Recognition with Bag-of-Words

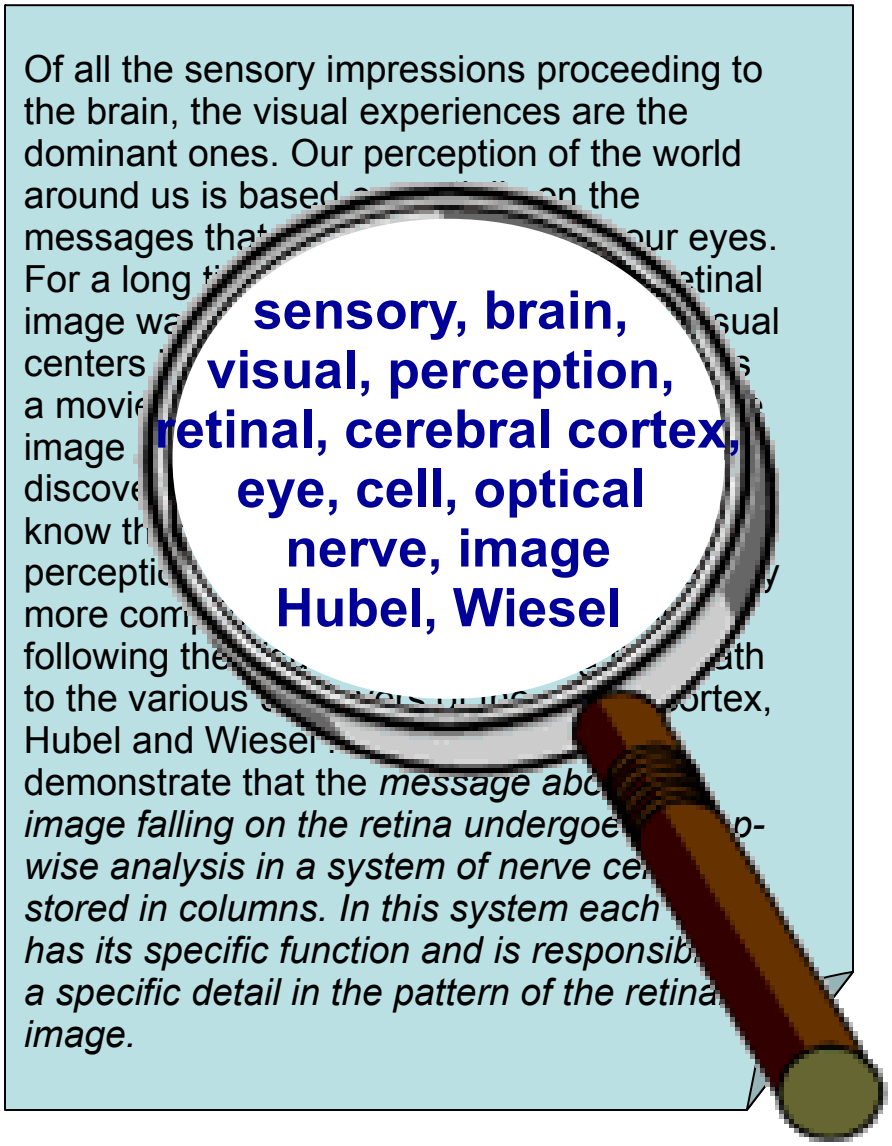
(Borrowing heavily from Tutorial Slides by Li Fei-fei)

Recognition

- So far, we've worked on recognizing edges
- Now, we'll work on recognizing objects
- We will use a bag-of-words approach

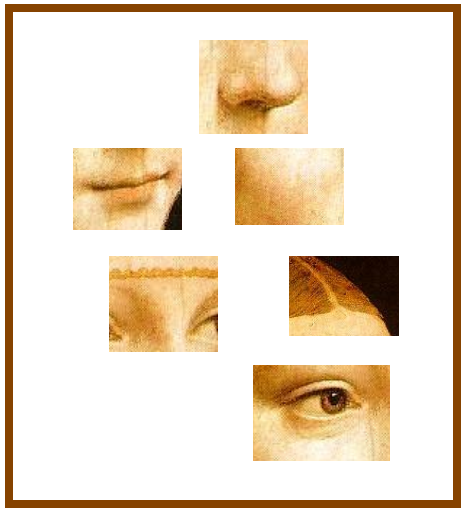
Object → Bag of ‘words’

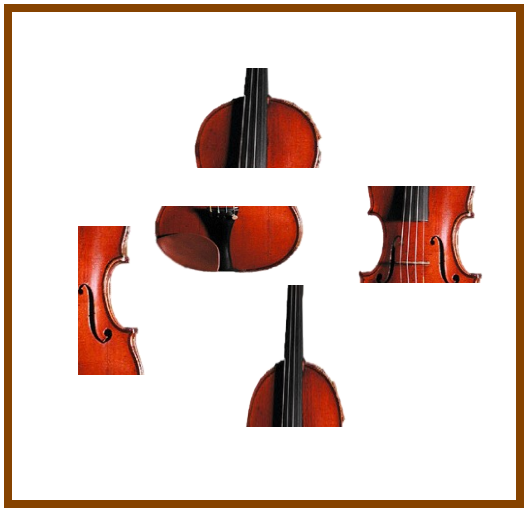
Analogy to documents



A clarification: definition of “BoW”

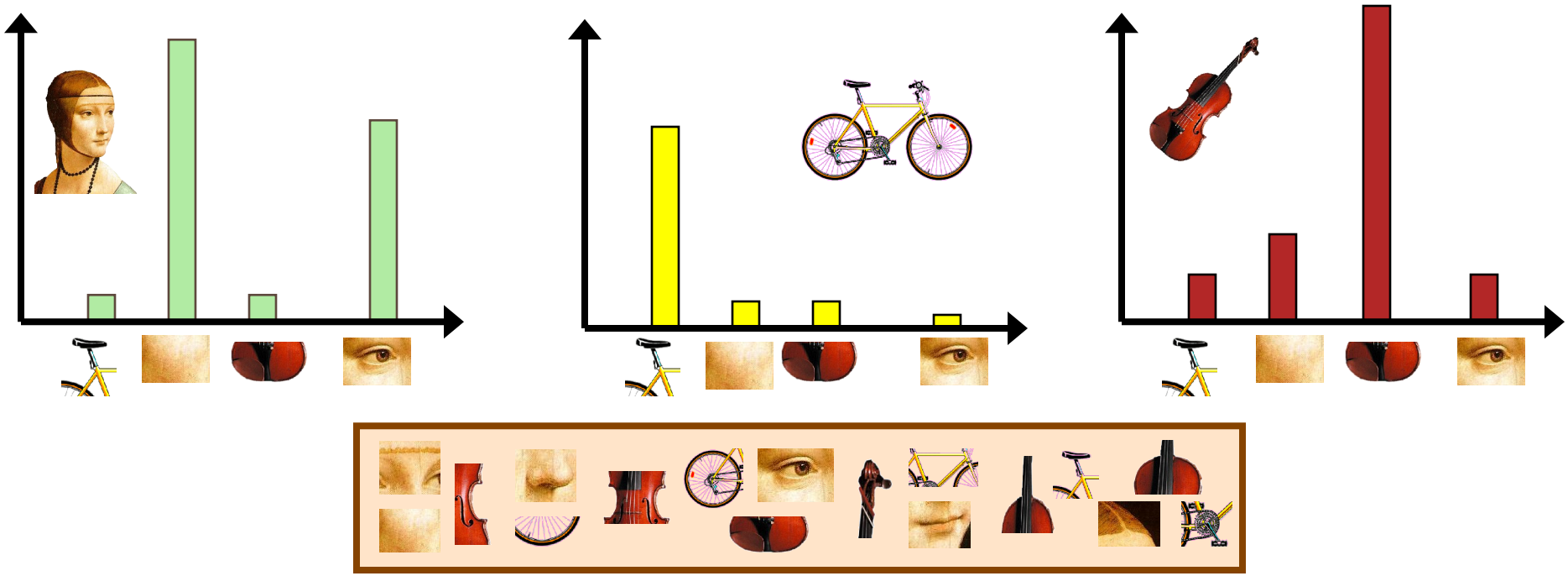
- Looser definition
  - Independent features



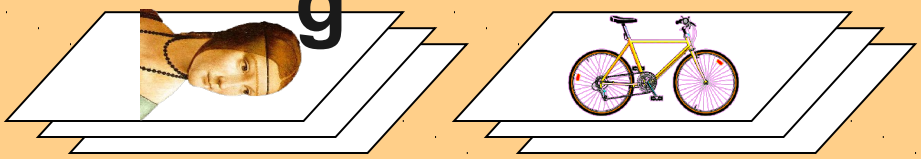


A clarification: definition of “BoW”

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



learnin
g



recognition

codewords dictionary

feature detection
& representation

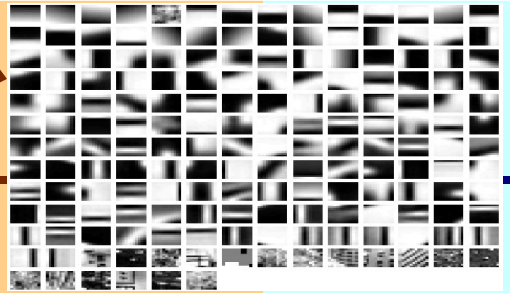
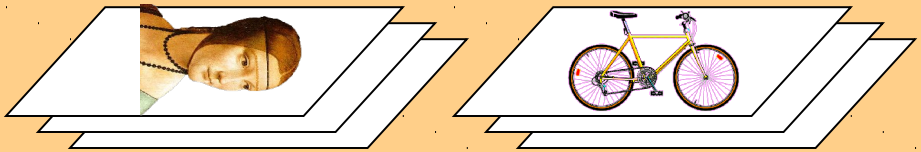


image representation

category models
(and/or) classifiers

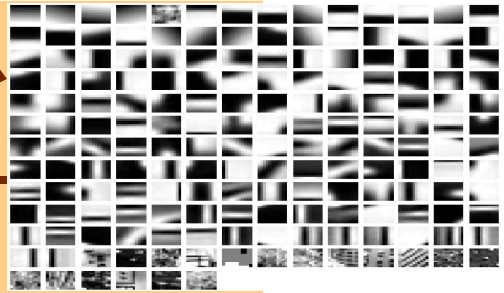
category
decision

Representation



2.

codewords dictionary



1. feature detection & representation

image representation

3.

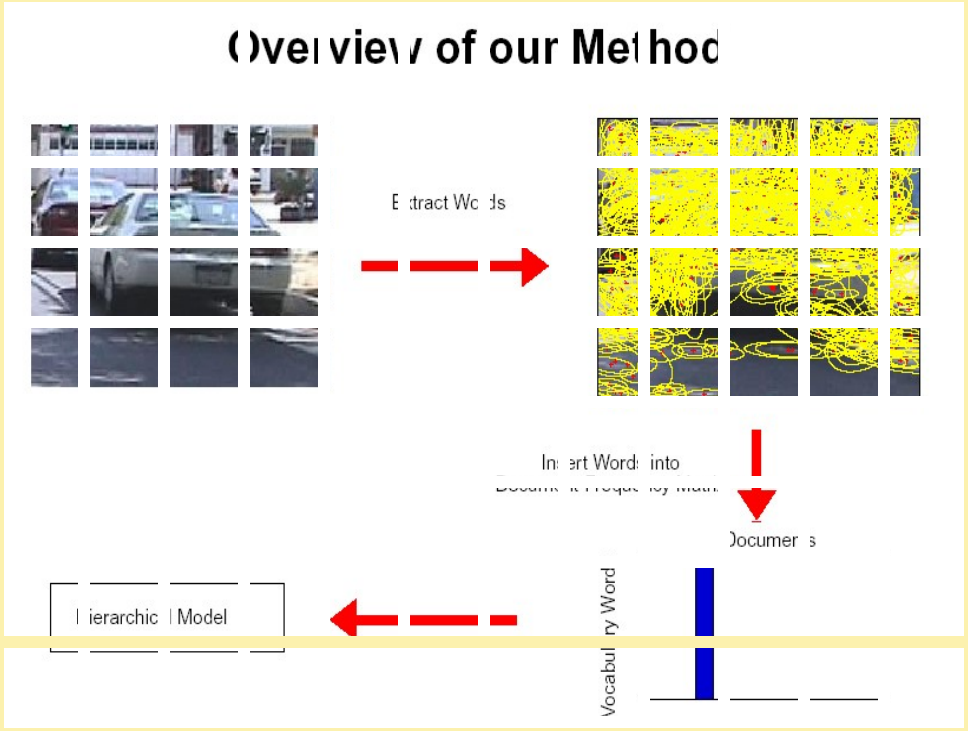
1.Feature detection and representation

1.Feature detection and representation

Regular grid

Vogel & Schiele, 2003

Fei-Fei & Perona, 2005



1.Feature detection and representation

Regular grid

Vogel & Schiele, 2003

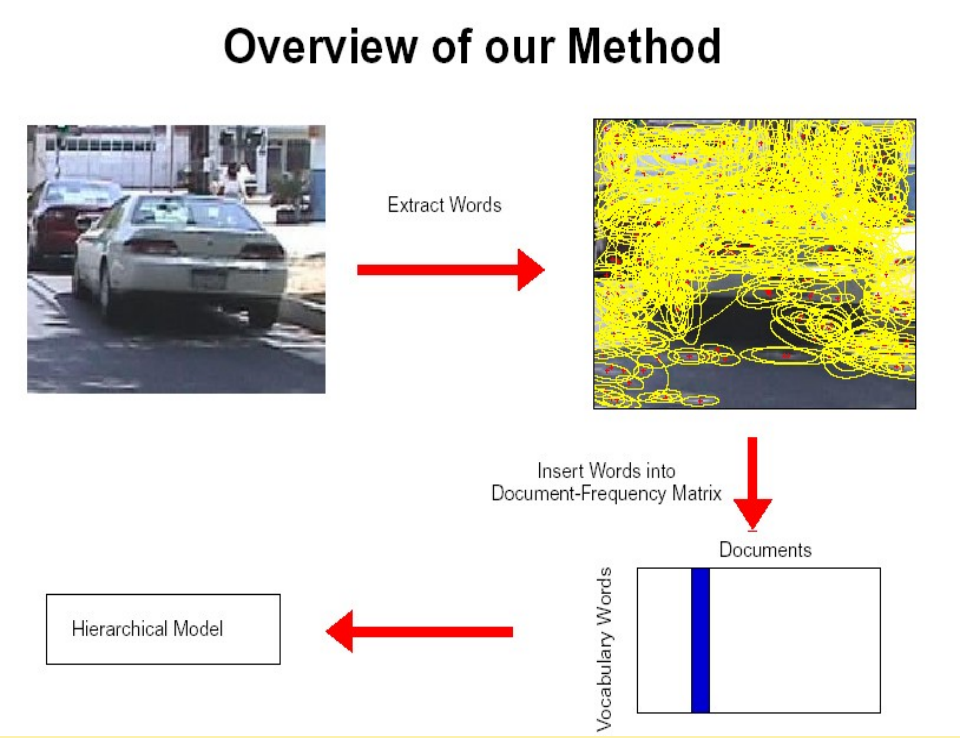
Fei-Fei & Perona, 2005

Interest point detector

Csurka, et al. 2004

Fei-Fei & Perona, 2005

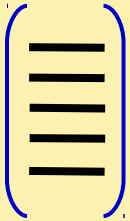
Sivic, et al. 2005



1.Feature detection and representation

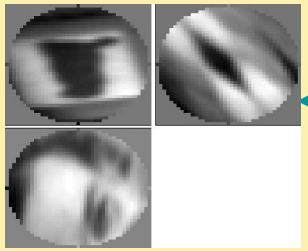
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka, Bray, Dance & Fan, 2004
  - Fei-Fei & Perona, 2005
  - Sivic, Russell, Efros, Freeman & Zisserman, 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation based patches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

1.Feature detection and representation



Compute SIFT descriptor

[Lowe'99]



Normalize patch

Detect patches

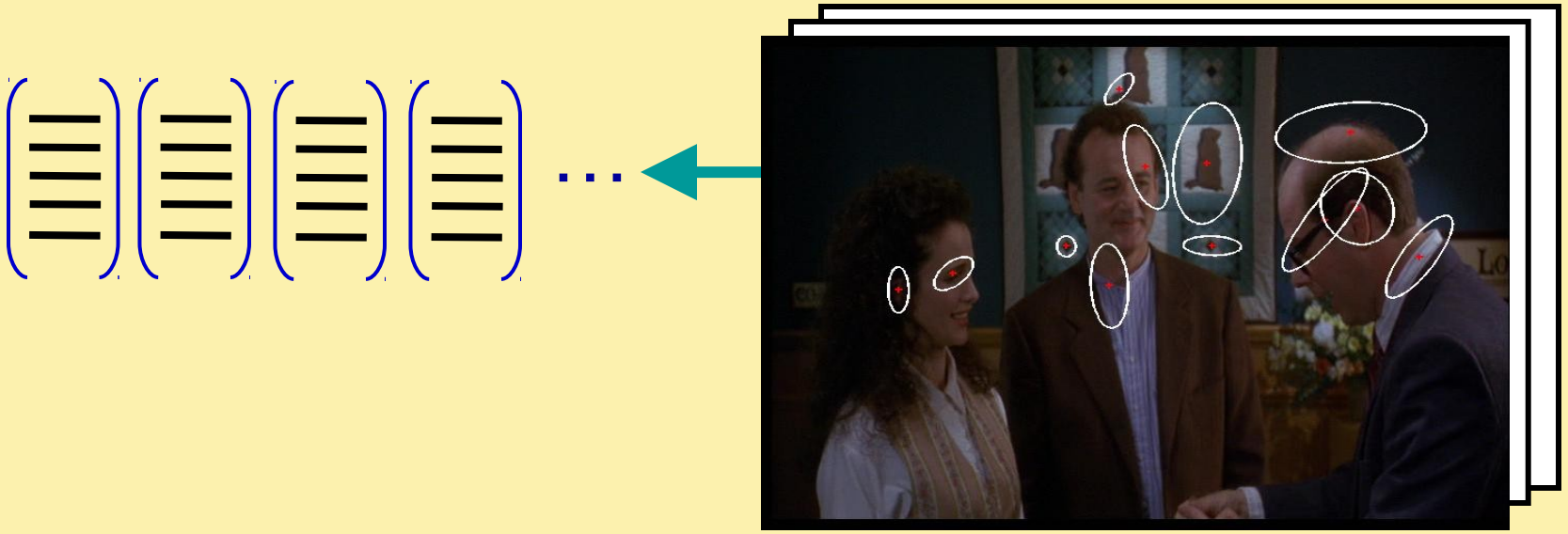
[Mikojaczyk and Schmid '02]

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

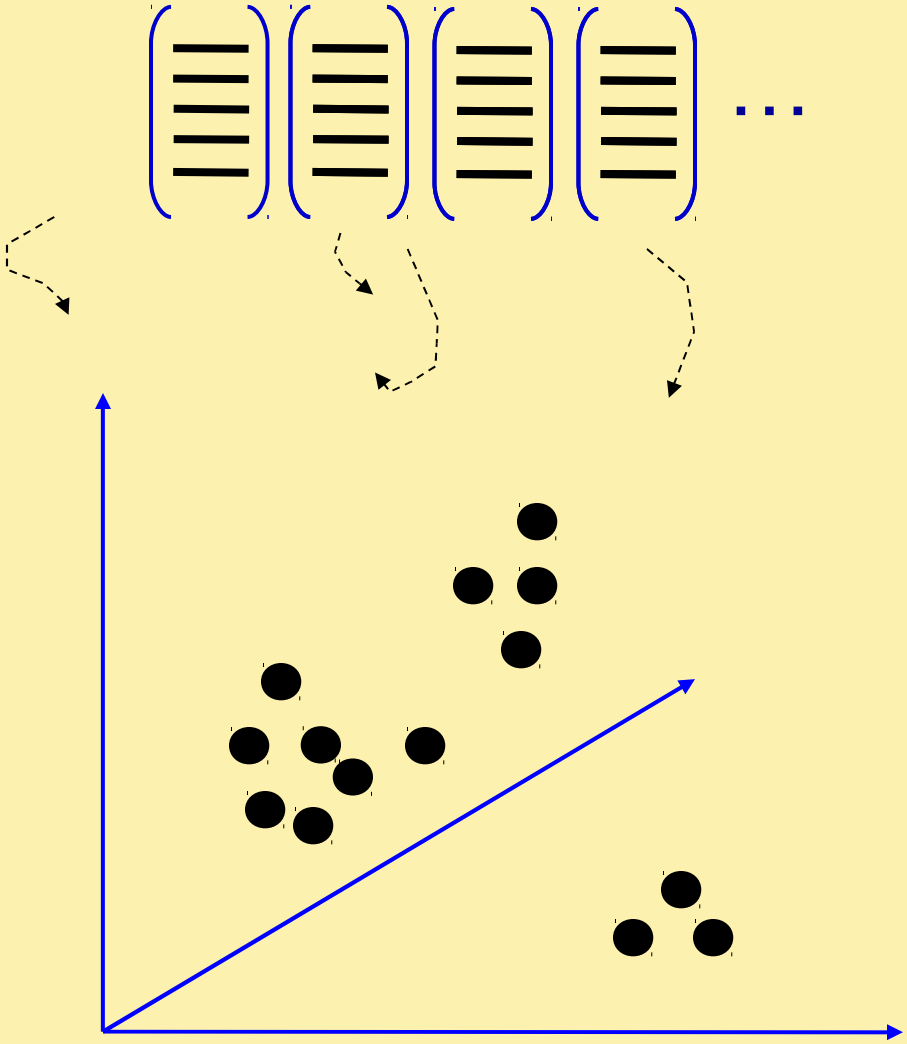
[Sivic & Zisserman, '03]

Slide credit: Josef Sivic

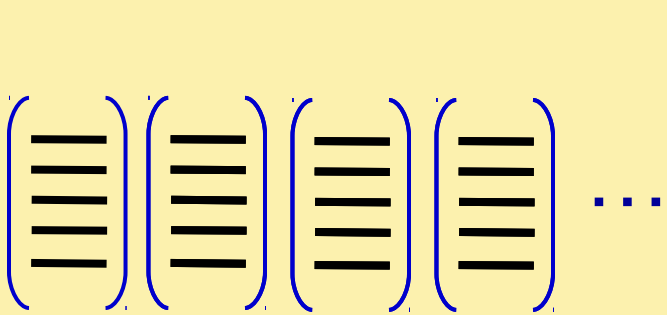
1.Feature detection and representation

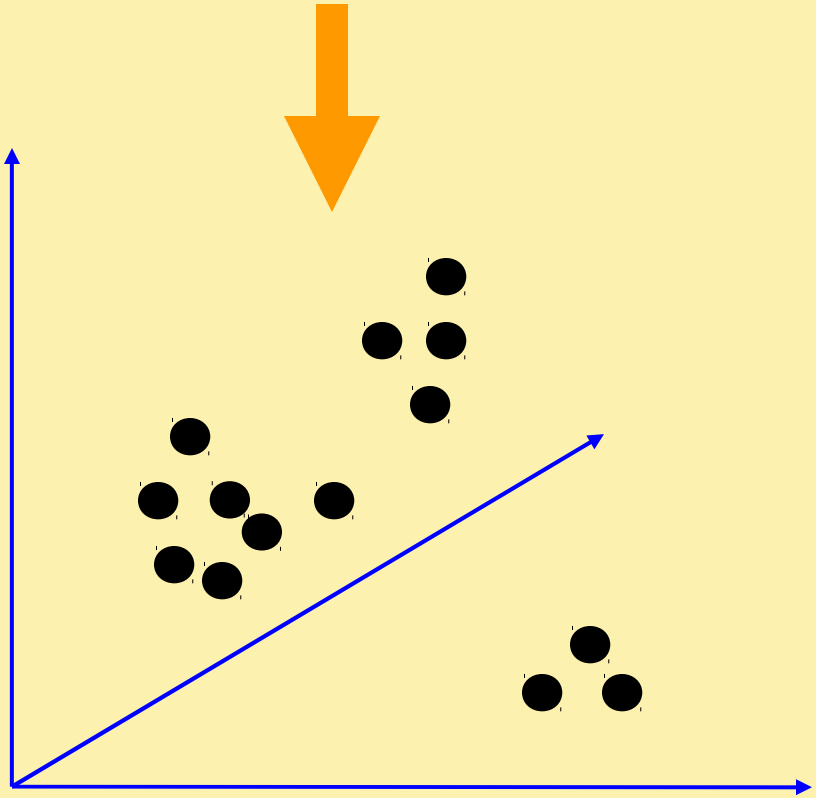


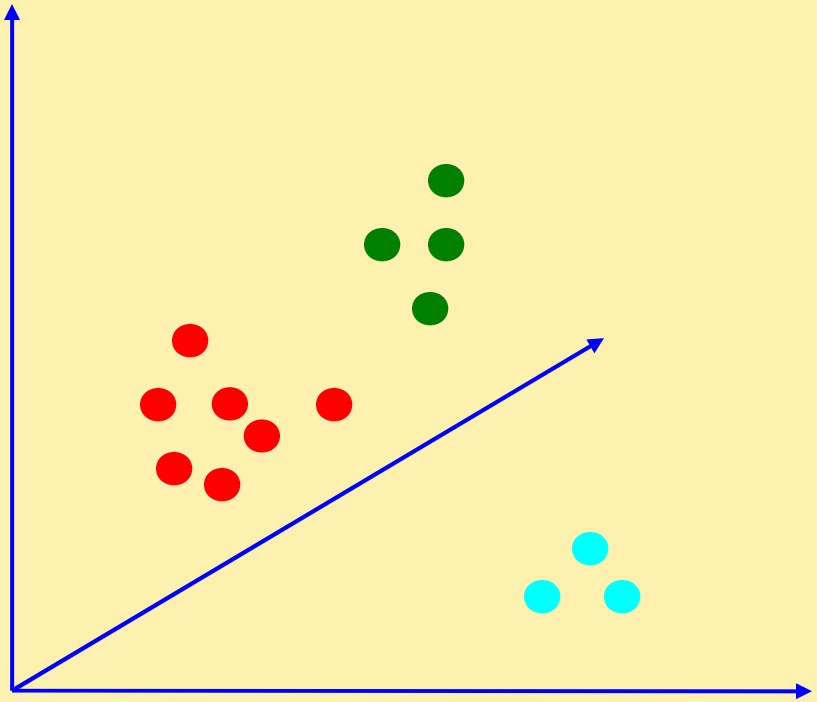
2. Codewords dictionary formation



2. Codewords dictionary formation



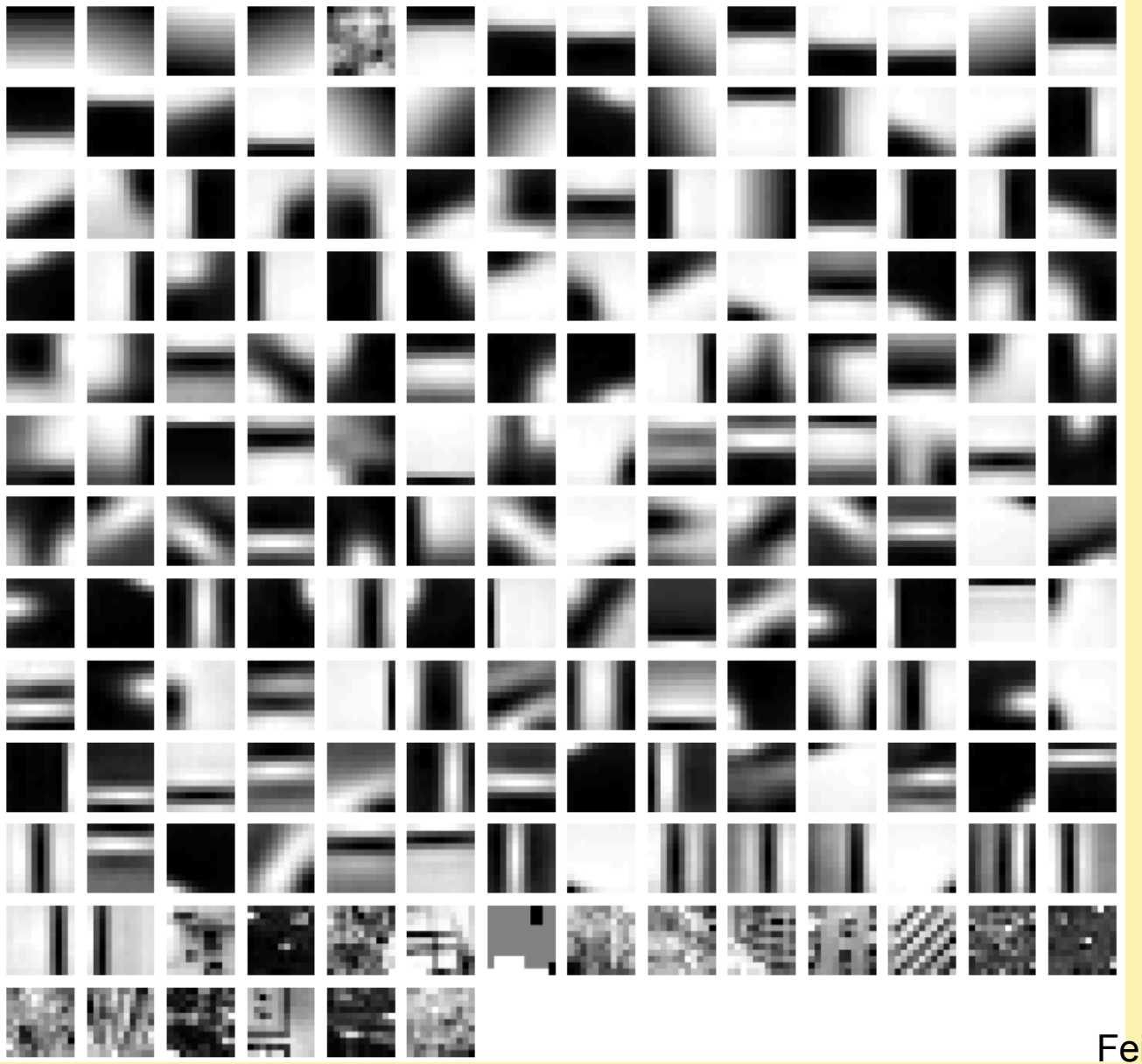




Vector quantization

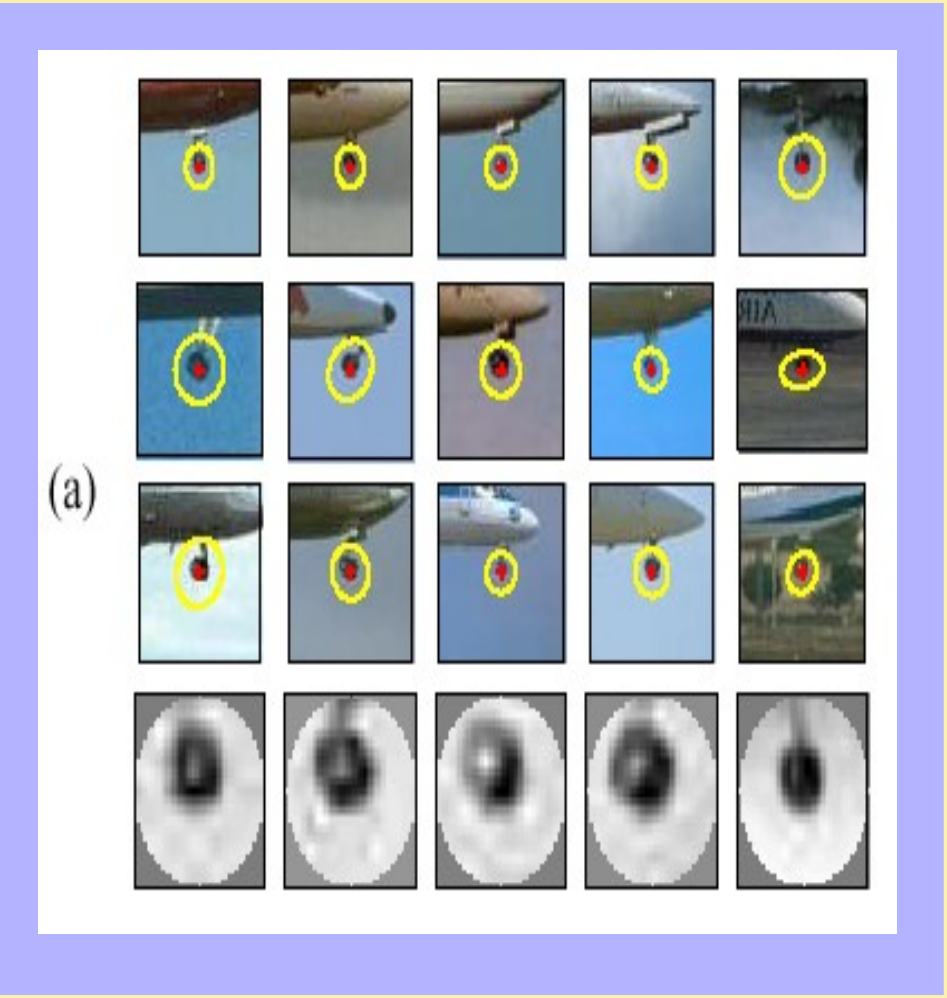
Slide credit: Josef Sivic

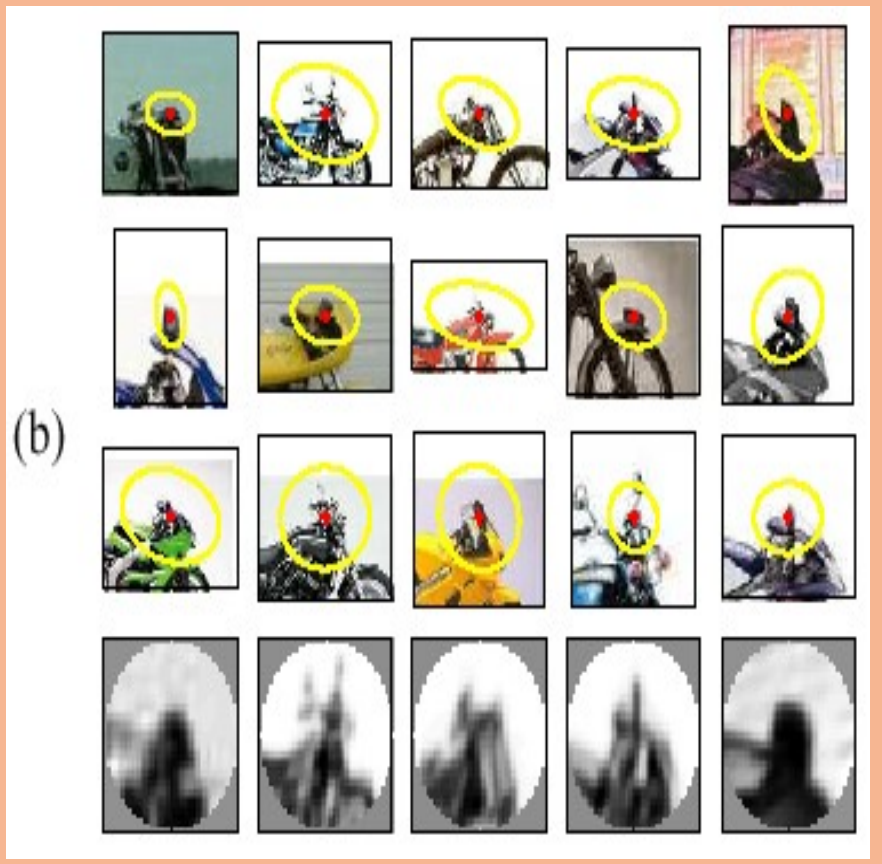
2. Codewords dictionary formation



Fei-Fei et al. 2005

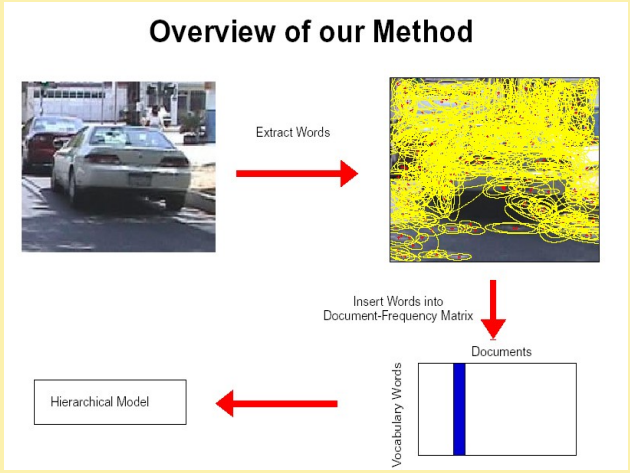
Image patch examples of codewords



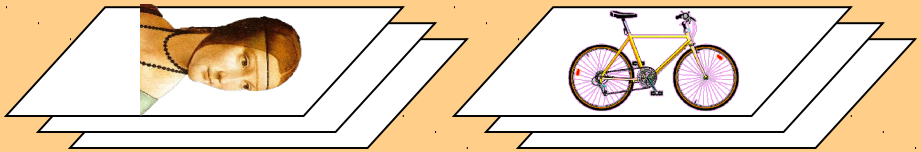


Sivic et al. 2005

3. Image representation

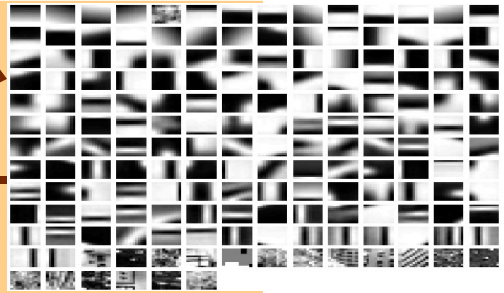


Representation



2.

codewords dictionary



1. feature detection & representation

image representation

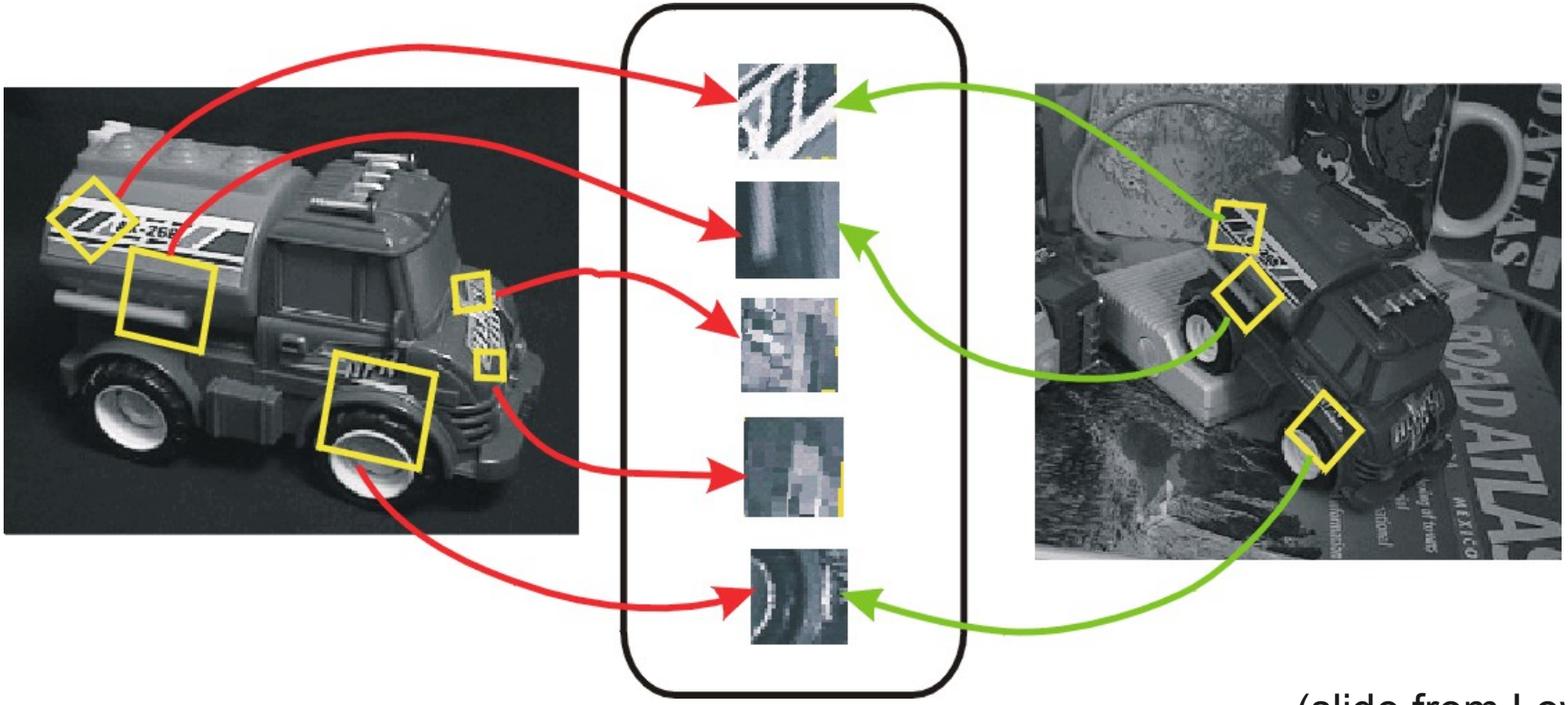
3.

One of the keys to success is a good representation of features

- Just pixels is a bad representation
- Pixel intensities are affected by a lot of different things
  - Rotation, scaling, perspective
  - Illumination changes
  - Reordering of scenes
- We want a good way of characterizing image patches that is somewhat robust to these different effects

Scale-Invariant Local Features

- Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters



SIFT Features

(slide from Lowe)

Advantages of invariant local features

- **Locality:** features are local, so robust to occlusion and clutter (no prior segmentation)
- **Distinctiveness:** individual features can be matched to a large database of objects
- **Quantity:** many features can be generated for even small objects
- **Efficiency:** close to real-time performance
- **Extensibility:** can easily be extended to wide range of differing feature types, with each adding robustness

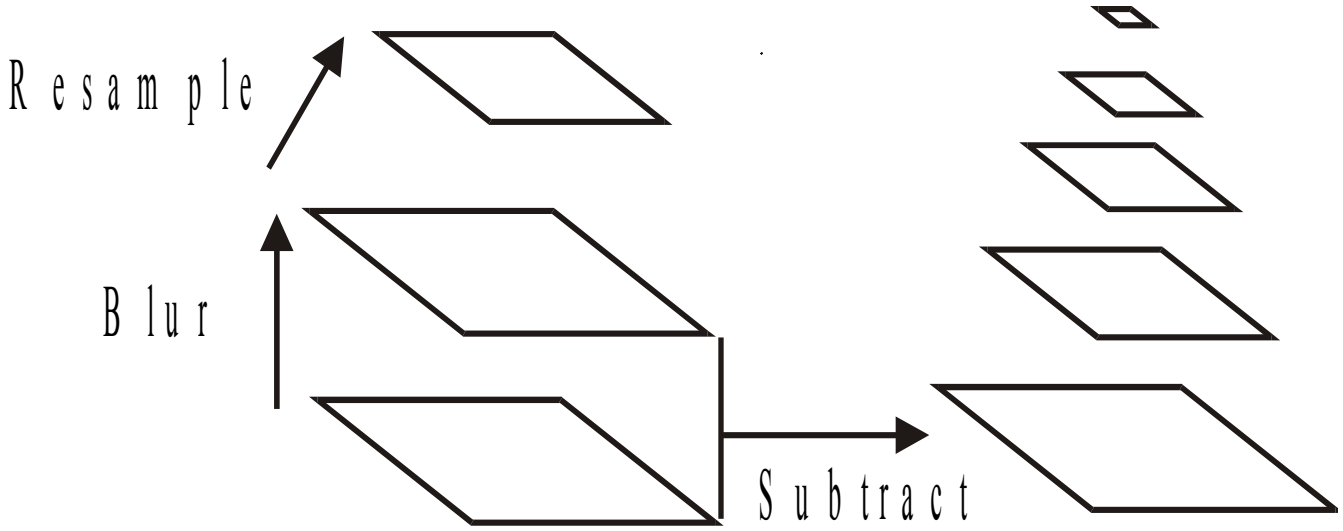
(slide from Lowe)

Think Back to Bag of Words - Two Key Problems

- Problem 1: What parts of the image do I look at?
- Problem 2: How do I represent the patches of pixels

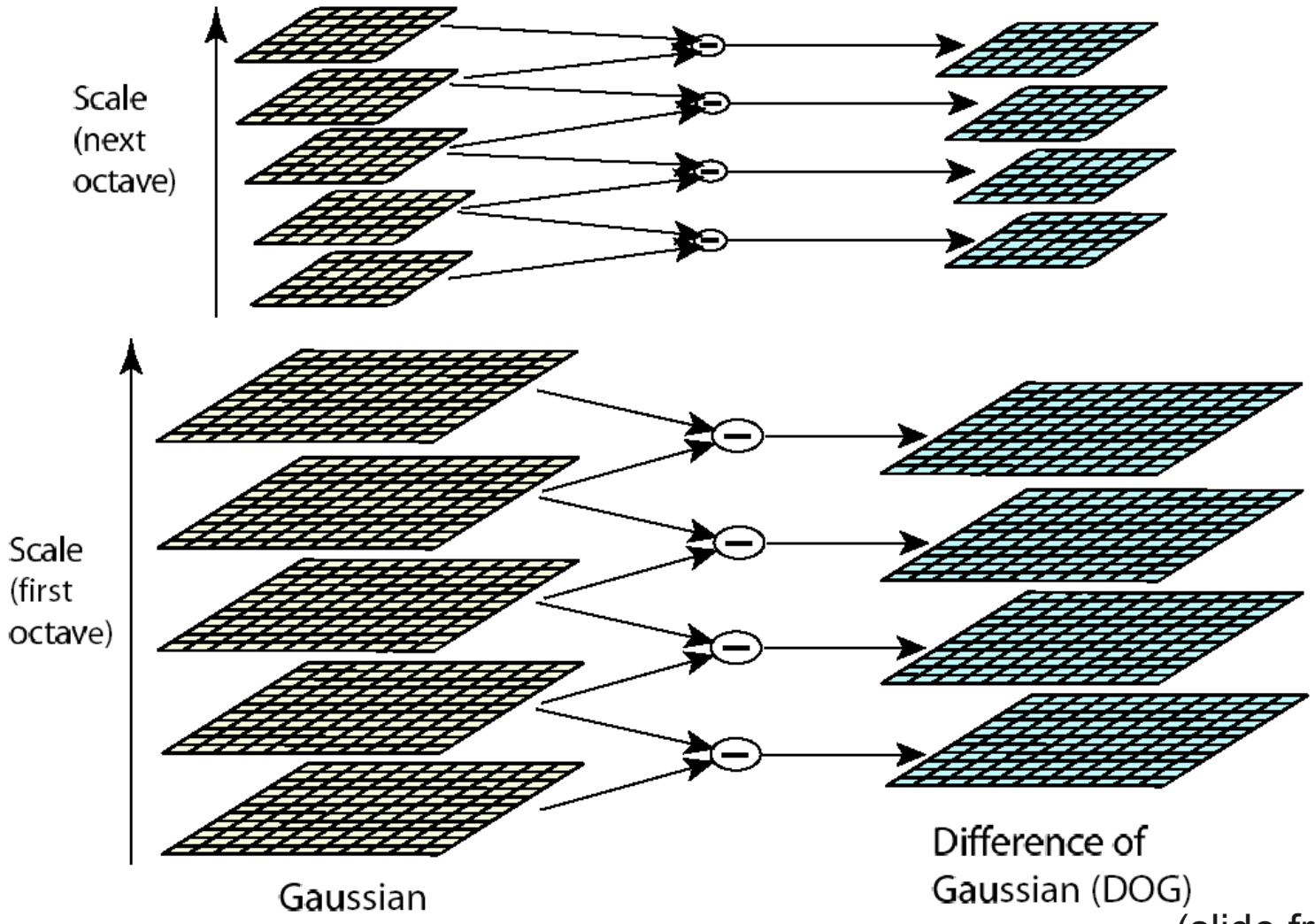
Build Scale-Space Pyramid

- All scales must be examined to identify scale-invariant features
- An efficient function is to compute the Difference of Gaussian (DOG) pyramid (Burt & Adelson, 1983)



(slide from Lowe)

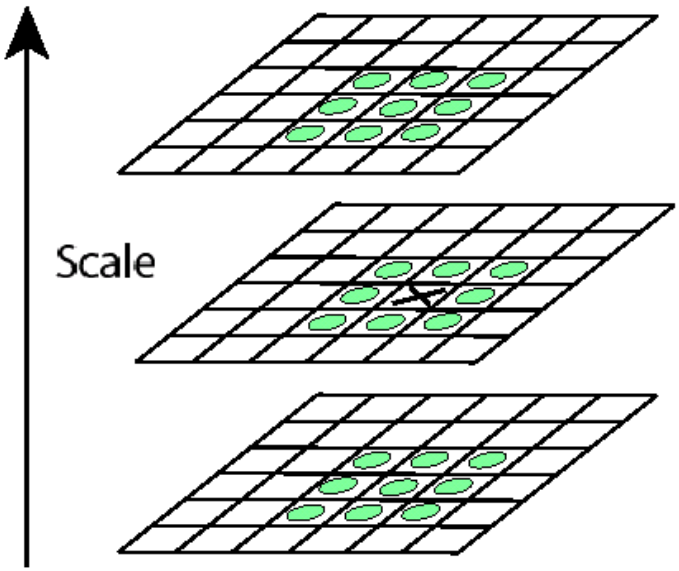
Scale space processed one octave at a time



(slide from Lowe)

Key point localization

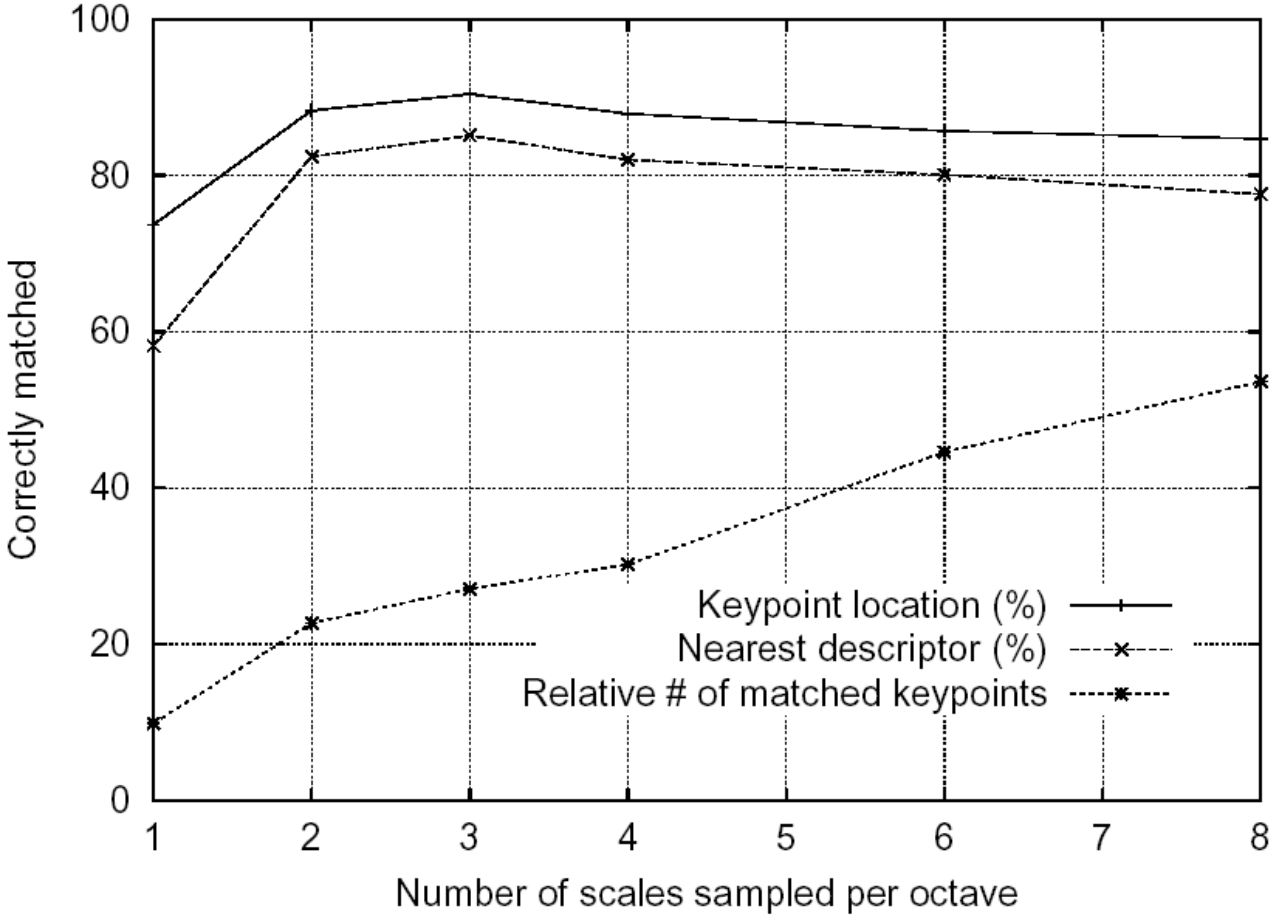
- Detect maxima and minima of difference-of-Gaussian in scale space



(slide from Lowe)

Sampling frequency for scale

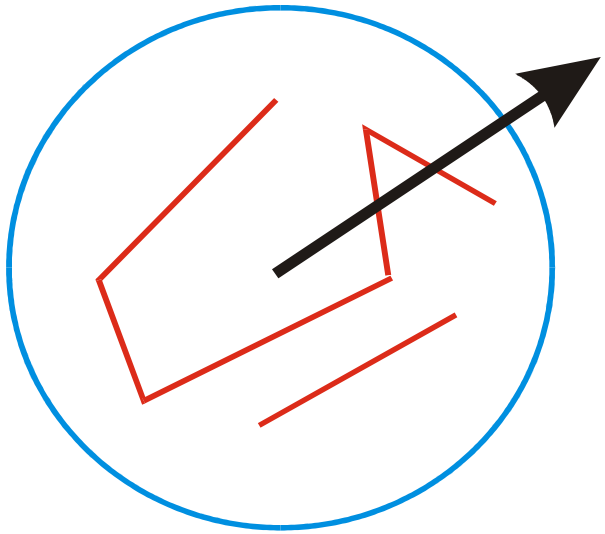
More points are found as sampling frequency increases, but accuracy of matching decreases after 3 scales/octave

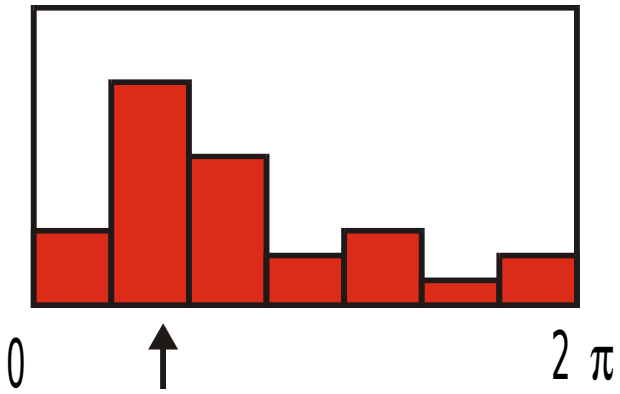


(slide from Lowe)

Select canonical orientation

- Create histogram of local gradient directions computed at selected scale
- Assign canonical orientation at peak of smoothed histogram
- Each key specifies stable 2D coordinates (x, y, scale, orientation)

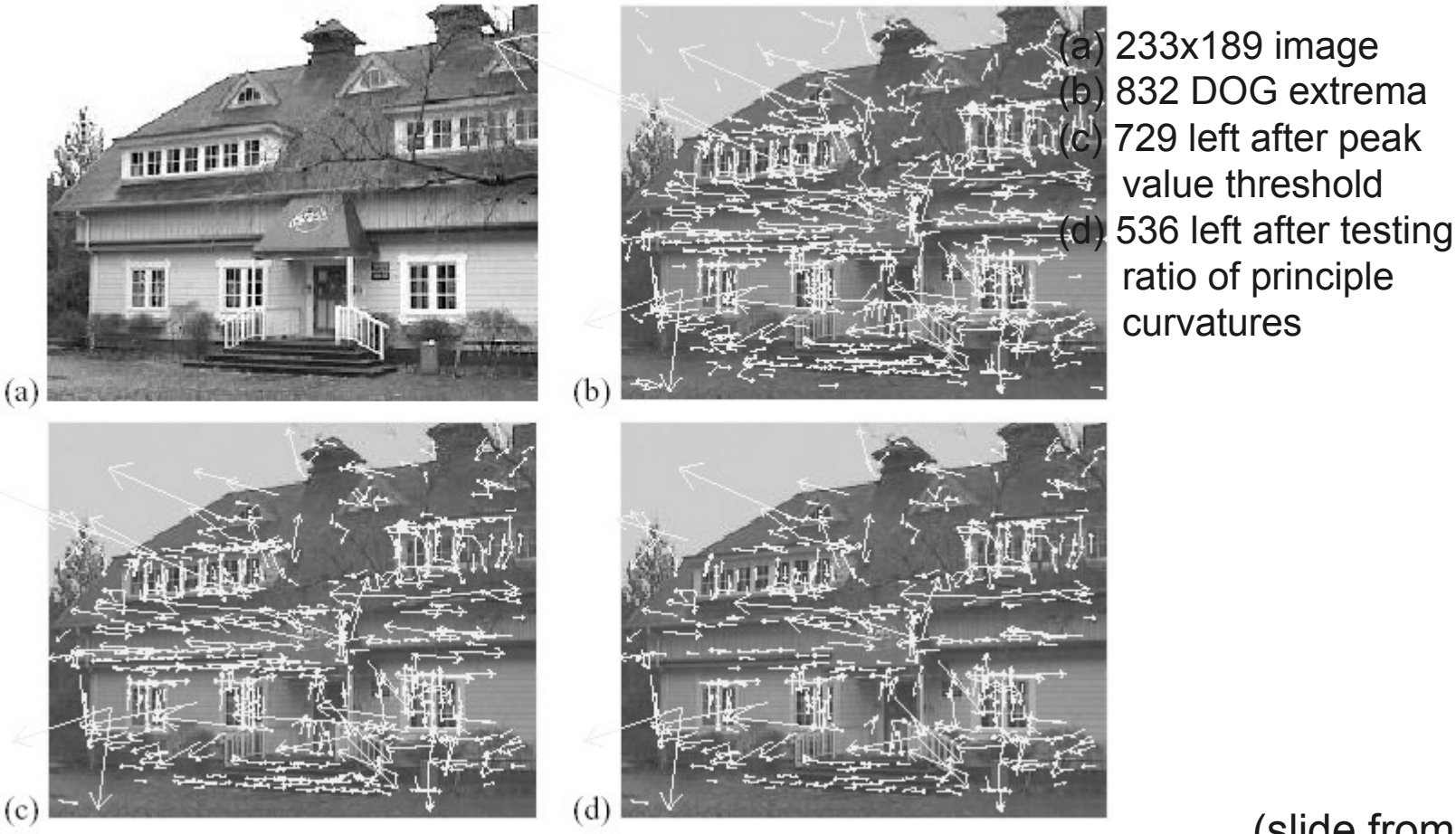




(slide from Lowe)

Example of keypoint detection

Threshold on value at DOG peak and on ratio of principle curvatures (Harris approach)



(a) 233x189 image
(b) 832 DOG extrema
(c) 729 left after peak value threshold
(d) 536 left after testing ratio of principle curvatures

(slide from Lowe)

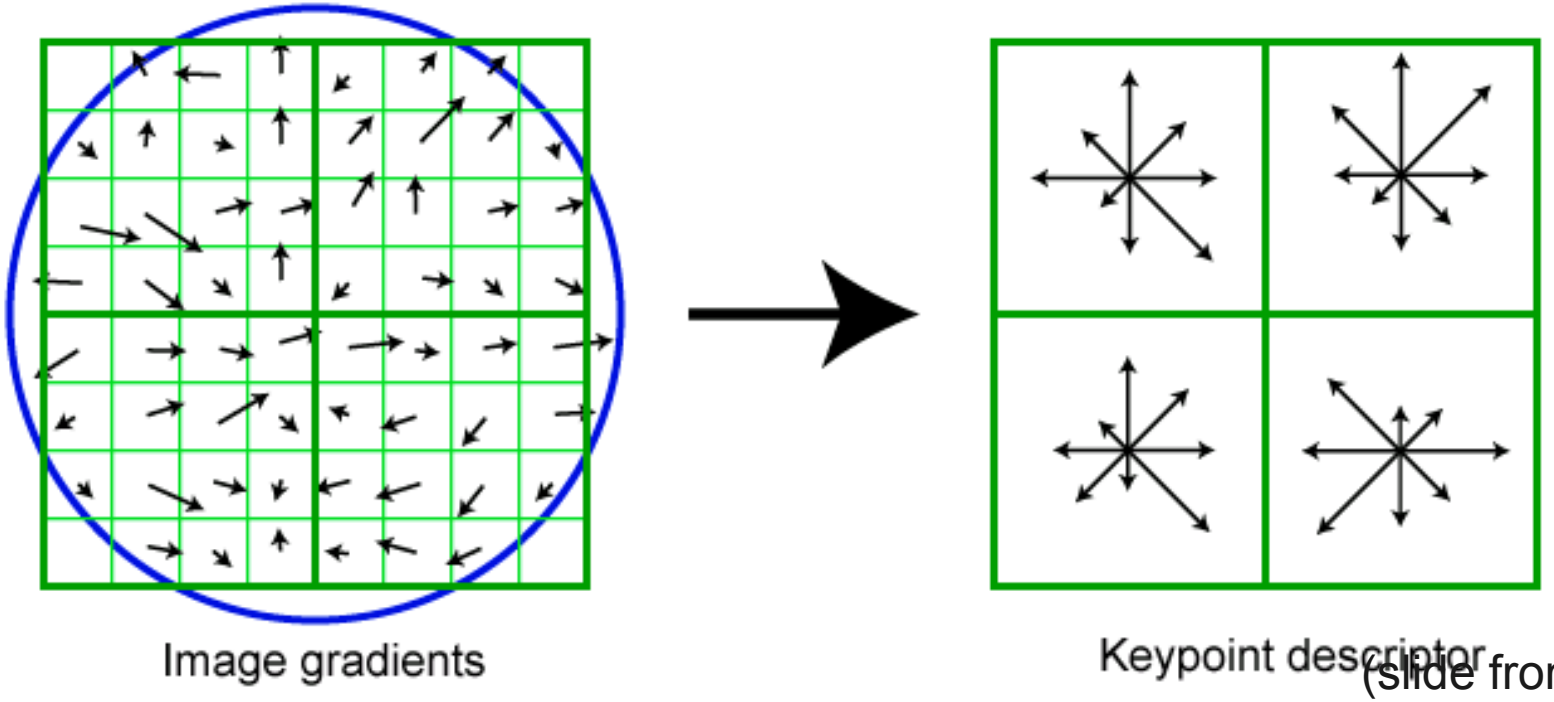
Detecting Keypoints is not always better

Descriptor	Grid	Random	Saliency [4]	DoG [7]
11×11 Pixel	64.0%	47.5%	45.5%	N/A
128-dim Sift	65.2%	60.7%	53.1%	52.5%

(From L. Fei-Fei and Perona)

SIFT vector formation

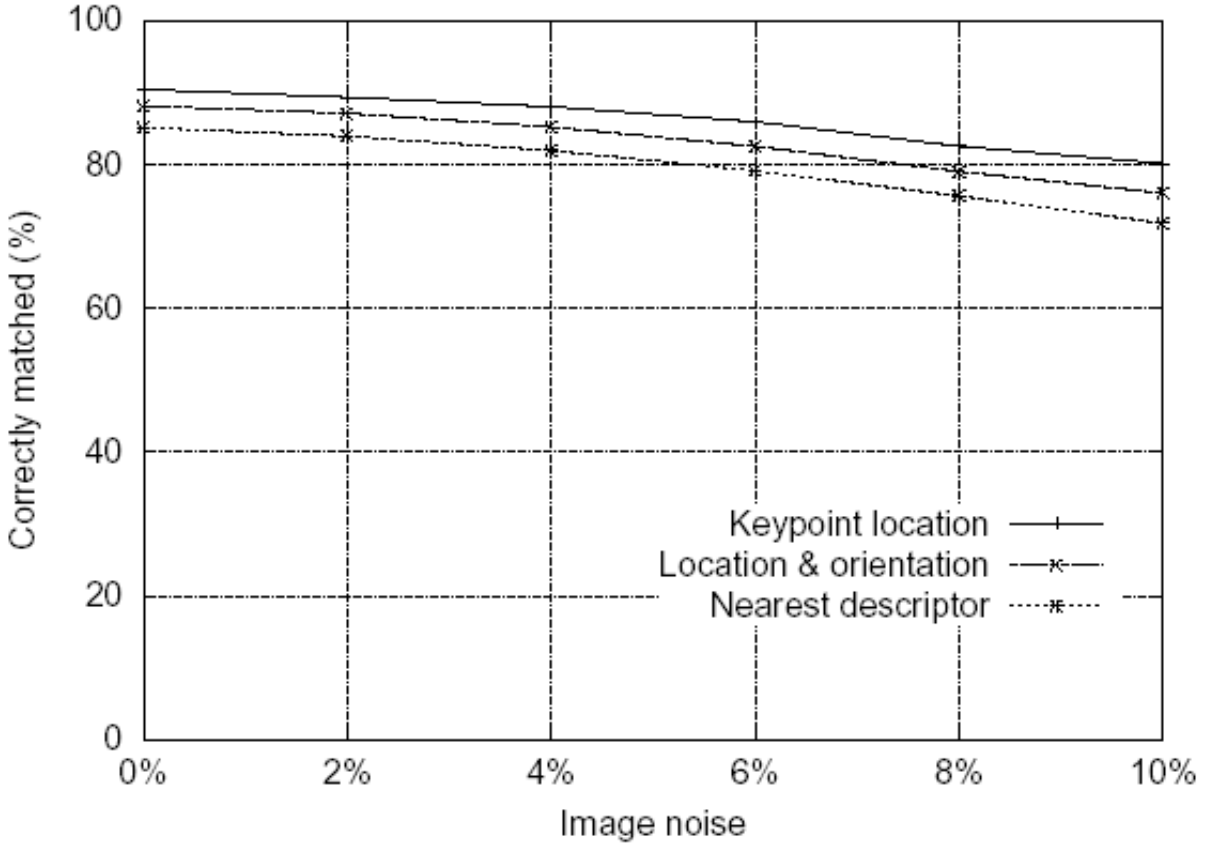
- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create array of orientation histograms
- 8 orientations x 4x4 histogram array = 128 dimensions



(slide from Lowe)

Feature stability to noise

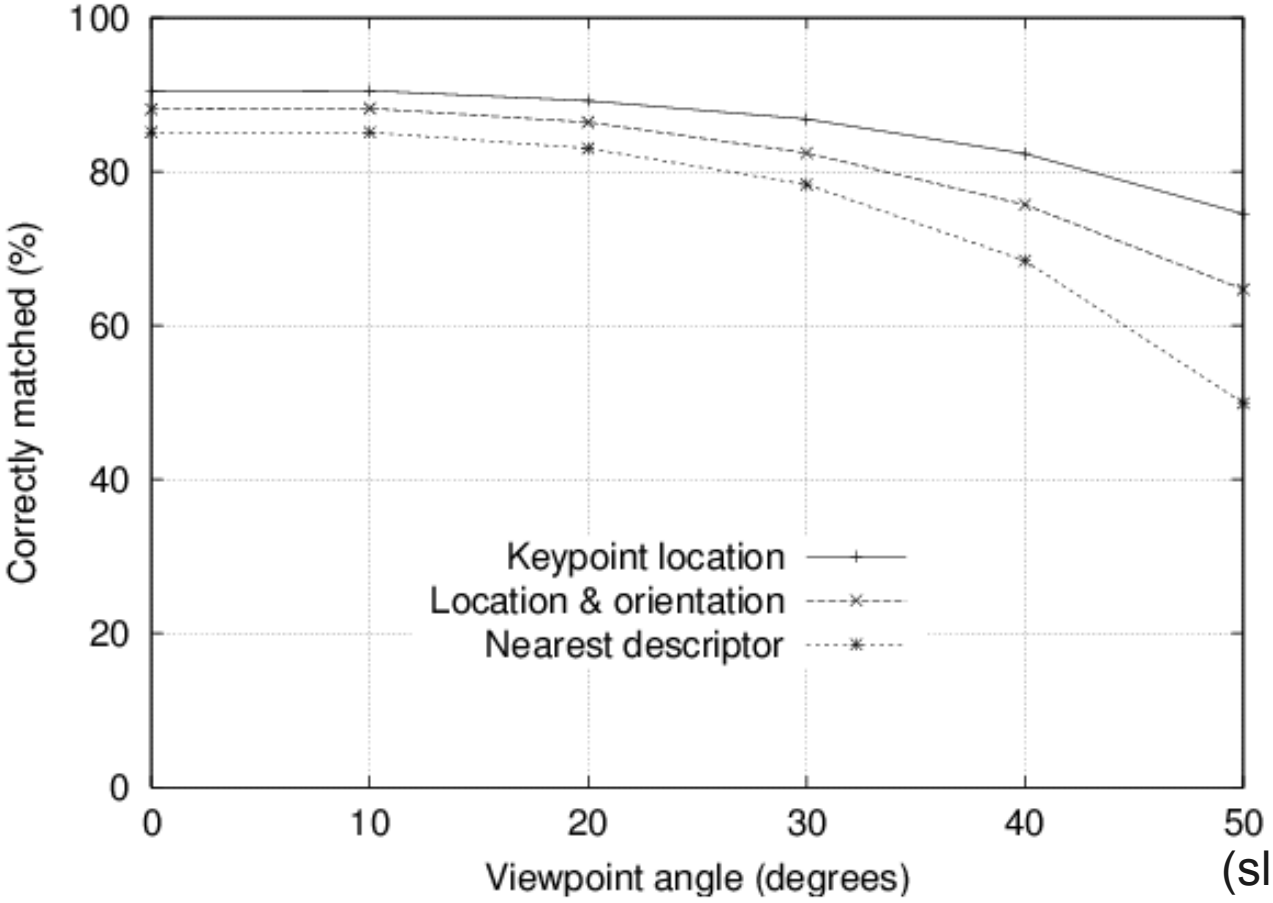
- Match features after random change in image scale & orientation, with differing levels of image noise
- Find nearest neighbor in database of 30,000 features



(slide from Lowe)

Feature stability to affine change

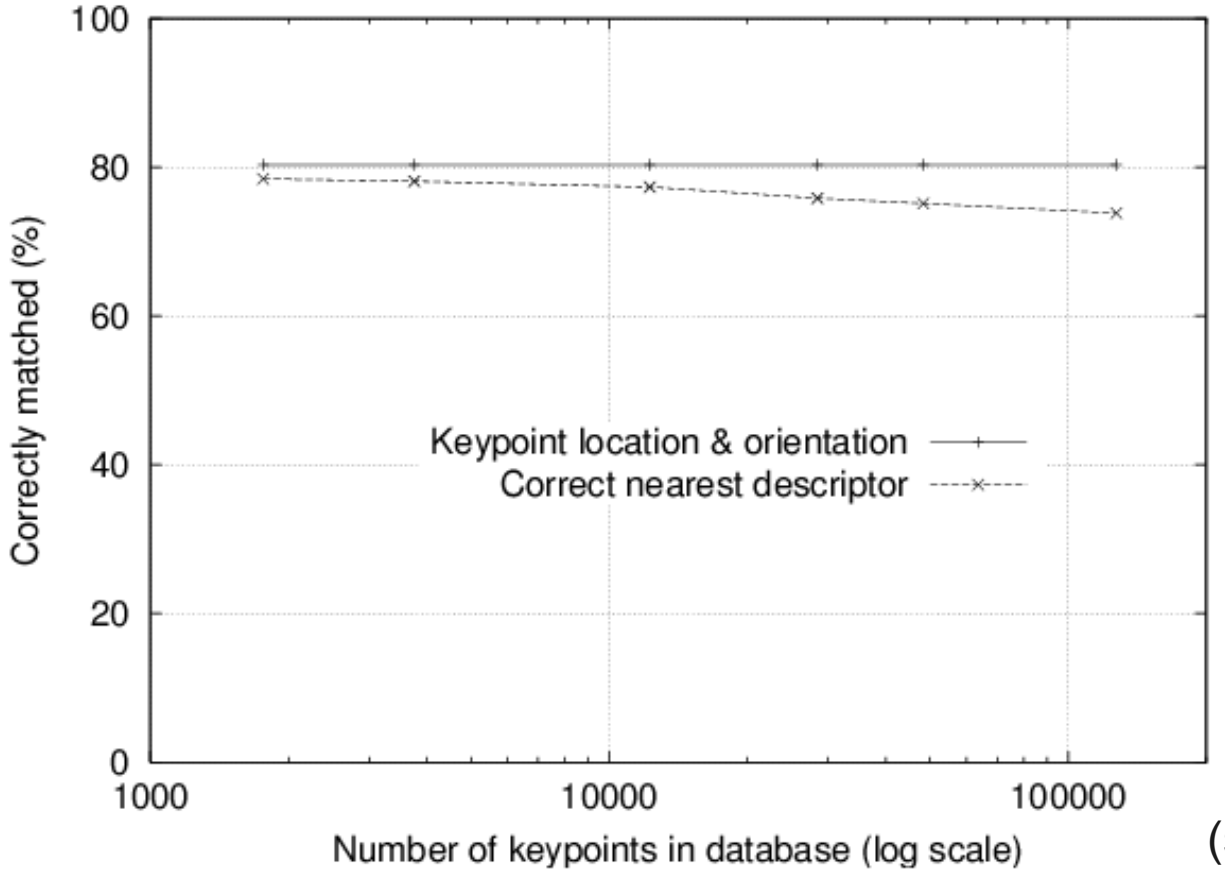
- Match features after random change in image scale & orientation, with 2% image noise, and affine distortion
- Find nearest neighbor in database of 30,000 features



(slide from Lowe)

Distinctiveness of features

- Vary size of database of features, with 30 degree affine change, 2% image noise
- Measure % correct for single nearest neighbor match



(slide from Lowe)

Sony Aibo (Evolution Robotics)

SIFT usage:

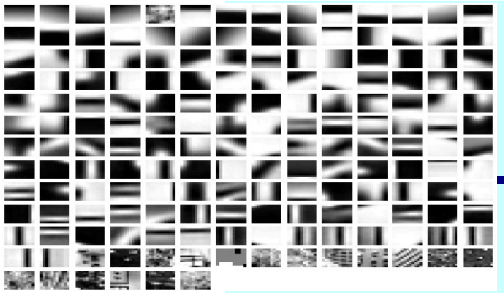
- Recognize charging station
- Communicate with visual cards

SIFT is just the beginning

- Authors have proposed more feature point detectors
  - Harris-Laplace,….
- Authors have proposed other feature descriptors
  - ColorSIFT
  - SURF
- The Koen executable implements many of this

Learning and Recognition

codewords dictionary

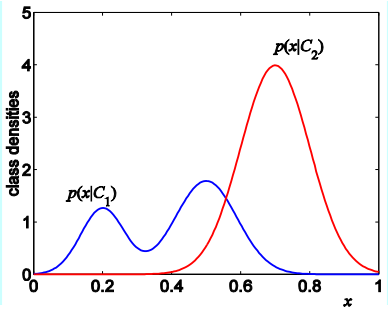


category models (and/or) classifiers

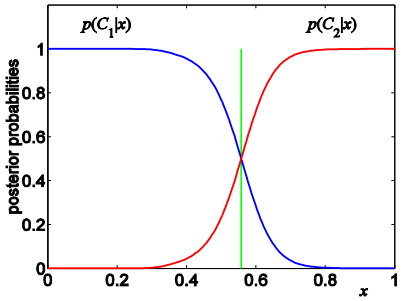
category decision

Learning and Recognition

1. Generative method:
   - graphical models



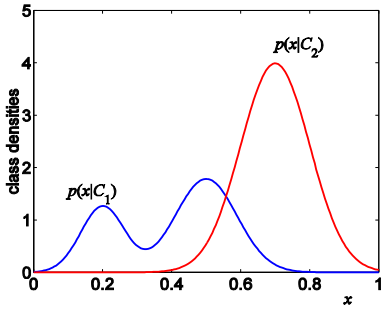
2. Discriminative method:
   - SVM



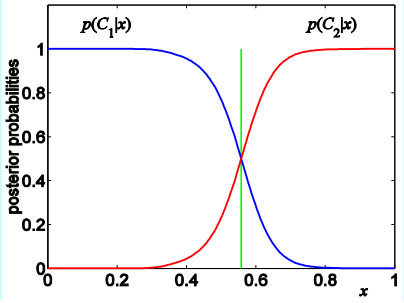
category models (and/or) classifiers

Learning and Recognition

1. Generative method:
   - graphical models

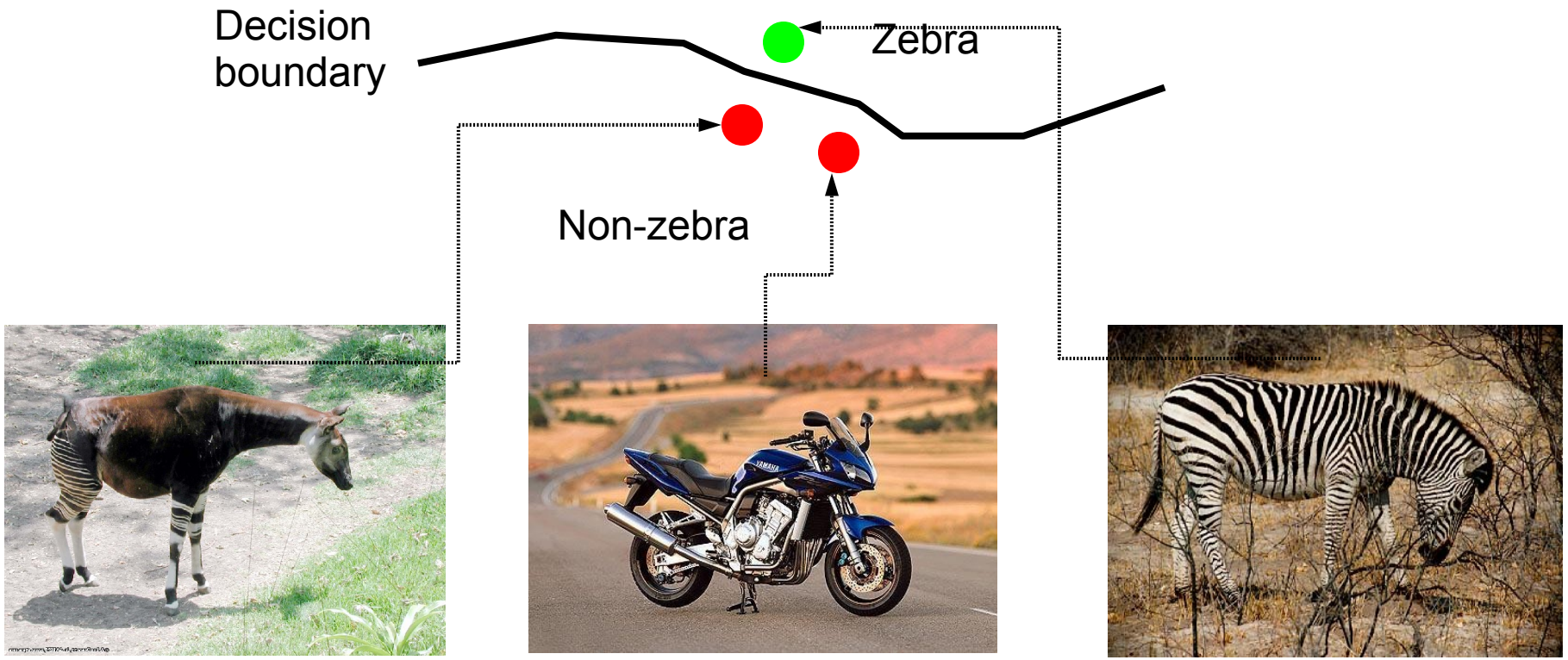


2. Discriminative method:
   - SVM



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:

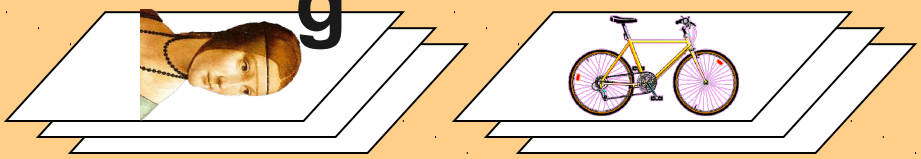
SVM w/ Pyramid Match kernels

Others

Csurka, Bray, Dance & Fan, 2004

Serre & Poggio, 2005

learnin
g



recognition

codewords dictionary

feature detection
& representation

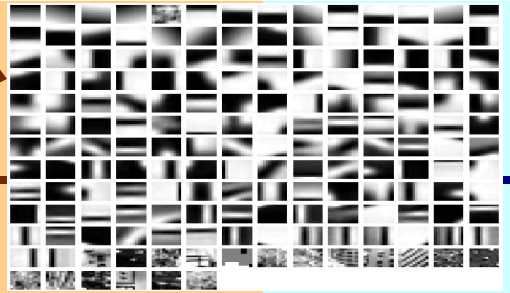
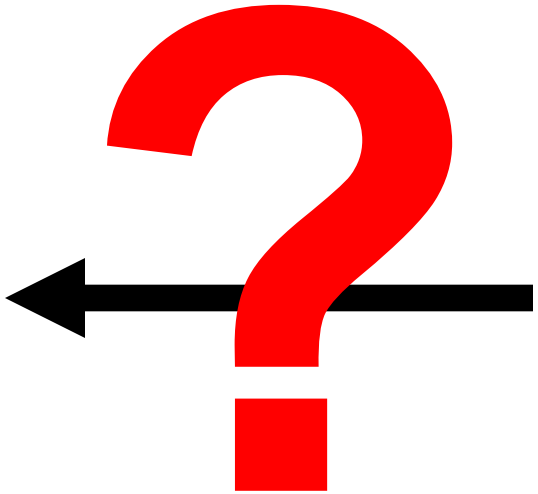


image representation

category models
(and/or) classifiers

category
decision

What about spatial info?



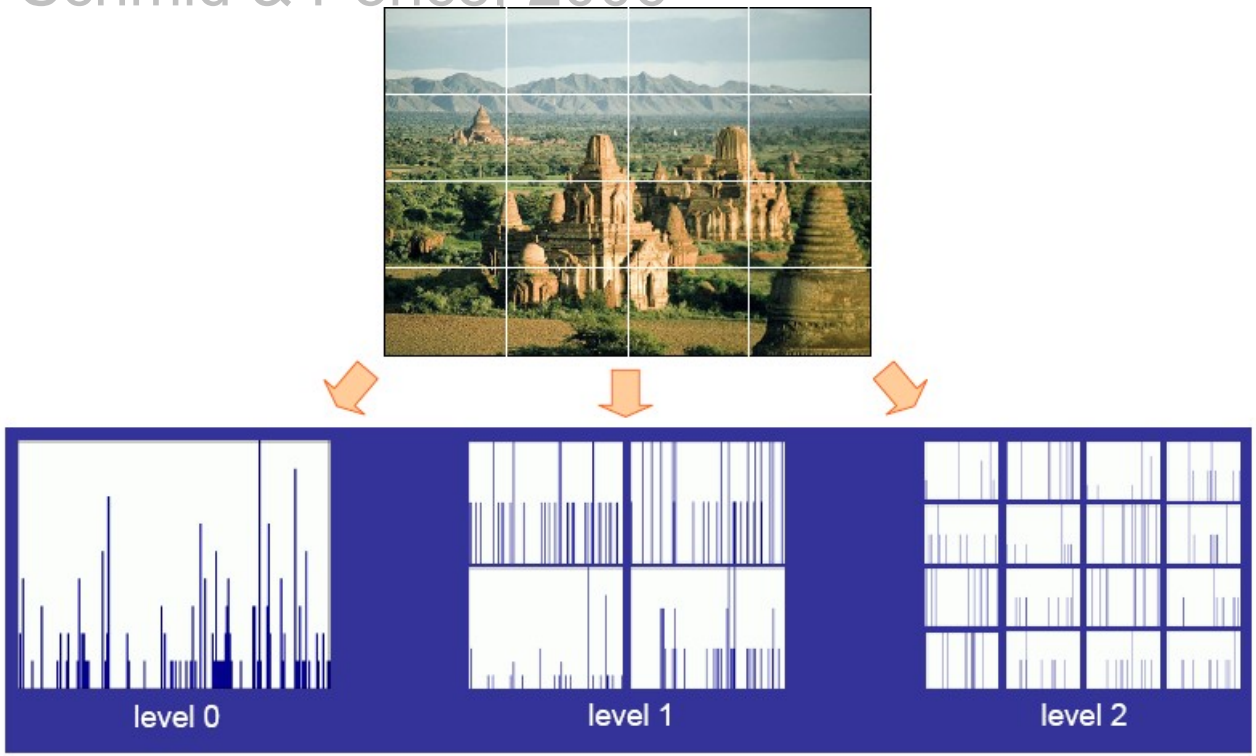
What about spatial info?

Feature level

Generative models

Discriminative methods

Lazebnik, Schmid & Ponce, 2006



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, 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

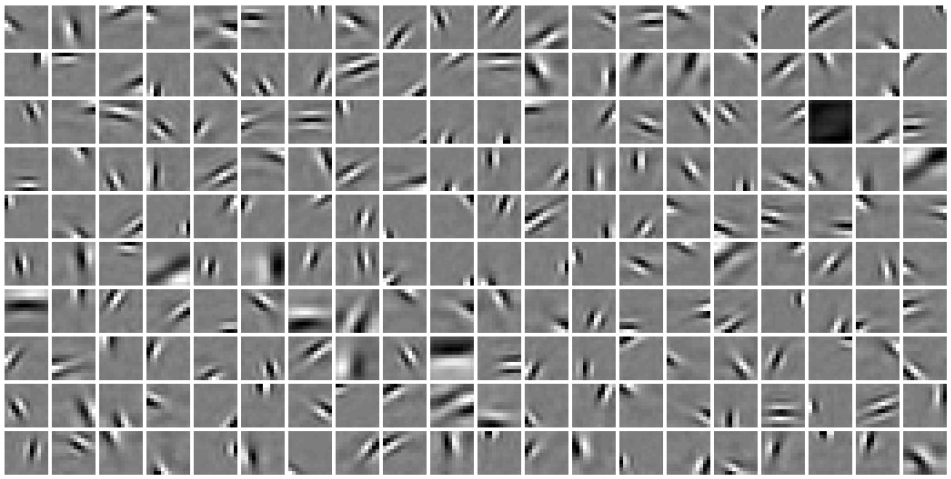
- Intuitive
  - Analogy to documents

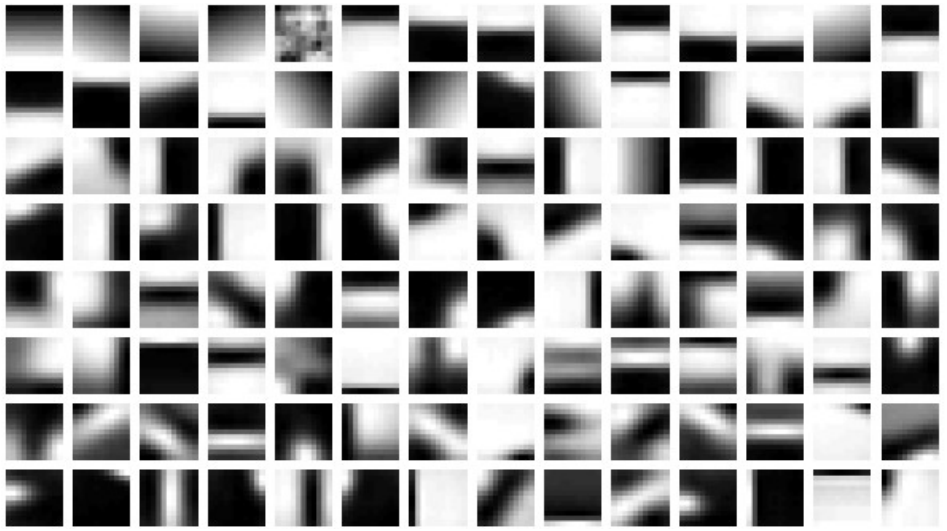
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that r… our eyes. For a long ti… retinal image wa… isual centers i… s a movie… image … discove… know th… perceptio… y more com… following the… ath to the various c… cortex, Hubel and Wiesel ha… demonstrate that the *message abo… image falling on the retina undergoes … wise analysis in a system of nerve cel… stored in columns. In this system each … has its specific function and is responsible … a specific detail in the pattern of the retinal … image.*

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

Model properties

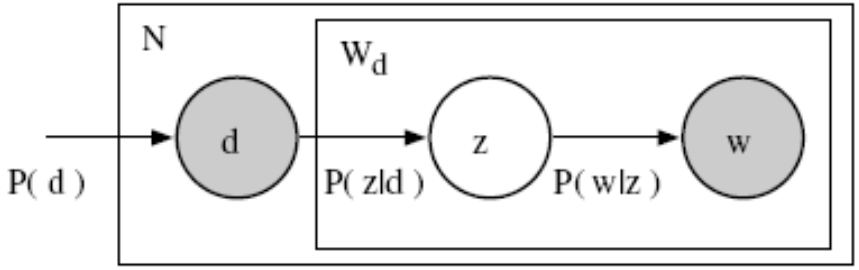
- Intuitive
  - Analogy to documents
  - Analogy to human vision





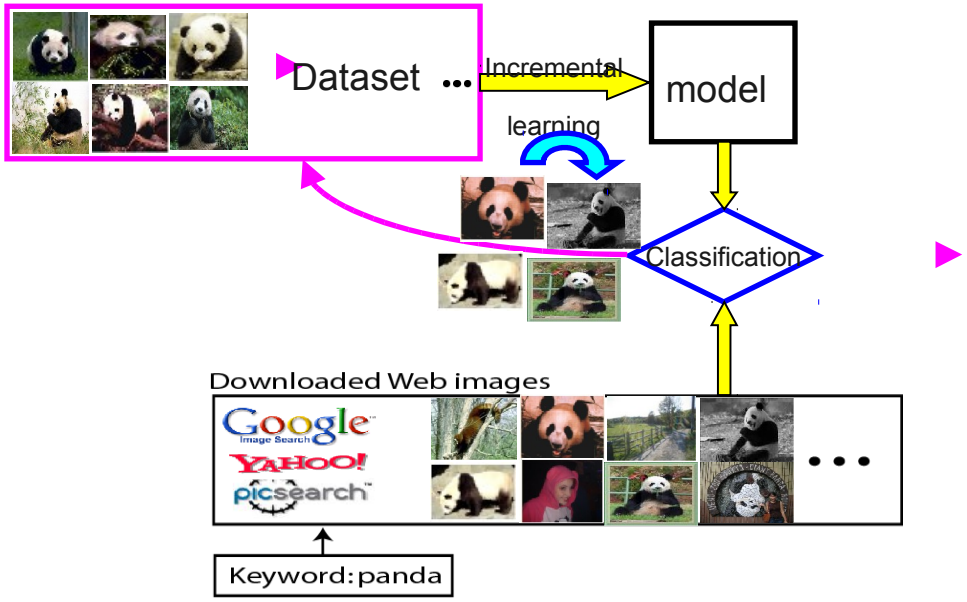
Olshausen and Field, 2004, Fei-Fei and Perona, 2005

Model properties



Sivic, Russell, Efros, Freeman, Zisserman, 2005

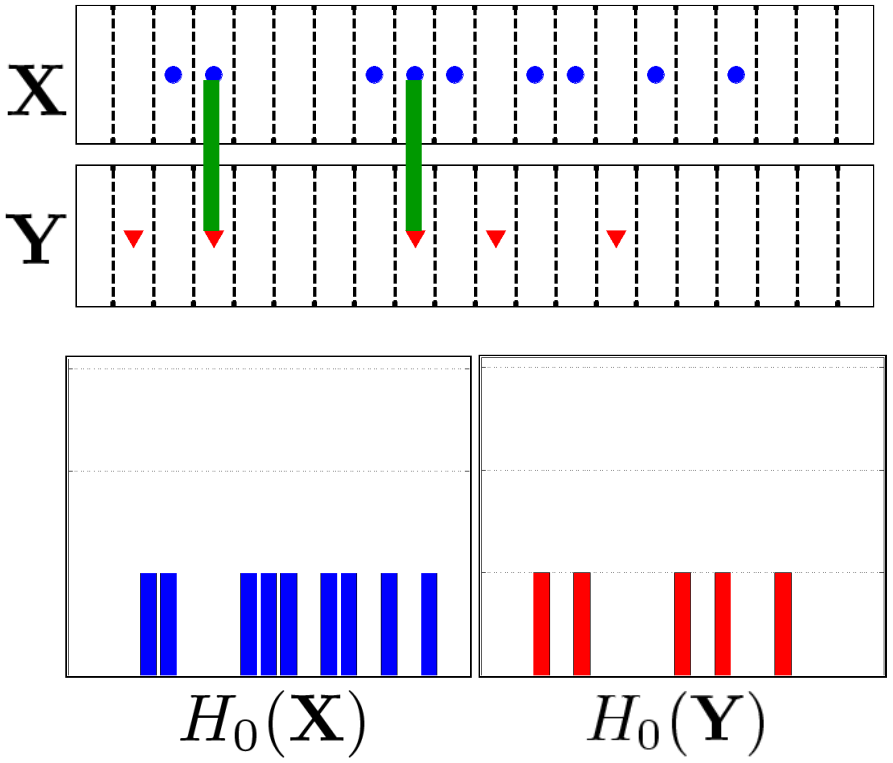
- Intuitive
- generative models
  - Convenient for weakly- or un-supervised, incremental training
  - Prior information
  - Flexibility (e.g. HDP)



Li, Wang & Fei-Fei, CVPR 2007

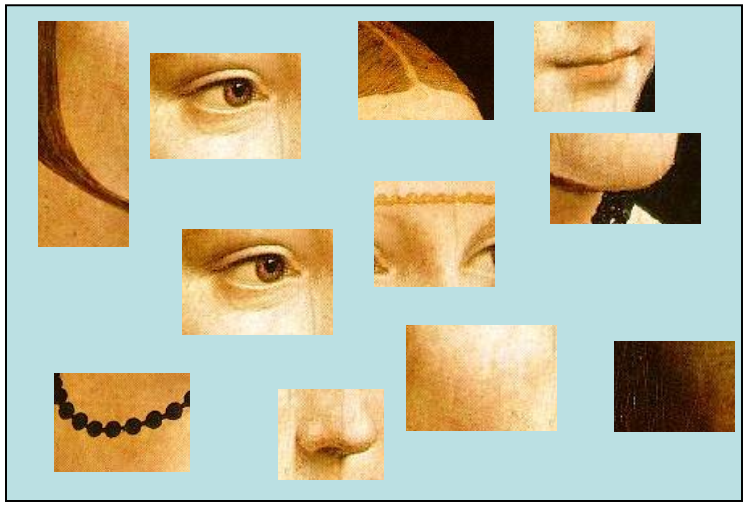
Model properties

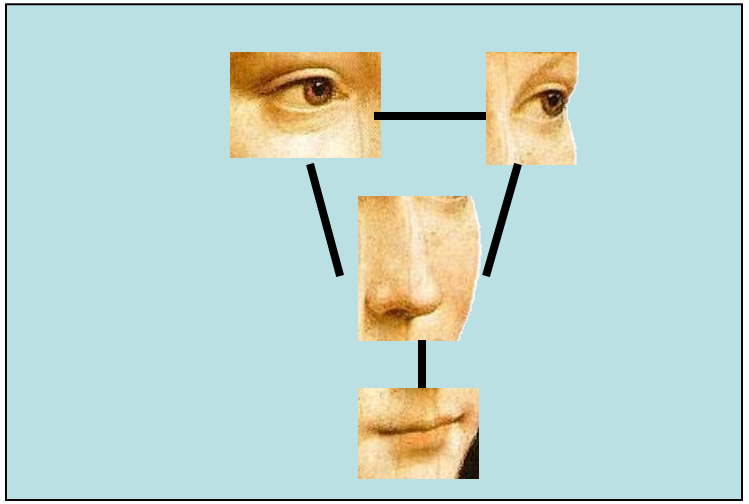
- Intuitive
- generative models
- Discriminative method
  - Computationally efficient



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