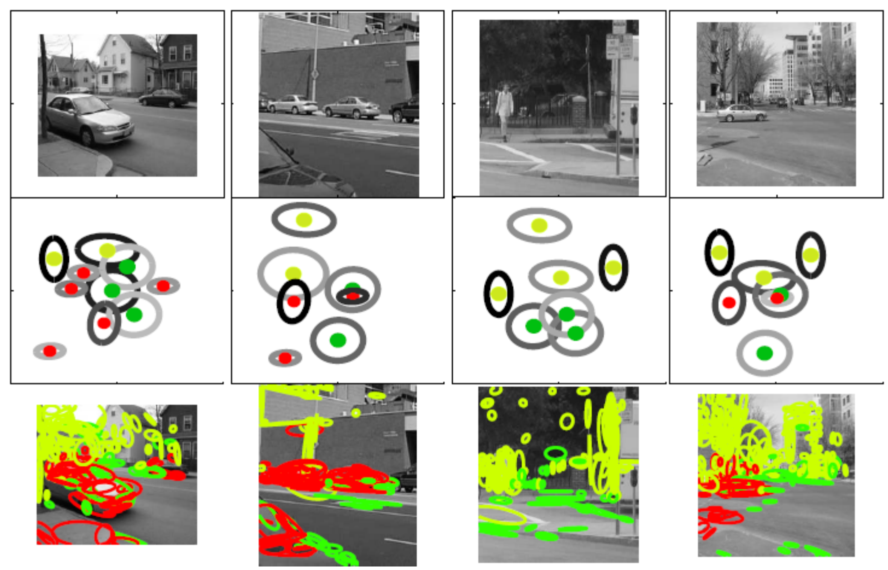


(a)

What about spatial info?

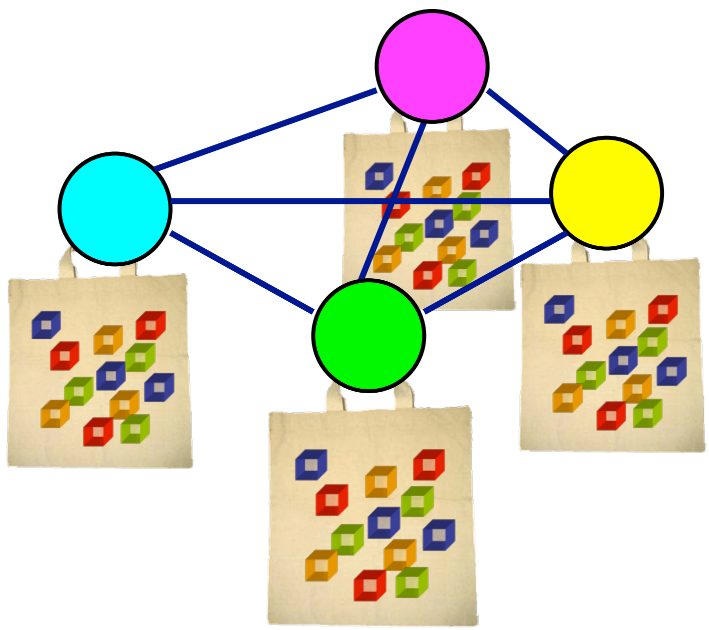
- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007

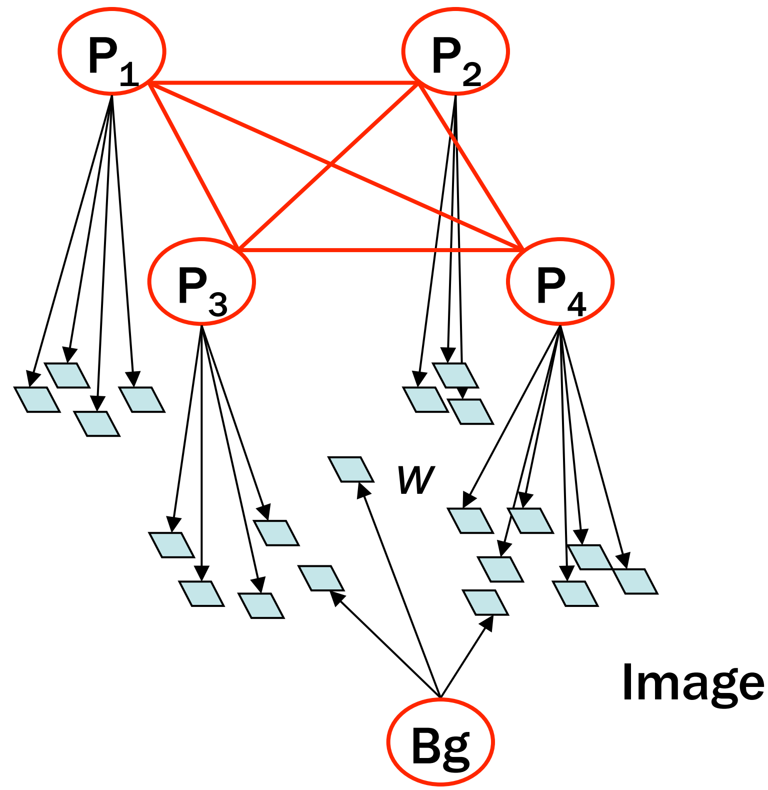




What about spatial info?

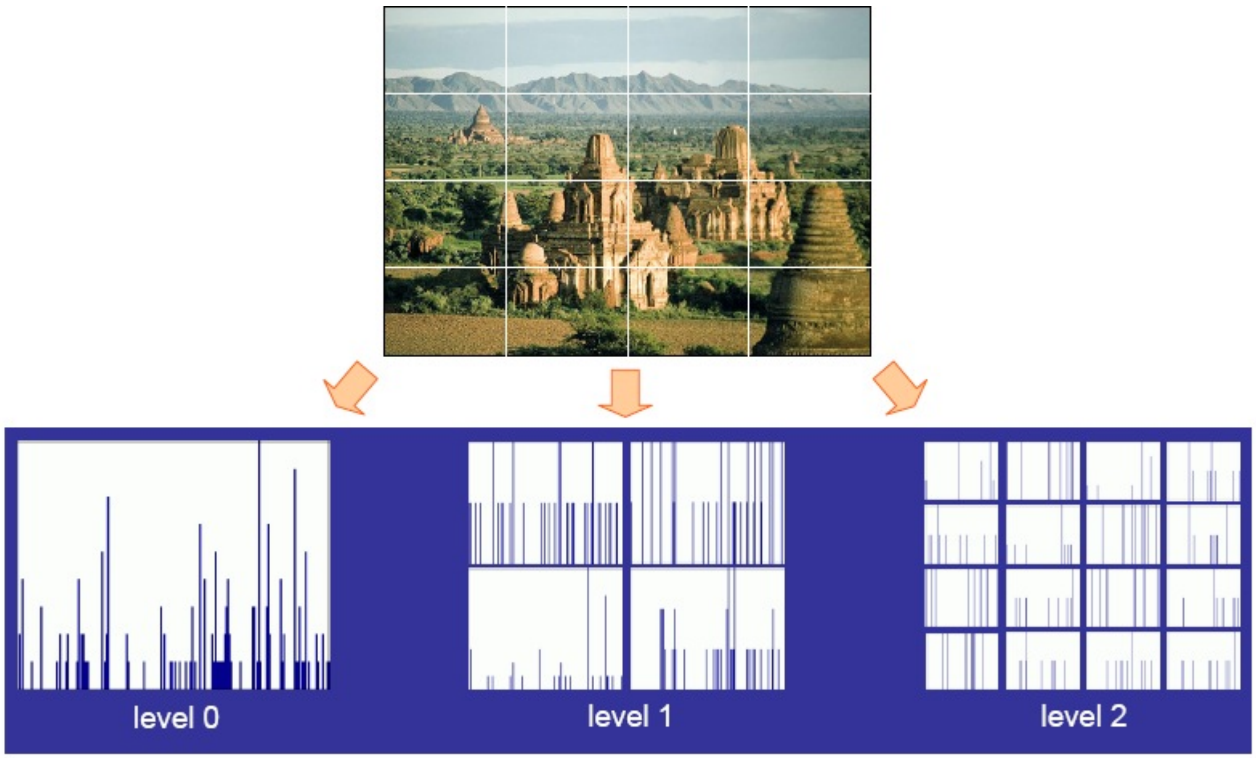
- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007





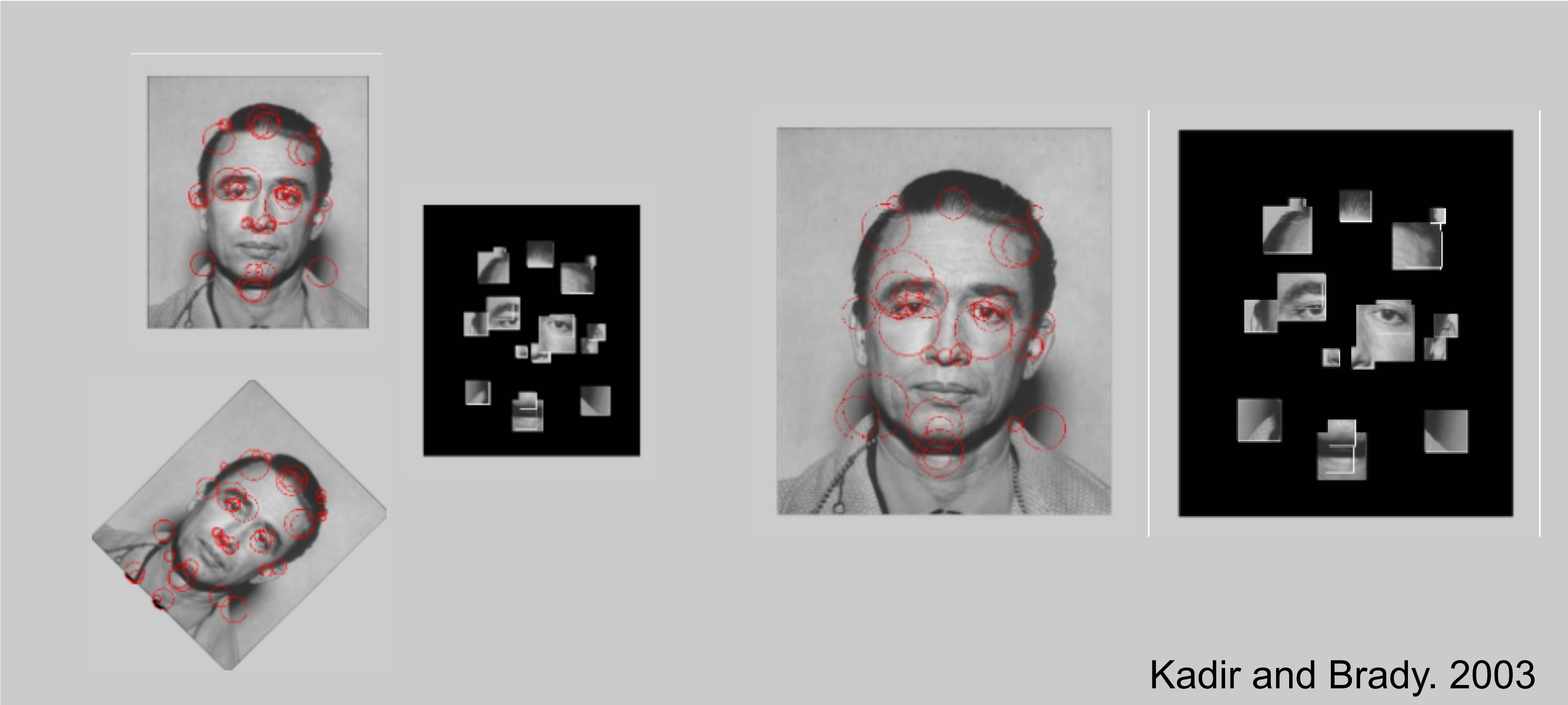
What about spatial info?

- 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

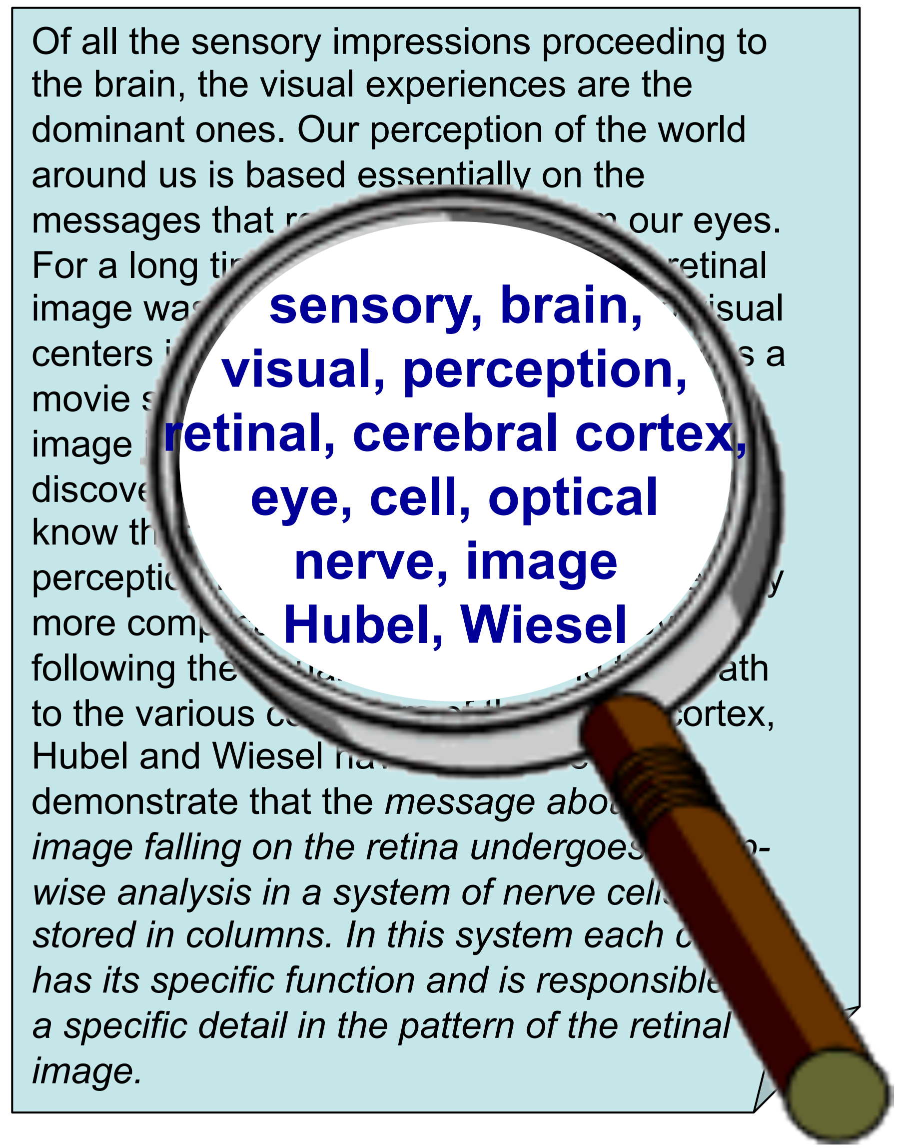
Invariance issues

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

Fergus, Fei-Fei, Perona & Zisserman, 2005

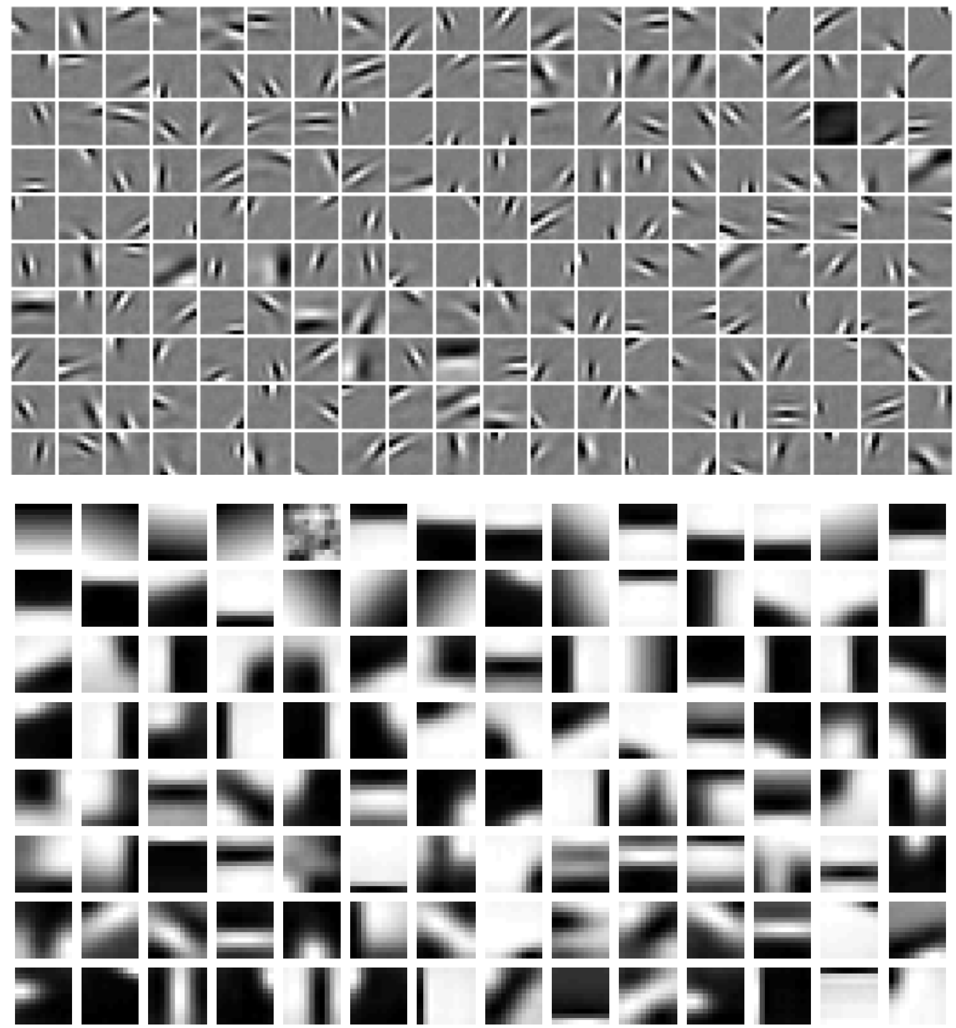
Model properties

- Intuitive
  - Analogy to documents



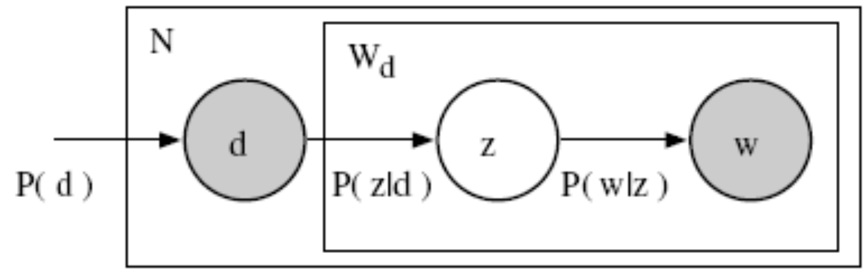
Model properties

- Intuitive
  - Analogy to documents
  - Analogy to human vision



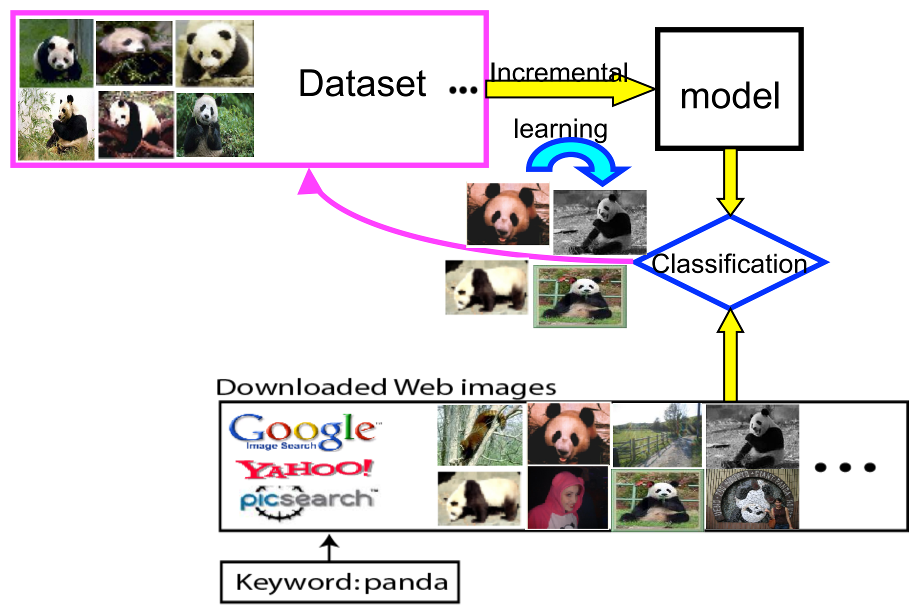
Olshausen and Field, 2004, Fei-Fei and Perona, 2005

Model properties



Sivic, Russell, Efros, Freeman, Zisserman, 2005

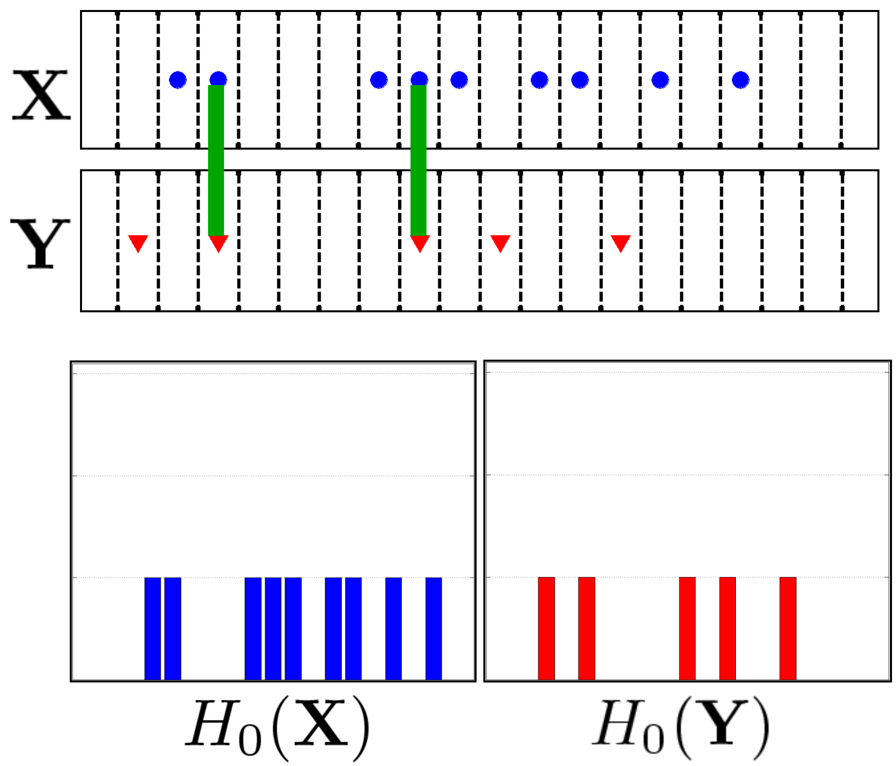
- Intuitive
- generative models
  - Convenient for weakly- or un-supervised, incremental training
  - Prior information
  - Flexibility (e.g. HDP)



Li, Wang & Fei-Fei, CVPR 2007

Model properties

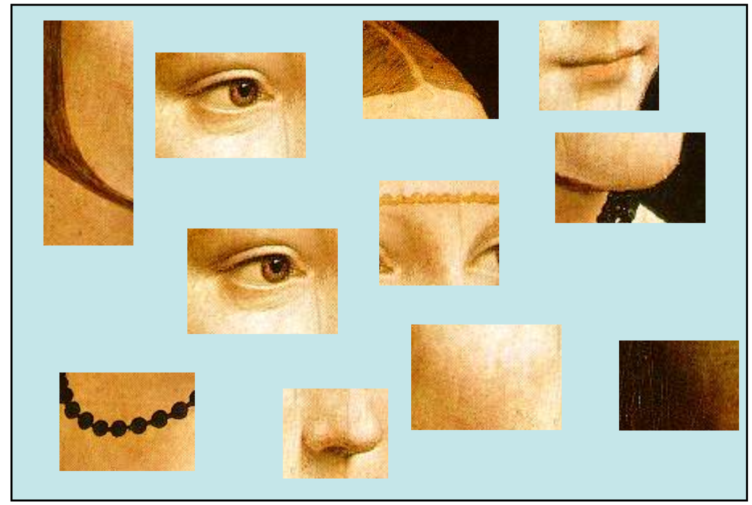
- Intuitive
- generative models
- Discriminative method
  - Computationally efficient

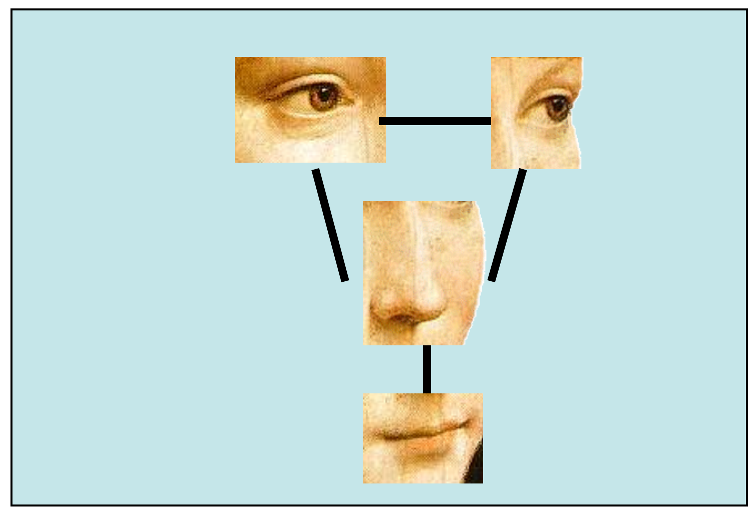


Grauman et al. CVPR 2005

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