Machine Learning in Computer Vision

Fei-Fei Li

What is (computer) vision?

- When we “see” something, what does it involve?
- Take a picture with a camera, it is just a bunch of colored dots (pixels)
- Want to make computers understand images
- Looks easy, but not really…

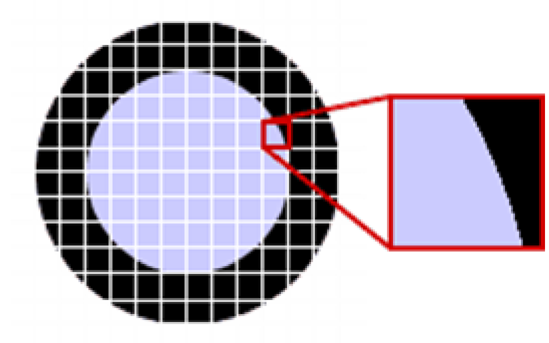
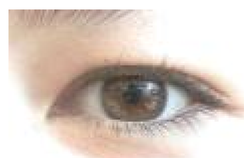


Image (or video)

Sensing device

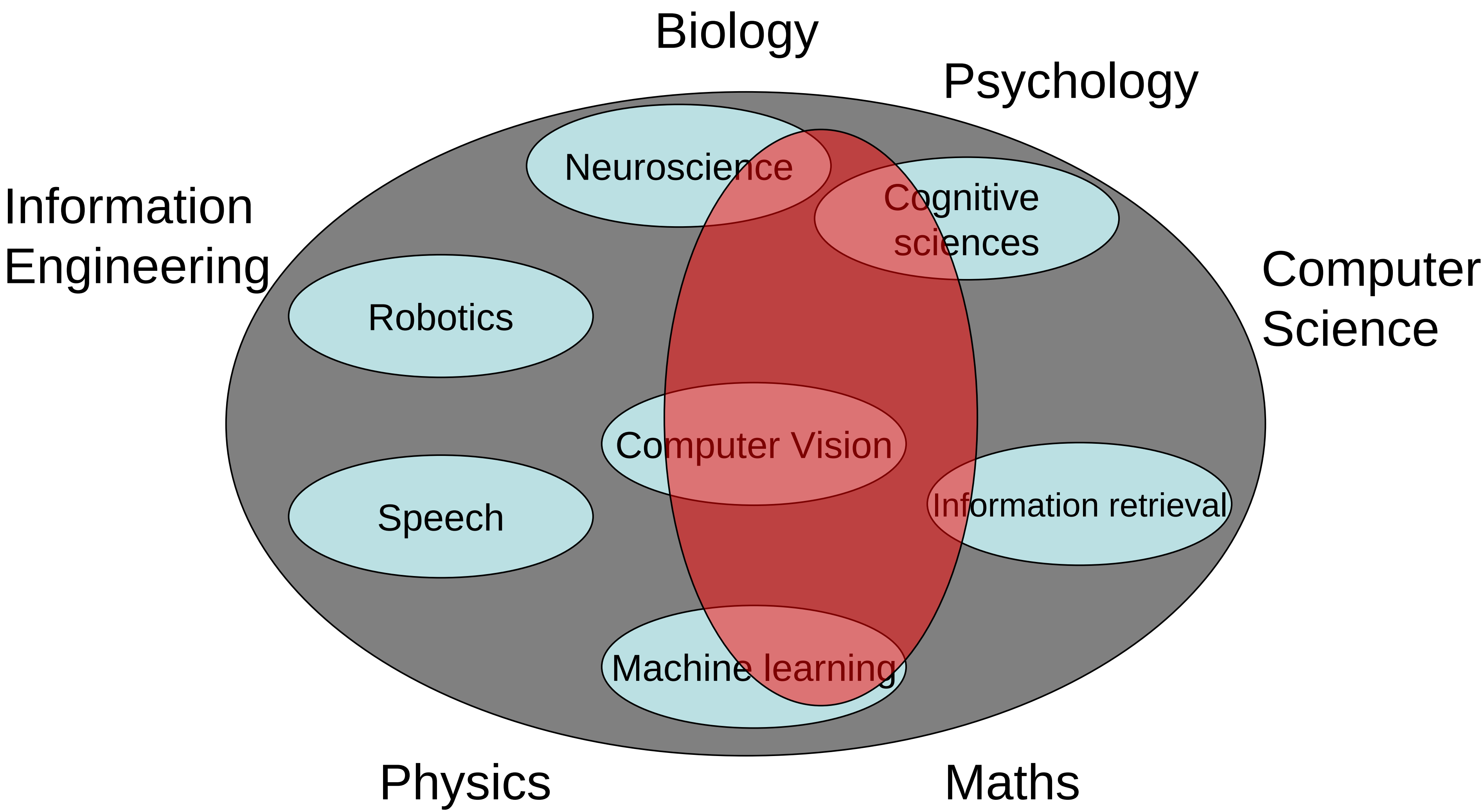


Interpreting device

Interpretations

Corn/mature corn
jn a cornfield/
plant/blue sky in
the background
Etc.

What is it related to?



Quiz?

What about this?

A picture is worth a thousand words.

--- Confucius

or *Printers' Ink* Ad (1921)

A picture is worth a thousand words.

--- Confucius

or *Printers' Ink* Ad (1921)

horizontal lines

textured

blue on the top

vertical

white

shadow to the left

porous

oblique

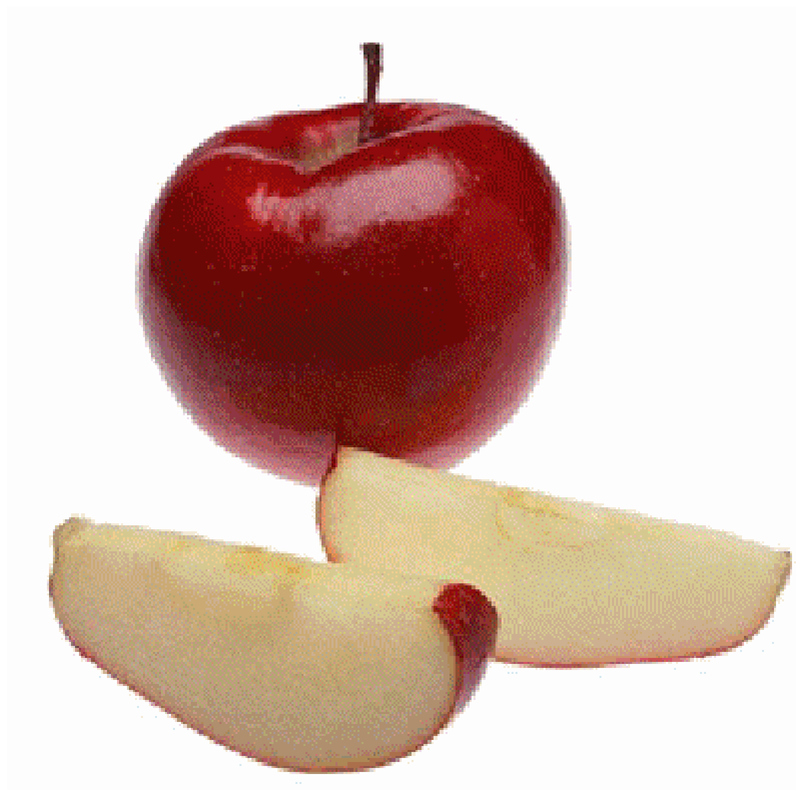
large green patches

A picture is worth a thousand words.

--- Confucius

or *Printers' Ink* Ad (1921)

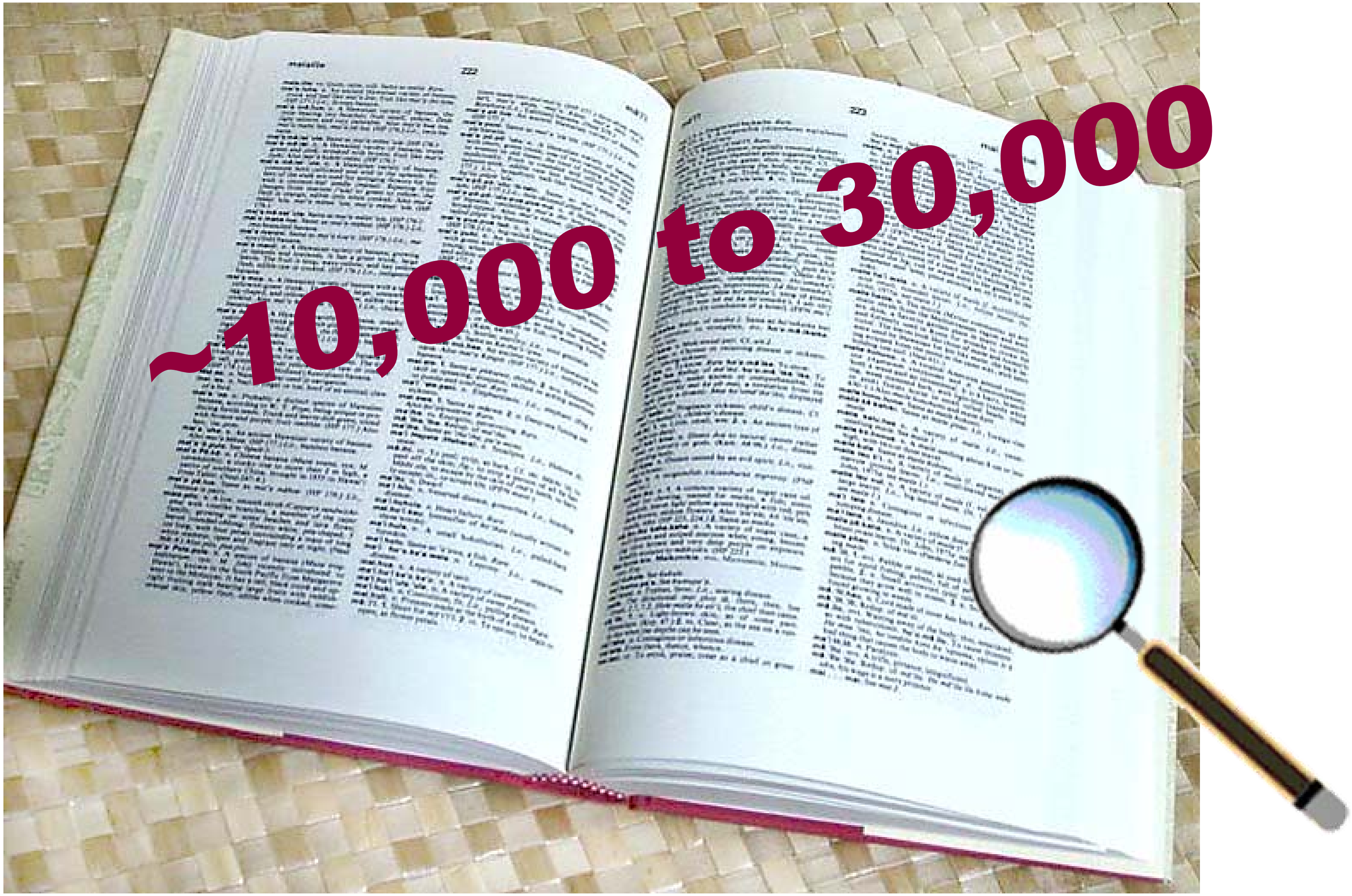
Today: machine learning methods for object recognition



outline

- Intro to object categorization
- Brief overview
  - Generative
  - Discriminative
- Generative models
- Discriminative models

How many object categories are there?



Biederman 1987

Challenges 1: view point variation

Michelangelo 1475-1564

Challenges 2: illumination

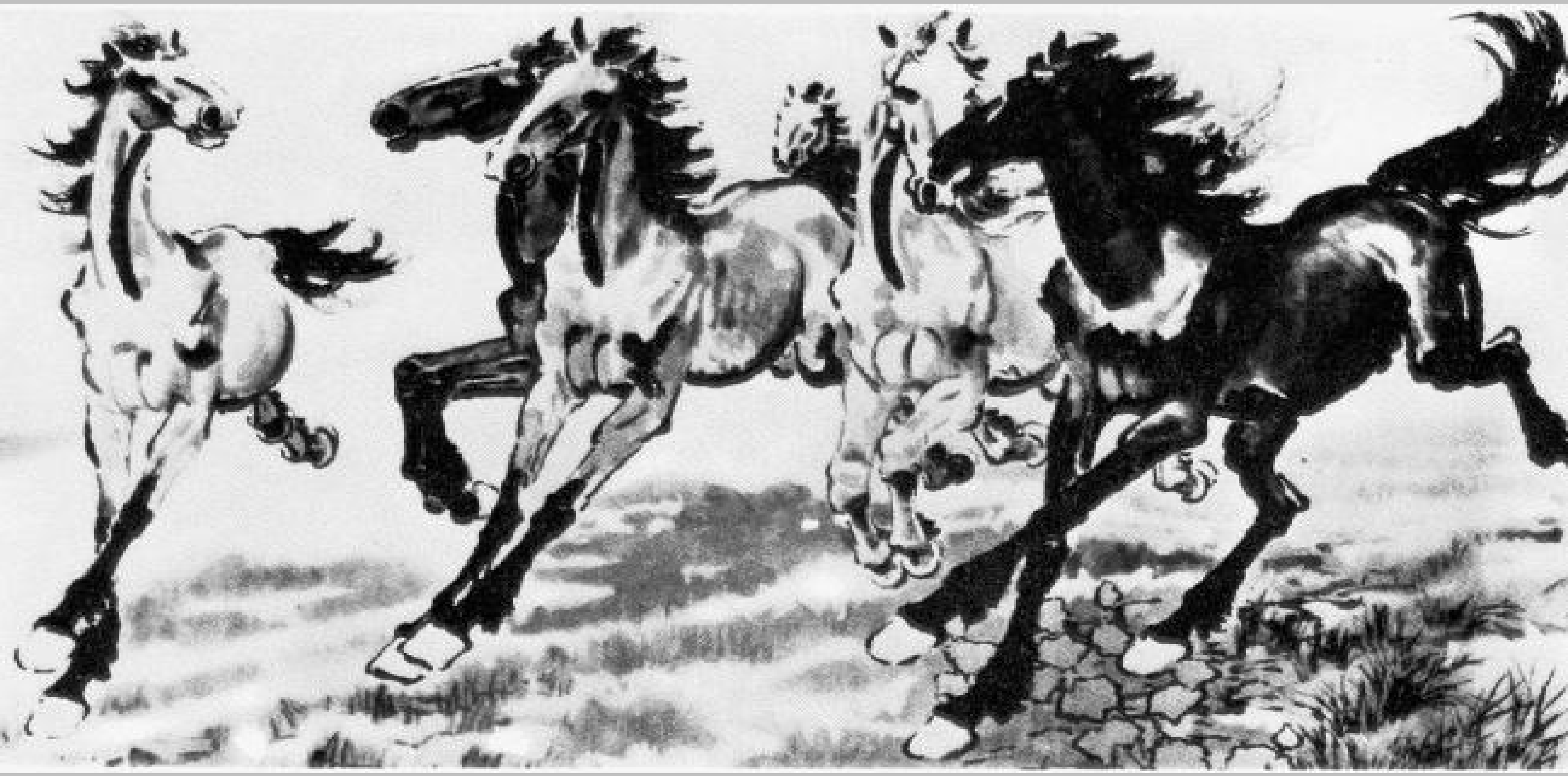

slide credit: S. Ullman

Challenges 3: occlusion

Magritte, 1957

Challenges 4: scale

Challenges 5: deformation



Xu, Beihong 1943

Challenges 6: background clutter

Klimt, 1913

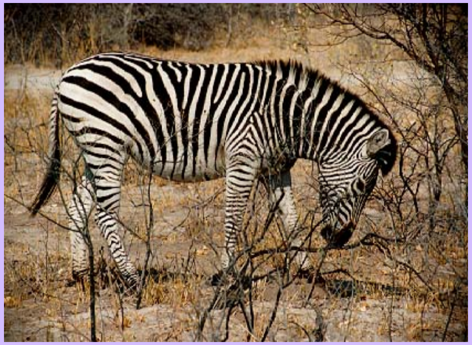
History: single object recognition

History: single object recognition

- Lowe, et al. 1999, 2003
- Mahamud and Herbert, 2000
- Ferrari, Tuytelaars, and Van Gool, 2004
- Rothganger, Lazebnik, and Ponce, 2004
- Moreels and Perona, 2005
- ...

Challenges 7: intra-class variation

Object categorization: the statistical viewpoint



$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

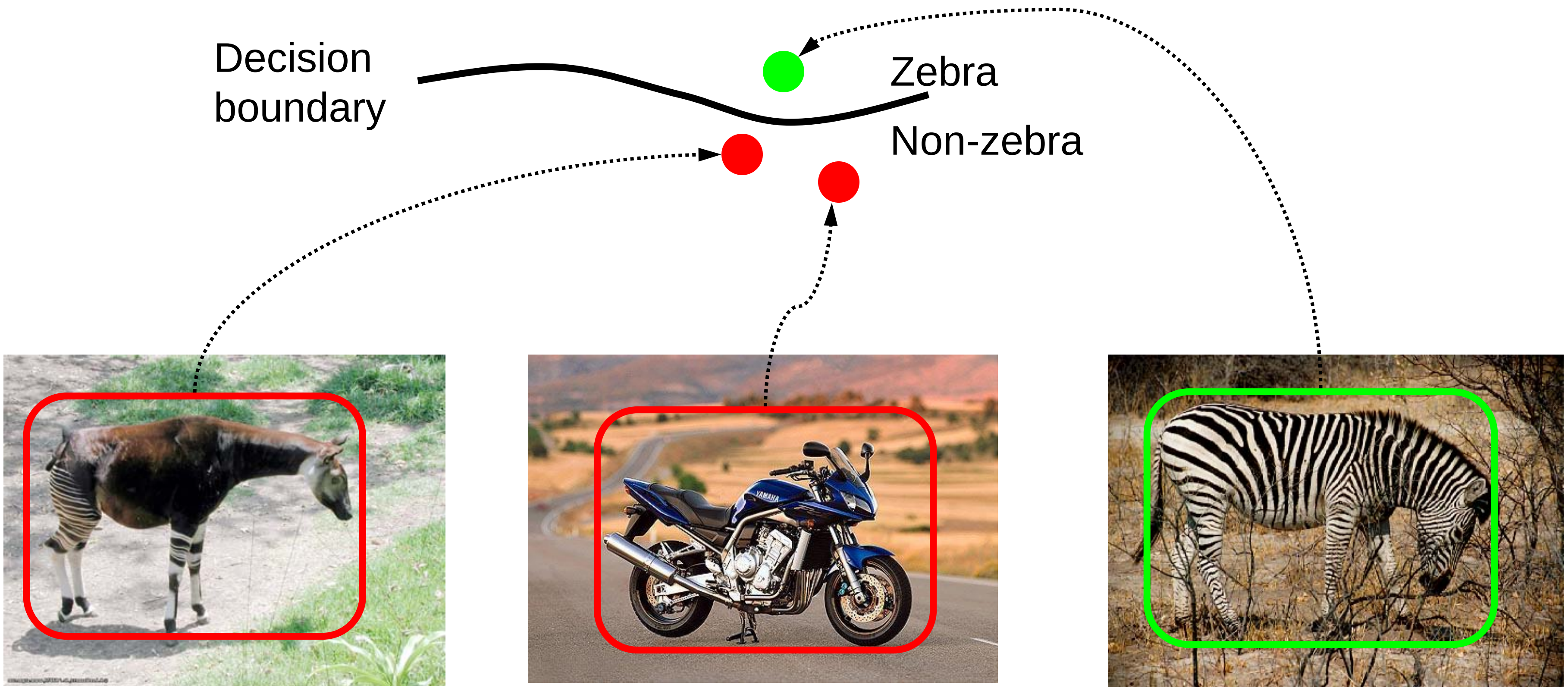
Object categorization: the statistical viewpoint

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

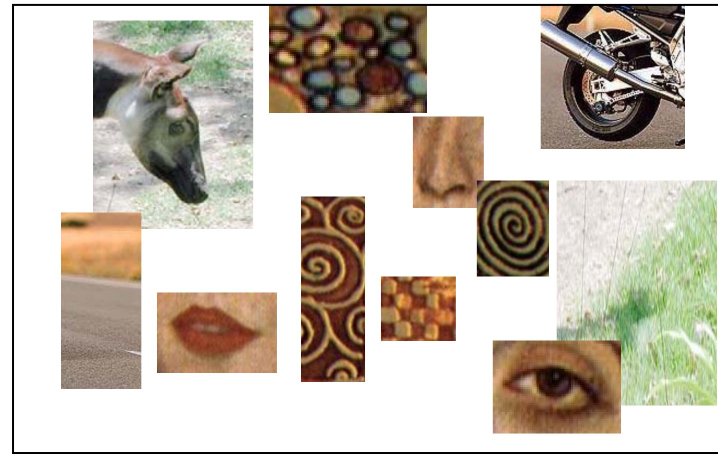
Discriminative

- Direct modeling of $$\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$$



Generative

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$



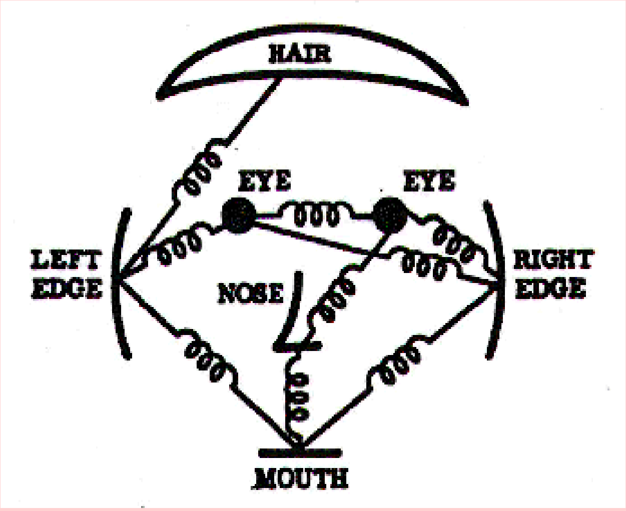
	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

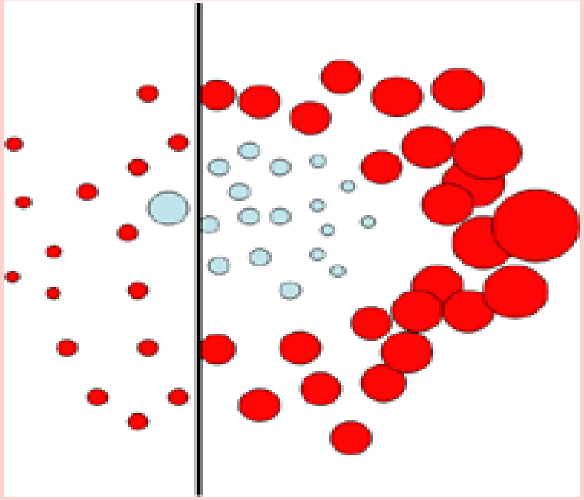
Three main issues

- Representation
  – How to represent an object category
- Learning
  – How to form the classifier, given training data
- Recognition
  – How the classifier is to be used on novel data

Representation

– Generative / discriminative / hybrid

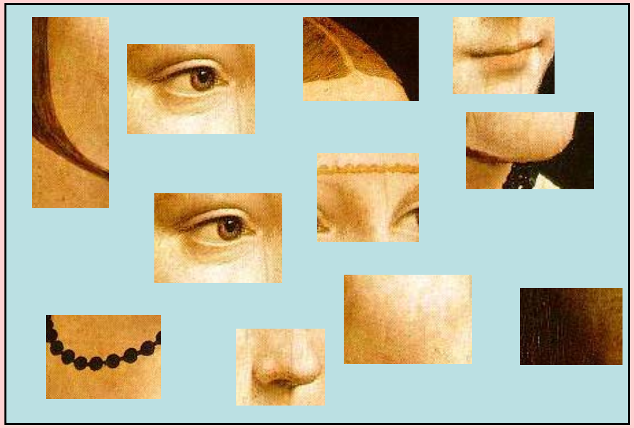


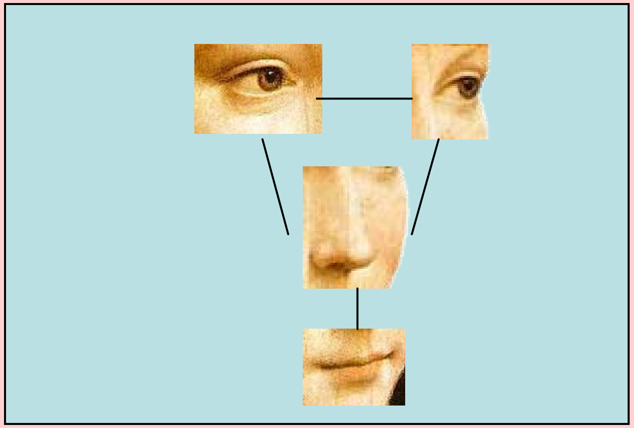


Representation

– Generative / discriminative / hybrid

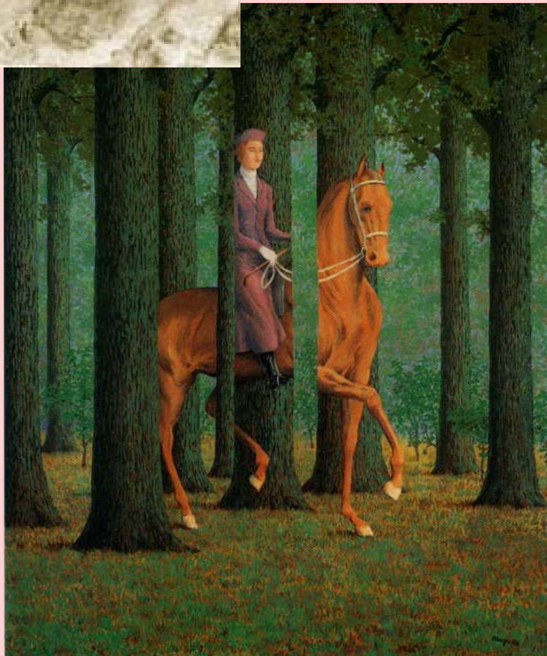
– Appearance only or location and appearance





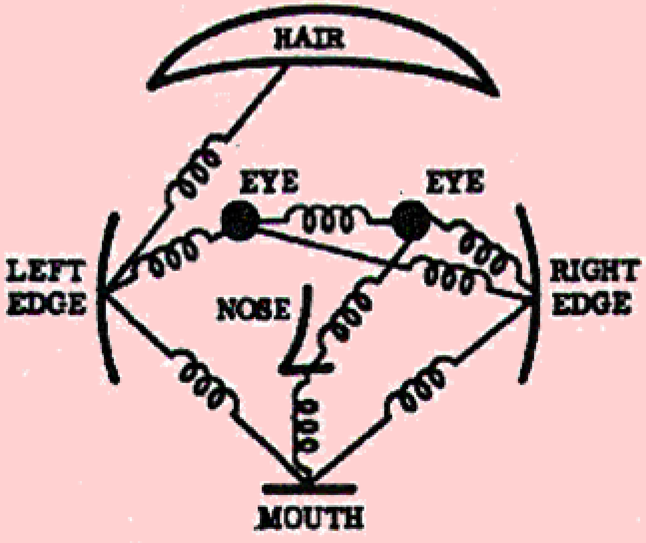
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.



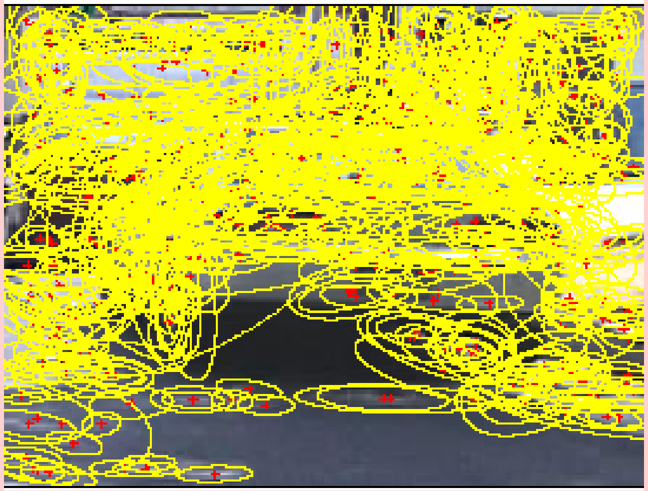
Representation

– Generative / discriminative / hybrid
– Appearance only or location and appearance
– invariances
– Part-based or global w/sub-window



Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- invariances
- Parts or global w/sub-window
- Use set of features or each pixel in image

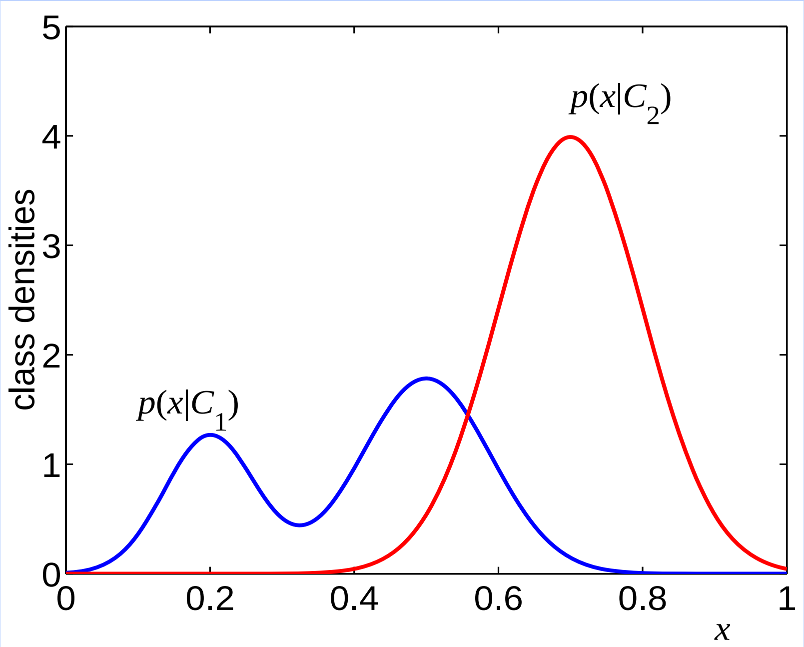


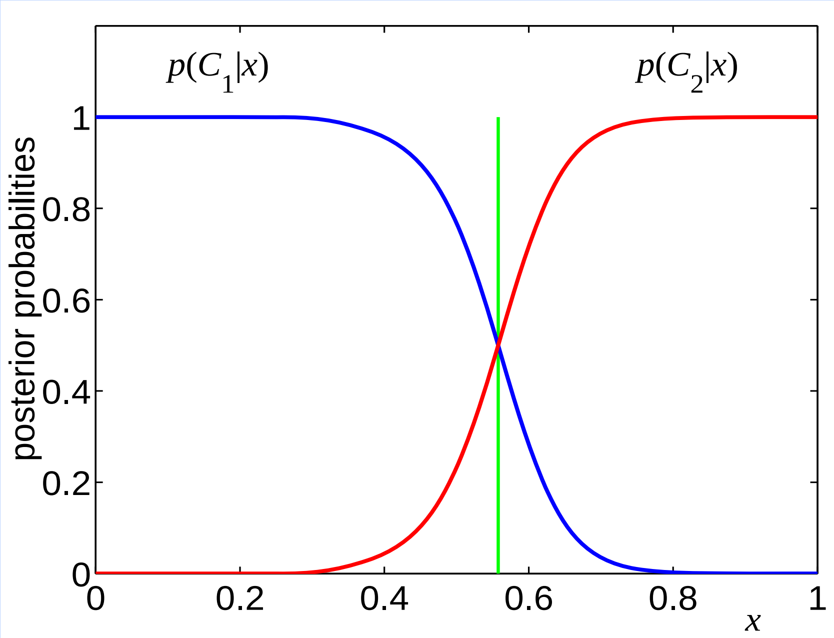
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
– Methods of training: generative vs. discriminative

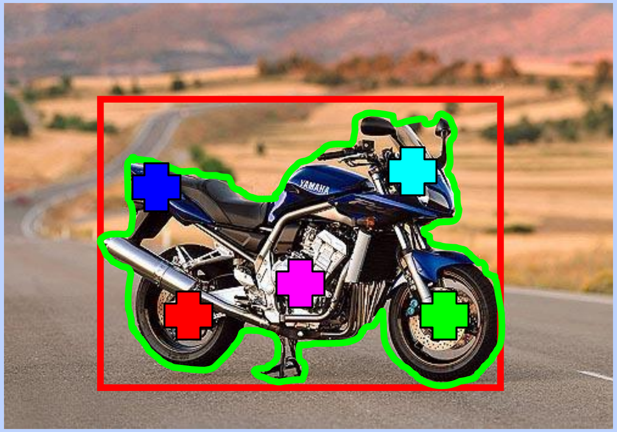




Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

Contains a motorbike



Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods
- Priors

Recognition

– Scale / orientation range to search over

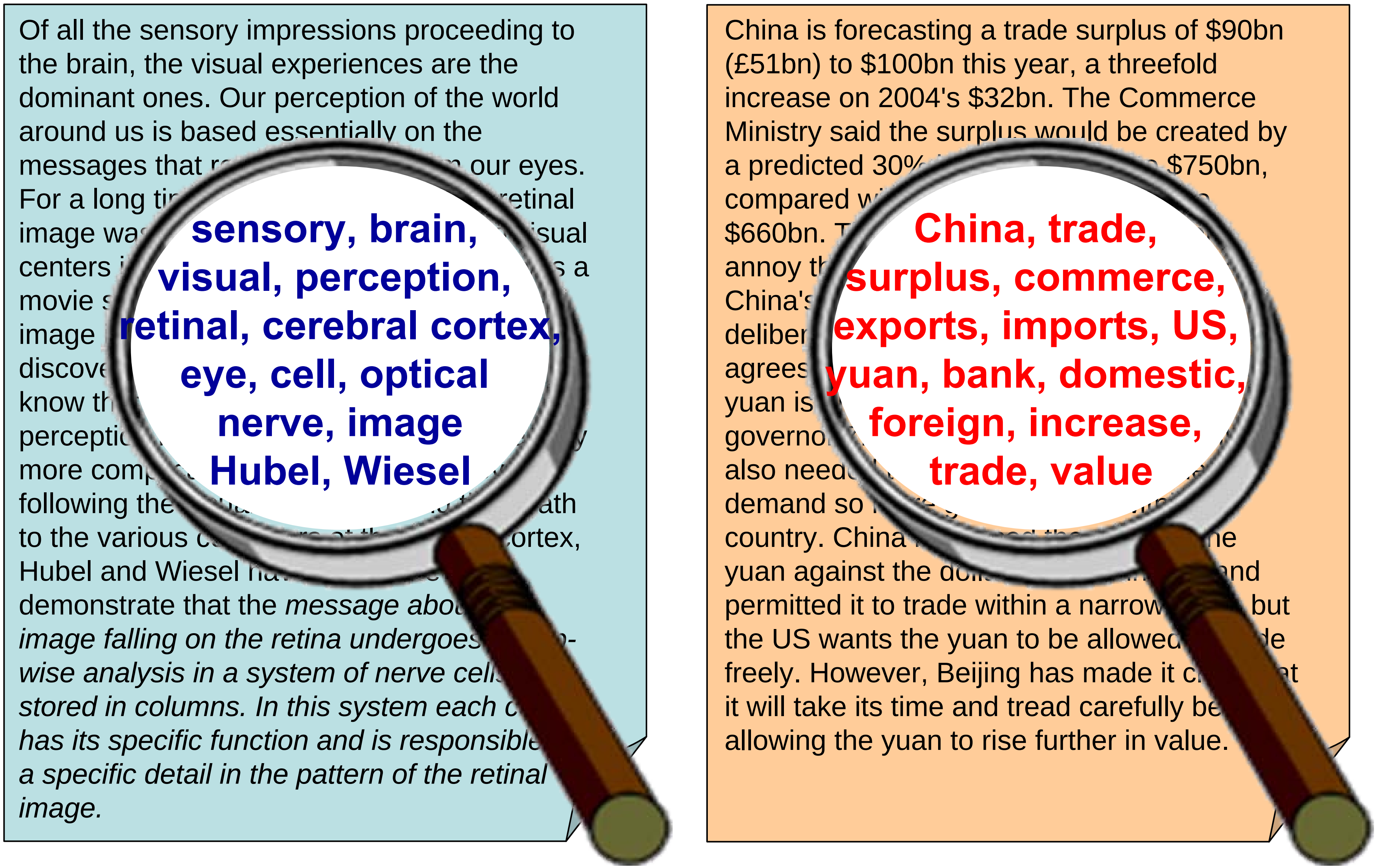
– Speed

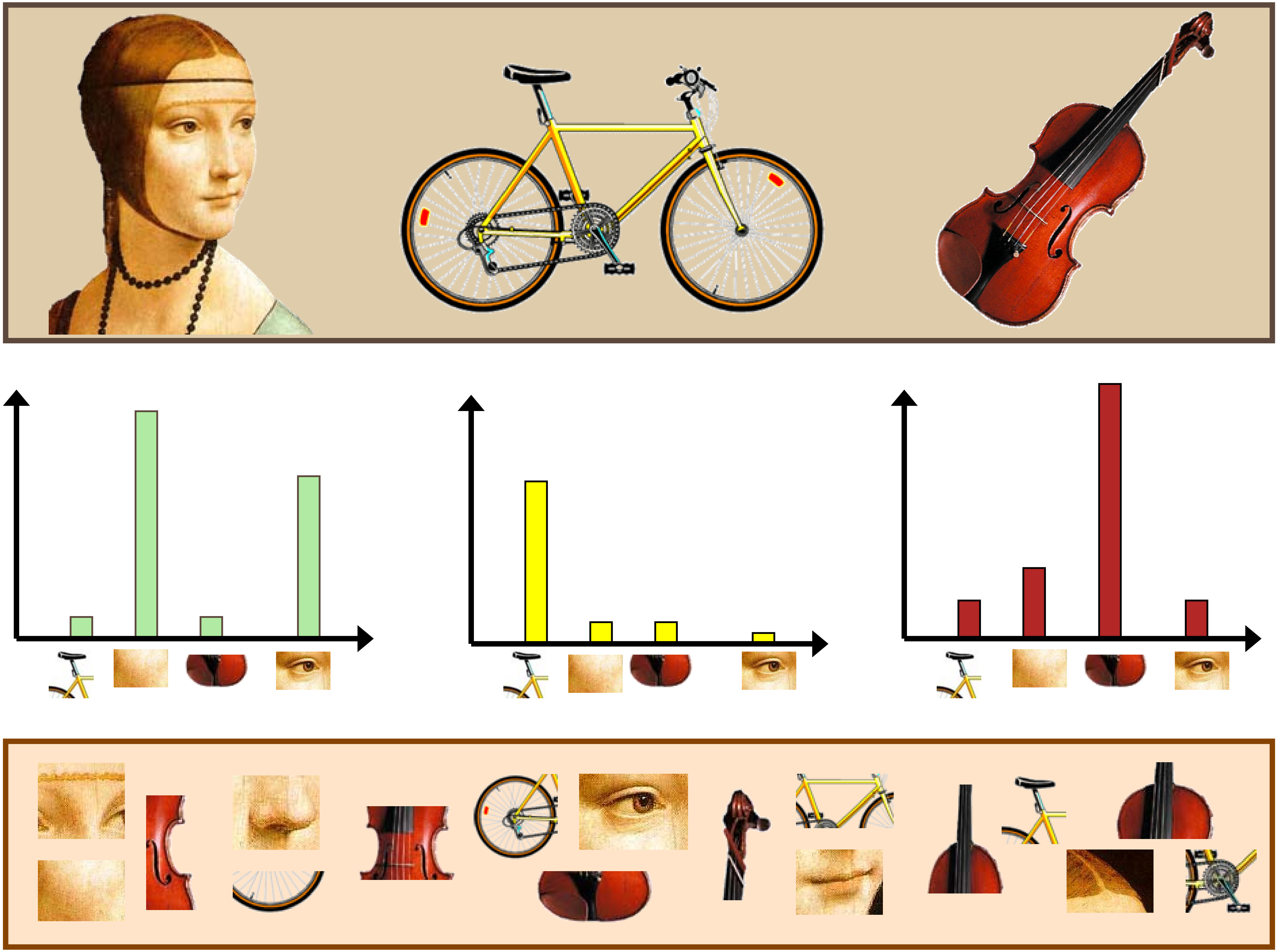
Bag-of-words models

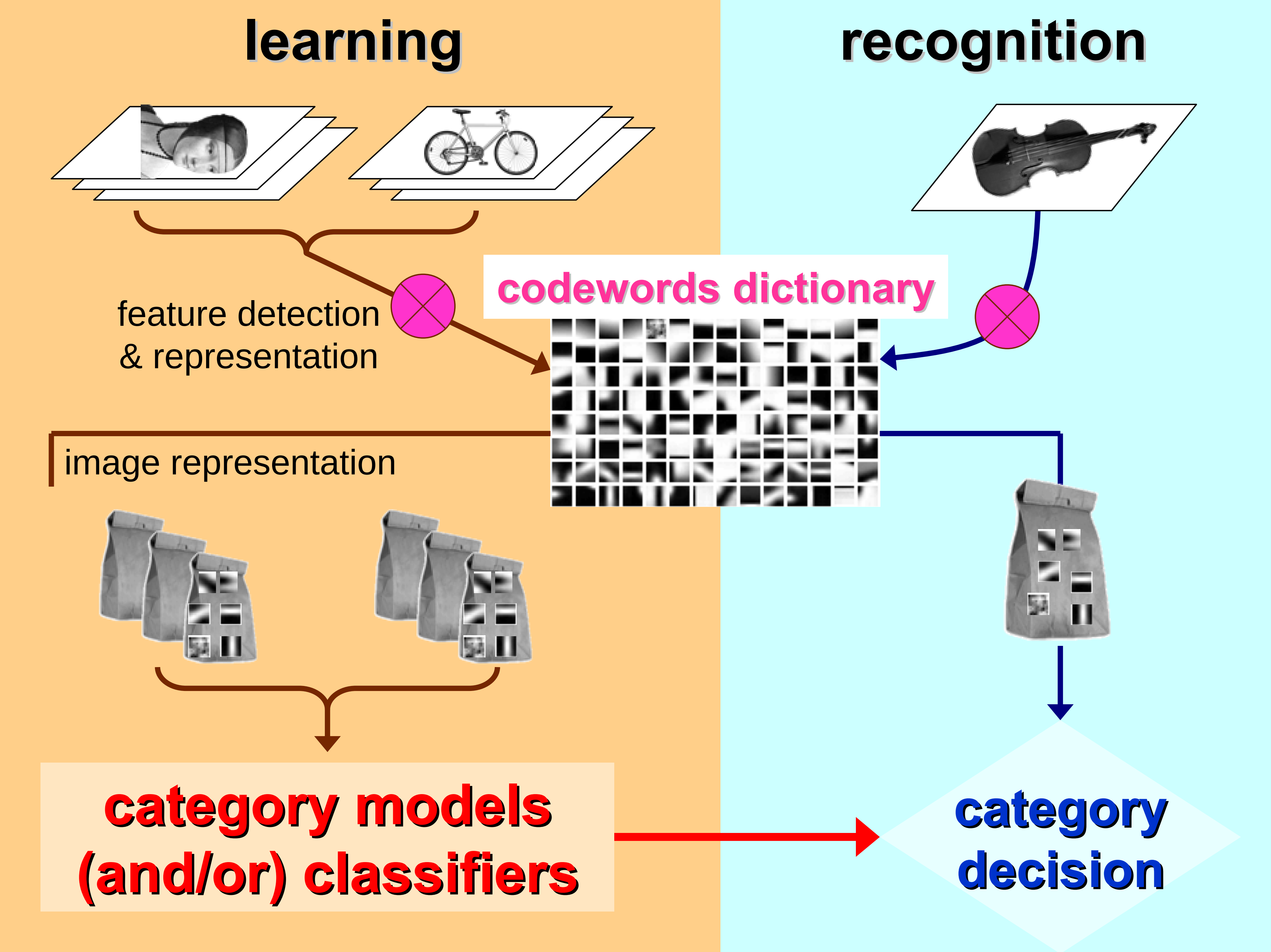
Object

Bag of ‘words’

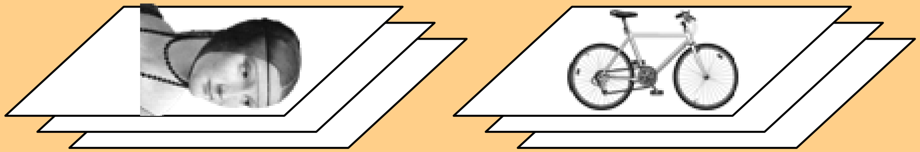
Analogy to documents






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

Representation



1. feature detection & representation

2. codewords dictionary

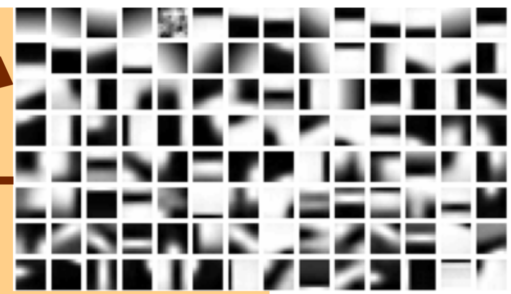
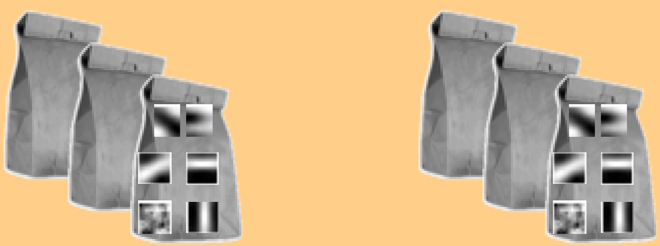
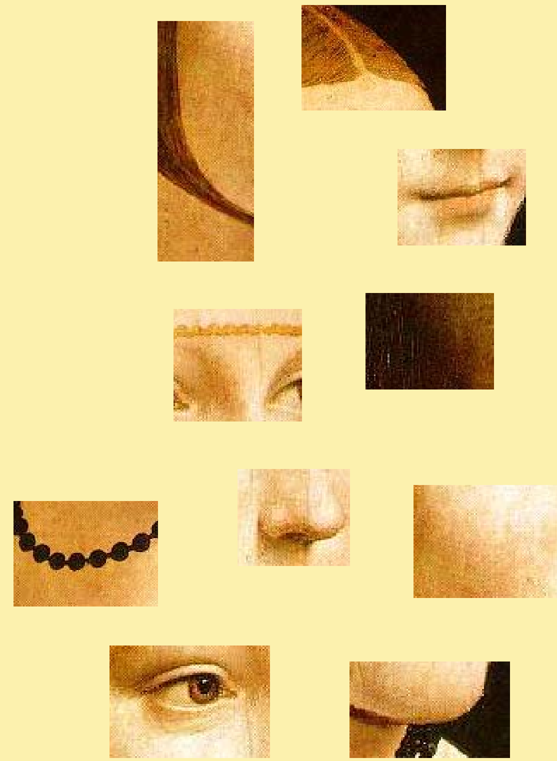


image representation

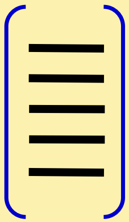
3.



1.Feature detection and representation

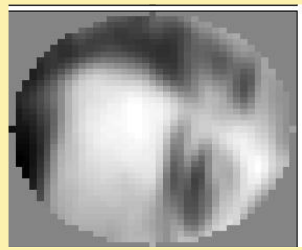


1.Feature detection and representation

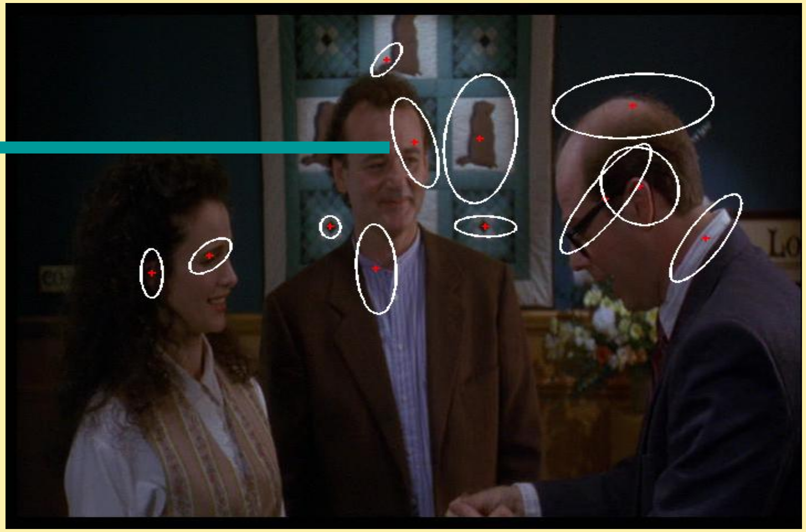


Compute
SIFT
descriptor

[Lowe'99]



Normalize
patch



Detect patches

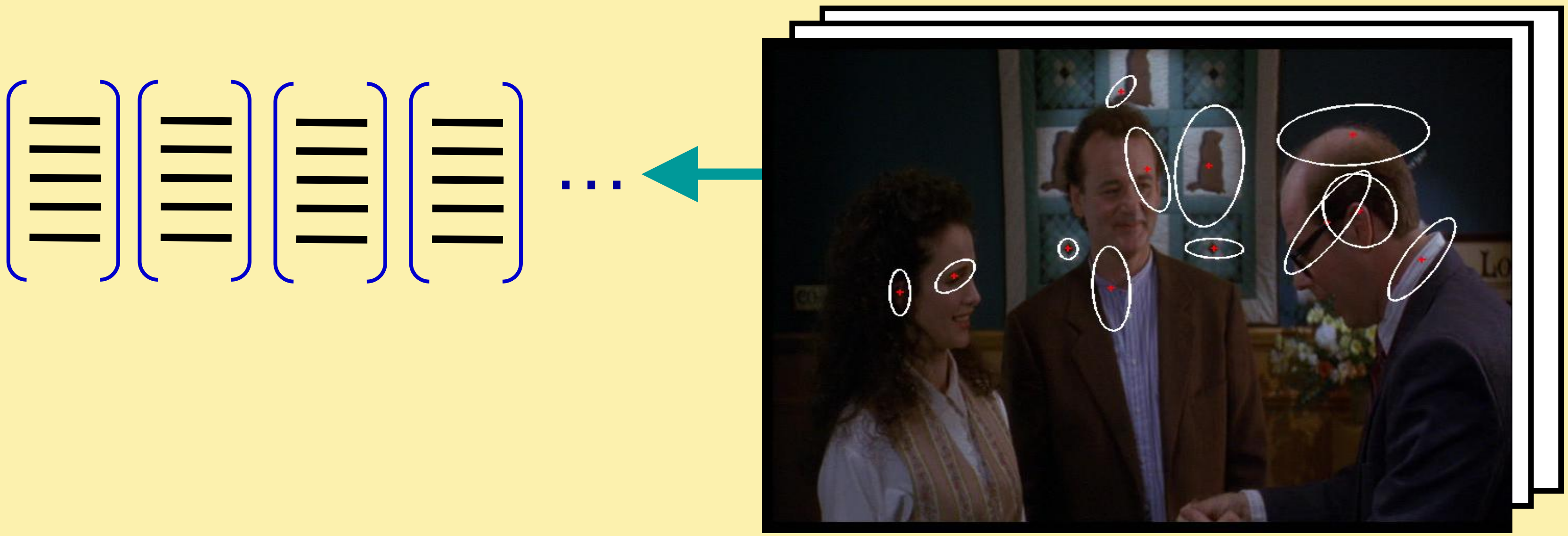
[Mikojaczyk and Schmid '02]

[Matas et al. '02]

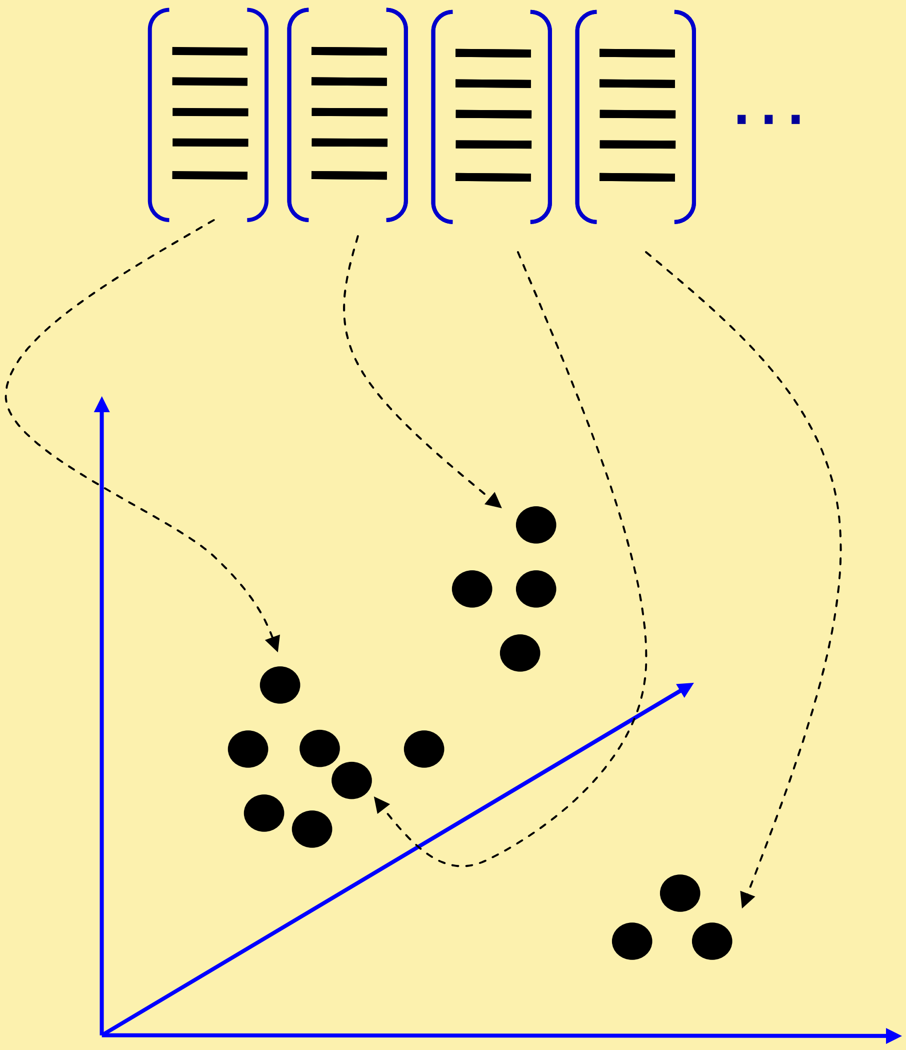
[Sivic et al. '03]

Slide credit: Josef Sivic

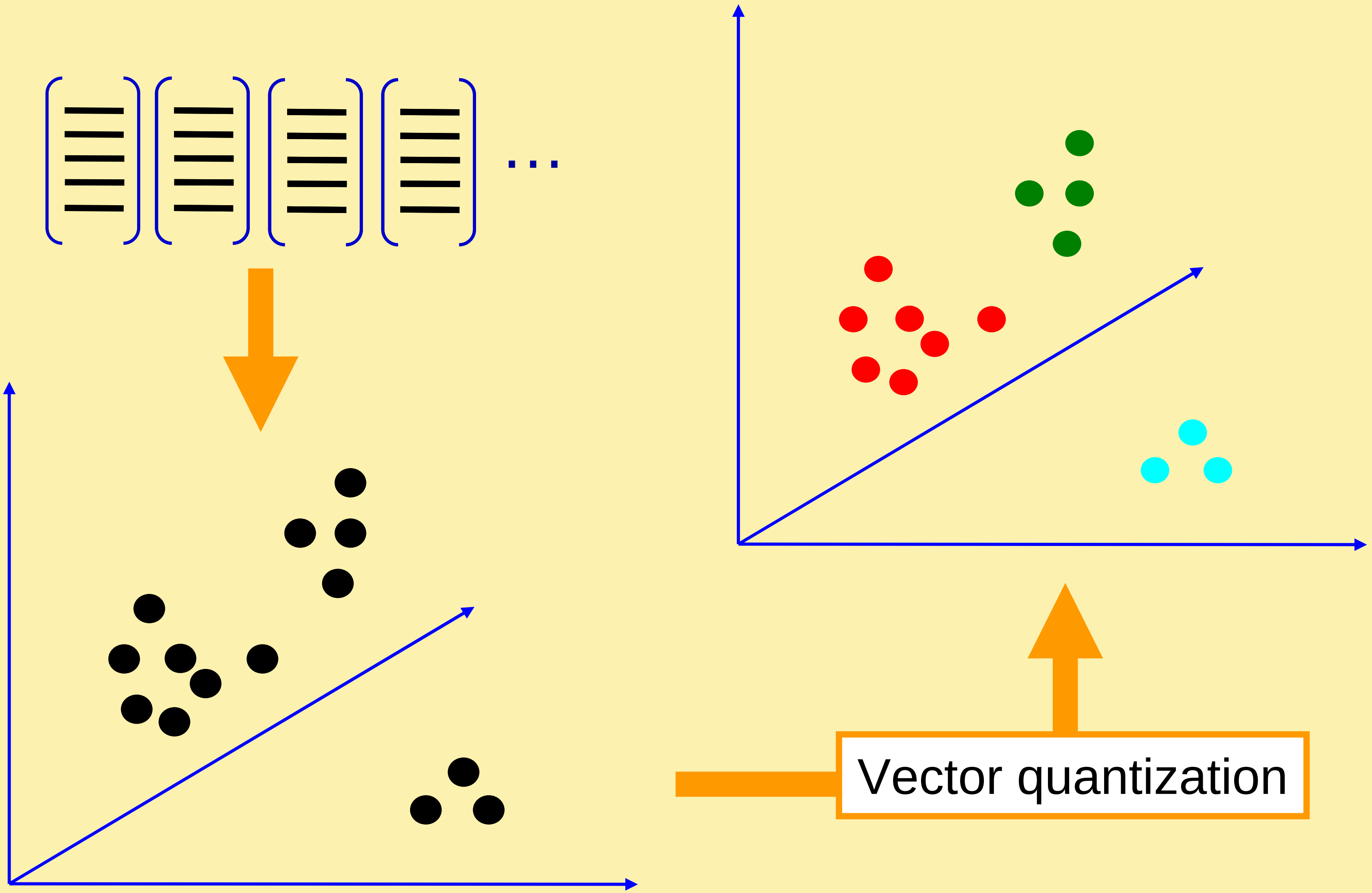
1.Feature detection and representation



2. Codewords dictionary formation

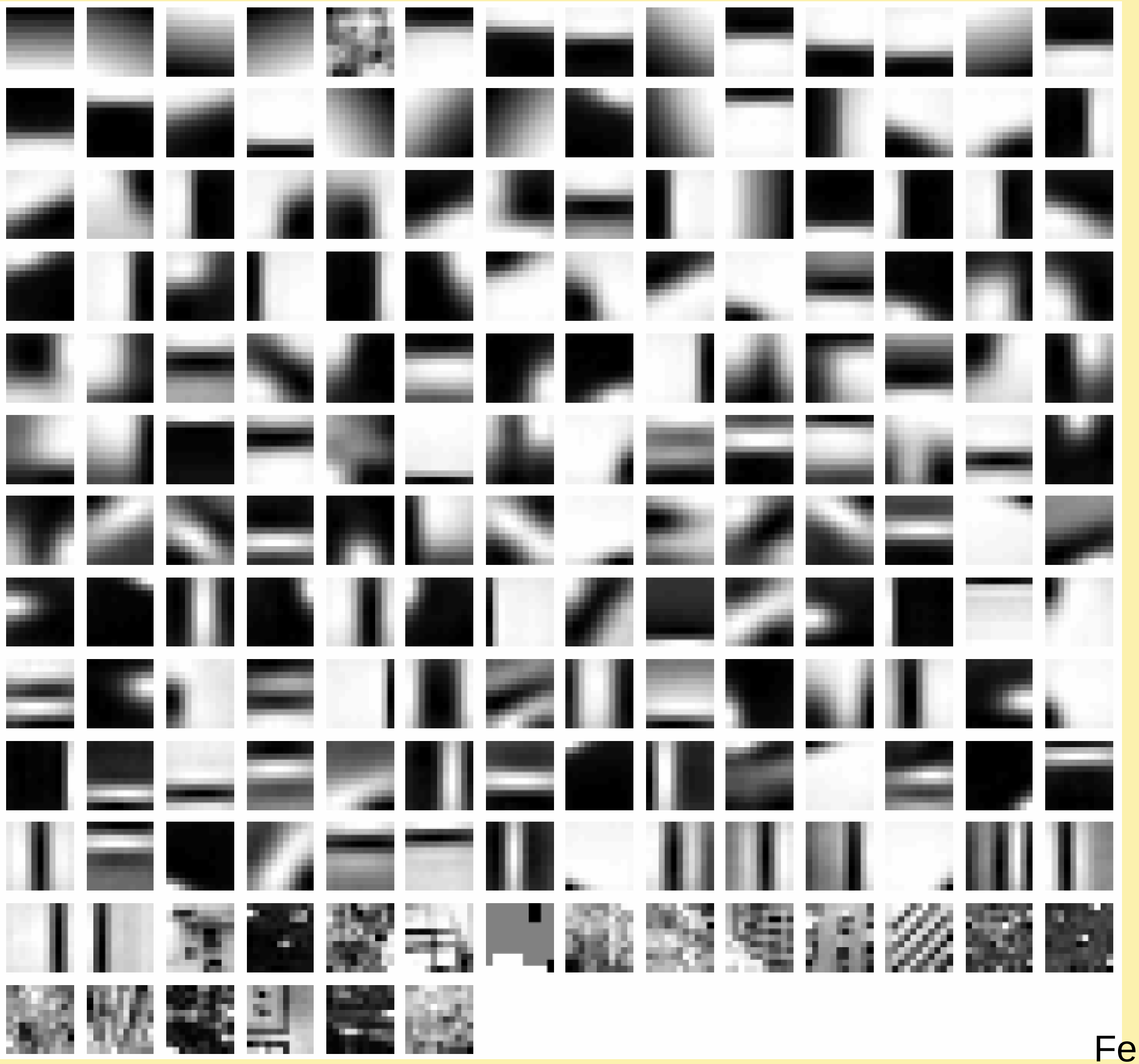


2. Codewords dictionary formation



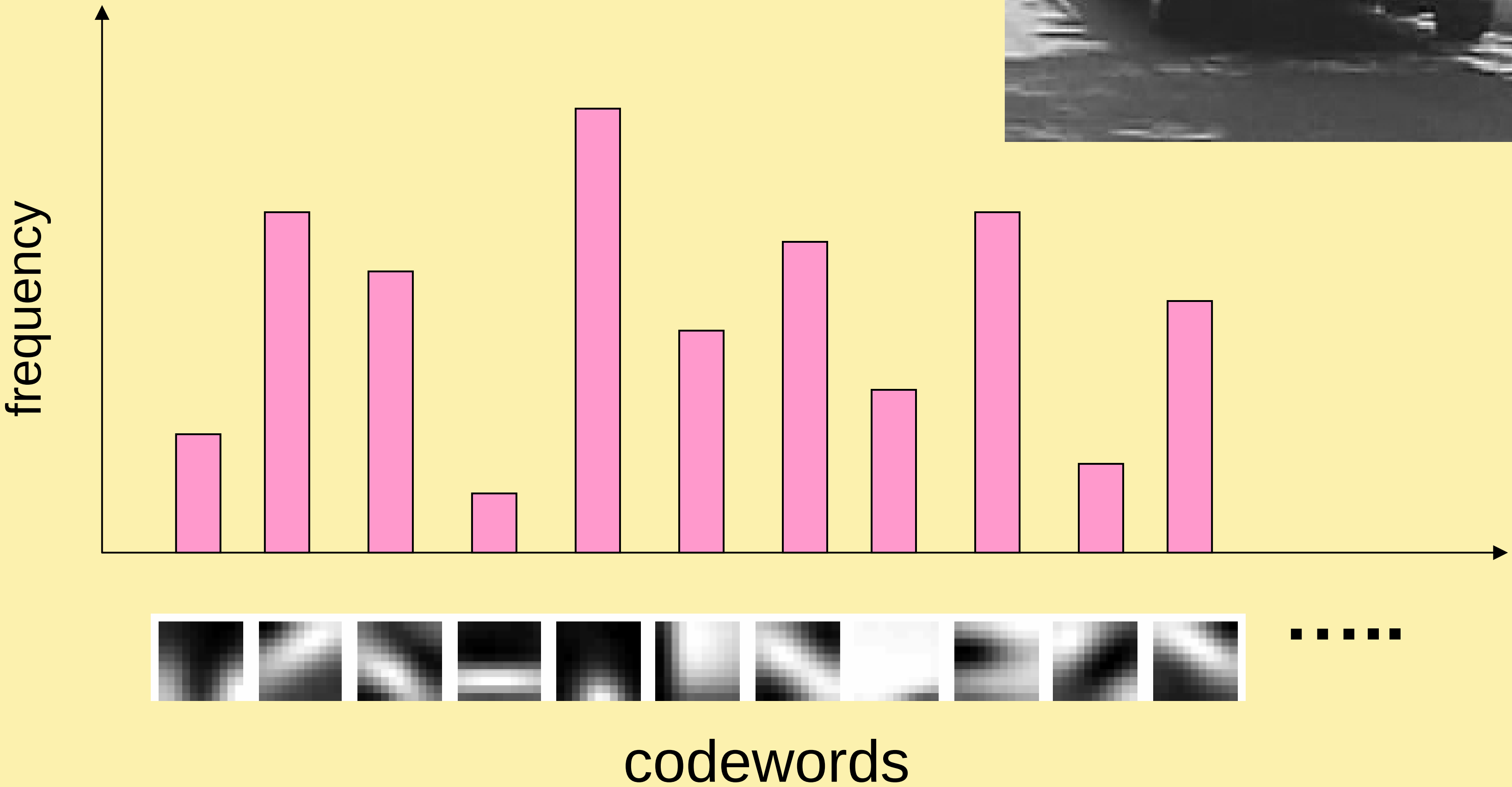
Slide credit: Josef Sivic

2. Codewords dictionary formation

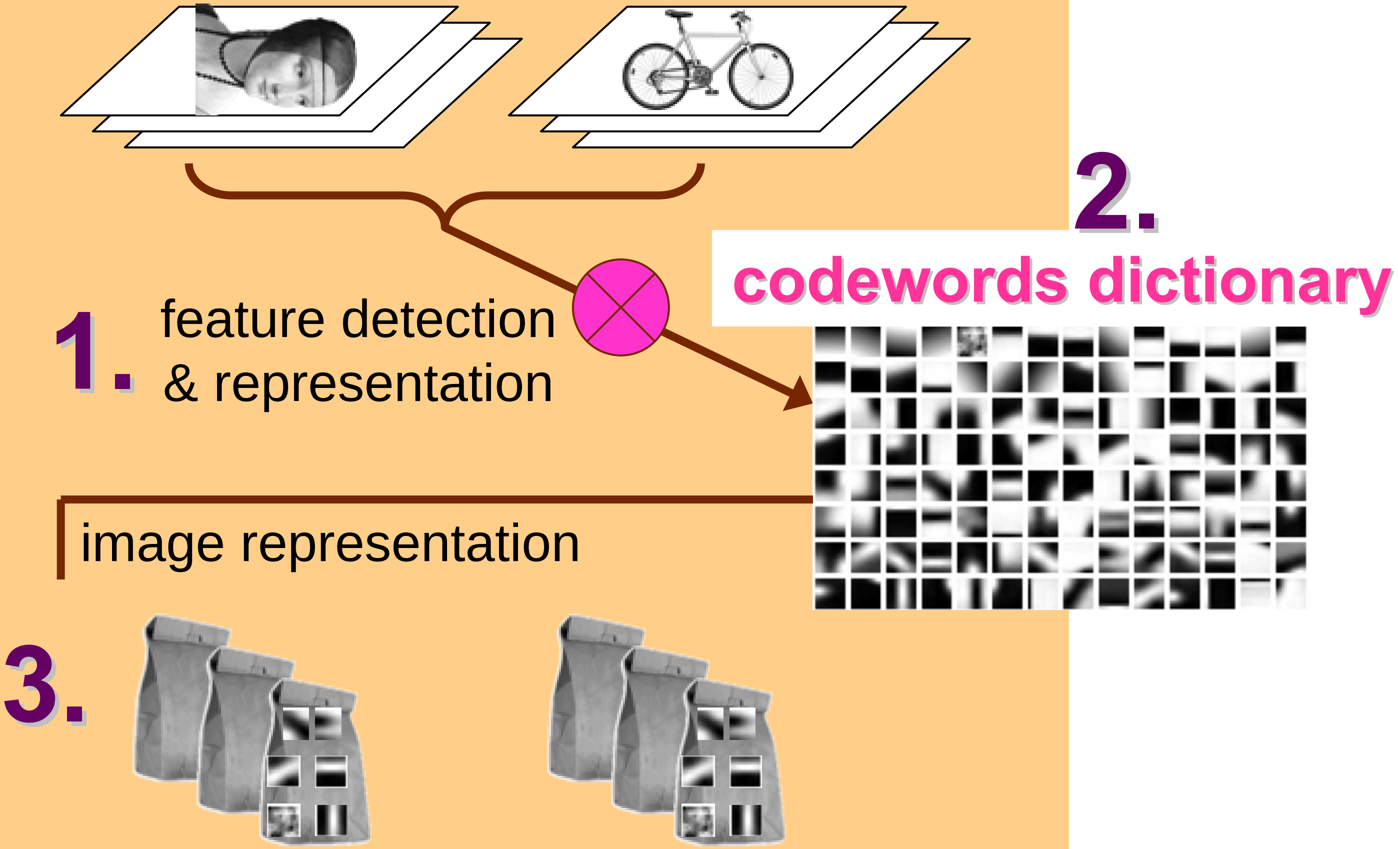


Fei-Fei et al. 2005

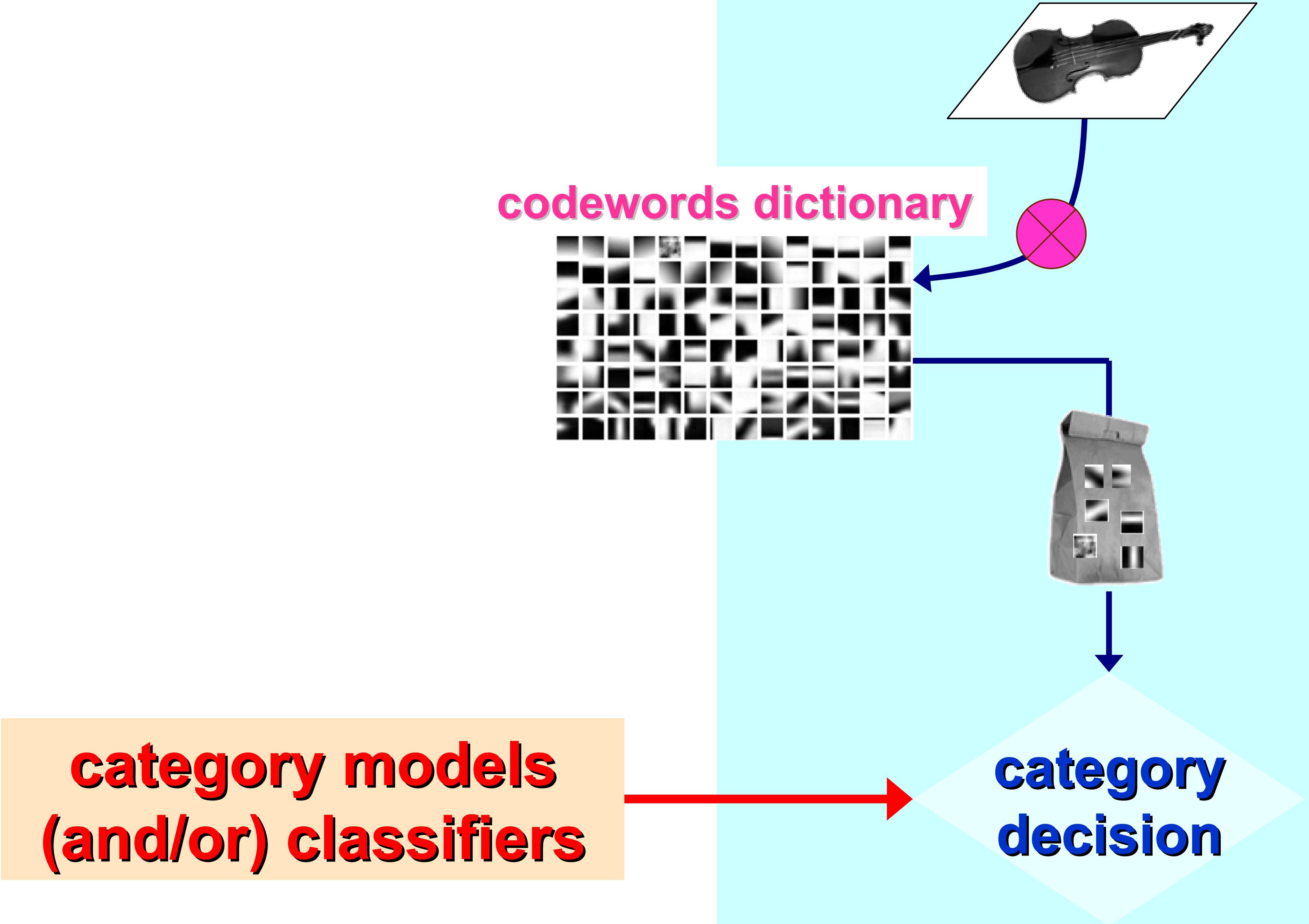
3. Image representation



Representation



Learning and Recognition



2 case studies

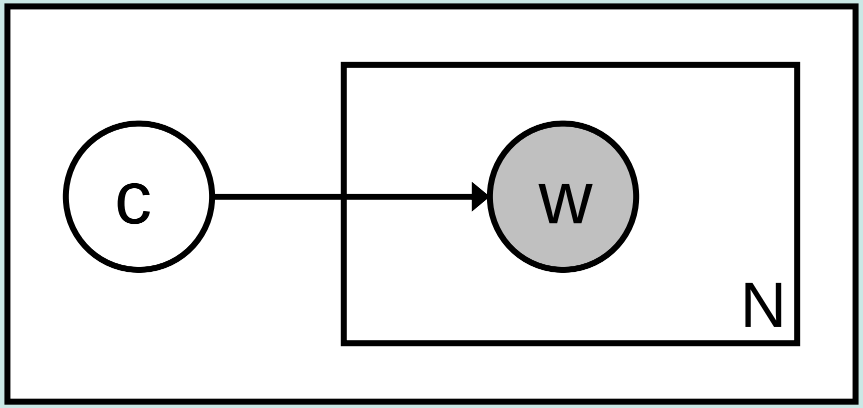
1. Naïve Bayes classifier
   – Csurka et al. 2004

2. Hierarchical Bayesian text models (pLSA and LDA)
   – Background: Hoffman 2001, Blei et al. 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
  - $w_n = [0,0,\ldots 1,\ldots,0,0]^T$
- **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

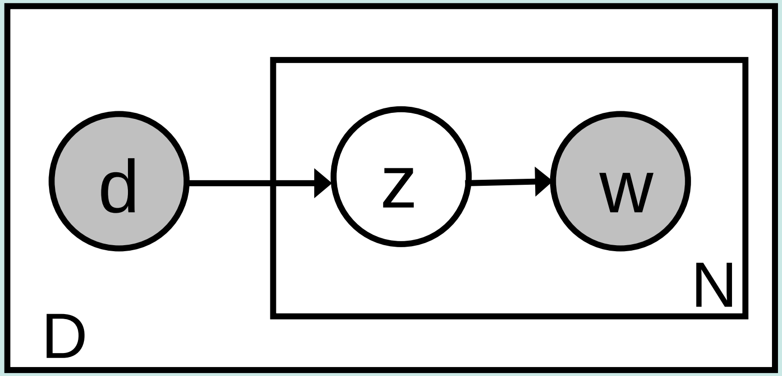
Prior prob. of the object classes

Image likelihood given the class

Csurka et al. 2004

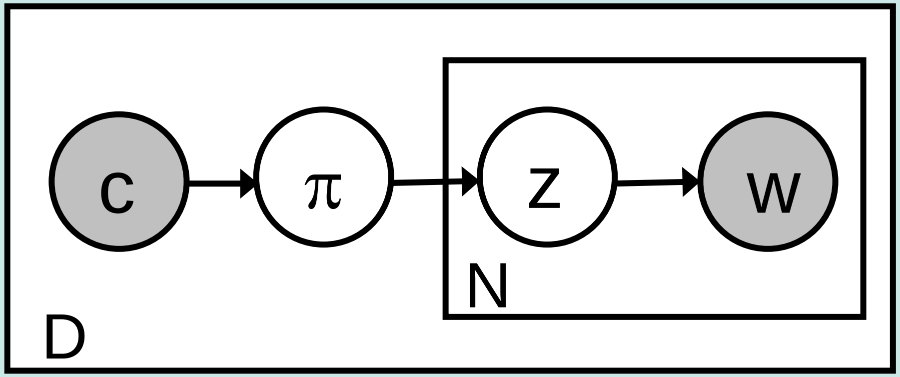
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 2001

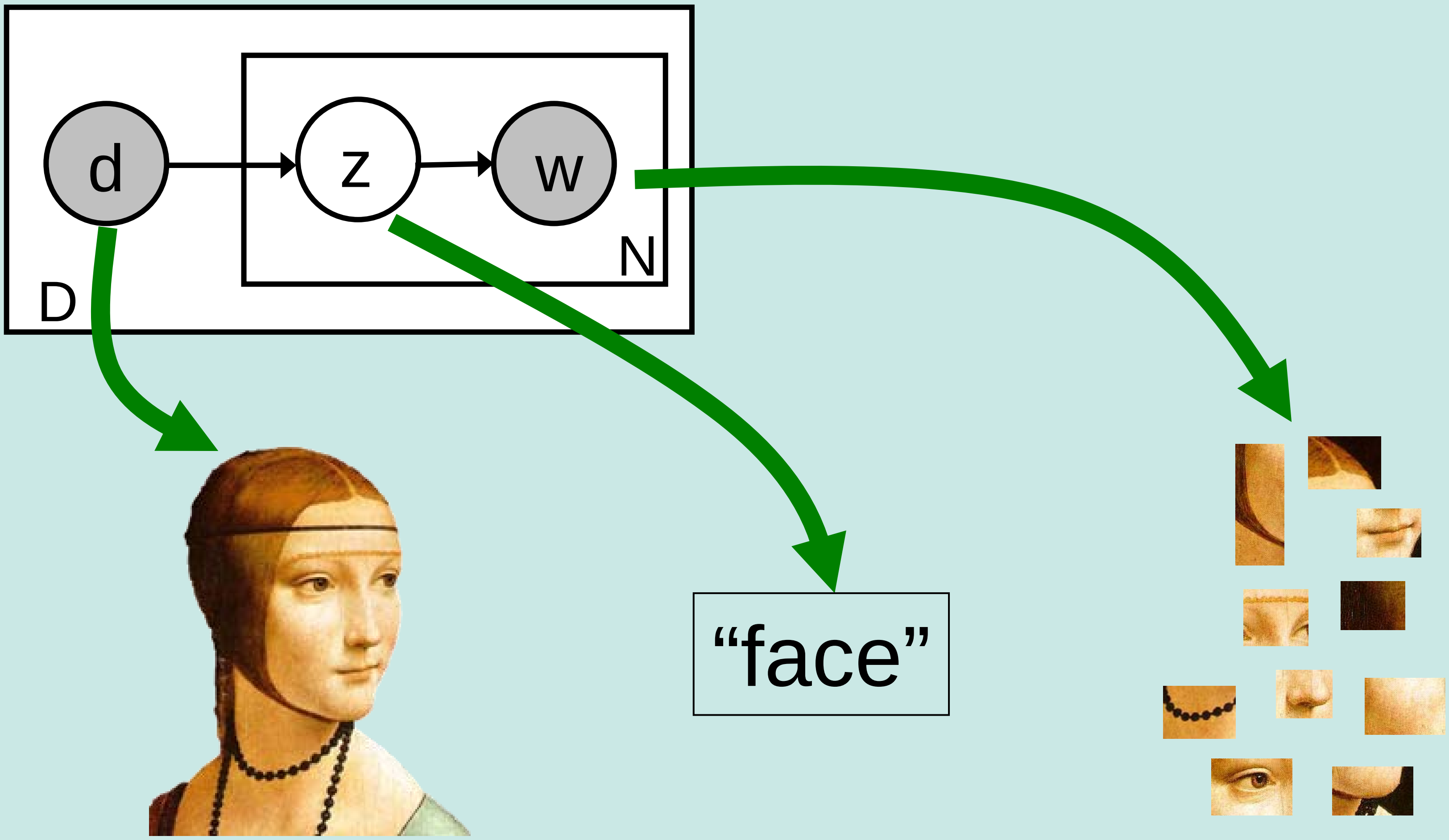
Latent Dirichlet Allocation (LDA)



Blei et al., 2001

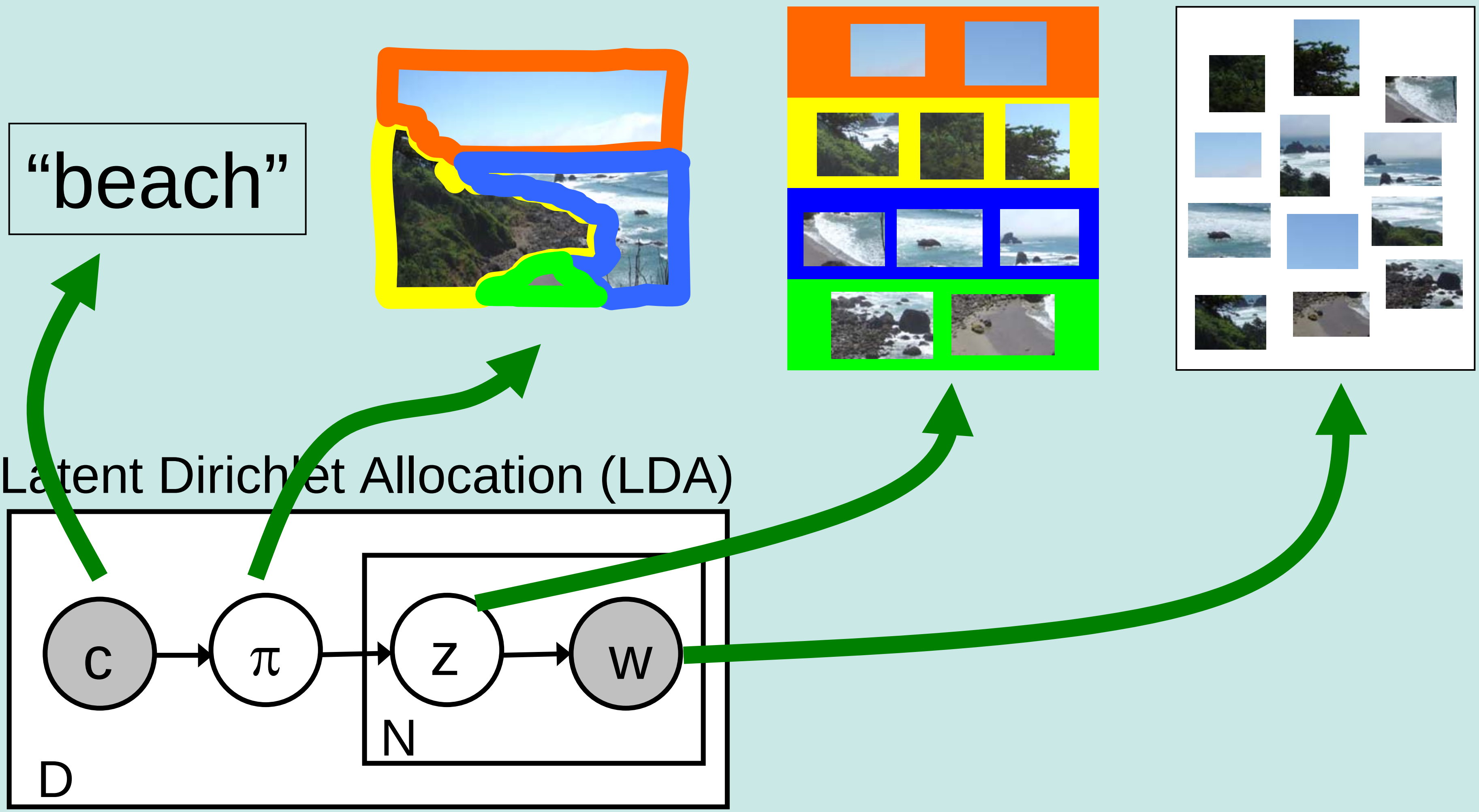
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Sivic et al. ICCV 2005

Case #2: Hierarchical Bayesian text models



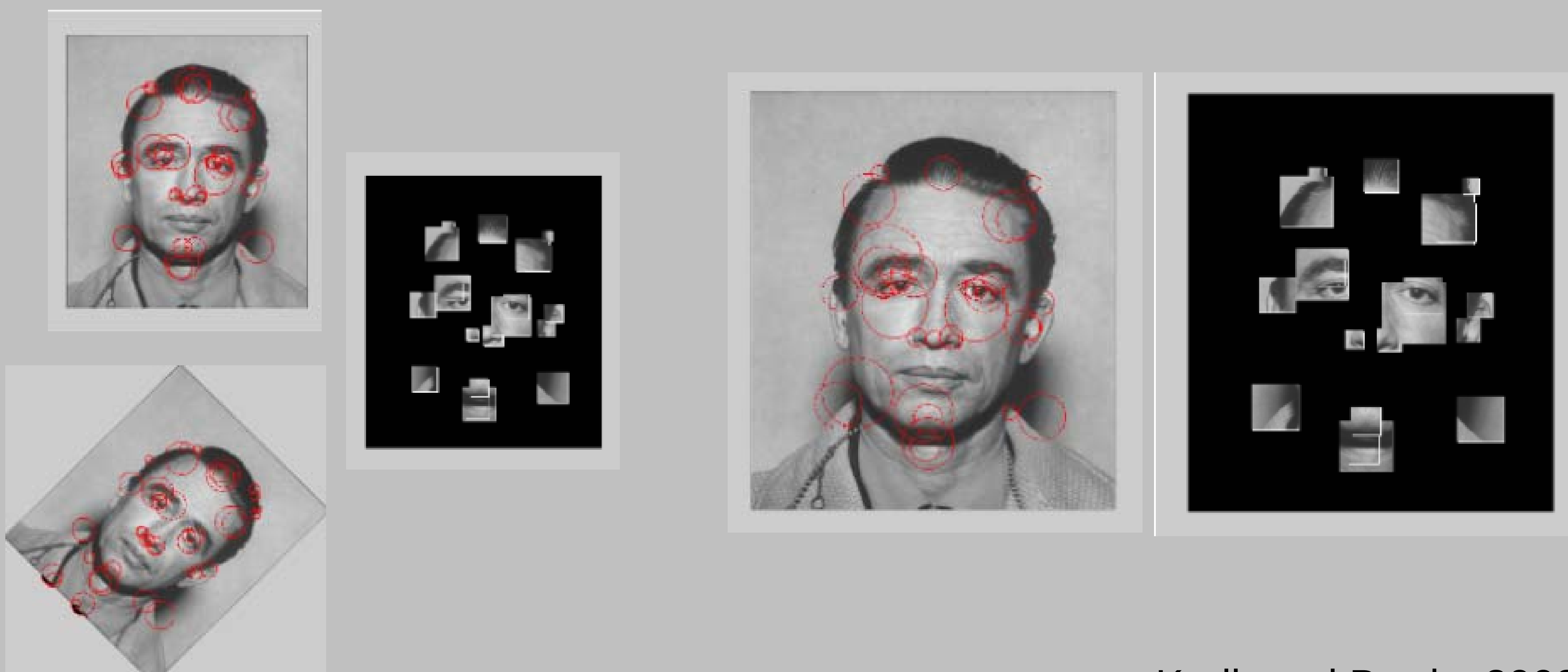
Fei-Fei et al. ICCV 2005

Another application

- Human action classification

Invariance issues

- Scale and rotation
  - Implicit
  - Detectors and descriptors



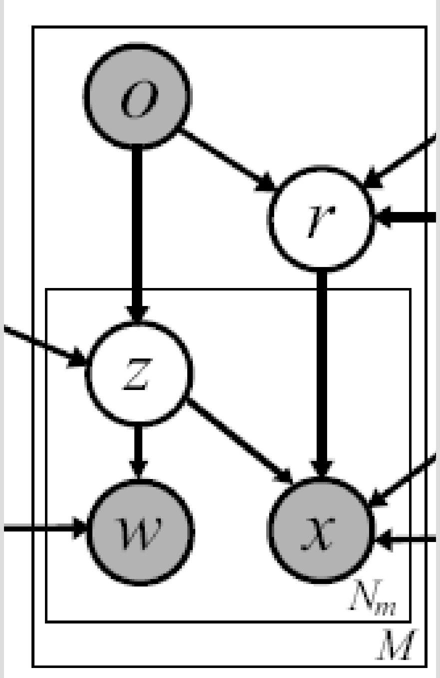
Kadir and Brady. 2003

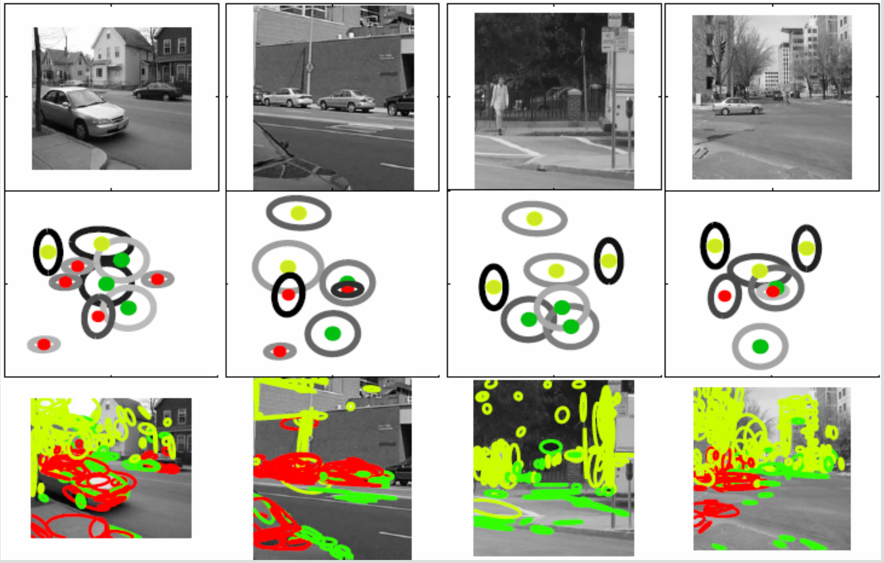
Invariance issues

- 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 et al. 2005

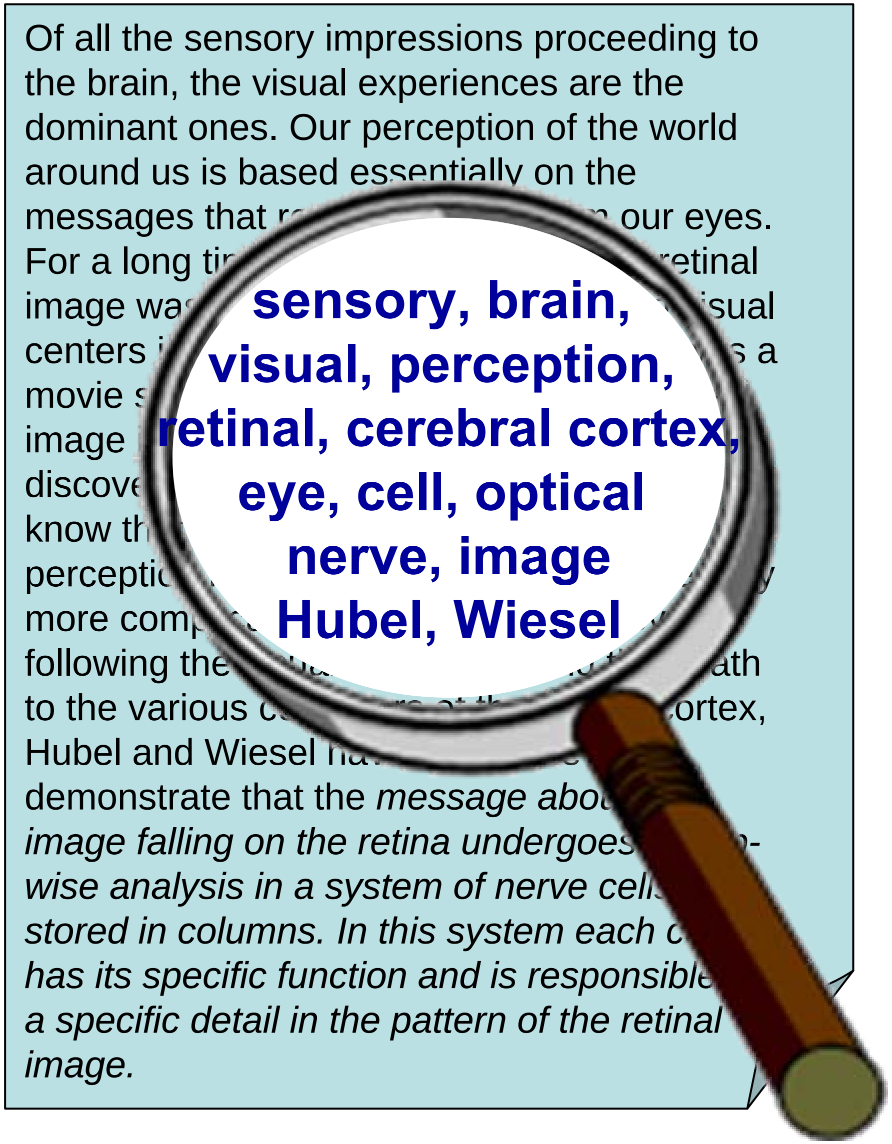
Invariance issues

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

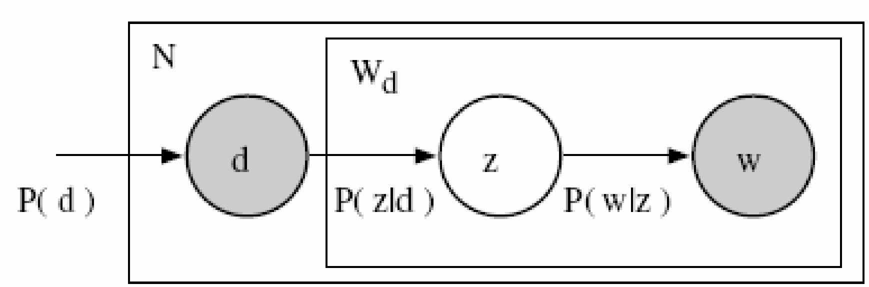
Fergus et al. 2005

Model properties

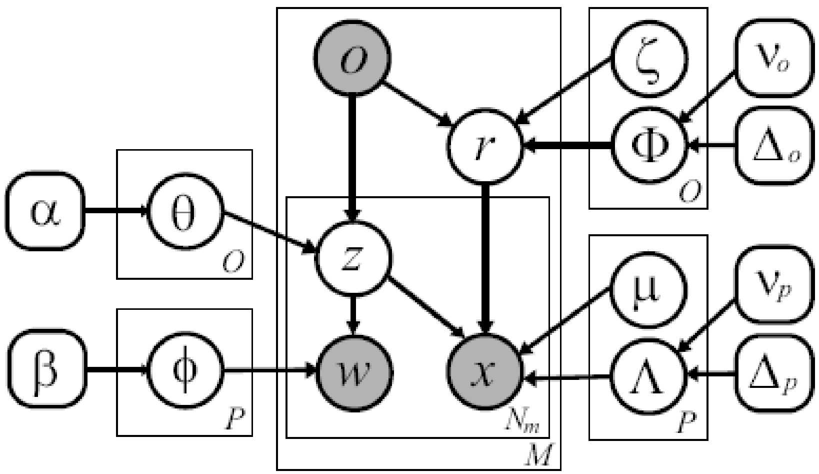
- Intuitive
  - Analogy to documents



Model properties

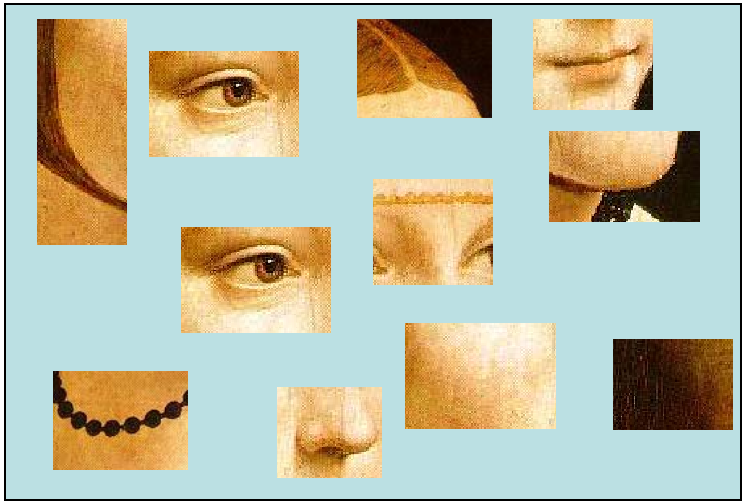


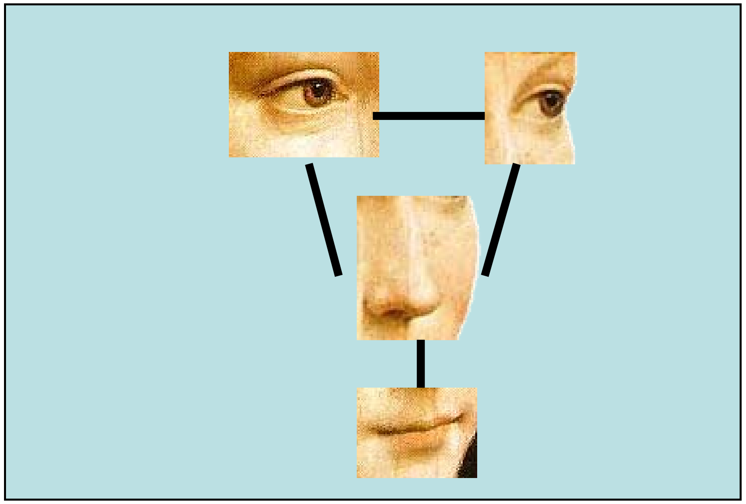
- Intuitive
- (Could use) generative models
  - Convenient for weakly- or un-supervised training
  - Prior information
  - Hierarchical Bayesian framework



Sivic et al., 2005, Sudderth et al., 2005

Model properties

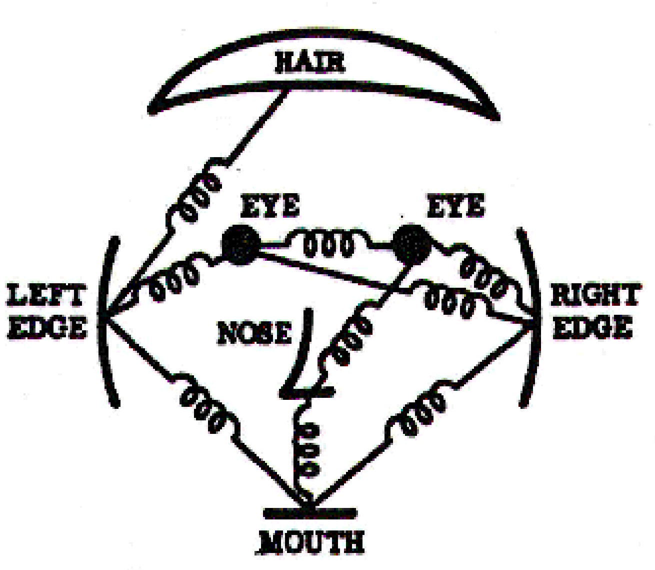




- Intuitive
- (Could use) generative models
- Learning and recognition relatively fast
  - Compare to other methods

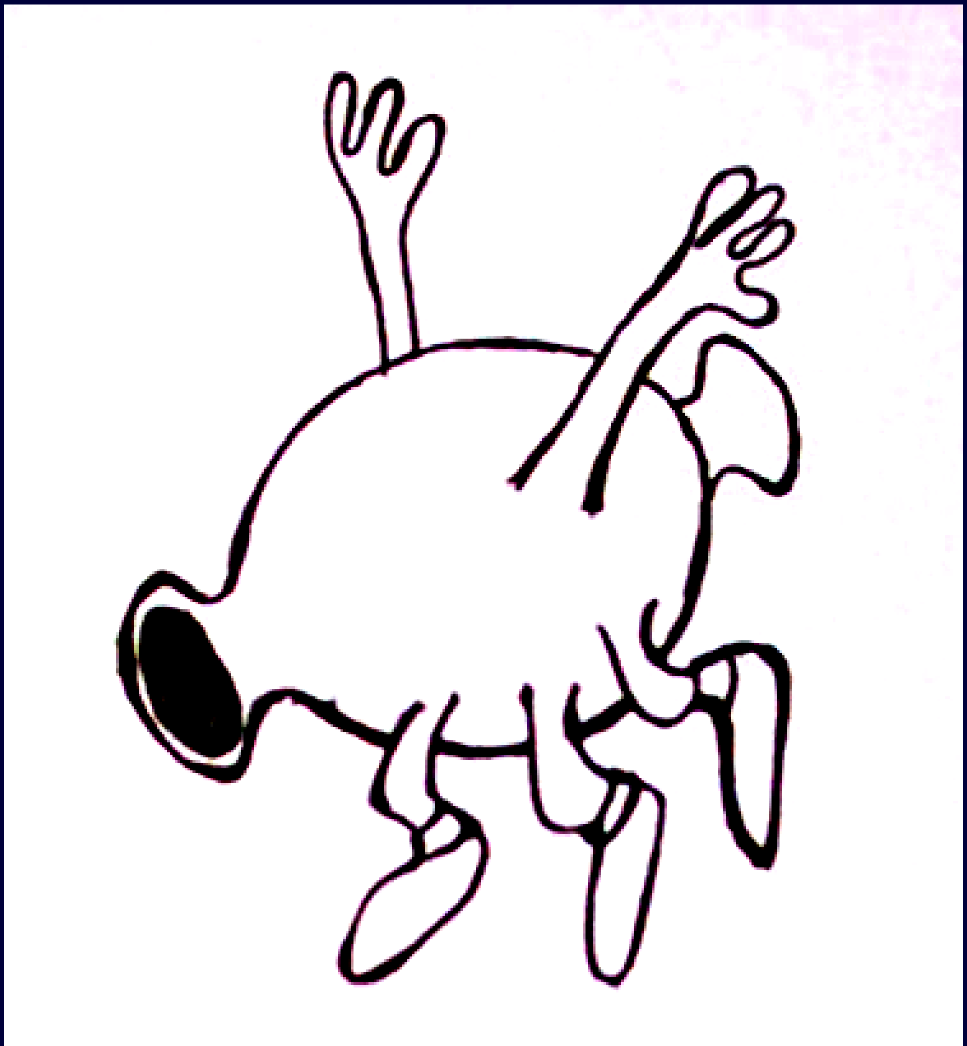
Weakness of the model

- 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



part-based models

Slides courtesy to Rob Fergus for “part-based models”



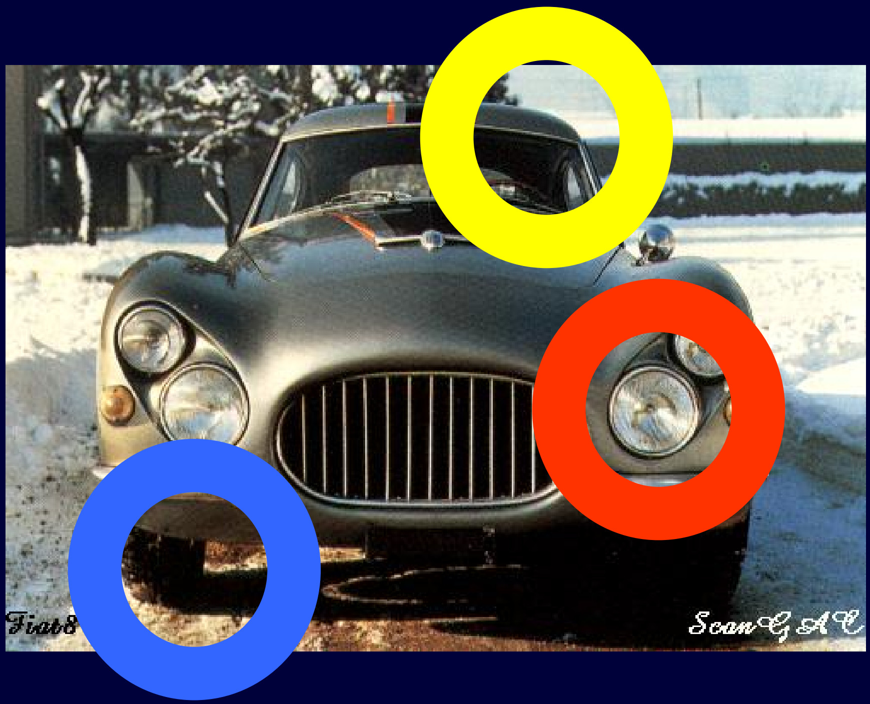
One-shot learning of object categories

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

One-shot learning of object categories

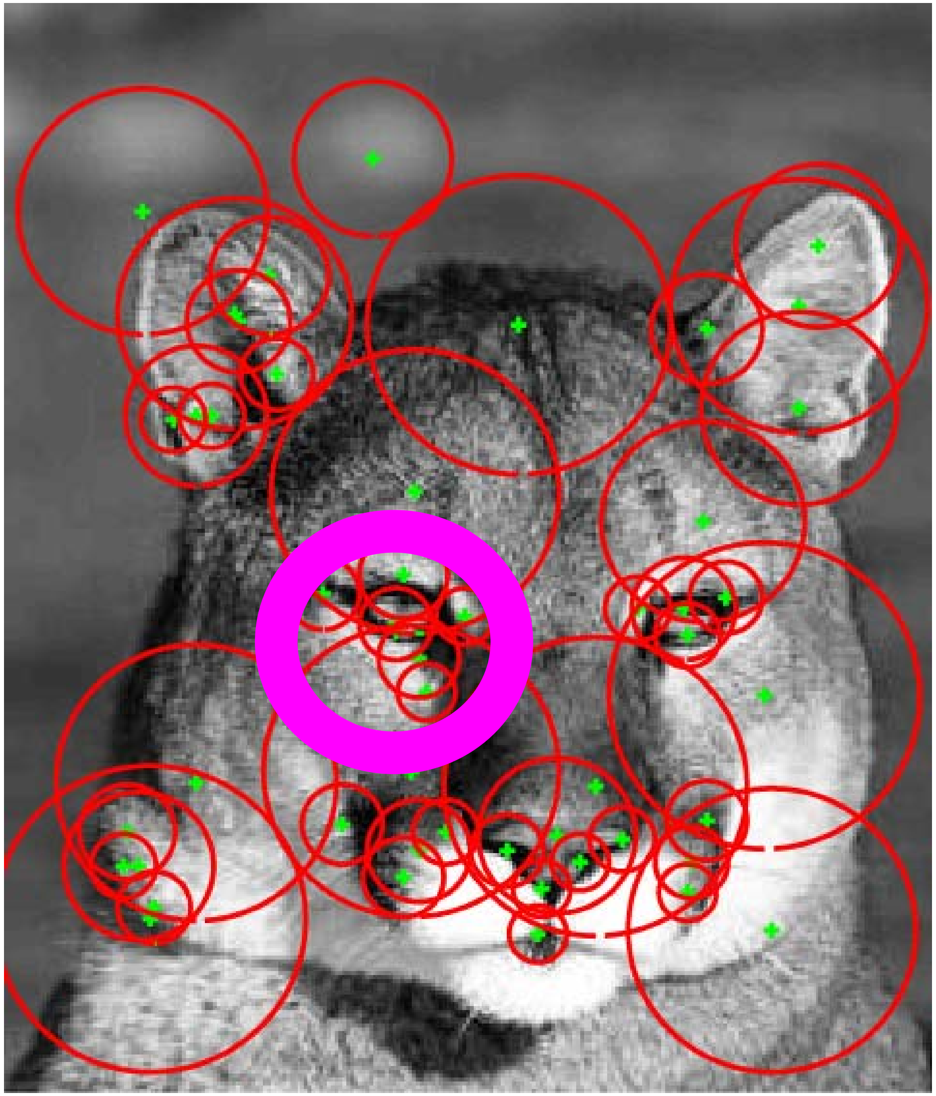
Fei-Fei et al. ‘03, ‘04, ‘06

model representation



One-shot learning of object categories

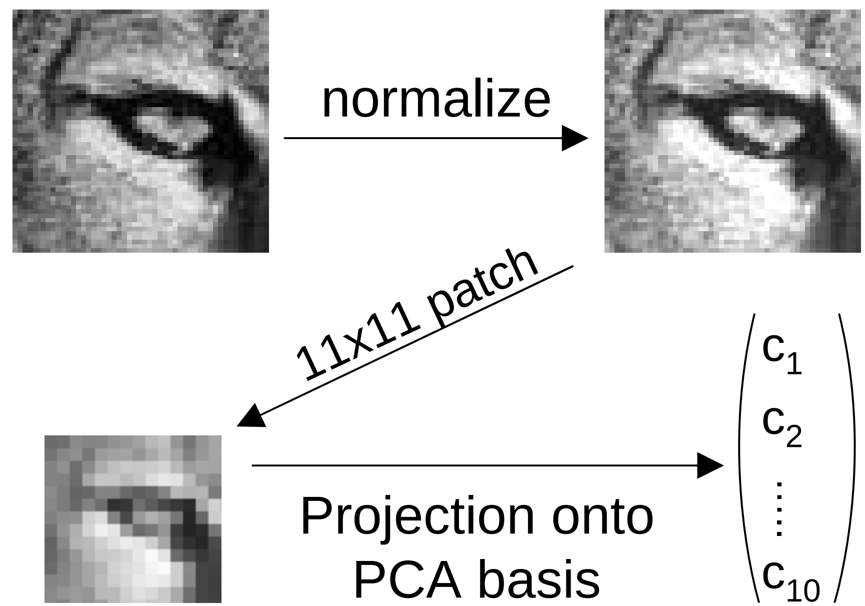
Fei-Fei et al. '03, '04, '06



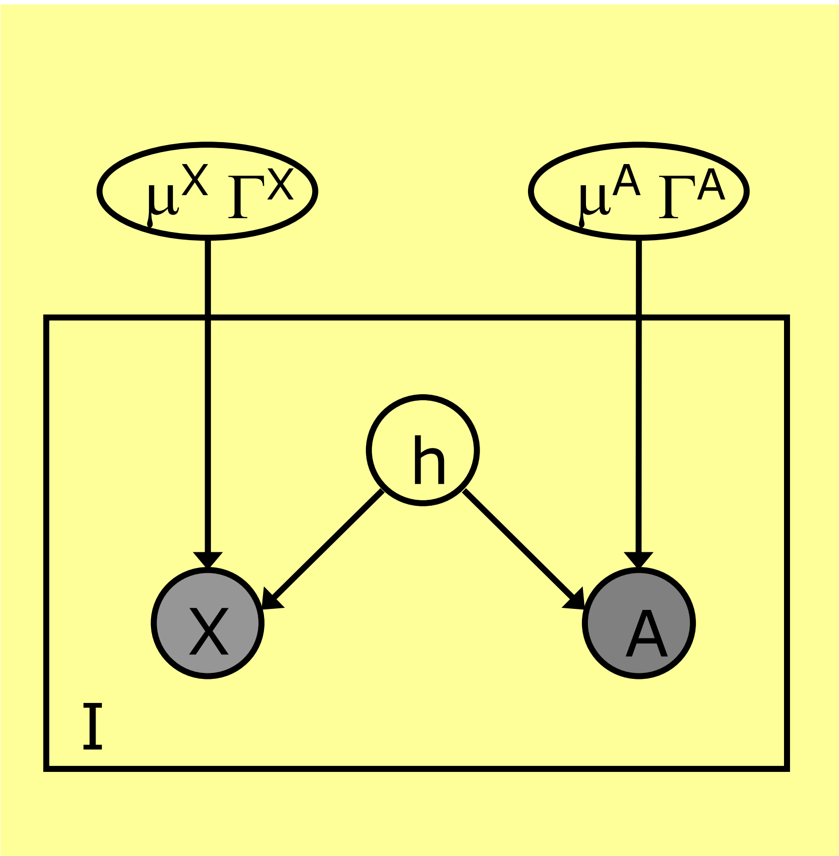
X (location)

(x,y) coords. of region center

A (appearance)



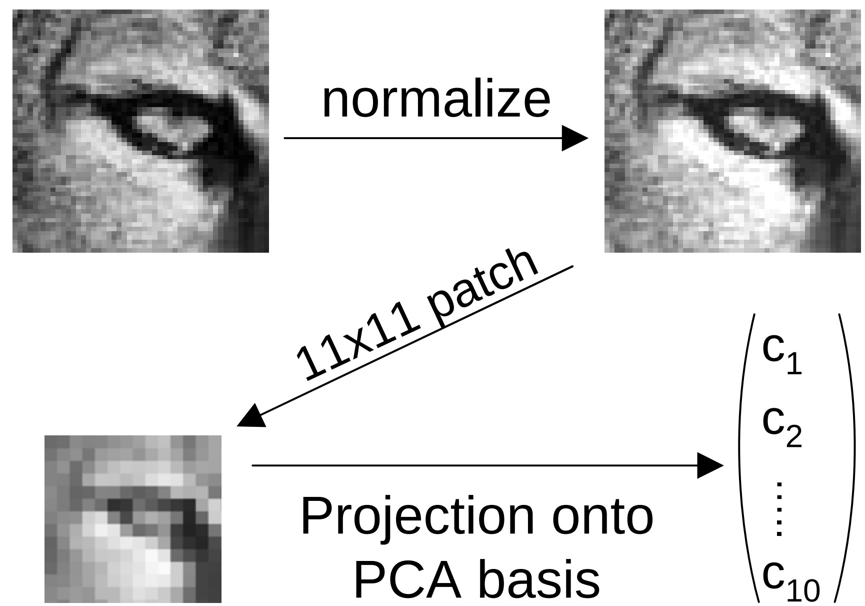
The Generative Model



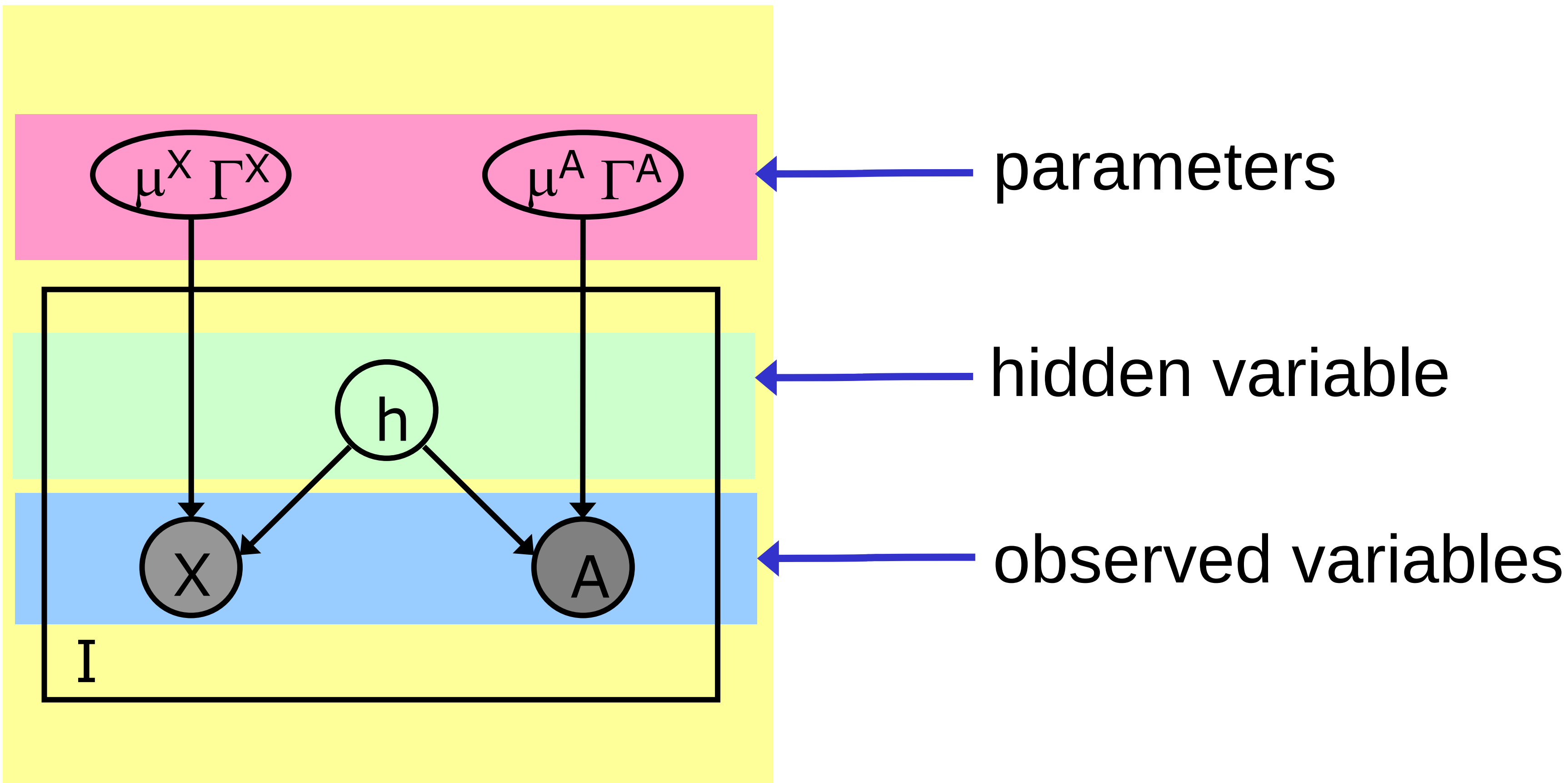
X (location)

(x,y) coords. of region center

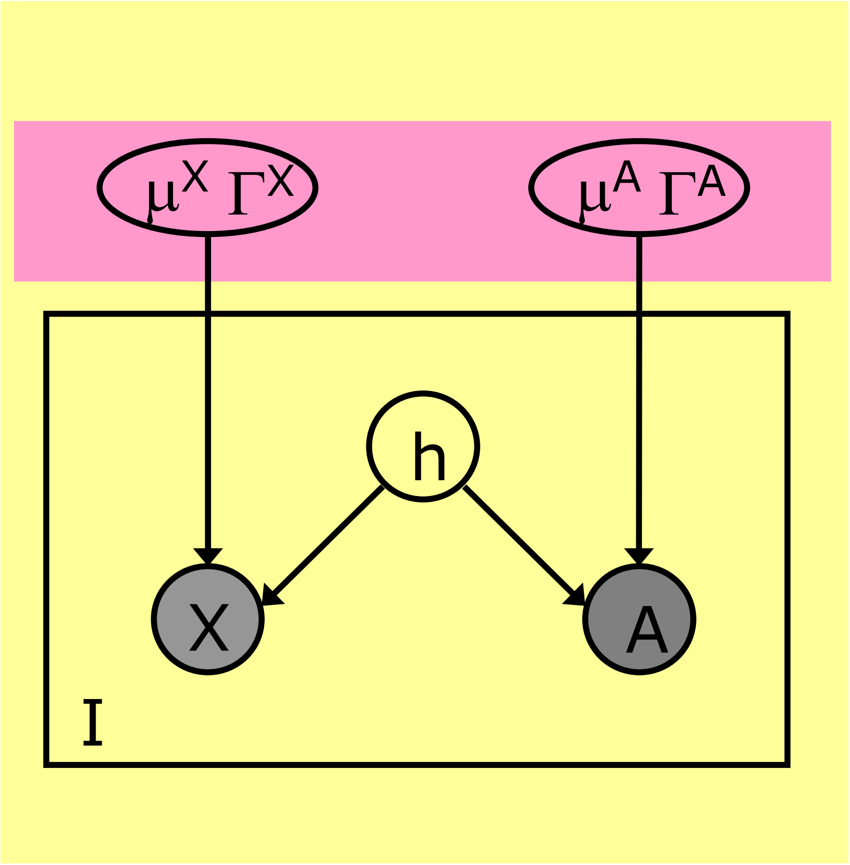
A (appearance)

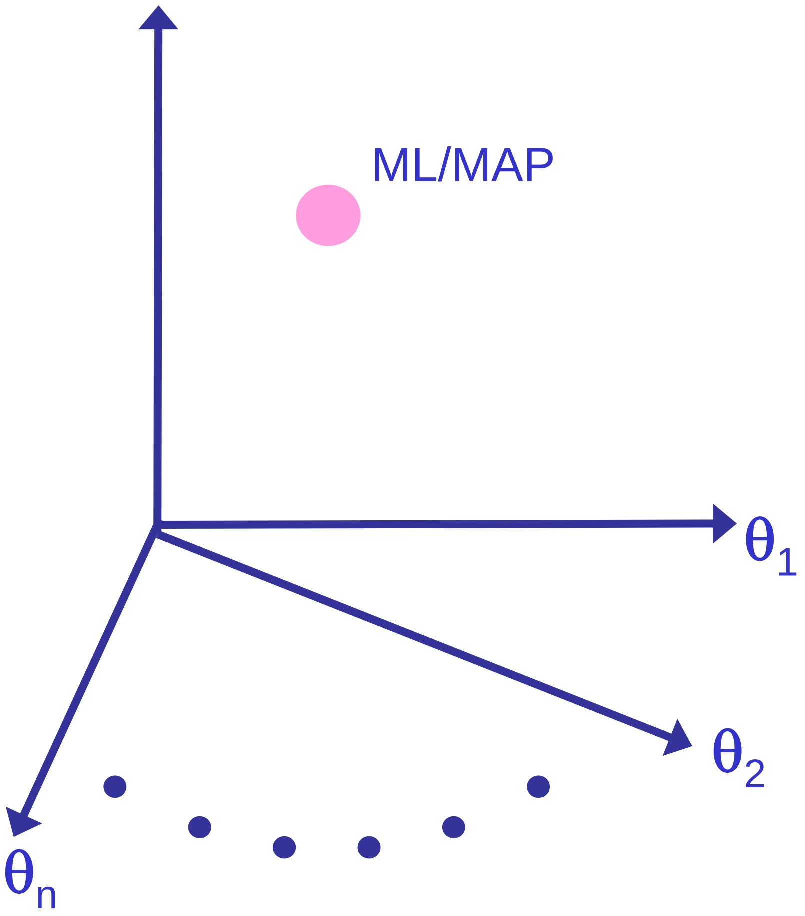


The Generative Model



The Generative Model

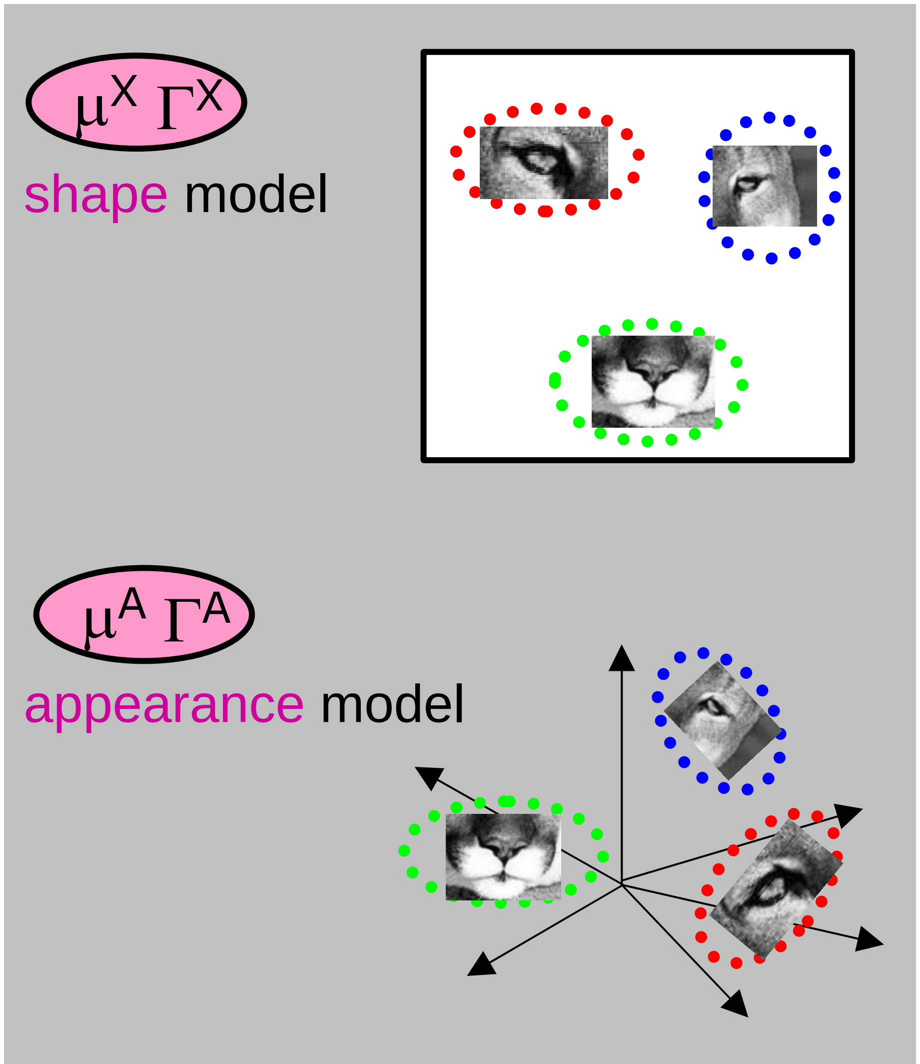


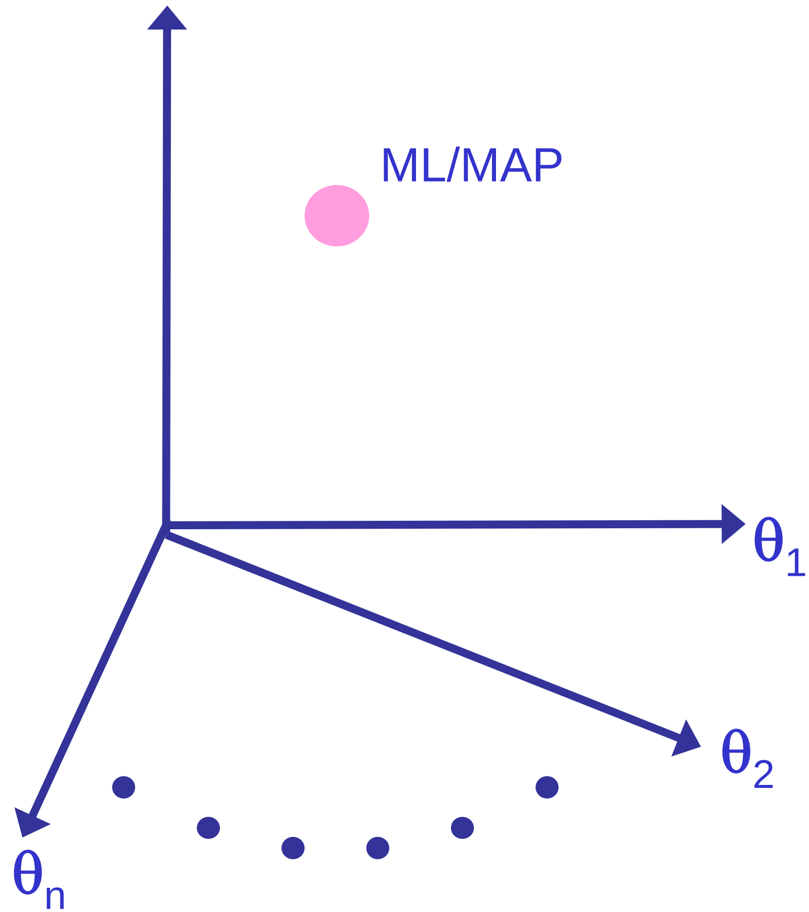


where θ = {μ^X, Γ^X, μ^A, Γ^A}

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

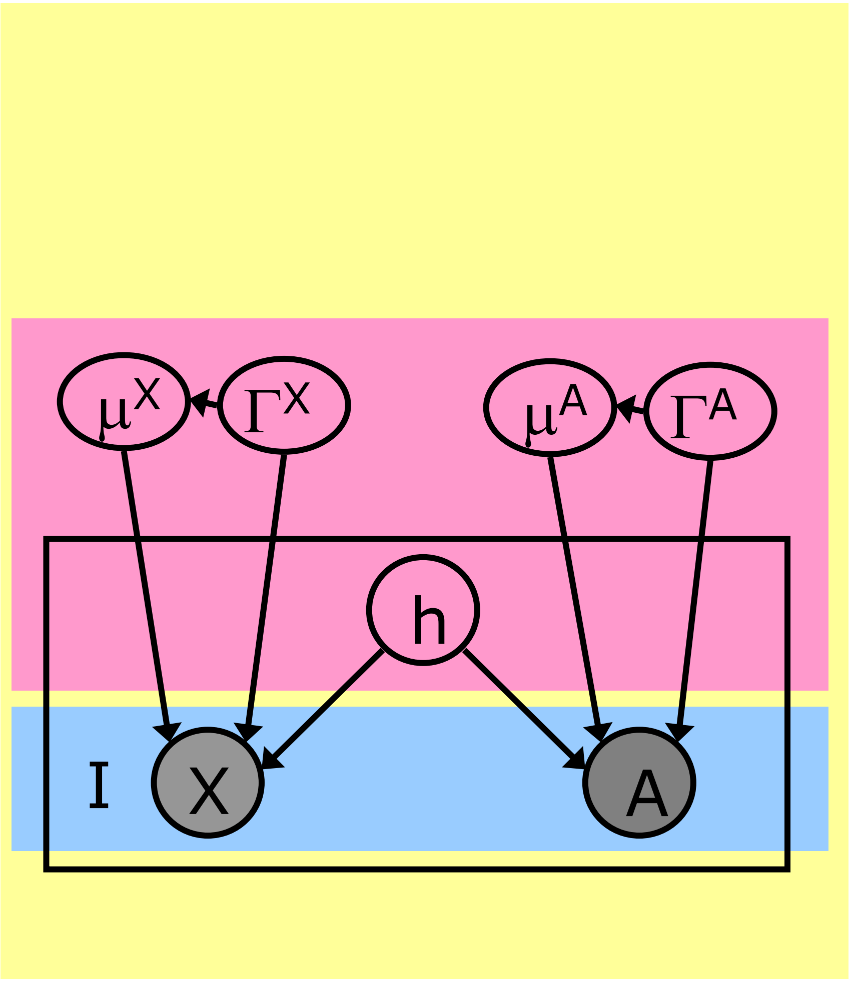
The Generative Model

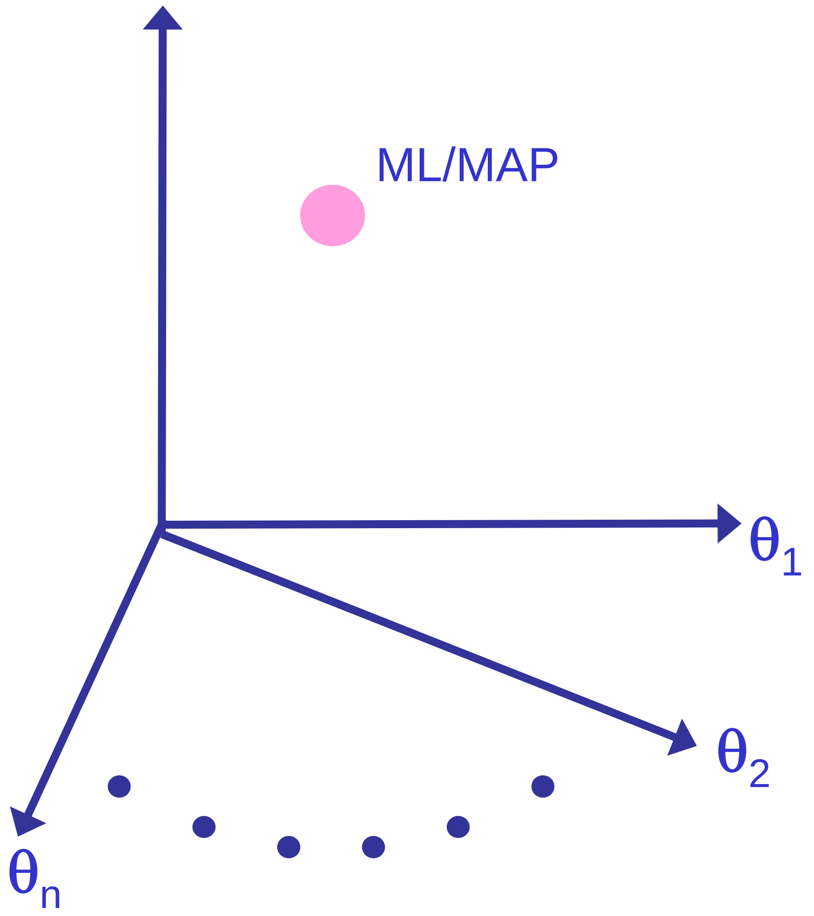




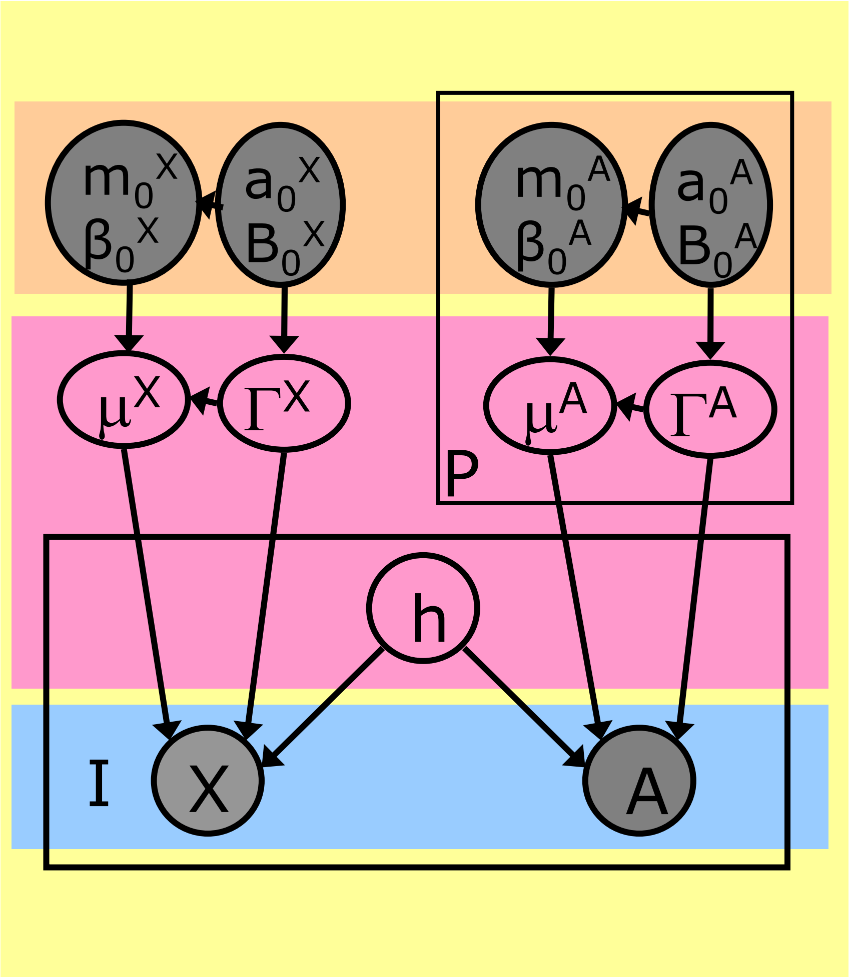
where $\theta = \{\mu^X, \Gamma^X, \mu^A, \Gamma^A\}$

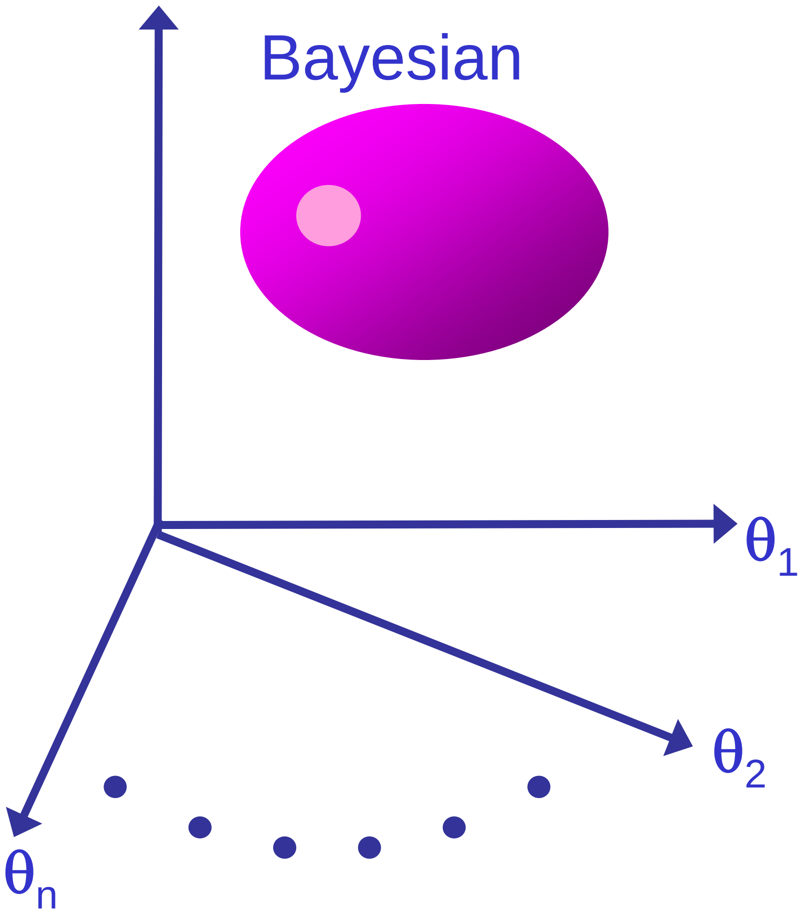
The Generative Model





The Generative Model

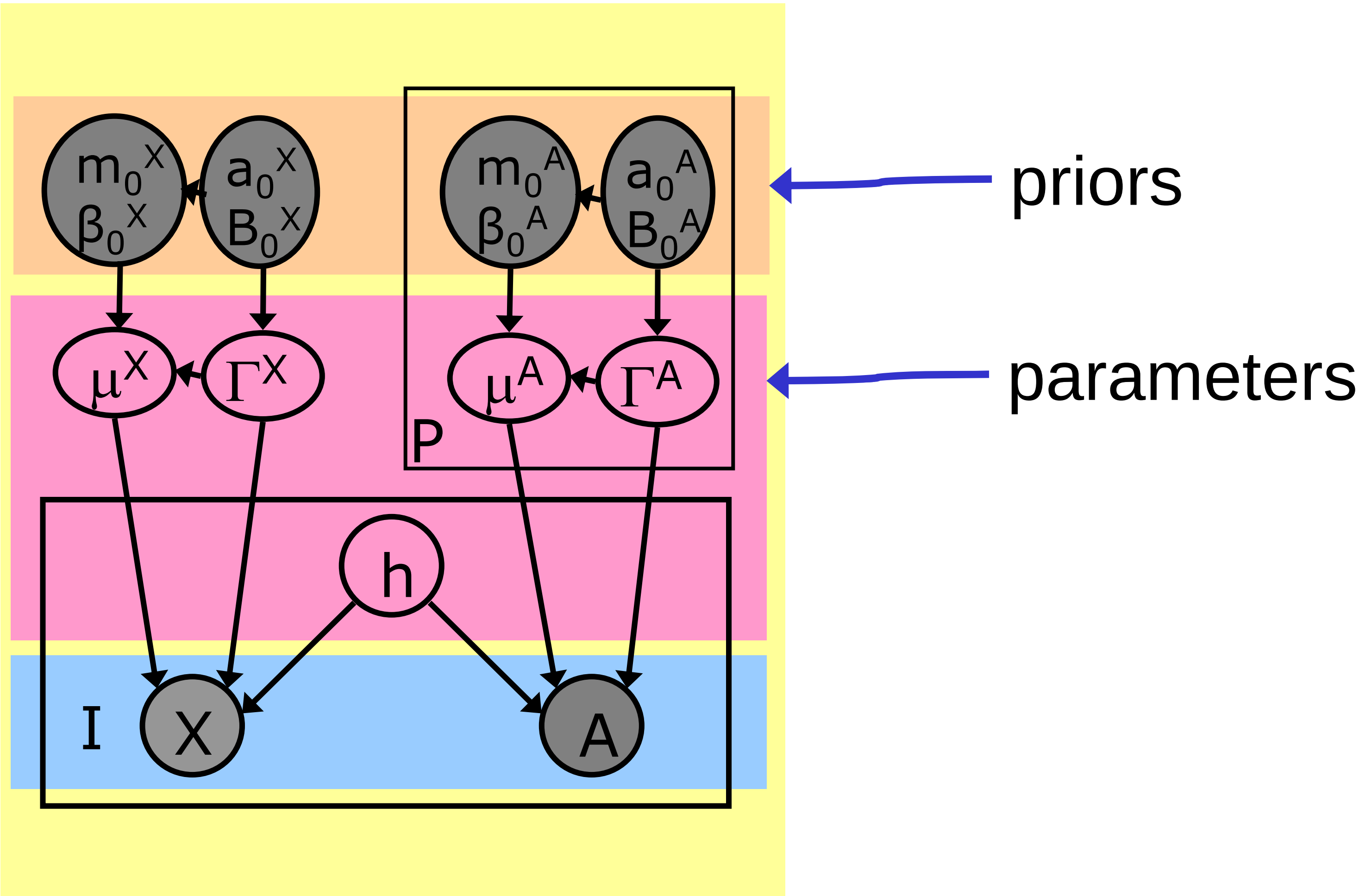




Fei-Fei et al. '03, '04, '06

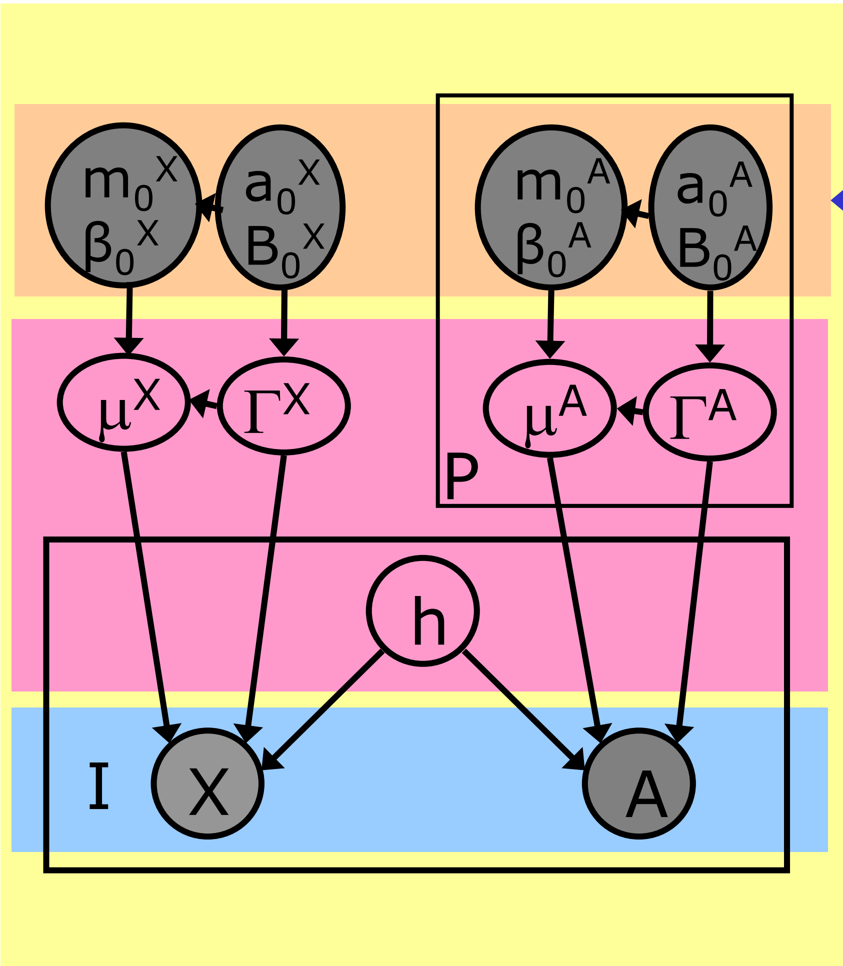
Parameters to estimate: $\{m^X, \beta^X, a^X, B^X, m^A, \beta^A, a^A, B^A\}$
i.e. parameters of Normal-Wishart distribution

The Generative Model

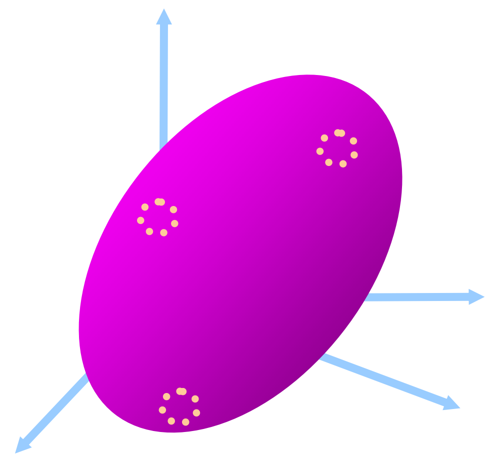


Fei-Fei et al. '03, '04, '06

The Generative Model



priors



Prior distribution

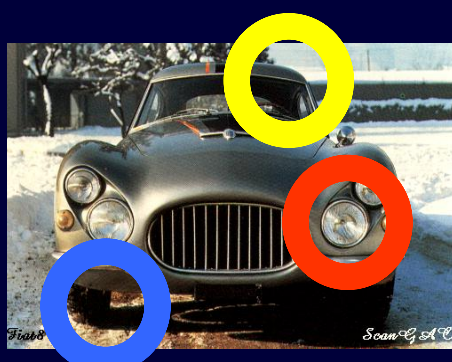
Fei-Fei et al. '03, '04, '06

1. human vision

3. learning
& inferences

One-shot learning of object categories

2. model representation



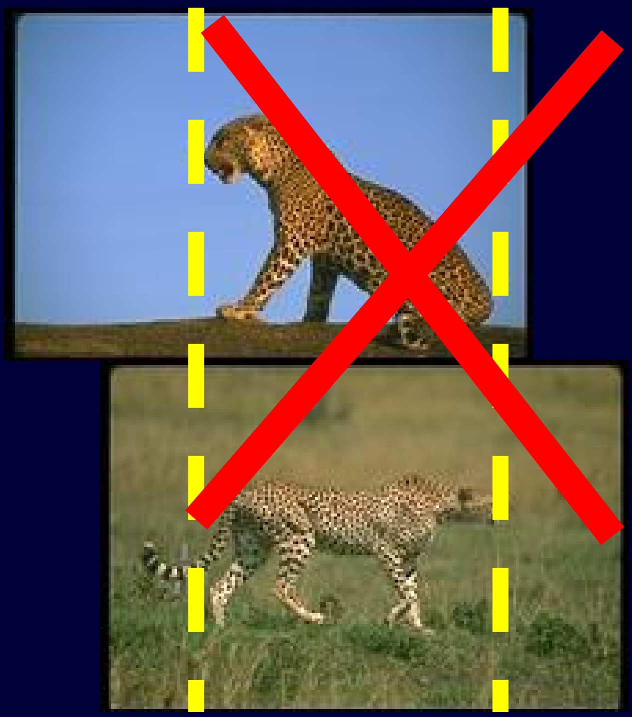
4. evaluation
& dataset
& application

learning & inferences

No labeling

No segmentation

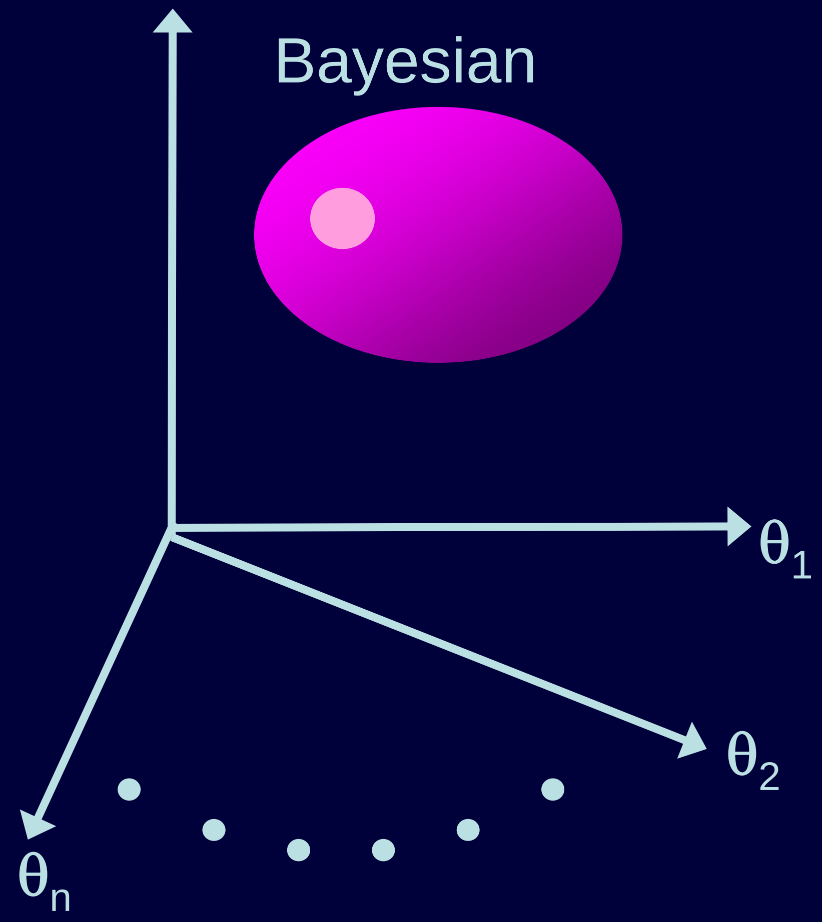
No alignment



One-shot learning of object categories

Fei-Fei et al. 2003, 2004, 2006

learning & inferences

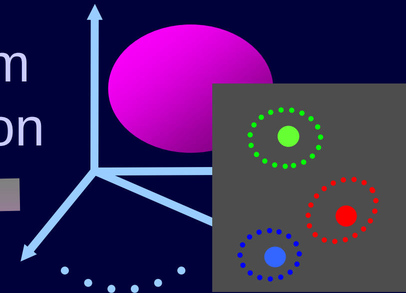


One-shot learning of object categories

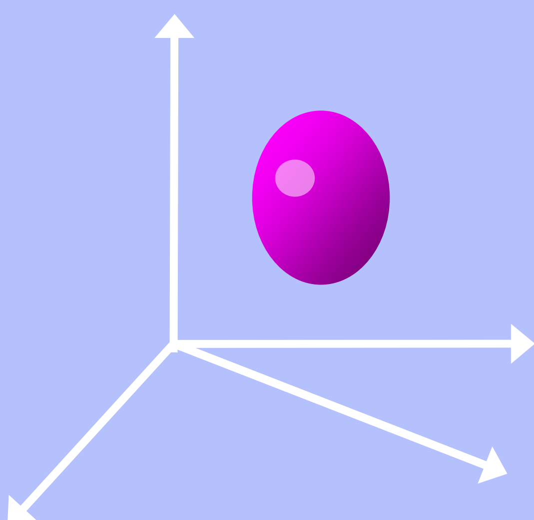
Fei-Fei et al. 2003, 2004, 2006

Variational EM

Random initialization

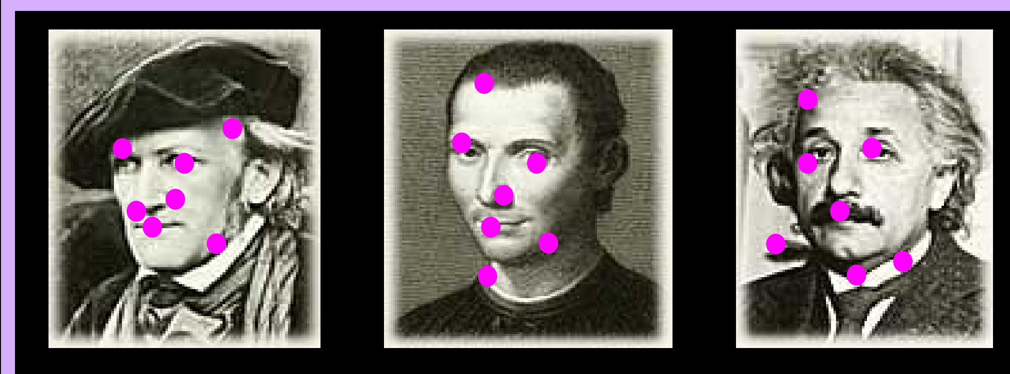


M-Step



new estimate of p(θ|train)

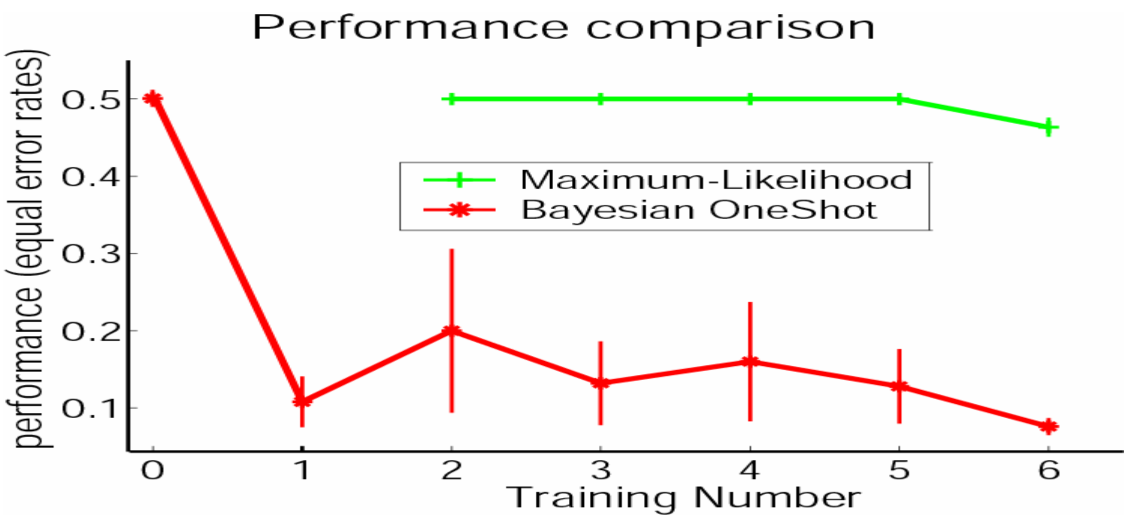
E-Step

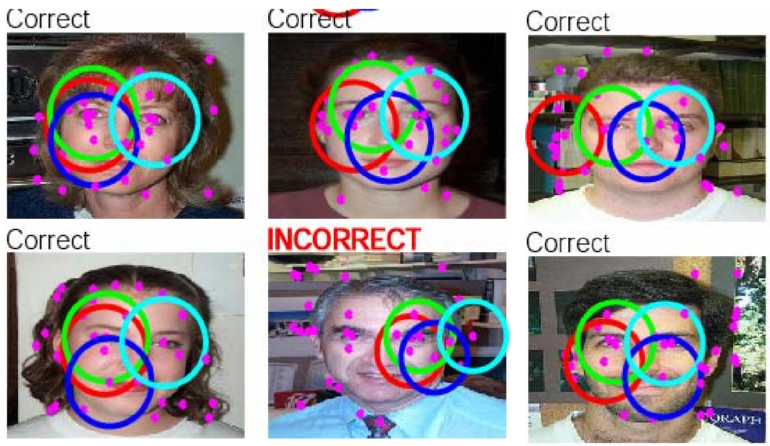


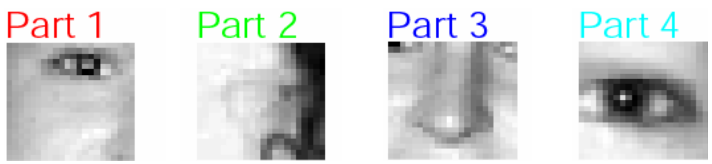
prior knowledge of p(θ)

Attias, Jordan, Hinton etc.

evaluation & dataset







Fei-Fei et al. 2004, 2006a, 2006b

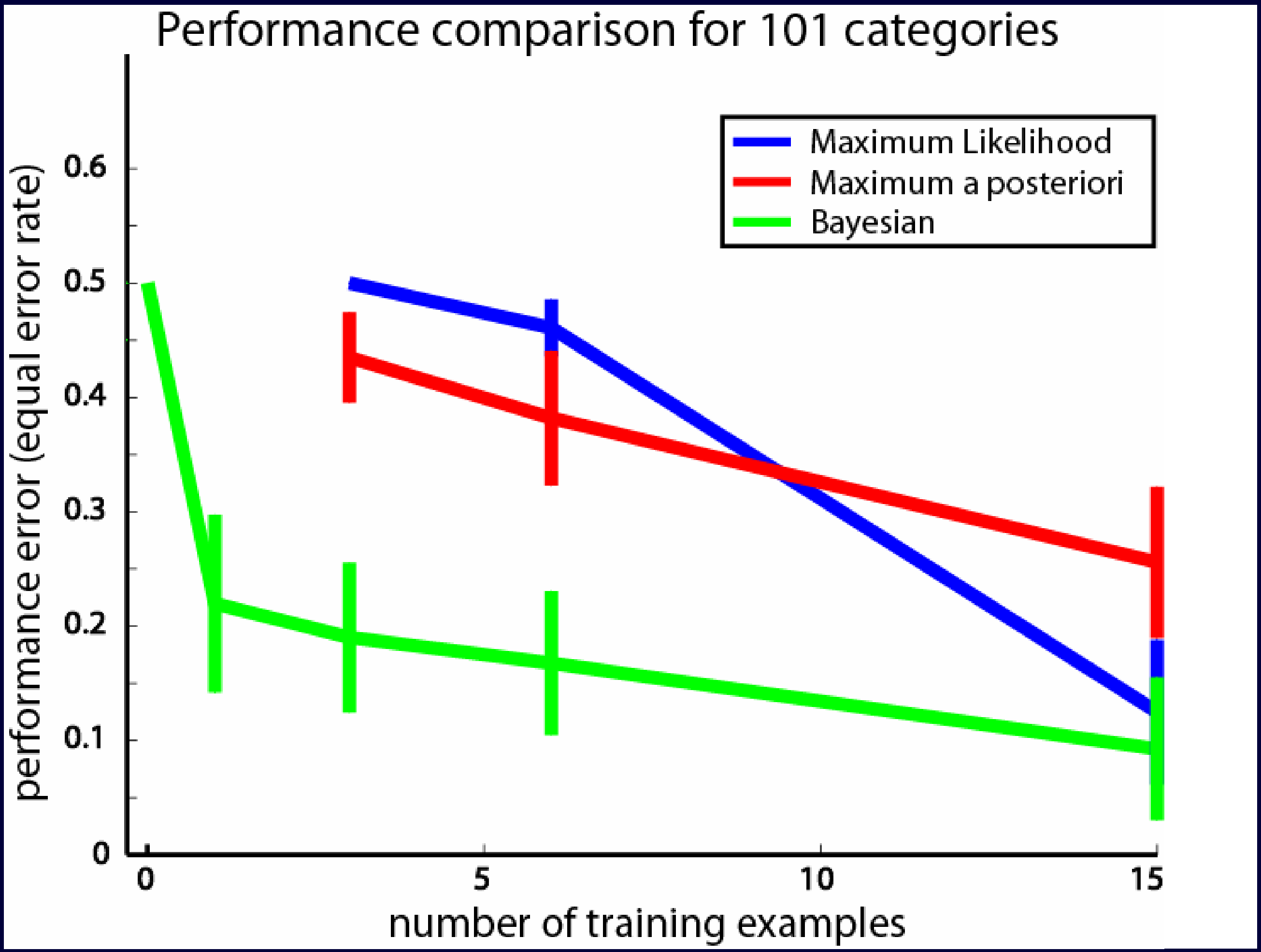
One-shot learning of object categories

evaluation & dataset -- Caltech 101 Dataset

One-shot learning of object categories

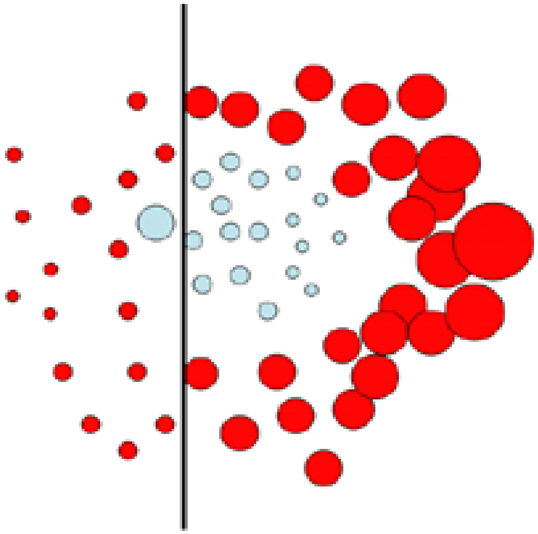
Fei-Fei et al. 2004, 2006a, 2006b

evaluation & dataset -- Caltech 101 Dataset



One-shot learning of object categories

Fei-Fei et al. 2004, 2006a, 2006b



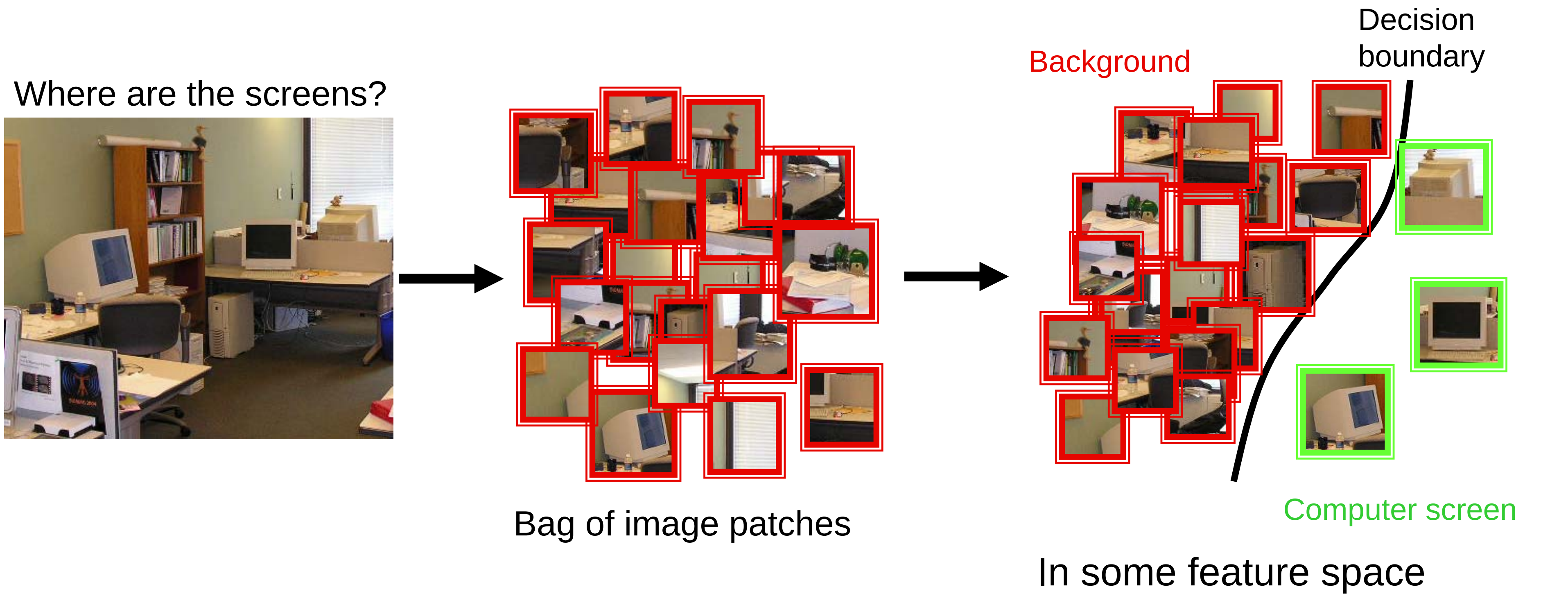
Part 3: discriminative methods

Discriminative methods

Object detection and recognition is formulated as a classification problem.

The image is partitioned into a set of overlapping windows

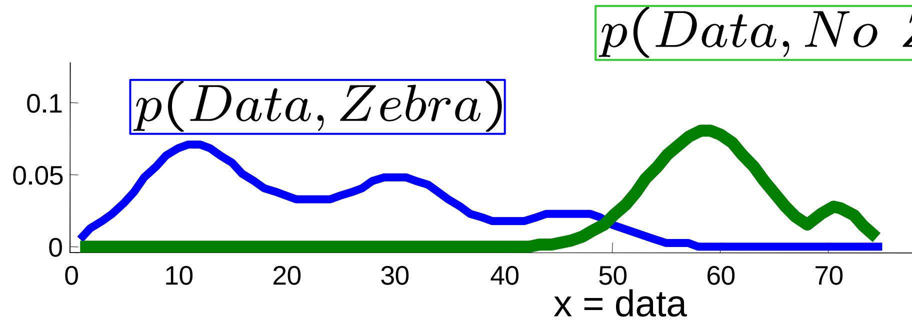
… and a decision is taken at each window about if it contains a target object or not.



Discriminative vs. generative

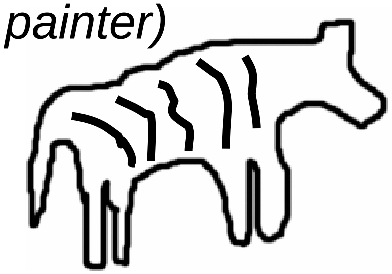
- Generative model

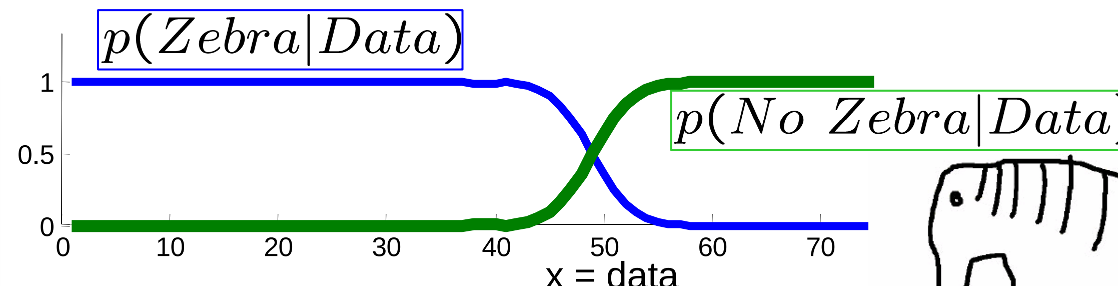
  (*The artist*)



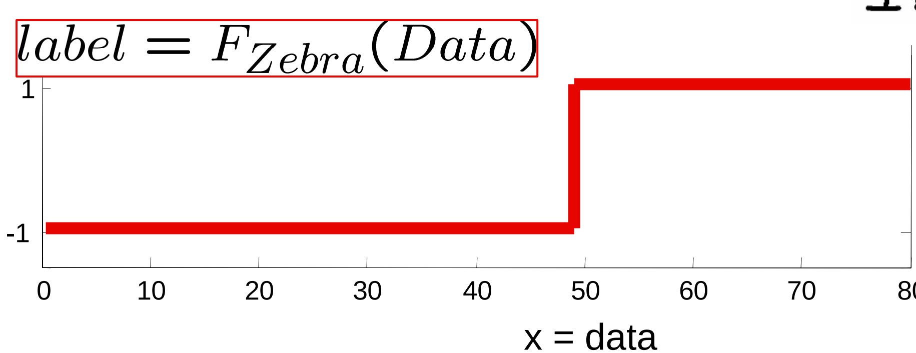
- Discriminative model

  (The lousy painter)





- Classification function



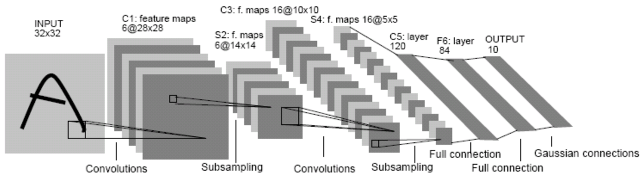
Discriminative methods

Nearest neighbor

10^6 examples

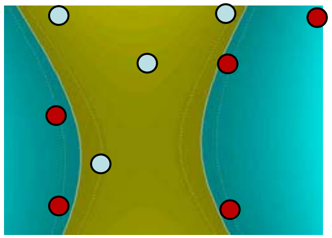
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005
…

Neural networks



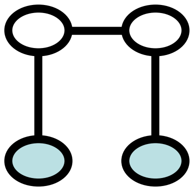
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001
…

Conditional Random Fields



McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003
…

Formulation

- Formulation: binary classification

Features	x =	x_1	x_2	x_3	$\cdots$	x_N	x_{N+1}	x_{N+2}	$\cdots$	x_{N+M}
Labels	y =	-1	+1	-1		-1	?	?		?

Training data: each image patch is labeled as containing the object or background

Test data

- Classification function

$\hat{y} = F(x)$ Where $F(x)$ belongs to some family of functions

- Minimize misclassification error

(Not that simple: we need some guarantees that there will be generalization)

Overview of section

- Object detection with classifiers

- **Boosting**
  - Gentle boosting
  - Weak detectors
  - Object model
  - Object detection

- Multiclass object detection

Why boosting?

- A simple algorithm for learning robust classifiers
  - Freund & Shapire, 1995
  - Friedman, Hastie, Tibshhirani, 1998

- Provides efficient algorithm for sparse visual feature selection
  - *Tieu & Viola, 2000*
  - *Viola & Jones, 2003*

- Easy to implement, not requires external optimization tools.

Boosting

- Defines a classifier using an additive model:

$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \dots$$

Strong classifier

Weak classifier

Weight

Features vector

Boosting

- Defines a classifier using an additive model:

$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \dots$$

Strong classifier

Weak classifier

Weight

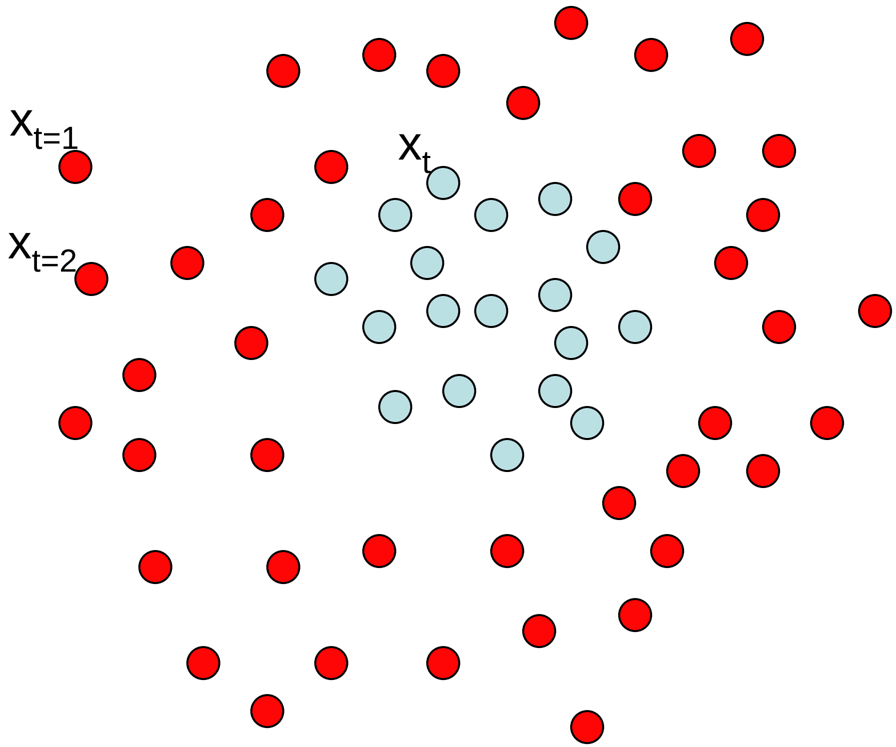
Features vector

- We need to define a family of weak classifiers

$f_k(x)$ from a family of weak classifiers

Boosting

- It is a sequential procedure:



Each data point has

a class label:

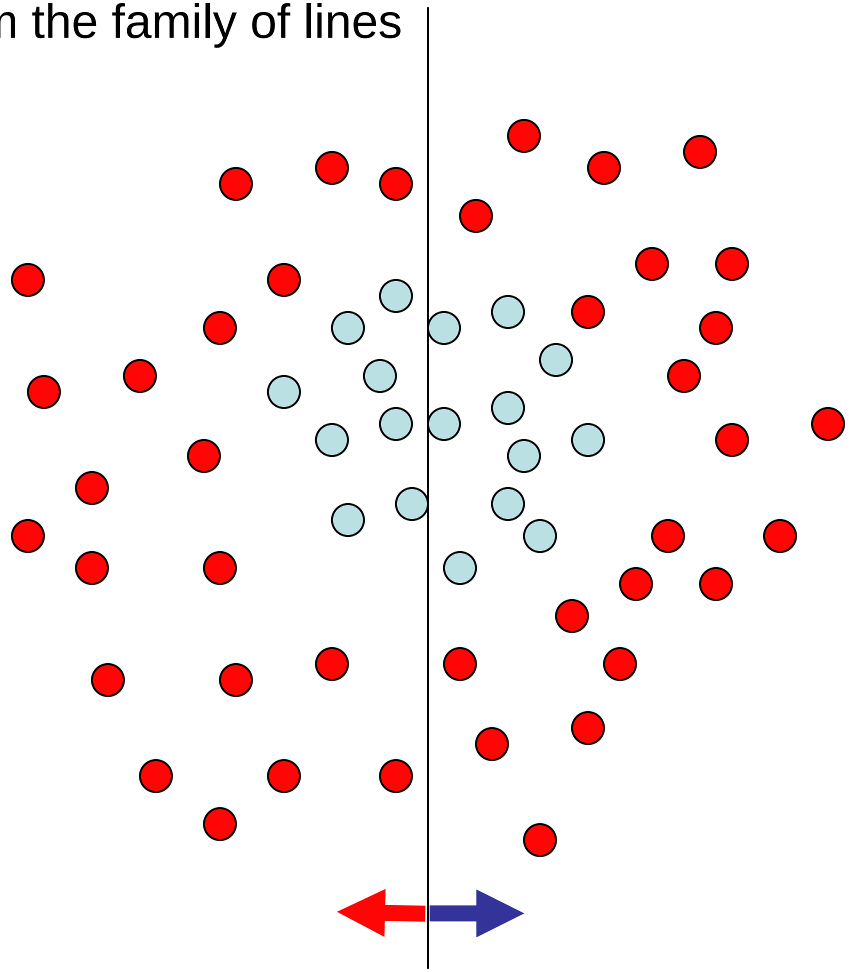
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

and a weight:

$$w_t = 1$$

Toy example

Weak learners from the family of lines



Each data point has

a class label:

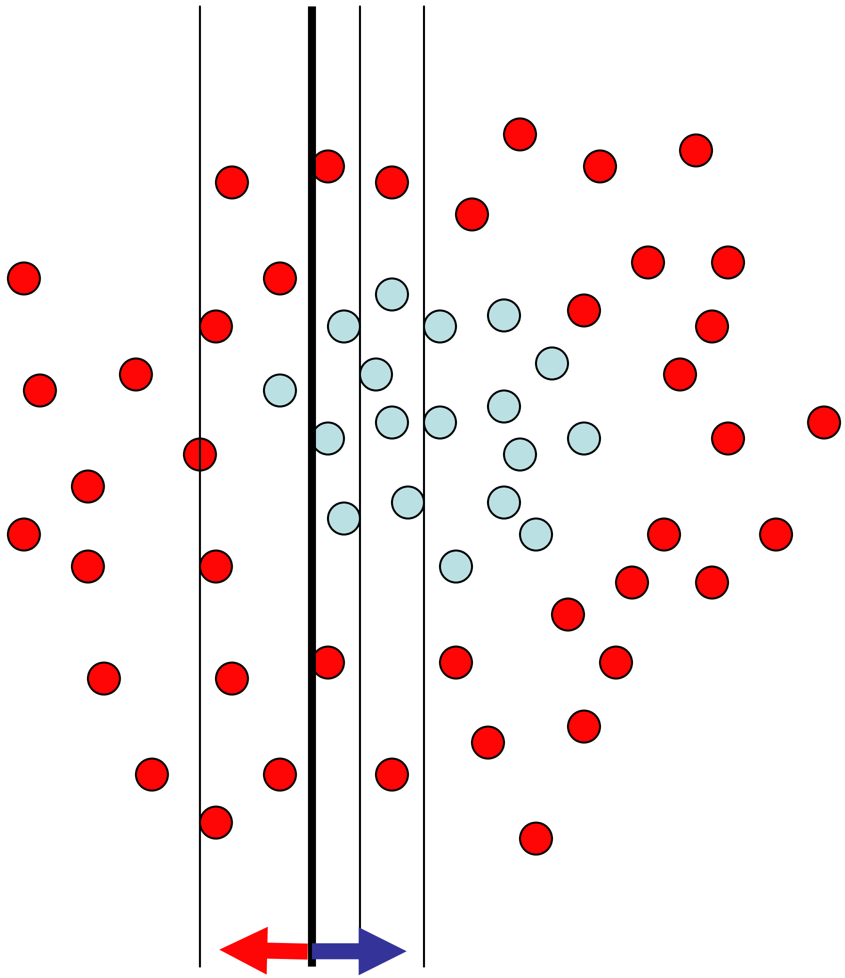
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

and a weight:

$w_t = 1$

h => p(error) = 0.5 it is at chance

Toy example



Each data point has

a class label:

$$y_t = \begin{cases} +1 \ (\bullet) \\ -1 \ (\circ) \end{cases}$$

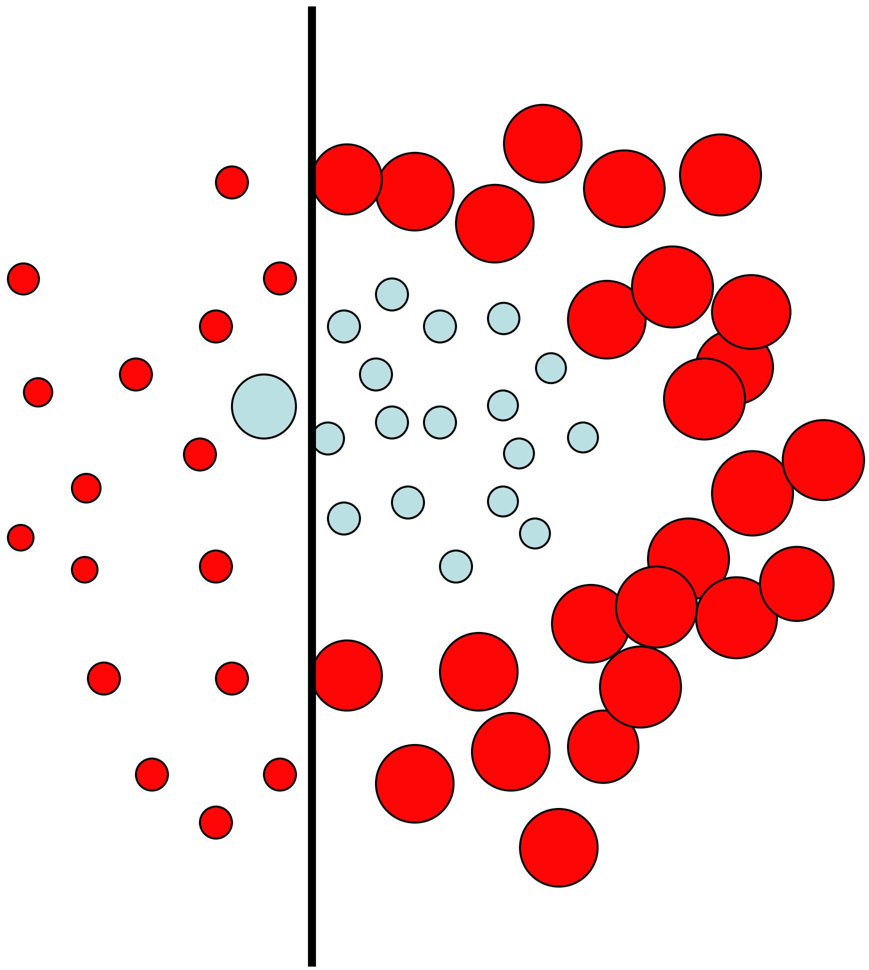
and a weight:

$w_t = 1$

This one seems to be the best

This is a ‘**weak classifier**’: It performs slightly better than chance.

Toy example



Each data point has

a class label:

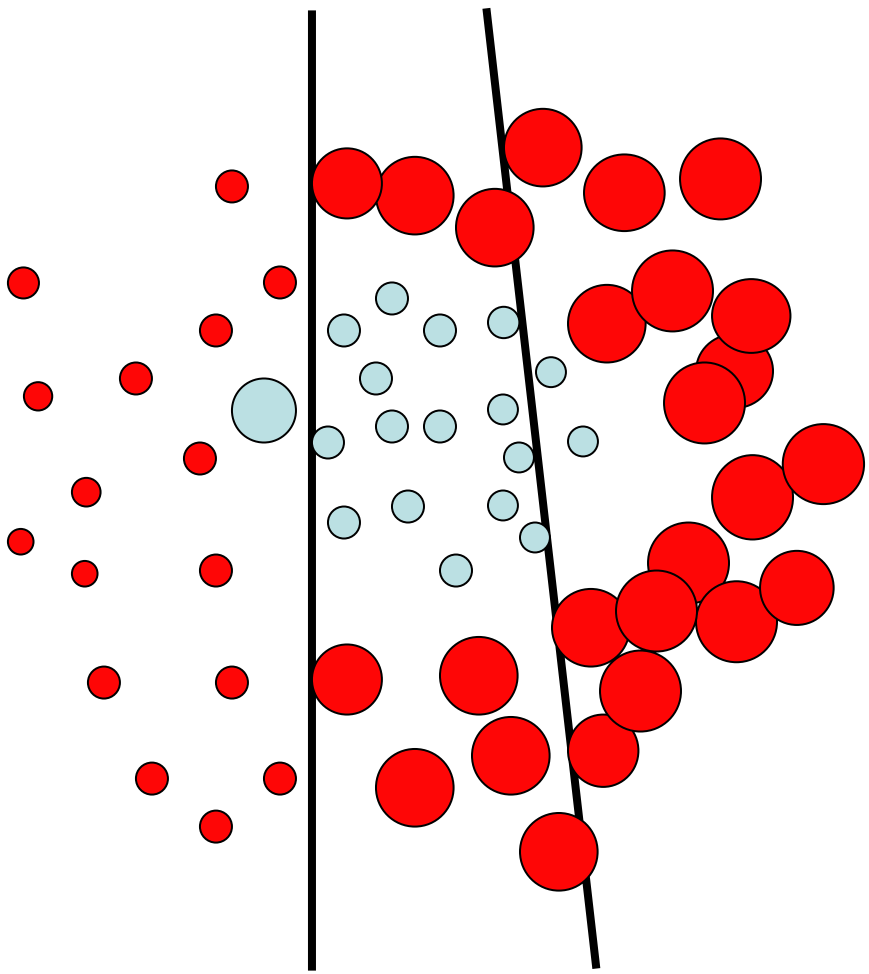
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

We set a new problem for which the previous weak classifier performs at chance again

Toy example



Each data point has

a class label:

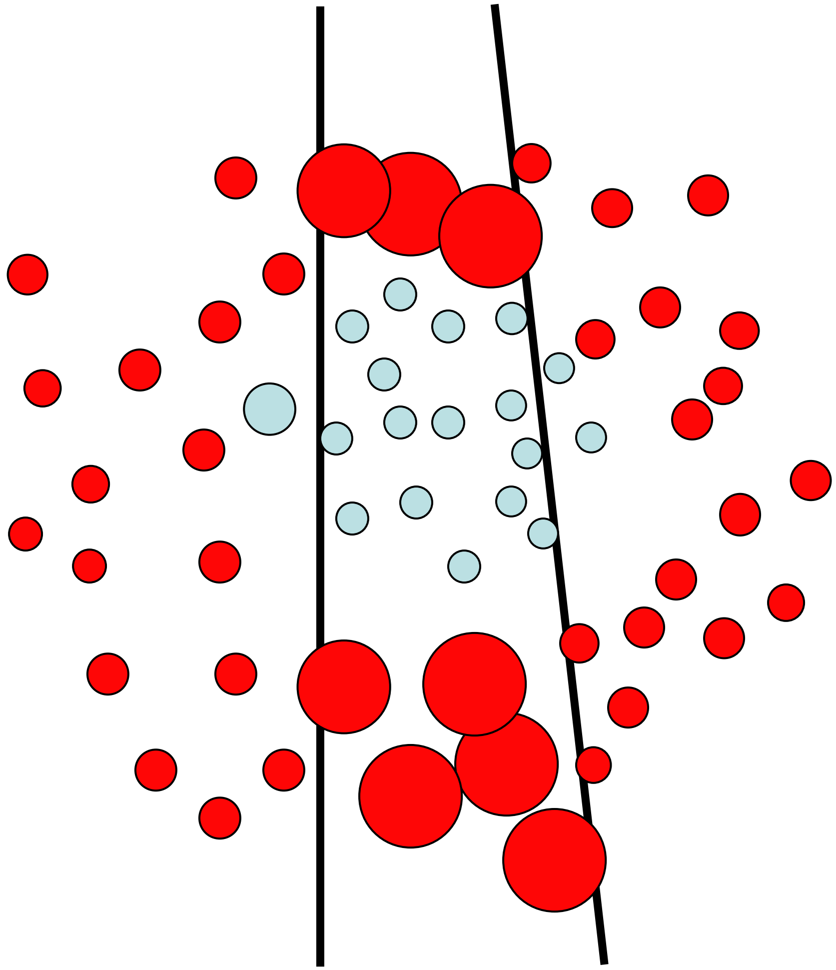
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

We set a new problem for which the previous weak classifier performs at chance again

Toy example



Each data point has

a class label:

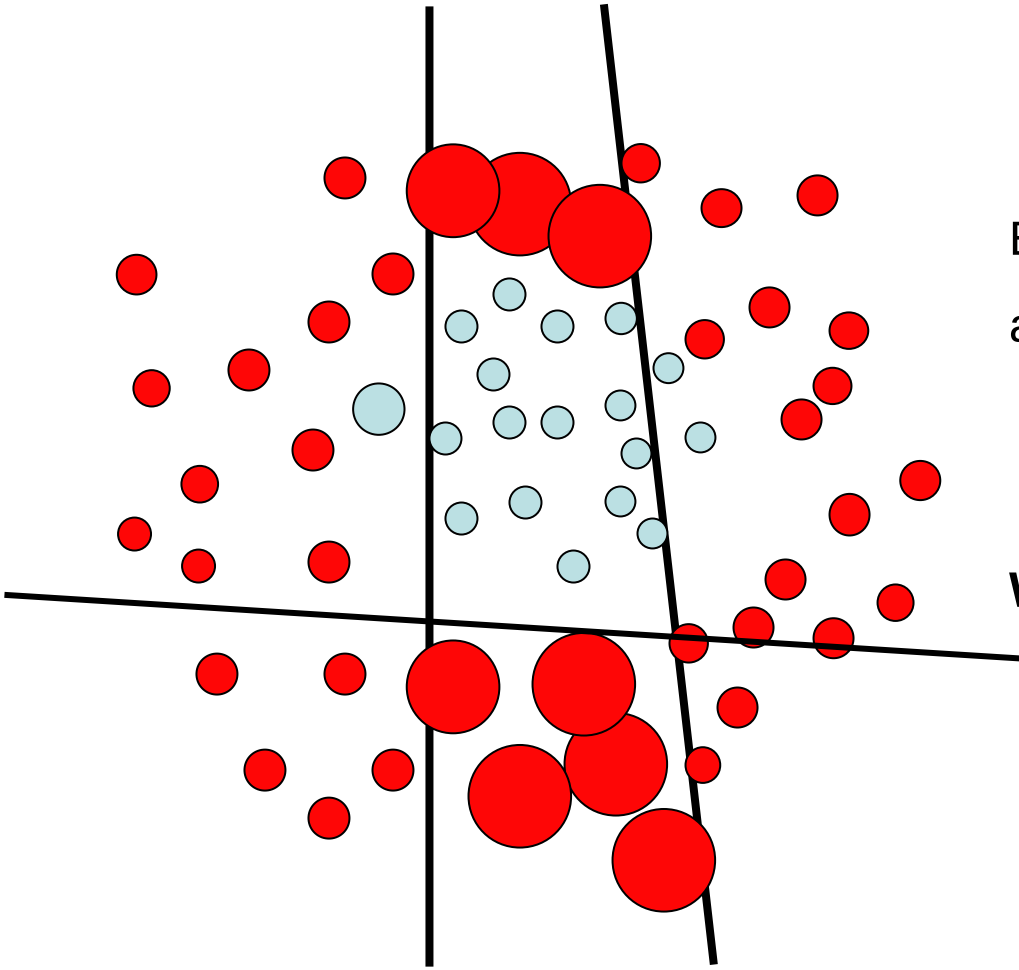
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

We set a new problem for which the previous weak classifier performs at chance again

Toy example



Each data point has

a class label:

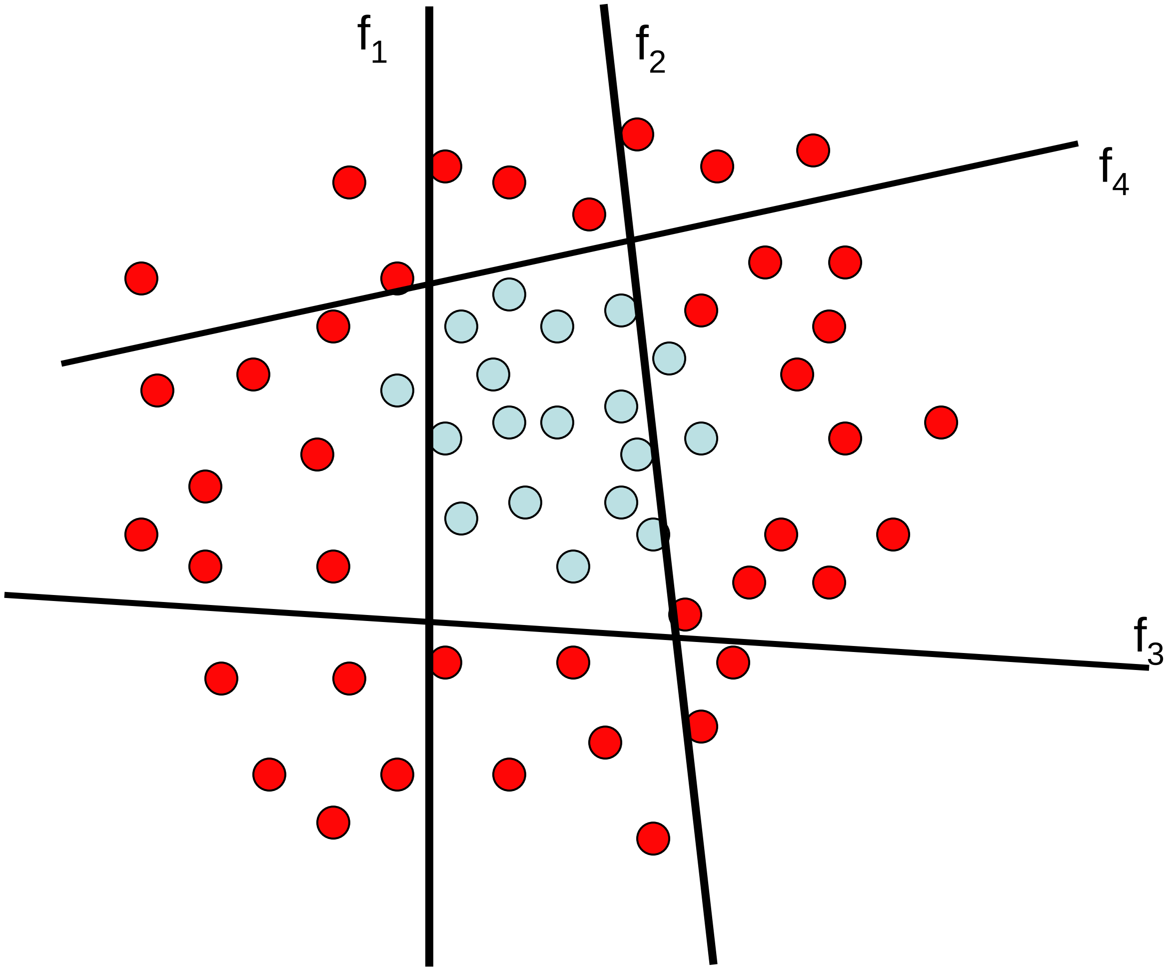
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

We set a new problem for which the previous weak classifier performs at chance again

Toy example



The strong (non- linear) classifier is built as the combination of all the weak (linear) classifiers.

From images to features: Weak detectors

We will now define a family of visual features that can be used as weak classifiers ("weak detectors")

$$\longrightarrow h_i(I, x, y) \longrightarrow$$

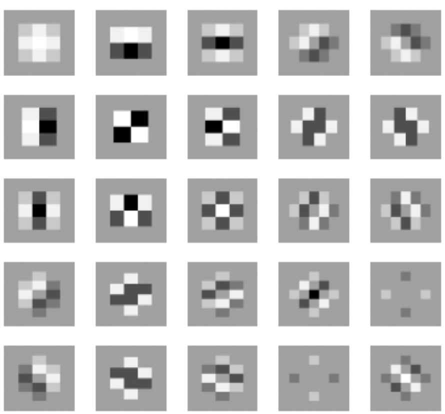
Takes image as input and the output is binary response.
The output is a weak detector.

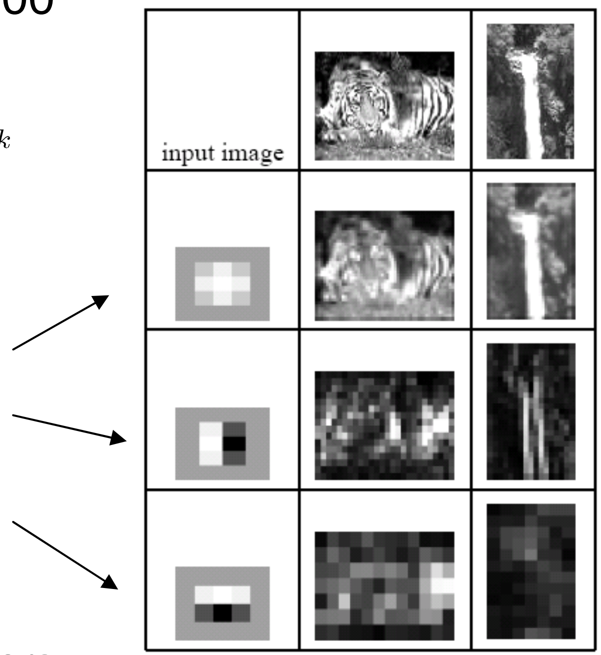
Weak detectors

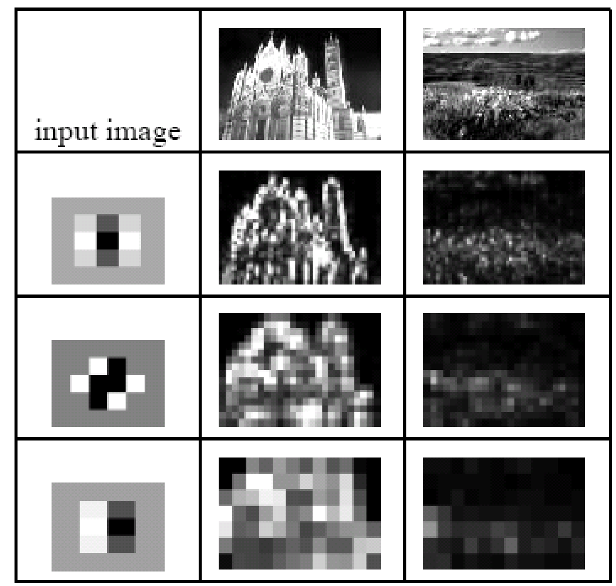
Textures of textures

Tieu and Viola, CVPR 2000

$$g_{i,j,k} = \sum_{pixels} ||I * f_i| \downarrow_2 * f_j| \downarrow_2 * f_k$$







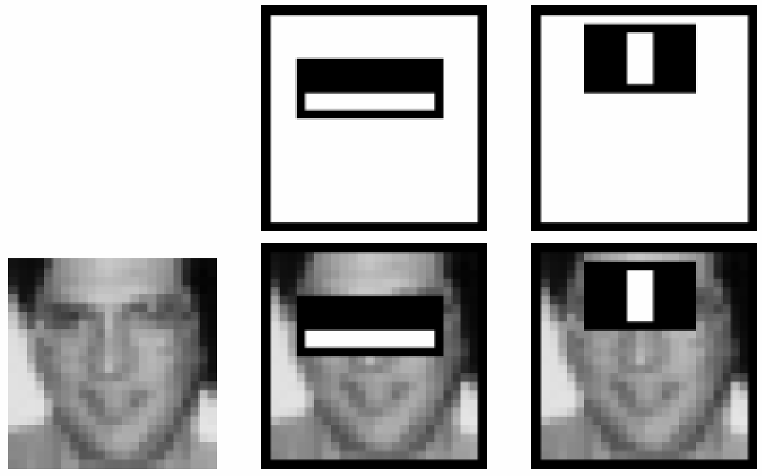
Every combination of three filters generates a different feature

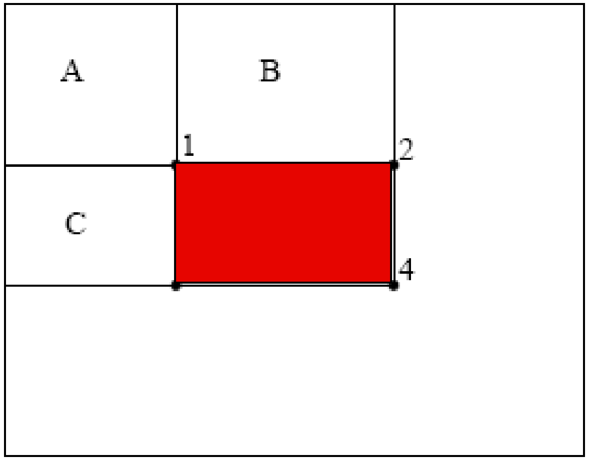
This gives thousands of features. Boosting selects a sparse subset, so computations on test time are very efficient. Boosting also avoids overfitting to some extend.

Weak detectors

Haar filters and integral image

Viola and Jones, ICCV 2001





The average intensity in the block is computed with four sums independently of the block size.

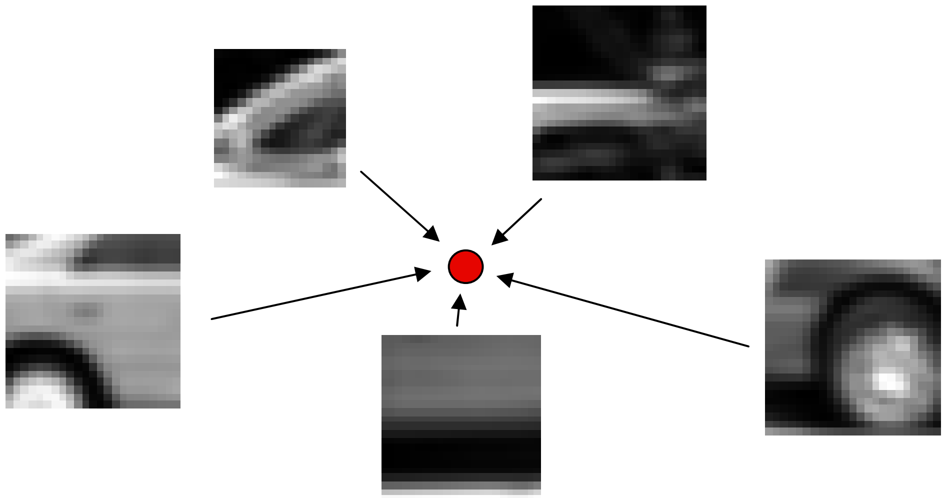
Weak detectors

Other weak detectors:

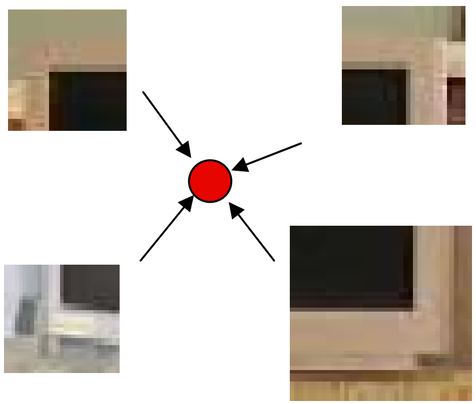
- Carmichael, Hebert 2004
- Yuille, Snow, Nitzbert, 1998
- Amit, Geman 1998
- Papageorgiou, Poggio, 2000
- Heisele, Serre, Poggio, 2001
- Agarwal, Awan, Roth, 2004
- Schneiderman, Kanade 2004
- …

Weak detectors

Part based: similar to part-based generative models. We create weak detectors by using parts and voting for the object center location



Car model



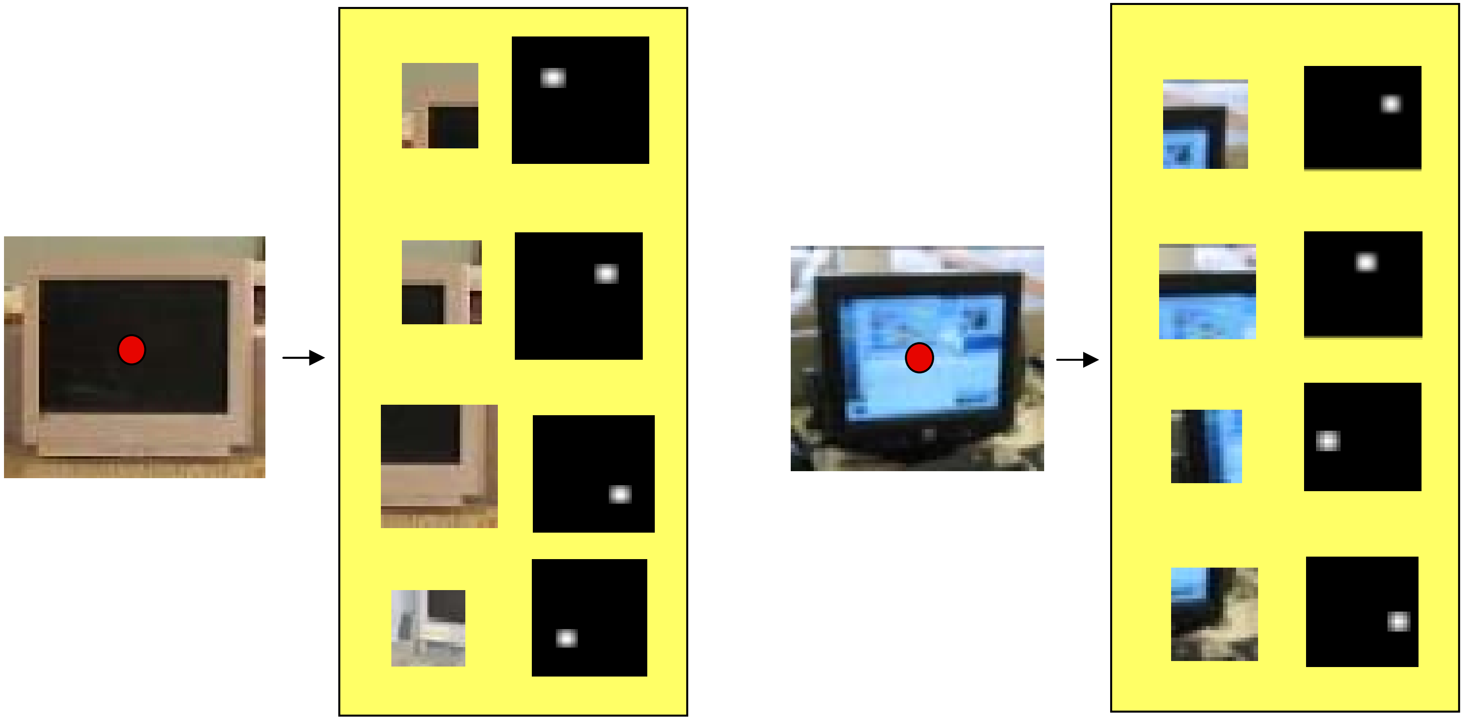
Screen model

These features are used for the detector on the course web site.

Weak detectors

First we collect a set of part templates from a set of training objects.

Vidal-Naquet, Ullman (2003)

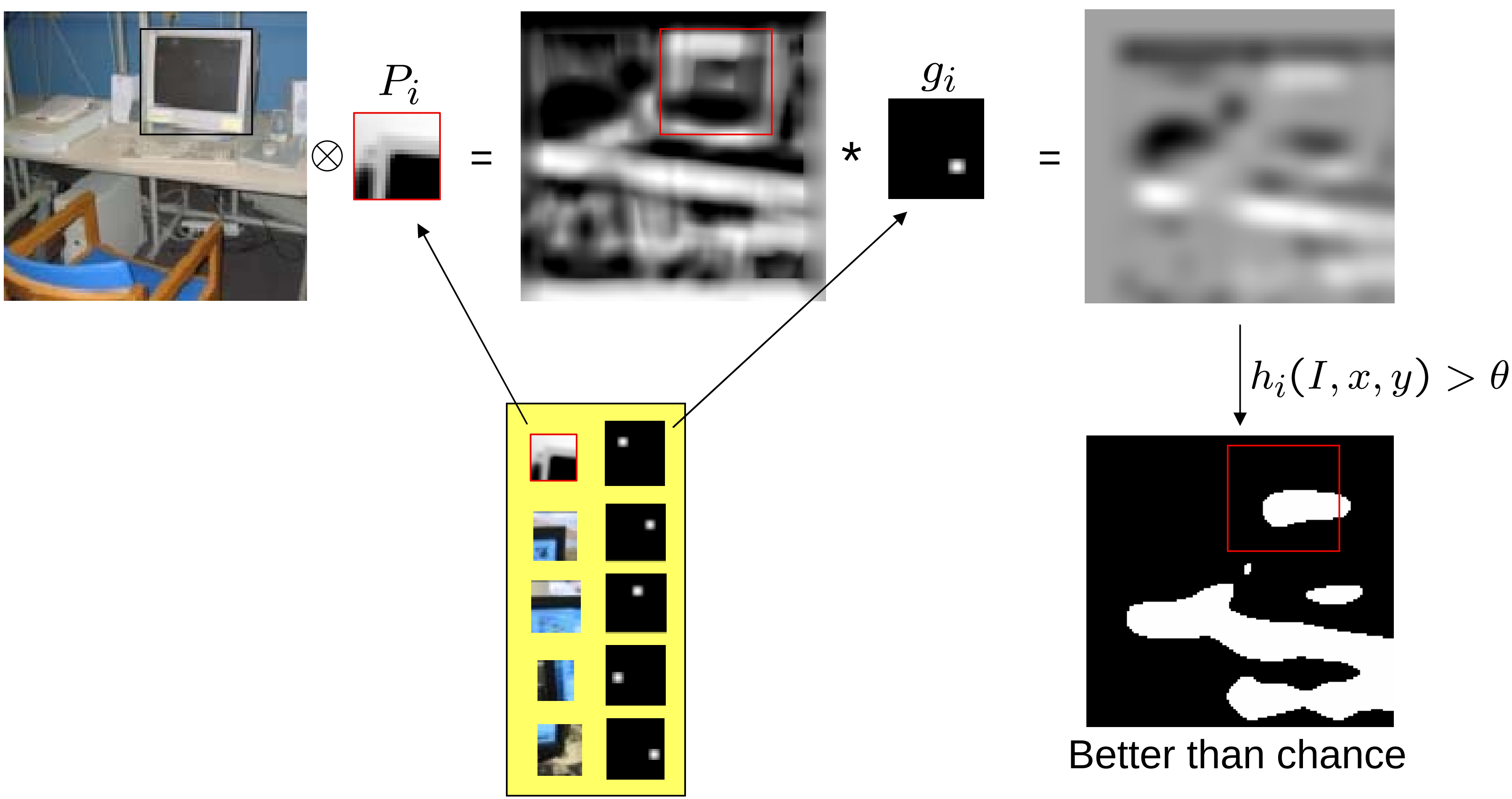


…

Weak detectors

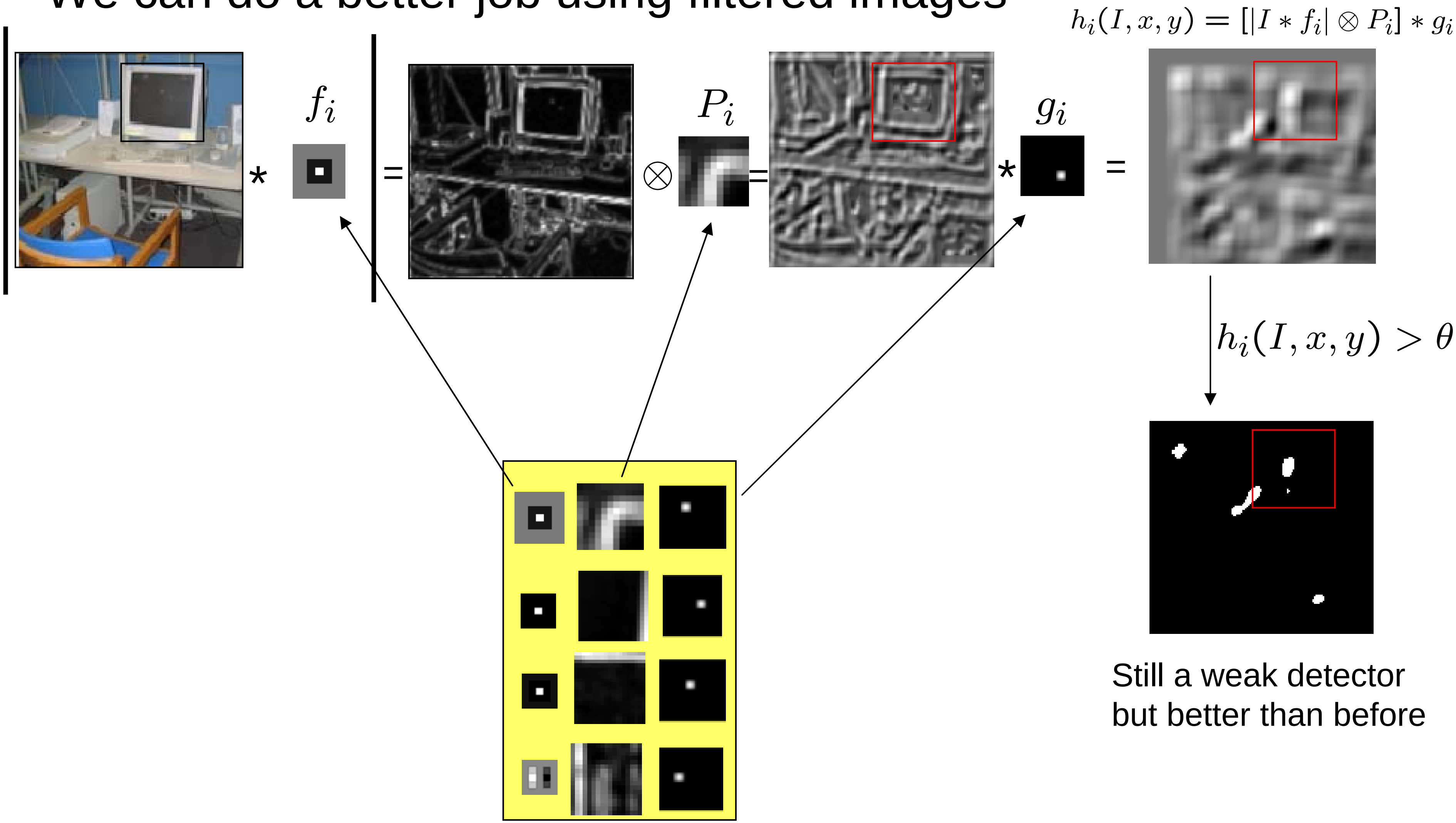
We now define a family of "weak detectors" as:

$$h_i(I, x, y) = [I \otimes P_i] * g_i$$



Weak detectors

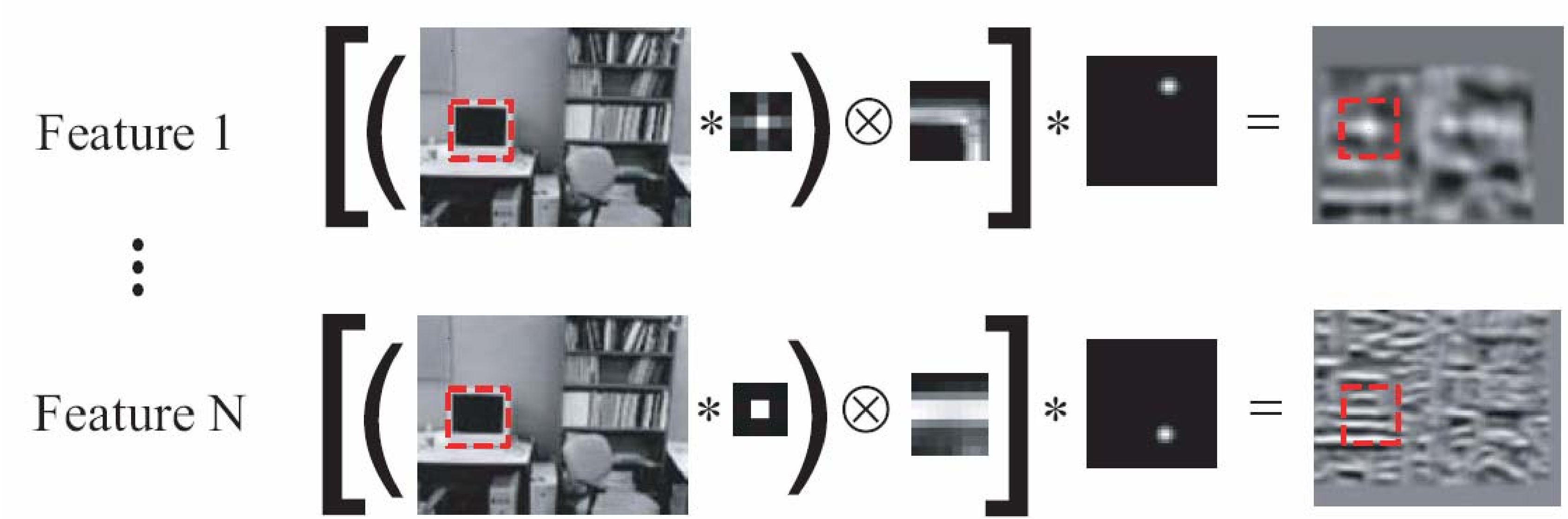
We can do a better job using filtered images



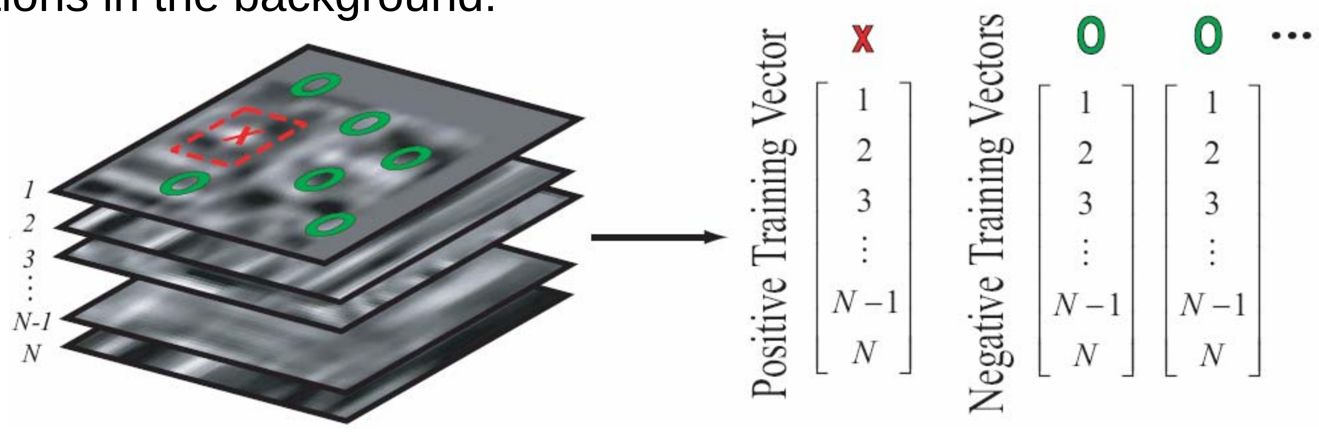
Still a weak detector
but better than before

Training

First we evaluate all the N features on all the training images.

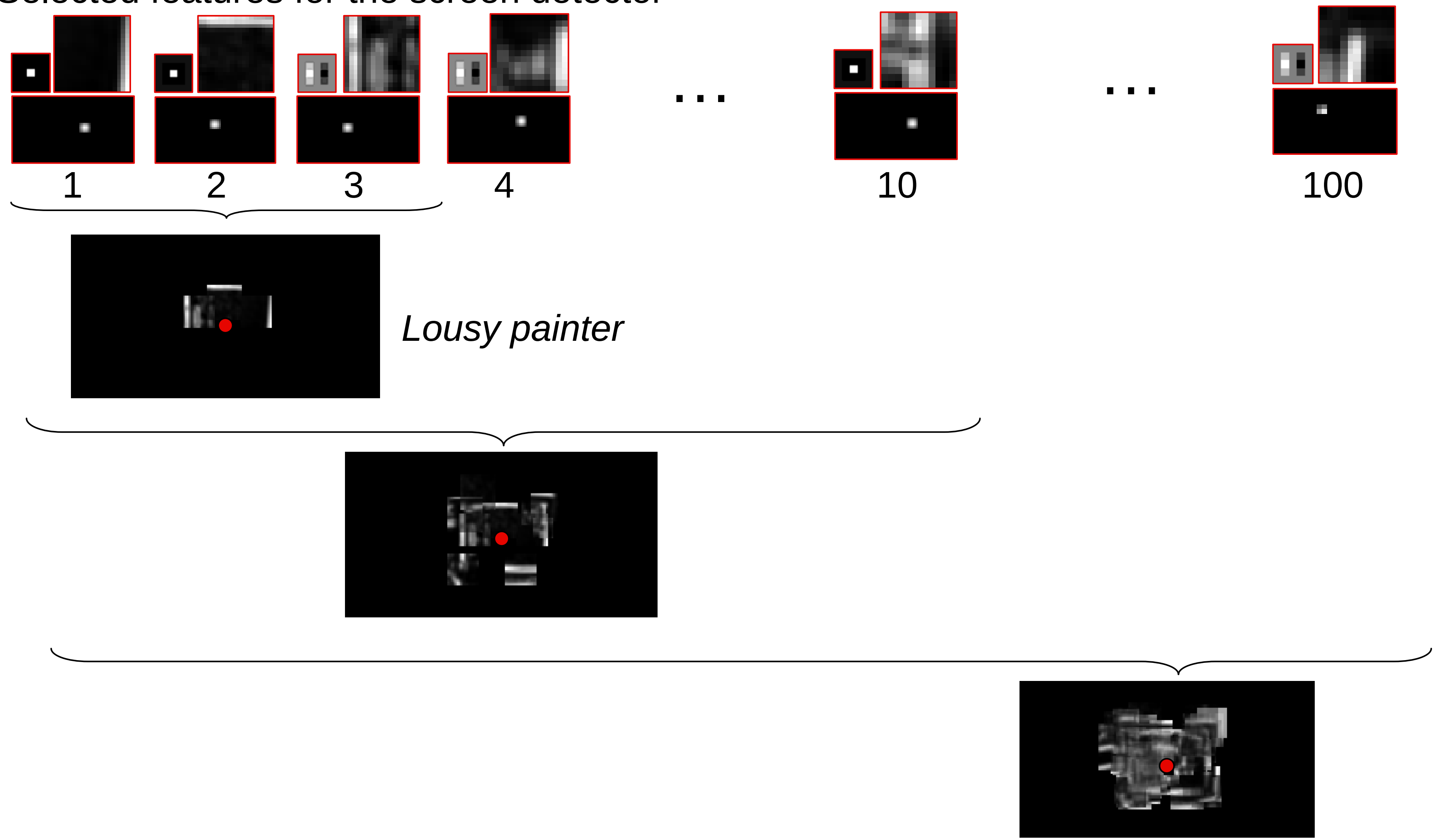


Then, we sample the feature outputs on the object center and at random locations in the background:



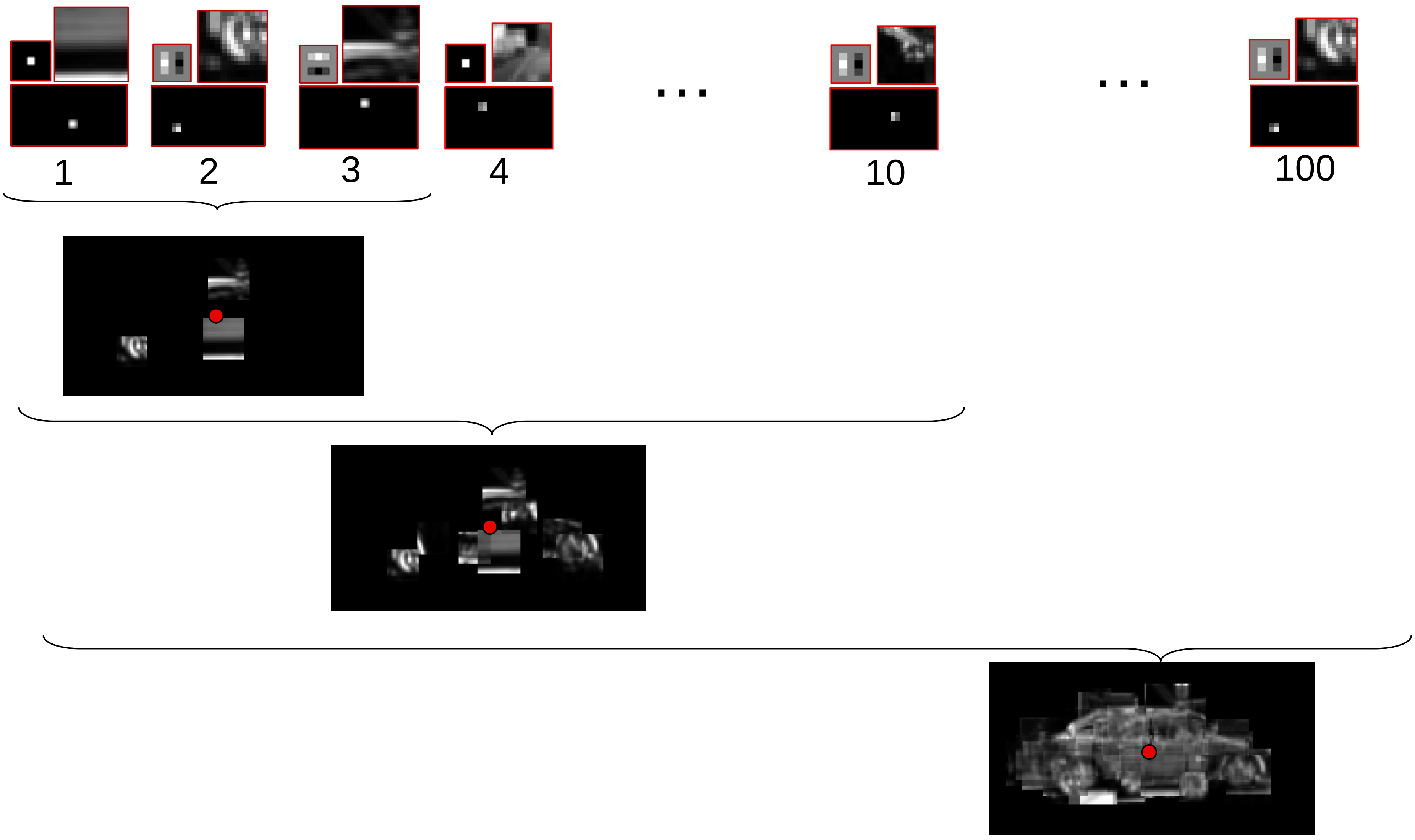
Representation and object model

Selected features for the screen detector



Representation and object model

Selected features for the car detector



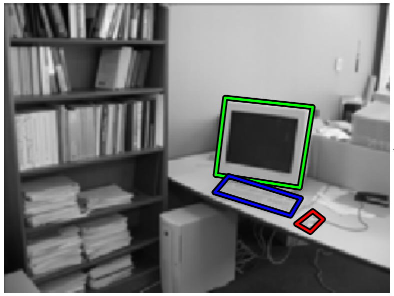
Overview of section

- Object detection with classifiers

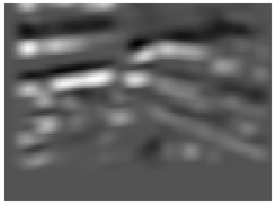
- Boosting
  - Gentle boosting
  - Weak detectors
  - Object model
  - **Object detection**

- Multiclass object detection

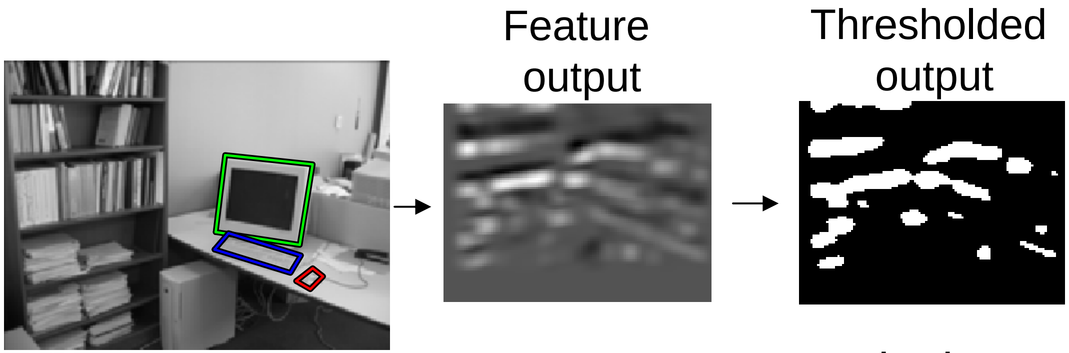
Example: screen detection



Feature output



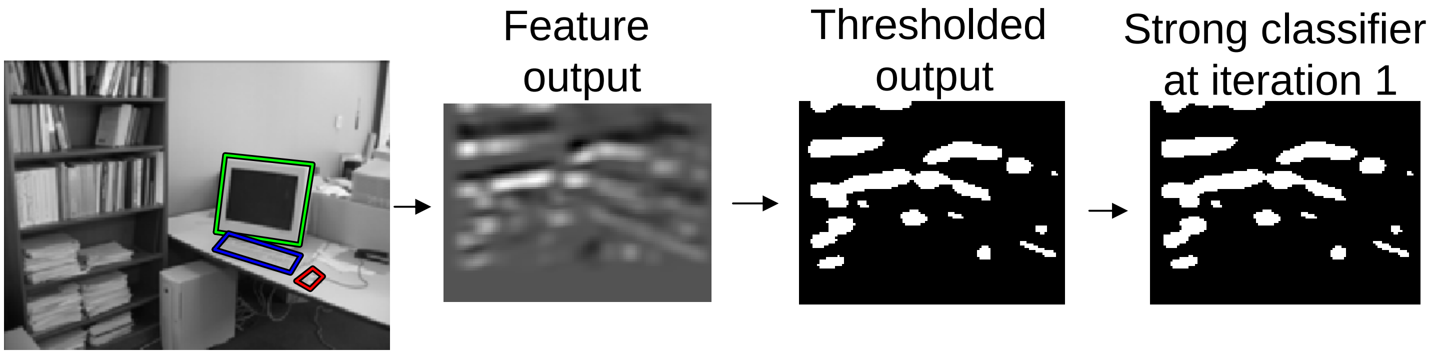
Example: screen detection



Weak ‘detector’

Produces many false alarms.

Example: screen detection

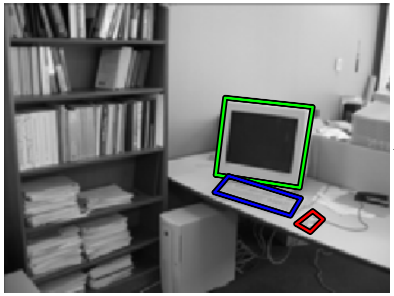


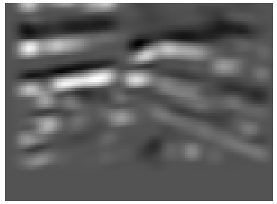
Example: screen detection

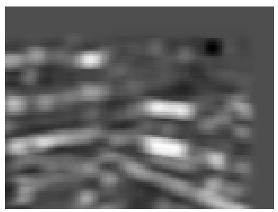
Feature output

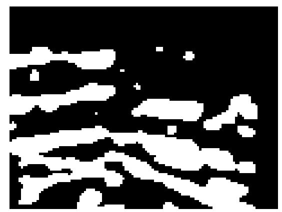
Thresholded output

Strong classifier





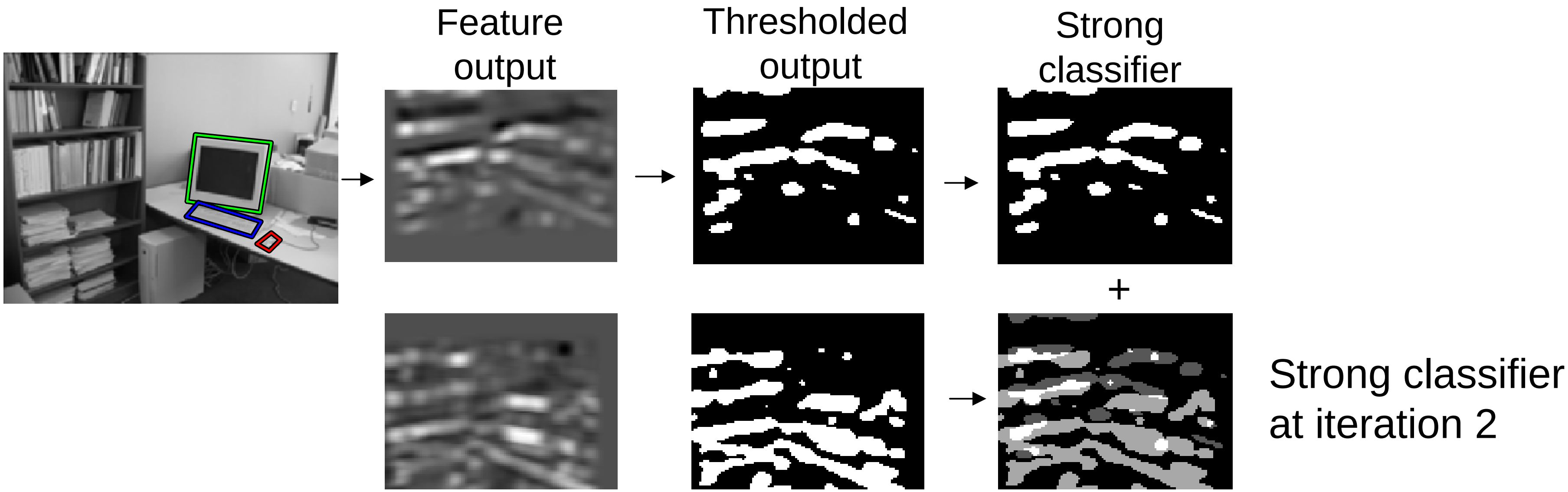




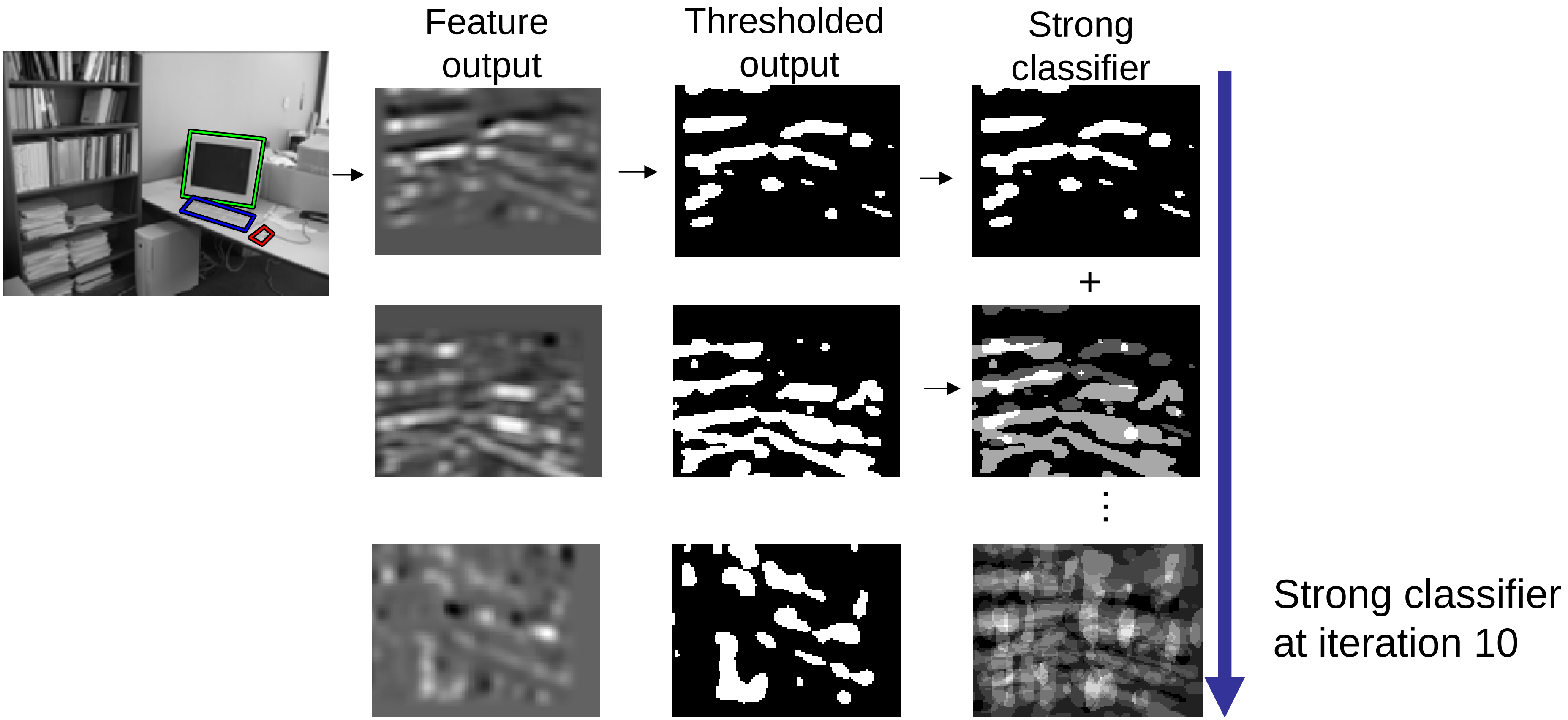
Second weak 'detector'

Produces a different set of false alarms.

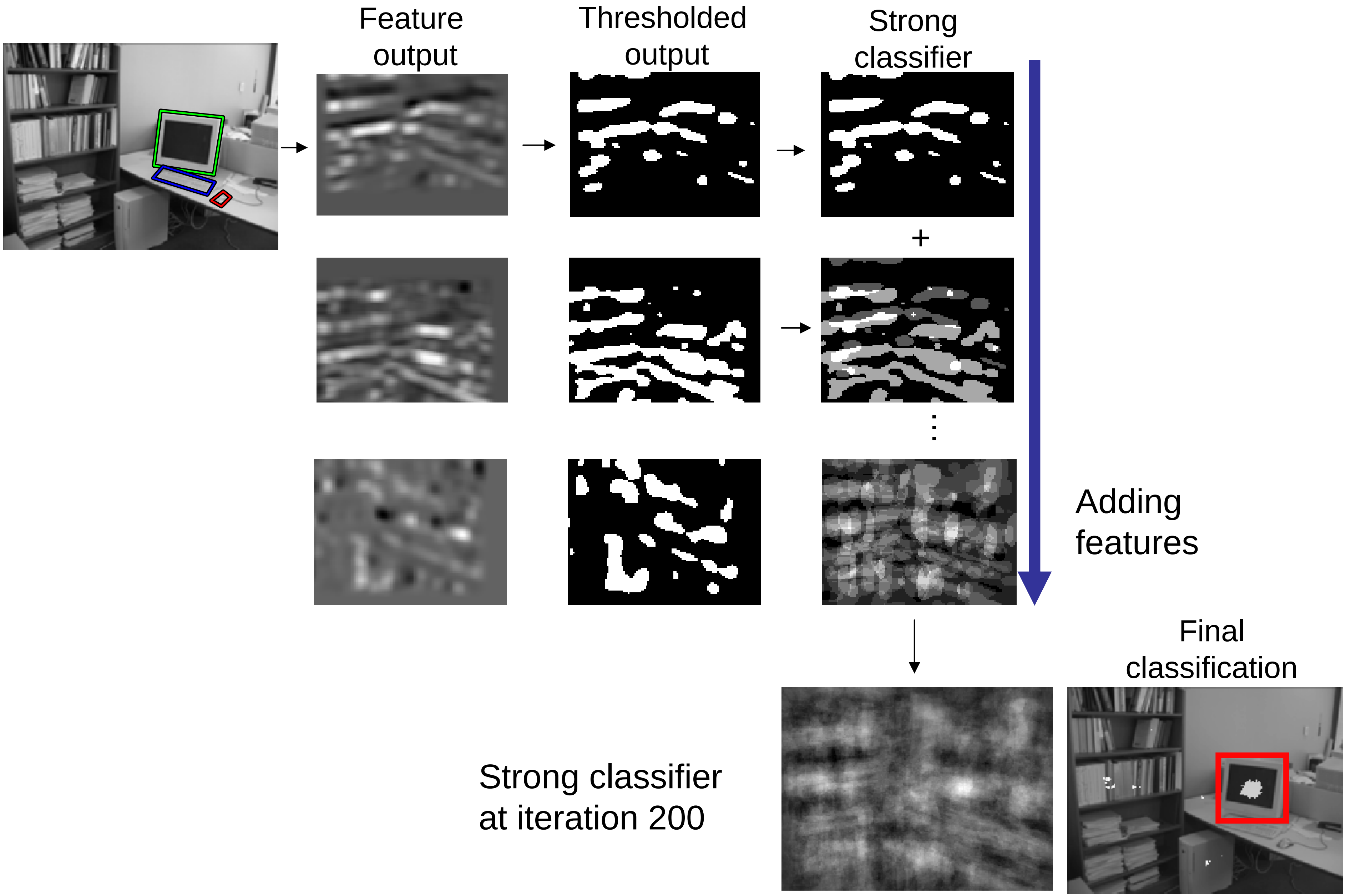
Example: screen detection



Example: screen detection



Example: screen detection



applications

BBC NEWS

UK version | International version | About the versions | L

Last Updated: Monday, 6 February 2006, 14:29 GMT

E-mail this to a friend | Printable version

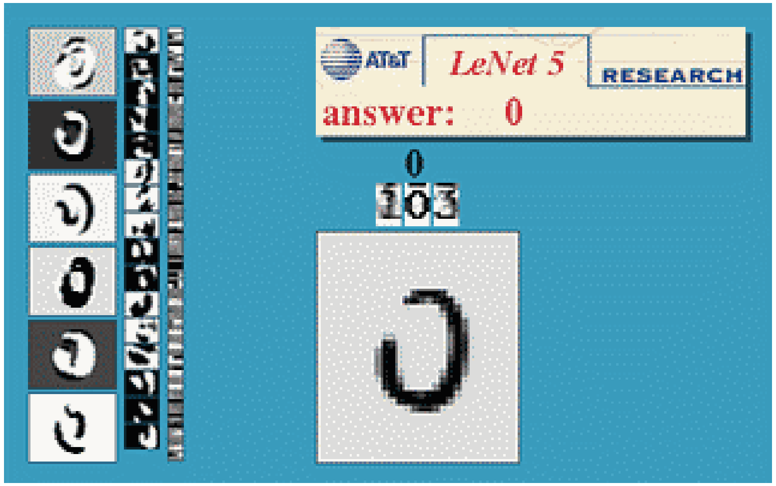
Face-hunting cameras boost Nikon

Japanese camera maker Nikon has tripled its profits on the back of strong sales of digital cameras that automatically focus on human faces.

Face recognition cameras like the Coolpix L1 are popular

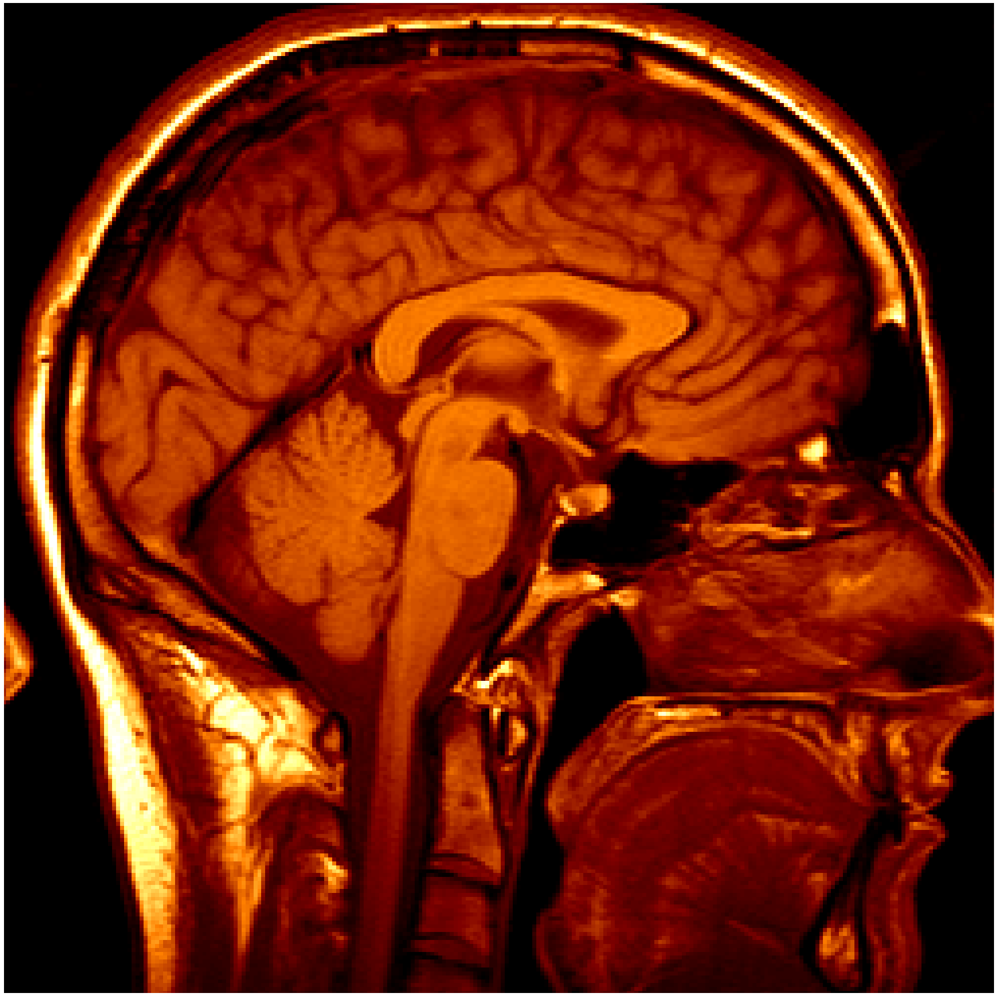
Sample image: Subject as seen on the COOLPIX 5900 camera's color LCD and when using Nikon's Face-priority AF function.

Document Analysis

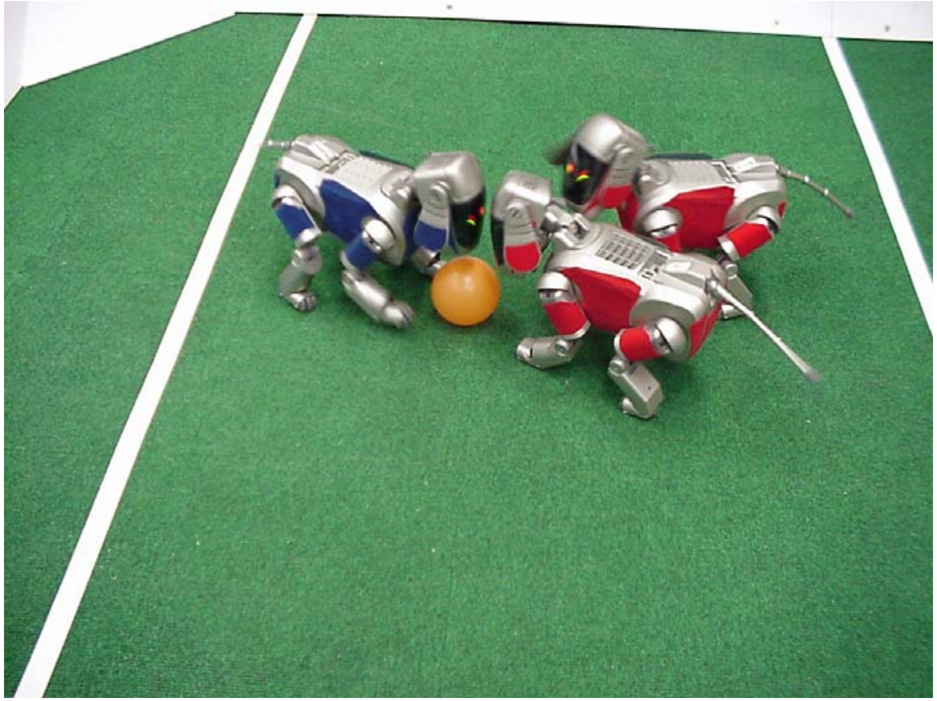


Digit recognition, AT&T labs
http://www.research.att.com/~yann/

Medical Imaging

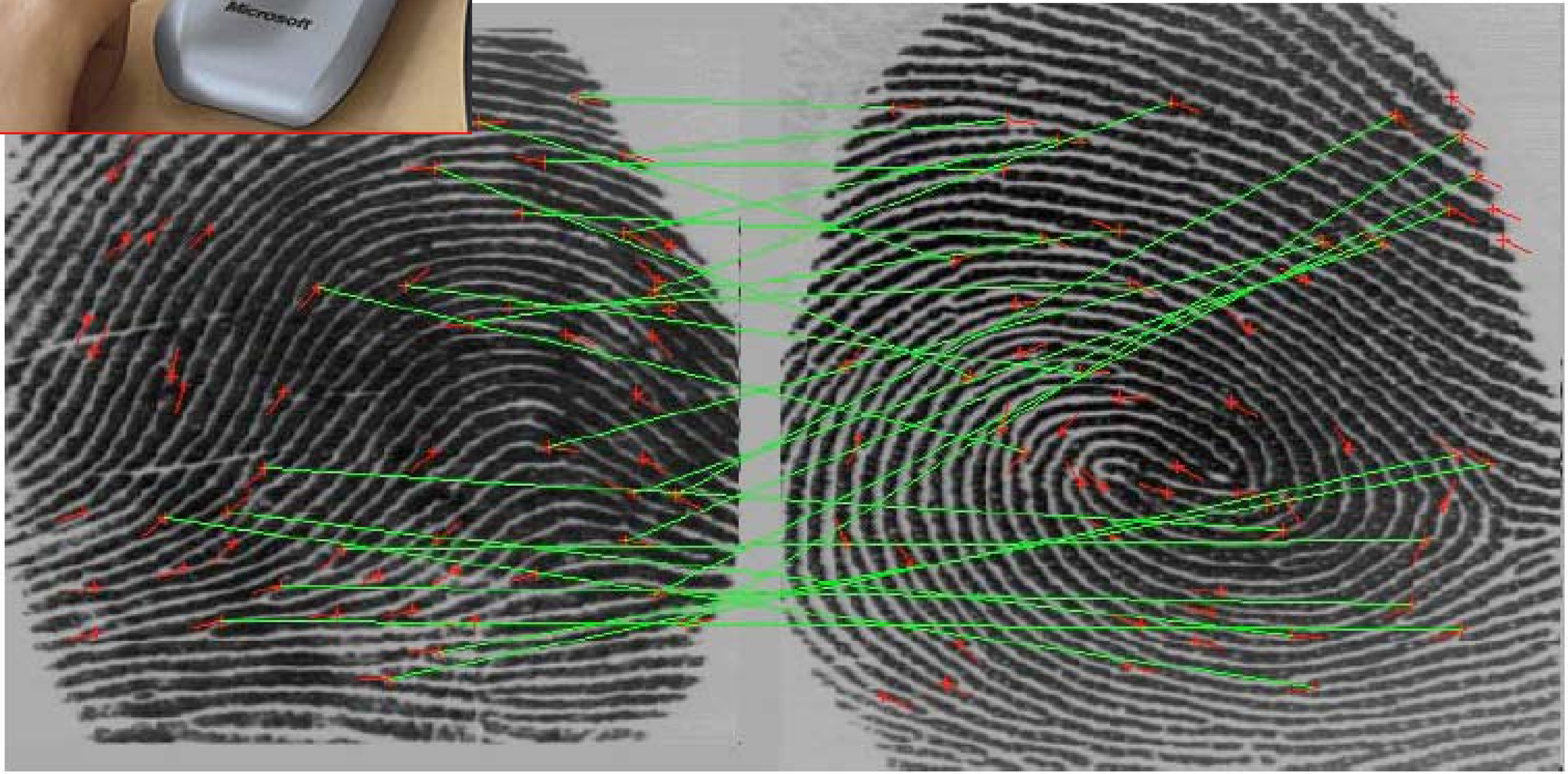


Robotics

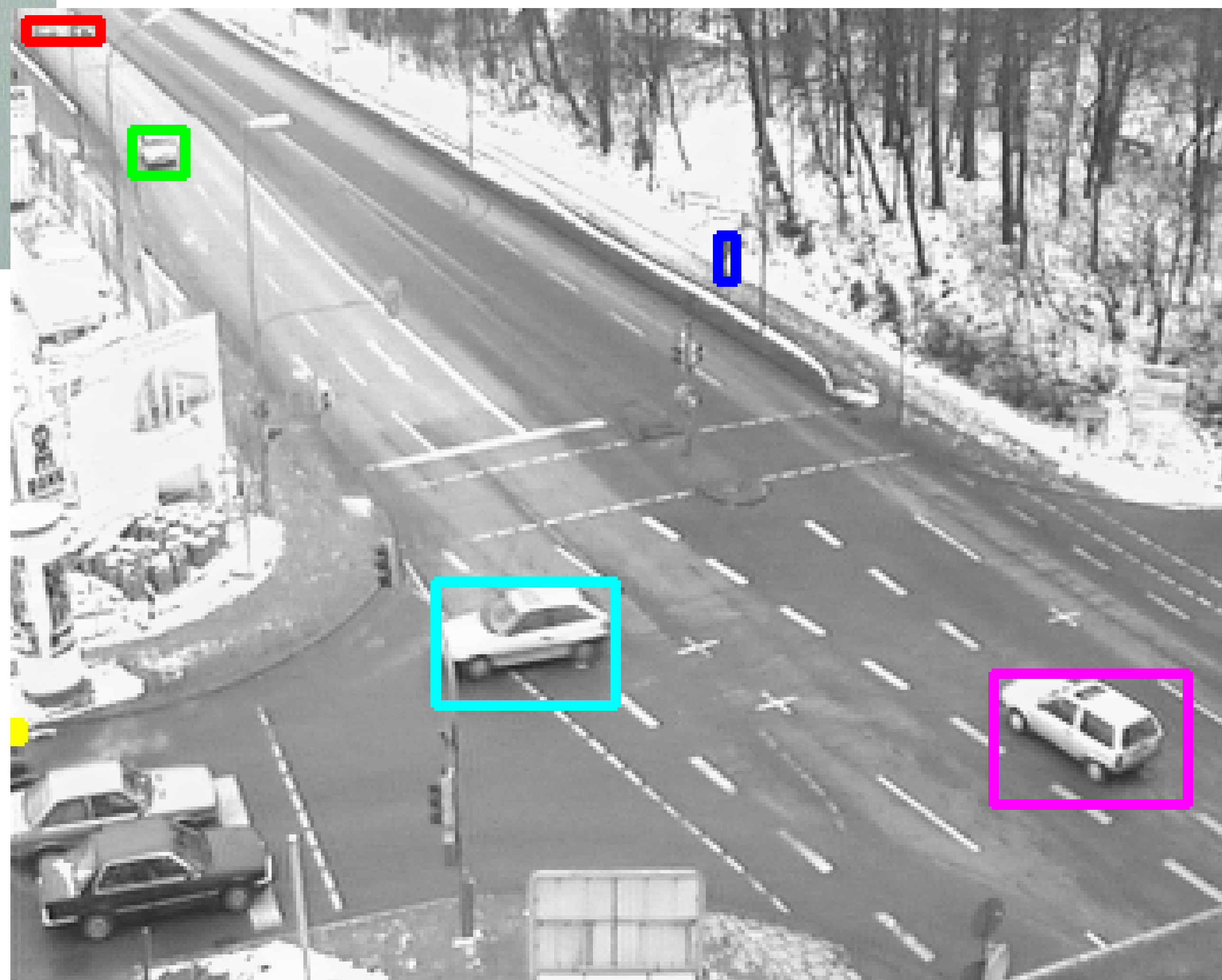


Toys and robots

Finger prints



Surveillance



Security

Searching the web