“Bag of Words”: when is object recognition, just texture recognition?

A quiet meditation on the importance of trying simple things first…

16-721: Advanced Machine Perception
A. Efros, CMU, Spring 2009

Adopted from Fei-Fei Li, S.c. Zhu, and L.Walker Renninger

What is Texture?

Texture depicts spatially repeating patterns

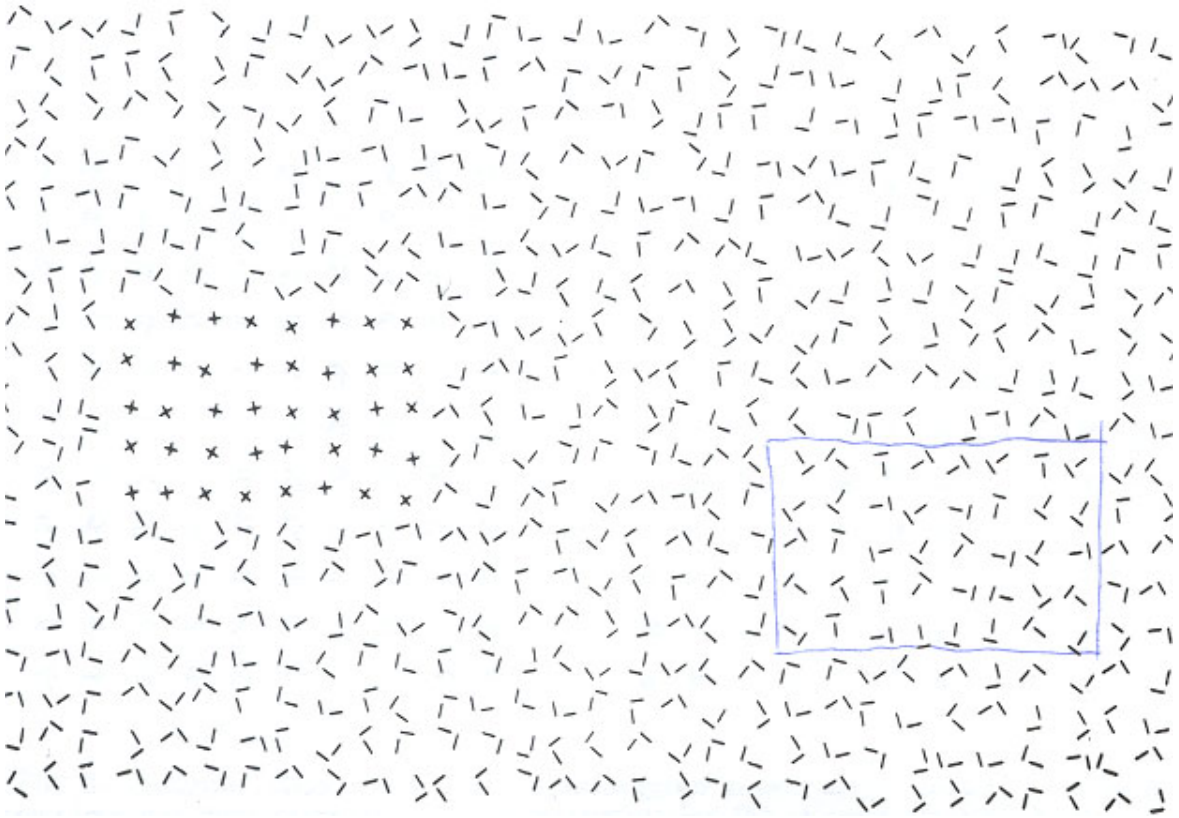
Many natural phenomena are textures

radishes

rocks

yogurt

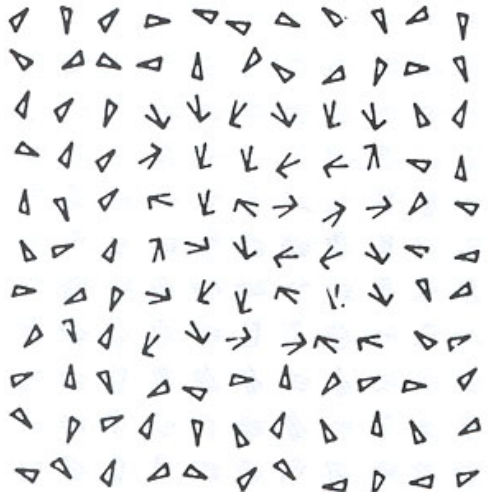
Texton Discrimination (Julesz)

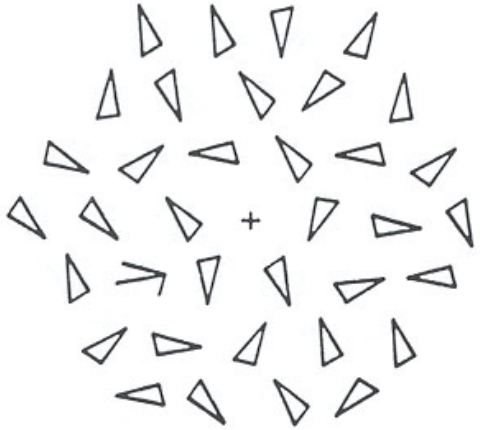


Human vision is sensitive to the difference of some types of elements and appears to be “numb” on other types of differences.

Search Experiment I

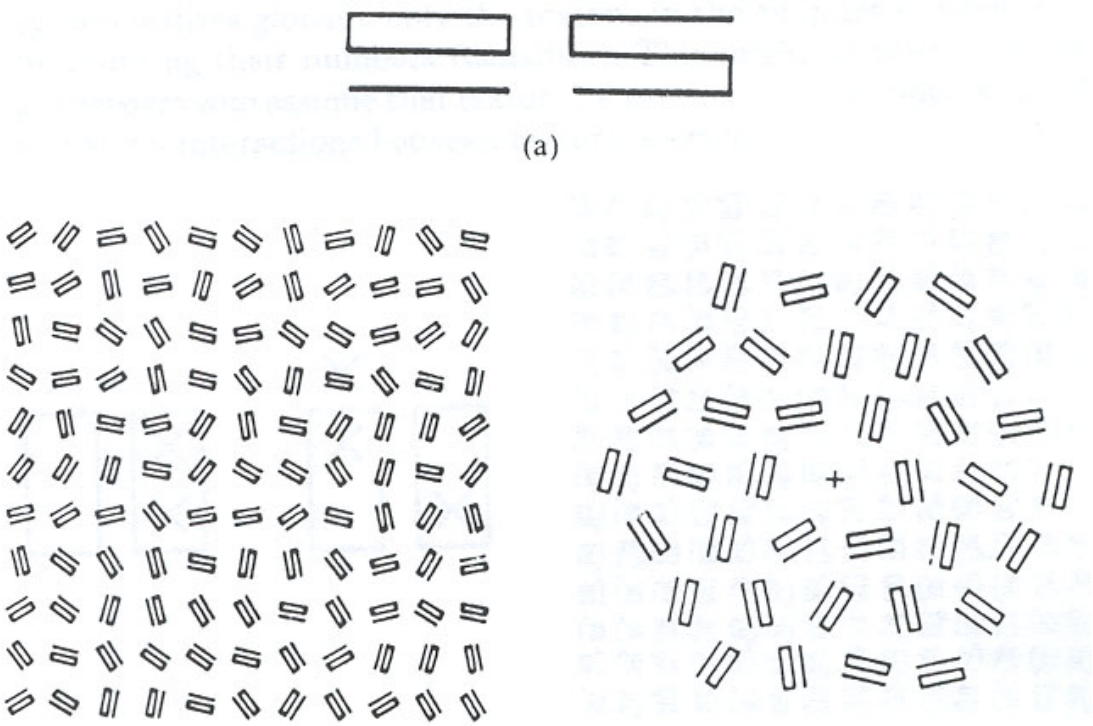
(a)





The subject is told to detect a target element in a number of background elements.
In this example, the detection time is independent of the number of background elements.

Search Experiment II



In this example, the detection time is proportional to the number of background elements, And thus suggests that the subject is doing element-by-element scrutiny.

Heuristic (Axiom) I

Julesz then conjectured the following axiom:

Human vision operates in two distinct modes:

1. Preattentive vision

   parallel, instantaneous *(~100--200ms)*, without scrutiny, independent of the number of patterns, covering a large visual field.

2. Attentive vision

   serial search by focal attention in 50ms steps limited to small aperture.

Then what are the basic elements?

Heuristic (Axiom) II

Julesz's second heuristic answers this question:

Textons are the fundamental elements in preattentive vision, including

1. Elongated blobs

   rectangles, ellipses, line segments with attributes

   color, orientation, width, length, flicker rate.

2. Terminators

   ends of line segments.

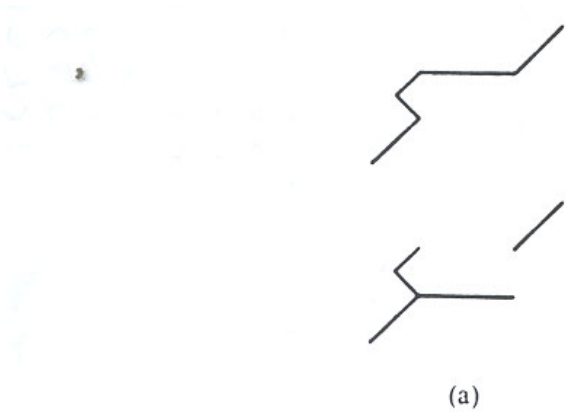
3. Crossings of line segments.

But it is worth noting that Julesz's conclusions are largely based by ensemble of artificial texture patterns. It was infeasible to synthesize natural textures for controlled experiments at that time.

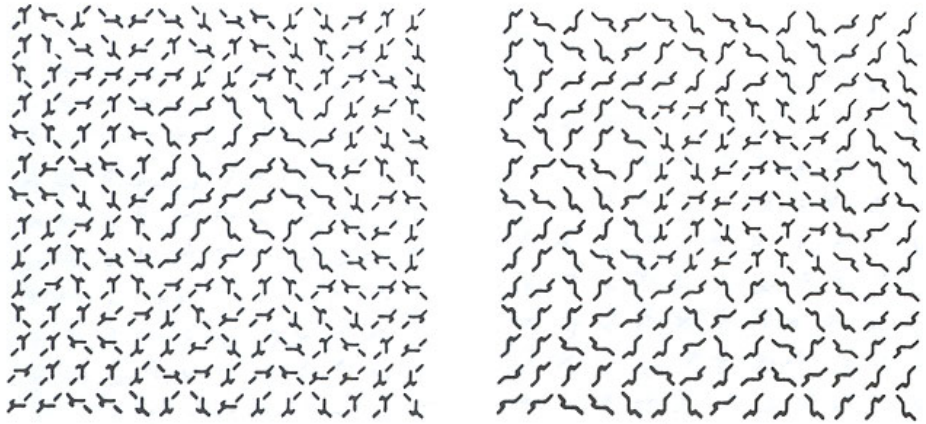
Examples

Pre-attentive vision is sensitive to size/width, orientation changes

Examples



(a)



Sensitive to number
of terminators

Left: fore-back
Right: back-fore

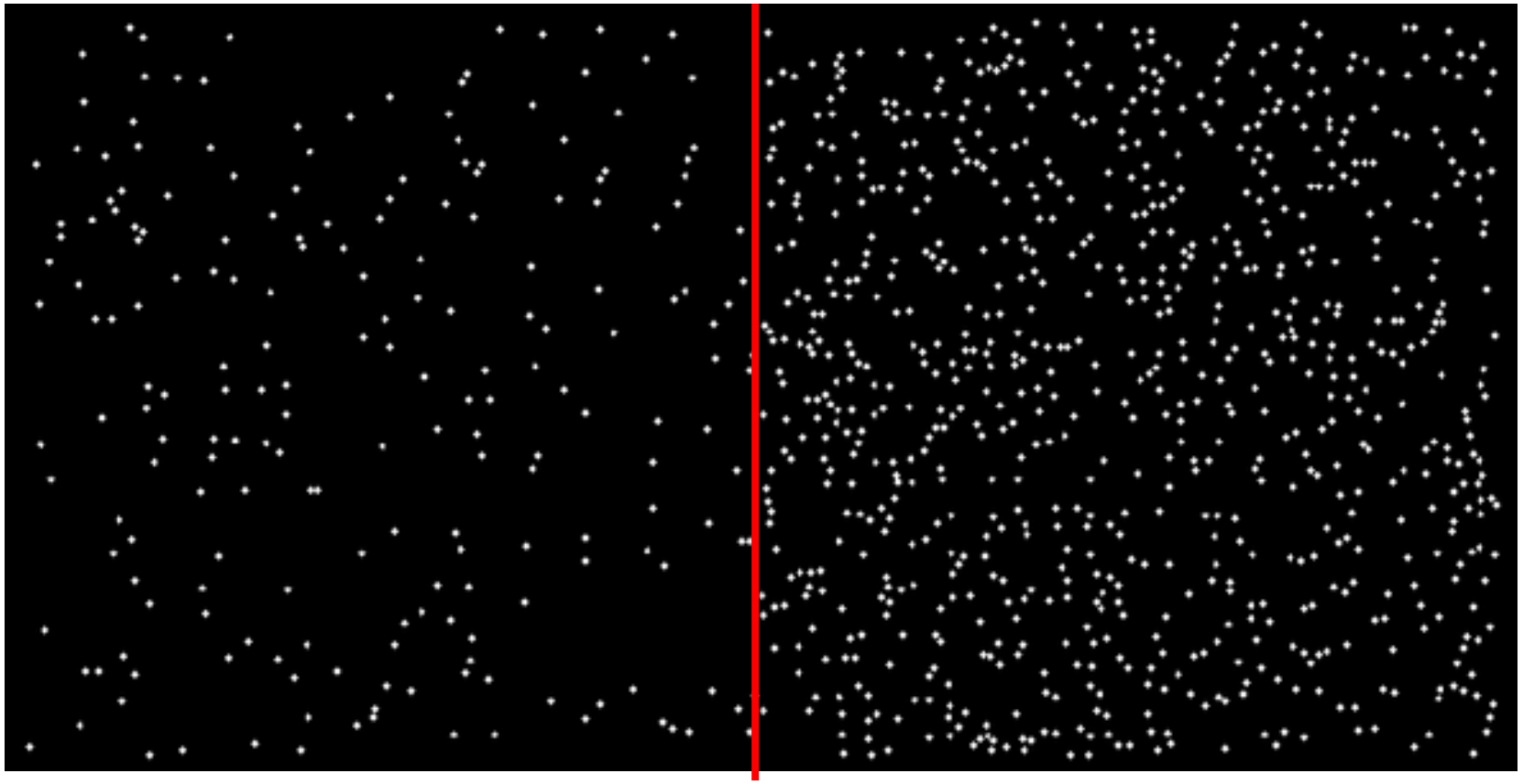
See previous examples
For cross and terminators

Julesz Conjecture

Textures cannot be spontaneously discriminated if they have the same first-order and second-order statistics and differ only in their third-order or higher-order statistics.

(later proved wrong)

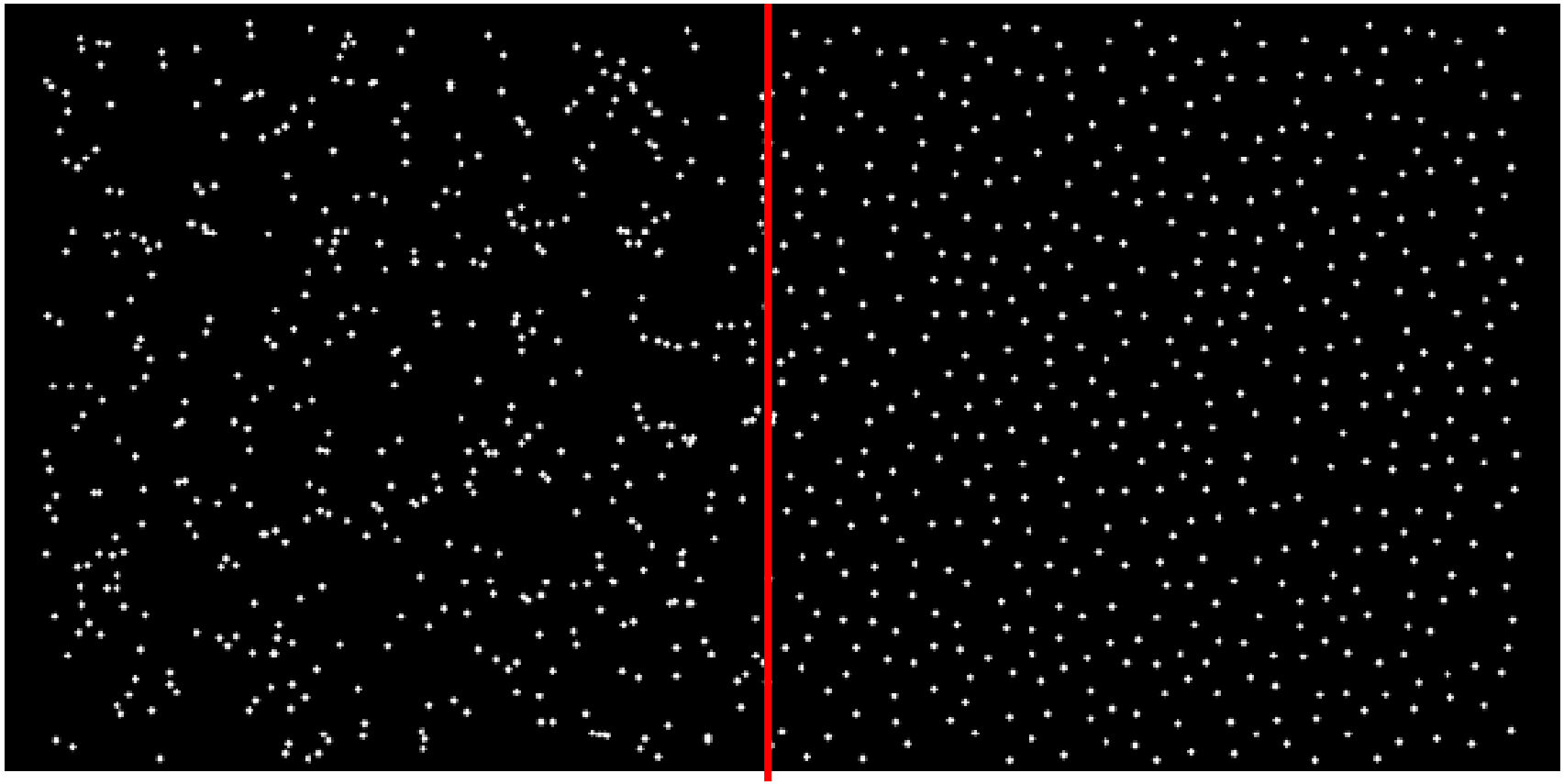
1st Order Statistics



5% white

20% white

2nd Order Statistics



10% white

Capturing the “essence” of texture

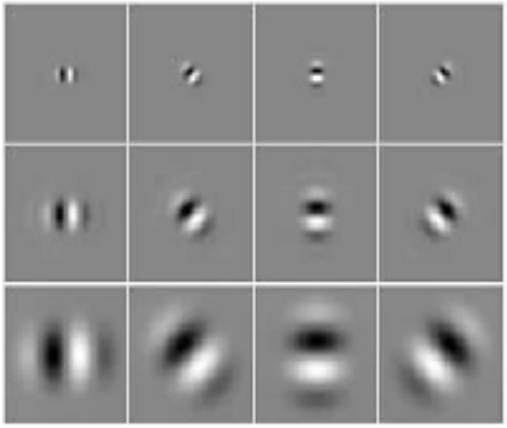
…for real images

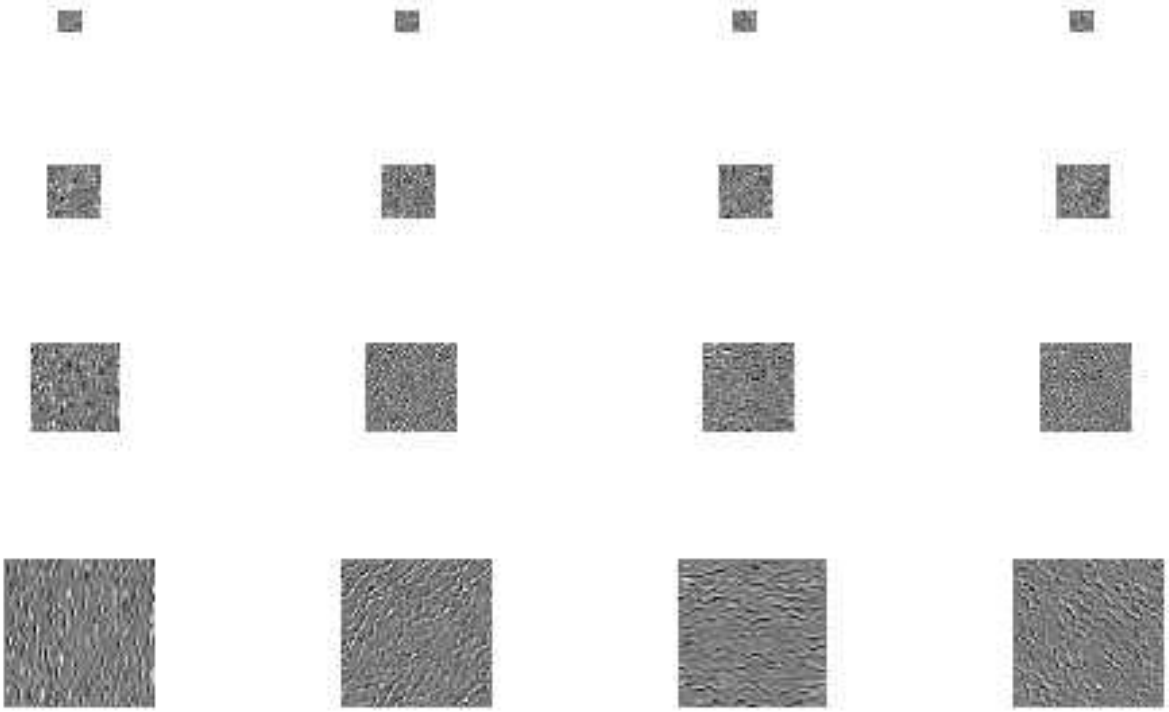
We don’t want an actual texture realization, we want a texture invariant

What are the tools for capturing <u>statistical</u> properties of some signal?

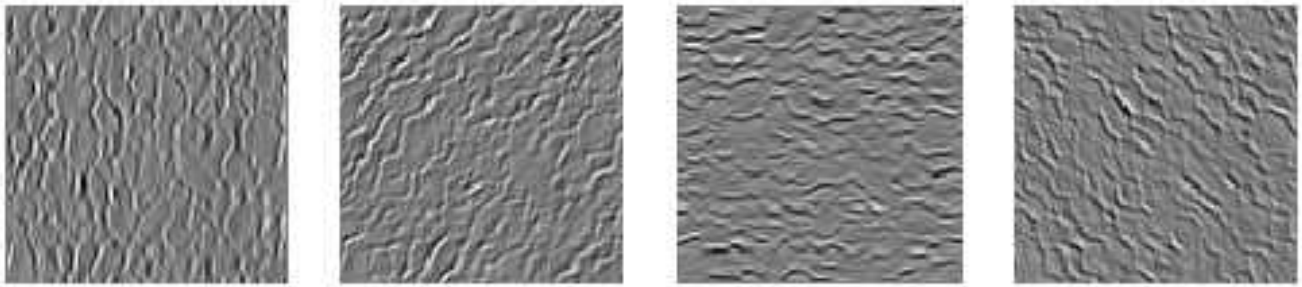
Multi-scale filter decomposition

Filter bank

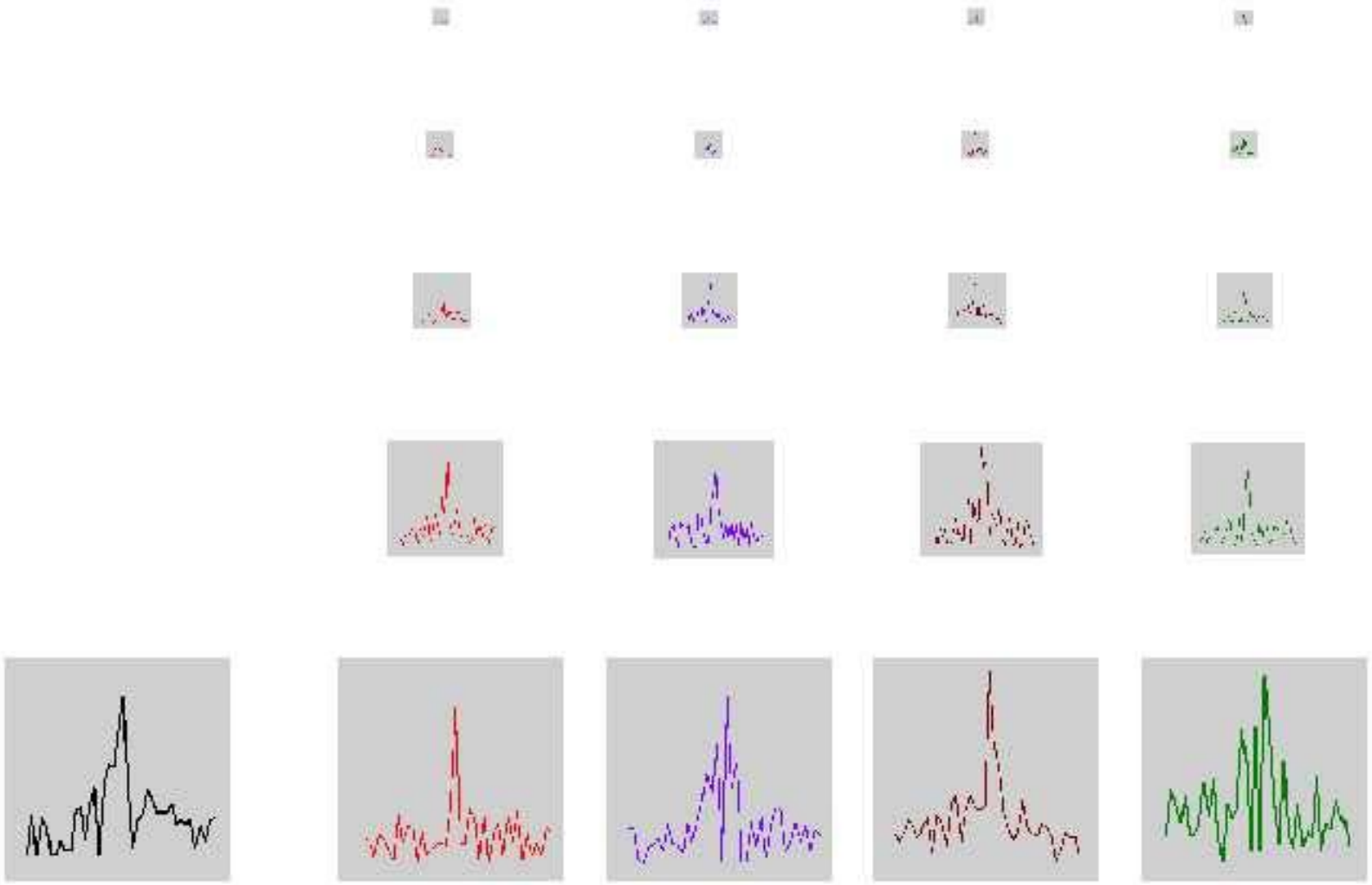




Input image



Filter response histograms



Heeger & Bergen ‘95

Start with a noise image as output

Main loop:

- Match pixel histogram of output image to input
- Decompose input and output images using multi-scale filter bank (Steerable Pyramid)
- Match subband histograms of input and output pyramids
- Reconstruct input and output images (collapse the pyramids)

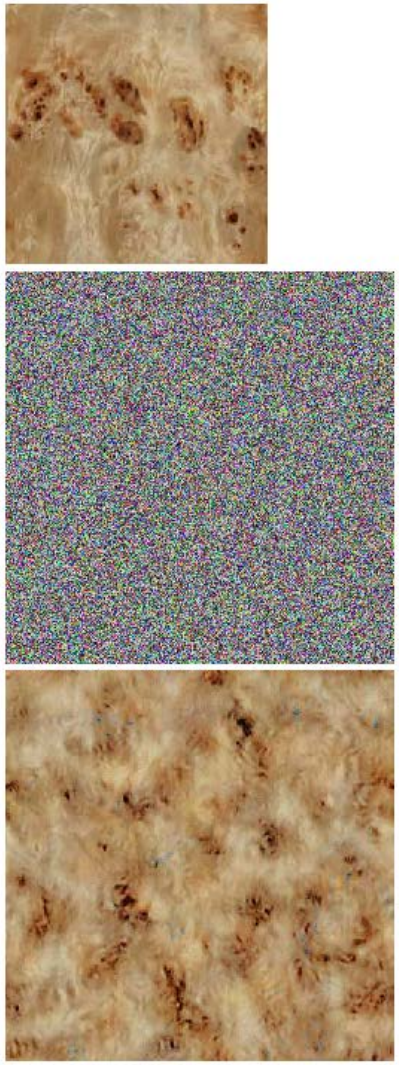
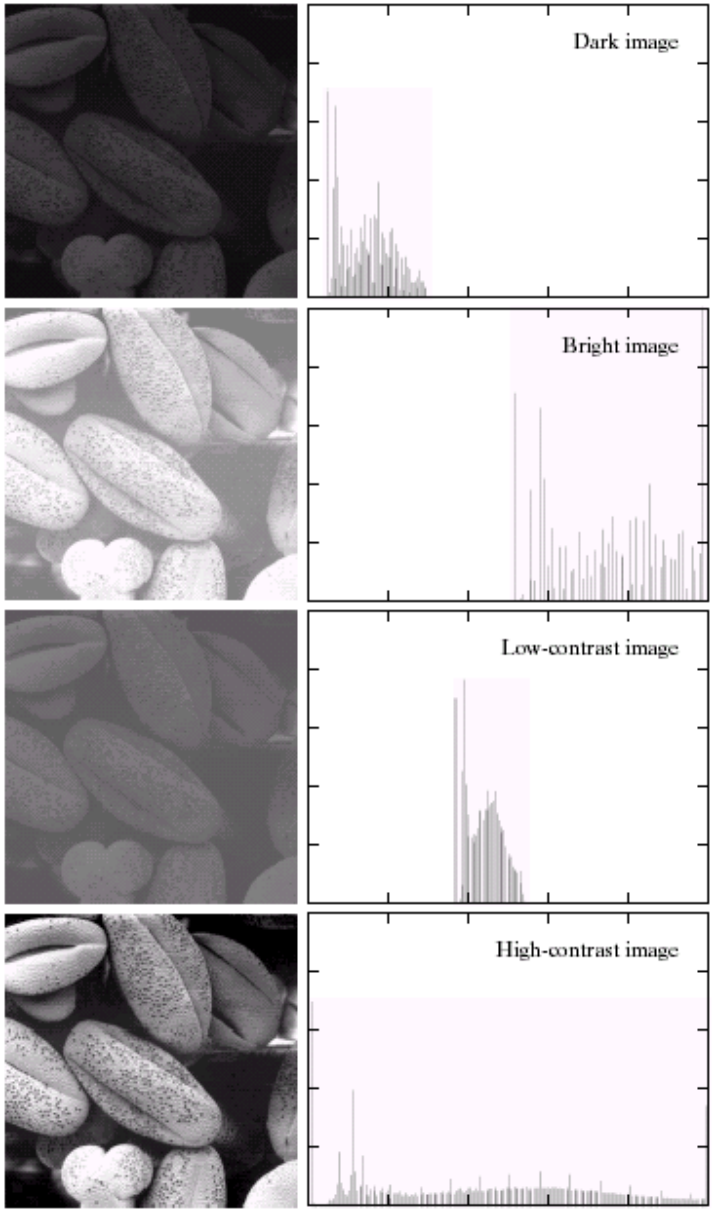
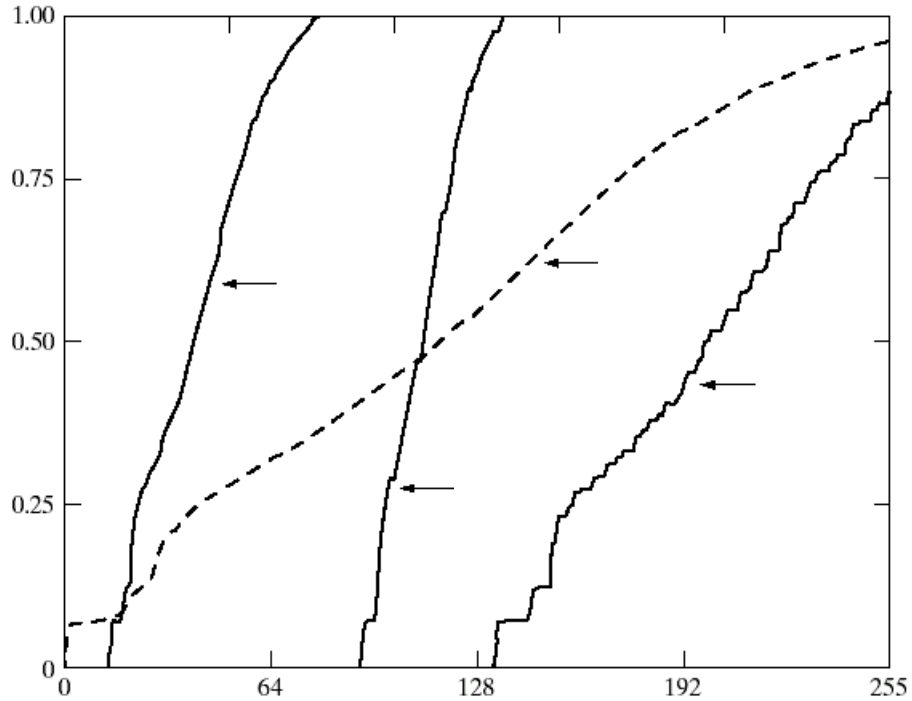


Image Histograms





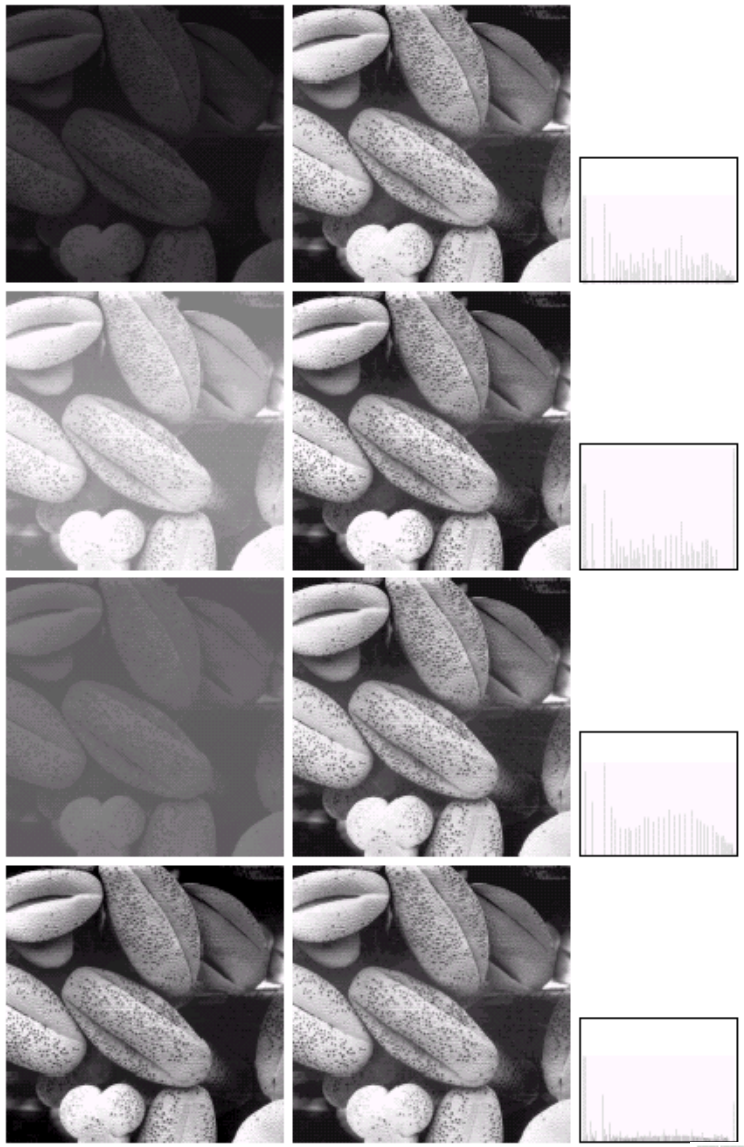
Cumulative Histograms

$s = T(r)$

a b

FIGURE 3.15 Four basic image types: dark, light, low contrast, high contrast, and their corresponding histograms. (Original image courtesy of Dr. Roger Heady, Research School of Biological Sciences, Australian National University, Canberra, Australia.)

Histogram Equalization



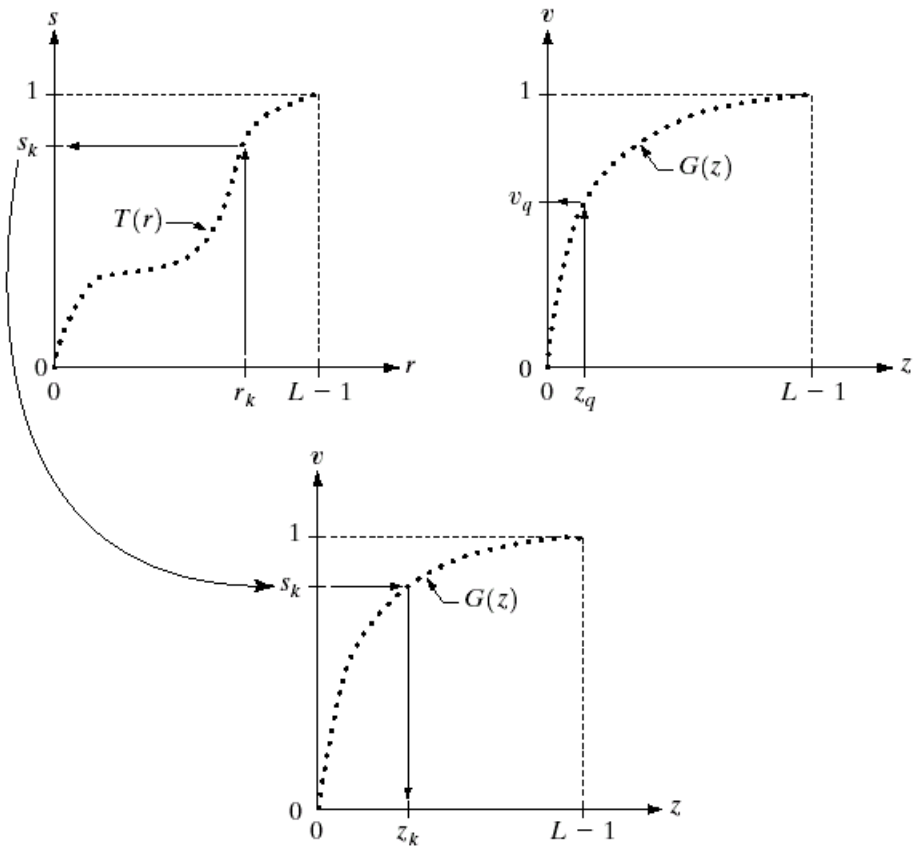
a b c

FIGURE 3.17 (a) Images from Fig. 3.15. (b) Results of histogram equalization. (c) Corresponding histograms.

Histogram Matching

a b
c

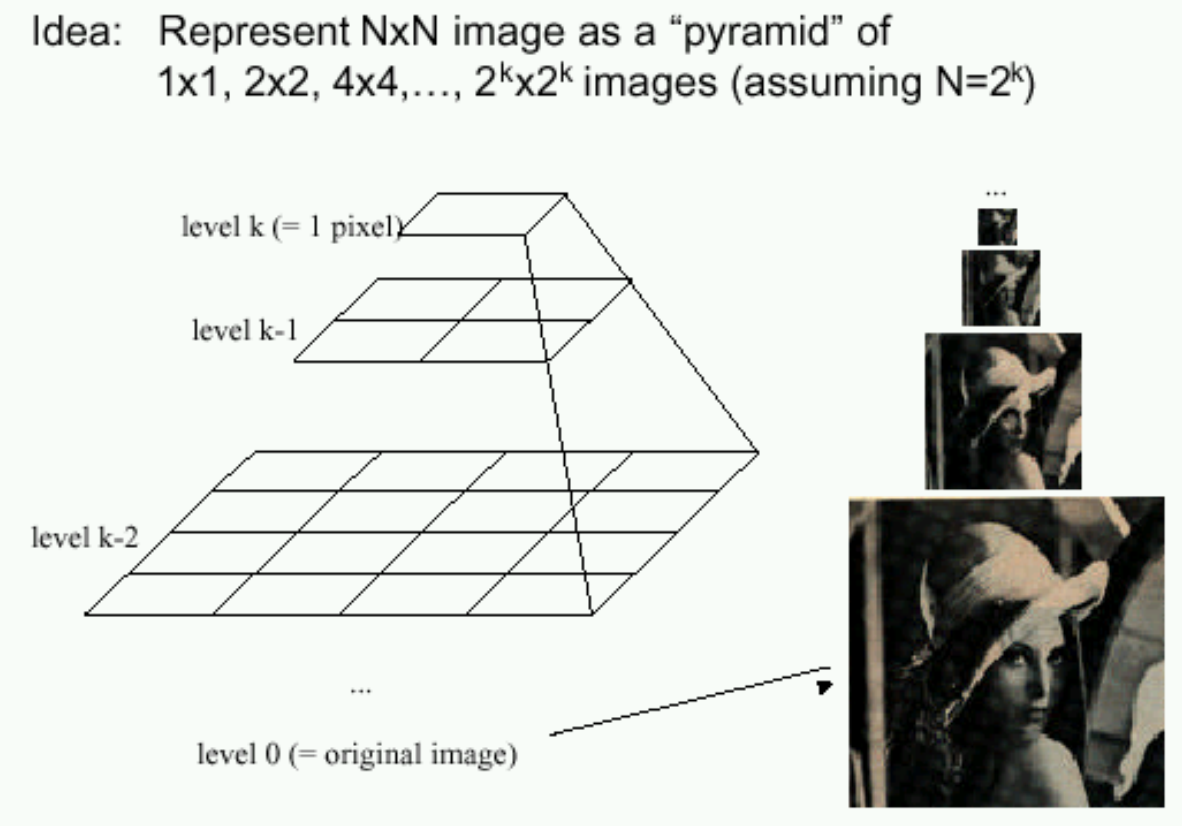
FIGURE 3.19
(a) Graphical interpretation of mapping from r_k to s_k via $T(r)$. (b) Mapping of z_q to its corresponding value v_q via $G(z)$. (c) Inverse mapping from s_k to its corresponding value of z_k.



Match-histogram code

```
Match-histogram (im1,im2)
  im1-cdf = Make-cdf(im1)
  im2-cdf = Make-cdf(im2)
  inv-im2-cdf = Make-inverse-lookup-table(im2-cdf)
  Loop for each pixel do
       im1[pixel] =
         Lookup(inv-im2-cdf,
                Lookup(im1-cdf,im1[pixel]))
```

Image Pyramids



Known as a **Gaussian Pyramid** [Burt and Adelson, 1983]

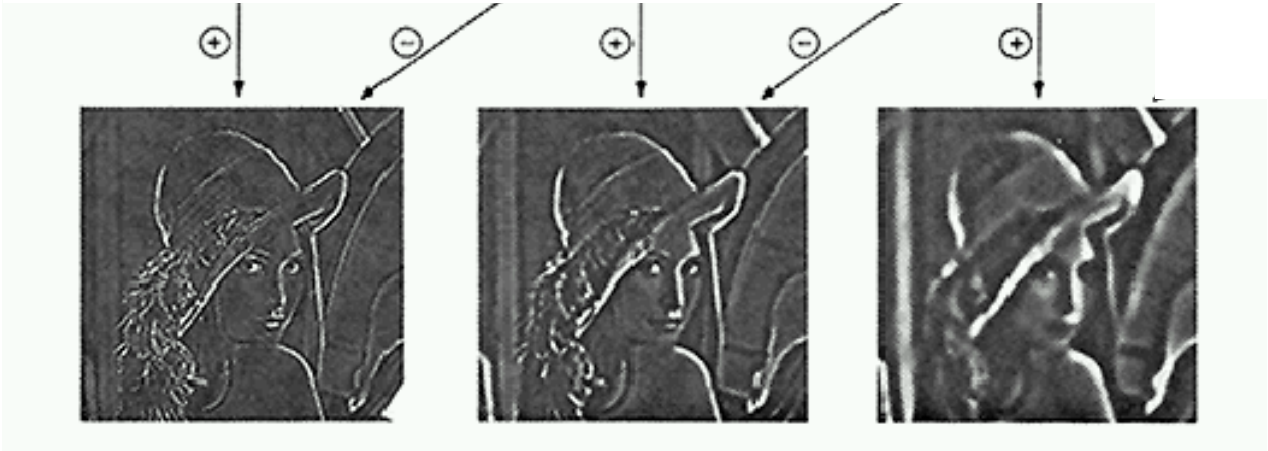
- In computer graphics, a *mip map* [Williams, 1983]
- A precursor to *wavelet transform*

Band-pass filtering

Gaussian Pyramid (low-pass images)

Laplacian Pyramid

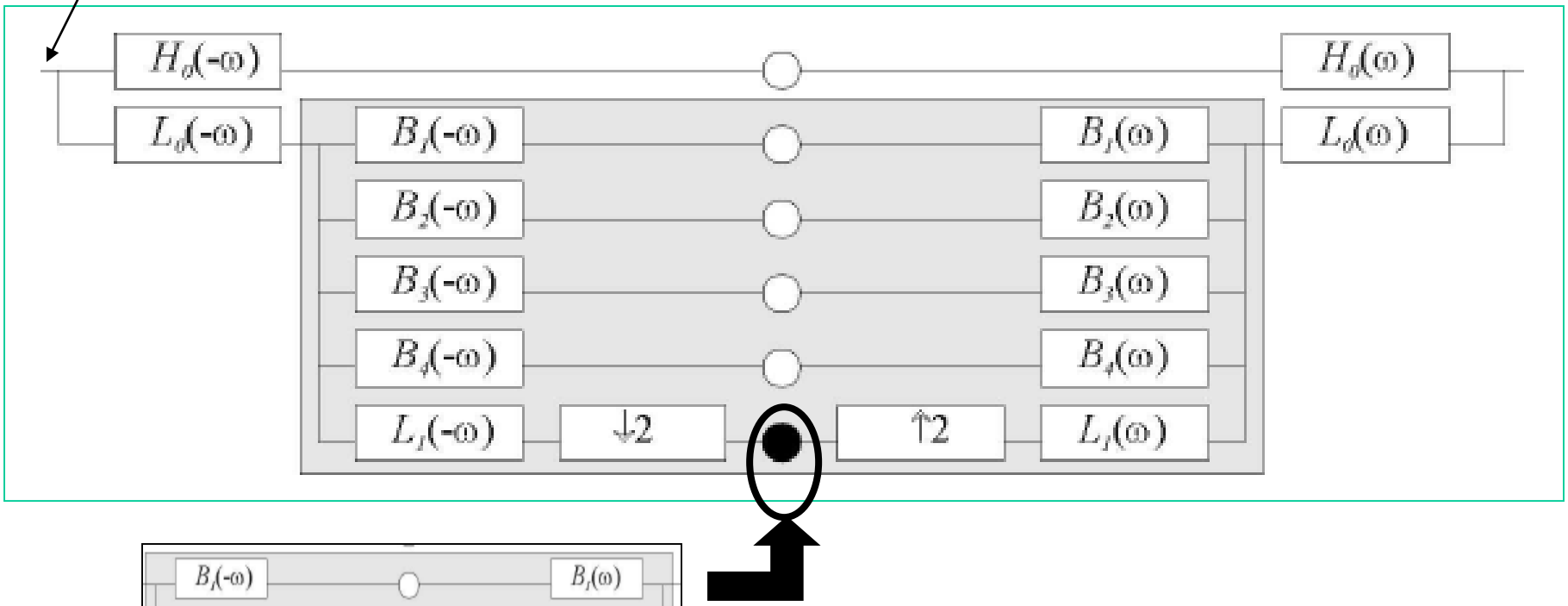
Original image

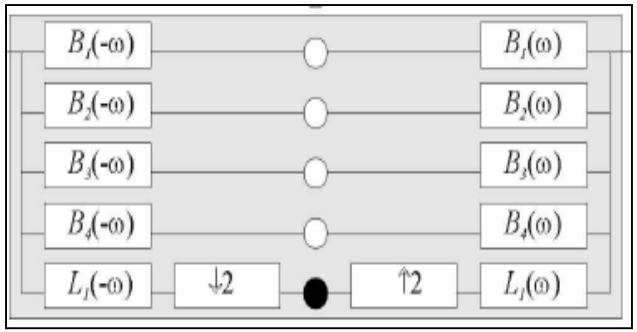


How can we reconstruct (collapse) this pyramid into the original image?

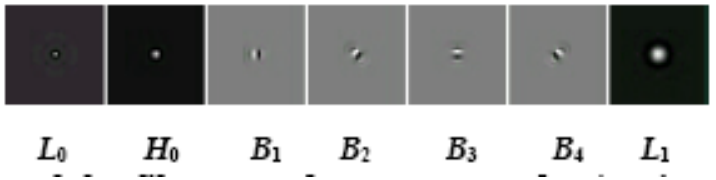
Steerable Pyramid

Input image





7 filters used:

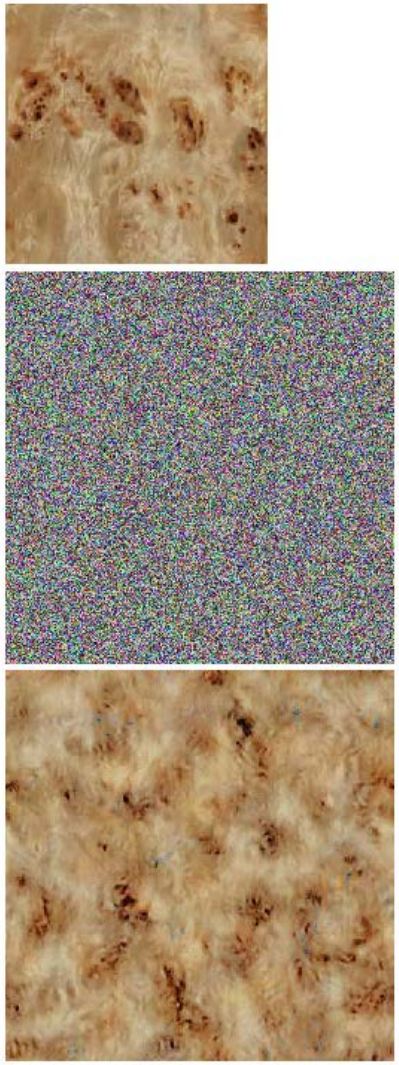


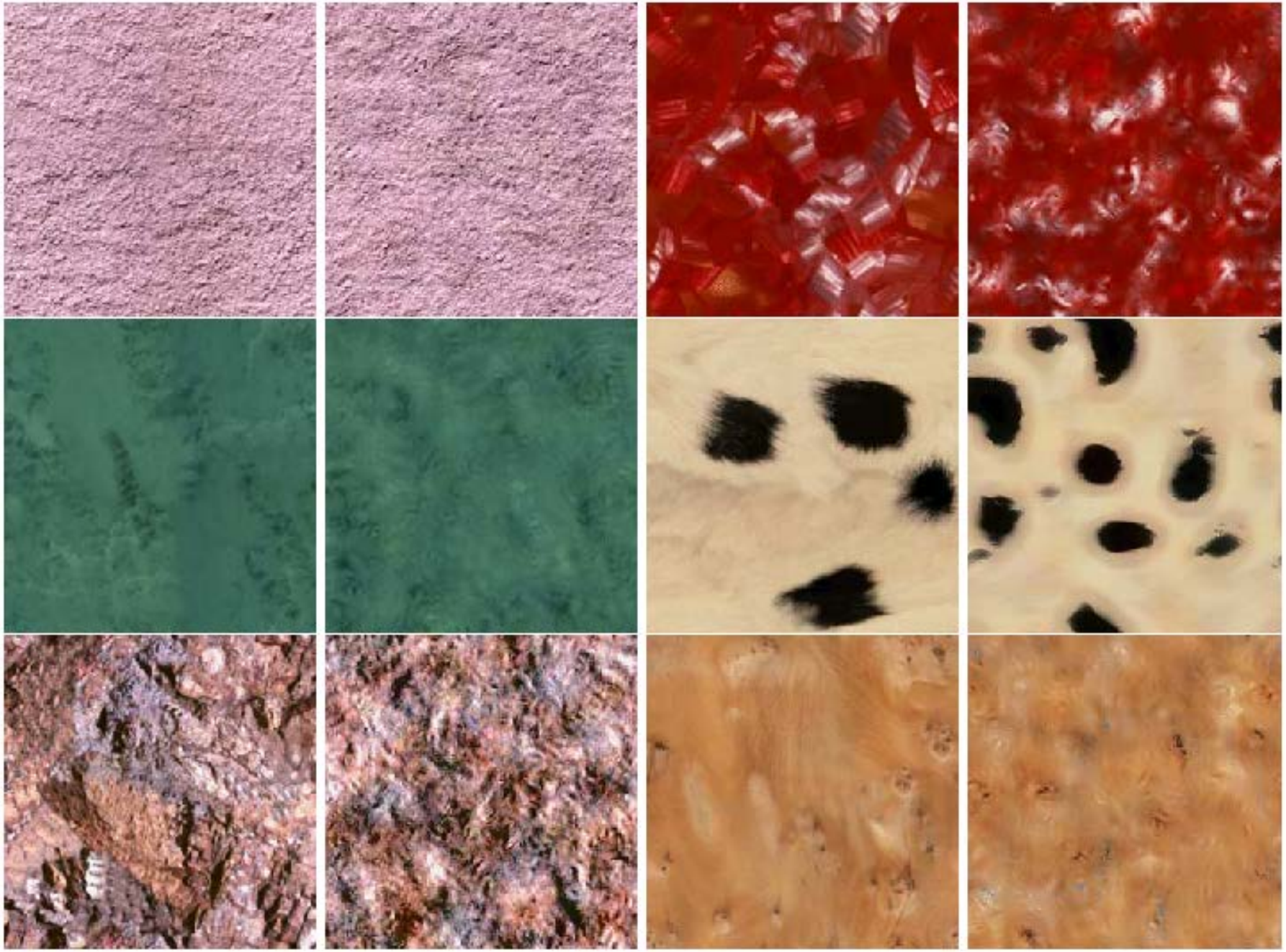
Heeger & Bergen ‘95

Start with a noise image as output

Main loop:

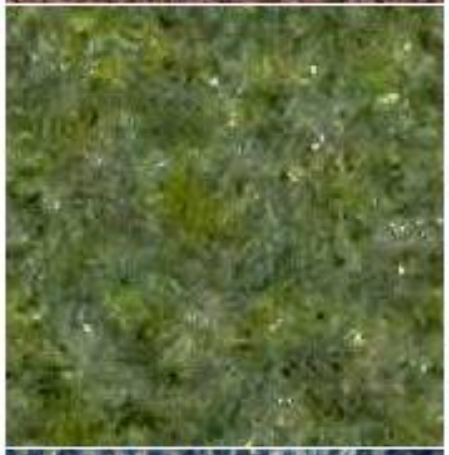
- Match pixel histogram of output image to input
- Decompose input and output images using multi-scale filter bank (Steerable Pyramid)
- Match subband histograms of input and output pyramids
- Reconstruct input and output images (collapse the pyramids)









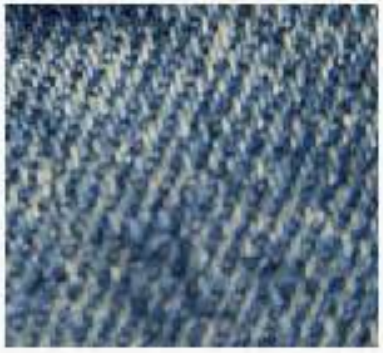
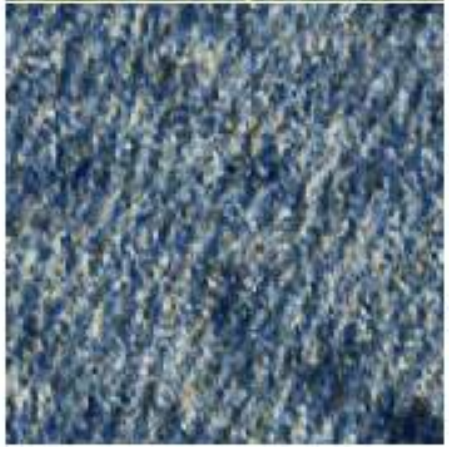




Figure 7: (Left pair) Inhomogeneous input texture produces blotchy synthetic texture. (Right pair) Homogenous input.

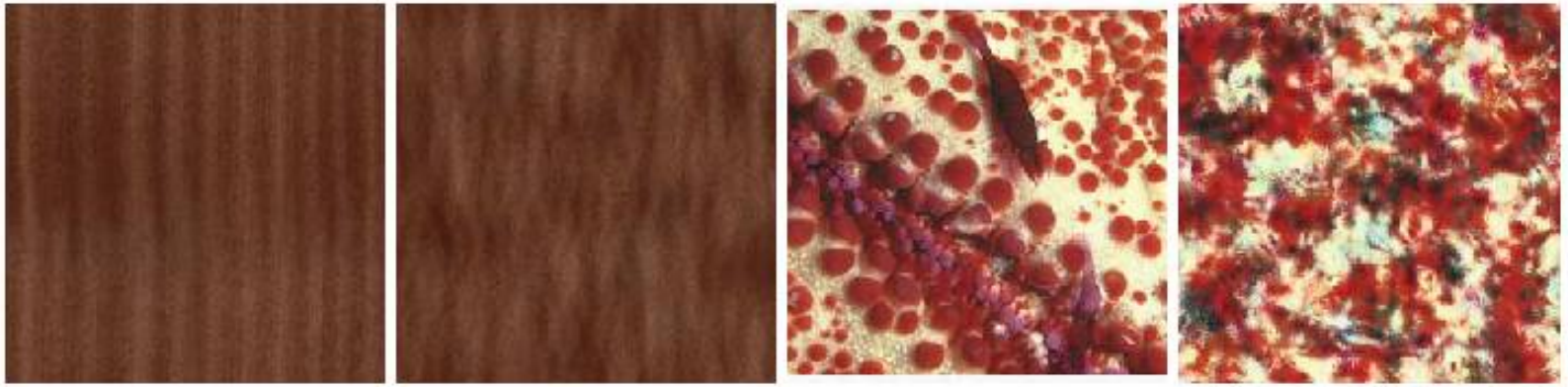


Figure 8: Examples of failures: wood grain and red coral.

Figure 9: More failures: hay and marble.

Simoncelli & Portilla ’98+

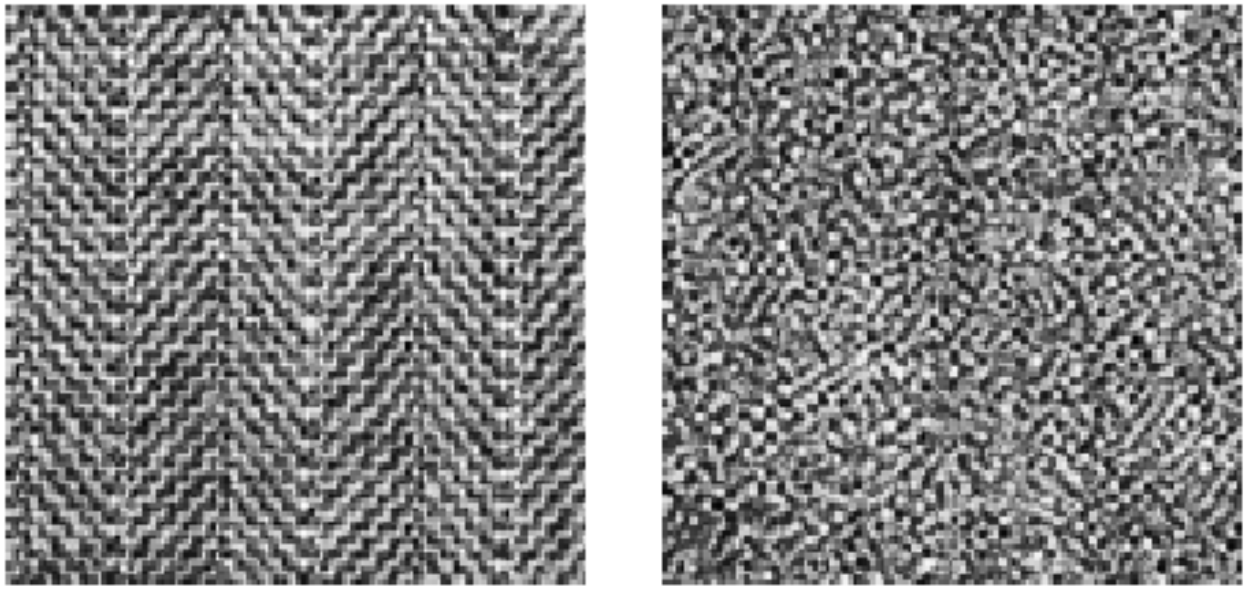


Figure 1. Textures with matching marginal statistics.

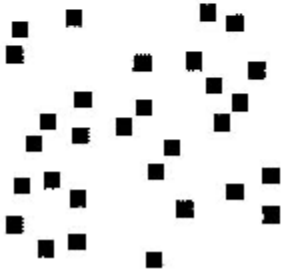
Marginal statistics are not enough

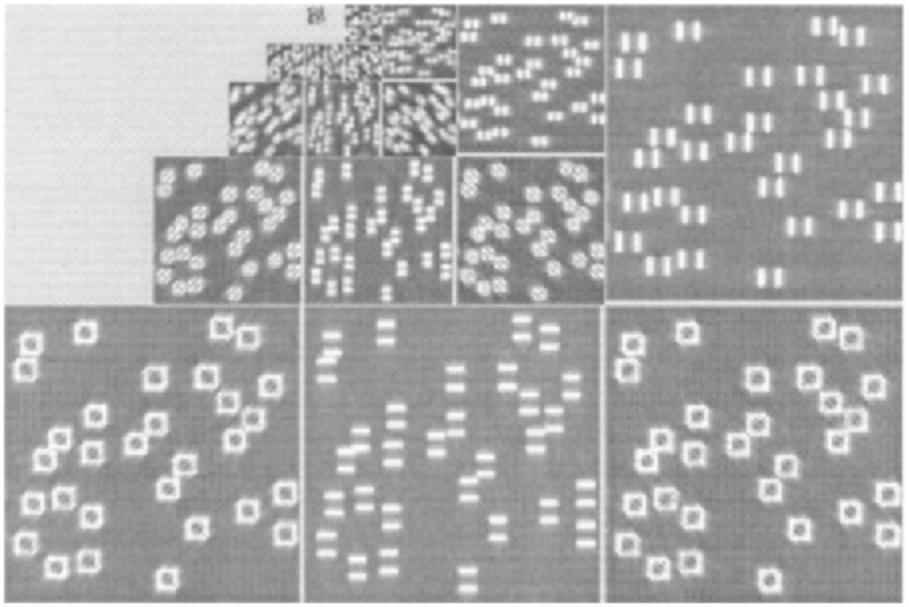
Neighboring filter responses are highly correlated

- an edge at low-res will cause an edge at high-res

Let’s match 2nd order statistics too!

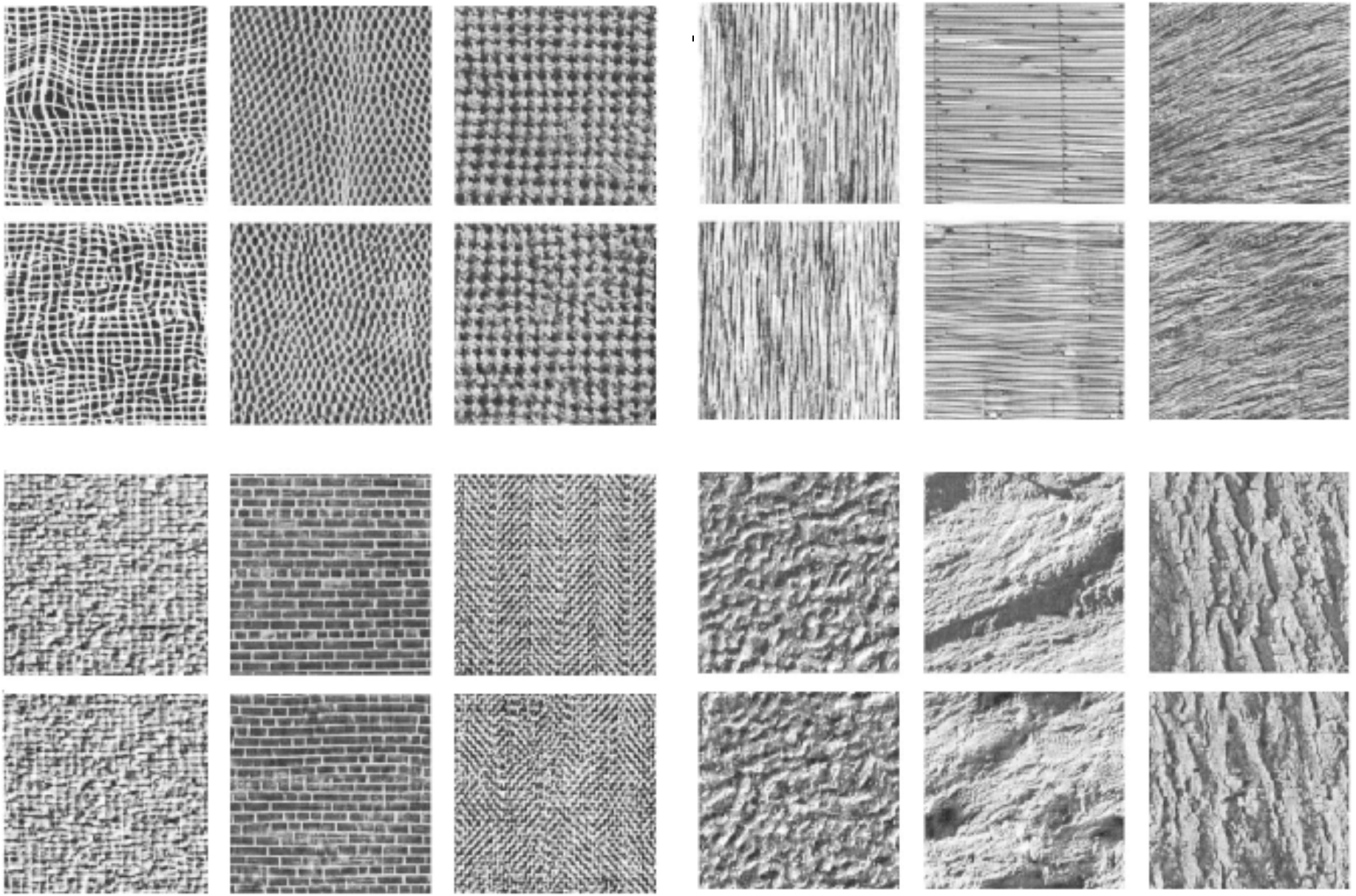
Simoncelli & Portilla ’98+





Match joint histograms of pairs of filter responses at adjacent spatial locations, orientations, and scales.

Optimize using repeated projections onto statistical constraint sufraces

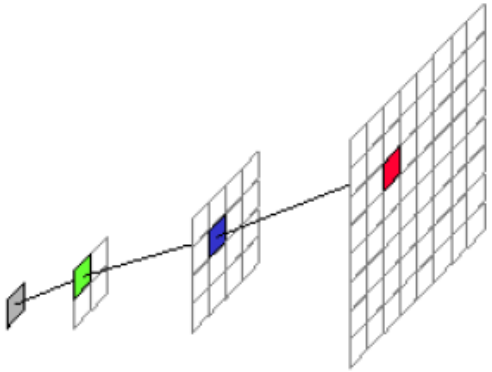


Texture for object recognition

A Cluster-Based Statistical Model for Object Detection

Thomas D. Rikert Michael J. Jones Paul Viola

A “jet”



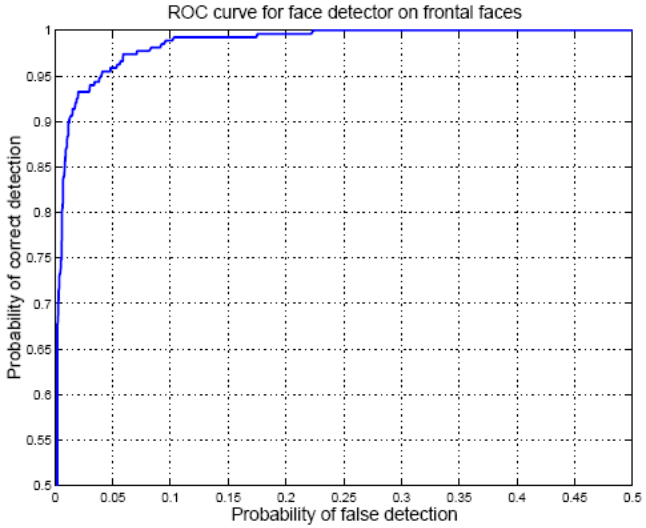


Figure 2: Texture synthesis examples using a set of

Figure 4: A cluster was chosen from the face model and compared to all parent vectors in each test image. The white boxes show the postion in each image where a parent vector was near the cluster. This cluster apparently represents a lip-corner feature.

Object → **Bag of ‘words’**

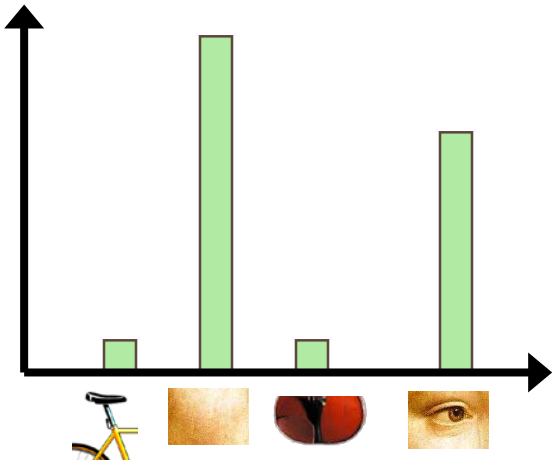
Analogy to documents

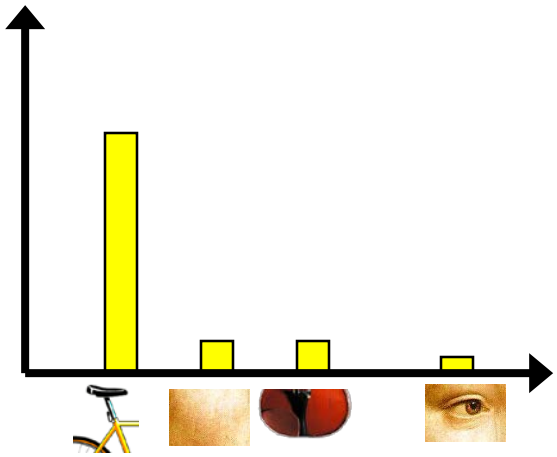
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that r… our eyes. For a long ti… retinal image wa… isual centers i… s a movie s… image … discove… know th… perceptio… more com…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

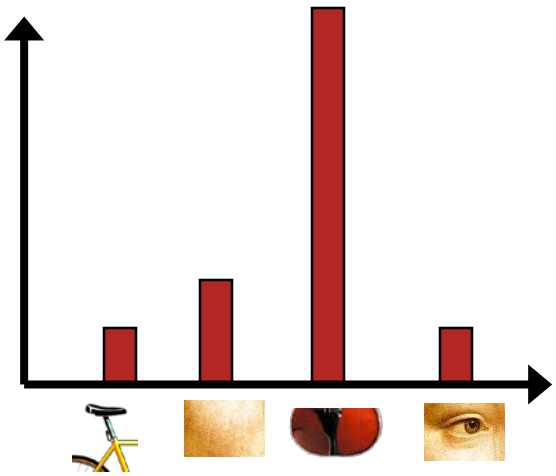
sensory, brain, visual, perception, retinal, cerebral cortex, eye, cell, optical nerve, image Hubel, Wiesel

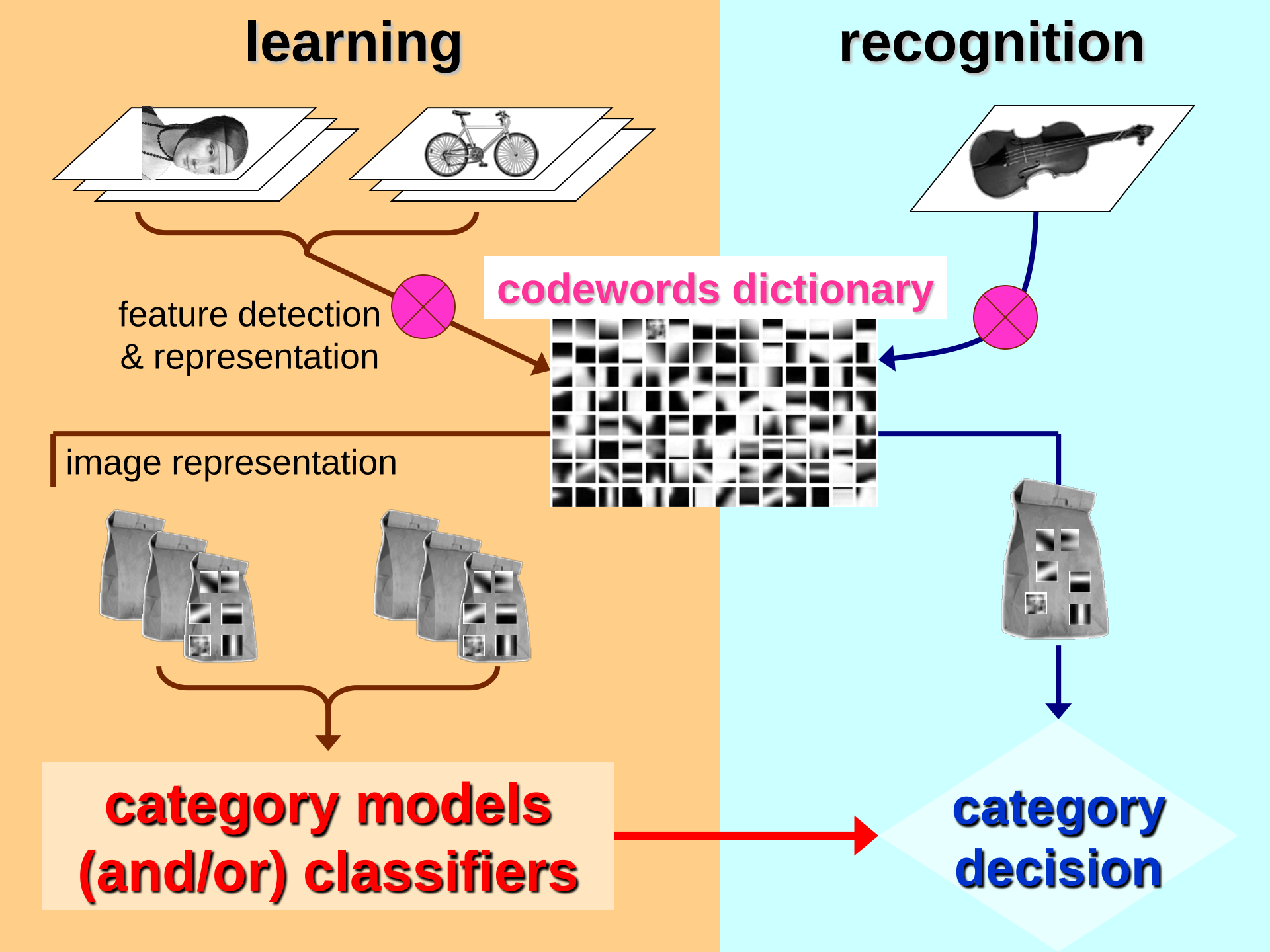
China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% … $750bn, compared wi… $660bn. T… annoy th… China's … deliber… agrees … yuan is … governo… also need… demand so … country. China … e yuan against the dol… nd permitted it to trade within a narrow … but the US wants the yuan to be allowed … de freely. However, Beijing has made it cl… t it will take its time and tread carefully be… allowing the yuan to rise further in value.

China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value

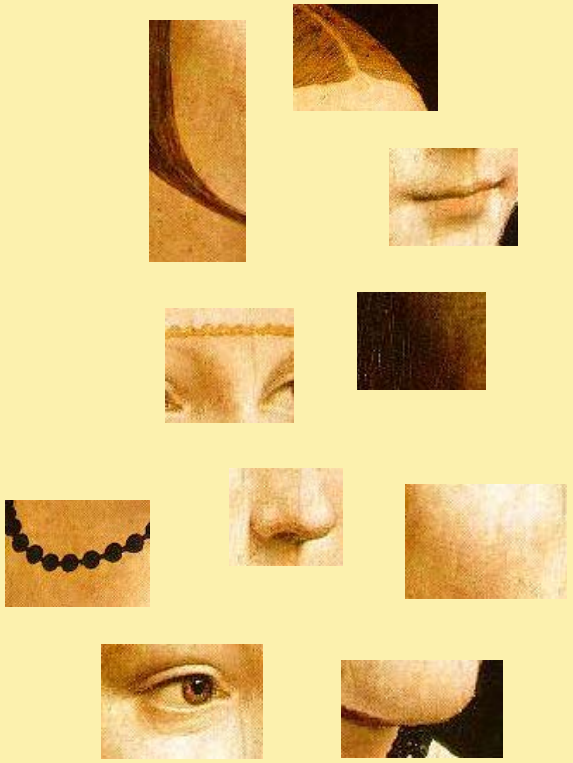







learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

1.Feature detection and representation



Feature detection

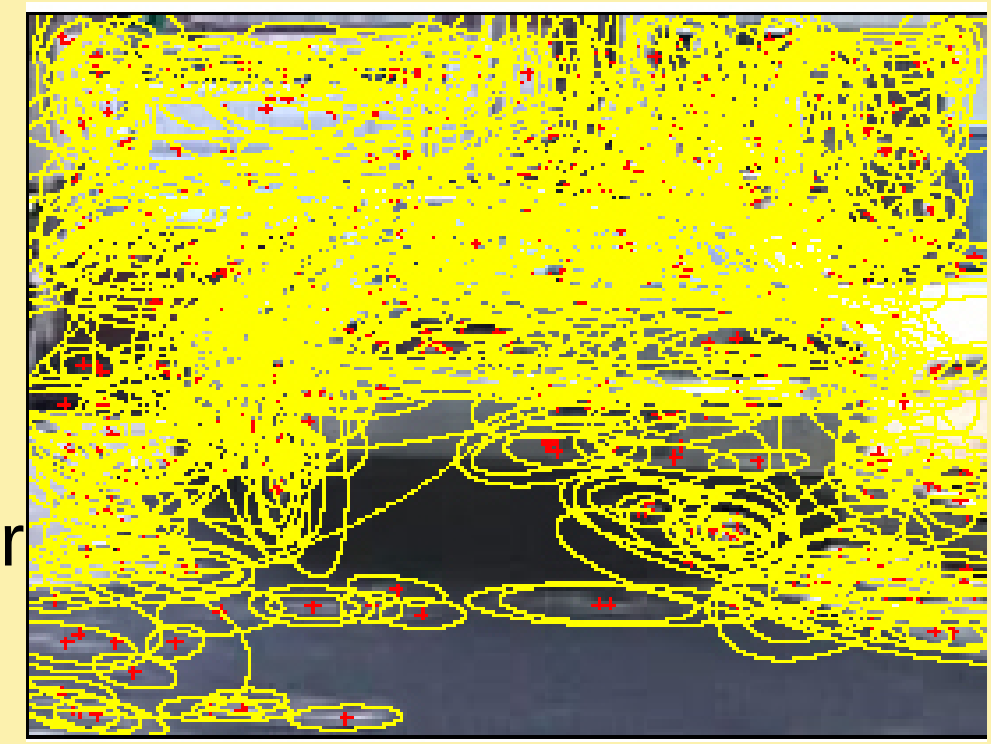
- Sliding Window
  - Leung et al, 1999
  - Viola et al, 1999
  - Renninger et al 2002

Feature detection

- Sliding Window
  - Leung et al, 1999
  - Viola et al, 1999
  - Renninger et al 2002
- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005

Feature detection

- Sliding Window
  - Leung et al, 1999
  - Viola et al, 1999
  - Renninger et al 2002
- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei et al. 2005
  - Sivic et al. 2005



Feature detection

- Sliding Window
  - Leung et al, 1999
  - Viola et al, 1999
  - Renninger et al 2002
- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei et al. 2005
  - Sivic et al. 2005
- Other methods
  - Random sampling (Ullman et al. 2002)
  - Segmentation based patches
    - Barnard et al. 2003, Russell et al 2006, etc.)

Feature Representation

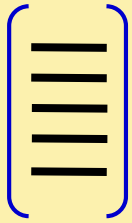
Visual words, aka textons, aka keypoints:

K-means clustered pieces of the image

- Various Representations:
  - Filter bank responses
  - Image Patches
  - SIFT descriptors

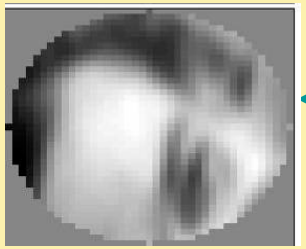
All encode more-or-less the same thing…

Interest Point Features

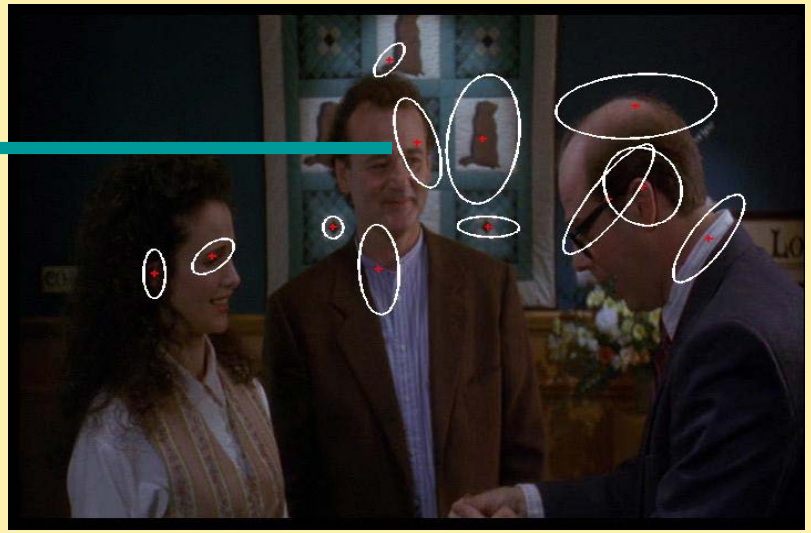


Compute SIFT descriptor

[Lowe'99]



Normalize patch



Detect patches

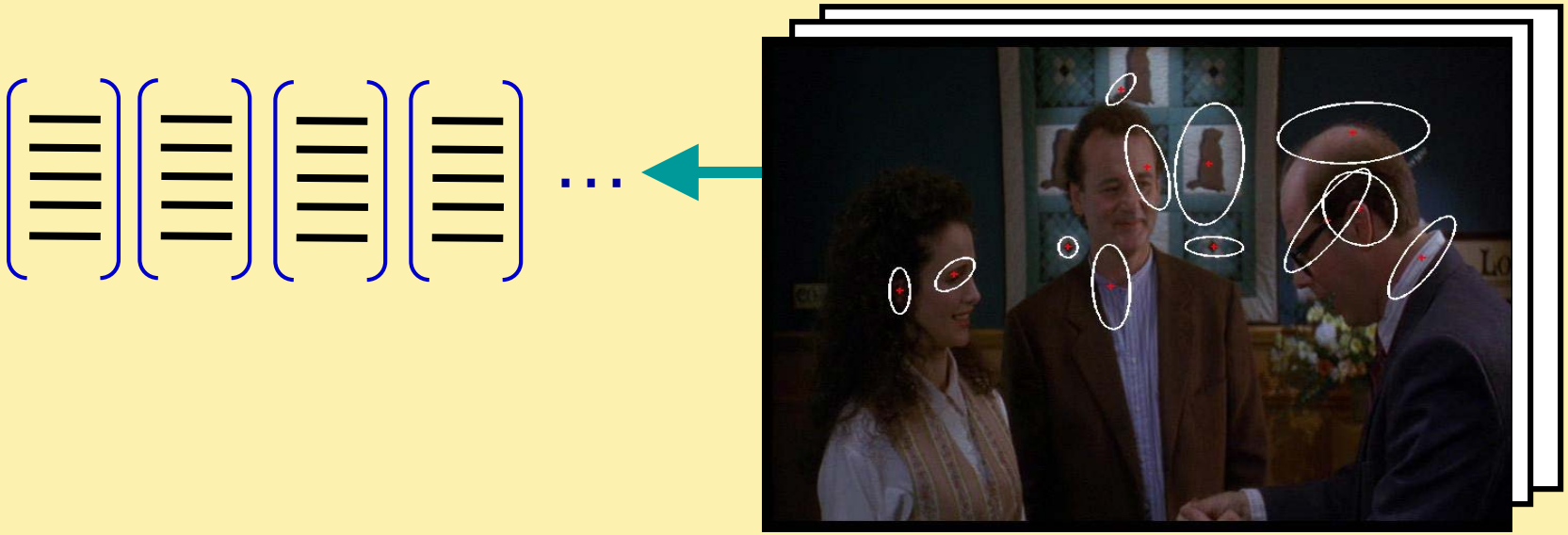
[Mikojaczyk and Schmid '02]

[Matas et al. '02]

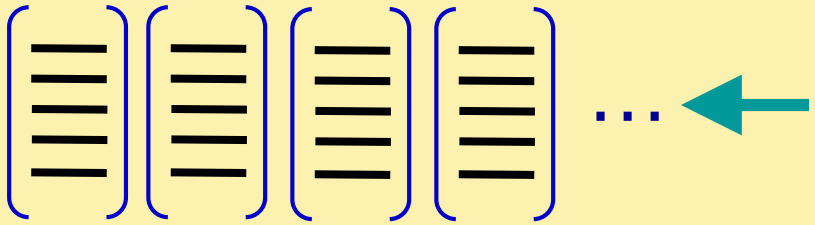
[Sivic et al. '03]

Slide credit: Josef Sivic

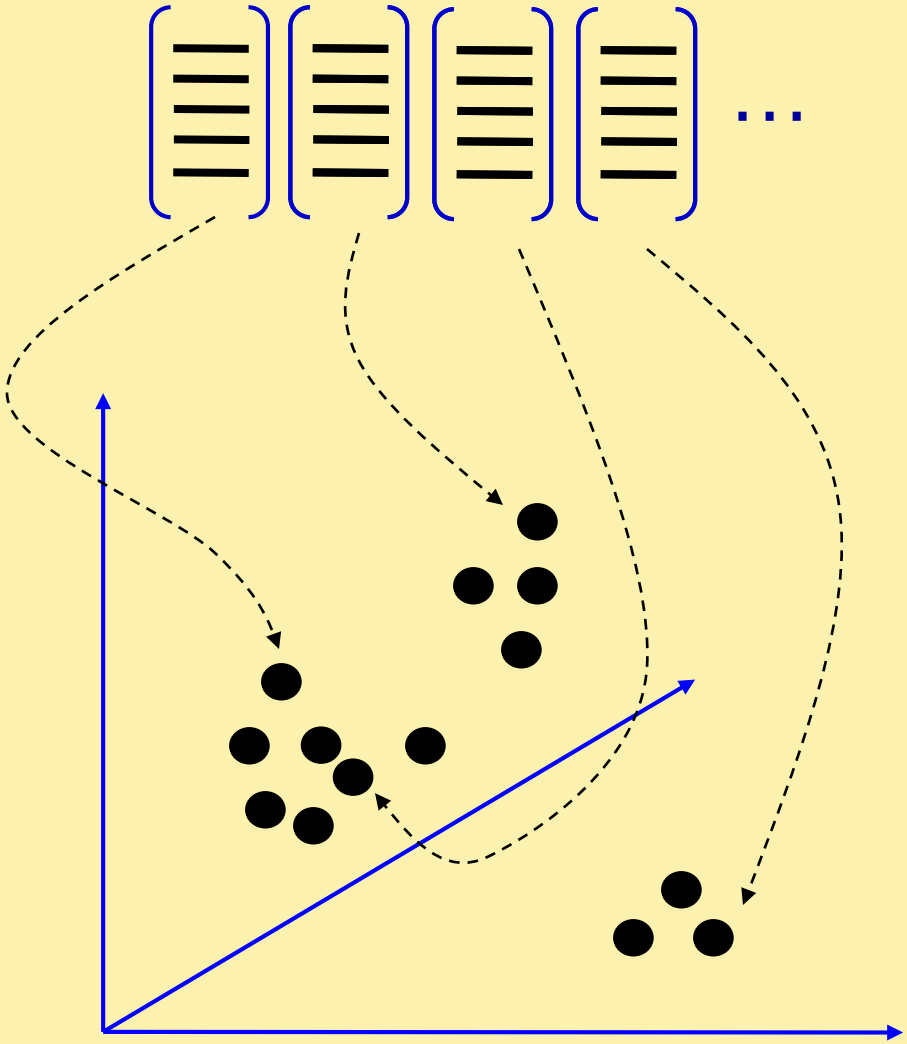
Interest Point Features



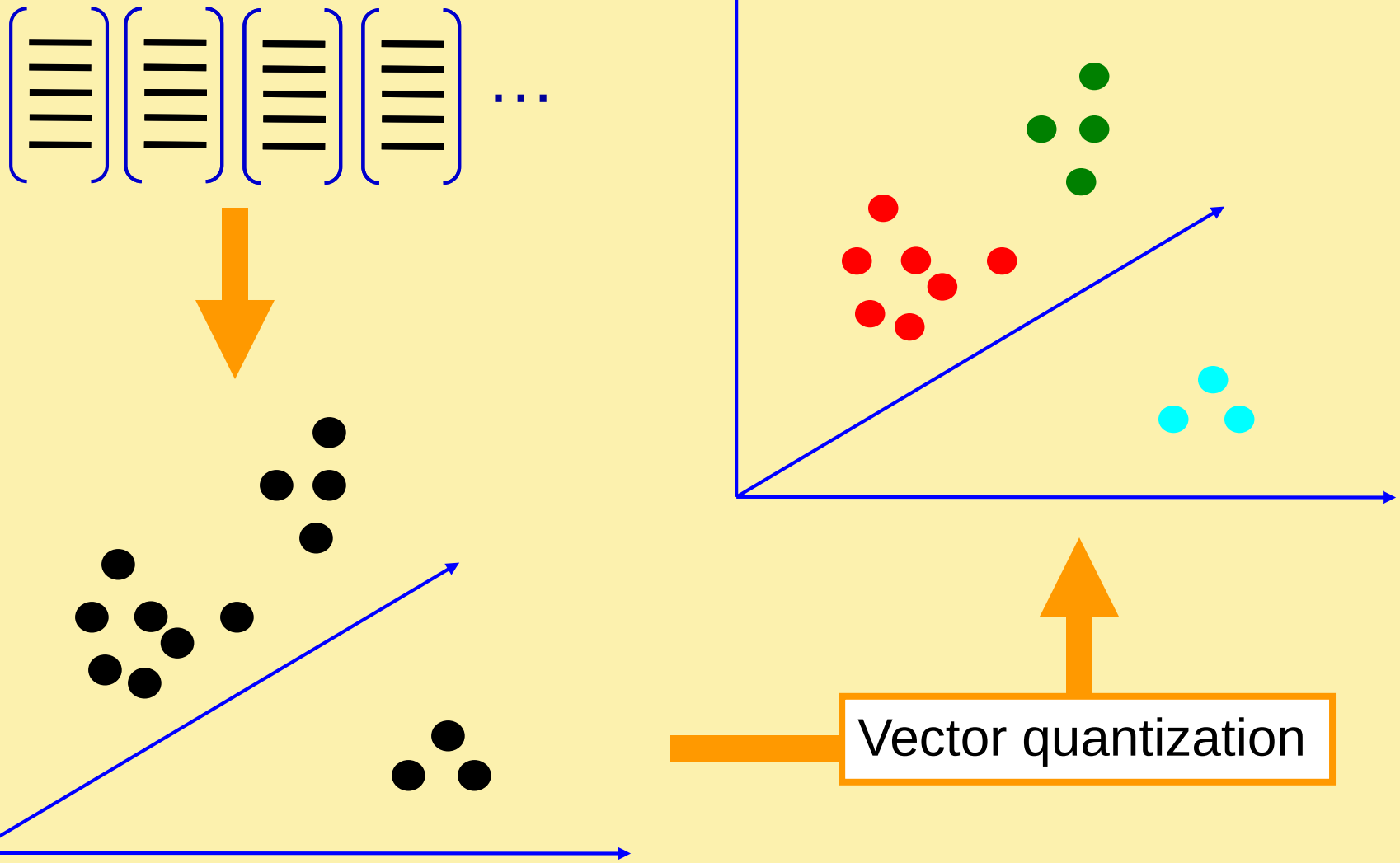
Patch Features



dictionary formation

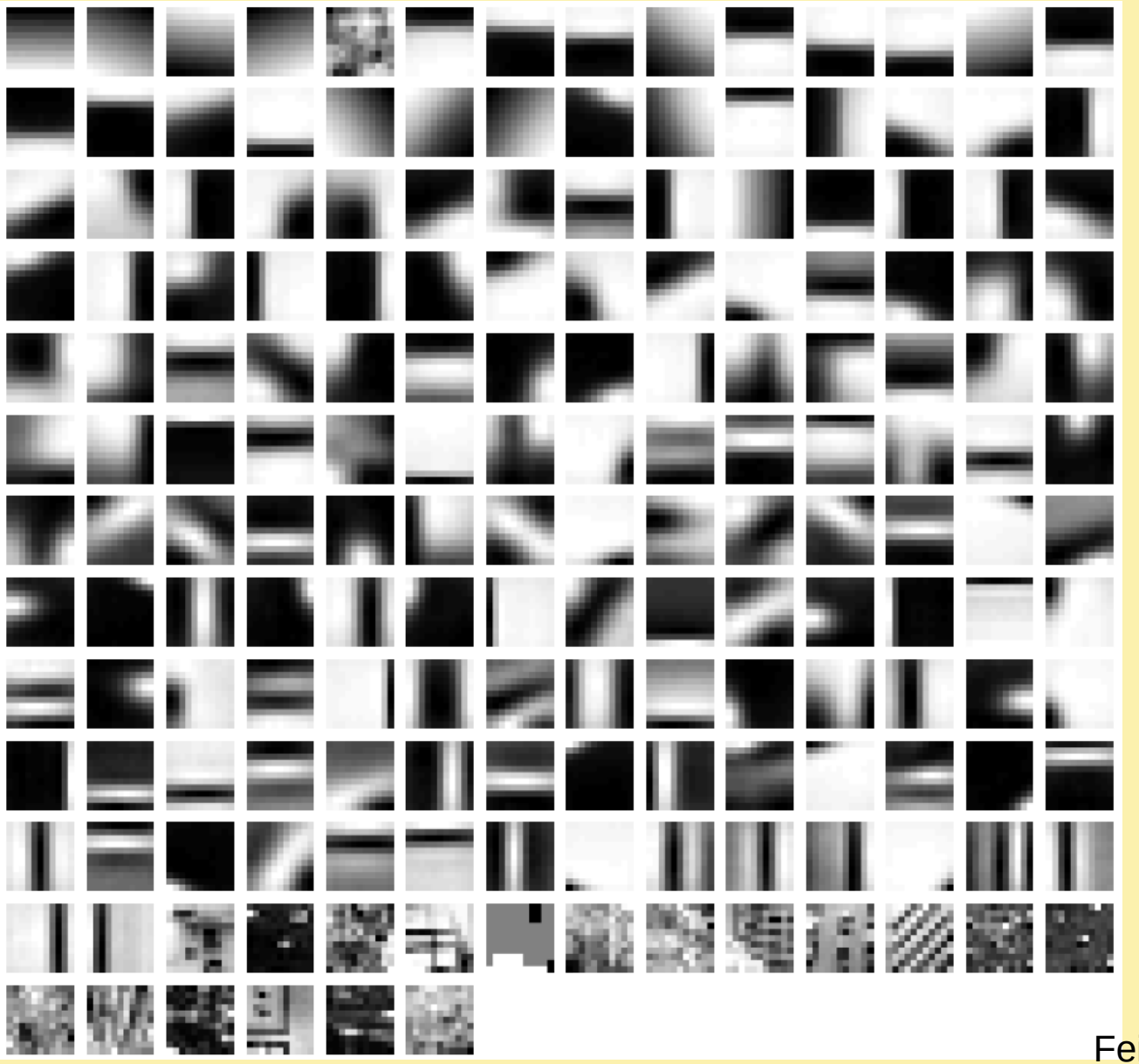


Clustering (usually k-means)



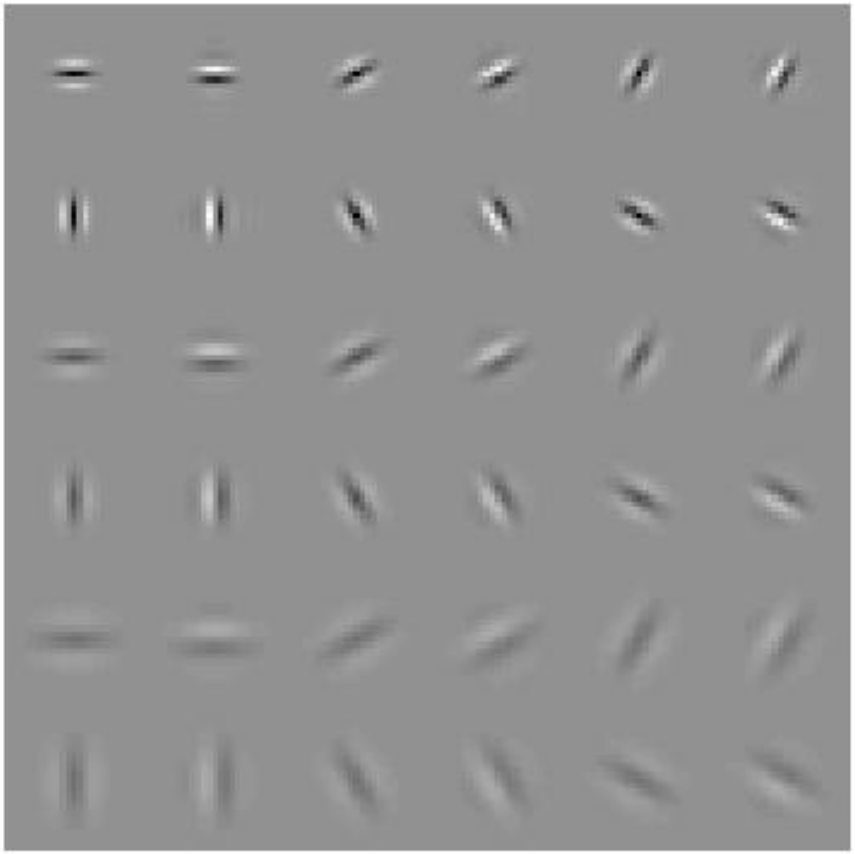
Slide credit: Josef Sivic

Clustered Image Patches



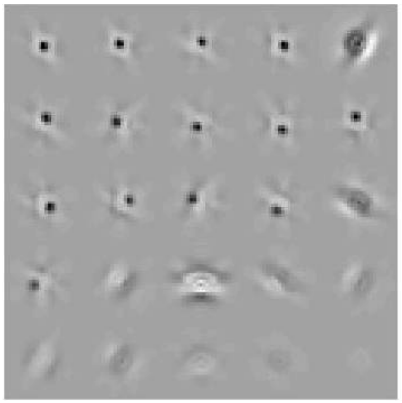
Fei-Fei et al. 2005

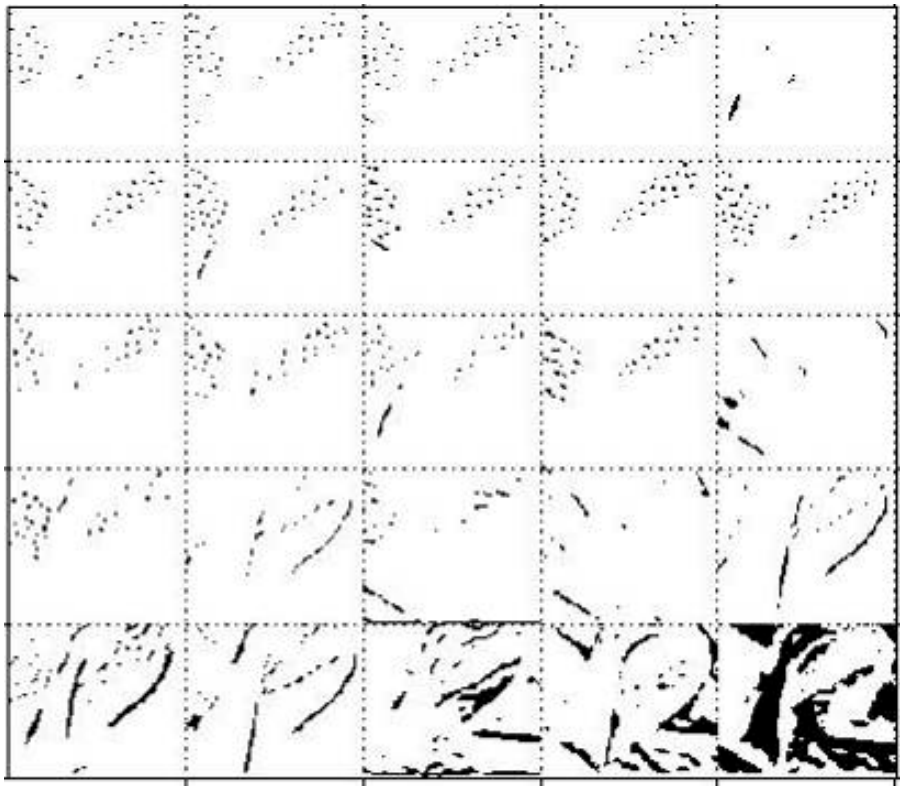
Filterbank



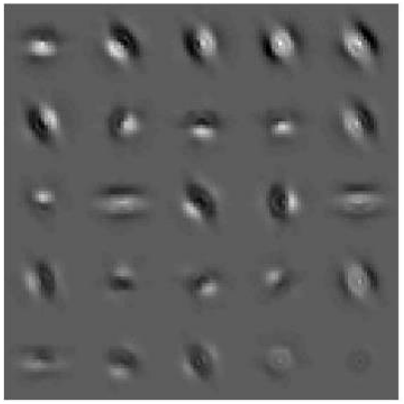
Textons (Malik et al, IJCV 2001)

- K-means on vectors of filter responses





Textons (cont.)



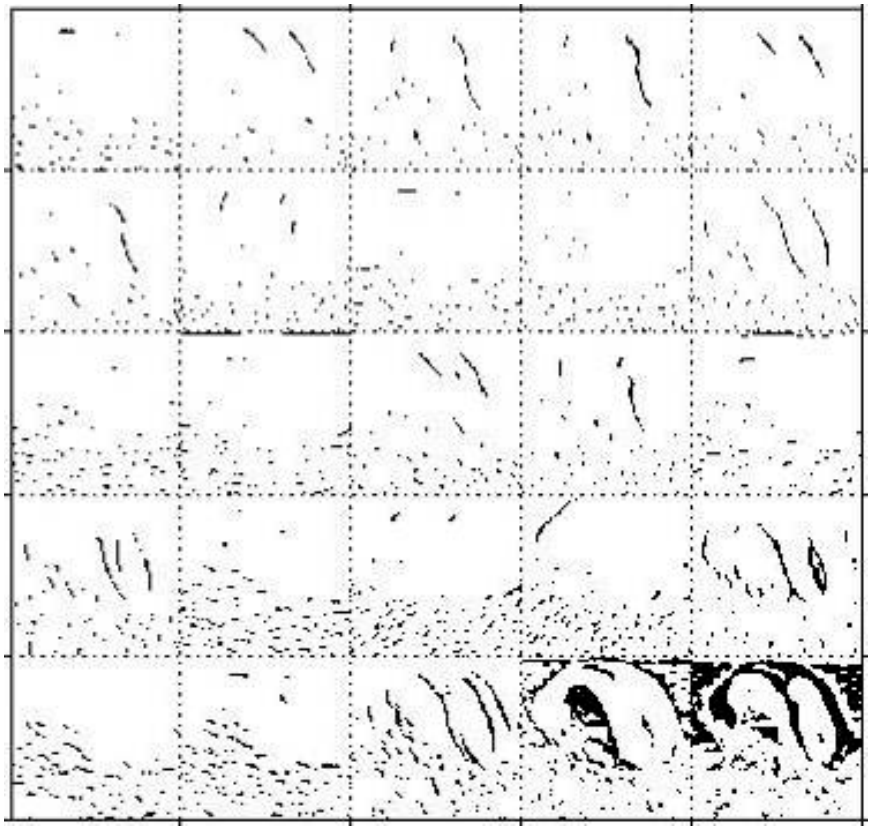
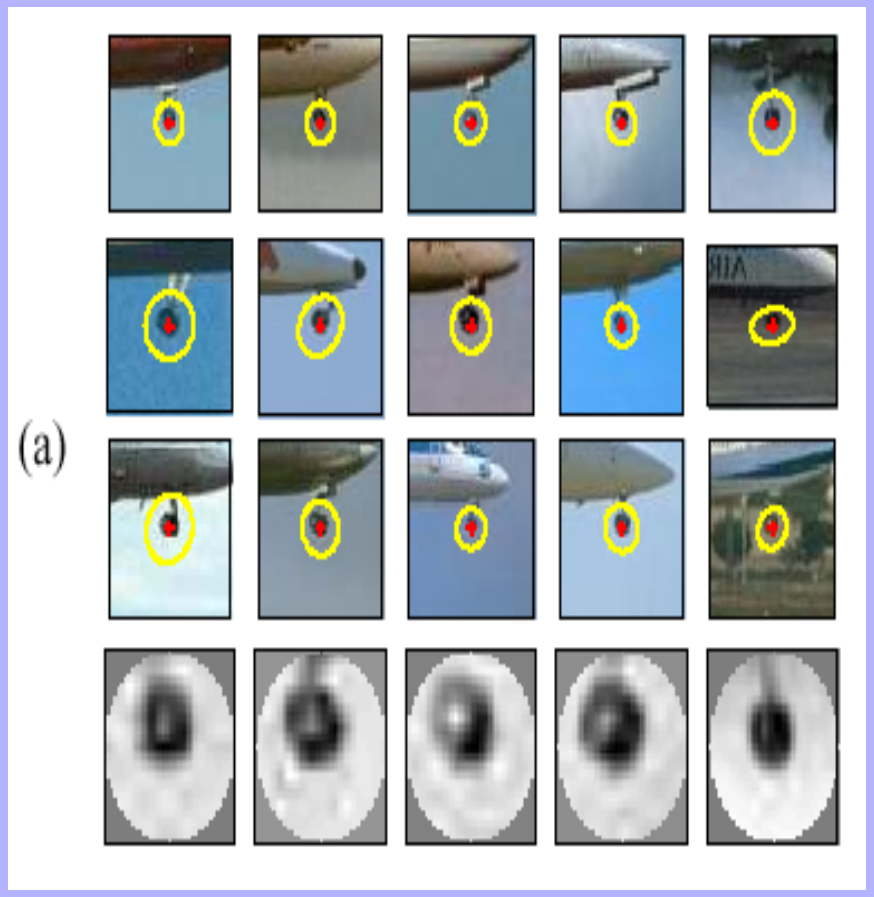
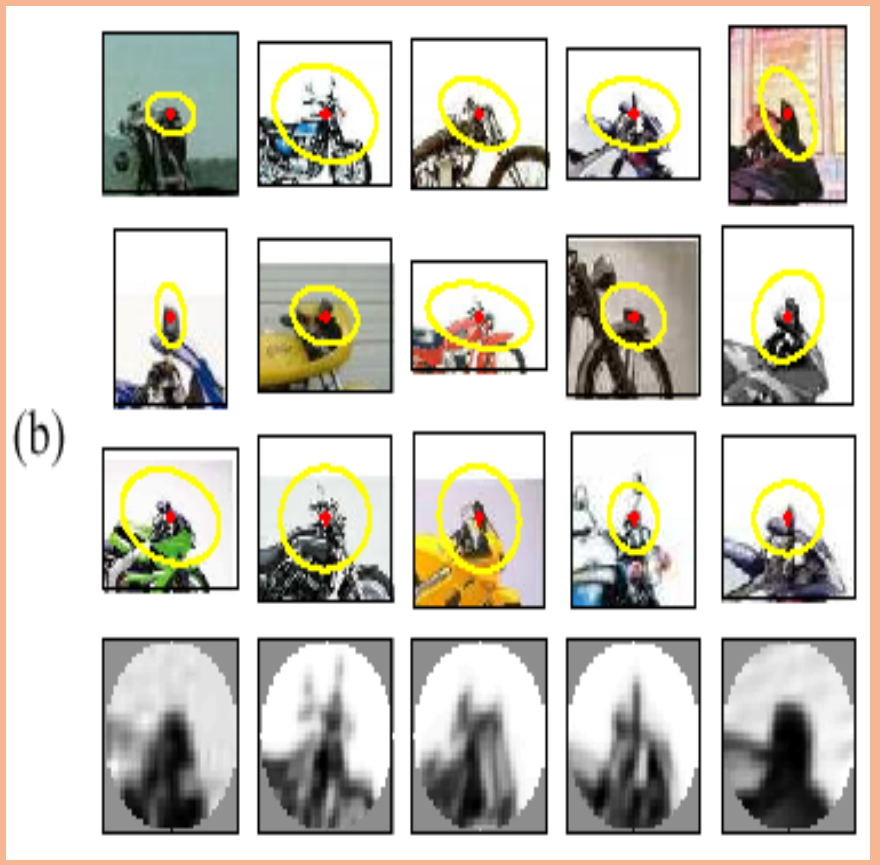


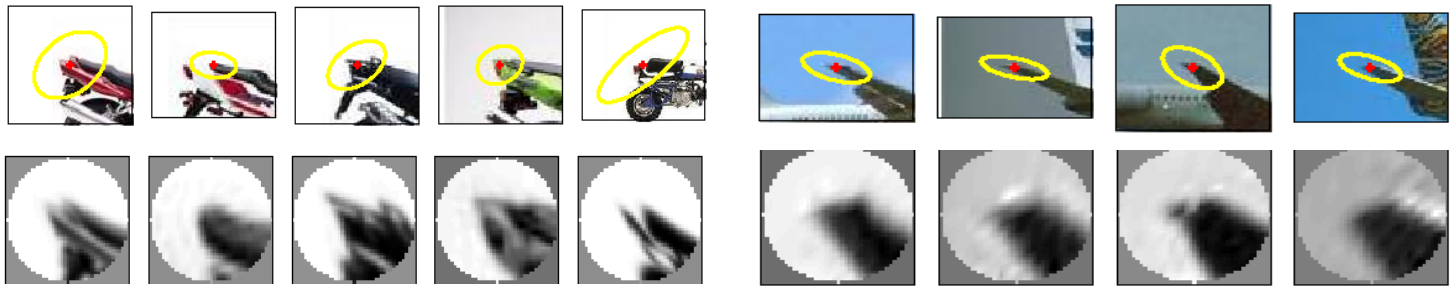
Image patch examples of codewords



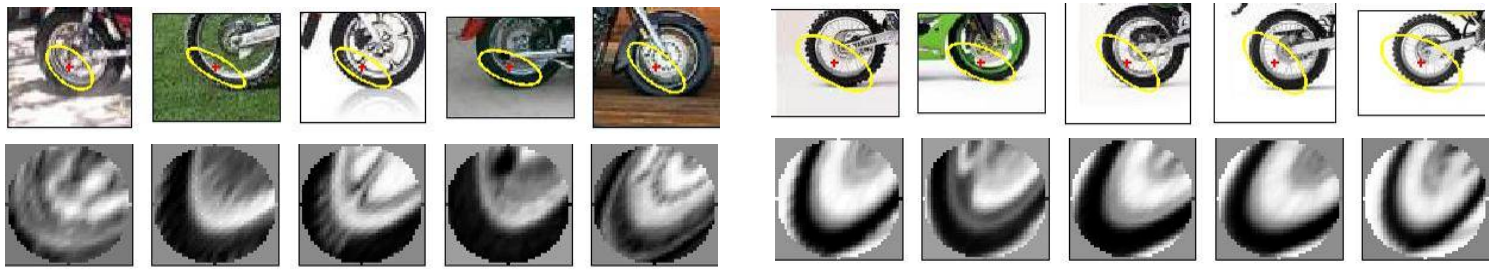


Sivic et al. 2005

Visual synonyms and polysemy

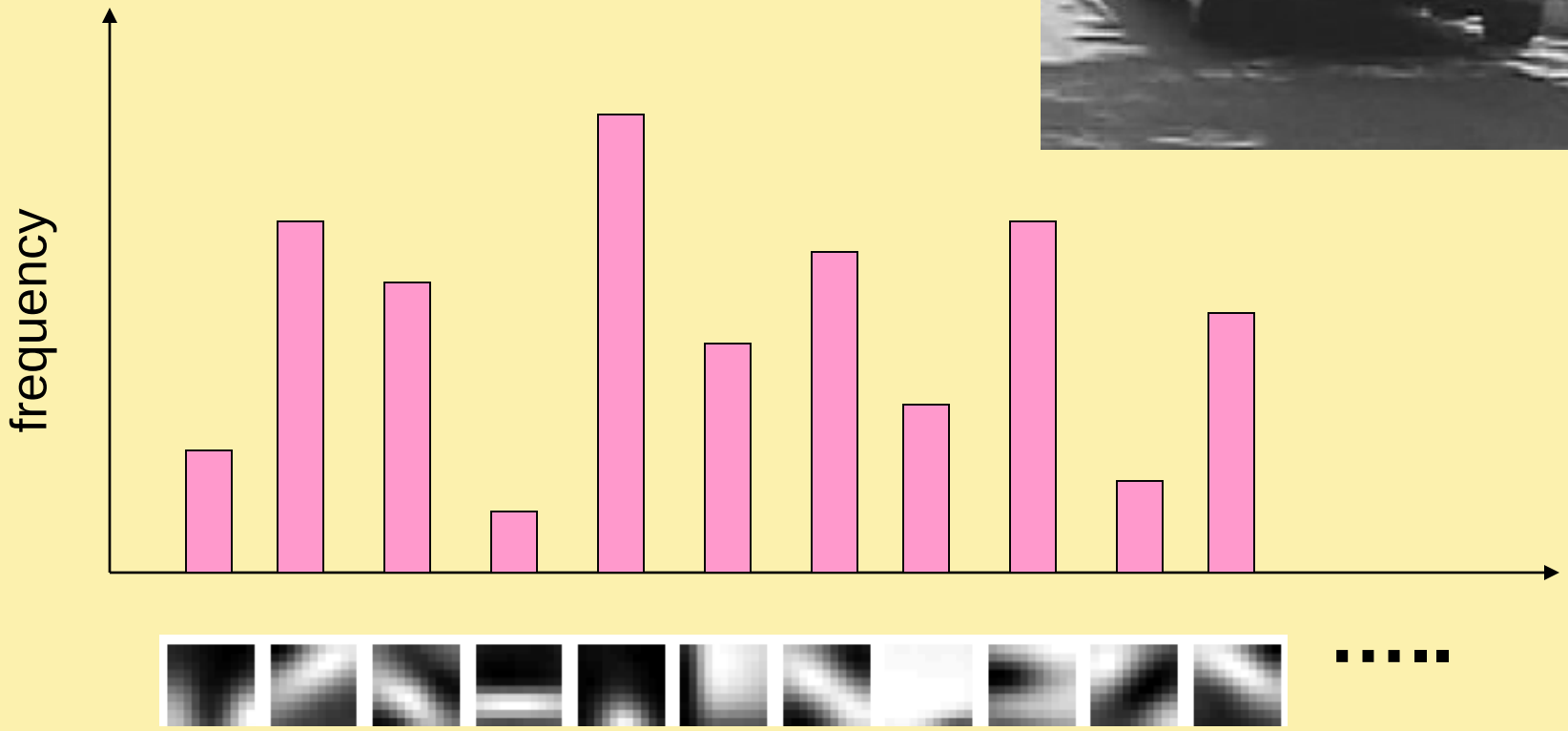


Visual Polysemy. Single visual word occurring on different (but locally similar) parts on different object categories.



Visual Synonyms. Two different visual words representing a similar part of an object (wheel of a motorbike).

Image representation



Scene Classification (Renninger & Malik)

beach

mountain

forest

city

street

farm

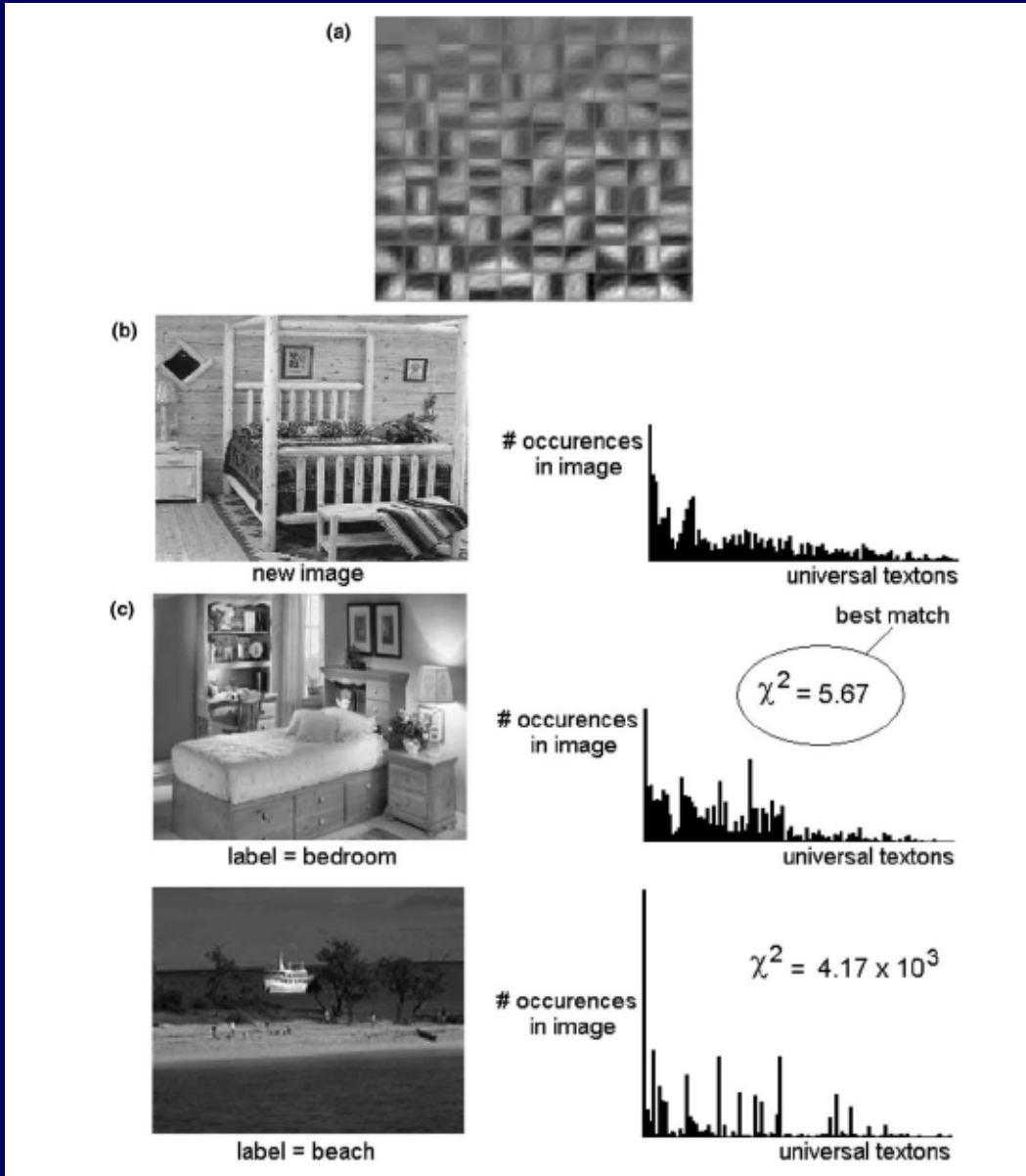
kitchen

livingroom

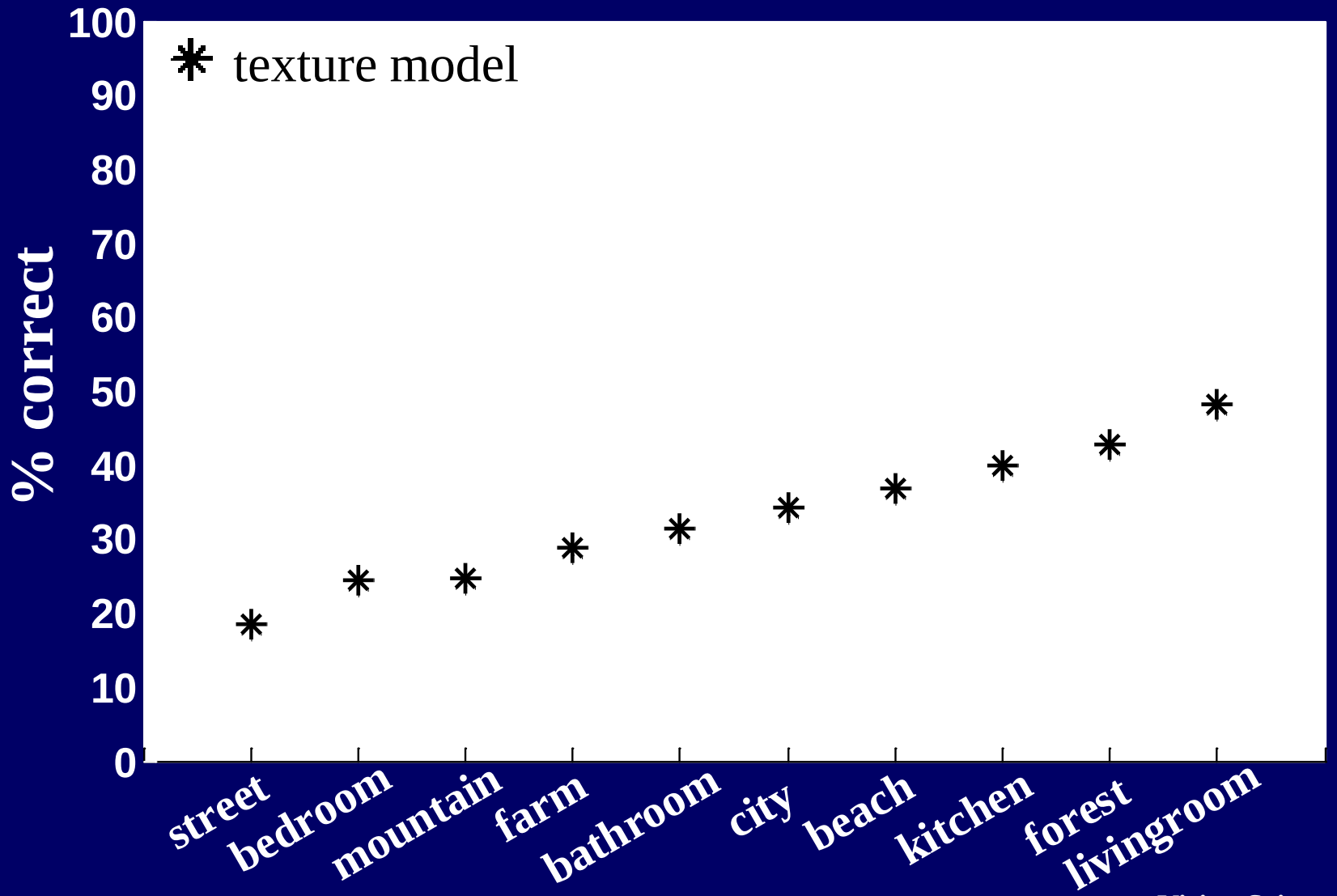
bedroom

bathroom

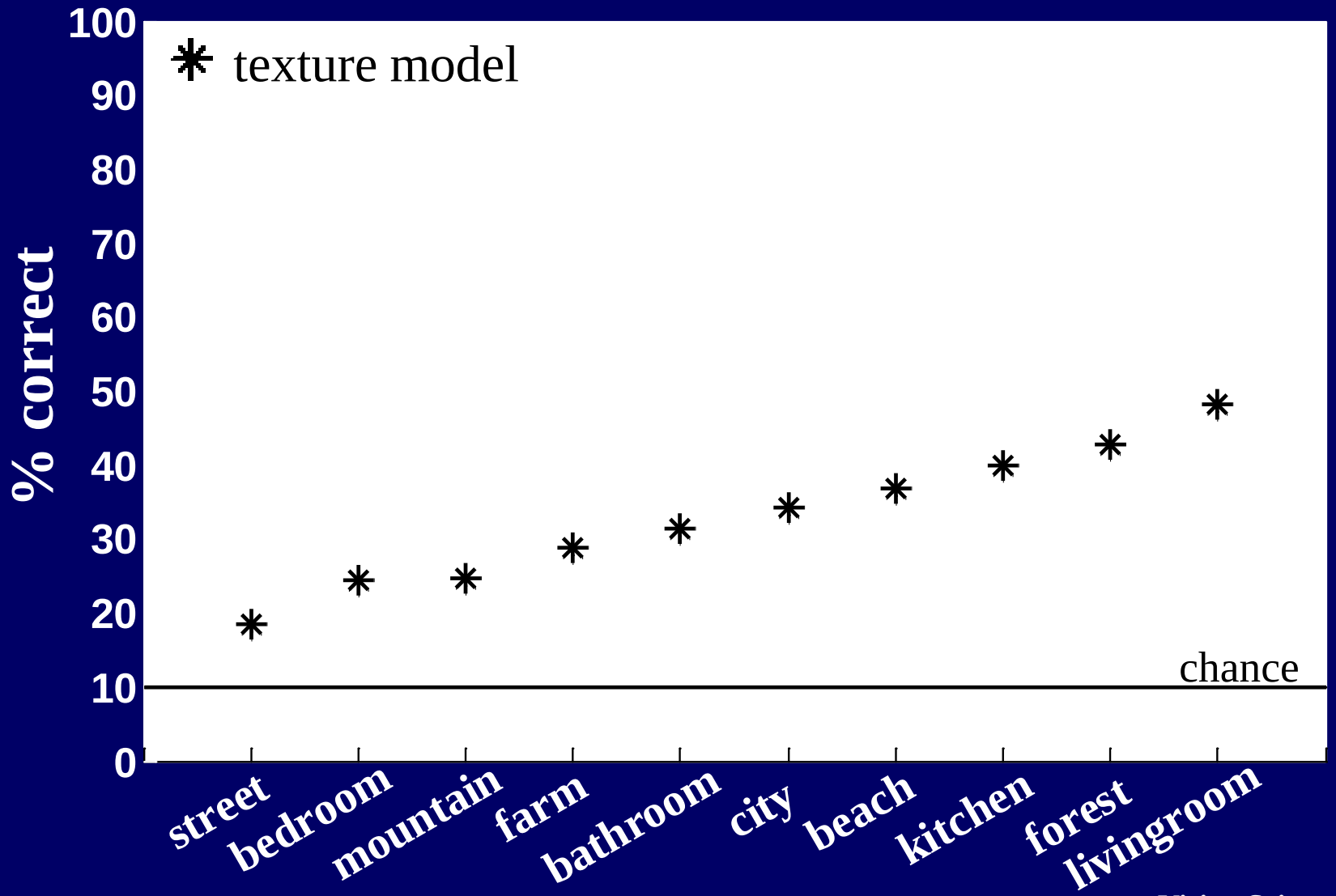
kNN Texton Matching



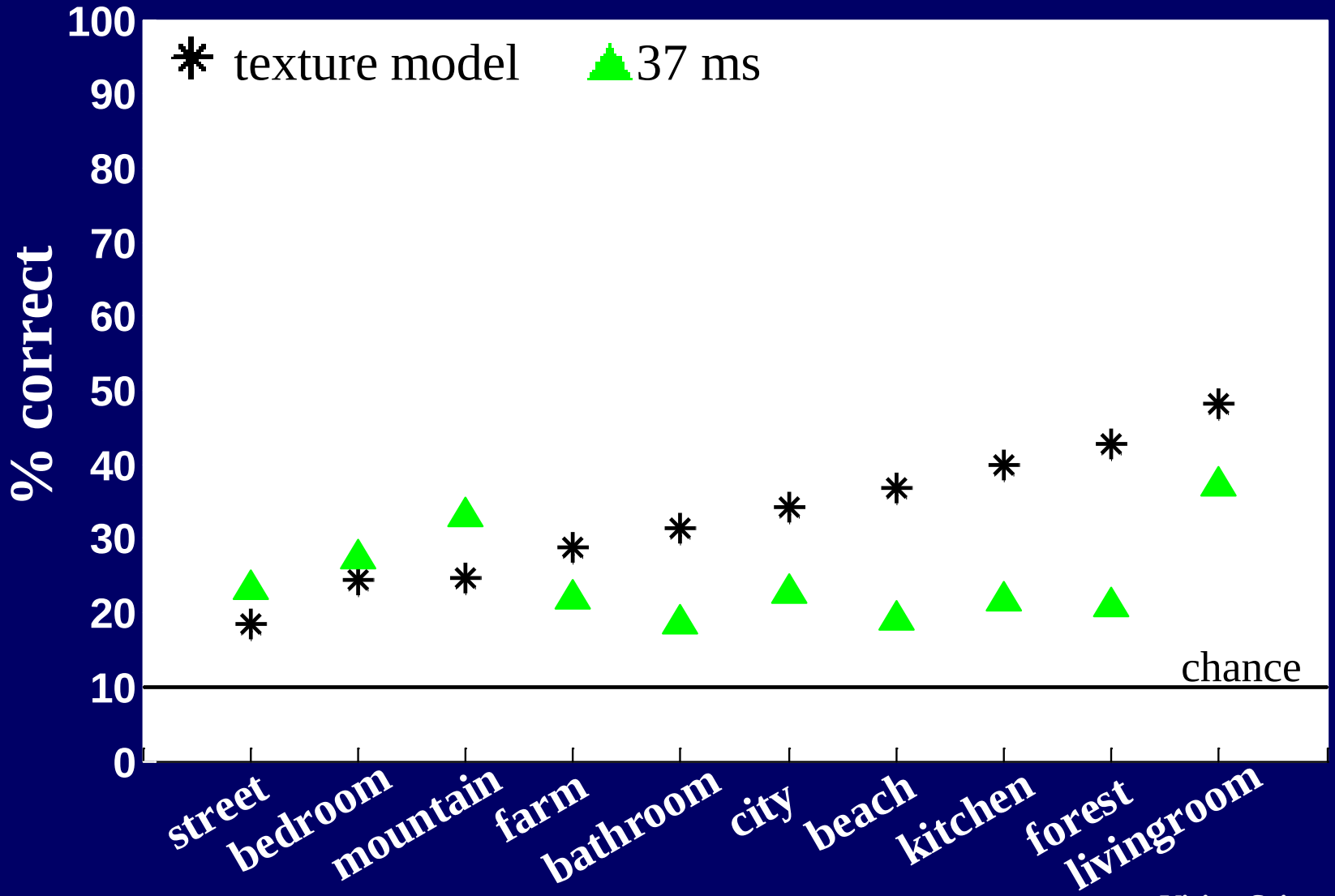
Discrimination of Basic Categories



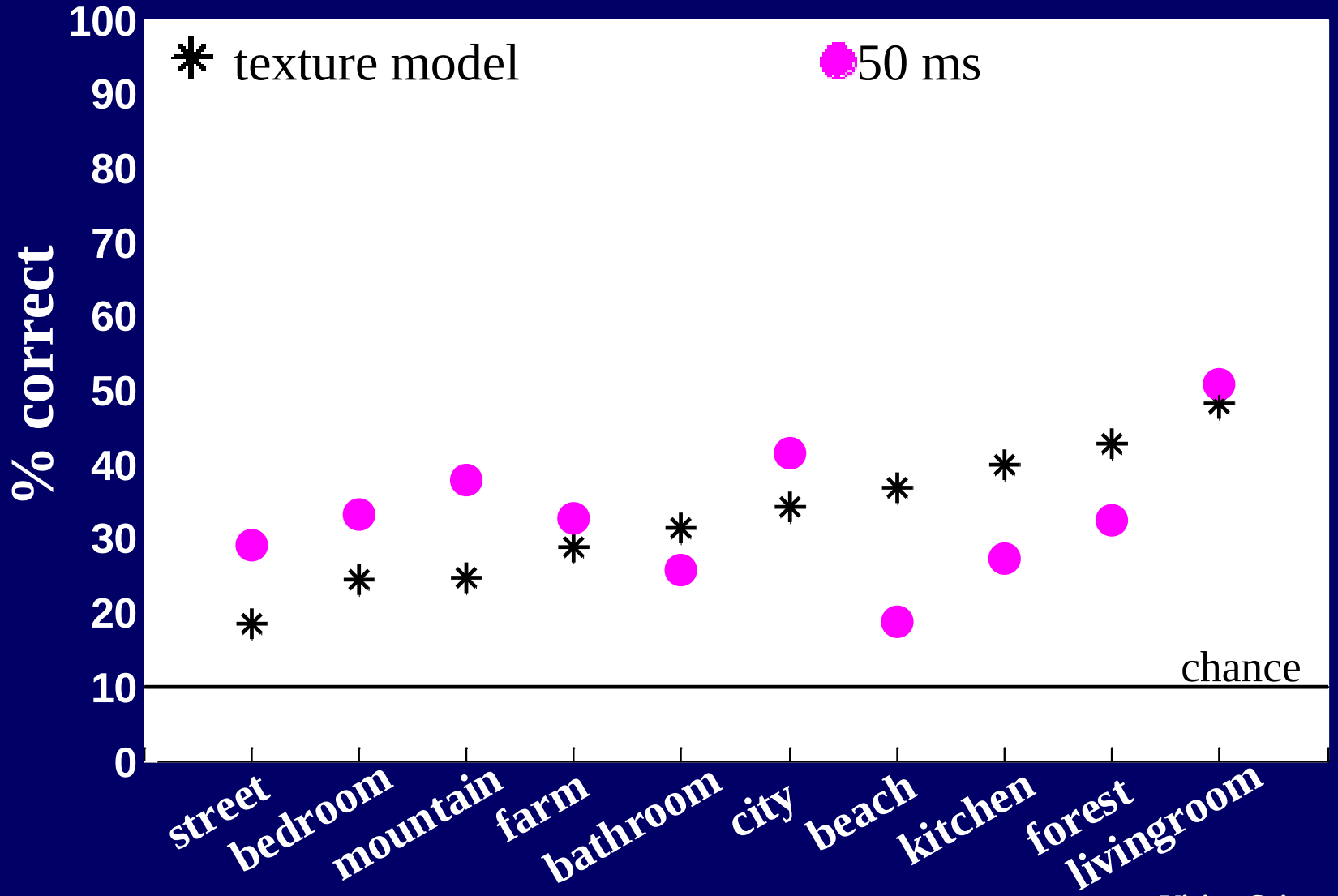
Discrimination of Basic Categories



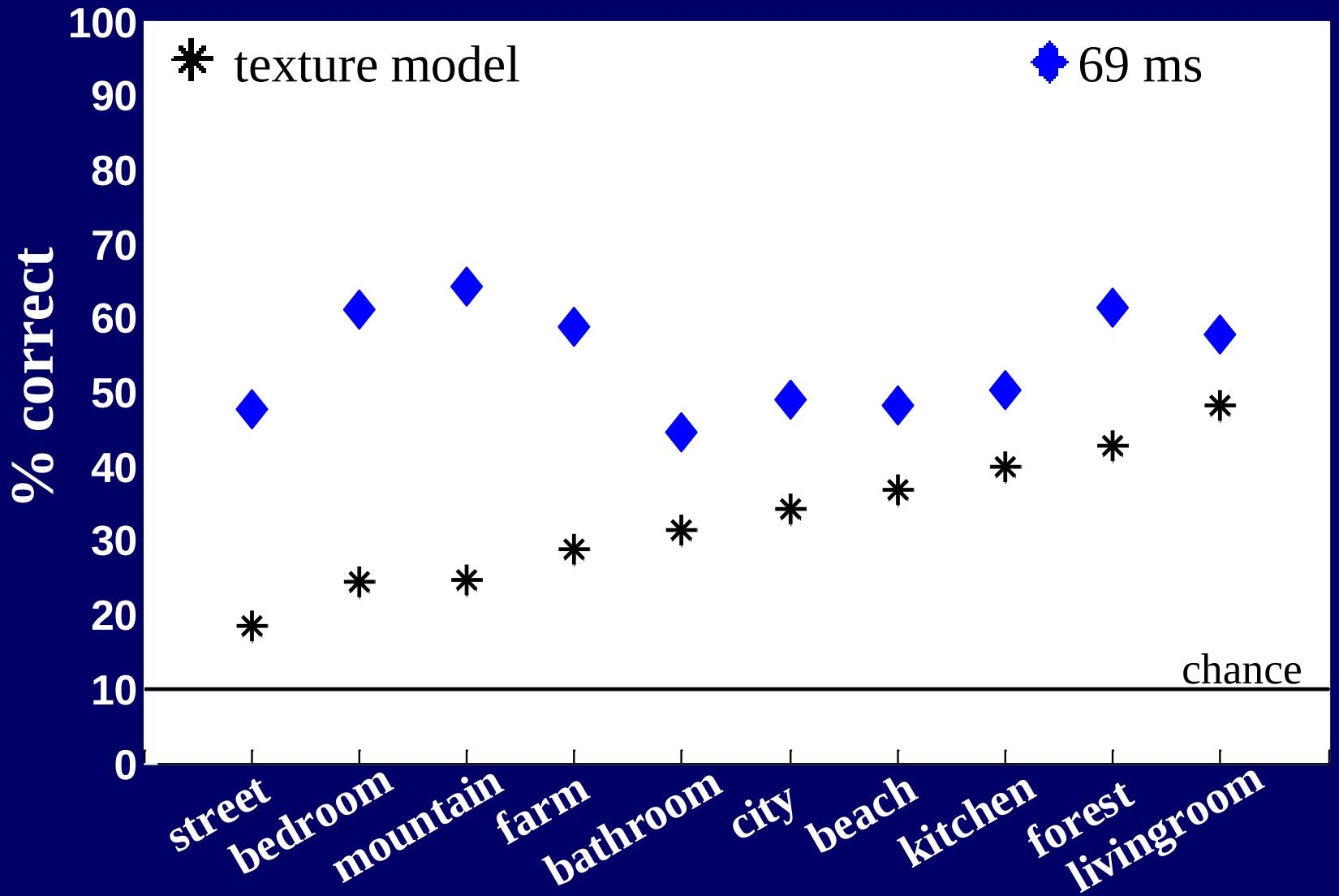
Discrimination of Basic Categories



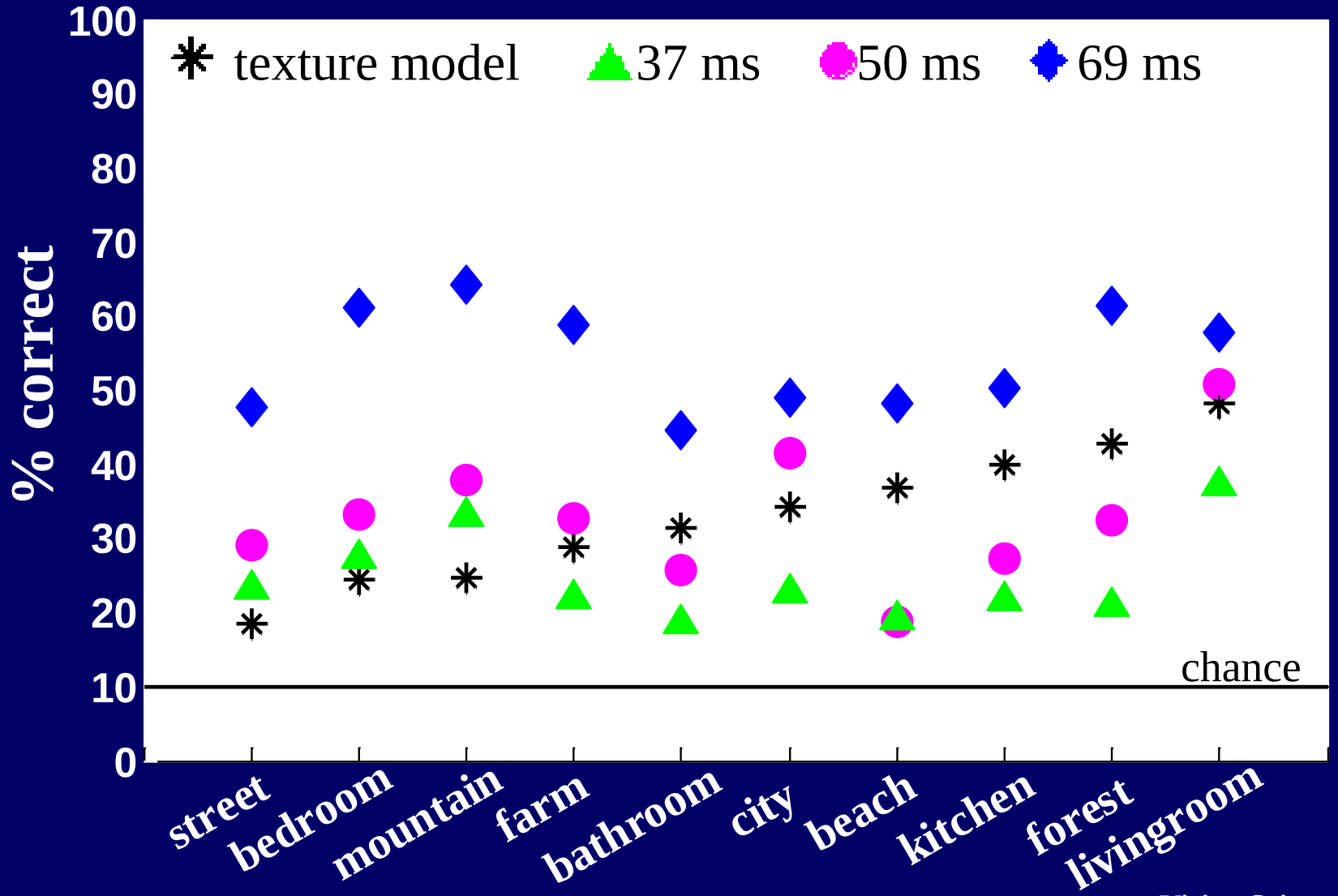
Discrimination of Basic Categories



Discrimination of Basic Categories



Discrimination of Basic Categories



Object Recognition using texture

Object Categorization by Learned Universal Visual Dictionary

J. Winn, A. Criminisi and T. Minka

Microsoft Research, Cambridge, UK – http://research.microsoft.com/vision/cambridge/recognition/

Learn texture model

representation:

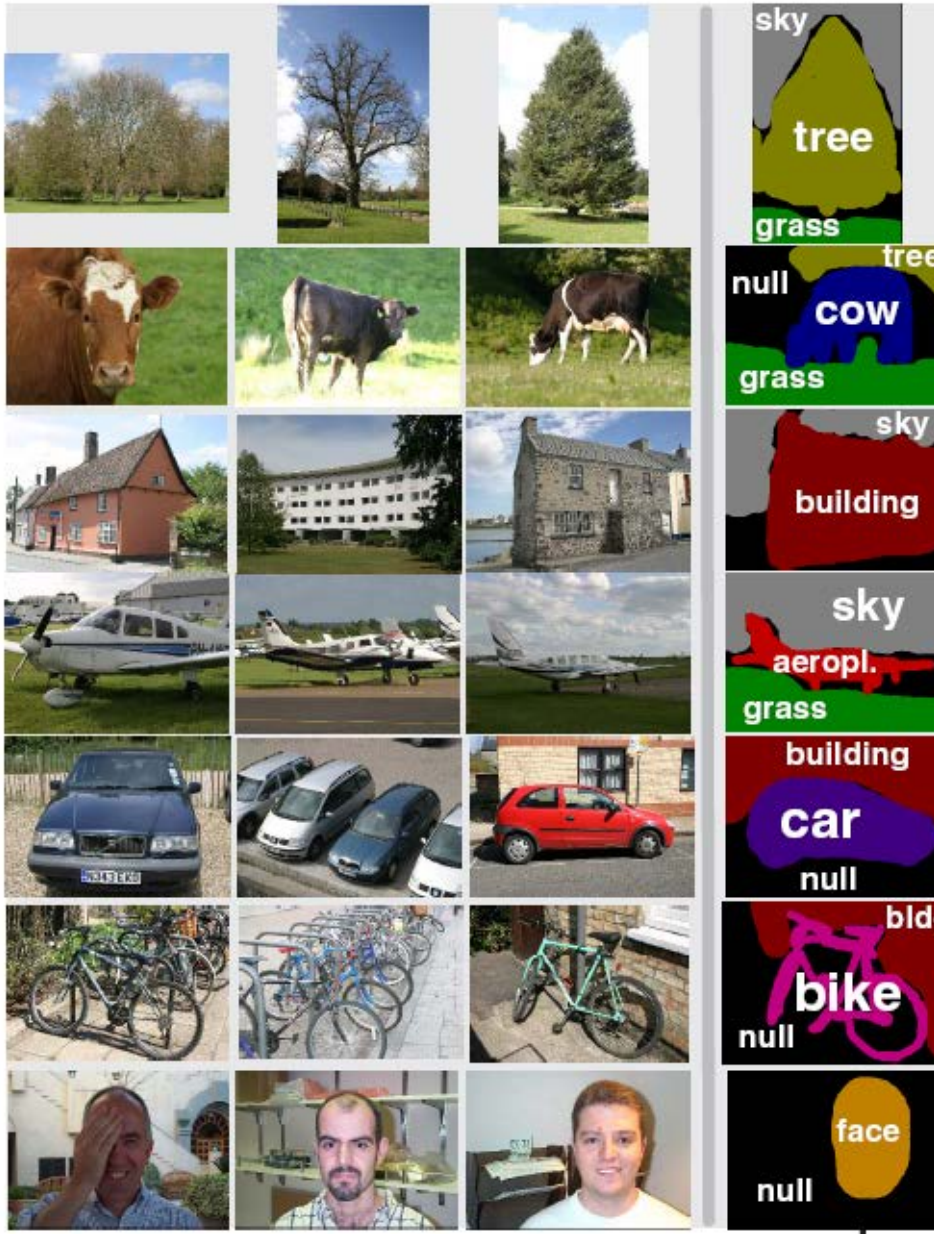
- Textons (rotation-variant)

Clustering

- K=2000
- Then clever merging
- Then fitting histogram with Gaussian

Training

- Labeled class data



Results movie

Simple is still the best!

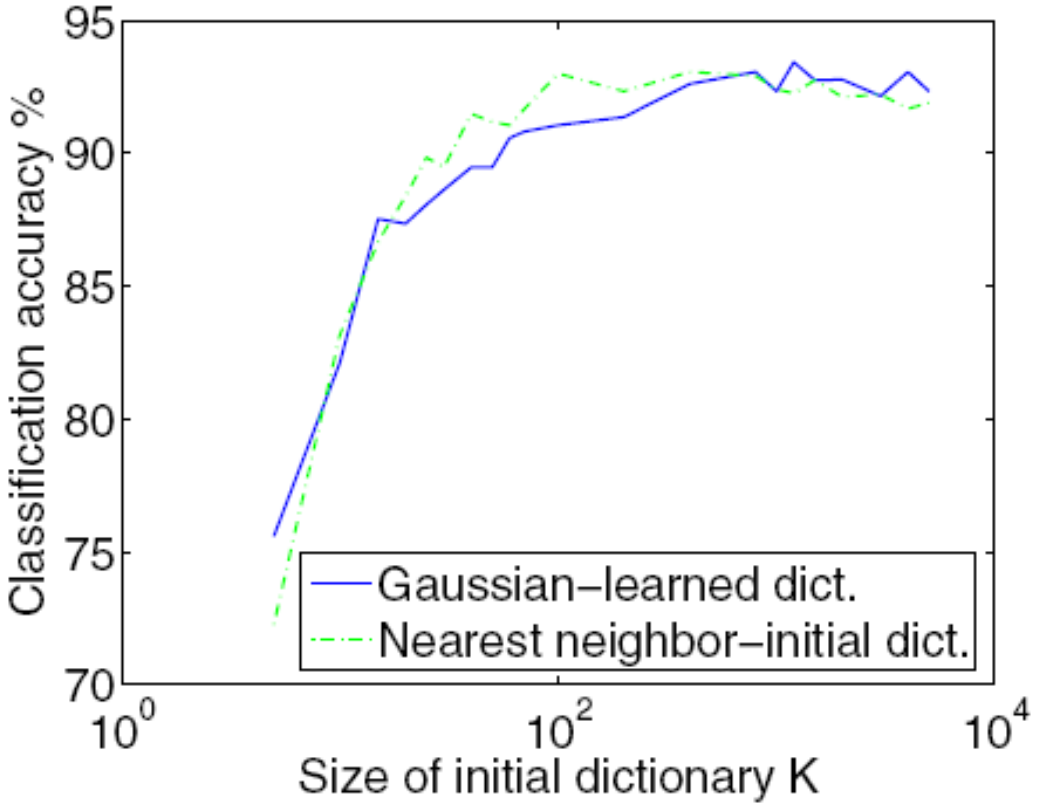


Figure 5: Comparing classification performance for Gaussian class models vs nearest neighbours classification.

Discussion

There seems to be no geometry (true/folse?), so why does it work so well?

Which sampling scheme is better you think?

Which patch representation is better (invariance vs. discriminability)?

What are the big challenges for this type of methods?

Analysis Project Grading

To get a B:

Have you met with me at least twice beforehand?

Have you done implementation/evaluation ahead of time and gotten some interesting results?

Have you presented the paper well enough to pass the speaking qualifier? Did you explain the “tricky” bits so that they make sense? Did you explain any of the prior work that might be relevant?

Have you followed up the questions in blog and in class?

Have you given me the ppt slides?

To get an A:

Have you done something creative that I didn’t ask you for?

Synthesis Project meetings

Bi-weekly

Proposed time: Tuesdays, 2-4pm

Sign up on blog