The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Fall 2022

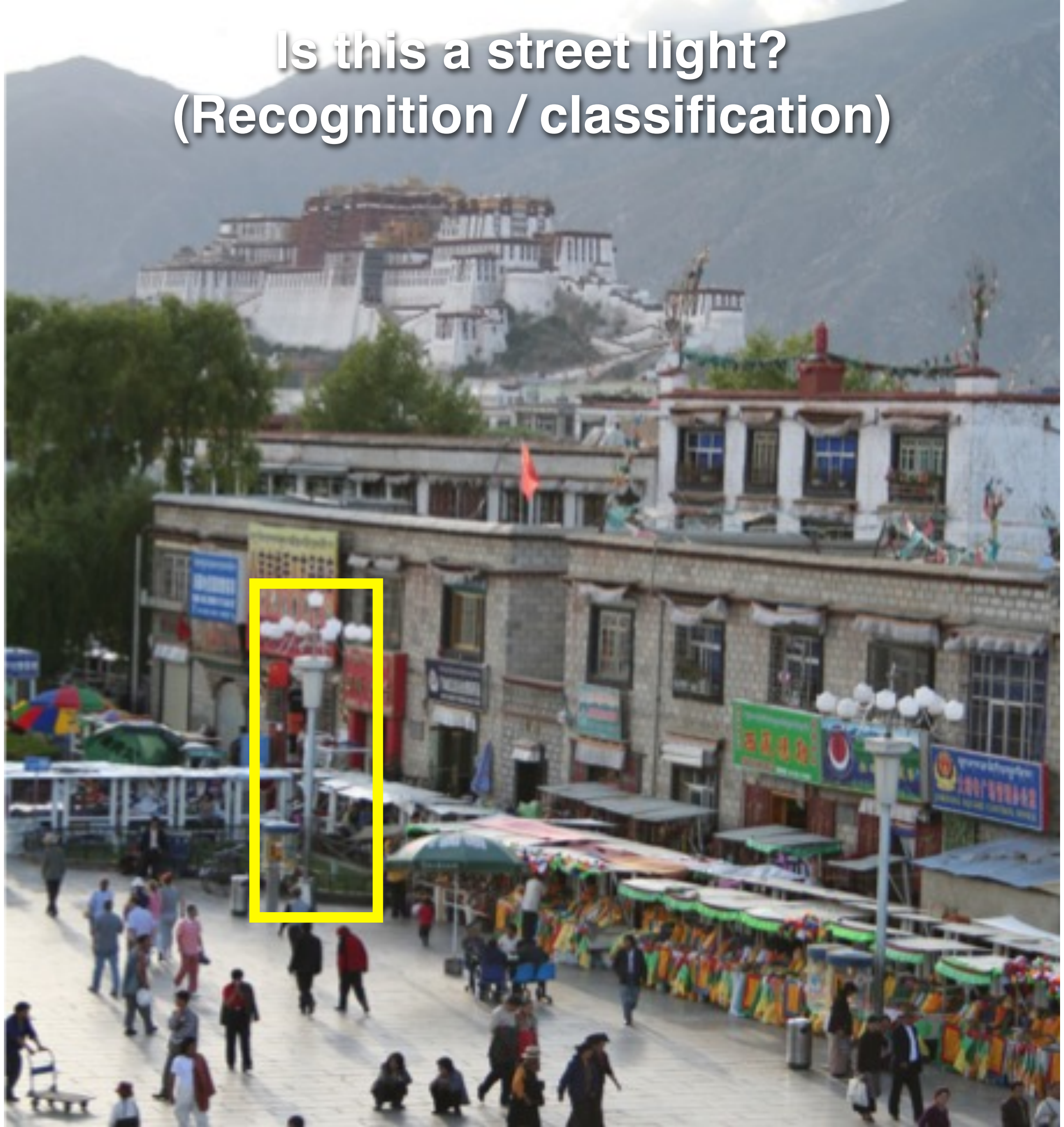
INTRODUCTION TO COMPUTER VISION

Atlas Wang
Assistant Professor, The University of Texas at Austin

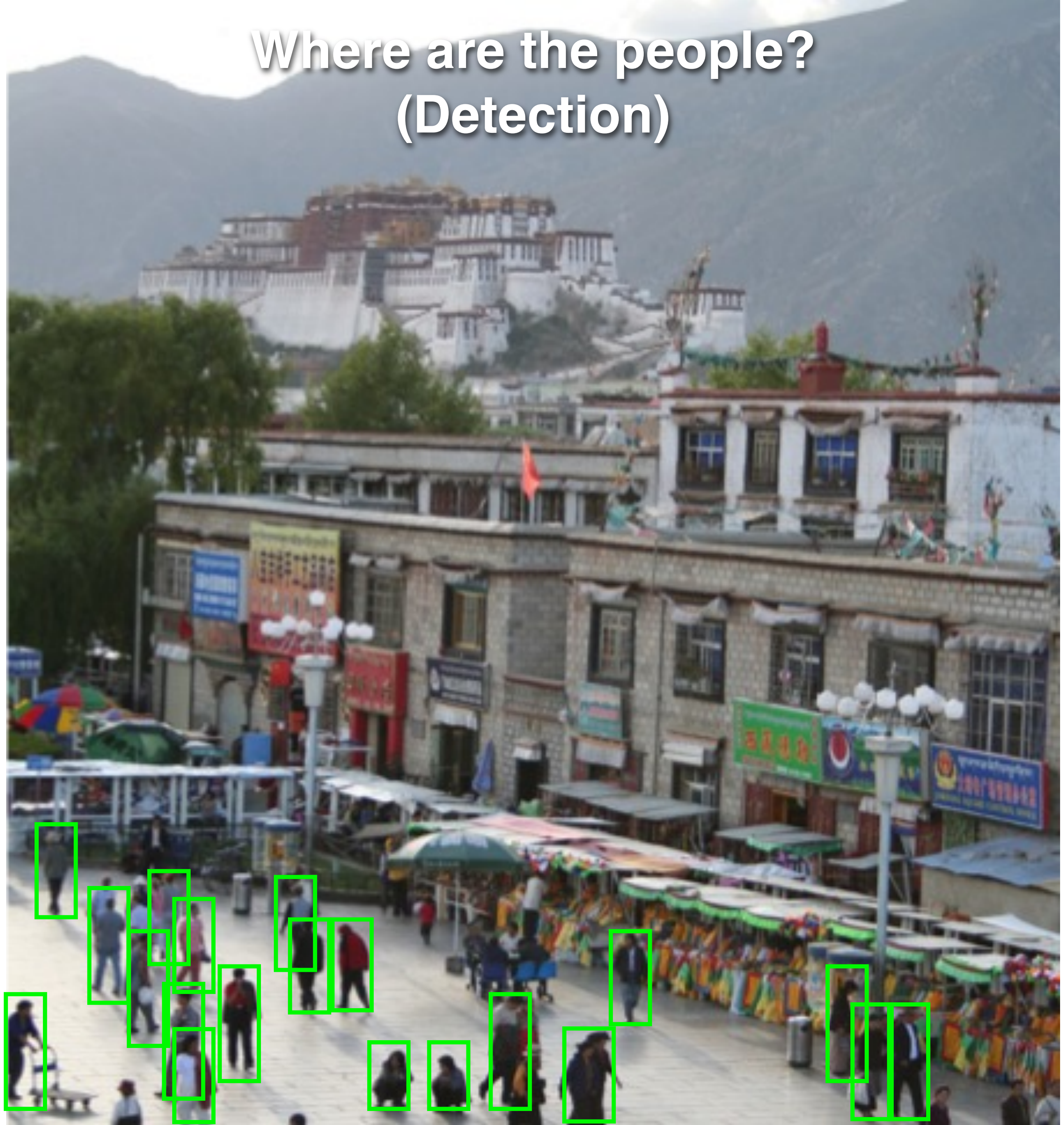
Visual Informatics Group@UT Austin
https://vita-group.github.io/

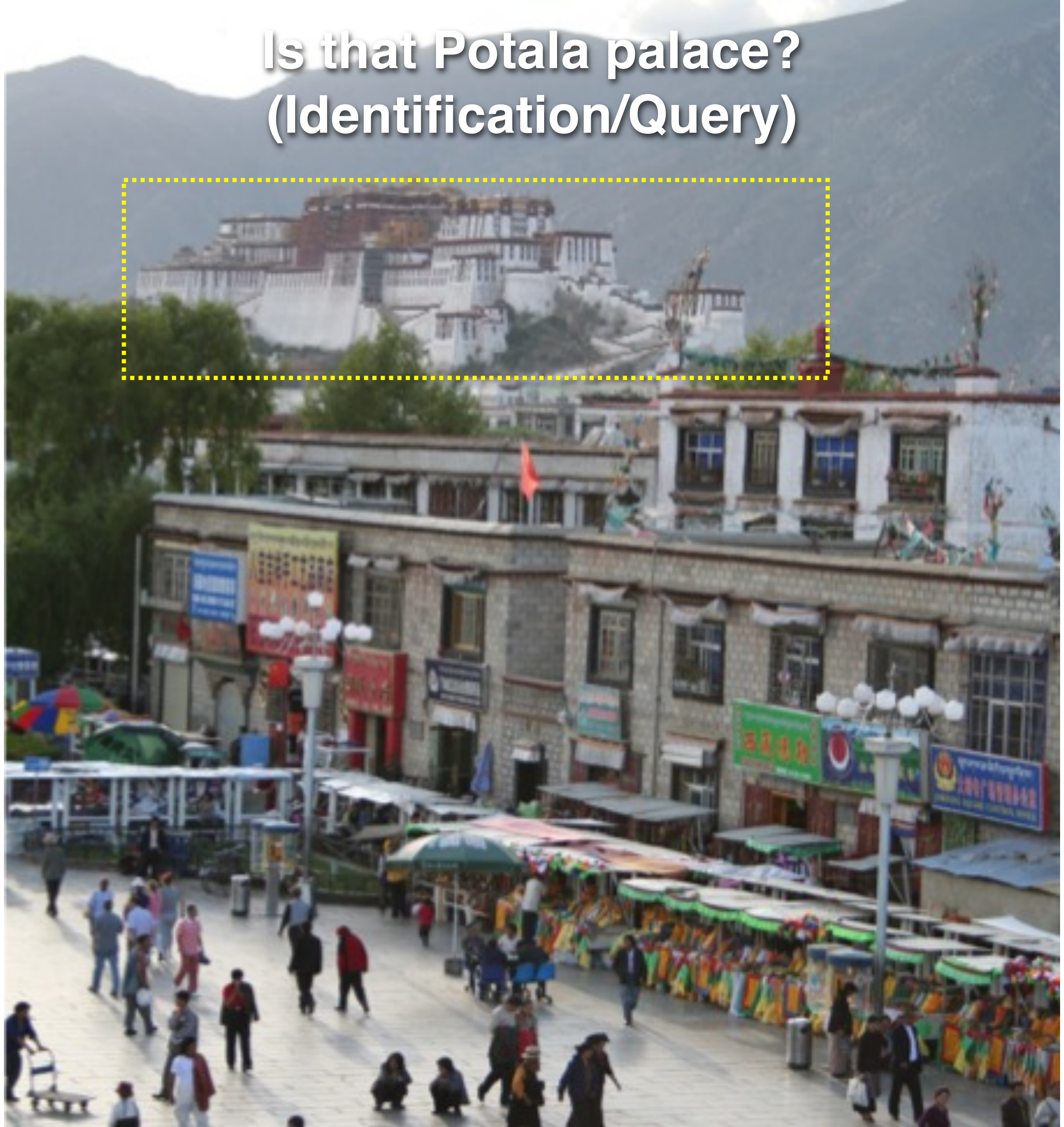
What do we mean by high-level vision or “semantic vision”?

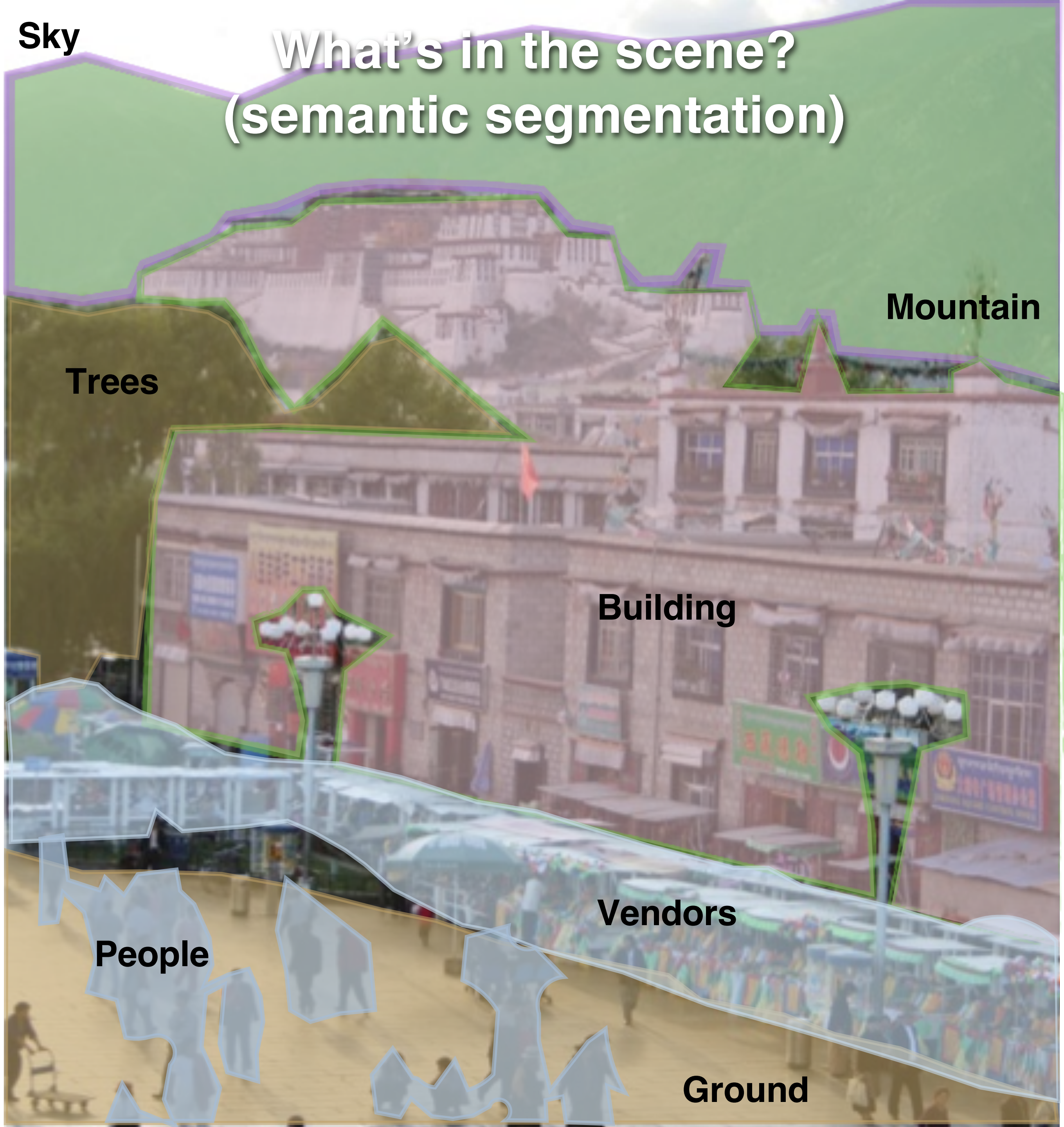
Is this a street light?
(Recognition / classification)



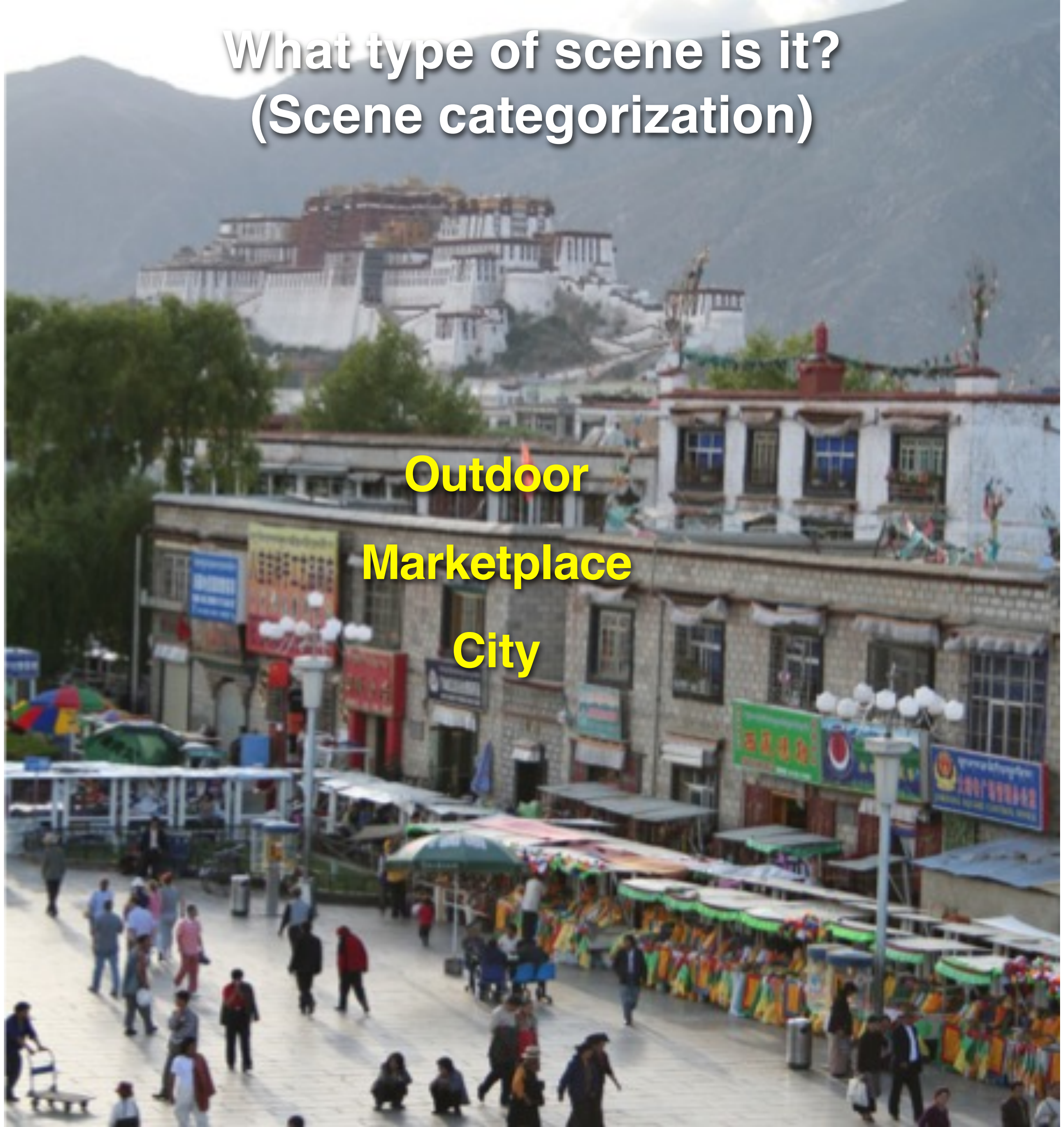
Where are the people?
(Detection)



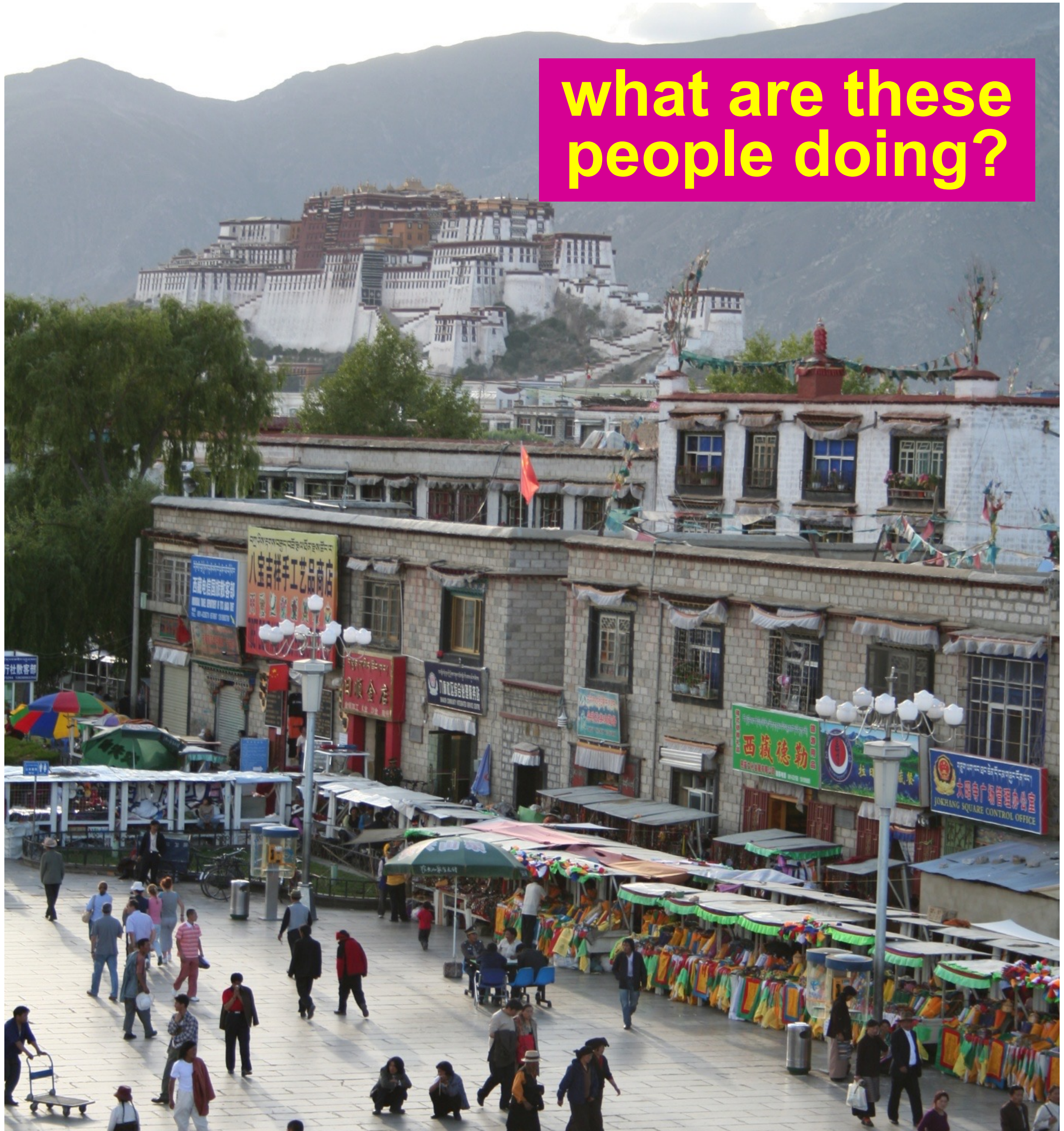

Is that Potala palace?
(Identification/Query)


Sky
What's in the scene?
(semantic segmentation)
Mountain
Trees
Building
Vendors
People
Ground

What type of scene is it?
(Scene categorization)



Activity / Event Recognition

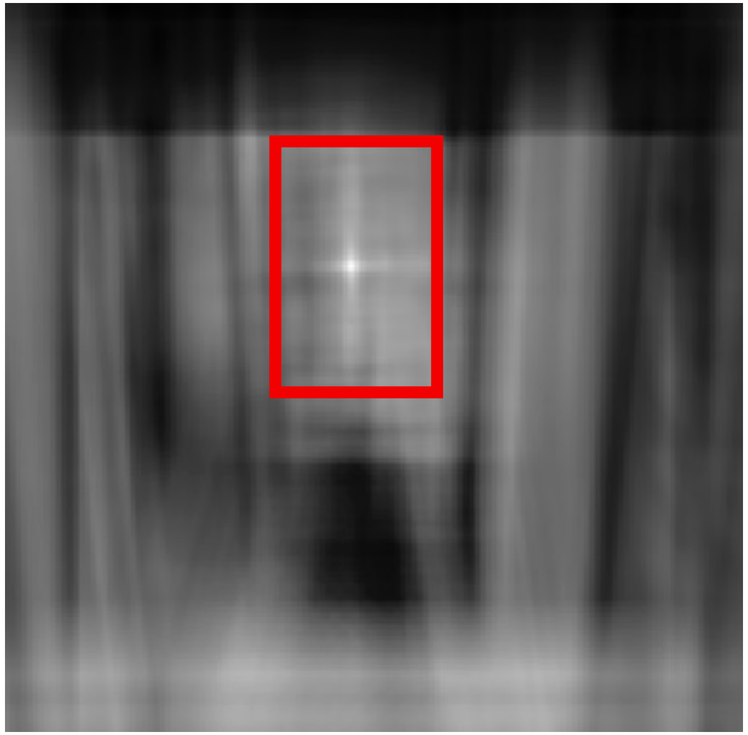


Object recognition
Is it really so hard?

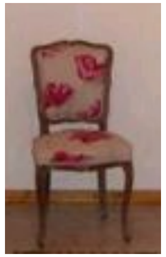
This is a chair

Find the chair in this image

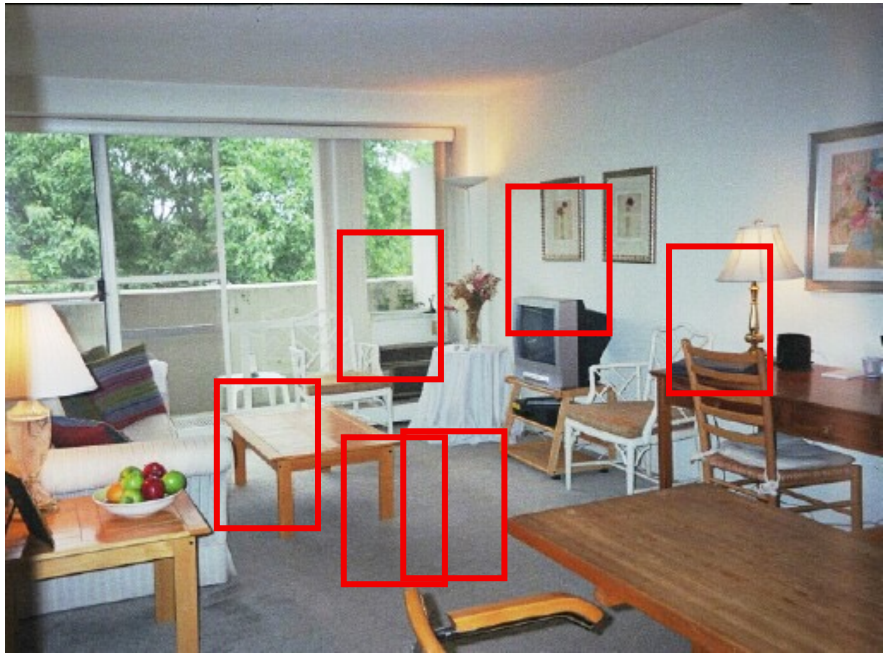
Output of normalized correlation

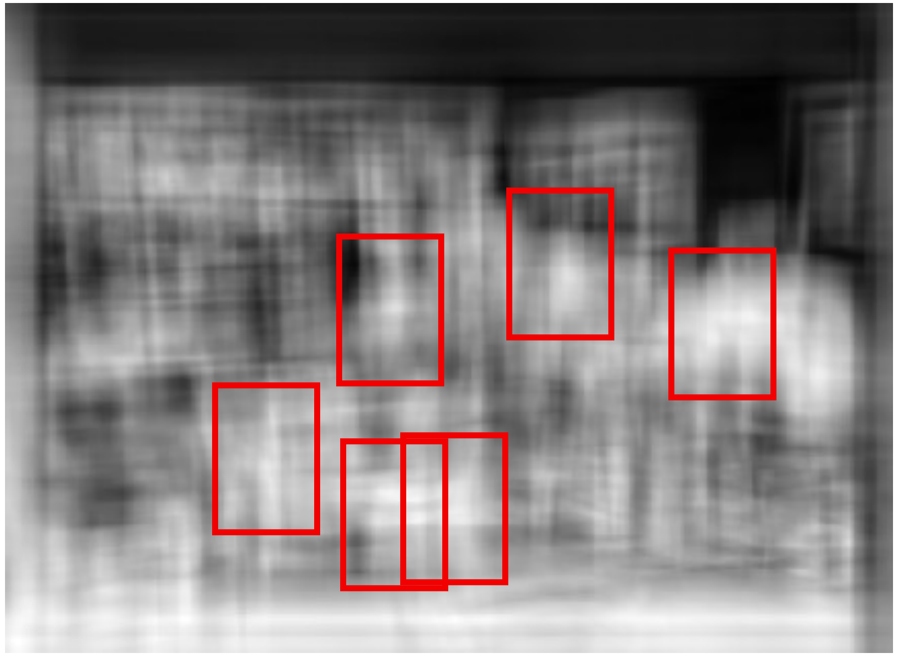


Object recognition
Is it really so hard?



Find the chair in this image

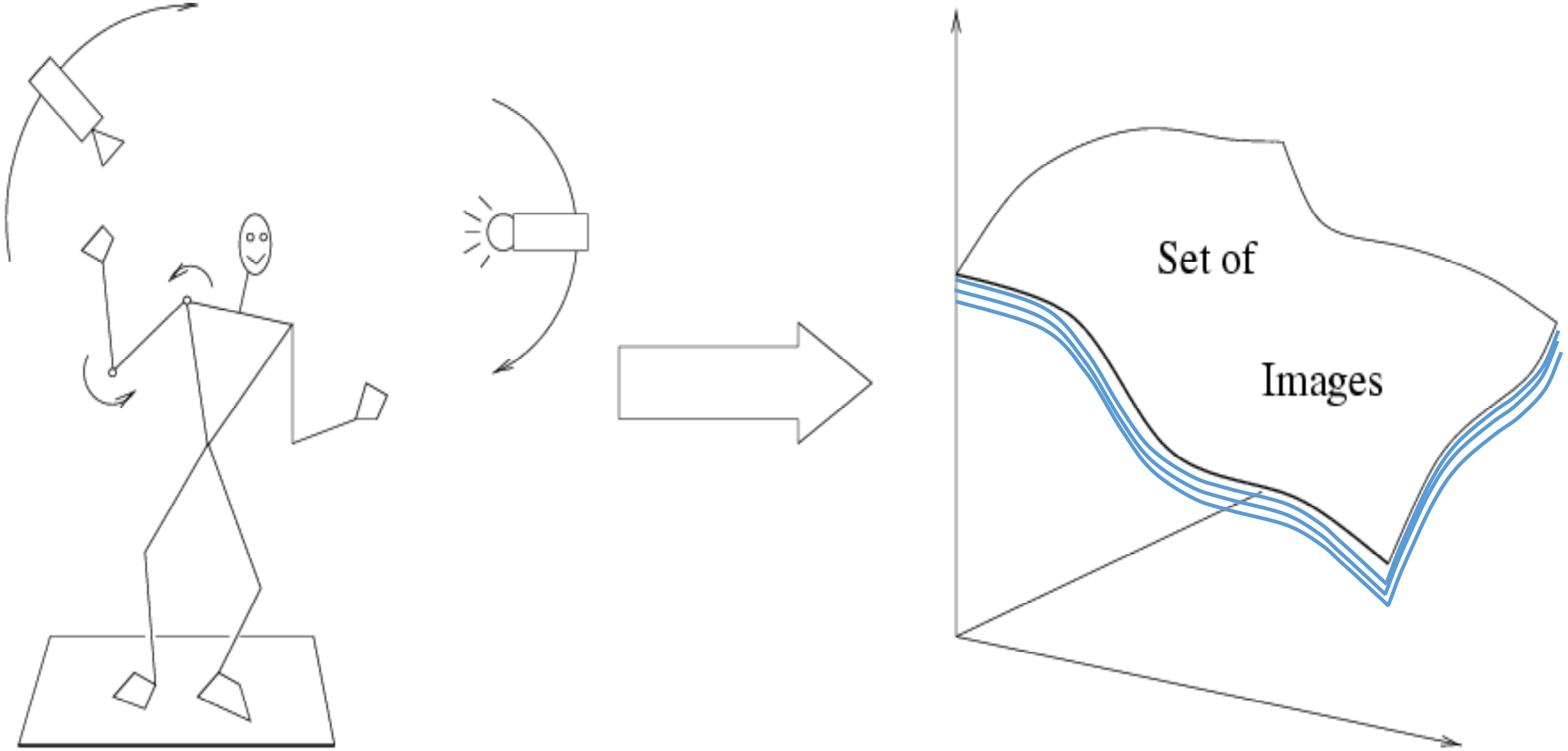




Pretty much garbage
Simple template matching is not going to make it

A "popular method is that of template matching, by point to point correlation of a model pattern with the image pattern. These techniques are inadequate for three-dimensional scene analysis for many reasons, such as occlusion, changes in viewing angle, and articulation of parts." Nivatia & Binford, 1977.

Why is this hard?



Variability: Camera position
Illumination
Shape parameters

How many object categories are there?

Challenge: variable viewpoint

Michelangelo 1475-1564

Challenge: variable illumination

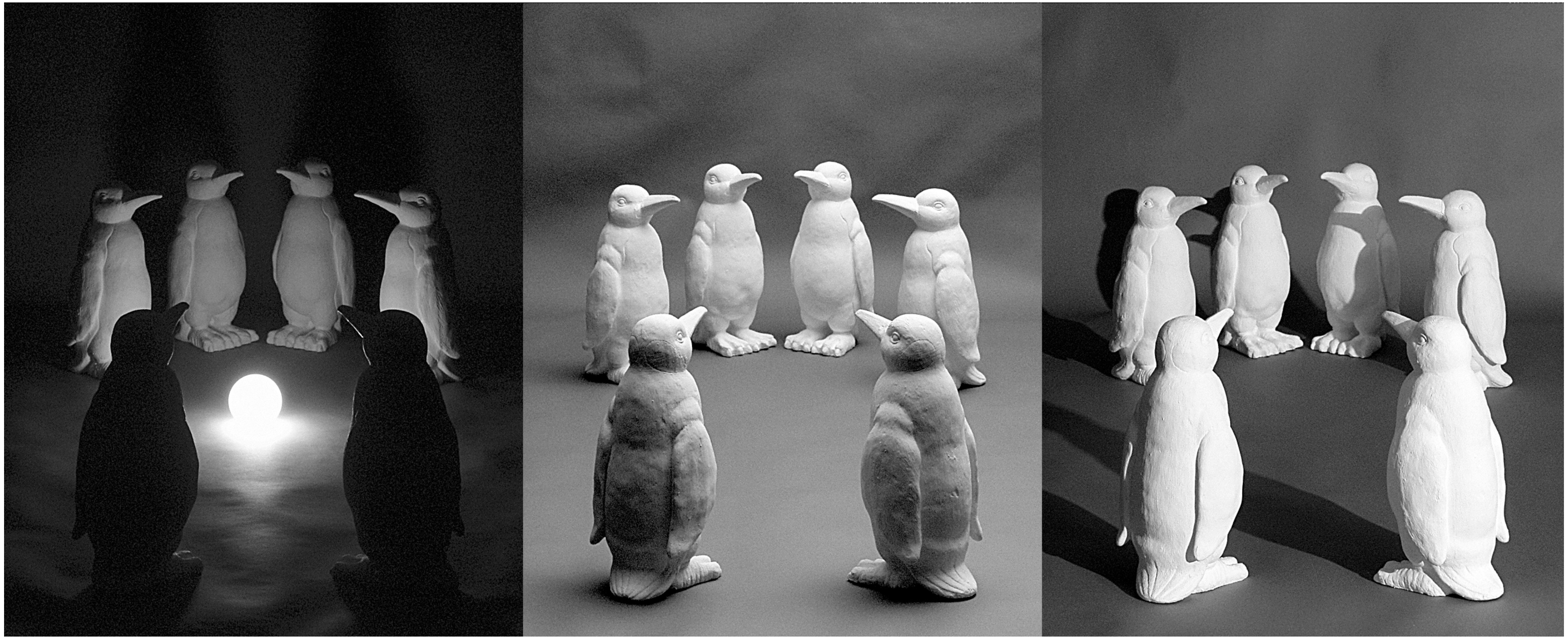


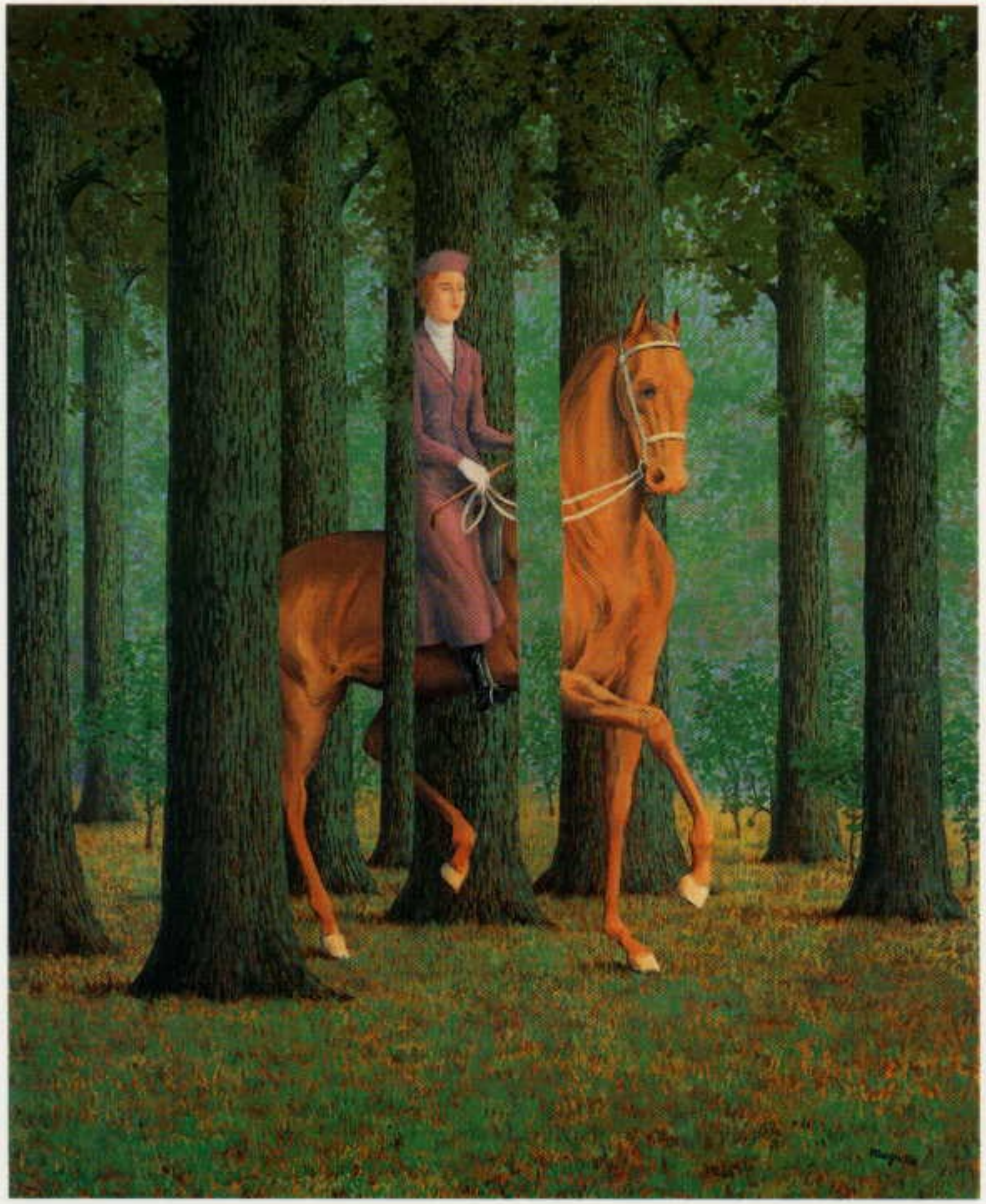
image credit: J. Koenderink

Challenge: scale

Challenge: deformation

Deformation

Challenge: Occlusion



Magritte, 1957

Challenge: background clutter

Challenge: intra-class variations





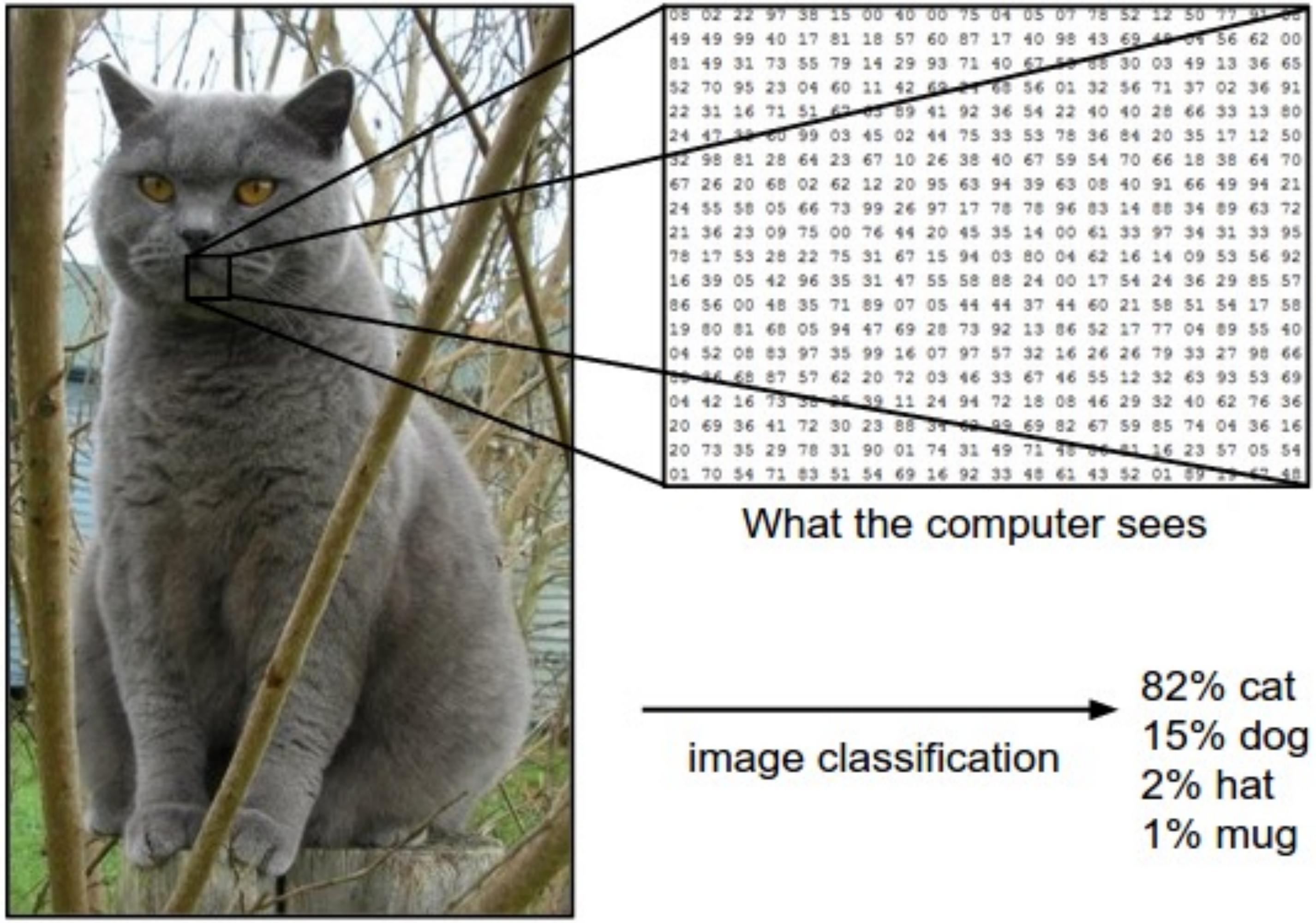

Svetlana Lazebnik

Image Classification

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

→ cat

Image Classification: Problem



Data-driven approach

- Collect a database of images with labels
- Use ML to train an image classifier
- Evaluate the classifier on test images

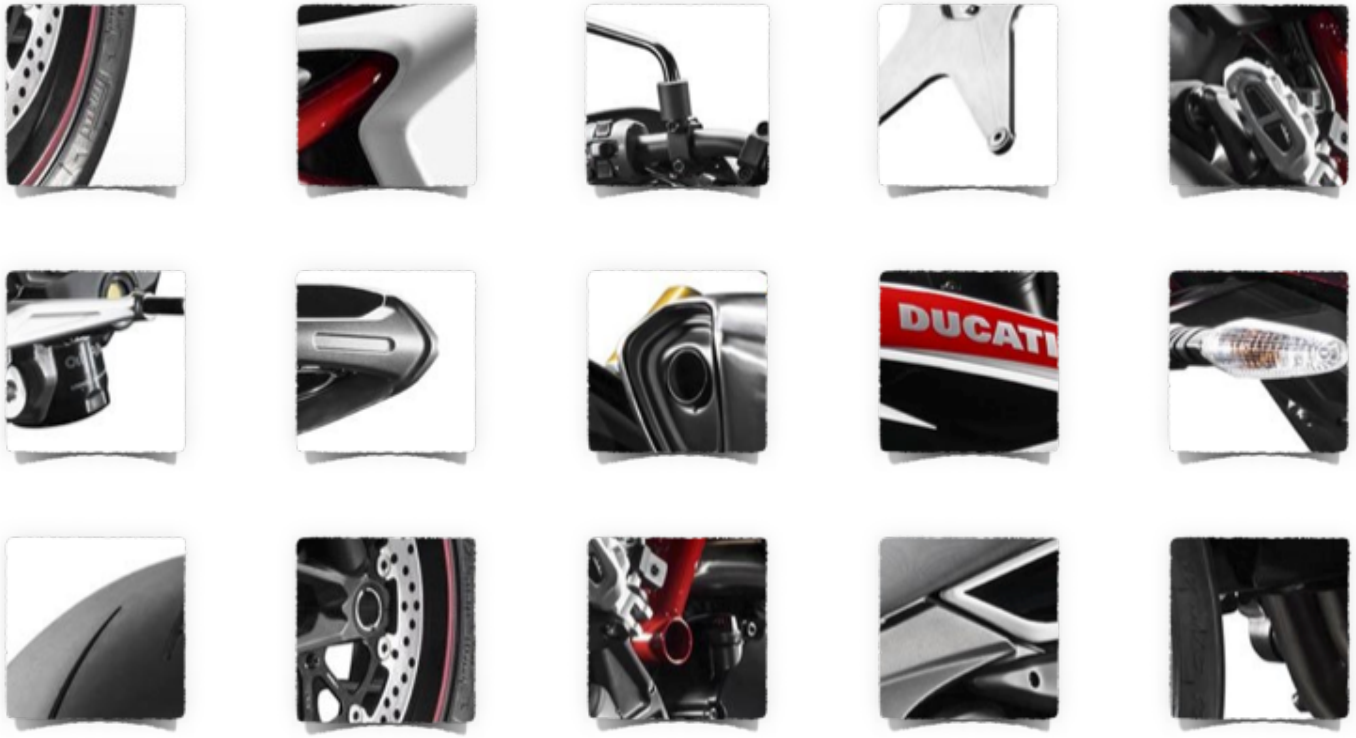
Example training set

History of Image Classfication

- 1960s –early 1990s: the geometric era
  - Recognition as an alignment problem: the simple “toy block” world …
- 1990s: appearance-based models
  - PCA (eigenface), color histogram …
- Mid-1990s: sliding window/template approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features (**Today**)
- *Present trends:* deep learning (we will get there)

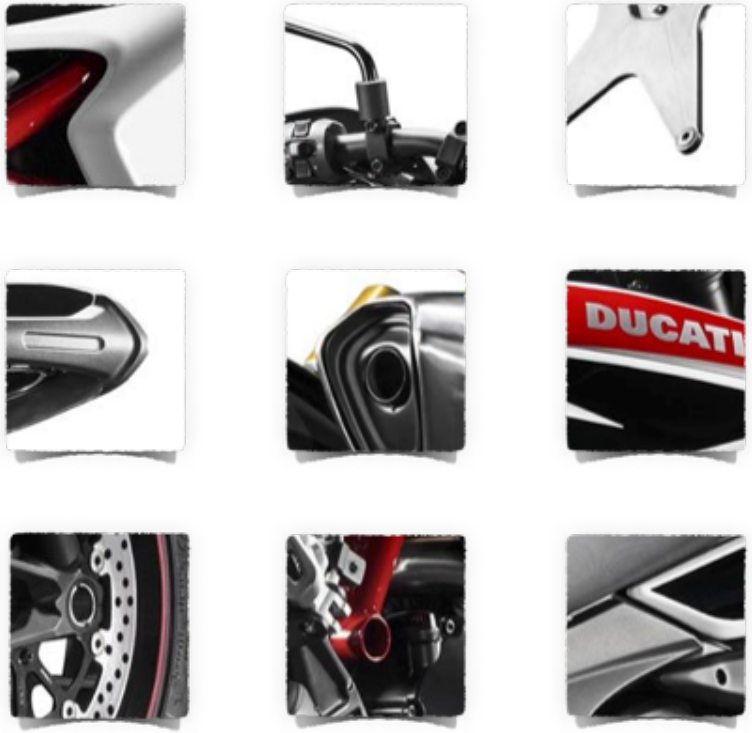
Bag of words

What object do these parts belong to?



Some local feature are very informative

An object as

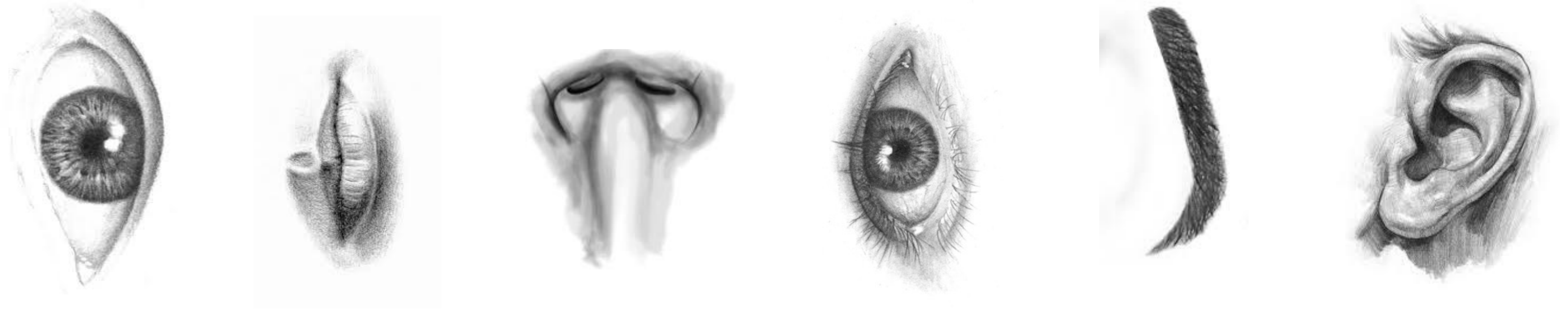


a collection of local features

(bag-of-features)

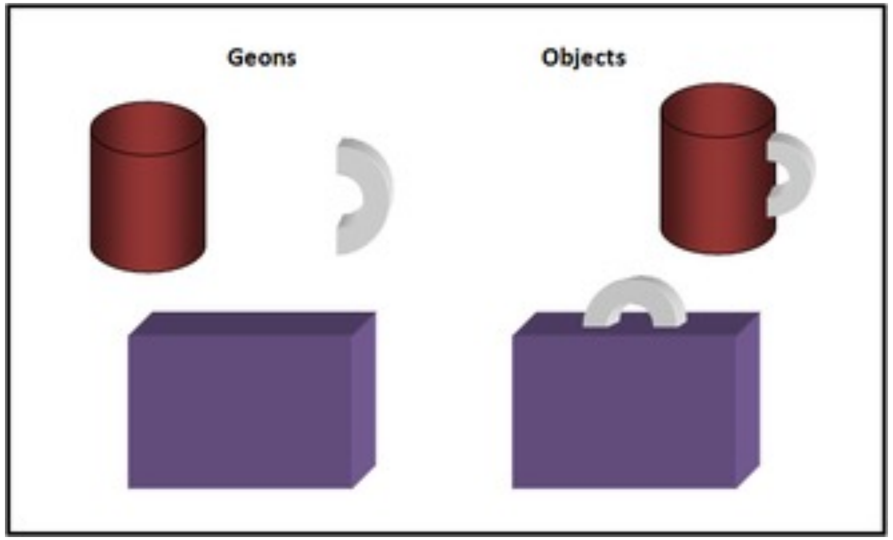
- deals well with occlusion
- scale invariant
- rotation invariant

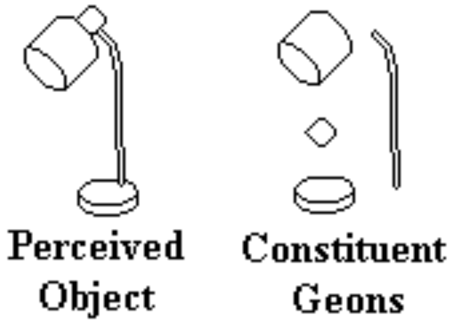
(not so) crazy assumption



spatial information of local features
can be ignored for object recognition (i.e., verification)

Recognition-by-Components (RBC) Theory (1987)





- A human learning theory to explain object recognition
- According to RBC theory, we are able to recognize objects by separating them into **geons** (the object’s main component parts).
- Geons are based on basic 3-dimensional shapes (cylinders, cones, etc.) that can be assembled in various arrangements to form a virtually unlimited number of objects.
- **Very impactful for computer vision recognition!**

Bag-of-features

represent a data item (document, texture, image)
as a histogram over features

an old idea

(e.g., texture recognition and information retrieval)

Vector Space Model

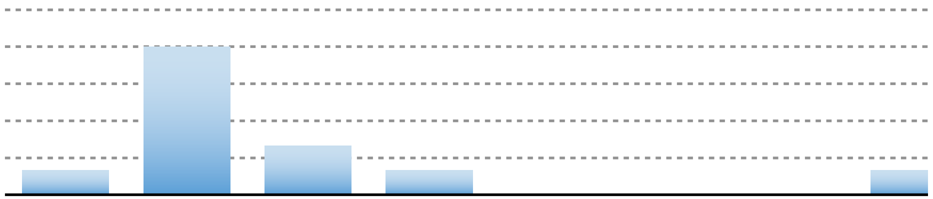
G. Salton. 'Mathematics and Information Retrieval' Journal of Documentation, 1979

The Newspa

Sunday, December 22, 2013

DARPA Selects Carnegie Me

The Tartan Rescue Team from Carnegie Mellon University's National Robotics Engineering Center ranked third among teams competing in the Defense Advanced Research Projects Agency (DARPA) Robotics Challenge Trials this weekend in Homestead, Fla., and was selected by the agency as one of eight teams eligible for DARPA funding to prepare for next December's finals. The team's four-limbed CMU Highly Intelligent Mobile Platform, or CHIMP, robot scored 18 out of a possible 32 points during the two-day trials. It demonstrated its ability to perform such tasks as removing debris, cutting a hole through a wall and closing a series of valves.



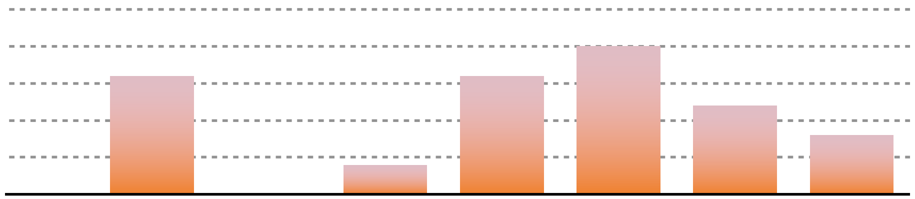
1	6	2	1	0	0	0	1
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor

Tartan Tim

Monday, January 20, 2014

Bio-Inspired Robotic Device

PITTSBURGH—A soft, wearable device that mimics the muscles, tendons and ligaments of the lower leg could aid in the rehabilitation of patients with ankle-foot disorders such as drop foot, said Yong-Lae Park, an assistant professor of robotics at Carnegie Mellon University. Park, working with collaborators at Harvard University, the University of Southern California, MIT and BioSensics, developed an active orthotic device using soft plastics and composite materials, instead of a rigid exoskeleton. The soft materials, combined with pneumatic artificial muscles (PAMs), lightweight sensors and advanced control software, made it possible for the robotic device to achieve natural motions in the ankle.



0	4	0	1	4	5	3	2
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor

A document (datapoint) is a vector of counts over each word (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

just a histogram over words

$n(\cdot)$ counts the number of occurrences

What is the similarity between two documents?

A document (datapoint) is a vector of counts over each word (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \;\; n(w_{2,d}) \;\; \cdots \;\; n(w_{T,d})]$$

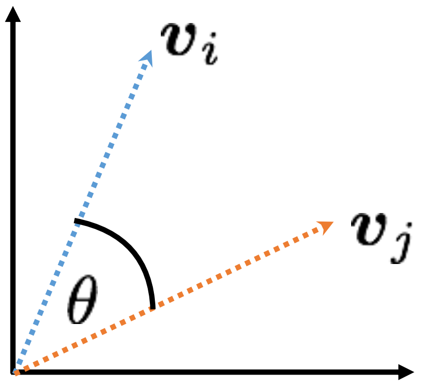
just a histogram over words

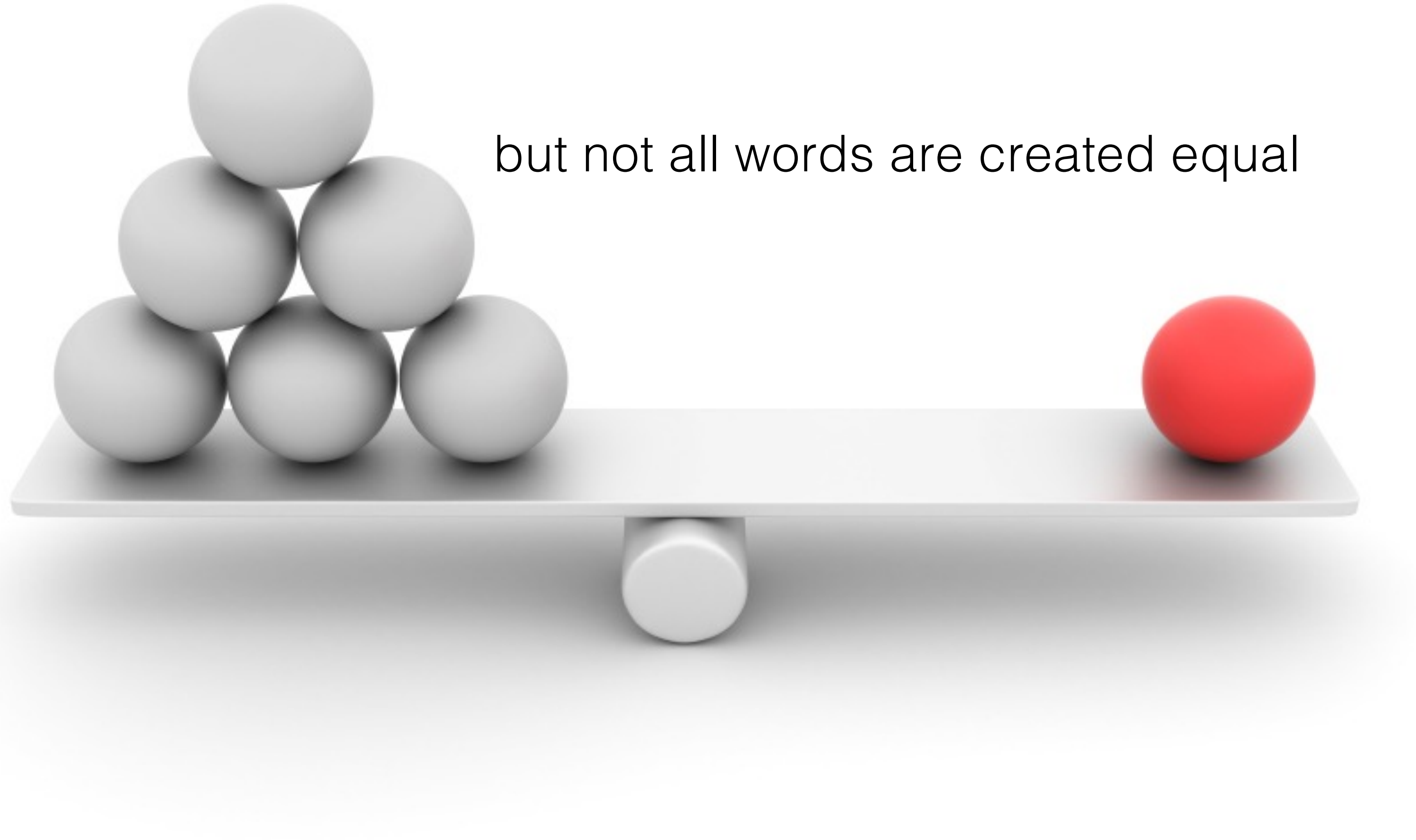
$n(\cdot)$ counts the number of occurrences

What is the similarity between two documents?

Use any distance you want but the cosine distance is fast.

$$d(\boldsymbol{v}_i, \boldsymbol{v}_j) = \cos\theta$$
$$= \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|}$$





but not all words are created equal

TF-IDF

Term **F**requency **I**nverse **D**ocument **F**requency

$$\boldsymbol{v}_d = \begin{bmatrix} n(w_{1,d}) & n(w_{2,d}) & \cdots & n(w_{T,d}) \end{bmatrix}$$

weigh each word by a heuristic

$$\boldsymbol{v}_d = \begin{bmatrix} n(w_{1,d})\alpha_1 & n(w_{2,d})\alpha_2 & \cdots & n(w_{T,d})\alpha_T \end{bmatrix}$$

$$n(w_{i,d})\alpha_i = \underbrace{n(w_{i,d})}_{\text{term frequency}} \underbrace{\log\left\{\frac{D}{\sum_{d'} \mathbf{1}[w_i \in d']}\right\}}_{\text{inverse document frequency}}$$

(down-weights **common** terms)

Standard BOW pipeline
(for image classification)

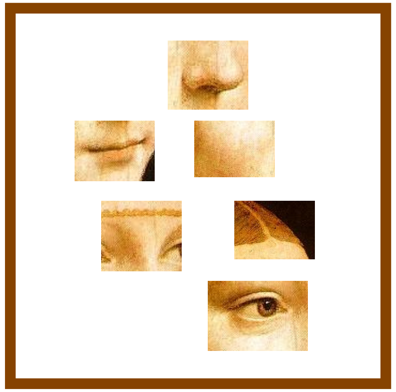
Dictionary Learning:
Learn Visual Words using clustering

Encode:
build Bags-of-Words (BOW) vectors
for each image

Classify:
Train and test data using BOWs

Dictionary Learning:
Learn Visual Words using clustering

1. extract features (e.g., SIFT) from images

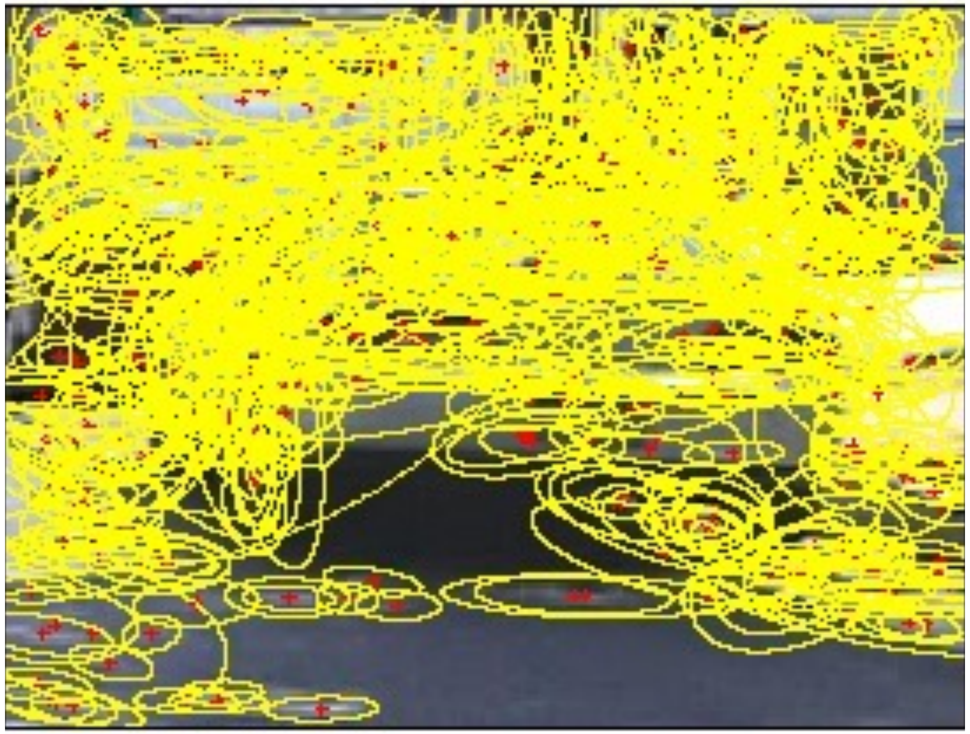
Dictionary Learning:

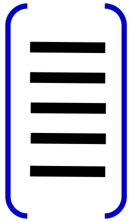
Learn Visual Words using clustering

2. Learn visual dictionary (e.g., K-means clustering)

What kinds of features can we extract?

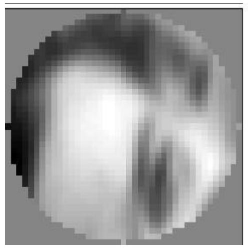
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic et al. 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation-based patches (Barnard et al. 2003)



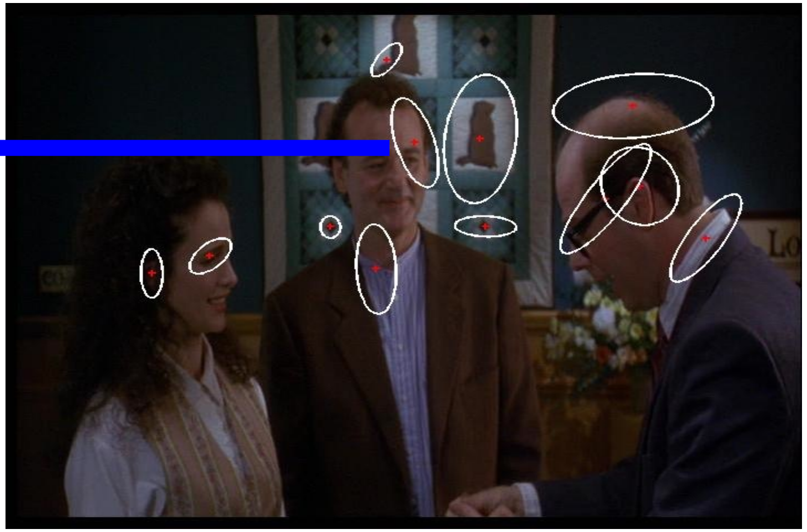


Compute SIFT descriptor

[Lowe’99]



Normalize patch

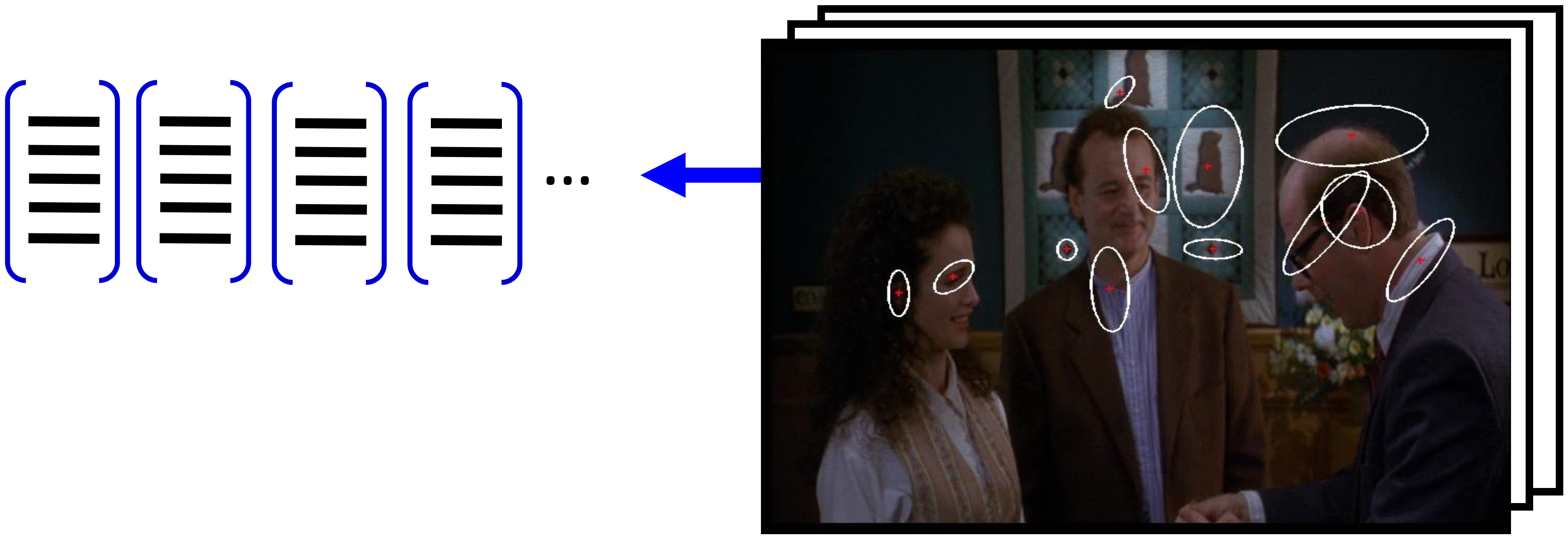


Detect patches

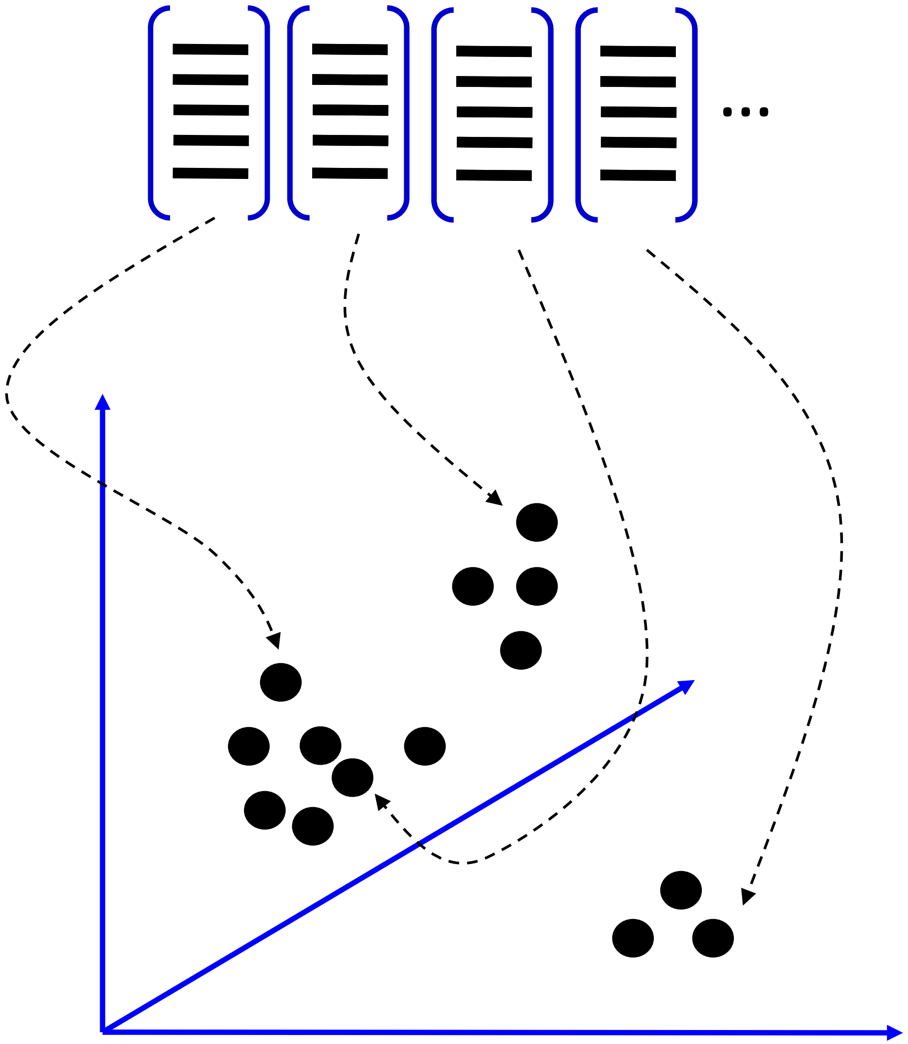
[Mikojaczyk and Schmid ’02]

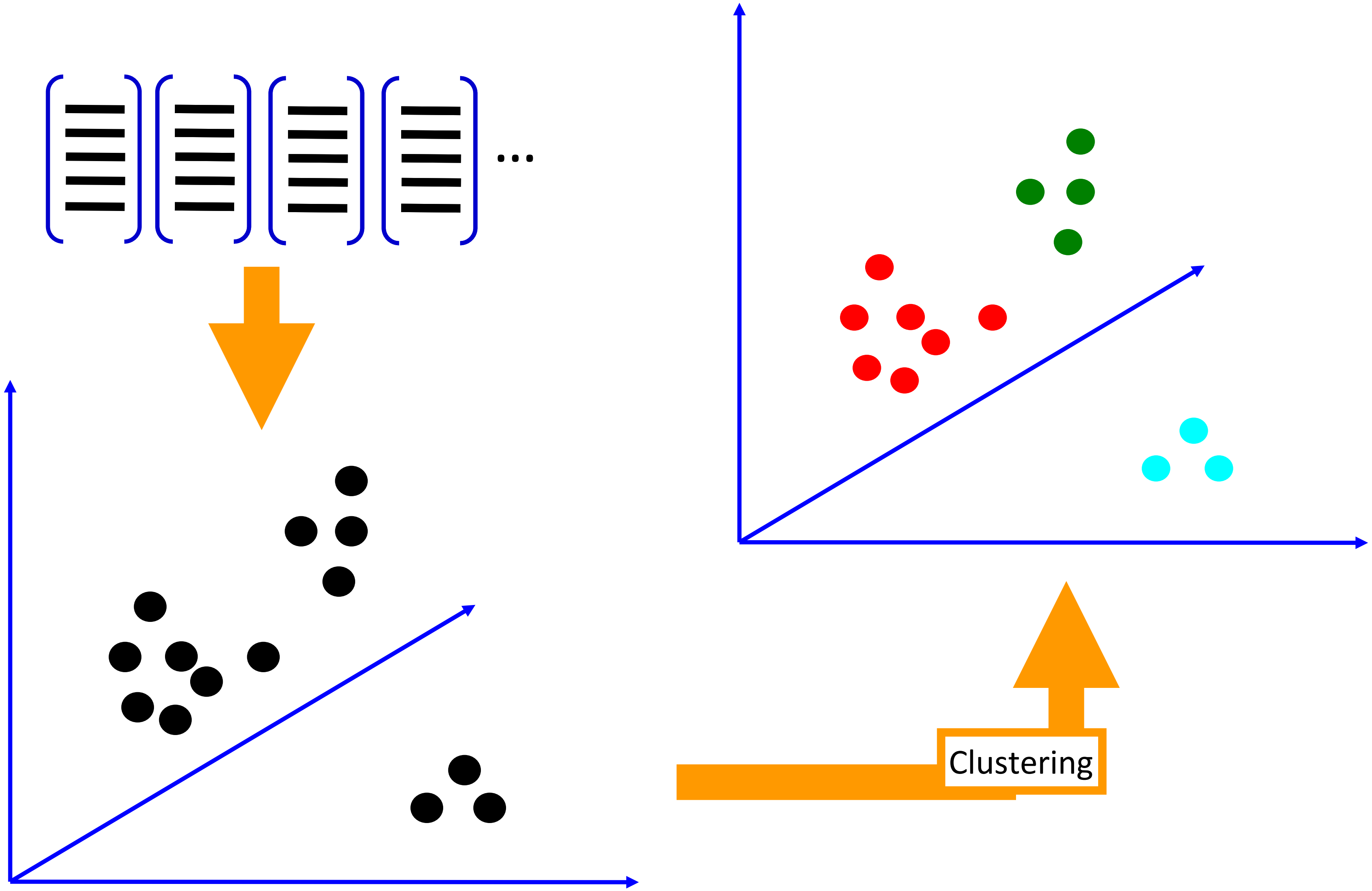
[Mata, Chum, Urban & Pajdla, ’02]

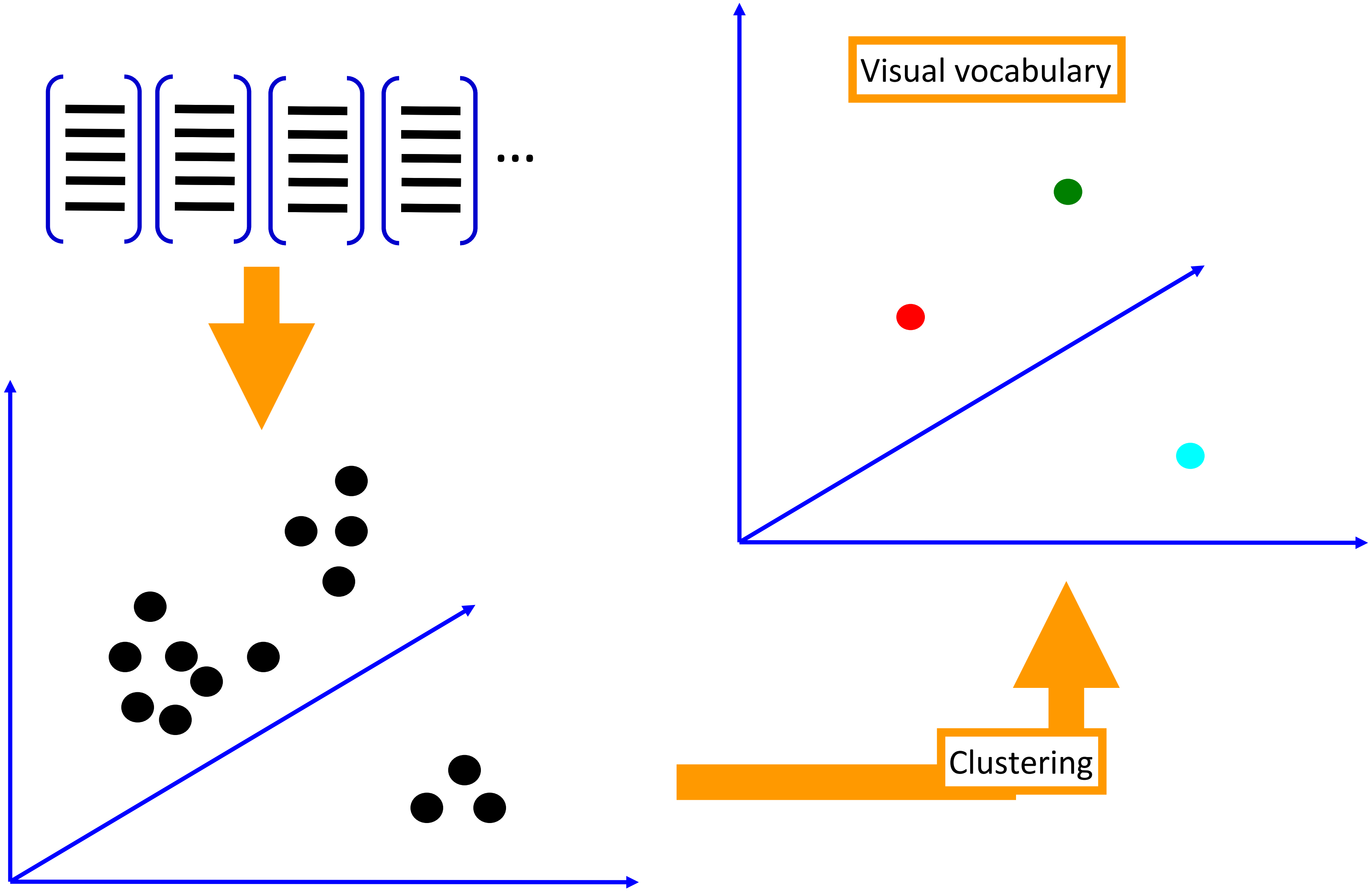
[Sivic & Zisserman, ’03]



How do we learn the dictionary?



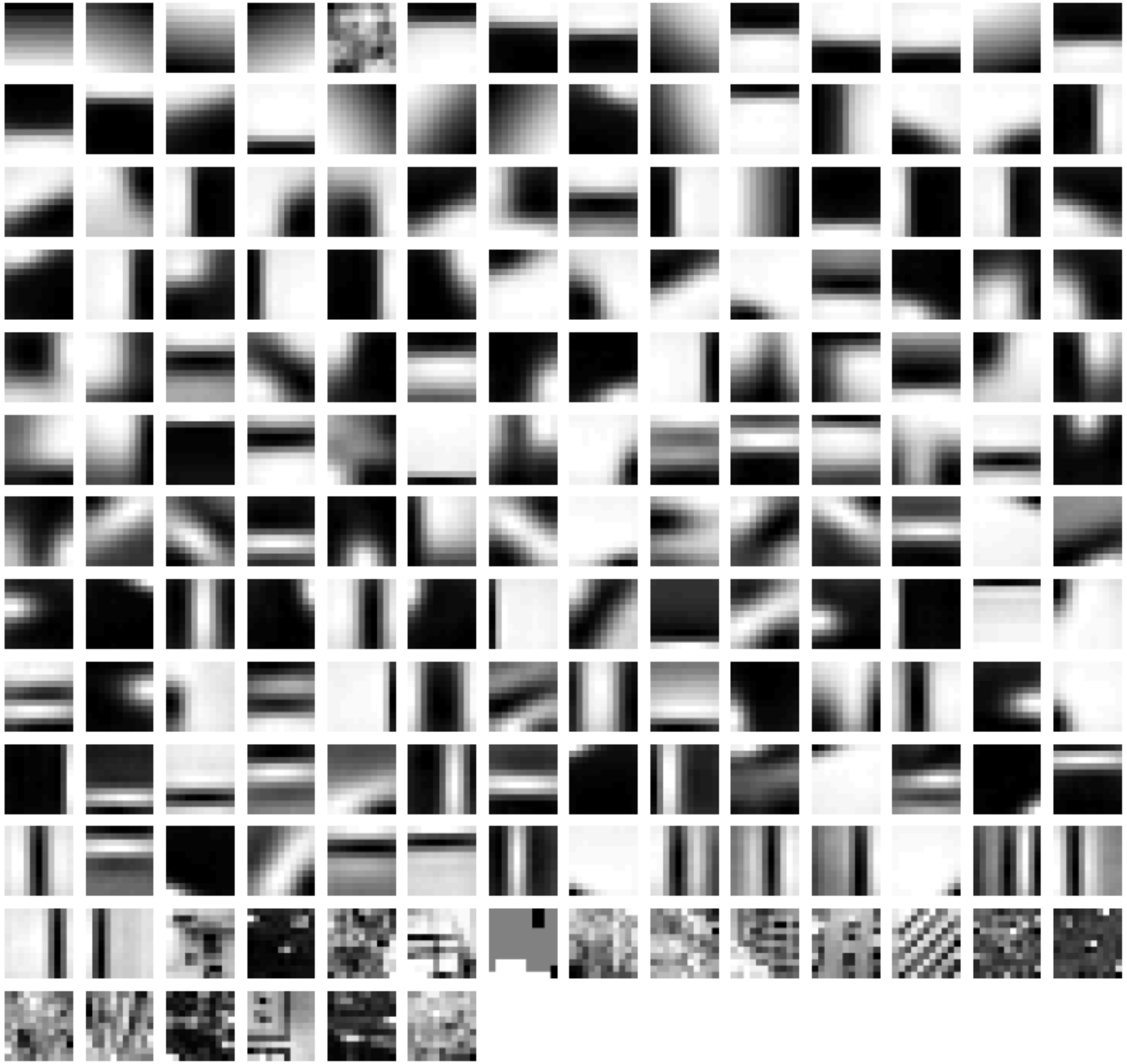

Clustering


Visual vocabulary
Clustering

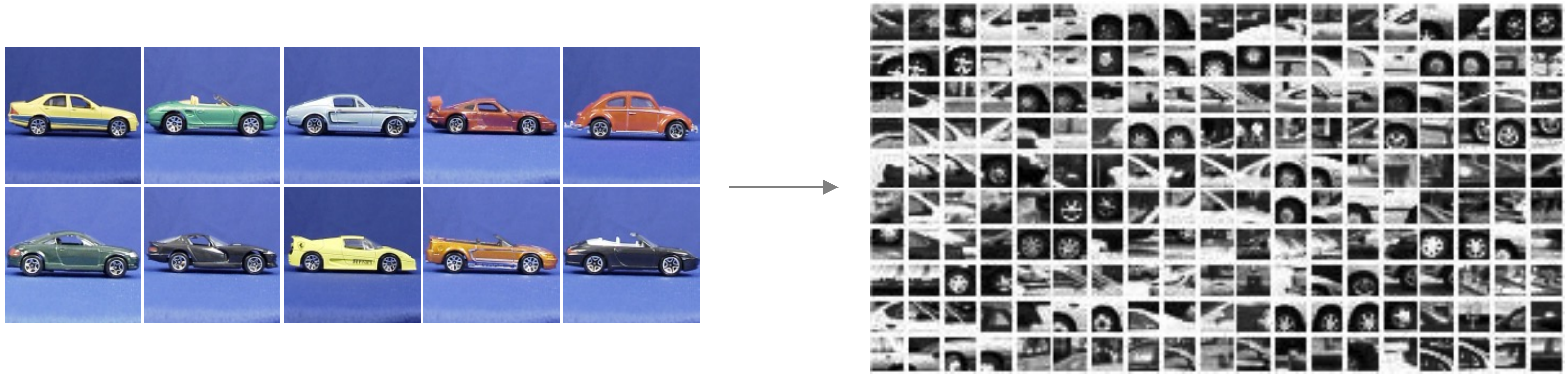
*From what **data** should I learn the dictionary?*

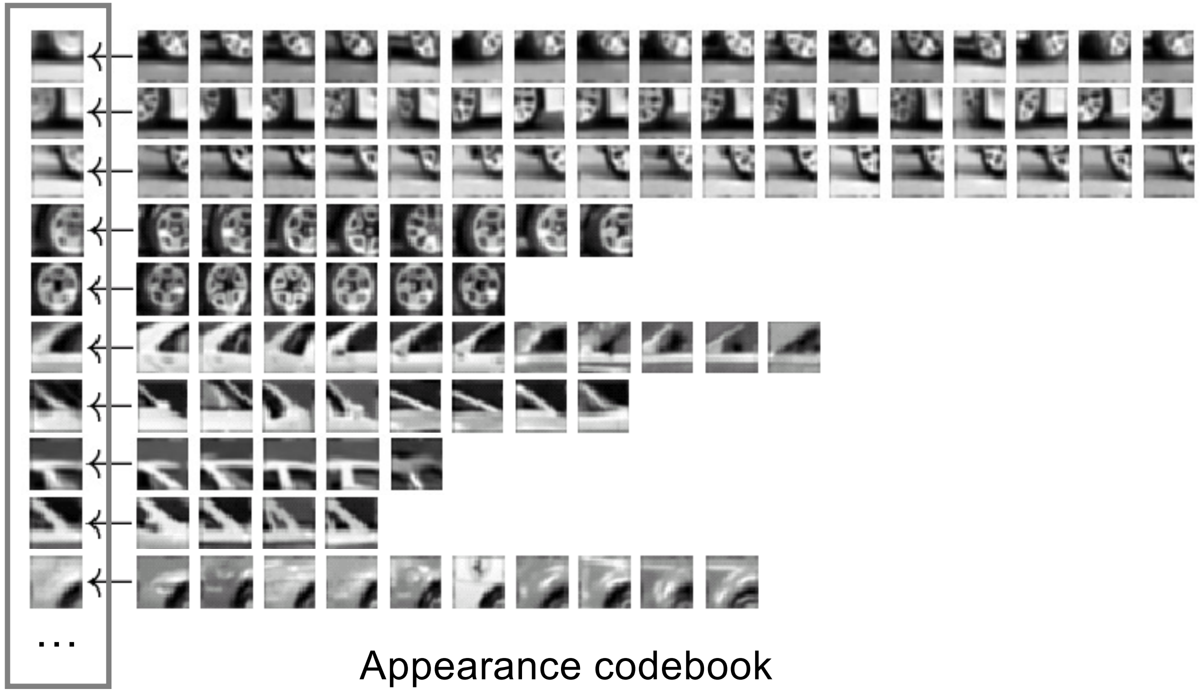
- Dictionary can be learned on separate training set
- Provided the training set is sufficiently representative, the dictionary will be “universal”

Example visual dictionary



Example dictionary

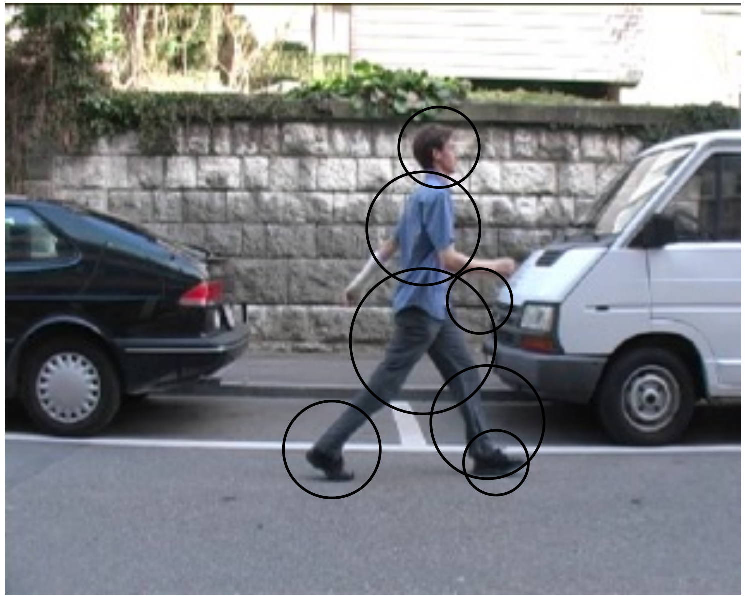


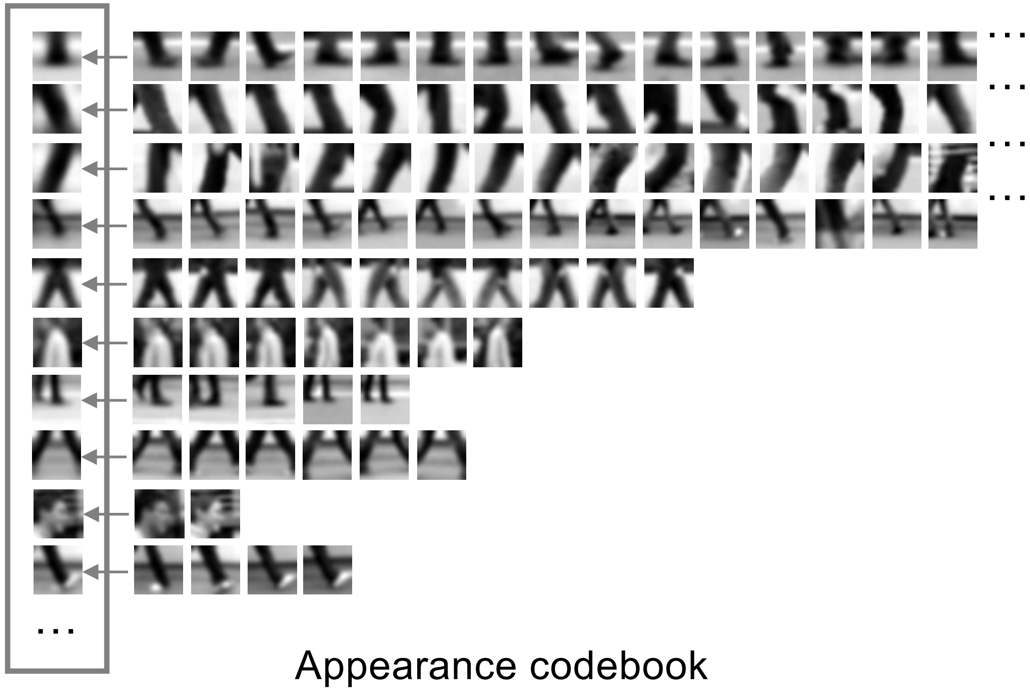


Appearance codebook

Source: B. Leibe

Another dictionary





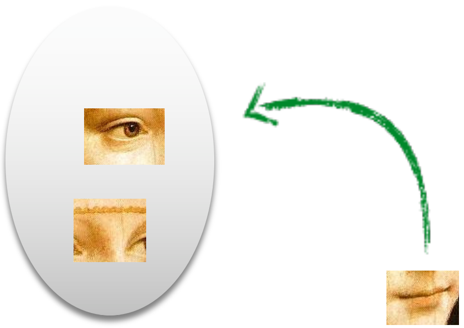
Appearance codebook

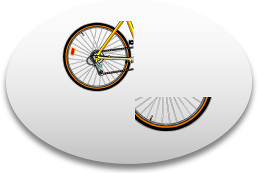
Source: B. Leibe

Dictionary Learning:
Learn Visual Words using clustering

Encode:
build Bags-of-Words (BOW) vectors
for each image

Classify:
Train and test data using BOWs

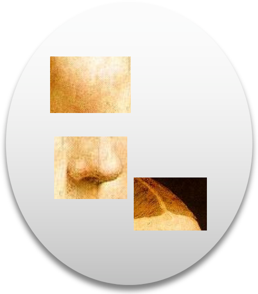




1. Quantization: image features gets associated to a visual word (nearest cluster center)

Encode:

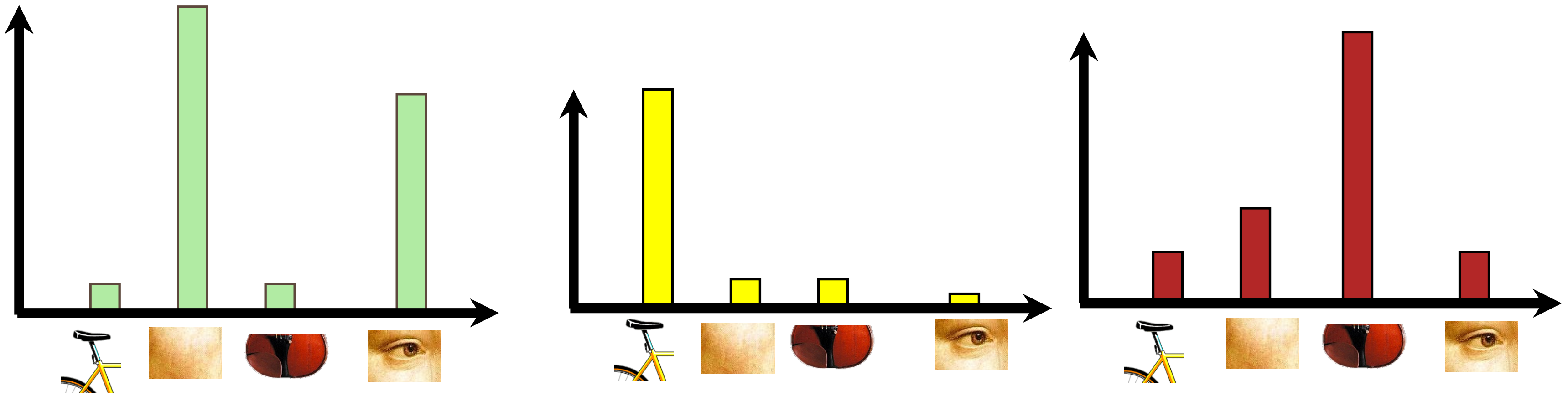
build Bags-of-Words (BOW) vectors for each image

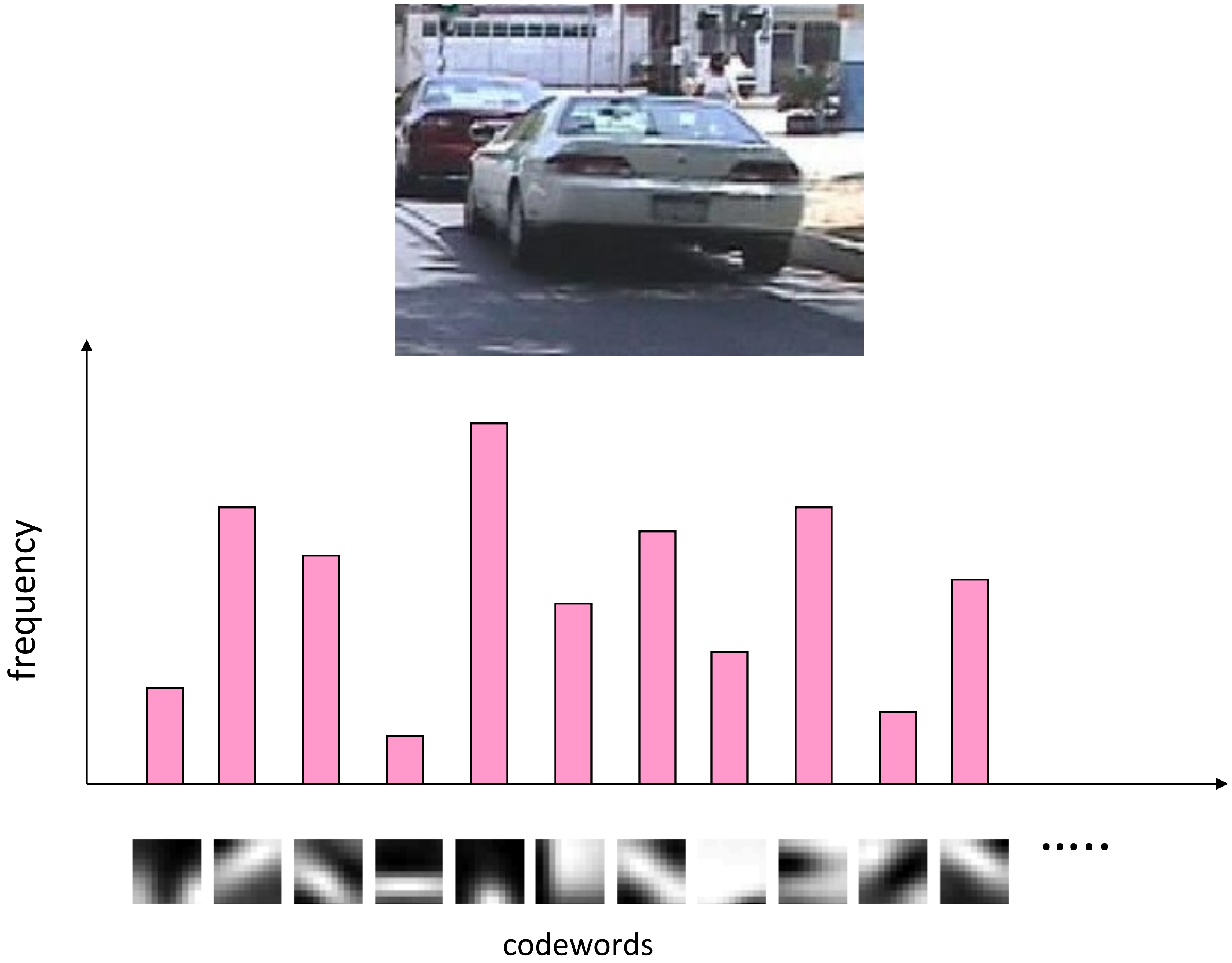


Encode:

build Bags-of-Words (BOW) vectors
for each image

2. Histogram: count the
number of visual word
occurrences

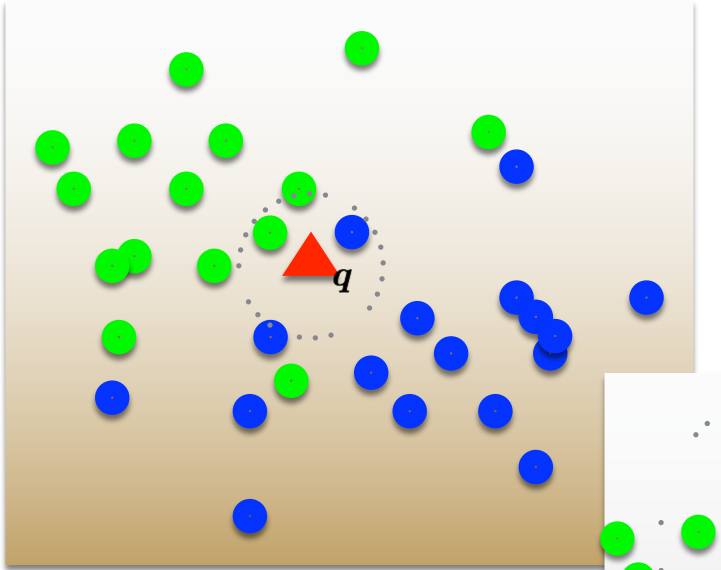



frequency
.....
codewords

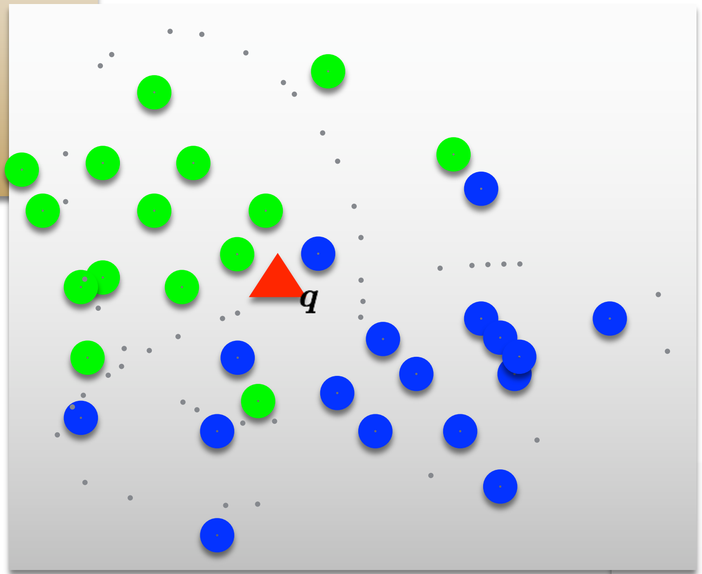
Dictionary Learning:
Learn Visual Words using clustering

Encode:
build Bags-of-Words (BOW) vectors
for each image

Classify:
Train and test data using BOWs

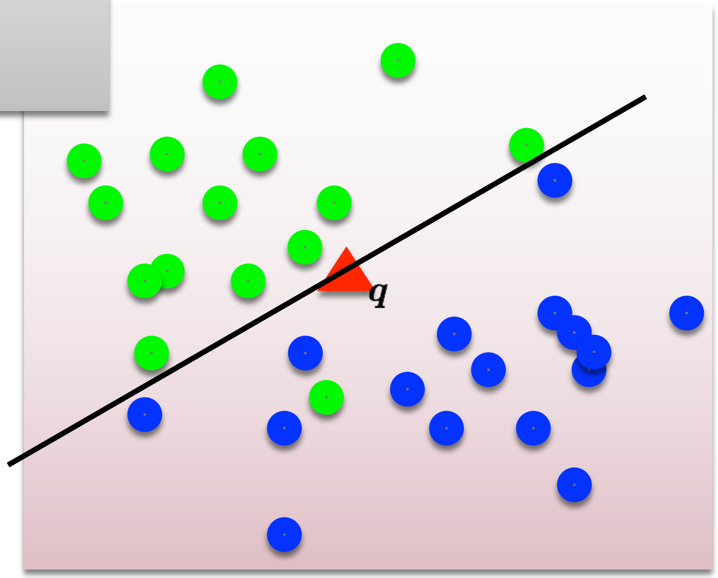


K nearest neighbors



Naïve Bayes

Support Vector Machine



The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering