Discovering Meaning in the Visual World

Fei-Fei Li

(publish under L. Fei-Fei)

A picture is worth a thousand words.

--- Confucius

or *Printers’ Ink* Ad (1921)

blue

rugged

white and red

bright

textured structure

green

solid

elongated shapes

grey

- To build intelligent visual algorithms for machines and robots

- To understand human visual intelligence by applying computational tools

Outline: it’s all about ‘categorization’

objects

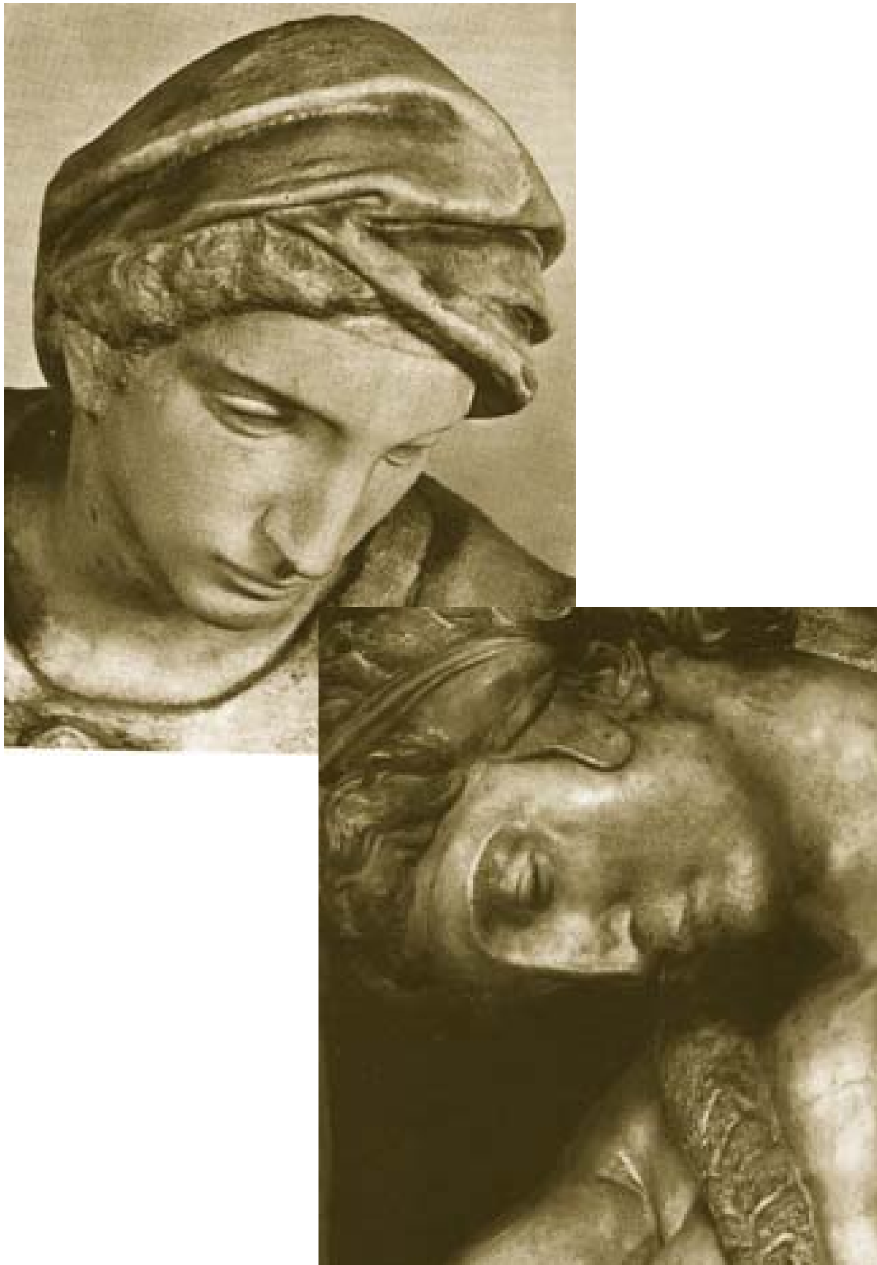
scenes

actions

events

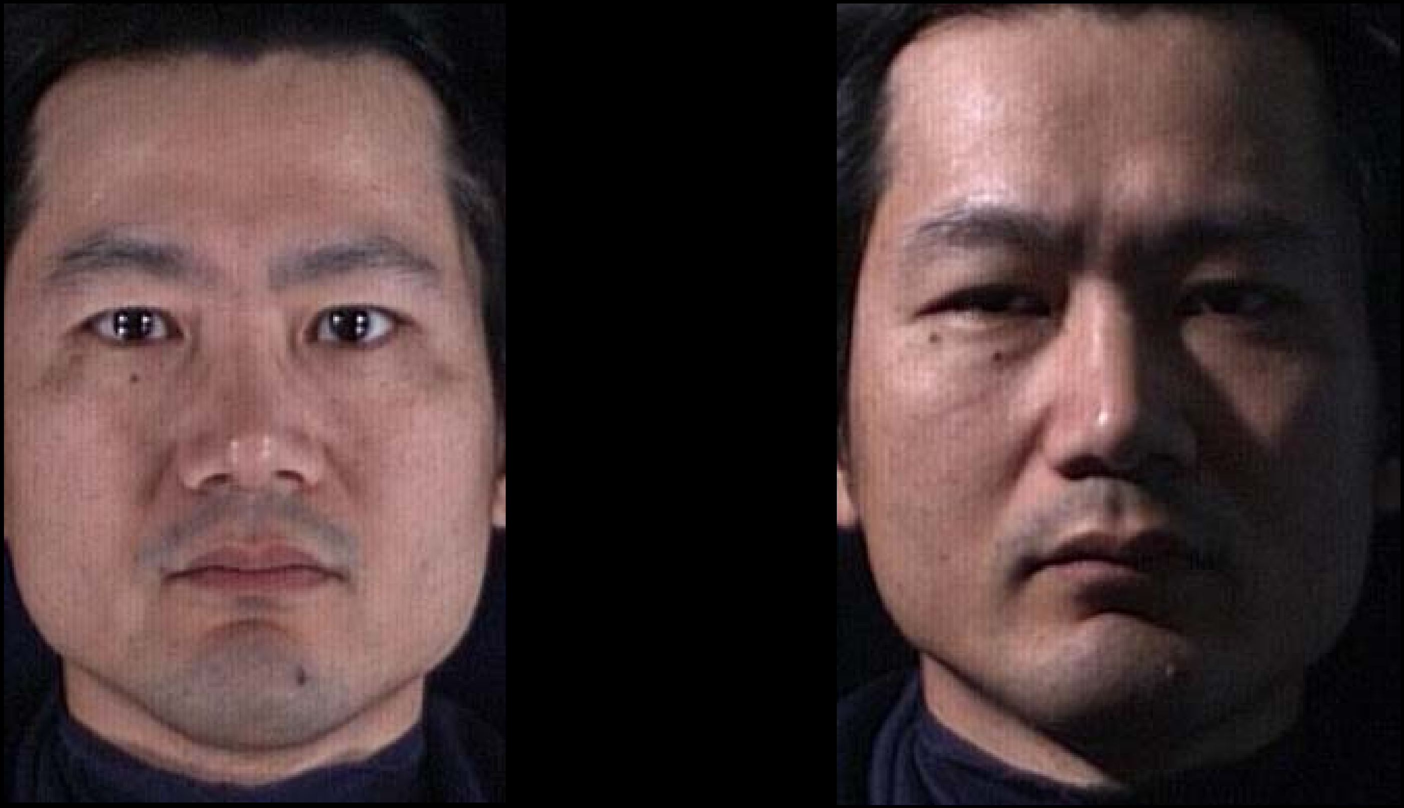
Objects are hard to recognize

- – View point



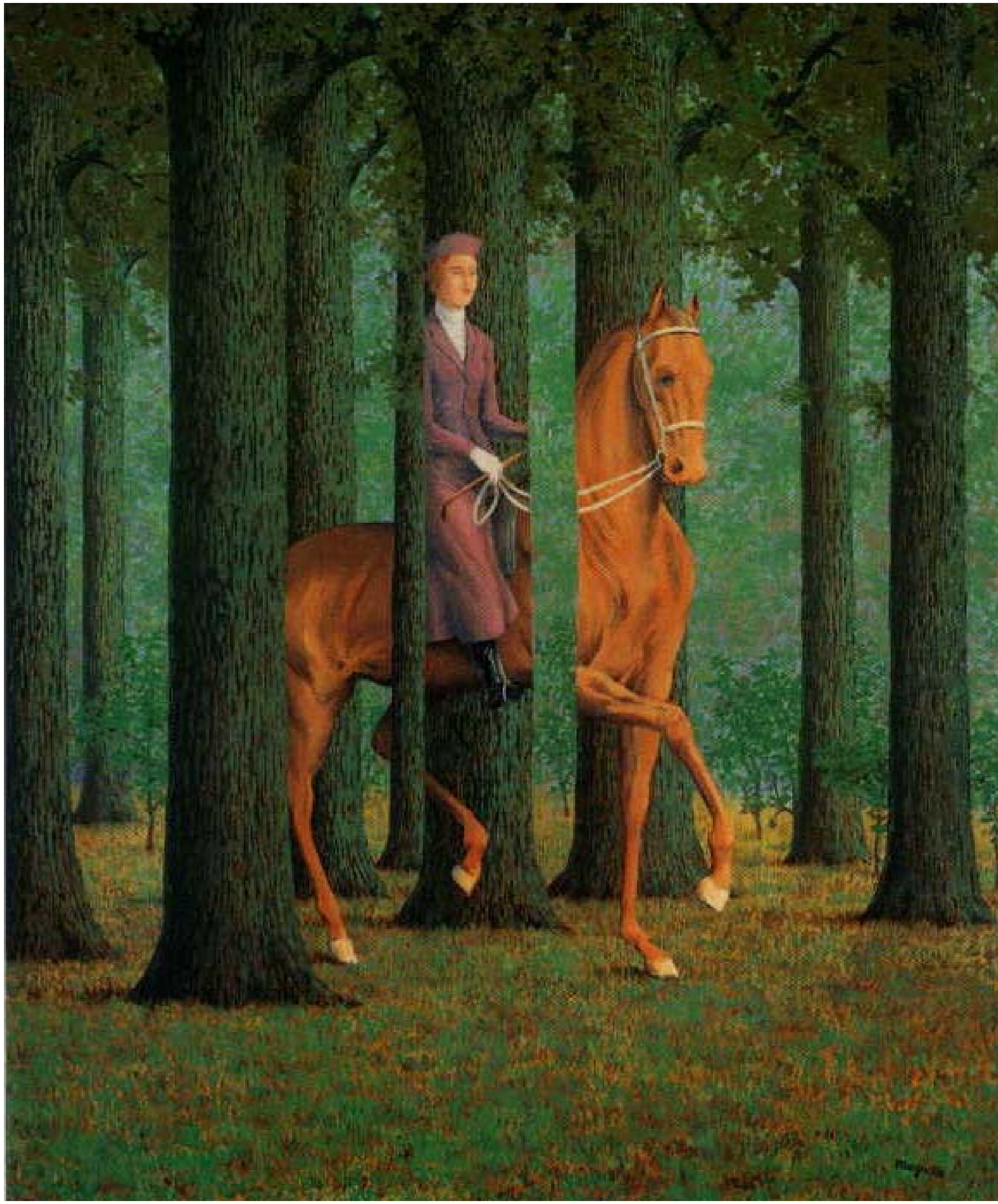
Objects are hard to recognize

- View point
- Illumination



Objects are hard to recognize

- View point
- Illumination
- Occlusion

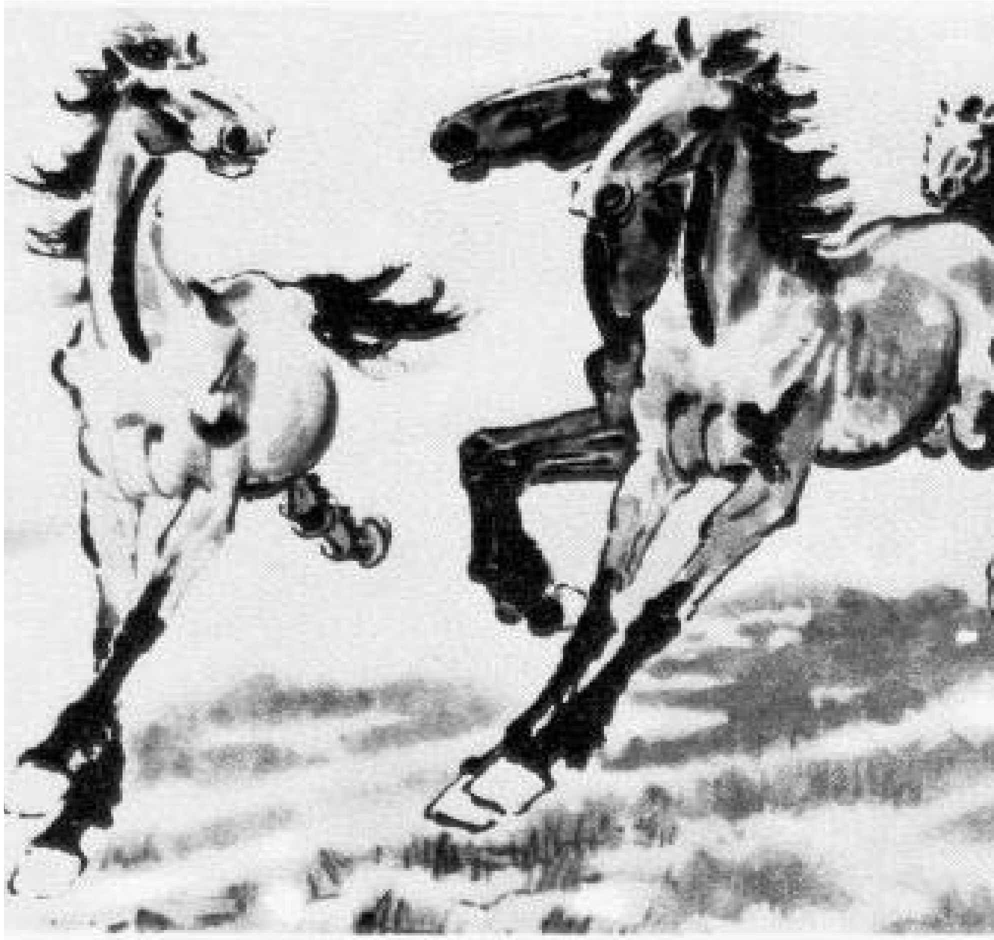


Objects are hard to recognize

- – View point
- – Illumination
- – Occlusion
- – Scale

Objects are hard to recognize

- – View point
- – Illumination
- – Occlusion
- – Scale
- – Deformation



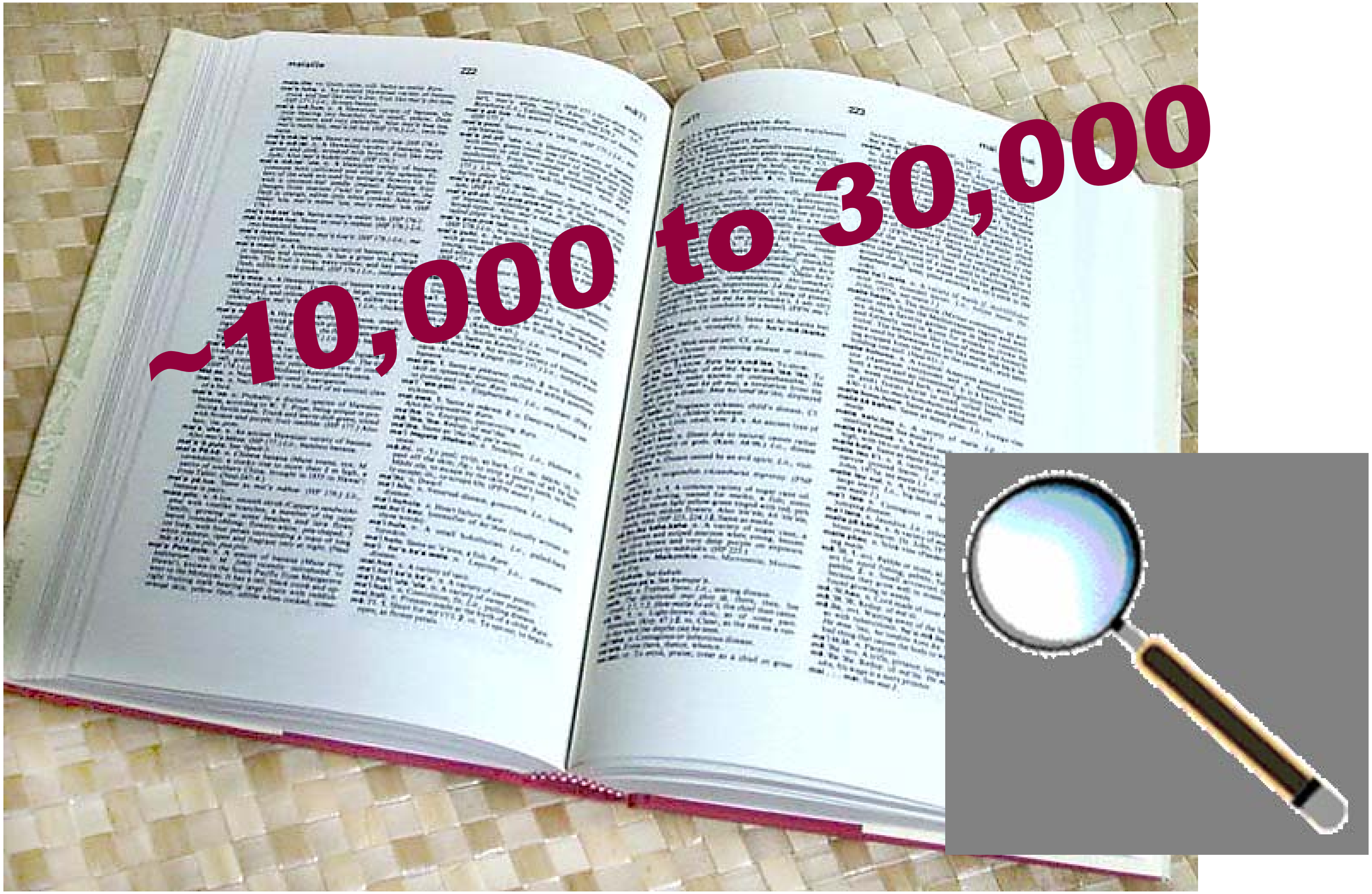
Objects are hard to recognize

- – View point
- – Illumination
- – Occlusion
- – Scale
- – Deformation
- – Clutter

Objects are hard to recognize

- – View point
- – Illumination
- – Occlusion
- – Scale
- – Deformation
- – Clutter
- – Intra-class variability

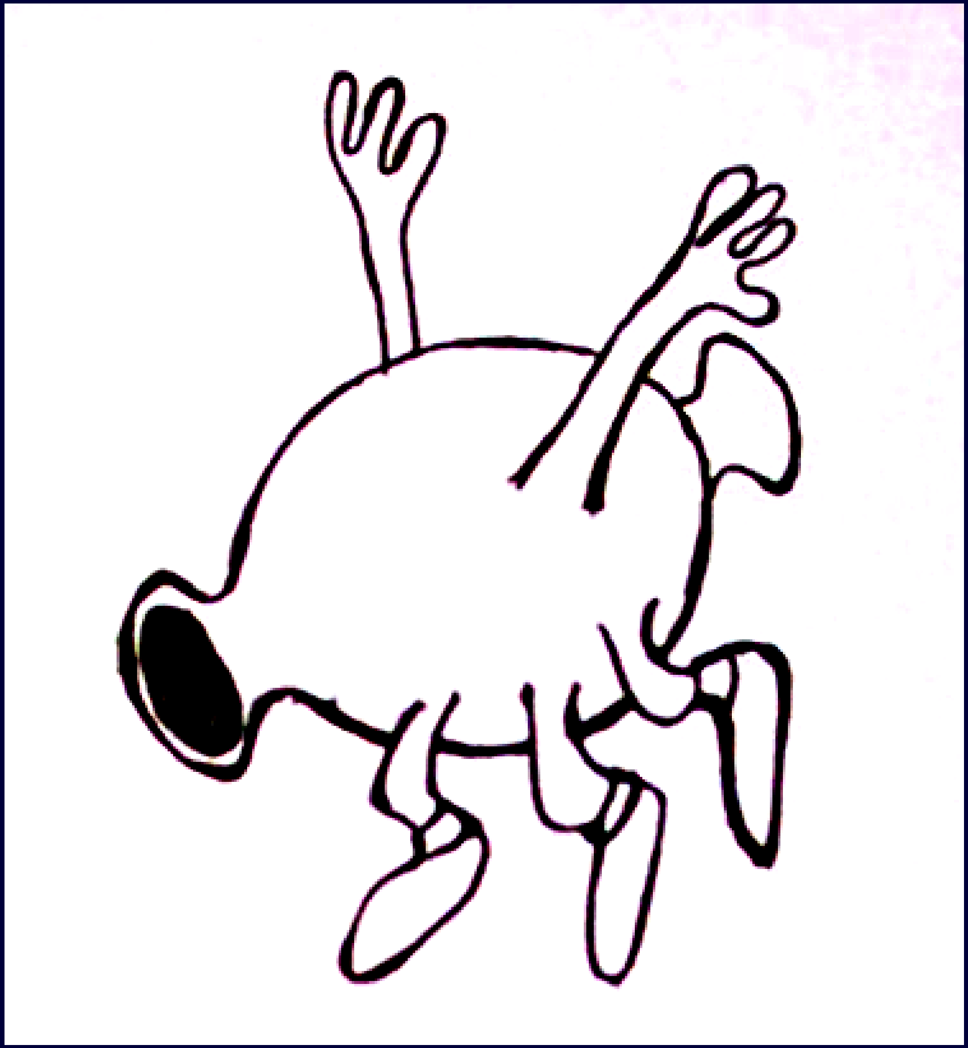
How many object categories are there?



~10,000 to 30,000

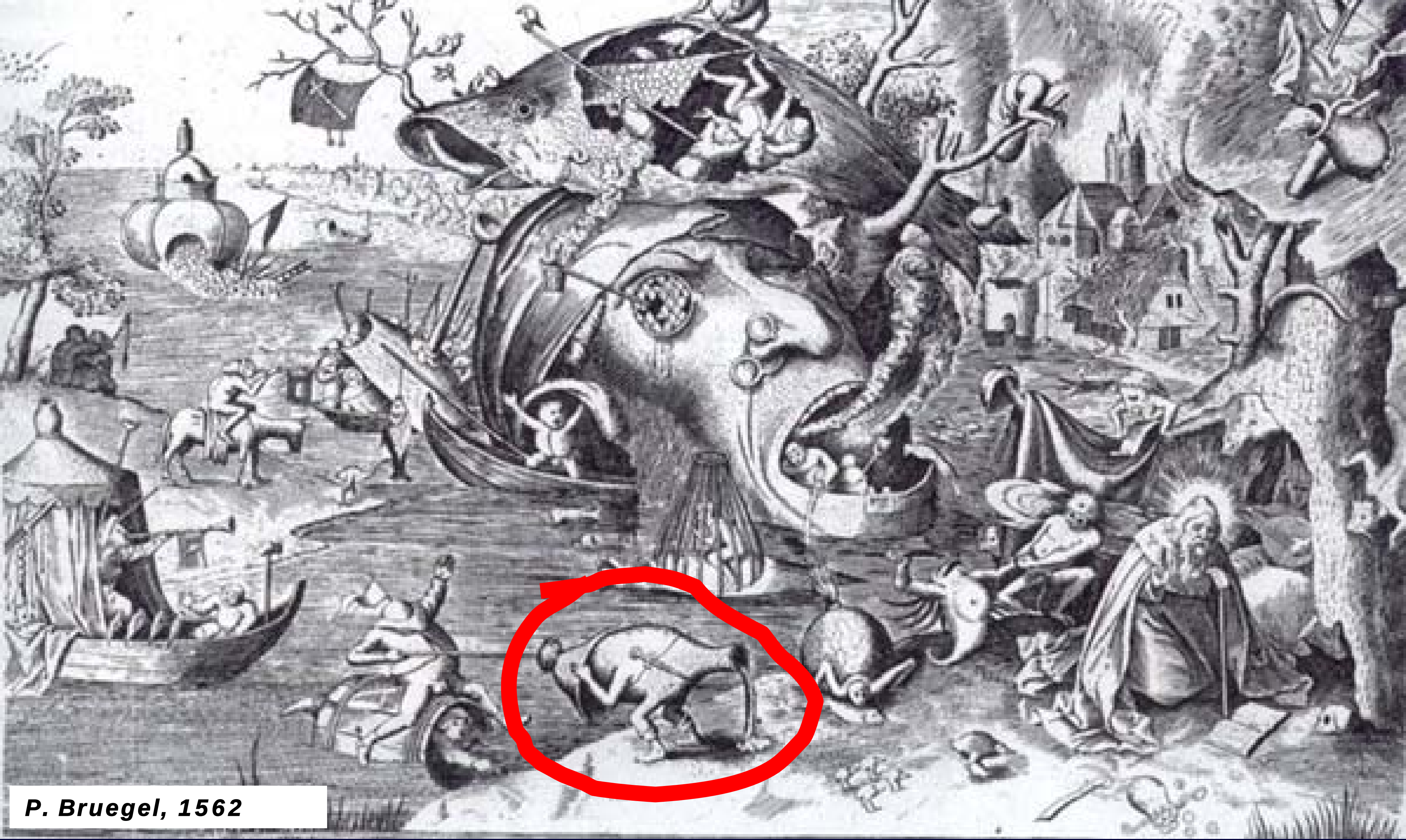
Biederman 1987

Algorithm	Training Examples	Categories
Rowley et al.	~500	Faces
Schneiderman, et al.	~2,000	Faces, Cars
Viola et al.	~10,000	Faces
Burl, et al. Weber, et al. Fergus, et al.	200 ~ 400	Faces, Motorbikes, Spotted cats, Airplanes, Cars



One-shot learning of object categories

Fei-Fei et al. ‘03, ‘04, ‘06



P. Bruegel, 1562

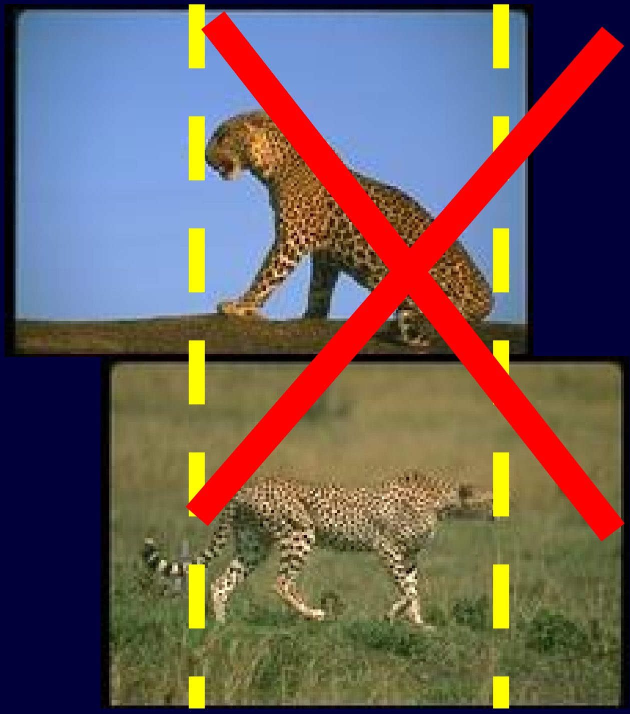
One-shot learning of object categories

Fei-Fei et al. ‘03, ‘04, ‘06

No labeling

No segmentation

No alignment

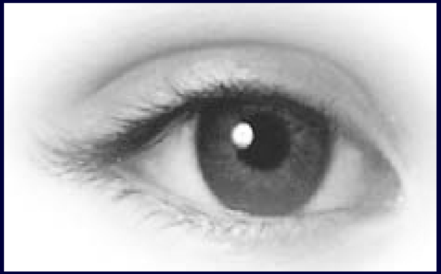


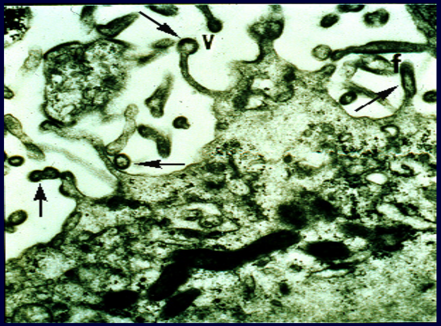
One-shot learning of object categories

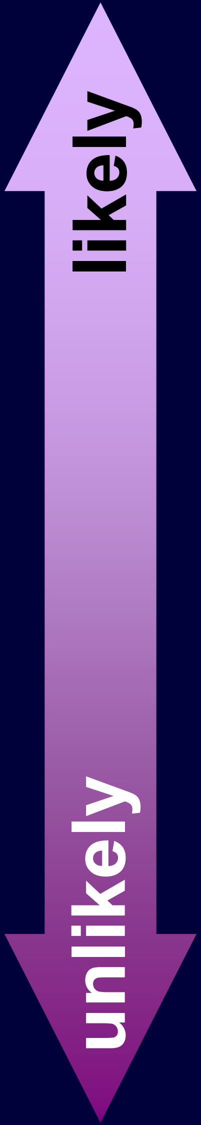
Fei-Fei et al. ‘03, ‘04, ‘06

Prior knowledge about objects

Appearance



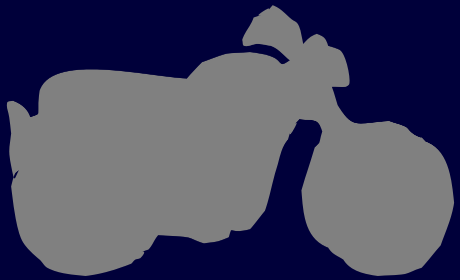




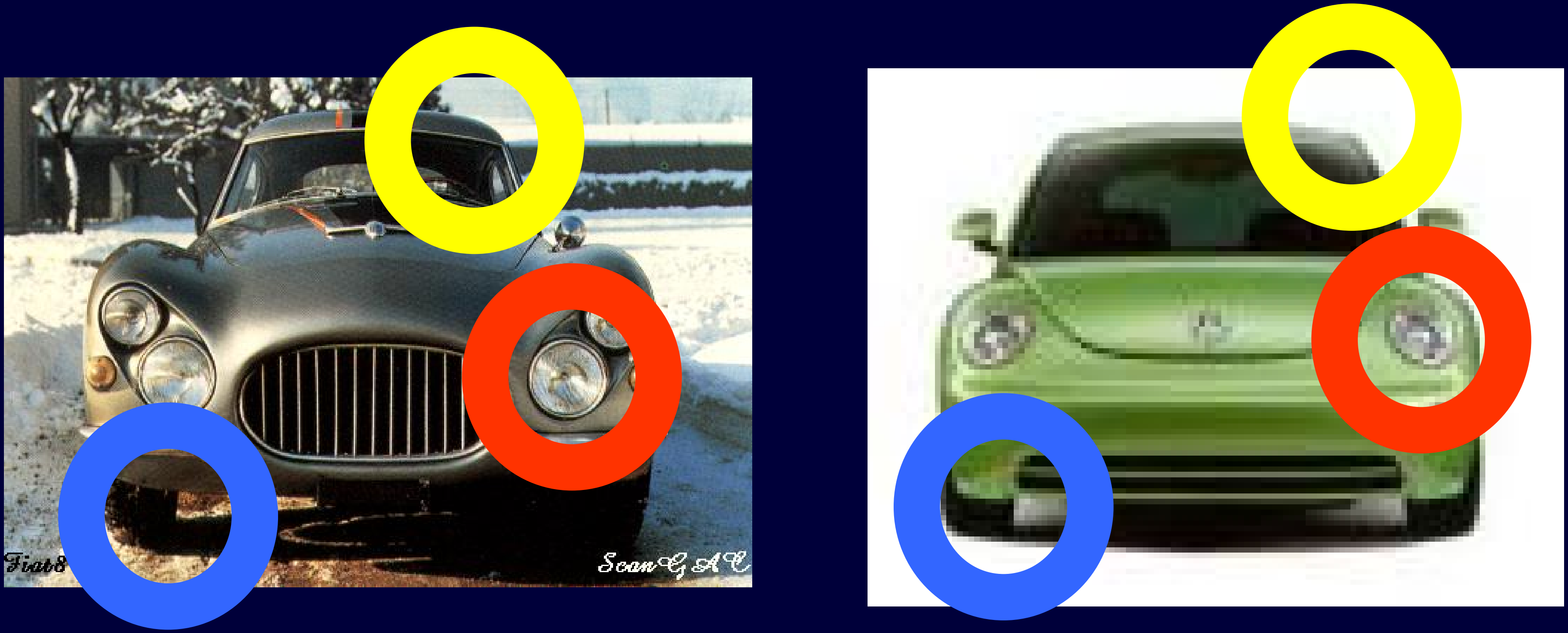
likely

unlikely

Shape

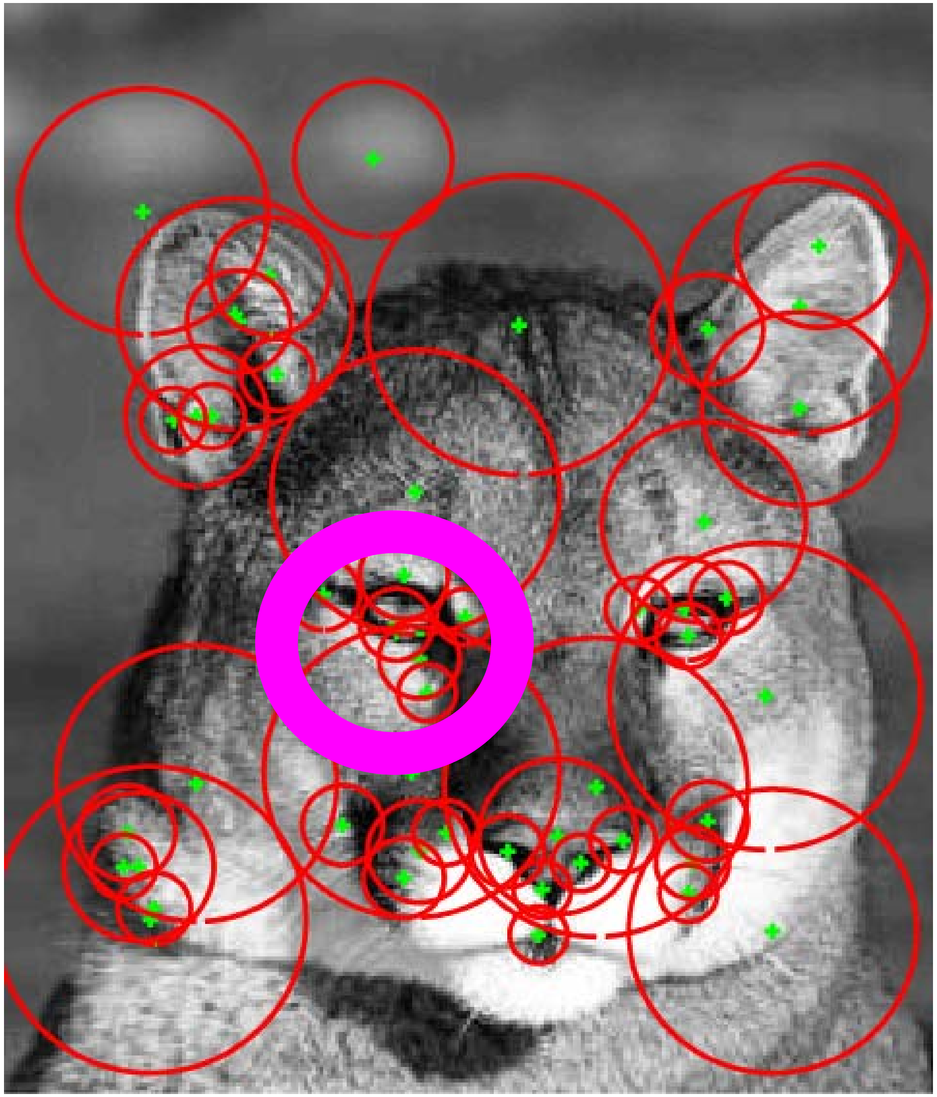


model representation



One-shot learning of object categories

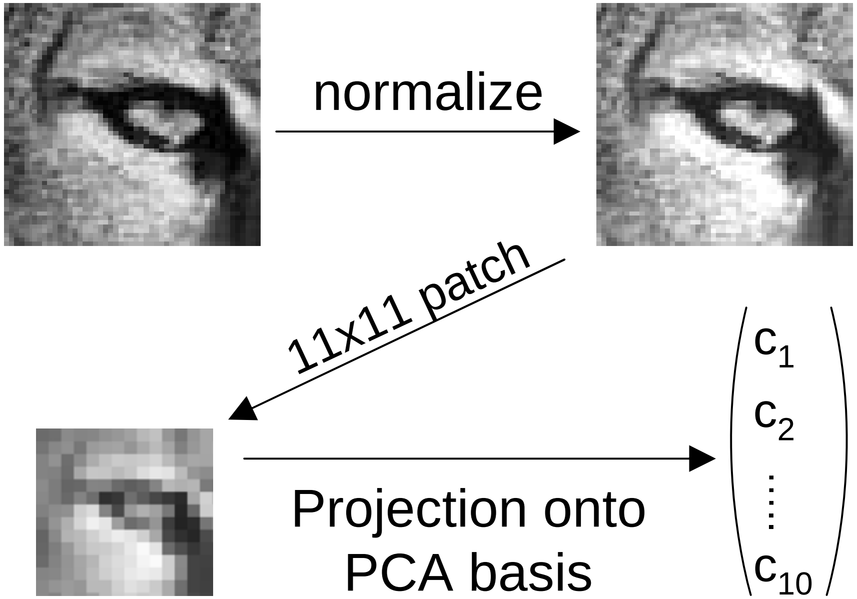
Fei-Fei et al. ‘03, ‘04, ‘06



X (location)

(x,y) coords. of region center

A (appearance)



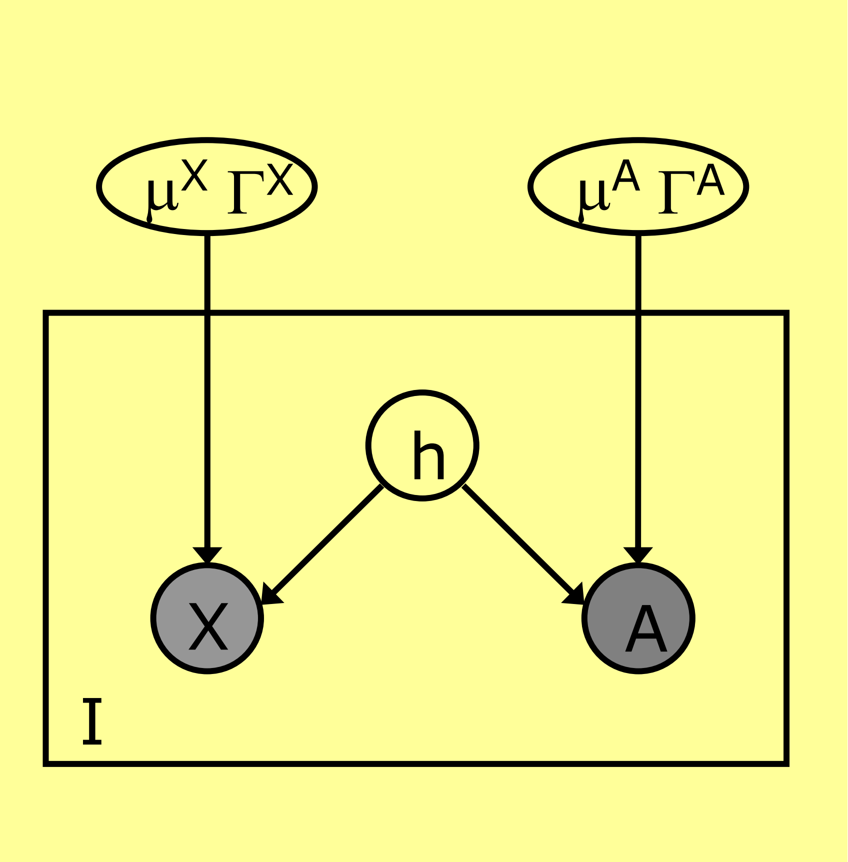
normalize

11x11 patch

Projection onto PCA basis

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{10} \end{pmatrix}$$

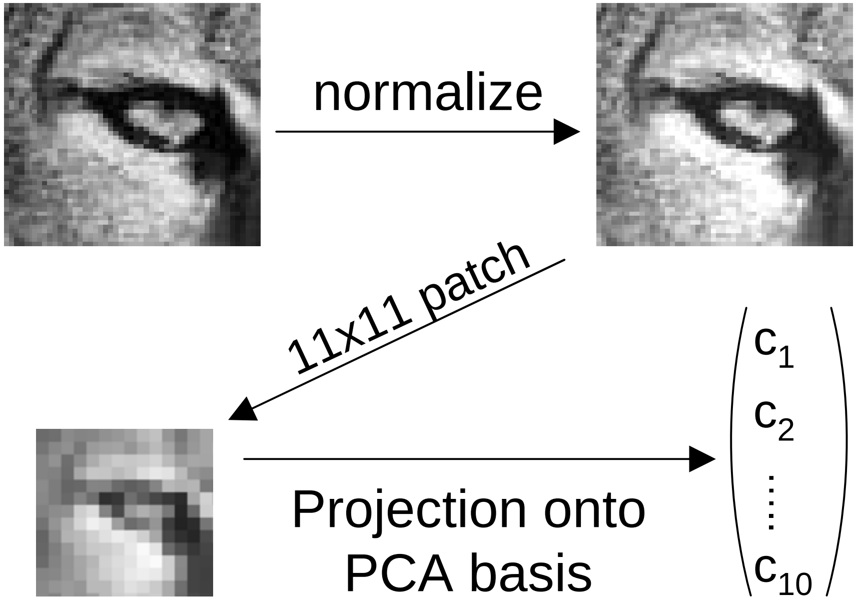
The Generative Model



X (location)

(x,y) coords. of region center

A (appearance)



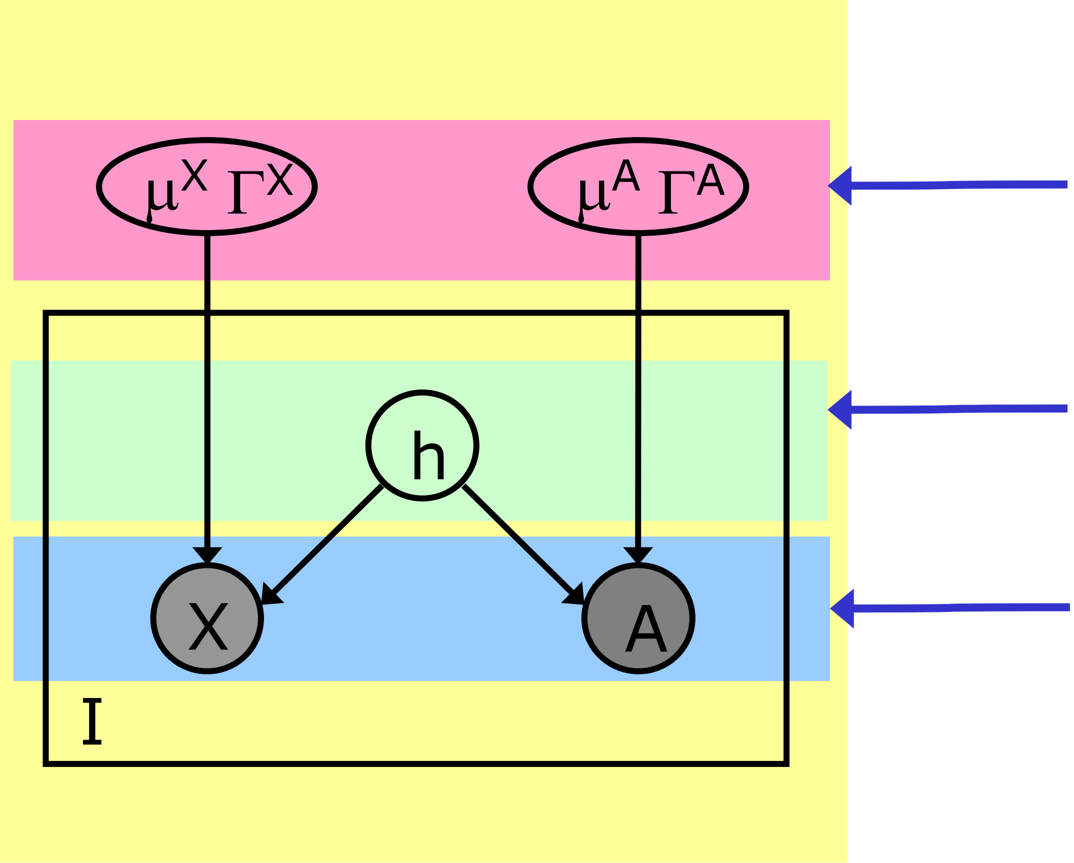
normalize

11x11 patch

Projection onto PCA basis

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{10} \end{pmatrix}$$

The Generative Model

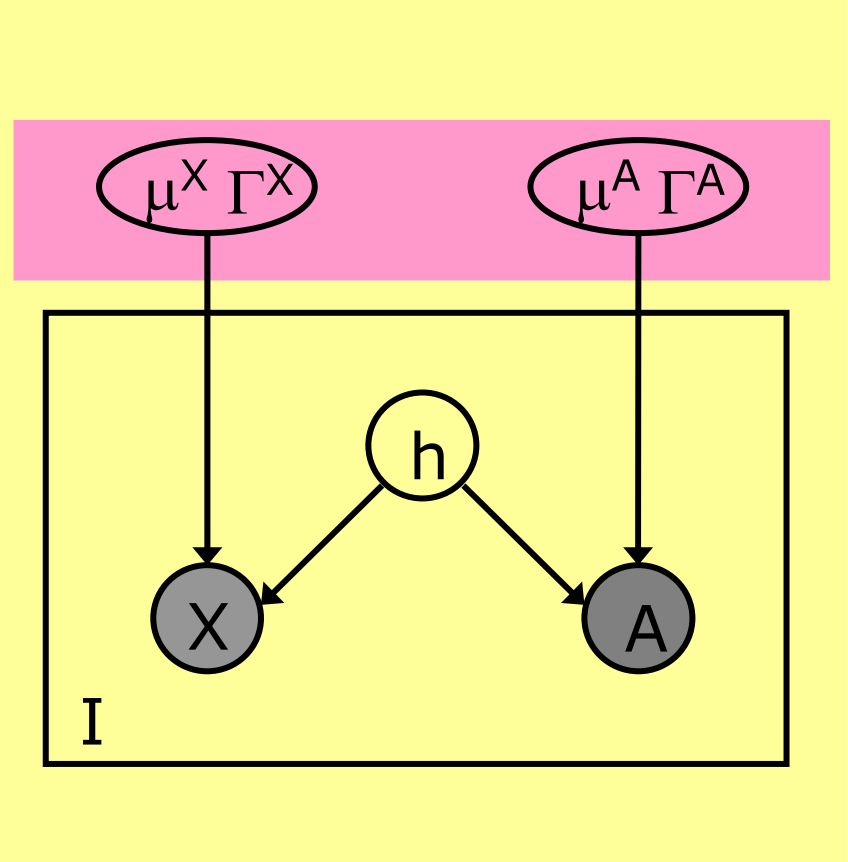


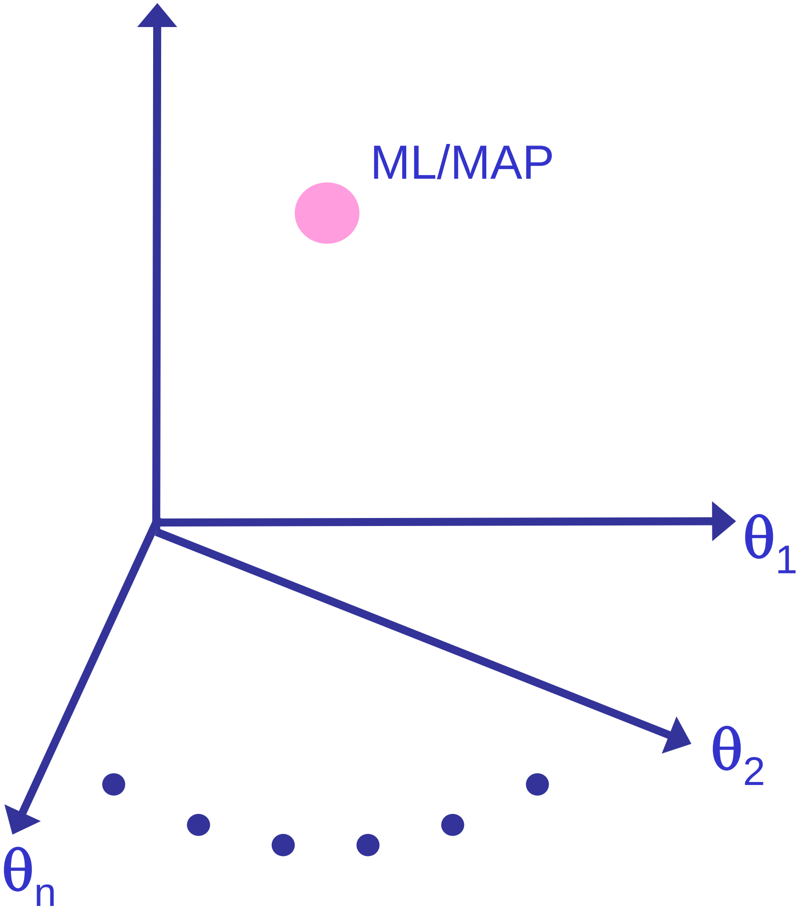
parameters

hidden variable

observed variables

The Generative Model

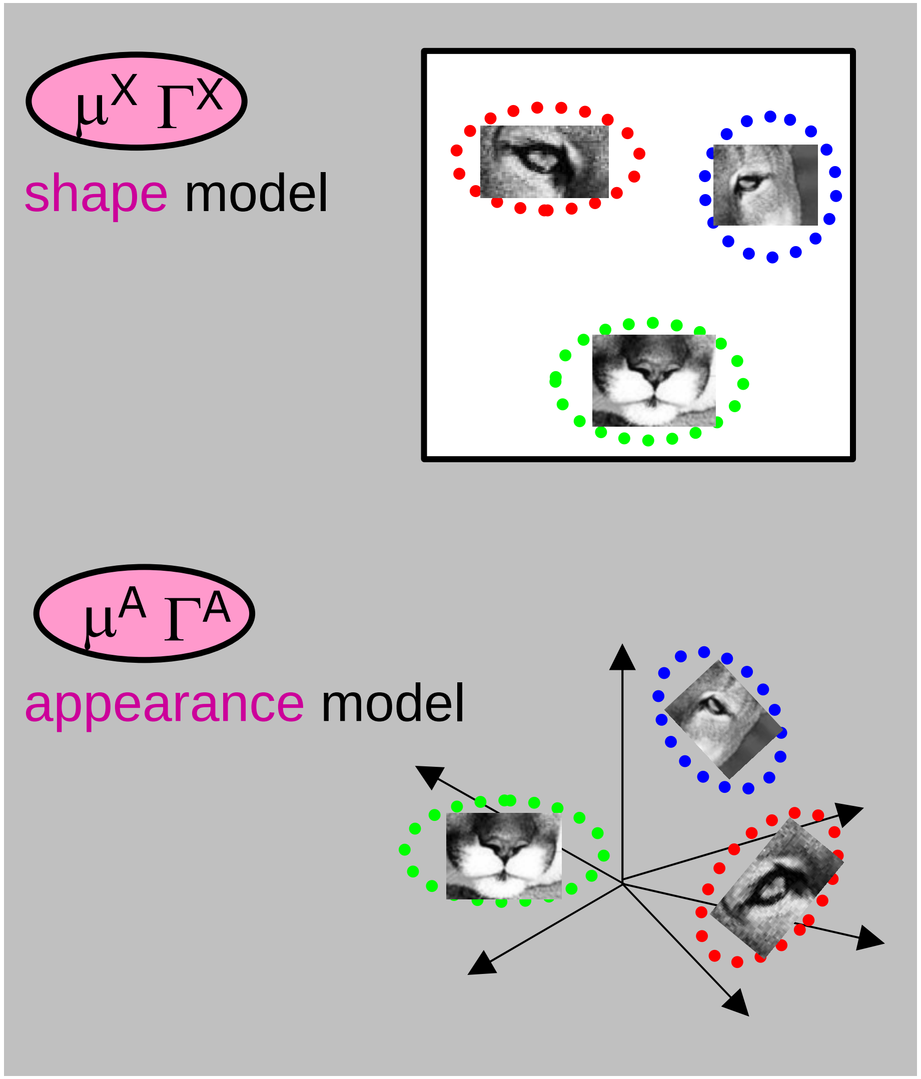




where $\theta = \{\mu^X, \Gamma^X, \mu^A, \Gamma^A\}$

Weber et al. '98 '00, Fergus et al. '03

The Generative Model

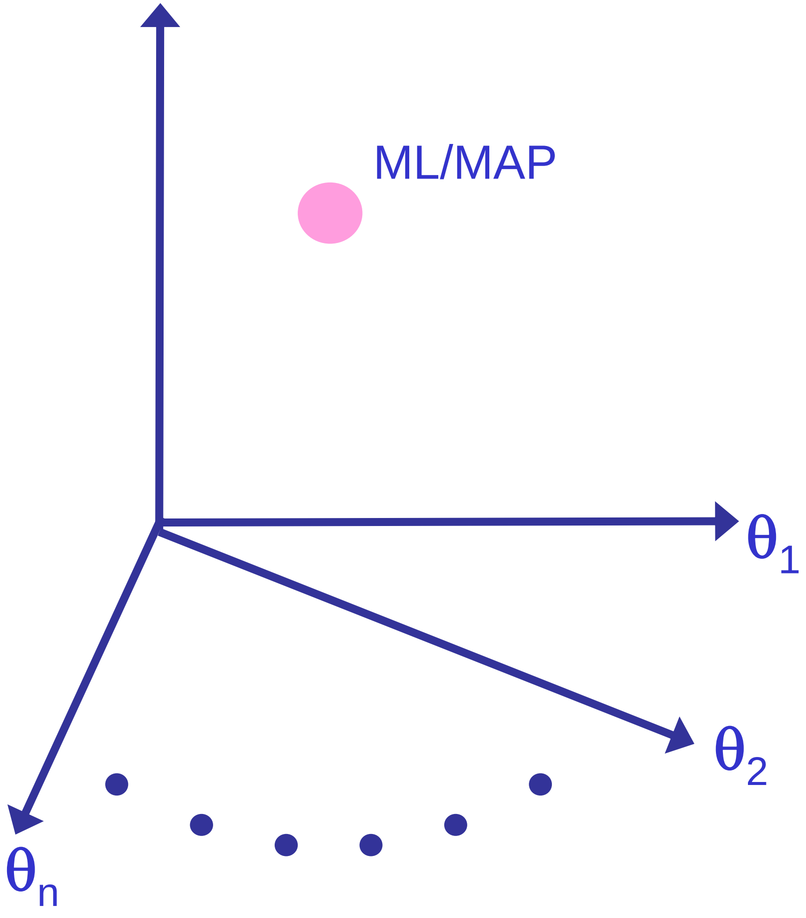


$\mu^X \; \Gamma^X$

shape model

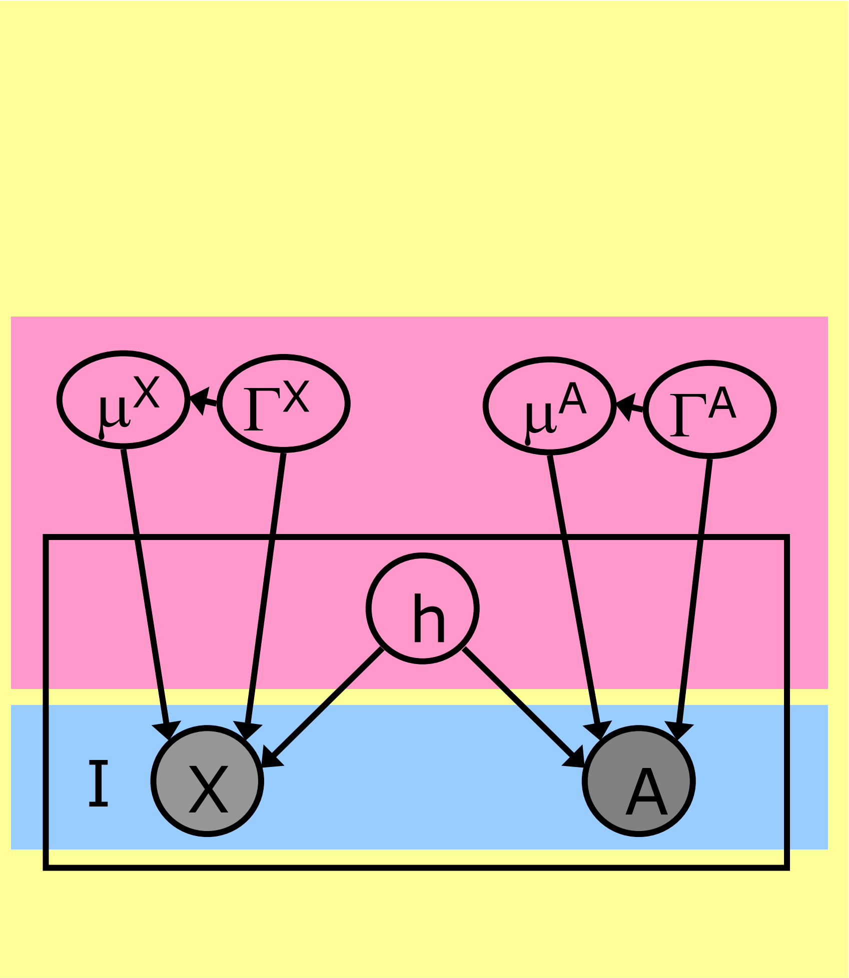
$\mu^A \; \Gamma^A$

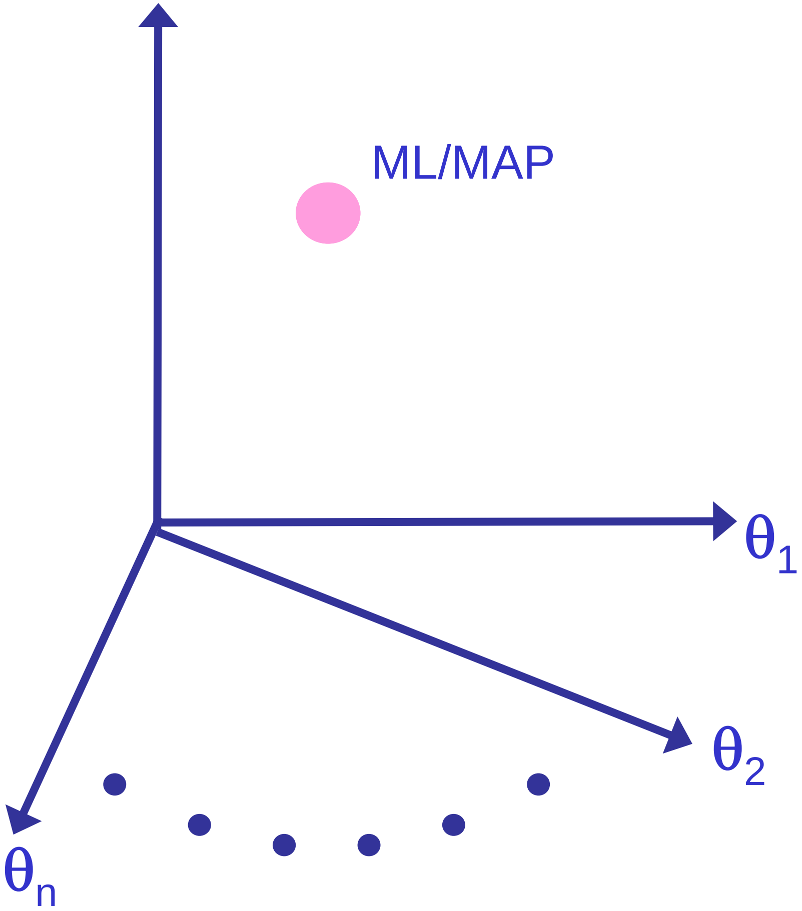
appearance model



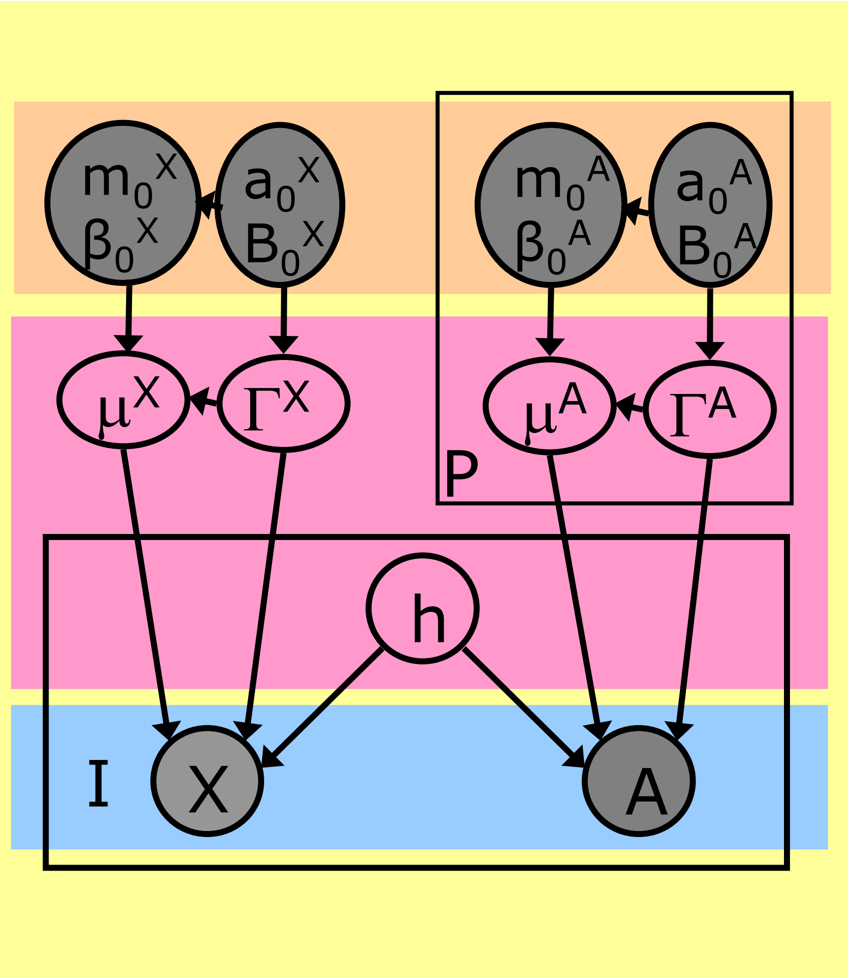
where $\theta = \{\mu^X, \Gamma^X, \mu^A, \Gamma^A\}$

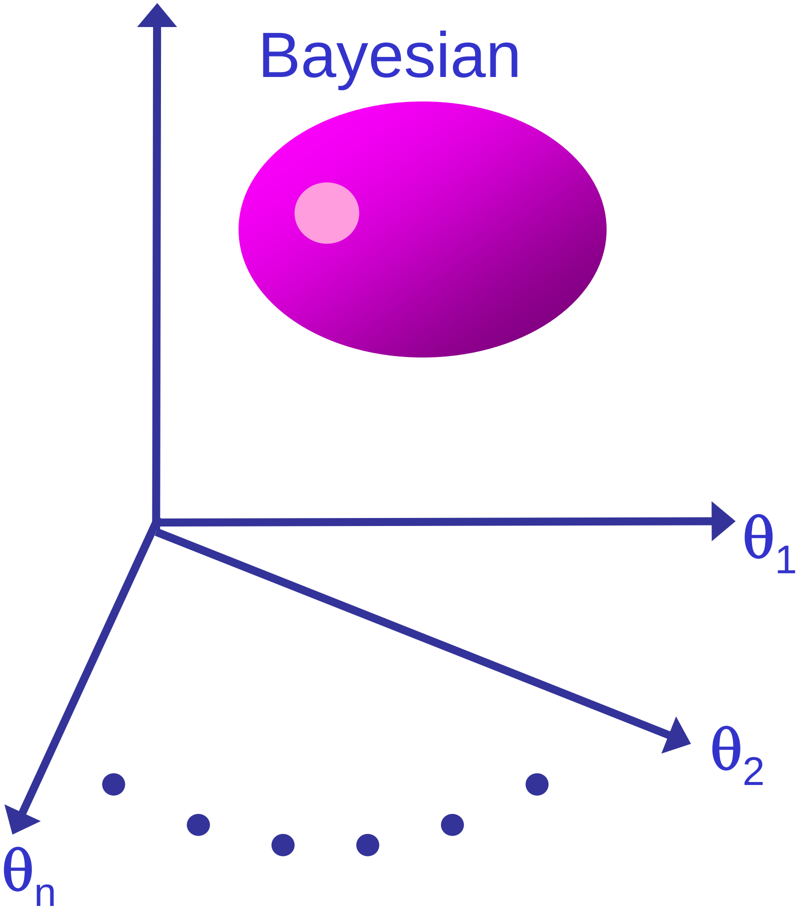
The Generative Model





The Generative Model

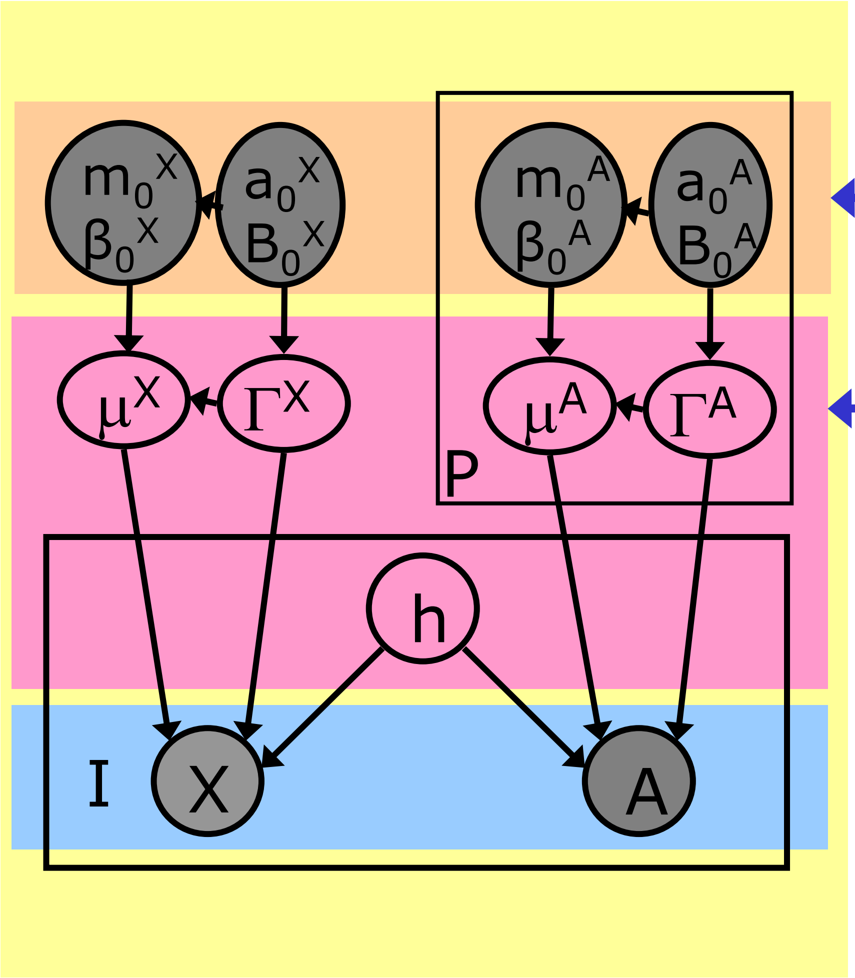




Fei-Fei et al. '03, '04, '06

Parameters to estimate: {m^X, β^X, a^X, B^X, m^A, β^A, a^A, B^A}
i.e. parameters of Normal-Wishart distribution

The Generative Model

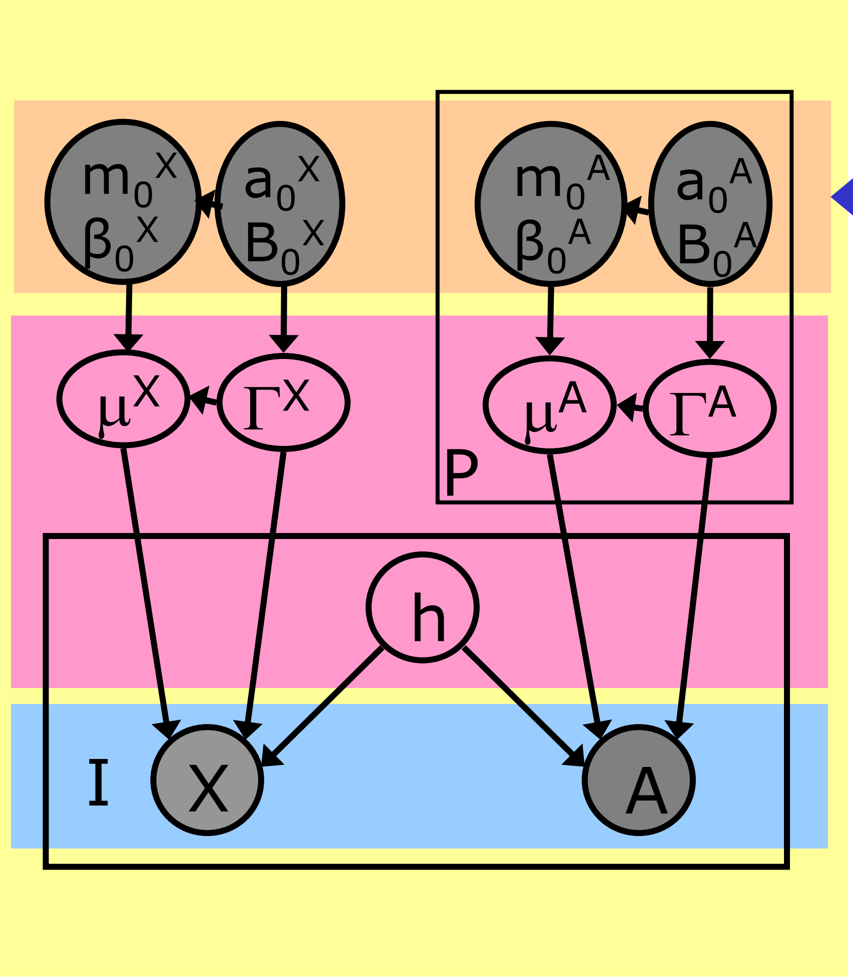


priors

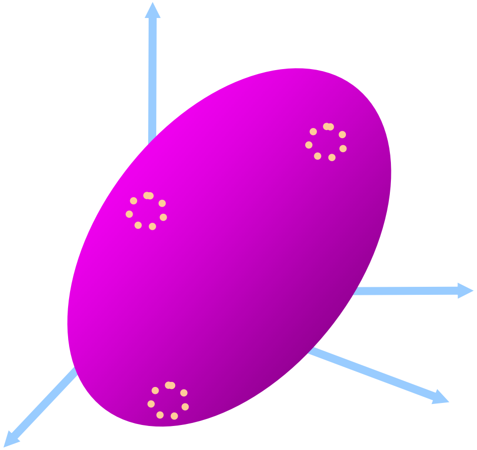
parameters

Fei-Fei et al. ‘03, ‘04, ‘06

The Generative Model

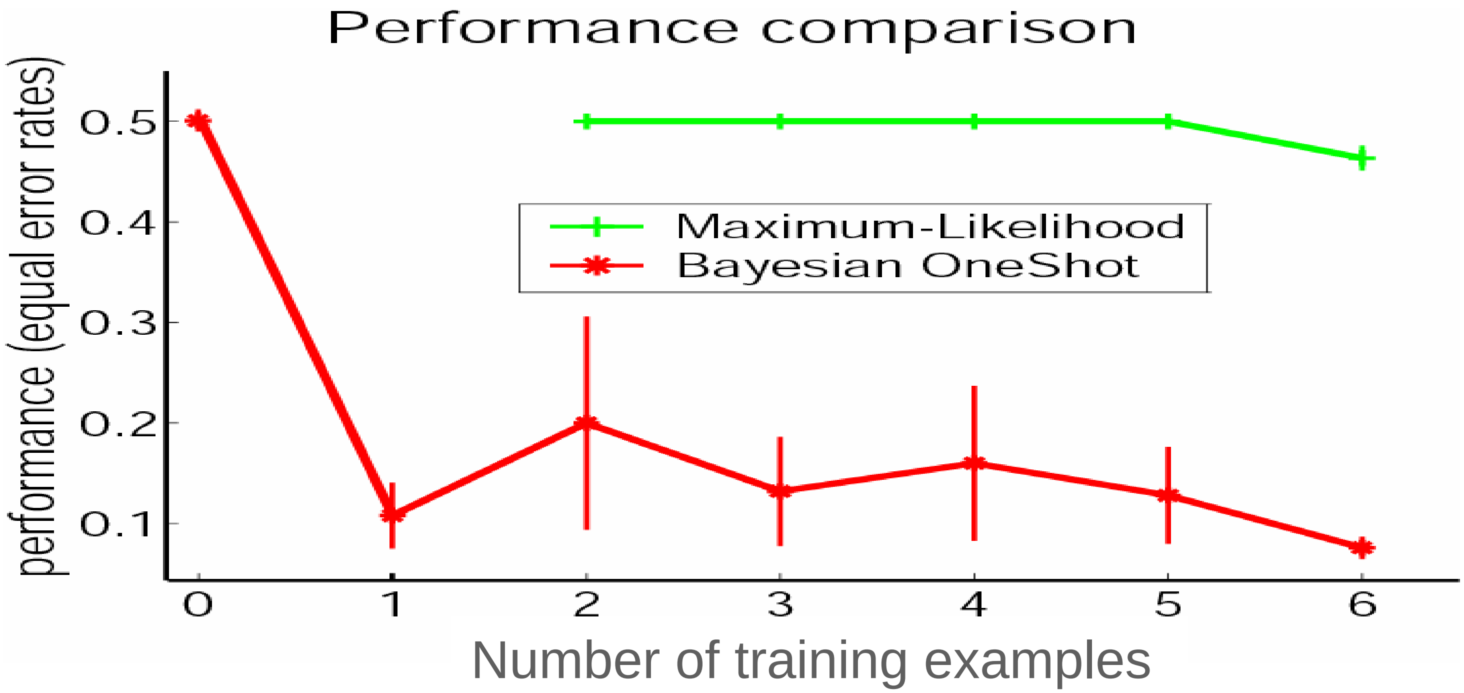


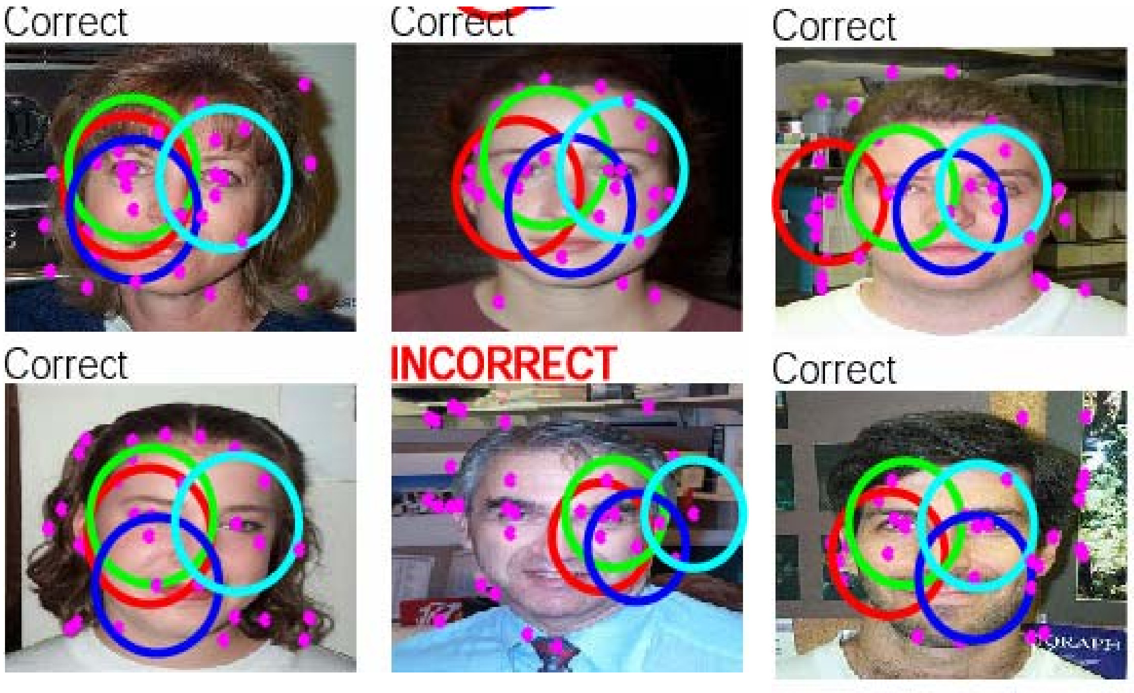
priors



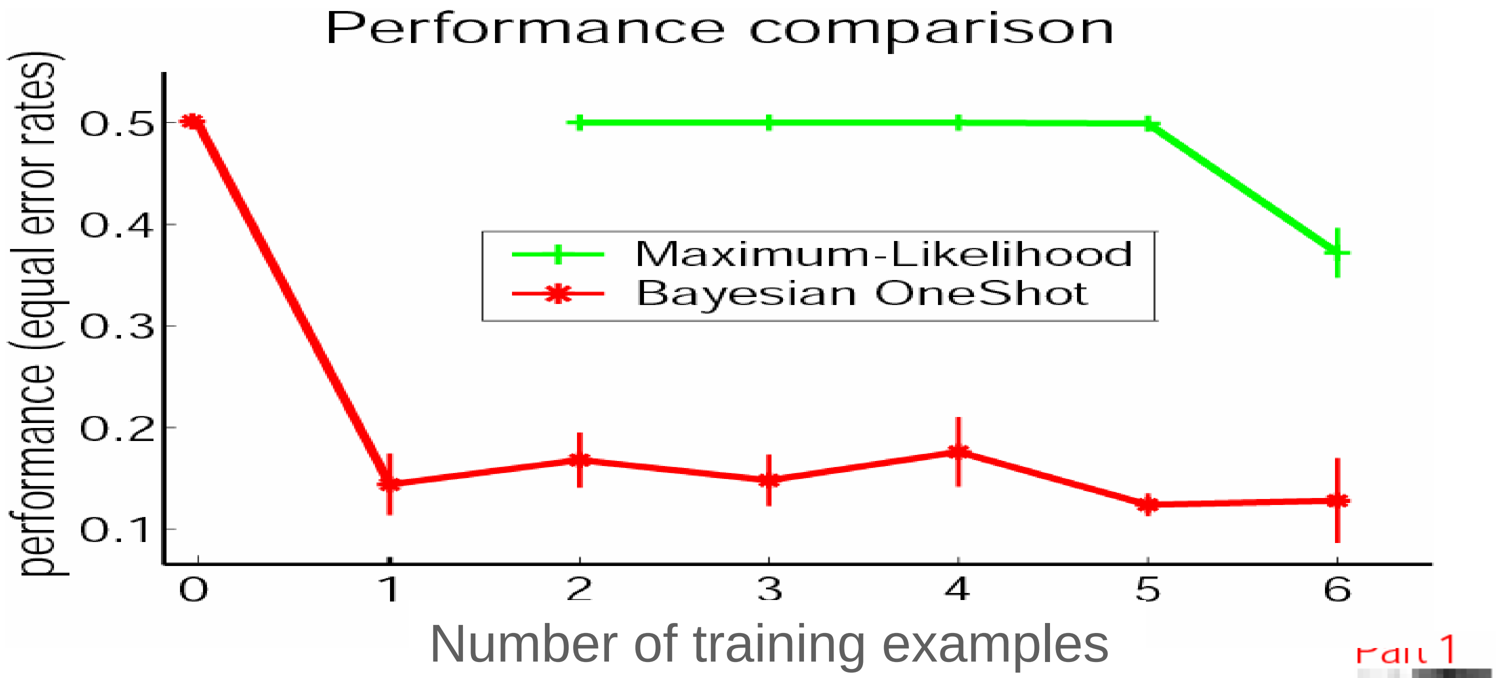
Prior distribution

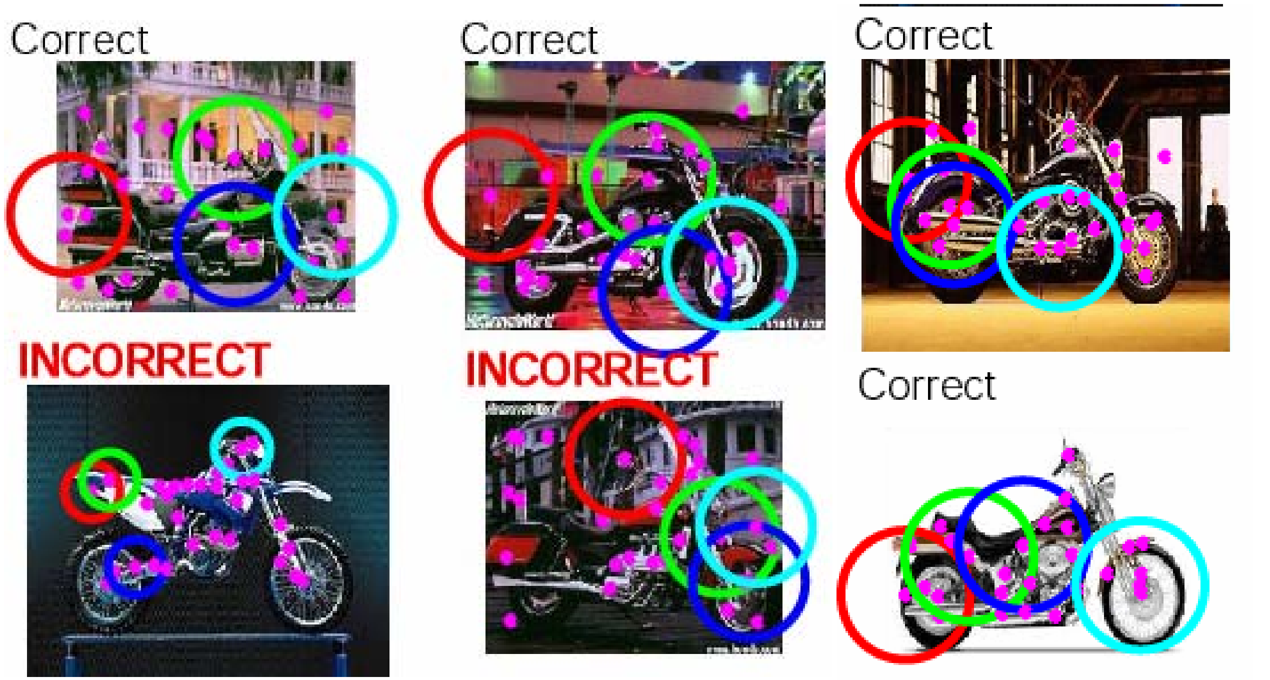
Fei-Fei et al. ‘03, ‘04, ‘06


Performance comparison
performance (equal error rates)
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
Number of training examples
Maximum-Likelihood
Bayesian OneShot


Correct
Correct
Correct
Correct
INCORRECT
Correct

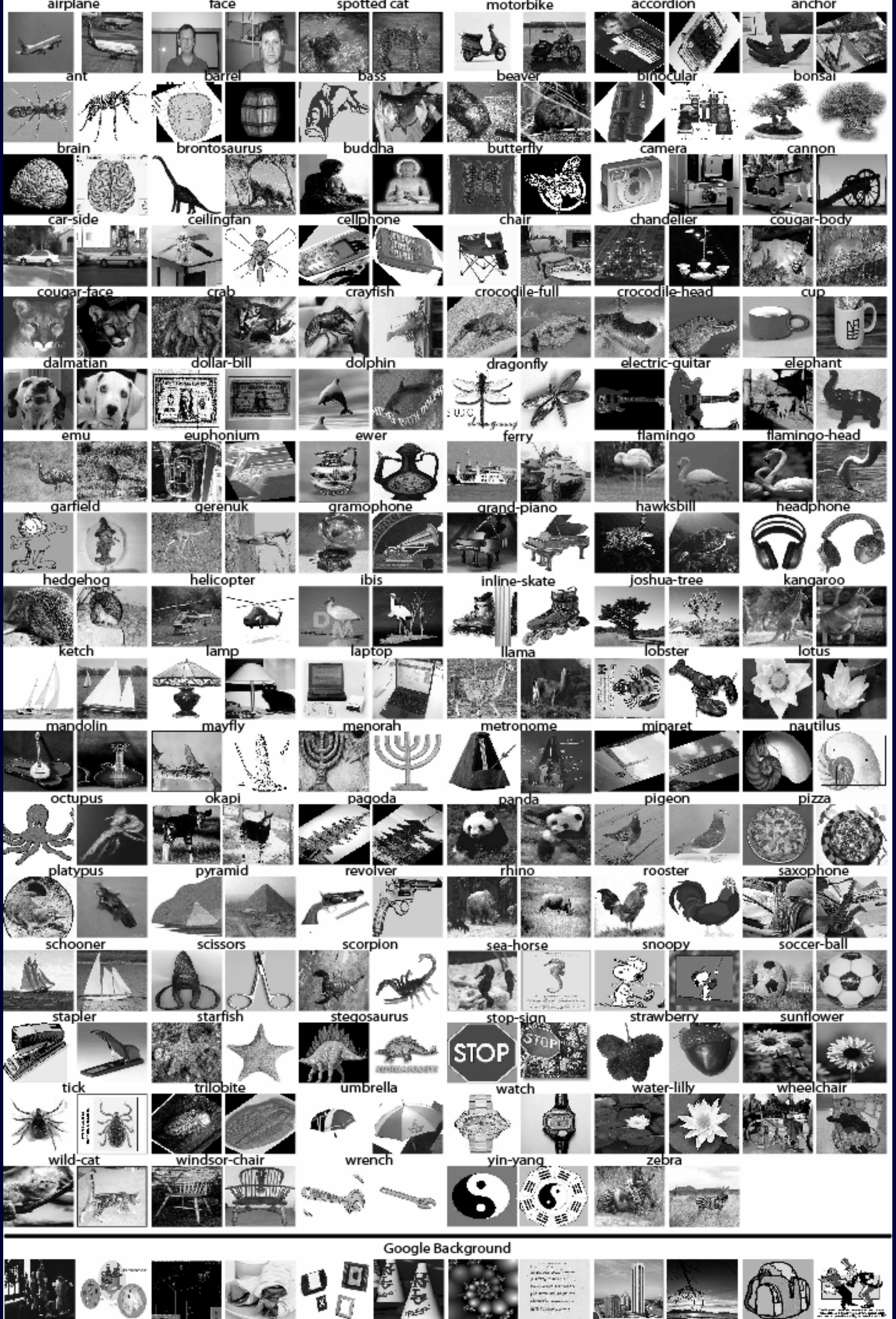
Part 1
Part 2
Part 3
Part 4


Performance comparison
performance (equal error rates)
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
Number of training examples
Maximum-Likelihood
Bayesian OneShot


Correct
Correct
Correct
INCORRECT
INCORRECT
Correct

Part 1
Part 2
Part 3
Part 4

Caltech101 dataset

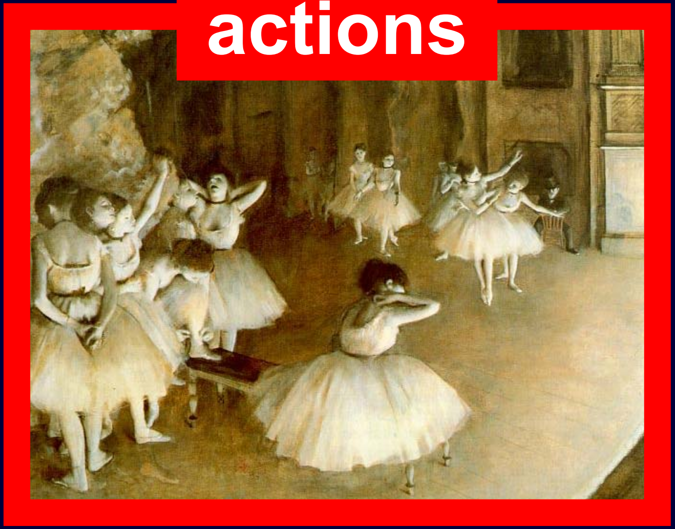


Fei-Fei et al. 2004

Outline: it’s all about ‘categorization’

objects

scenes



actions

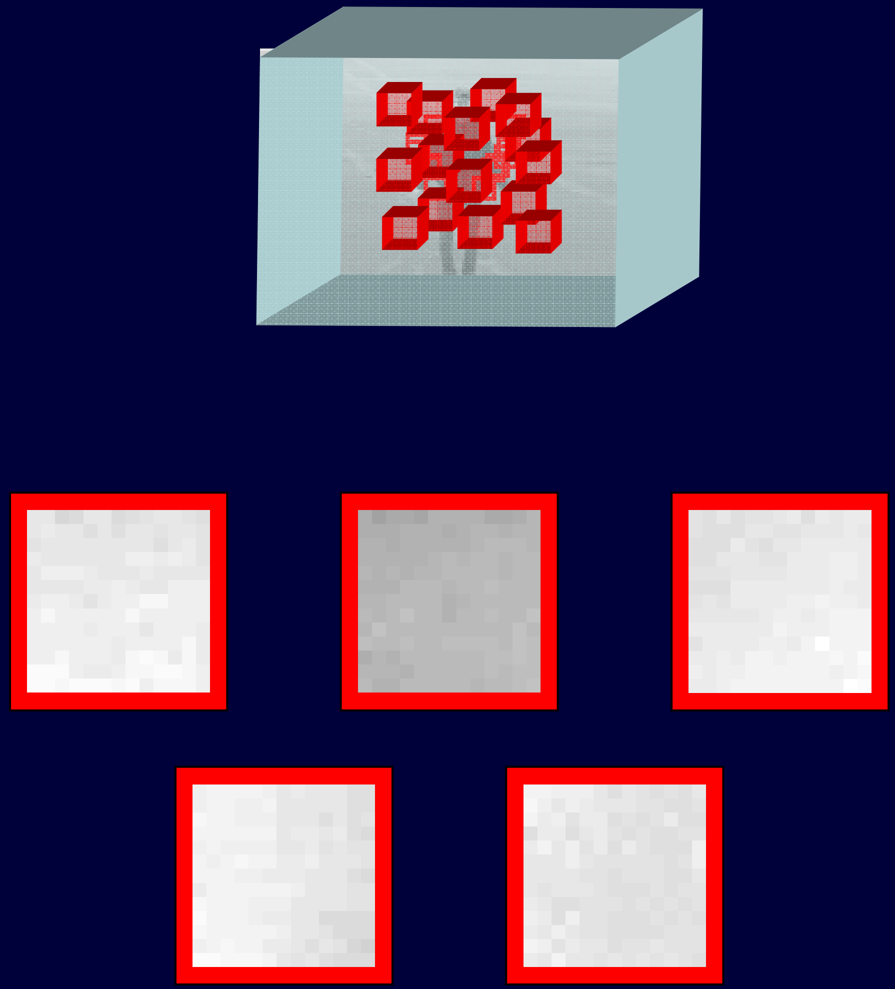
events

Human Action Classification

Challenges:

- **Camera Motion**
- **Complex Background**
- **Viewpoint Change**

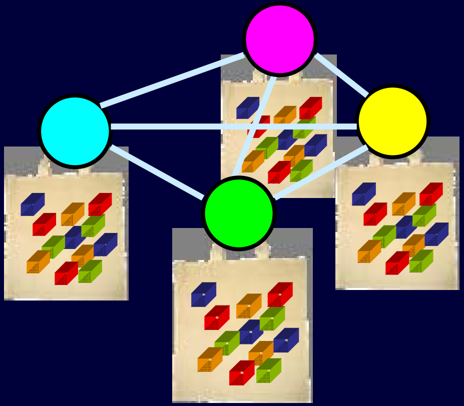
Spatial-Temporal Interest Points



[Dollar et al ’05]

Unsupervised learning of human action categories using spatial-temporal words.

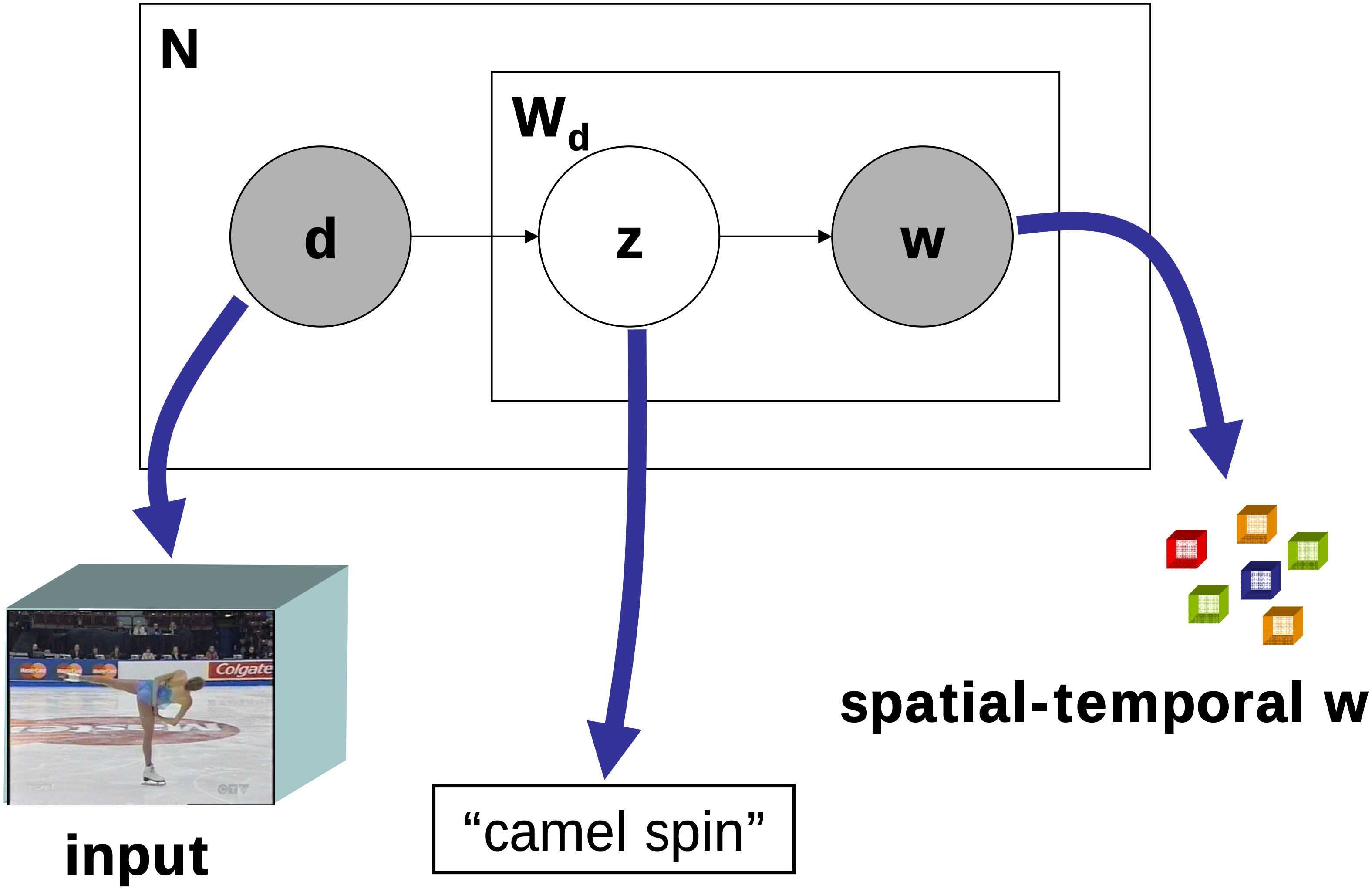
by J.C. Niebles, H. Wang, and L. Fei-Fei, BMVC 2006



A hierarchical model of shape and appearance for human action classification.

by J.C. Niebles, and L. Fei-Fei, CVPR 2007

Unsupervised learning using pLSA



N

W_d

d

z

w

spatial-temporal word

input

“camel spin”

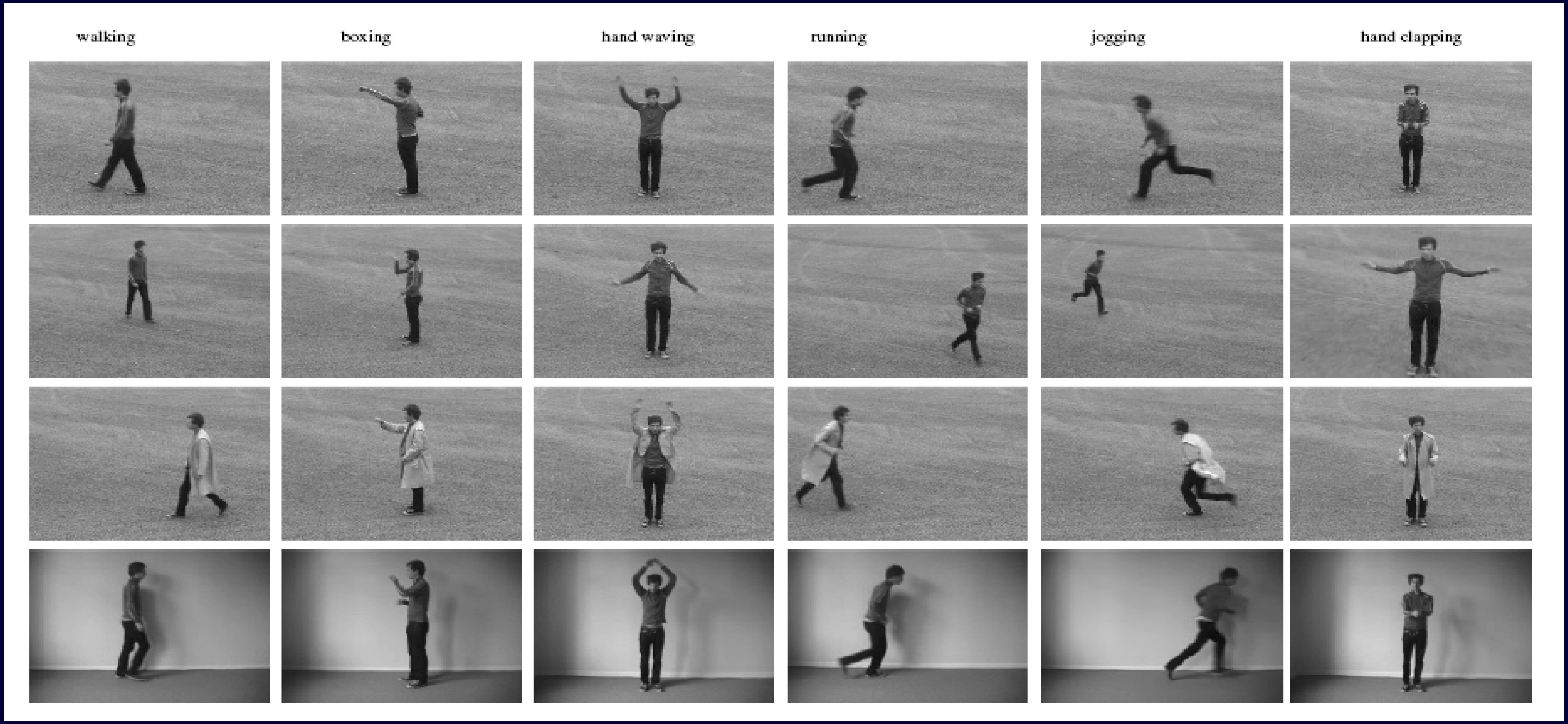
action category

Niebles, Wang & Fei-Fei, *BMVC* 2006

Experiment I:

KTH dataset

[Schuldt et al., 2004]:

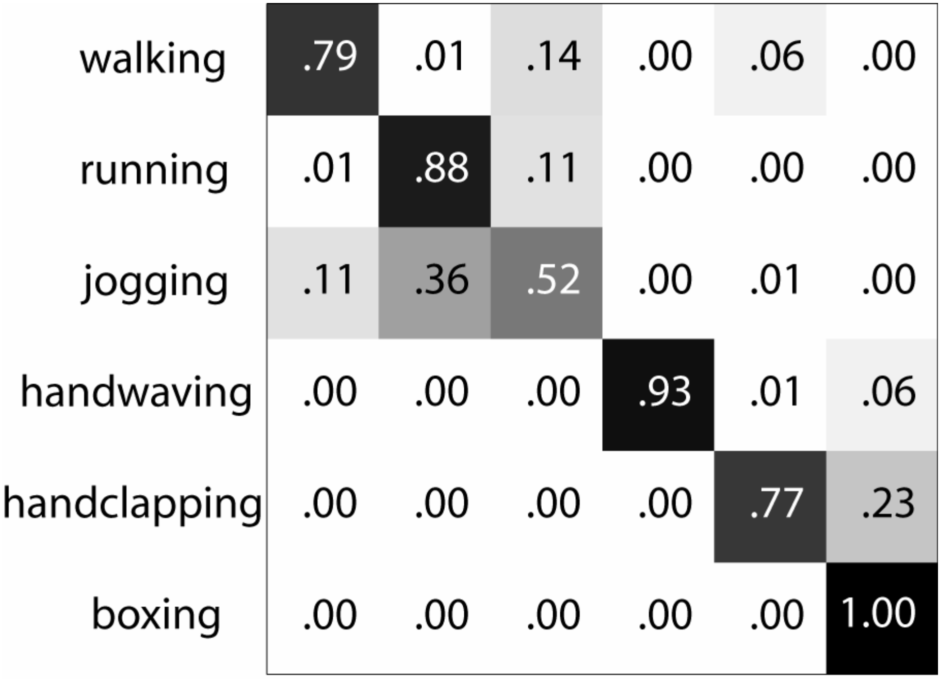


25 persons, indoors and outdoors, 4 long sequences per person

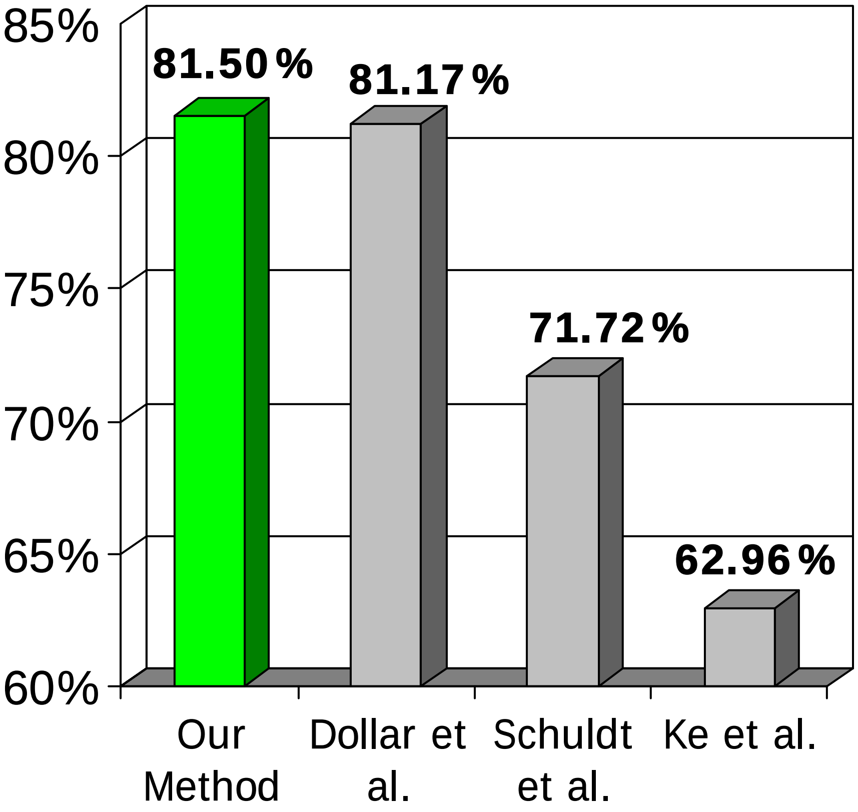
Experiment I: Performance

- Leave-one person out cross validation
- Average performance: 81.50%

- Unsupervised training
- Handle multiple motions

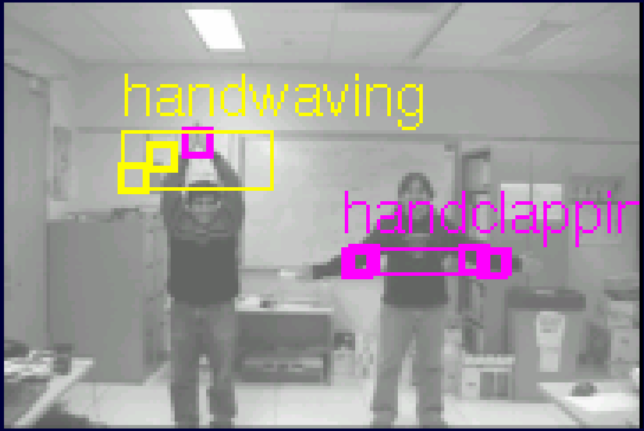


walking	.79	.01	.14	.00	.06	.00
running	.01	.88	.11	.00	.00	.00
jogging	.11	.36	.52	.00	.01	.00
handwaving	.00	.00	.00	.93	.01	.06
handclapping	.00	.00	.00	.00	.77	.23
boxing	.00	.00	.00	.00	.00	1.00



Niebles, Wang & Fei-Fei, *BMVC* 2006

Experiment I: Multiple motions



🟪 handclapping

🟨 handwaving

Trained with the KTH data

Tested with our own data

Niebles, Wang & Fei-Fei, *BMVC* 2006

Experiment I: A longer sequence

🟥 walking

🟩 running

Trained with the KTH data

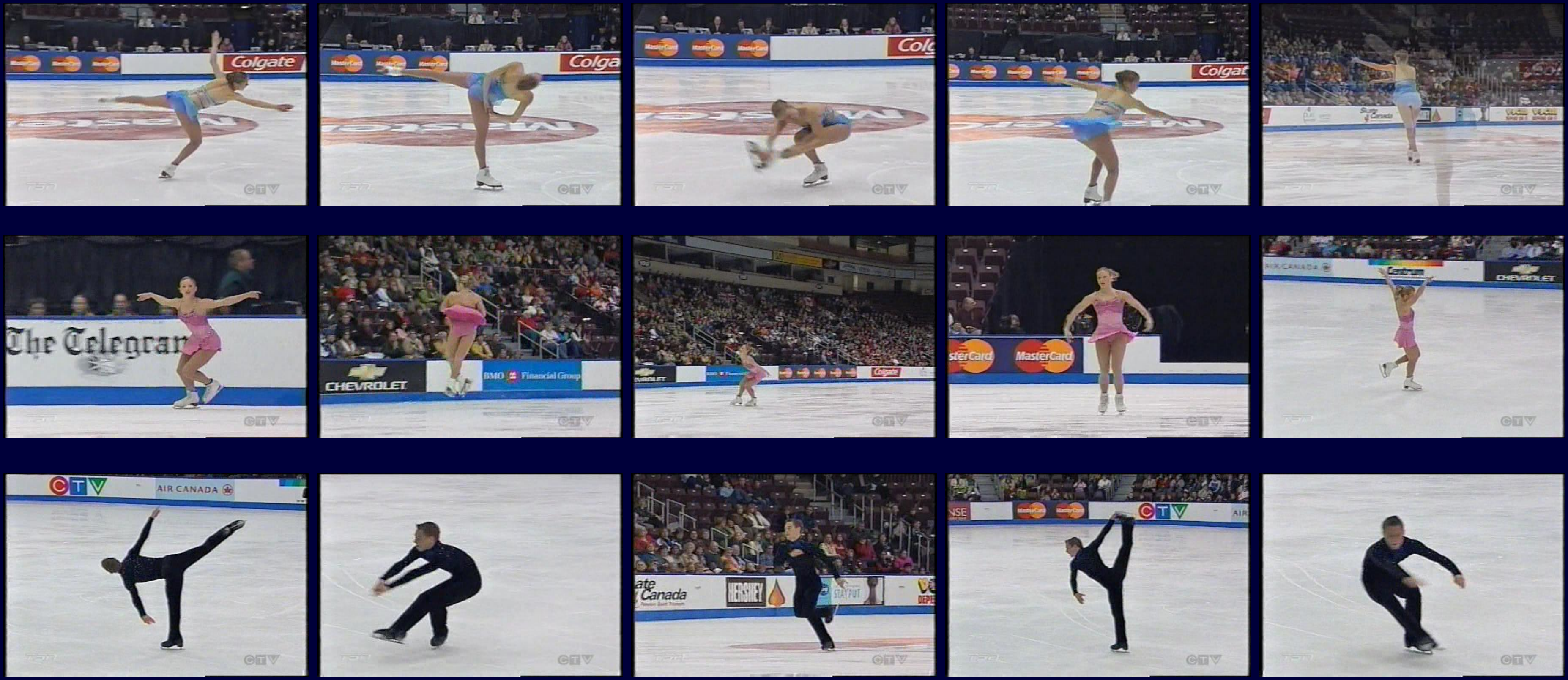
Tested with our own data

Niebles, Wang & Fei-Fei, *BMVC* 2006

Experiment II:

Figure Skating data set:

[Y.Wang, G.Mori et al, CVPR 2006]



7 persons, 3 action classes: camel spin, stand spin, sit spin

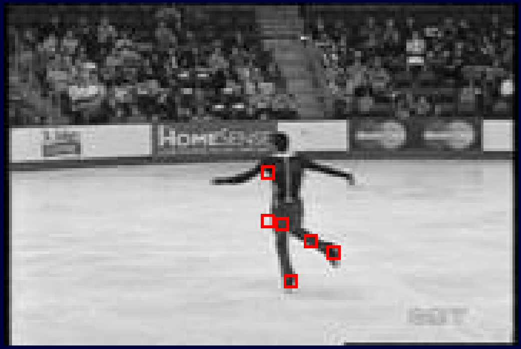
Niebles, Wang & Fei-Fei, *BMVC* 2006

Experiment II: Examples

Figure skating actions

Camel spin

Sit spin



Stand spin

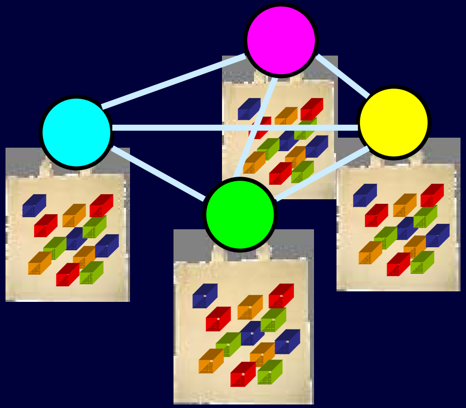
Niebles, Wang & Fei-Fei, *BMVC* 2006

Experiment II: Long Sequences

Niebles, Wang & Fei-Fei, *BMVC* 2006

Unsupervised learning of human action categories using spatial-temporal words.

by J.C. Niebles, H. Wang, and L. Fei-Fei, BMVC 2006

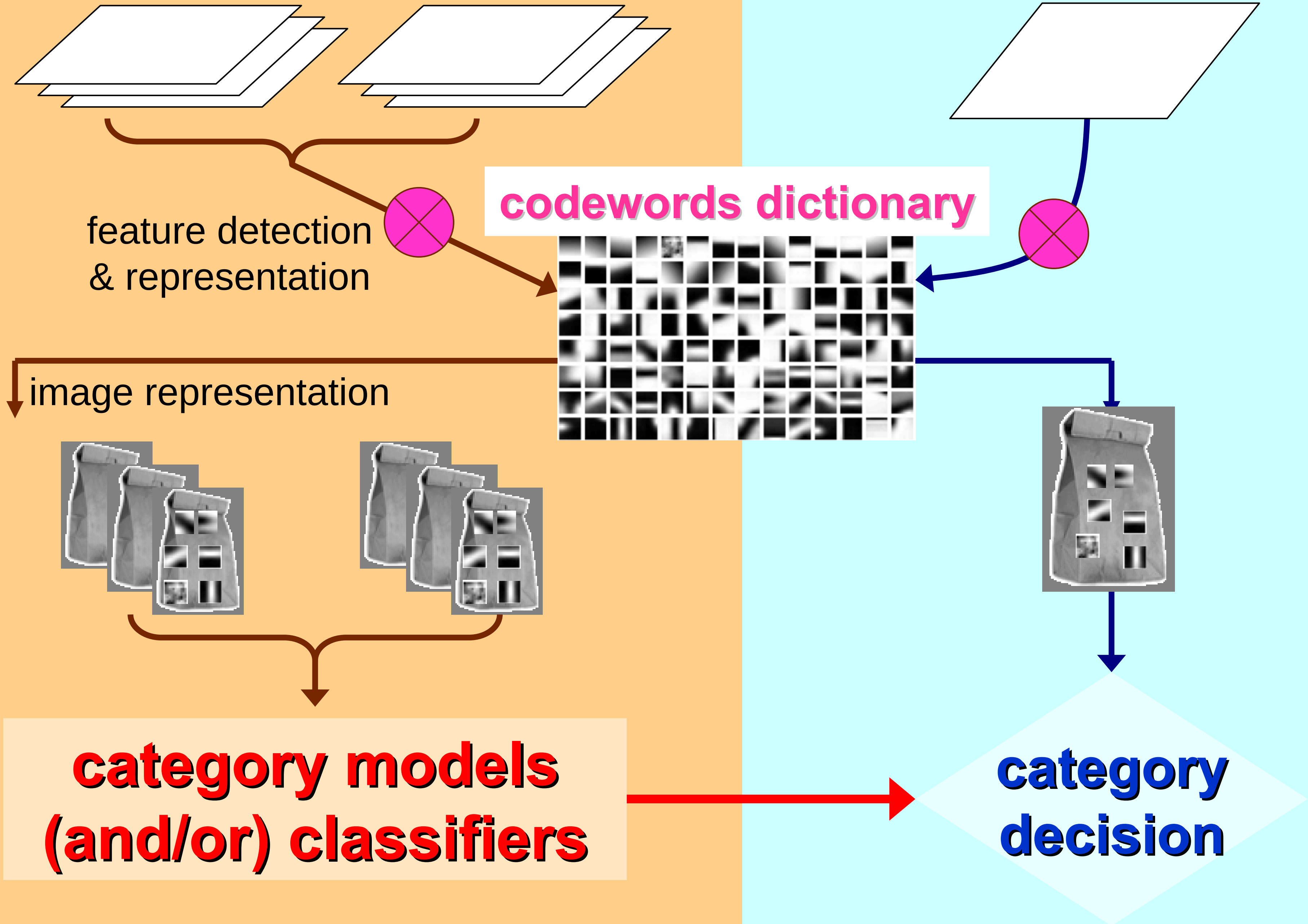


A hierarchical model of shape and appearance for human action classification.

by J.C. Niebles, and L. Fei-Fei, CVPR 2007

learning

recognition



feature detection
& representation

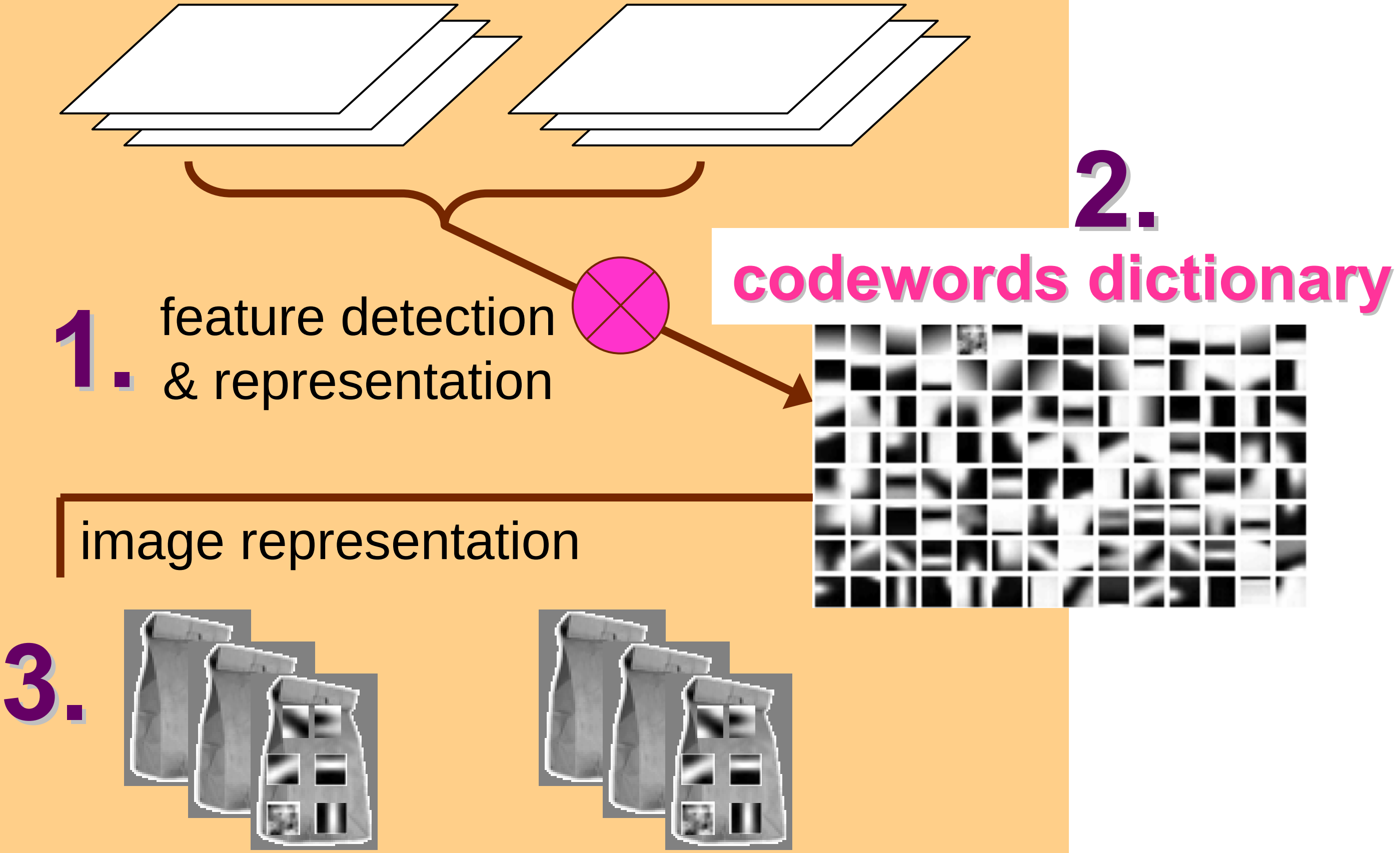
codewords dictionary

image representation

category models
(and/or) classifiers

category
decision

Representation



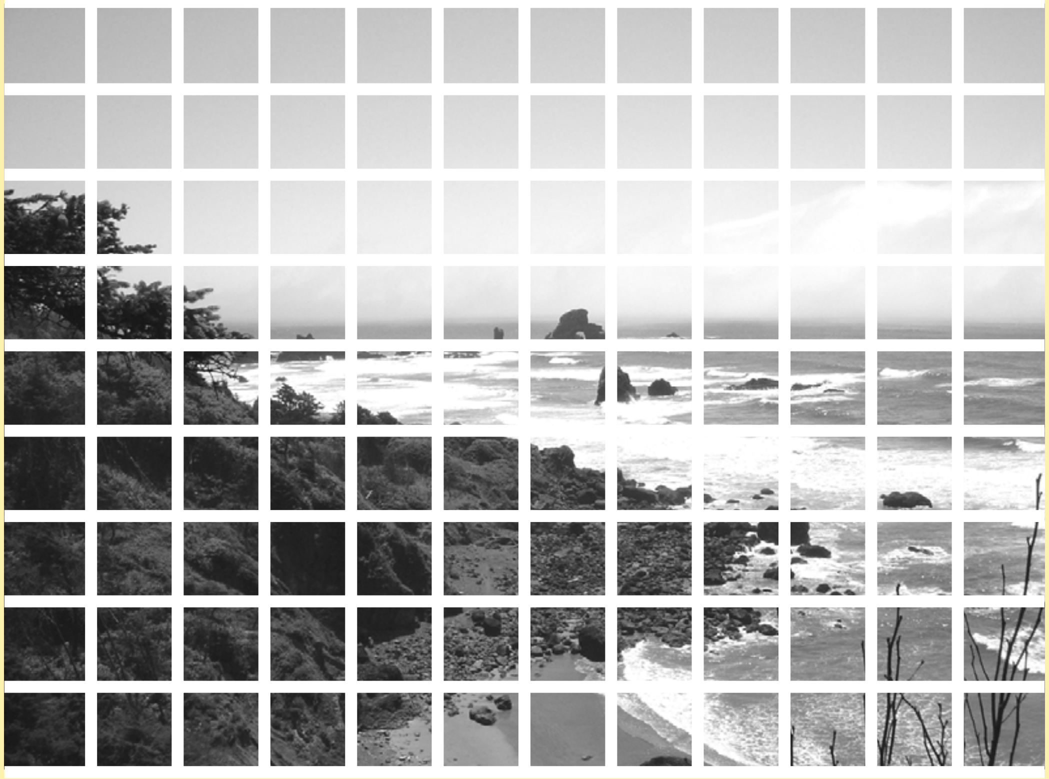
1. feature detection & representation

2. codewords dictionary

image representation

3.

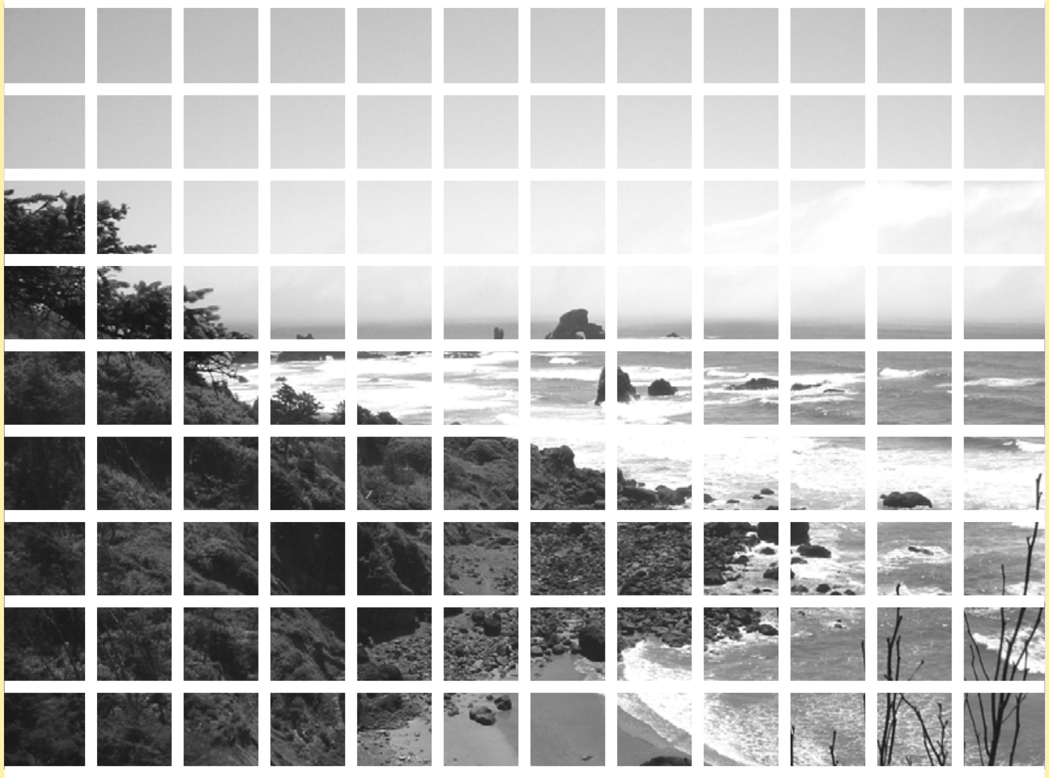
1.Feature detection and representation



extract interest points

- DoG
- Saliency detector (Kadir and Brady)
- grid

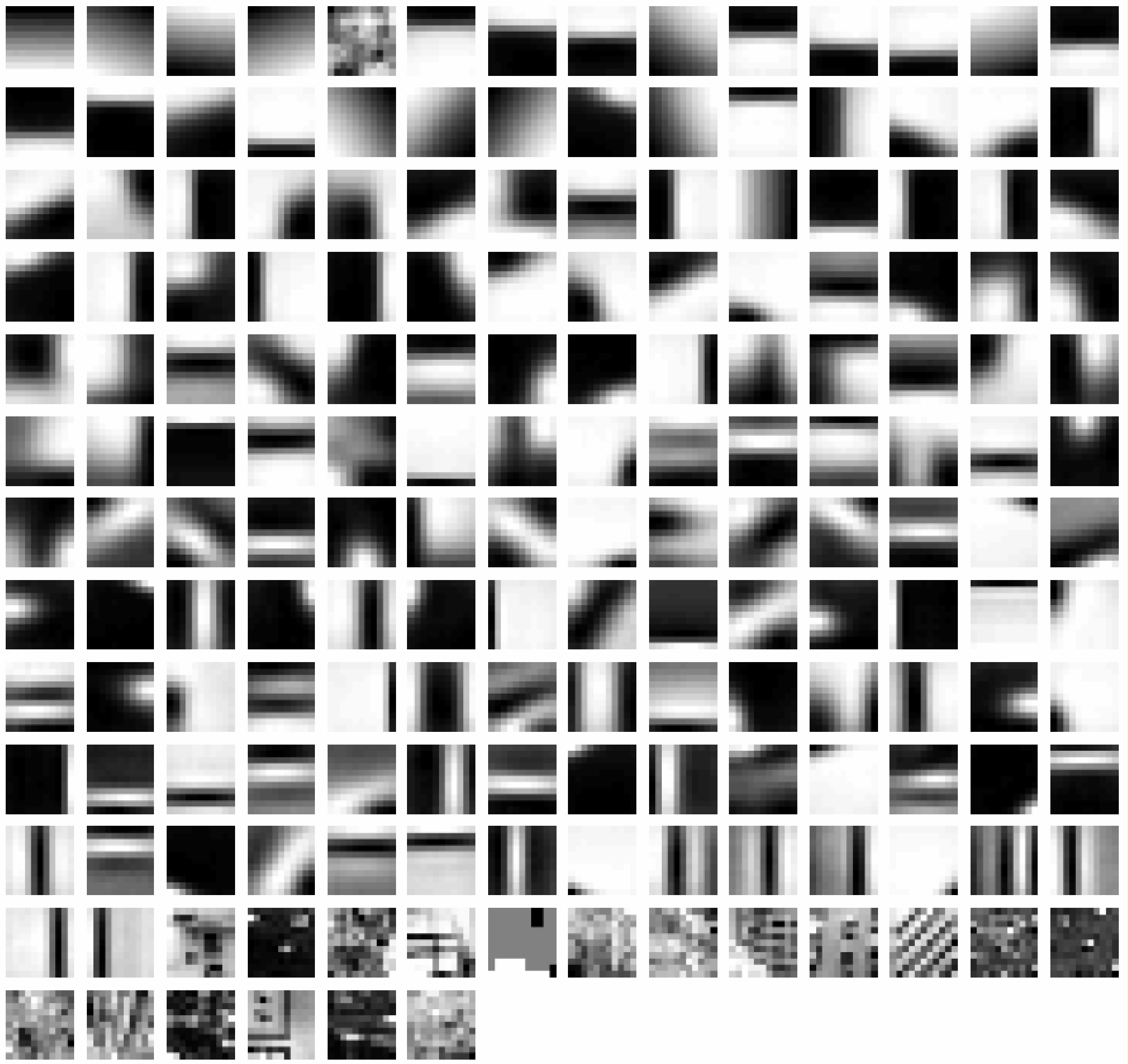
1.Feature detection and representation



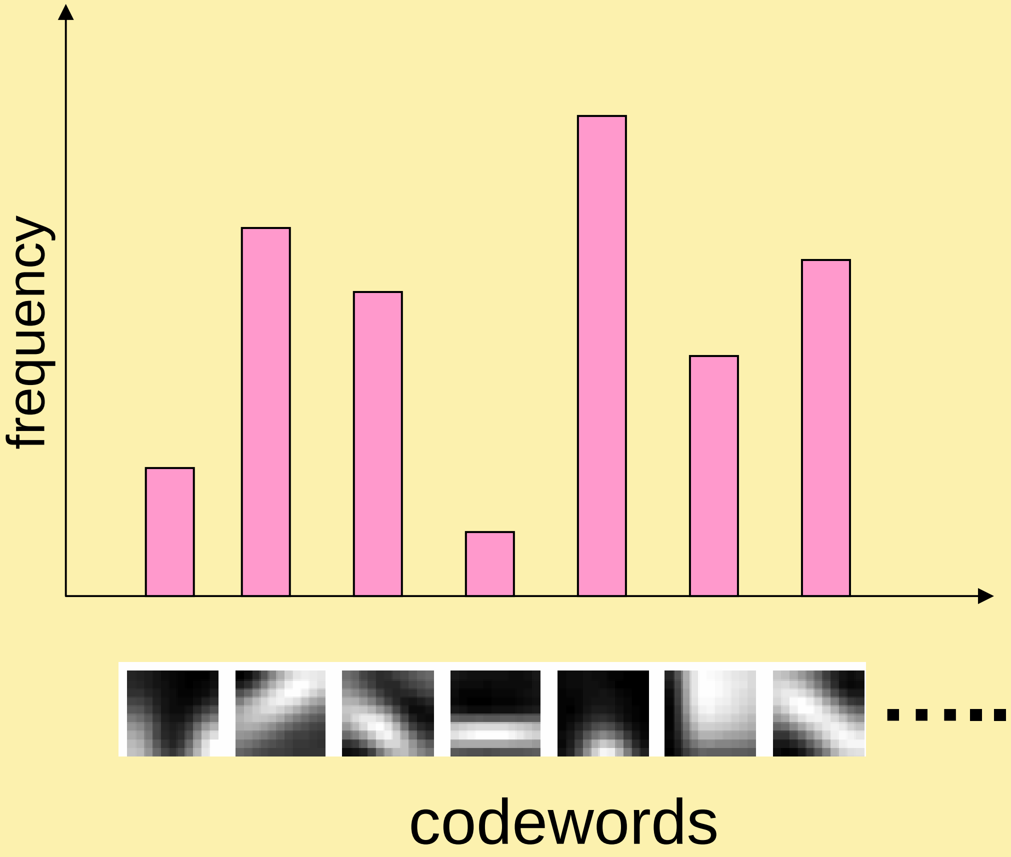
represent interest points

- SIFT (Lowe ’99)
- gray scale values

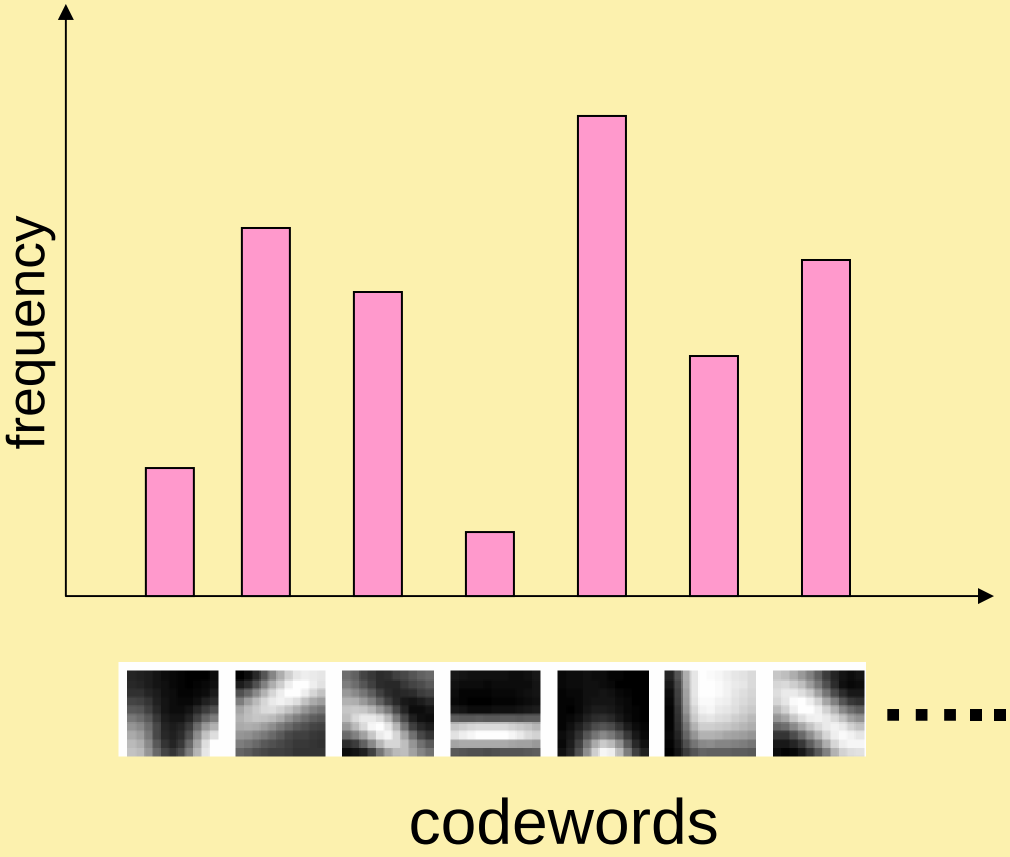
2. Codewords dictionary formation



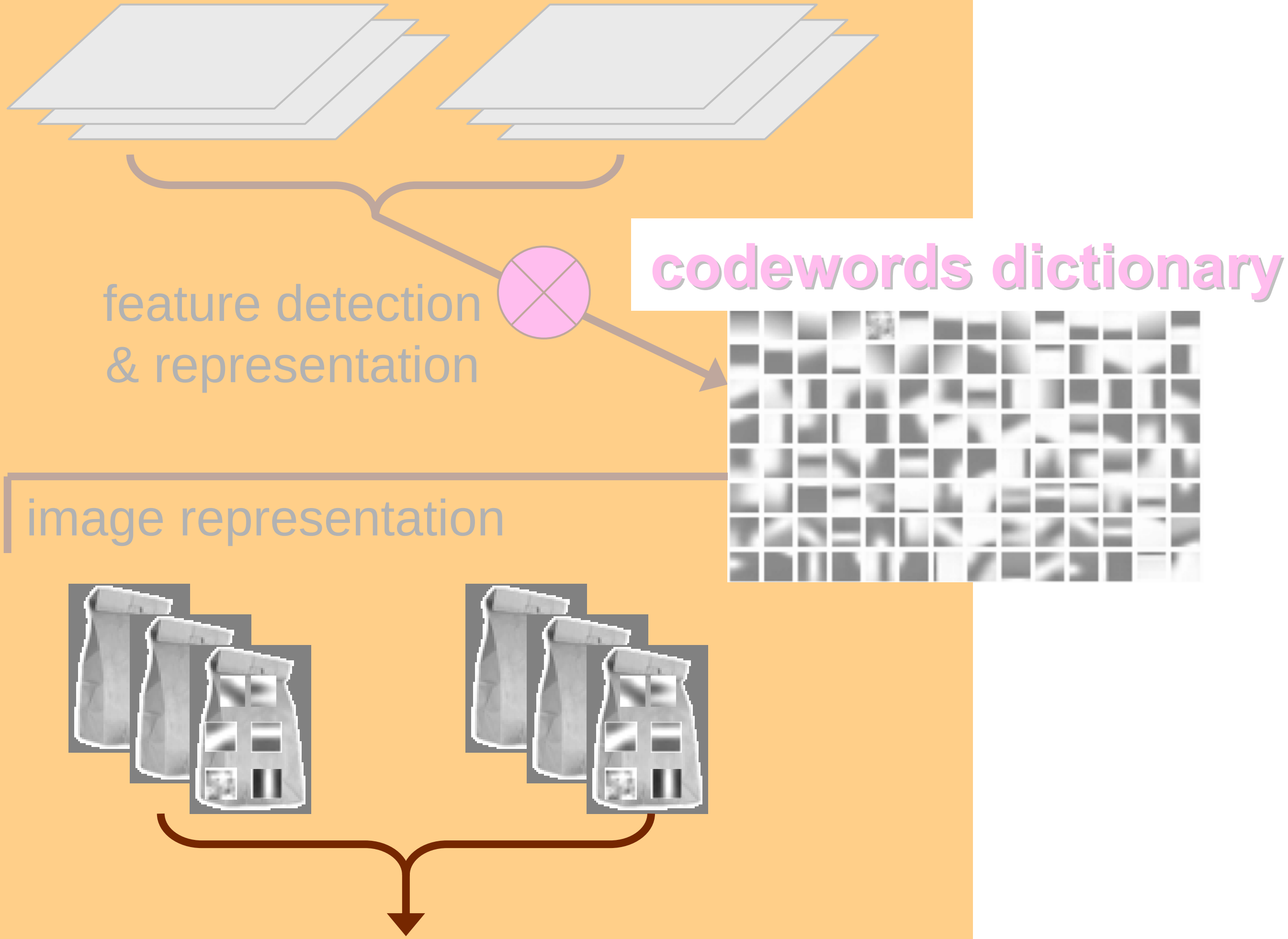
3. Image representation



3. Image representation



learning



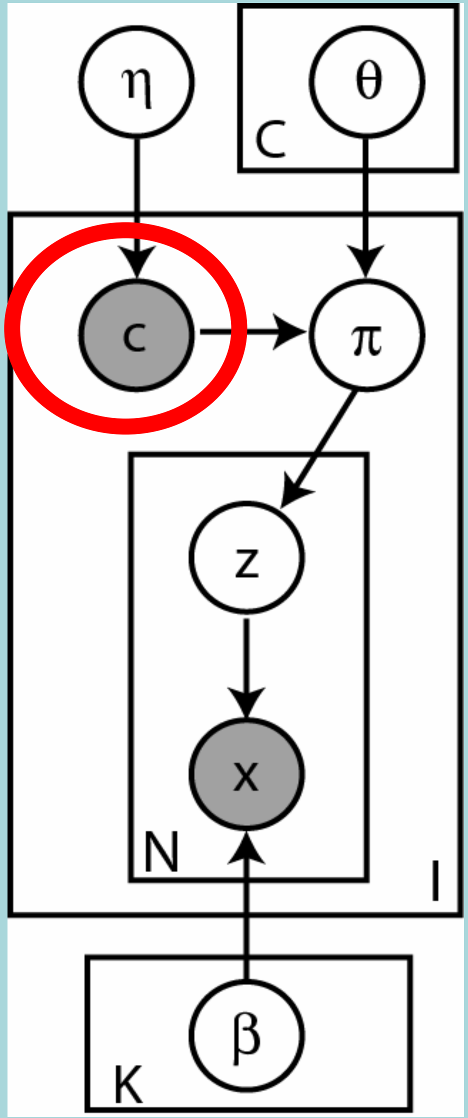
feature detection & representation

codewords dictionary

image representation

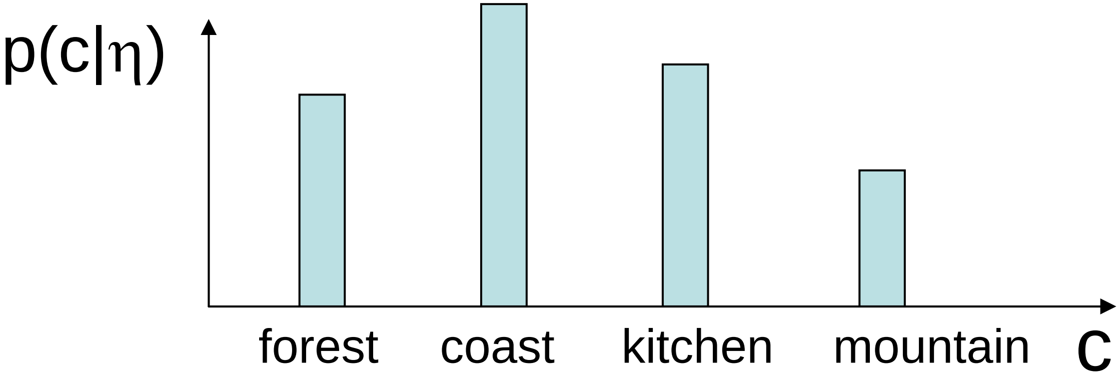
category models (and/or) classifiers

A Generative Model

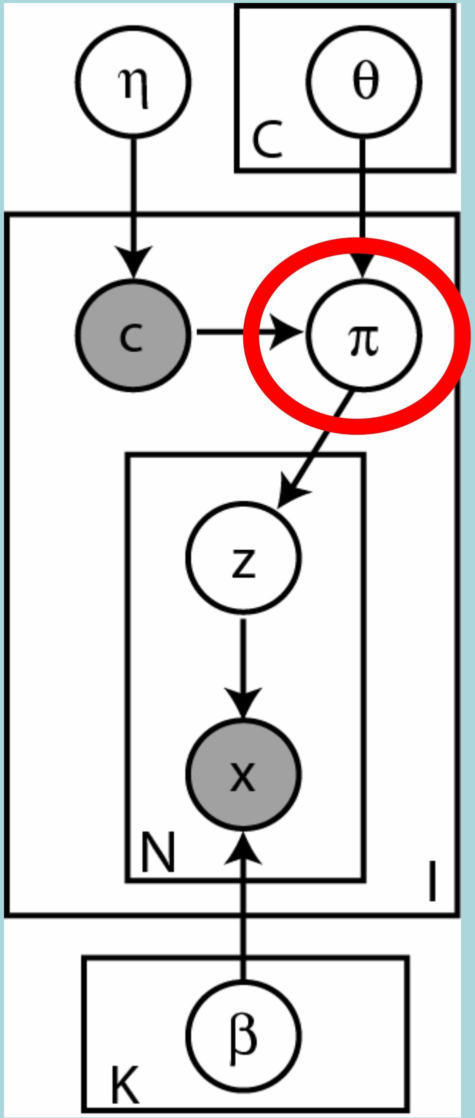


scene category

discrete variable: $c \sim p(c|\eta)$



A Generative Model



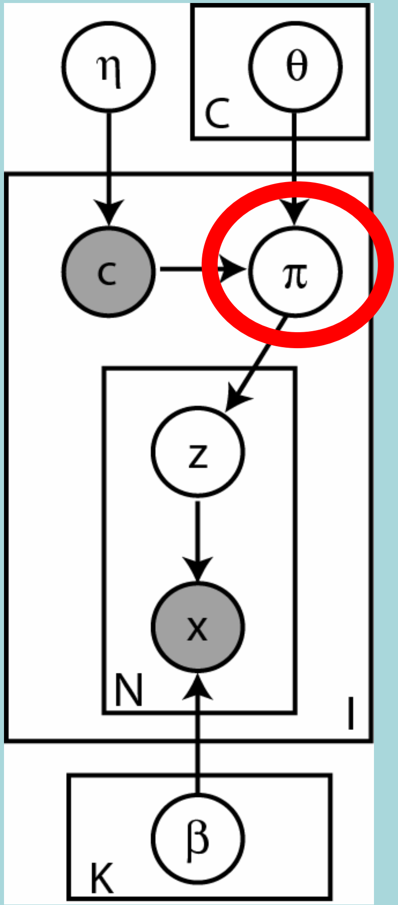
mixing parameter for the latent topics

$$\pi \sim p(\pi | c, \theta)$$

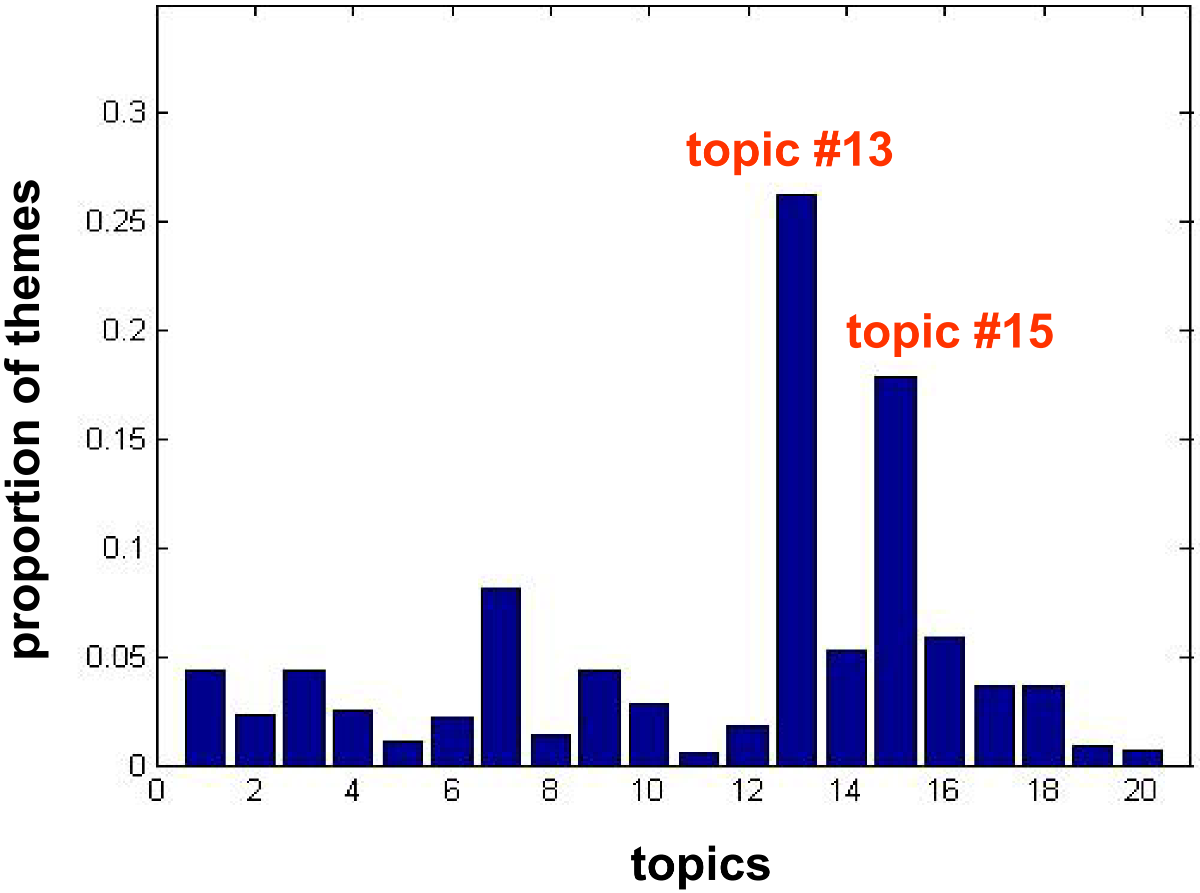
$$\sim Dir(\pi | c, \theta)$$

$$\sum_{k=1}^{K} \pi_k = 1$$ K~ total number of topics

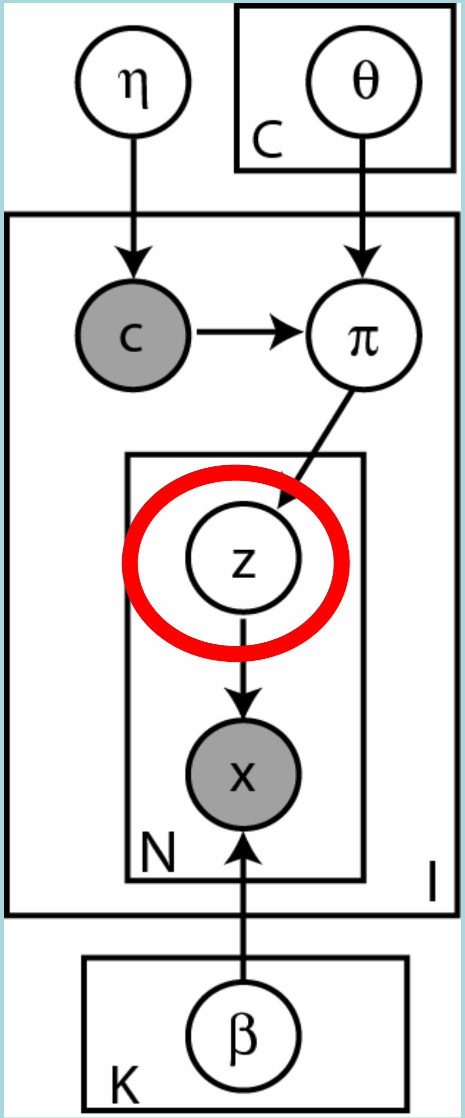
details of a learnt model - coast



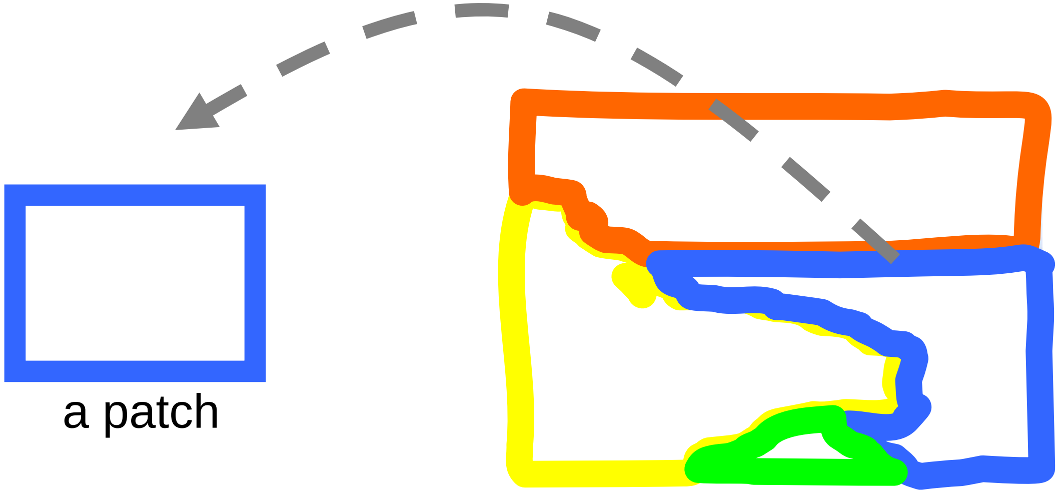
expected value of π given ‘coast’



A Generative Model



topic label



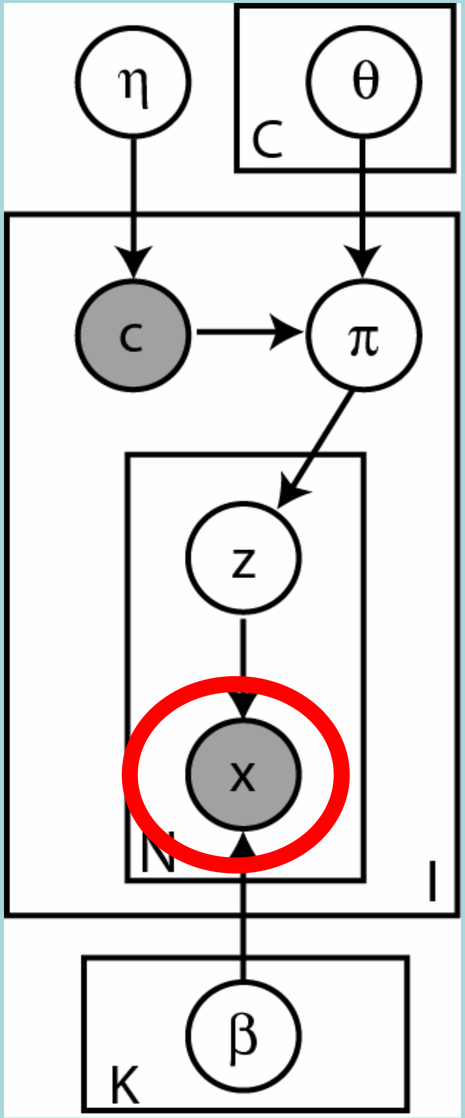
discrete variable:

$$z \sim p\left(z \middle| \pi\right)$$

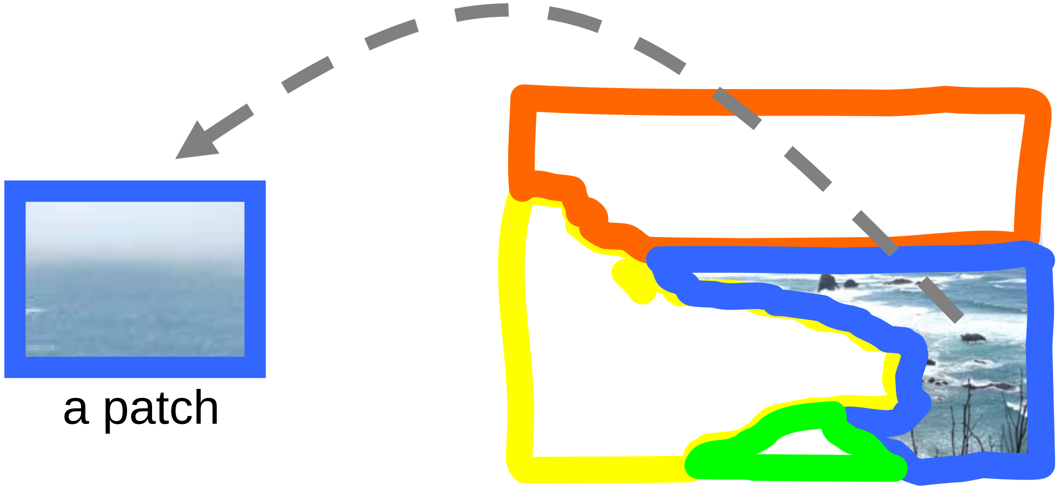
$$\sim Mult\left(z \middle| \pi\right)$$

$z = \{1, \ldots, K\}$ K~ total number of topic

A Generative Model



patch label



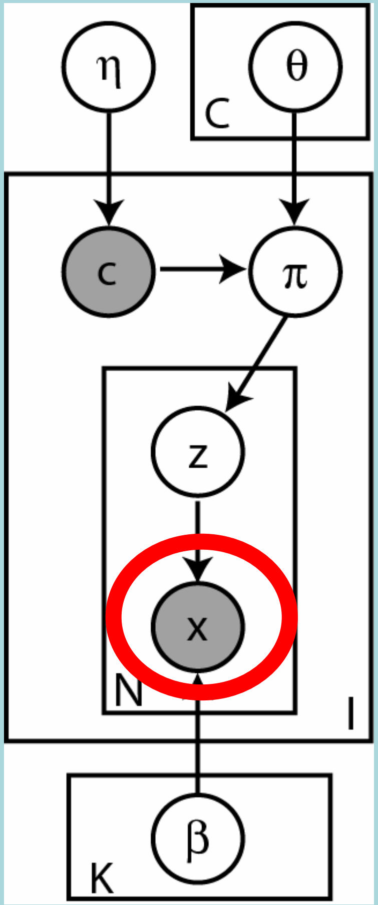
discrete variable:

$$x \sim p(x \mid z, \beta)$$

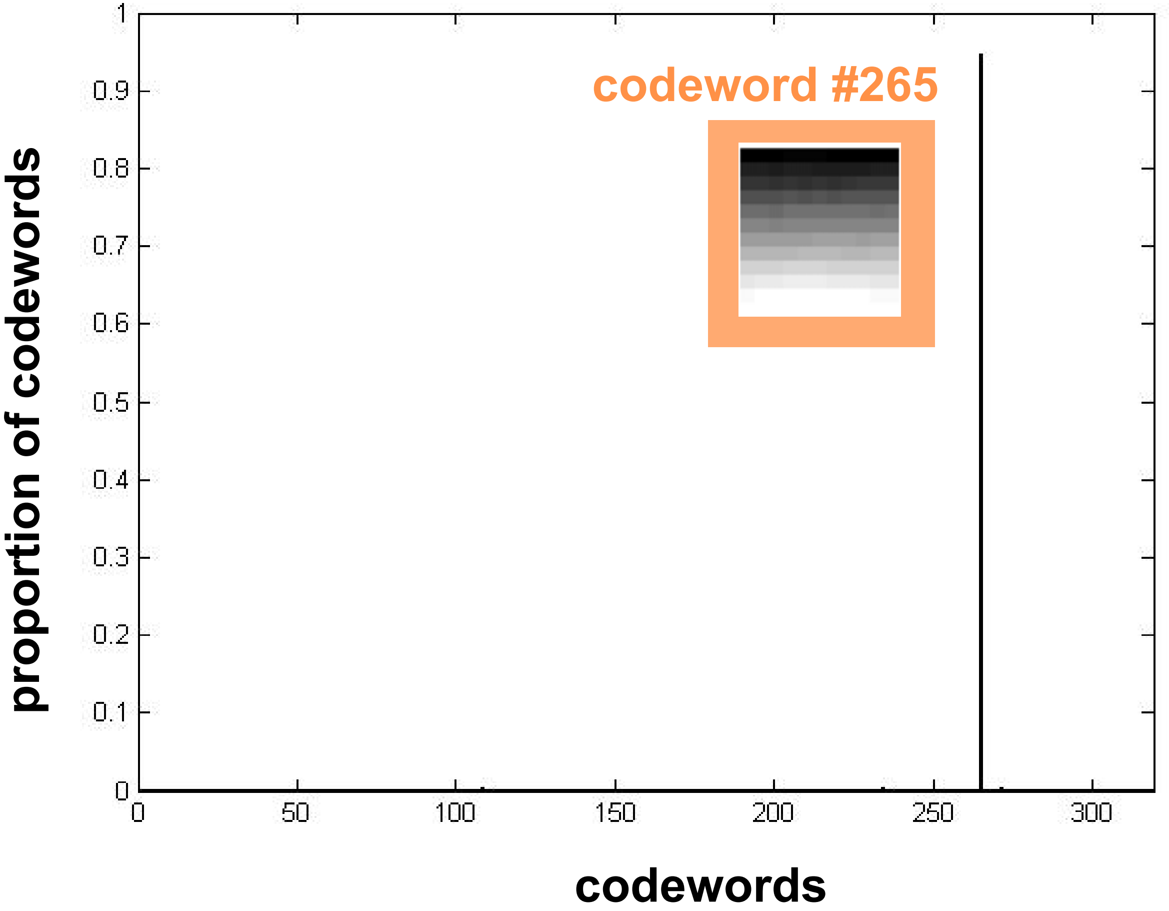
$$\sim \mathit{Mult}\ (x \mid z, \beta)$$

$x = \{1, \ldots, T\}$ T~ total number of codewords

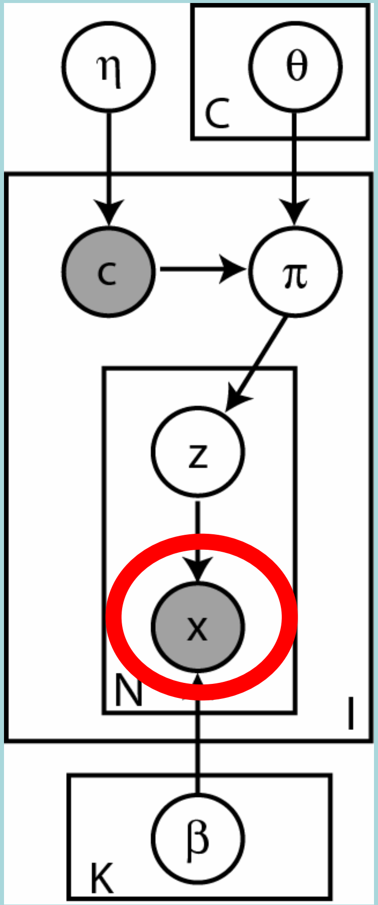
details of a learnt model - coast



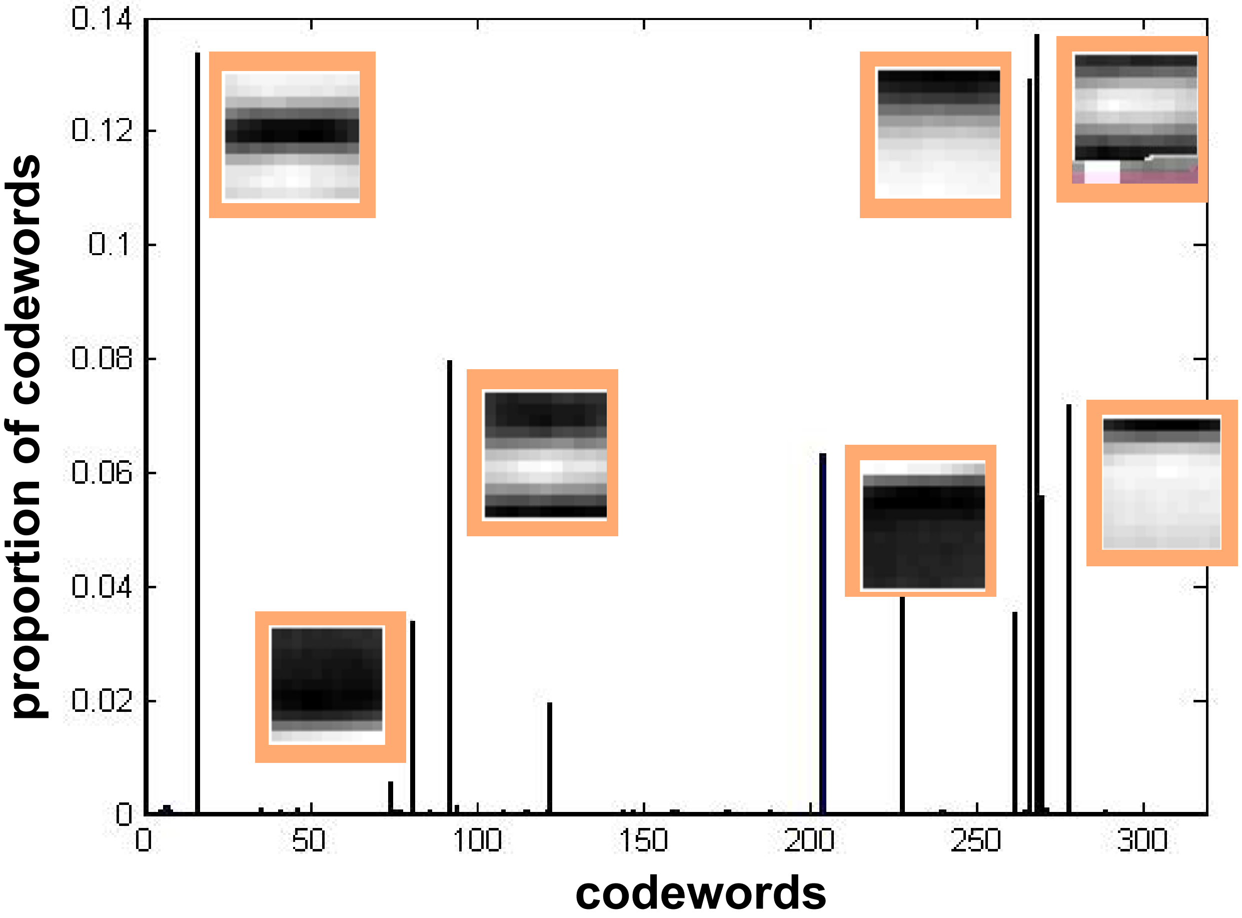
expected value of β given ‘z=13’



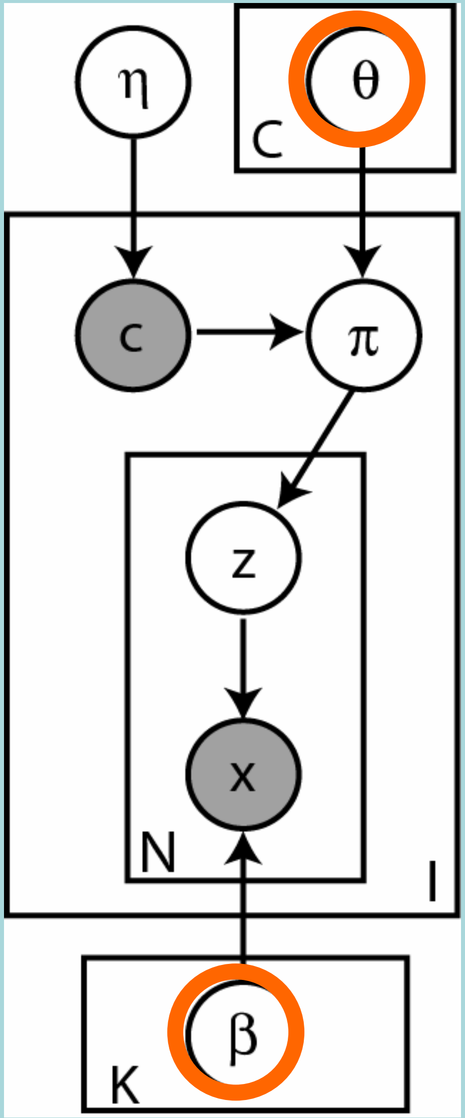
details of a learnt model - coast



expected value of β given 'z=15'



A Generative Model



learning

Find the ‘best’ θ and β

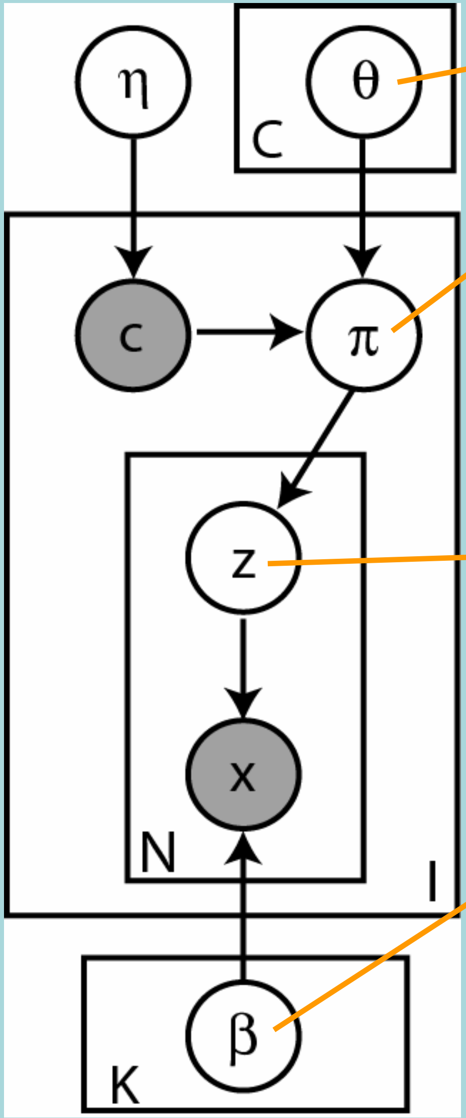
joint probability

$$p(x,z,\pi|\theta,\beta,c) = p(\pi|c,\theta)\prod_{n}^{N} p(z_n|\pi)p(x_n|z_n,\beta)$$

$$p(x|\theta,\beta,c) = \int p(\pi|c,\theta)\left(\prod_{n}^{N}\sum_{z_n} p(z_n|\pi)p(x_n|z_n,\beta)\right)d\pi$$

- exact inference is intractable
- use Variational Inference

A Generative Model



Variational Inference

Maximum Likelihood estimation (Minka 2000)

$$\gamma_{ck} = \theta^0_{ck} + \sum_n^N \left\langle \delta\left(z_n^k = 1\right)\right\rangle$$

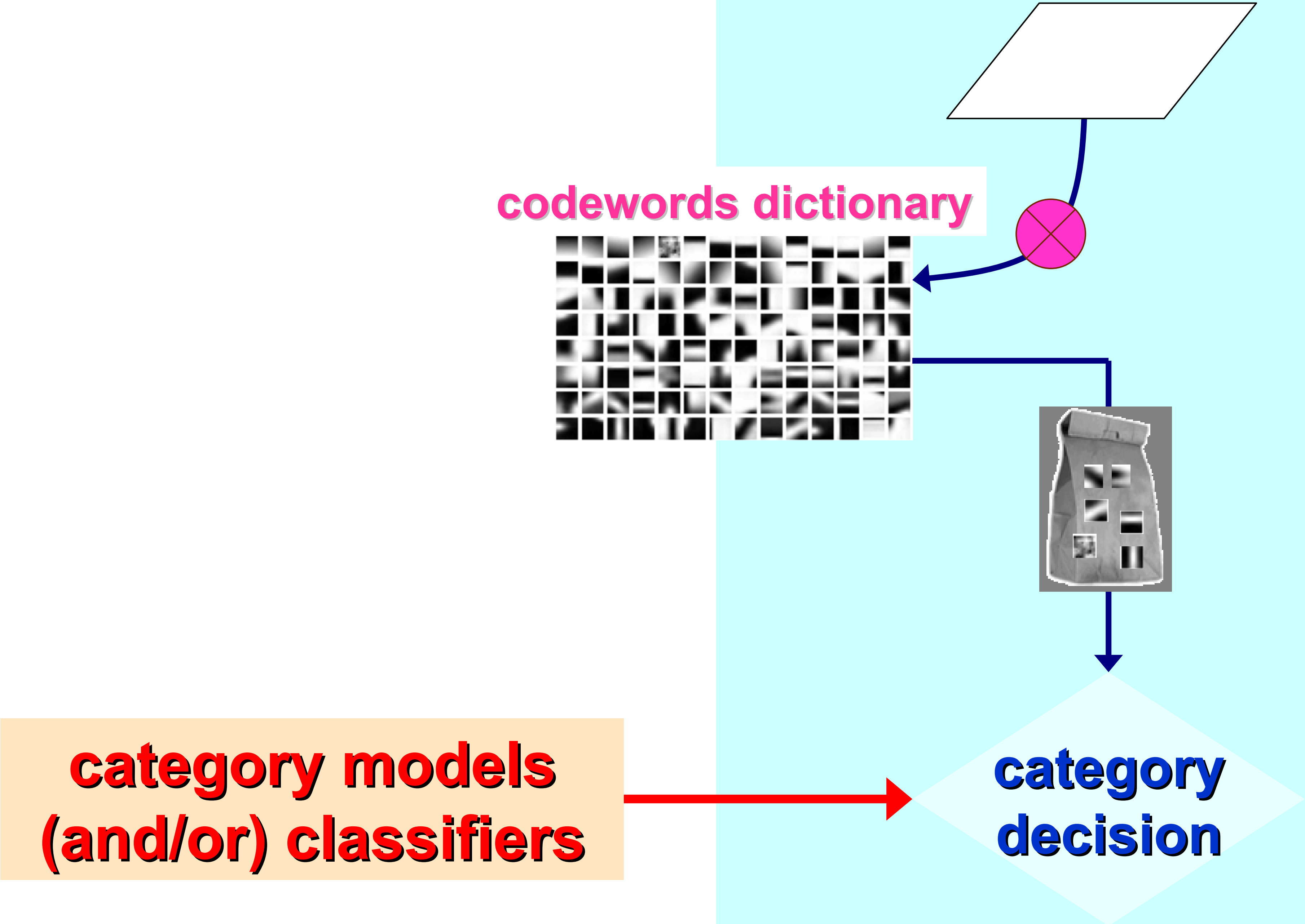
$$\left\langle \log \pi_{ck}\right\rangle = \Psi\left(\gamma_{ck}\right) - \Psi\left(\sum_k \gamma_{ck}\right)$$

$$\left\langle \delta\left(z_n^k = 1\right)\right\rangle = \exp\left\{\left\langle \log \pi_{ck}\right\rangle + \sum_t^T \left\langle \log \beta_{kt}\right\rangle \delta\left(x_n^t = 1\right)\right\}$$

$$\xi_{kt} = \zeta^0 + \sum_i^I \sum_n^N \left\langle \delta\left(z_{i,n}^k = 1\right)\right\rangle \delta\left(x_{i,n}^t = 1\right)$$

$$\left\langle \log \beta_{kt}\right\rangle = \Psi\left(\xi_{kt}\right) - \Psi\left(\sum_t \xi_{kt}\right)$$

Recognition

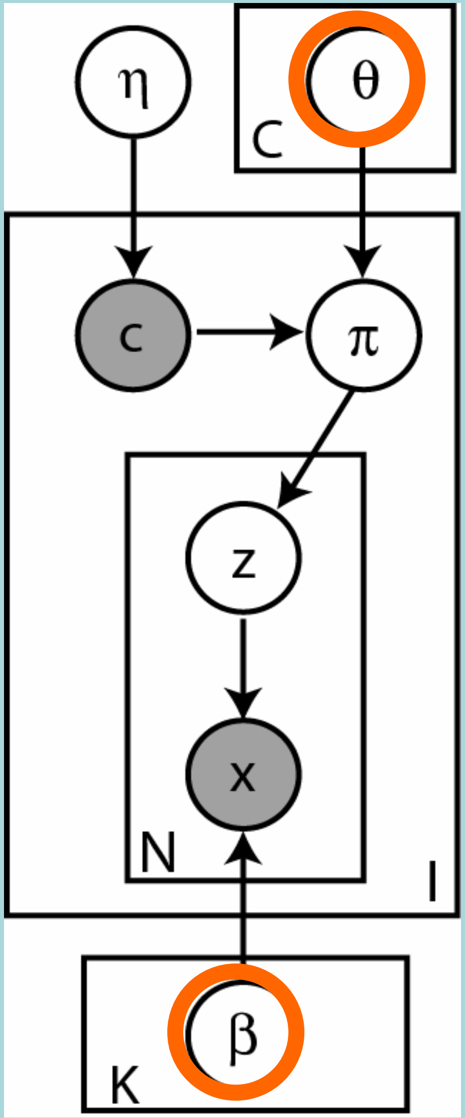


codewords dictionary

category models (and/or) classifiers

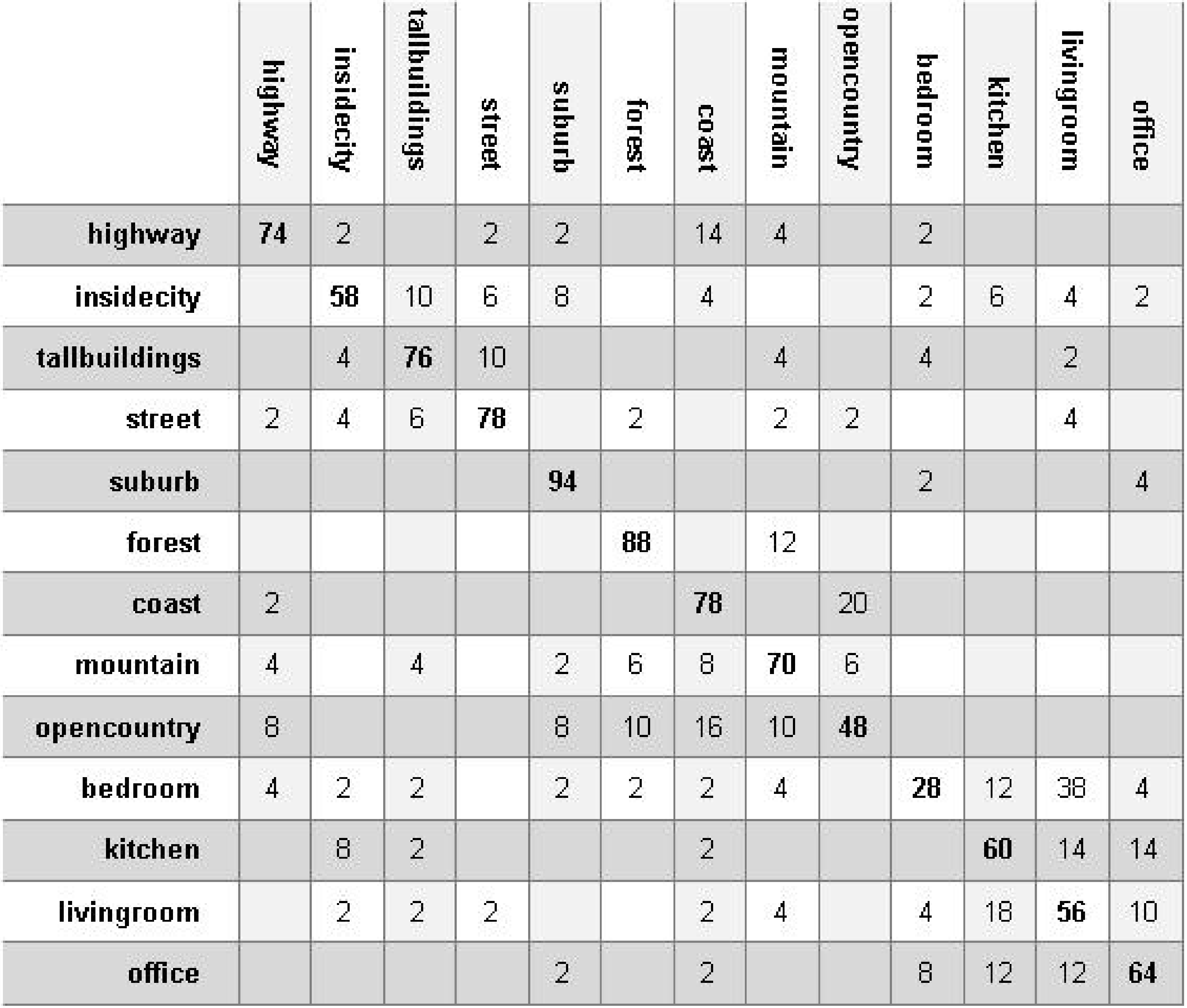
category decision

A Generative Model

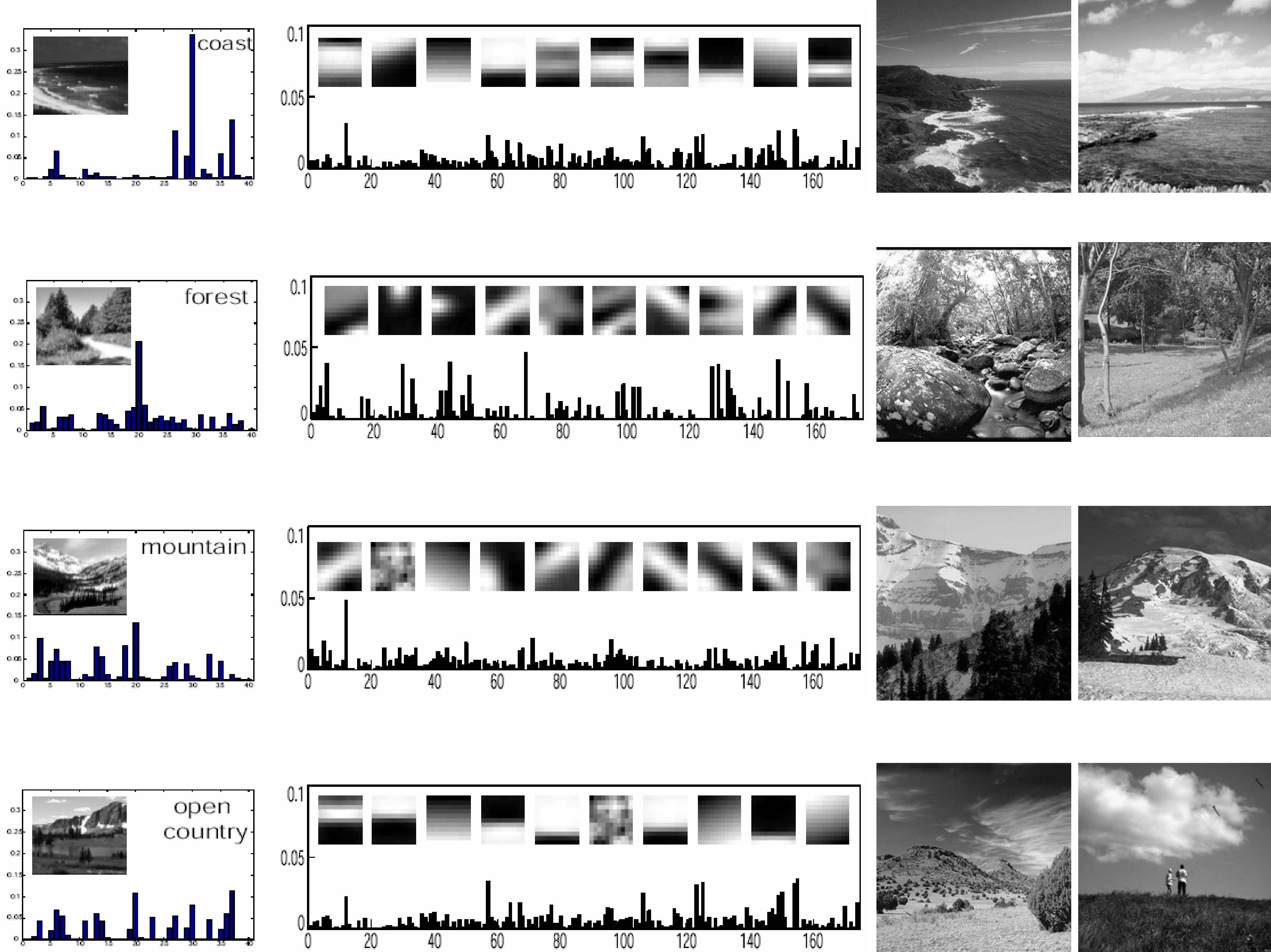


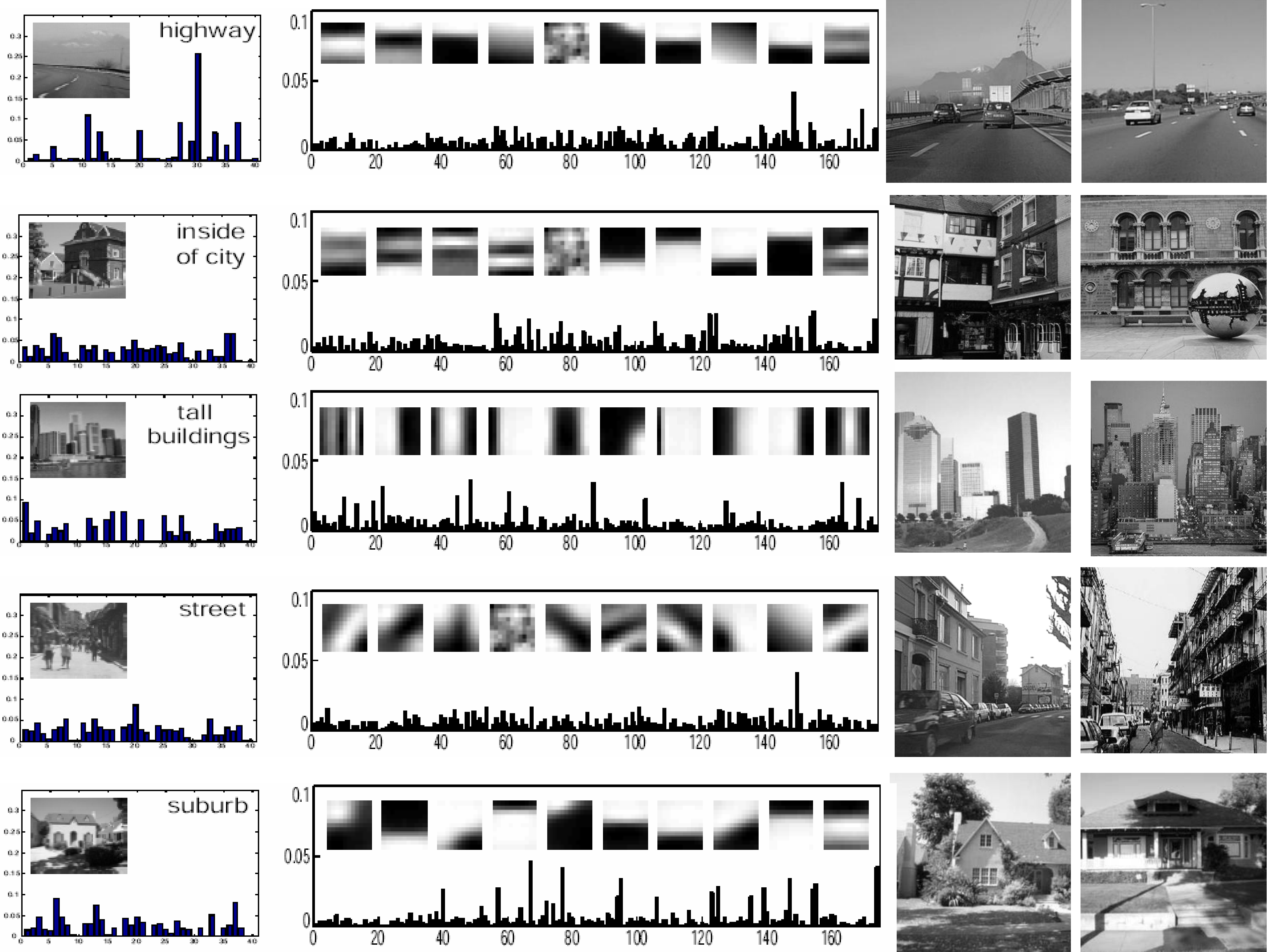
testing (inference)

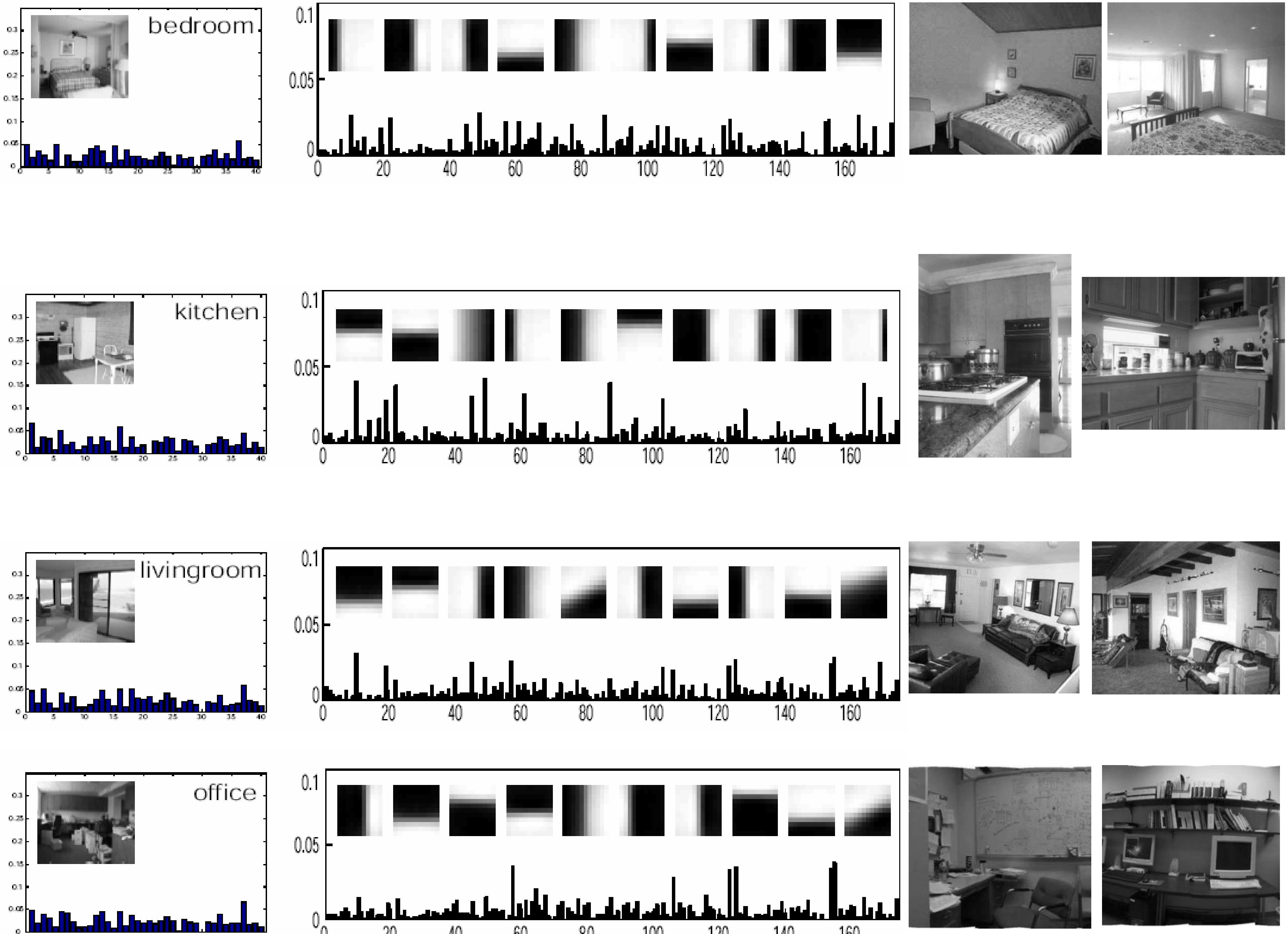
$$c = \arg\max_{c} \; p\left(x \mid c, \theta, \beta\right)$$



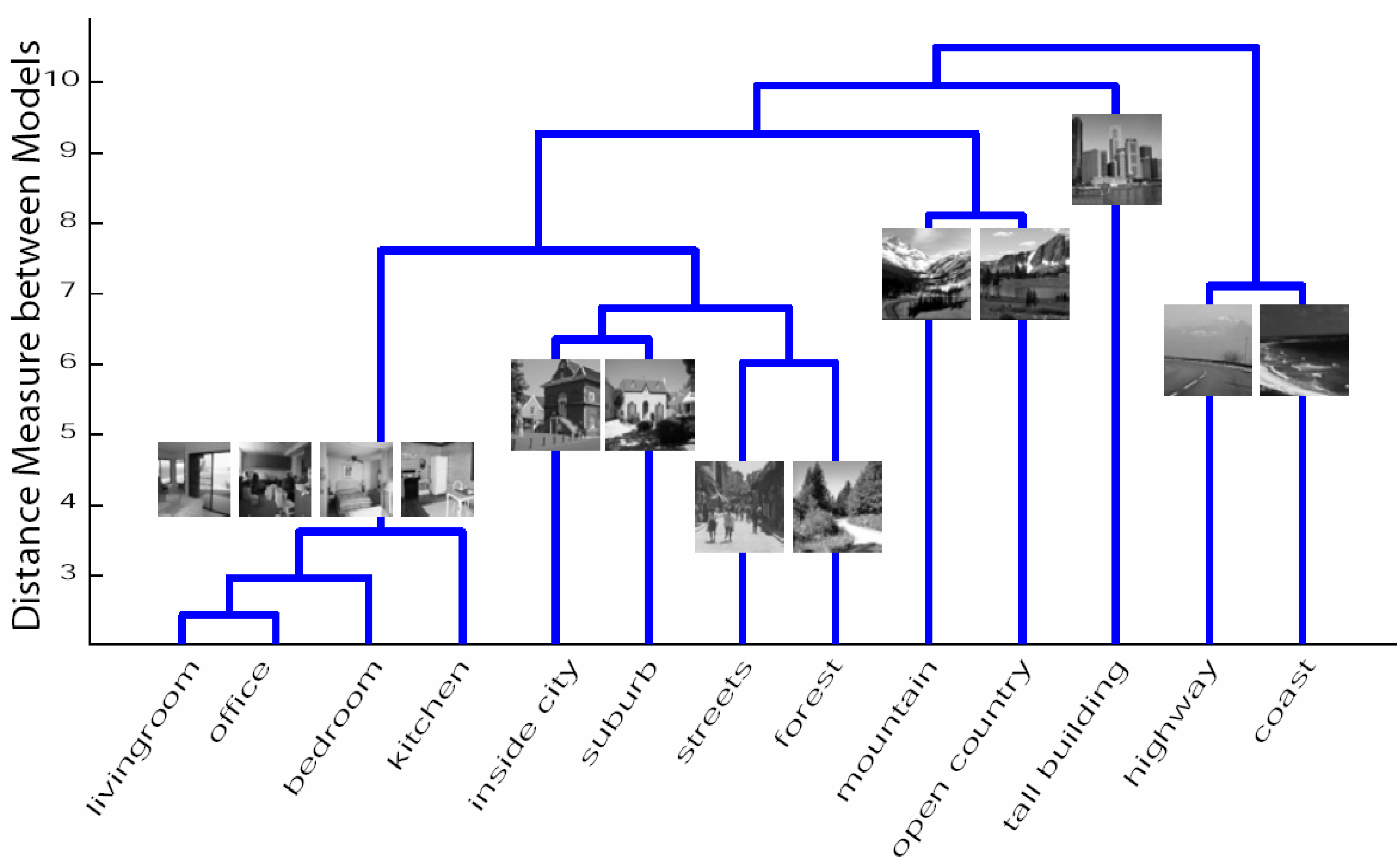
	highway	insidecity	tallbuildings	street	suburb	forest	coast	mountain	opencountry	bedroom	kitchen	livingroom	office
highway	**74**	2		2	2		14	4		2			
insidecity		**58**	10	6	8		4			2	6	4	2
tallbuildings		4	**76**	10				4		4		2	
street	2	4	6	**78**		2		2	2			4	
suburb					**94**					2			4
forest						**88**		12					
coast	2						**78**		20				
mountain	4		4		2	6	8	**70**	6				
opencountry	8				8	10	16	10	**48**				
bedroom	4	2	2		2	2	2	4		**28**	12	38	4
kitchen		8	2				2				**60**	14	14
livingroom		2	2	2			2	4		4	18	**56**	10
office					2		2			8	12	12	**64**


coast
forest
mountain
open
country
0
0.05
0.1
0.15
0.2
0.25
0.3
5
10
15
20
25
30
35
40
0.1
0.05
0
0
20
40
60
80
100
120
140
160


highway
inside of city
tall buildings
street
suburb
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
35
40
0.1
0.05
0
0
20
40
60
80
100
120
140
160


bedroom
kitchen
livingroom
office
0
0.05
0.1
0.15
0.2
0.25
0.3
5
10
15
20
25
30
35
40
0.05
0.1
20
40
60
80
100
120
140
160

model distance based on theme distribution



Thank you!

- Collaborators:
  - Pietro Perona, Silvio Savarese, Rob Fergus
- Students:
  - **Juan Carlos Niebles**

  - **Li-Jia Li**

http://vision.cs.princeton.edu