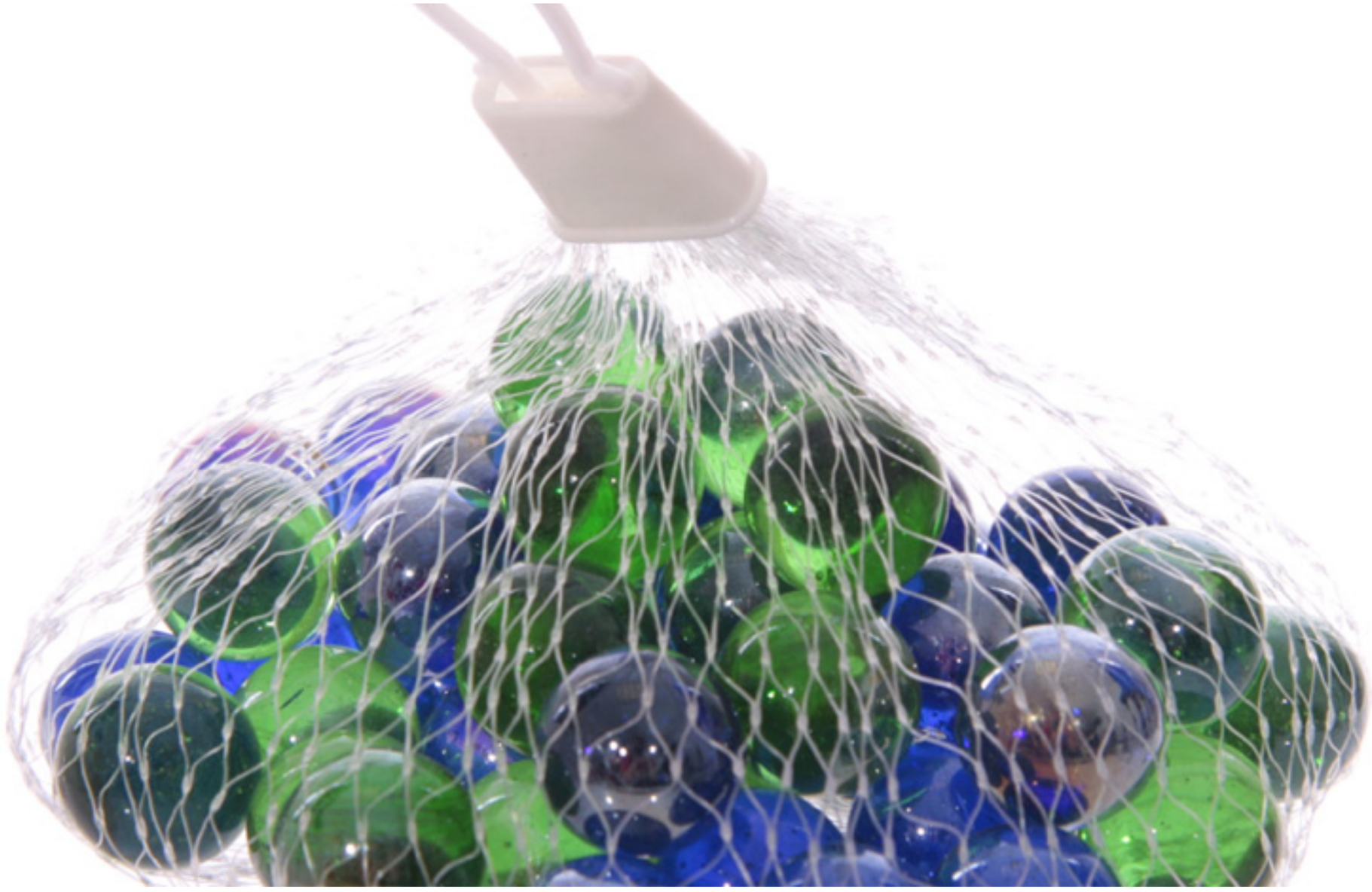
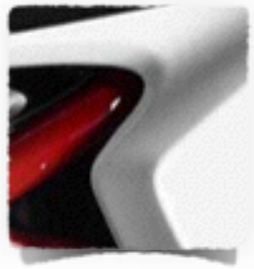


Bag-of-Visual-Words

16-385 Computer Vision (Kris Kitani)

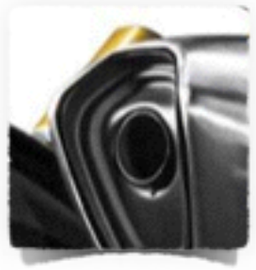
Carnegie Mellon University

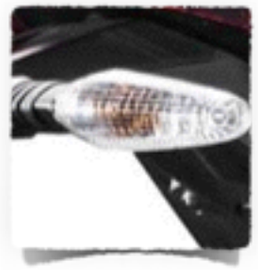
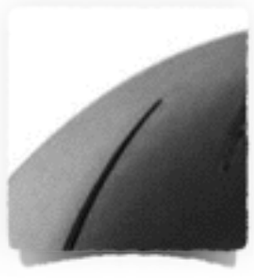
What object do these parts belong to?

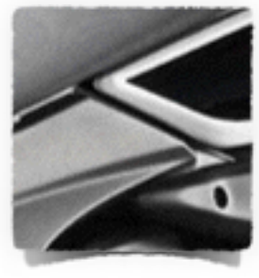










Some local feature are very informative

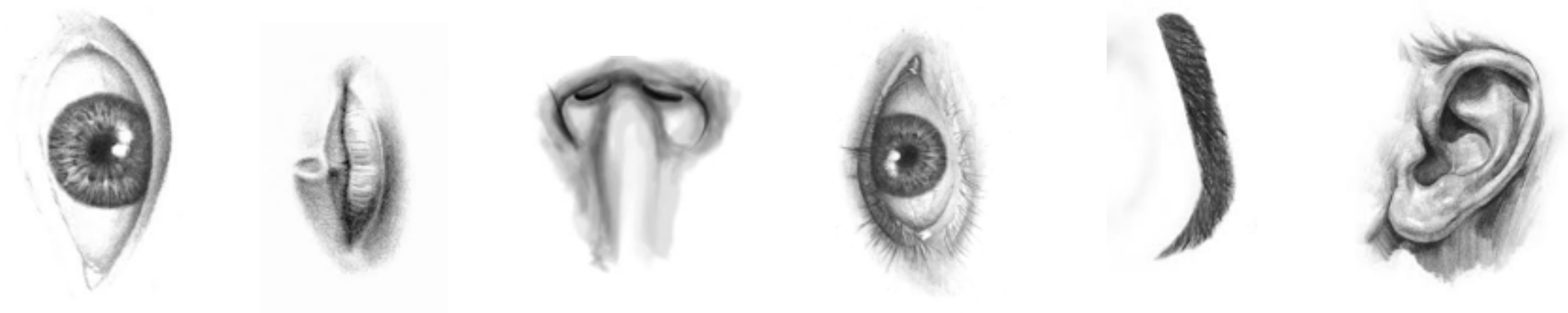
An object as

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)

CalTech6 dataset






class	bag of features	bag of features	Parts-and-shape model
	Zhang et al. (2005)	Willamowski et al. (2004)	Fergus et al. (2003)
airplanes	**98.8**	97.1	90.2
cars (rear)	98.3	**98.6**	90.3
cars (side)	**95.0**	87.3	88.5
faces	**100**	99.3	96.4
motorbikes	**98.5**	98.0	92.5
spotted cats	**97.0**	—	90.0

Works pretty well for image-level classification

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

Bag-of-features

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

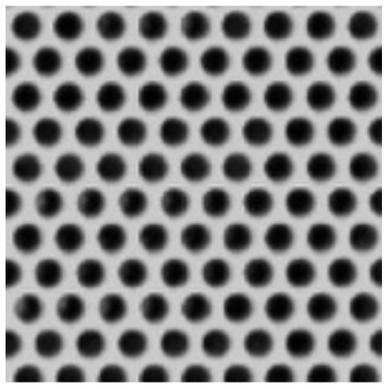
Bag-of-features

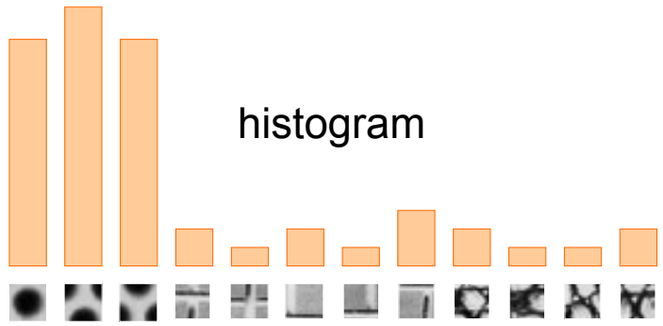
represent a data item (document, texture, image)
as a histogram over features

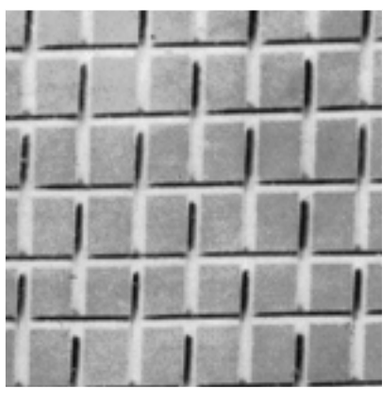
an old idea

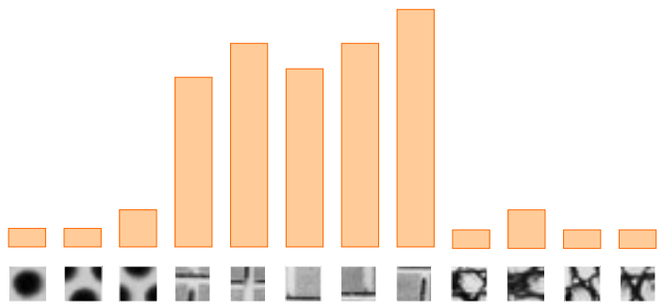
(e.g., texture recognition and information retrieval)

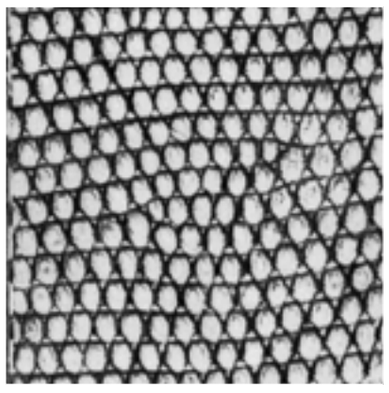
Texture recognition

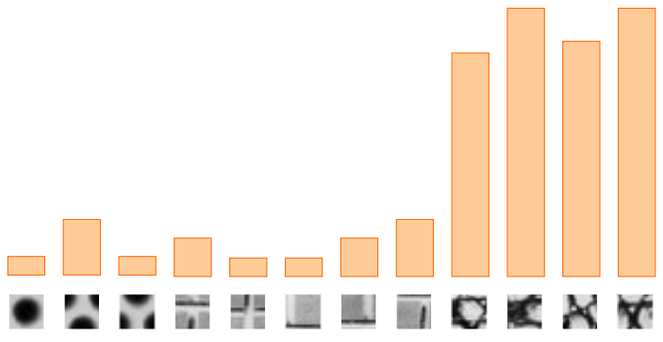












Julesz, 1981

Mori, Belongie and Malik, 2001

Vector Space Model

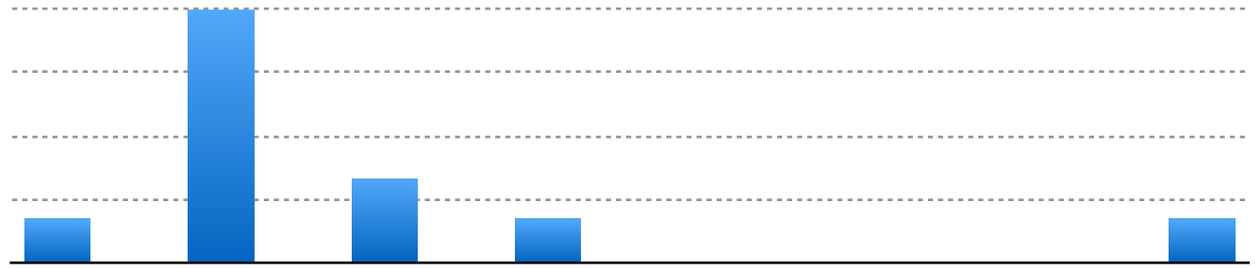
G. Salton. ‘Mathematics and Information Retrieval’ Journal of Documentation,1979


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.



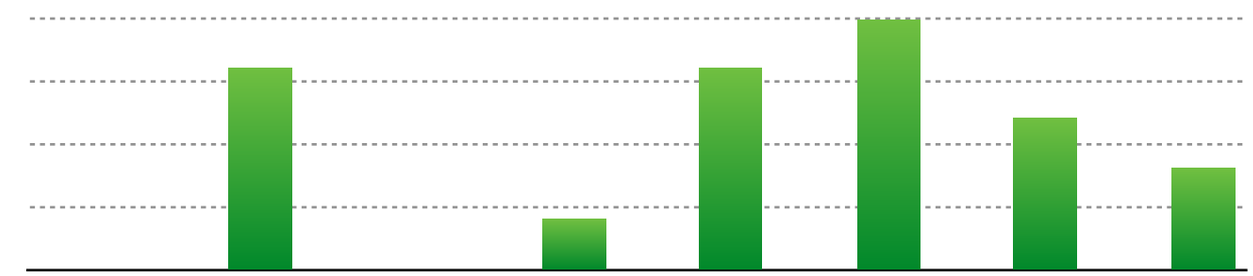
1	6	2	1	0	0	0	1
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor

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

http://www.fodey.com/generators/newspaper/snippet.asp

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)

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

just a histogram over words

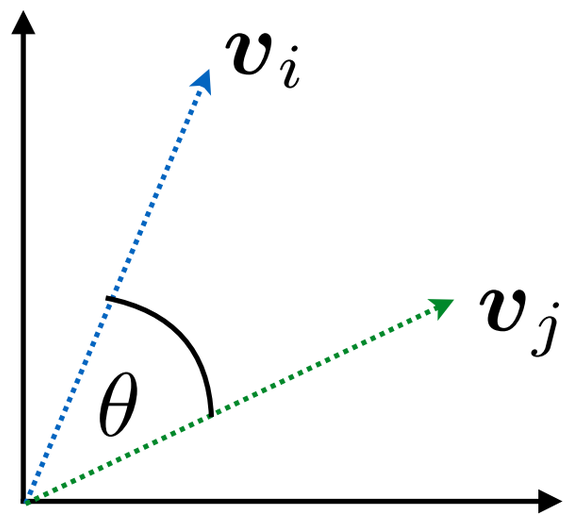
$n(\cdot)$ counts the number of occurrences

What is the similarity between two documents?

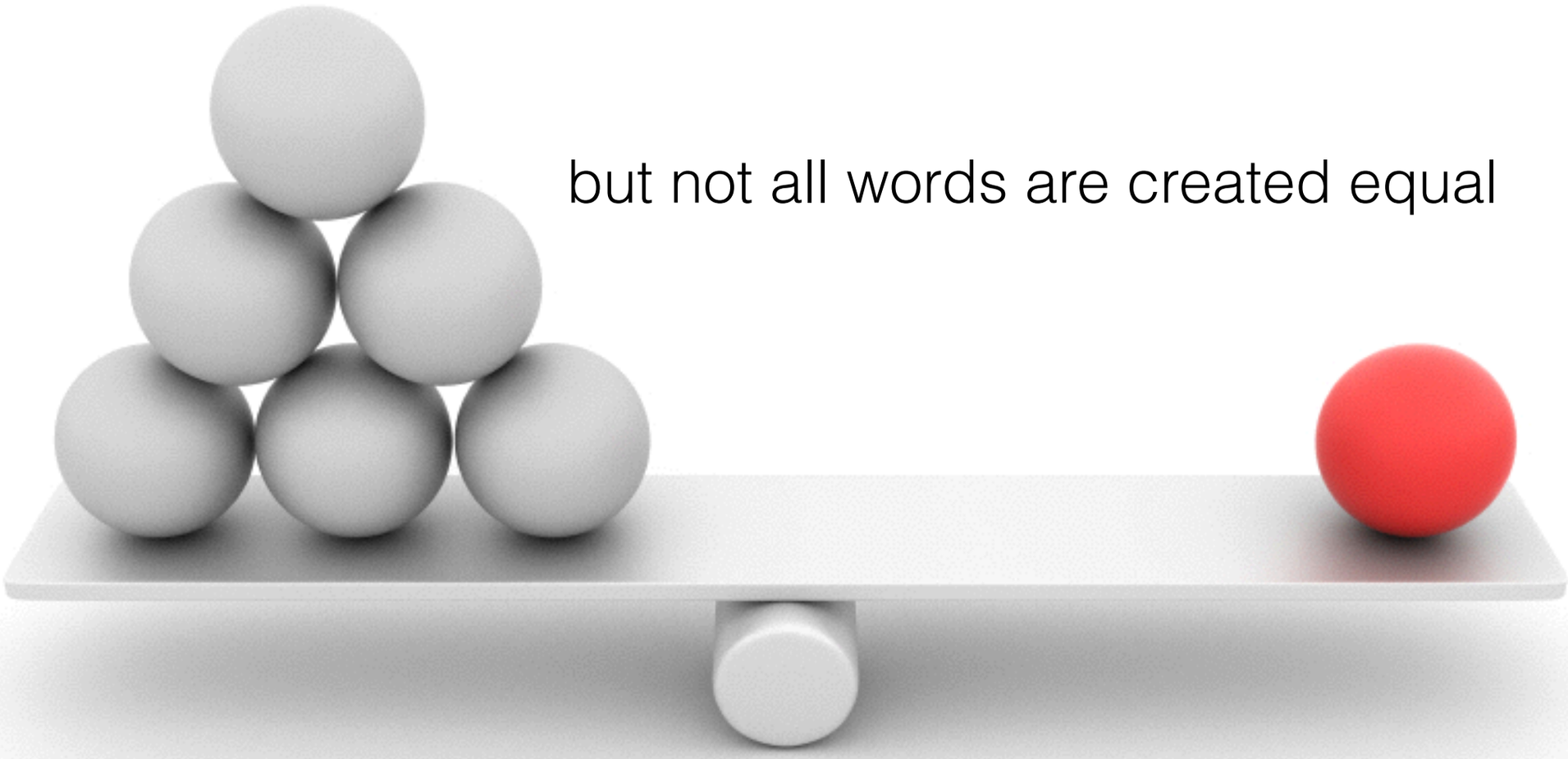
Use any distance you want but the cosine distance is fast.

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

$$= \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\| \|\boldsymbol{v}_j\|}$$



but not all words are created equal



TF-IDF

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

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

weigh each word by a heuristic

$$\boldsymbol{v}_d = [n(w_{1,d})\alpha_1 \quad n(w_{2,d})\alpha_2 \quad \cdots \quad n(w_{T,d})\alpha_T]$$

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

(down-weights **common** terms)

Standard BOW pipeline

(for image classification)

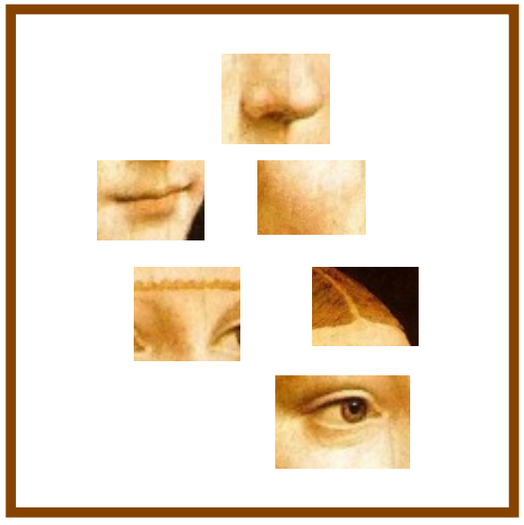
Dictionary Learning:
Learn Visual Words using clustering

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)

Dictionary Learning:

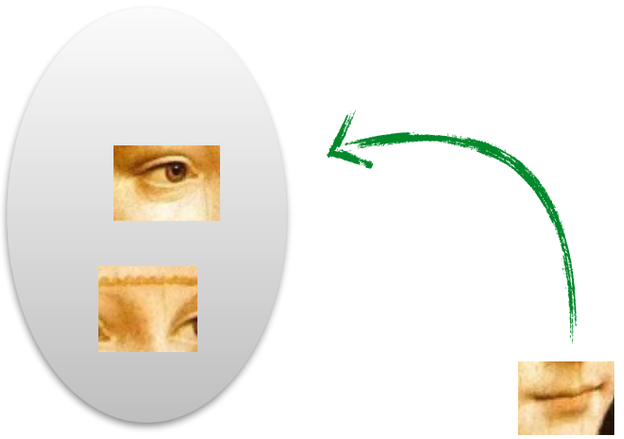
Learn Visual Words using clustering

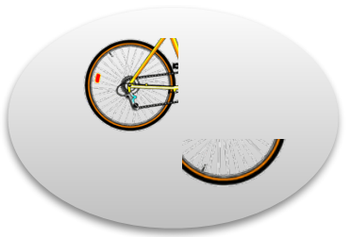
Encode:

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

Classify:

Train and test data using BOWs

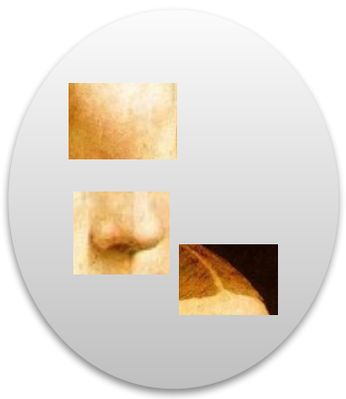




1. Quantization: image features gets associated to a visual word (nearest cluster center)

Encode:

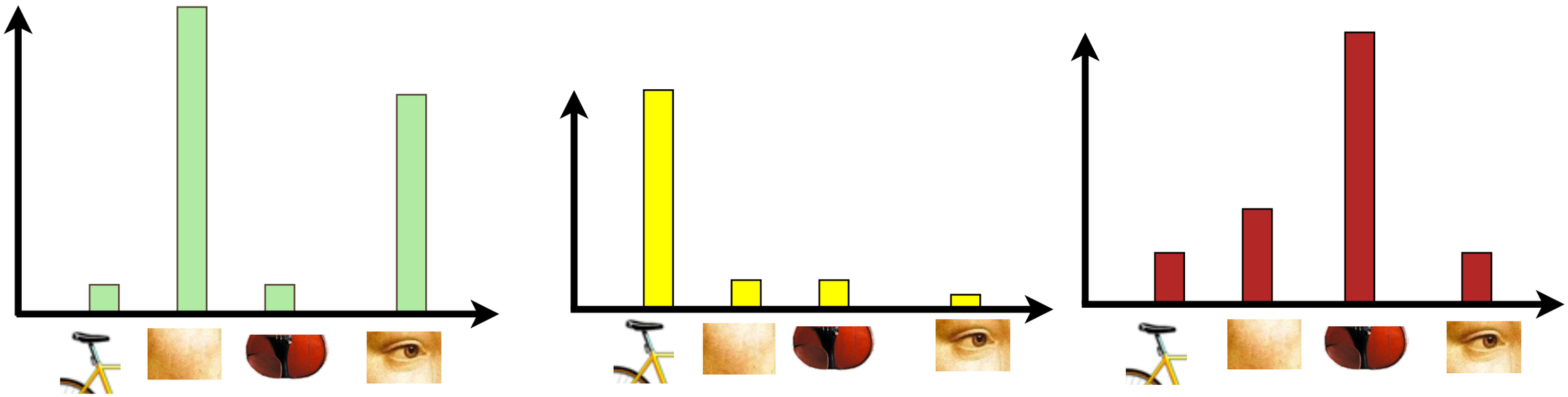
build Bags-of-Words (BOW) vectors for each image



Encode:

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

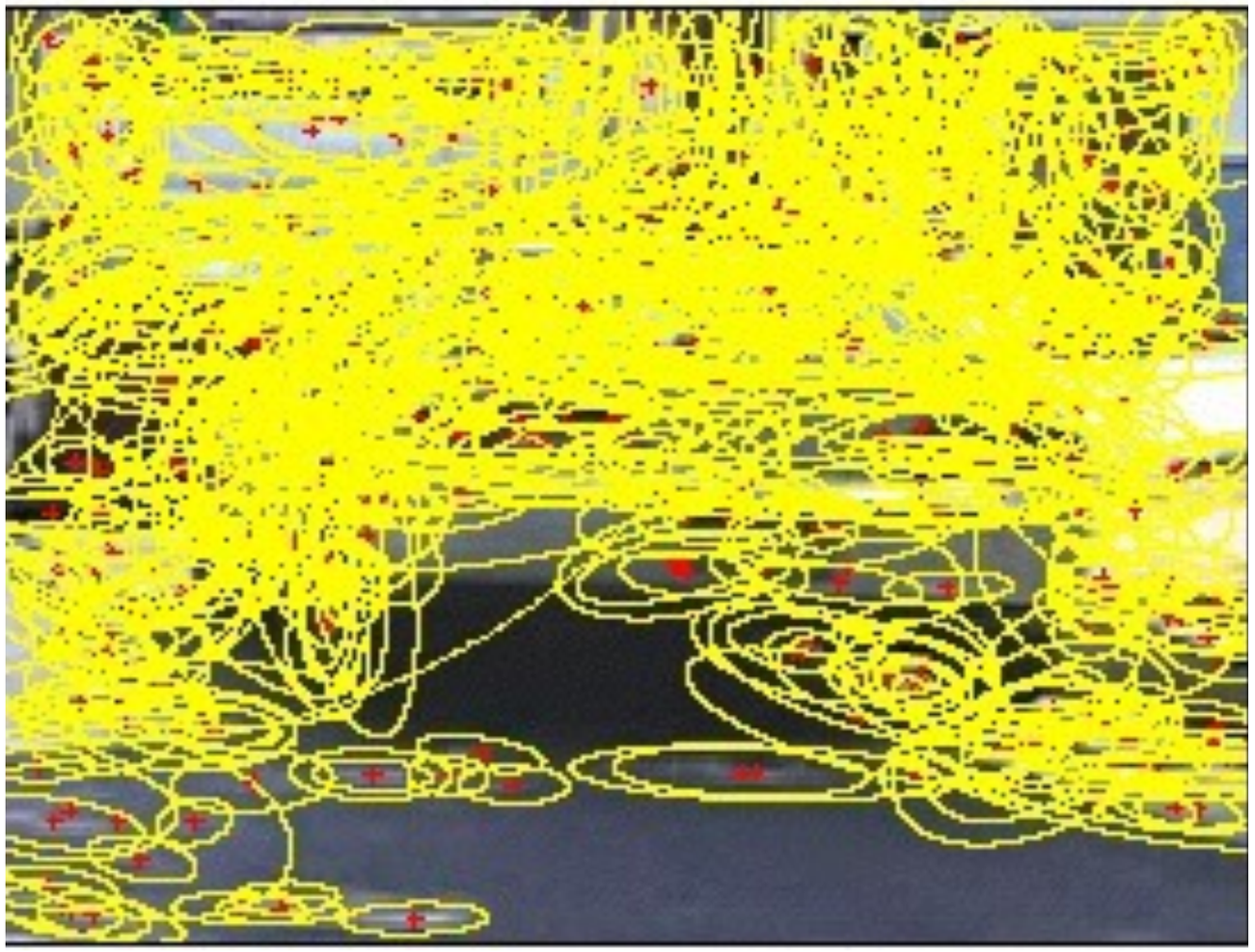
2. Histogram: count the number of visual word occurrences

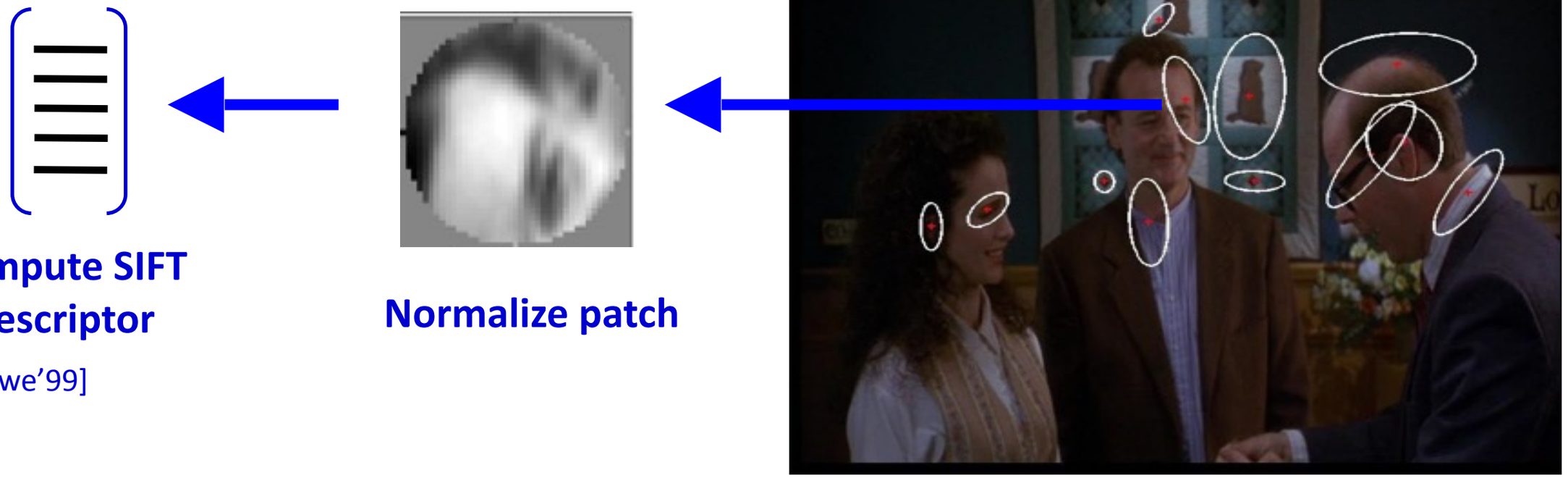


Feature Extraction

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)



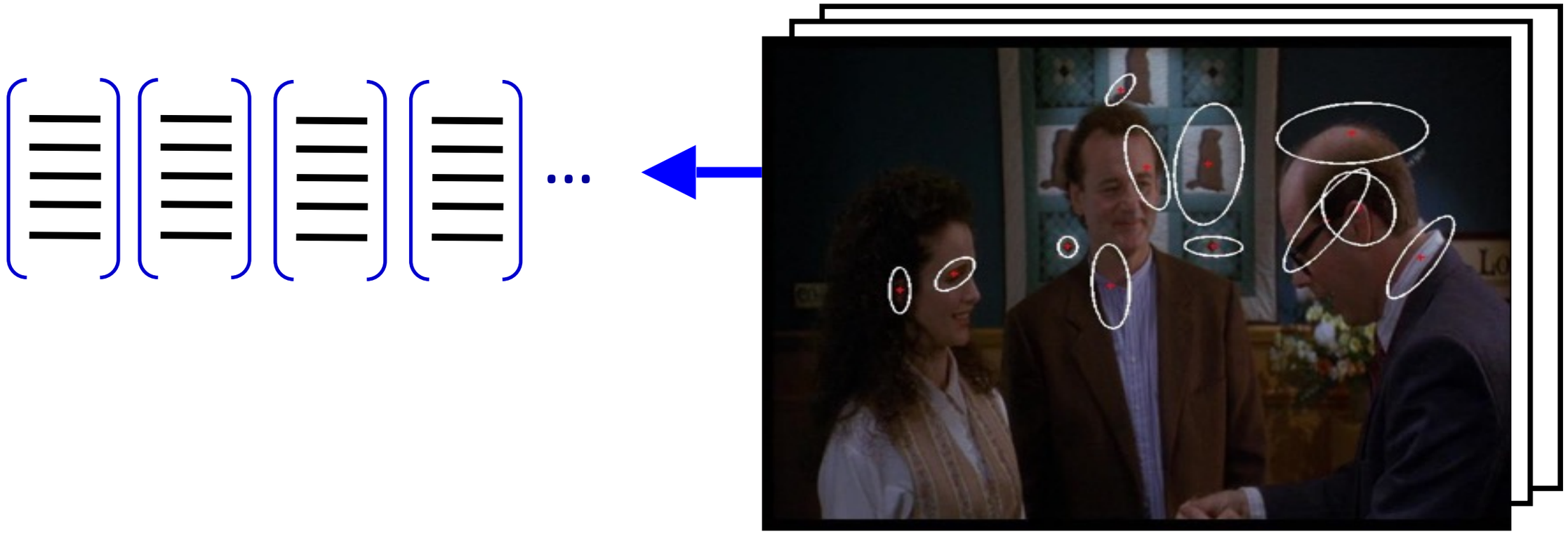


Detect patches

[Mikojaczyk and Schmid '02]

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

[Sivic & Zisserman, '03]



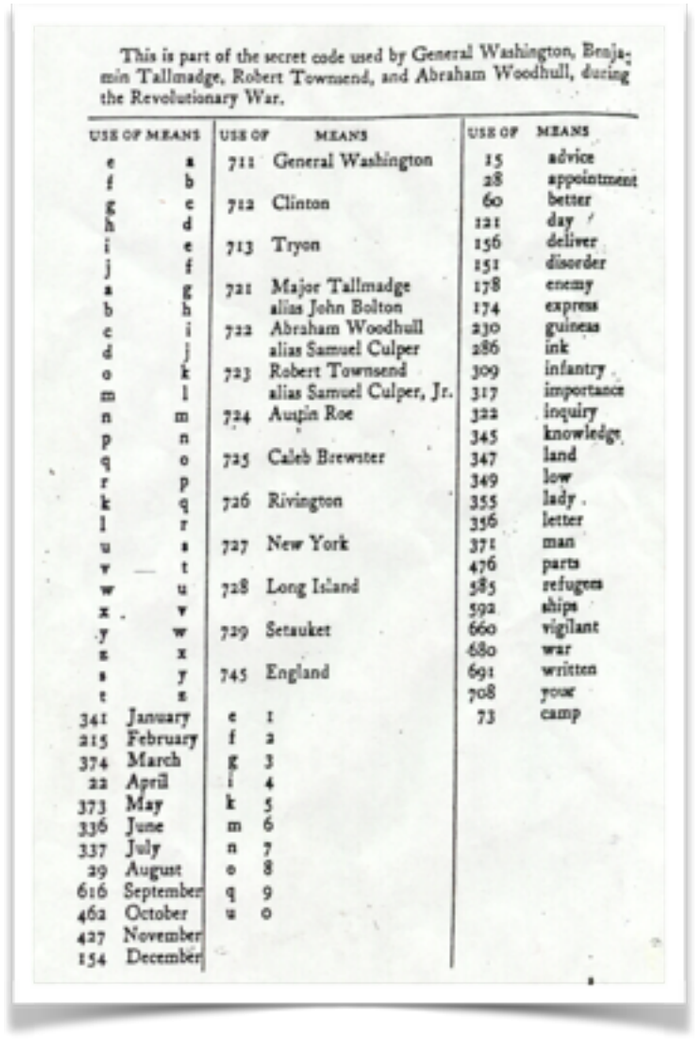
Visual Vocabulary

(coding and vector quantization)

Alternative perspective…

visual vocabulary = code book

visual word = code vector

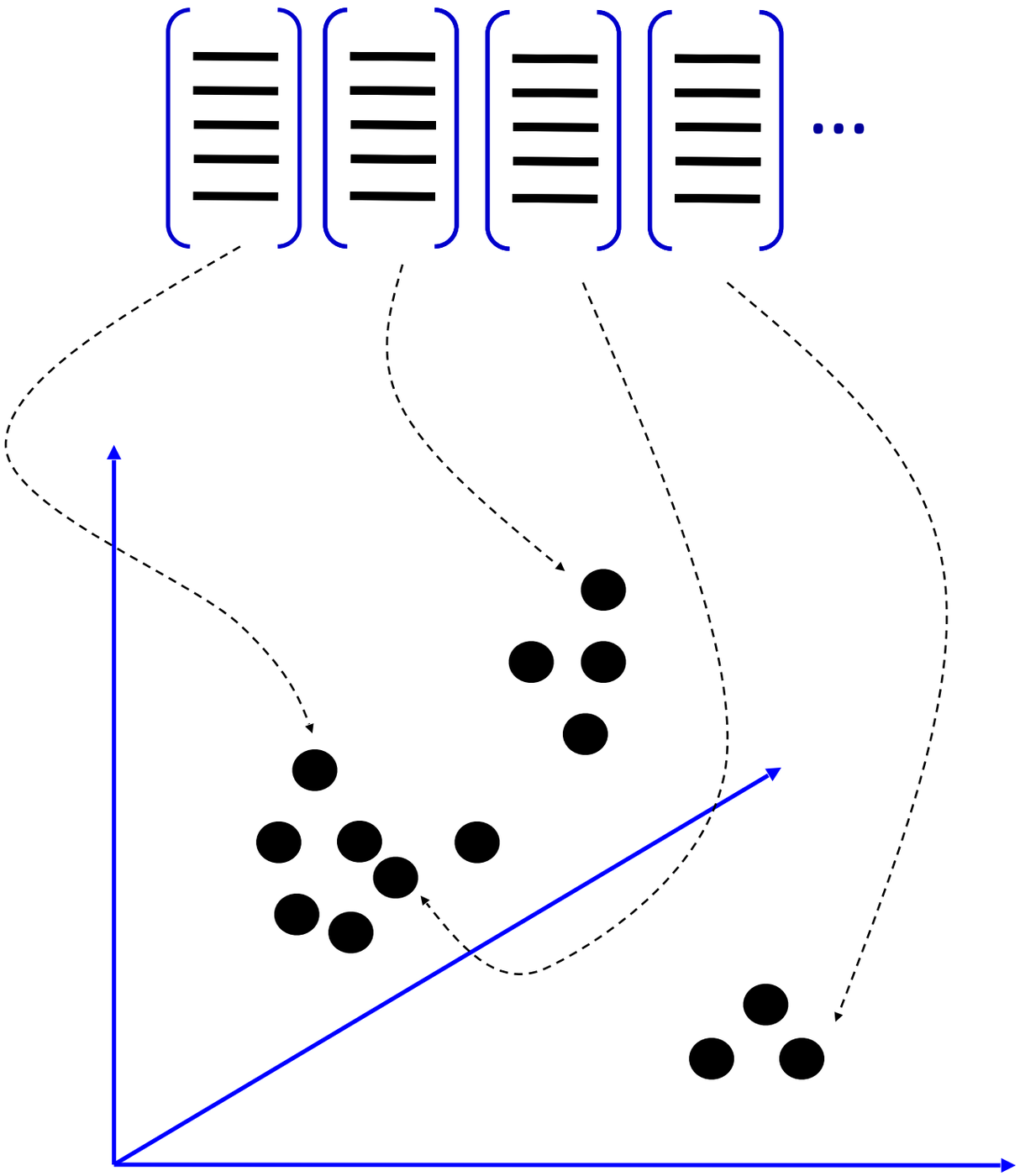


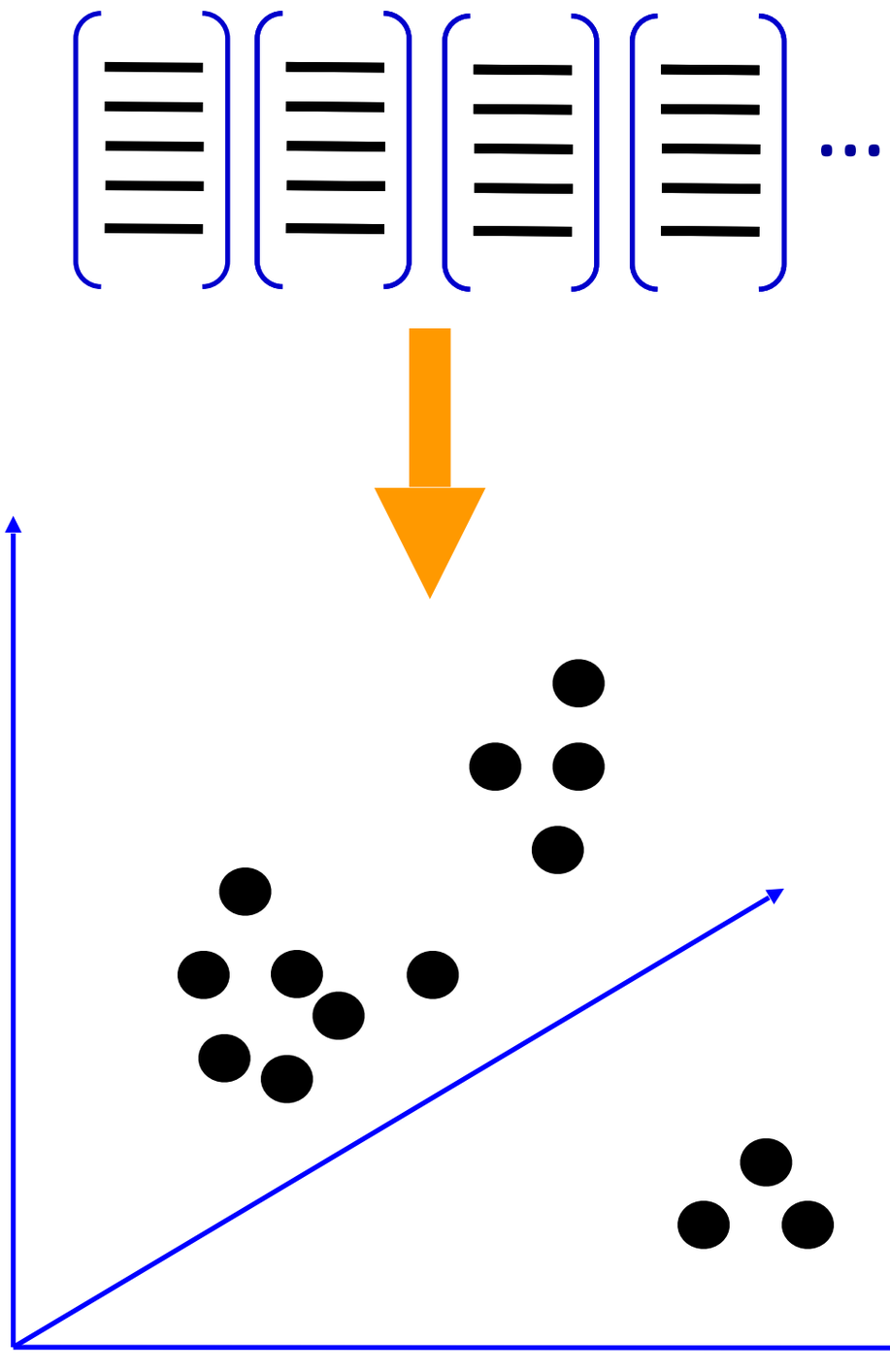
This is part of the secret code used by General Washington, Benjamin Tallmadge, Robert Townsend, and Abraham Woodhull, during the Revolutionary War.

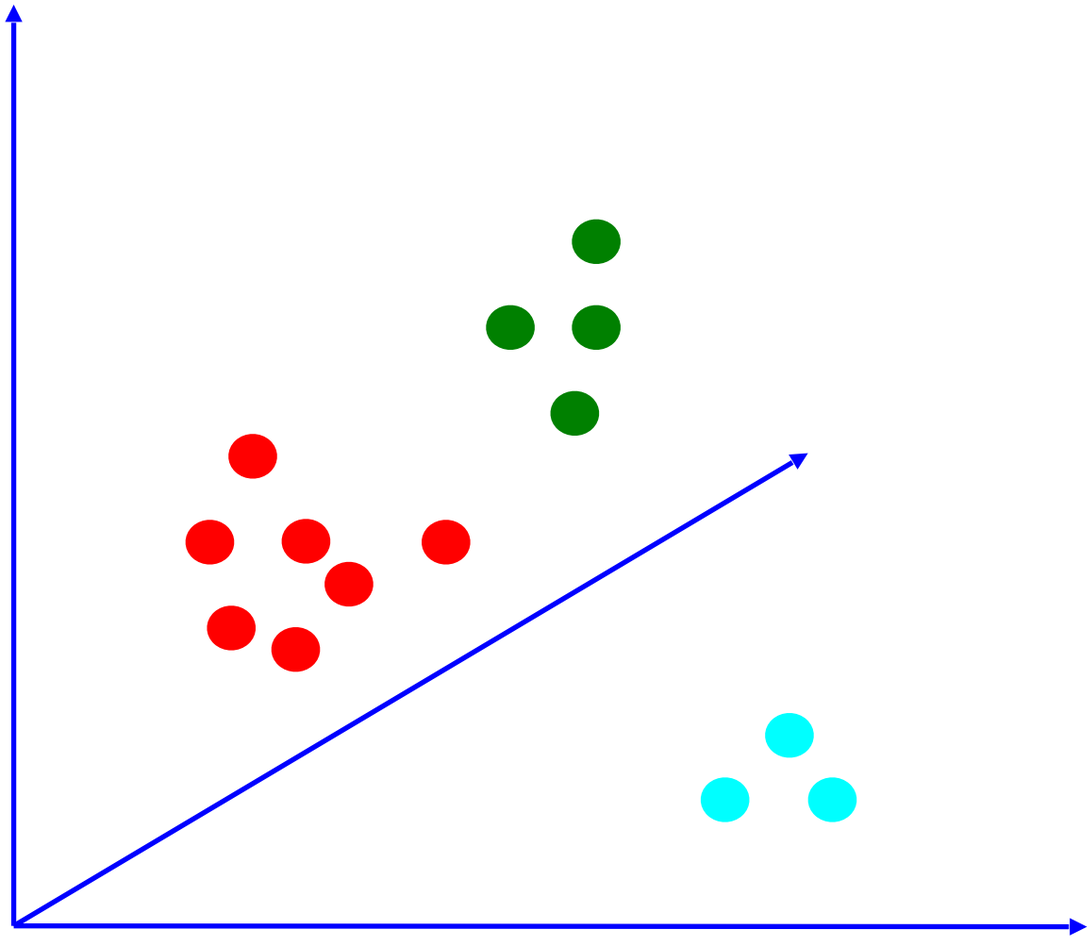
USE OF	MEANS	USE OF	MEANS	USE OF	MEANS
e	a	711	General Washington	15	advice
f	b			28	appointment
g	c	712	Clinton	60	better
h	d			121	day
i	e	713	Tryon	156	deliver
j	f			151	disorder
a	g	721	Major Tallmadge alias John Bolton	178	enemy
b	h			174	express
c	i	722	Abraham Woodhull alias Samuel Culper	230	guineas
d	j			286	ink
o	k	723	Robert Townsend alias Samuel Culper, Jr.	309	infantry
m	l			317	importance
n	m	724	Austin Roe	322	inquiry
p	n			345	knowledge
q	o	725	Caleb Brewster	347	land
r	p			349	low
k	q	726	Rivington	355	lady
l	r			356	letter
u	s	727	New York	371	man
v	t			476	parts
w	u	728	Long Island	585	refugees
x	v			592	ships
y	w	729	Setauket	660	vigilant
z	x			680	war
s	y	745	England	691	written
t	z			708	your
341	January	e	1	73	camp
215	February	f	2		
374	March	g	3		
22	April	i	4		
373	May	k	5		
336	June	m	6		
337	July	n	7		
29	August	o	8		
616	September	q	9		
462	October	u	0		
427	November				
154	December				

The **codebook** is used for quantizing features

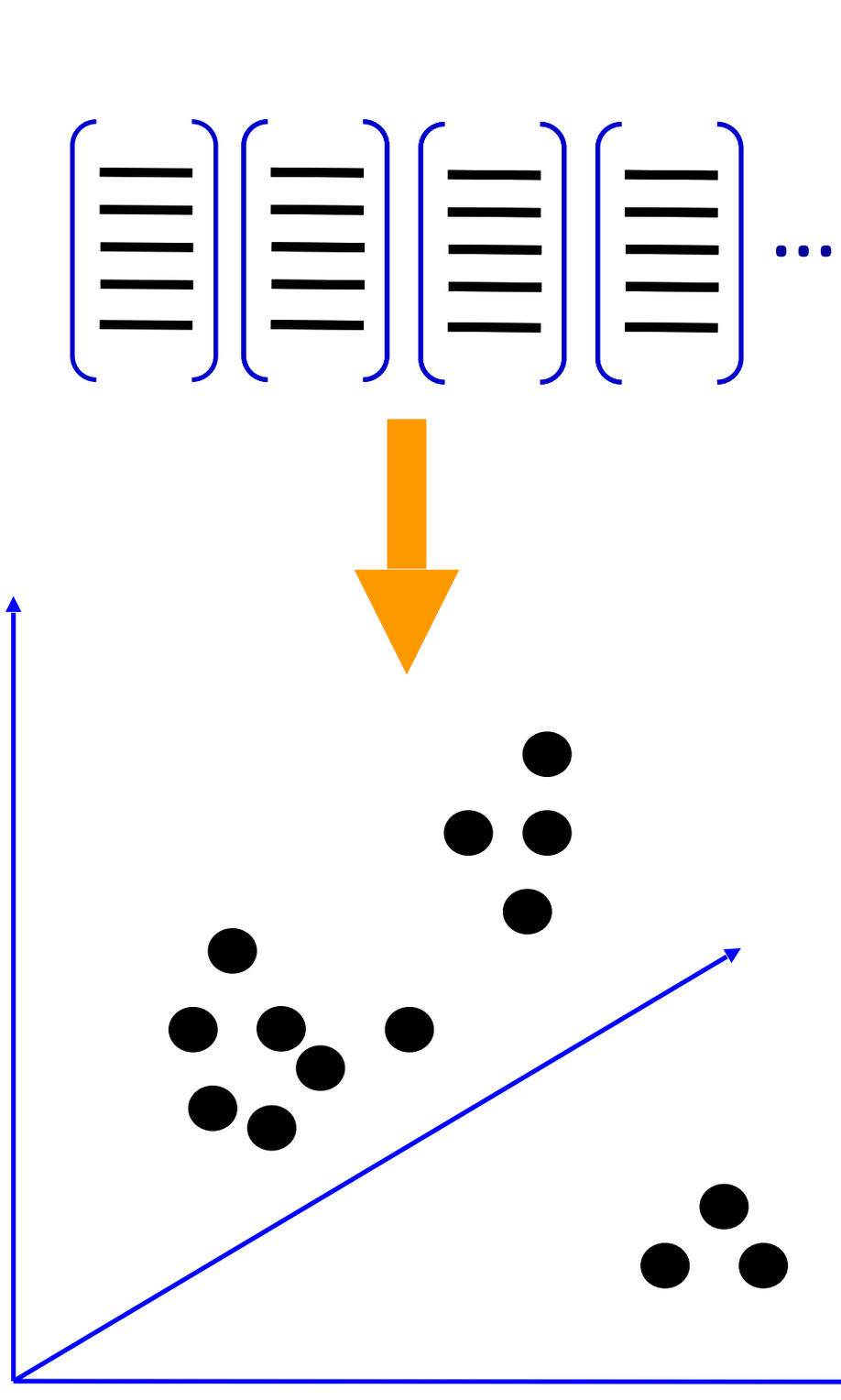
A **vector quantizer** takes a feature vector and maps it to the index of the nearest code vector in a codebook

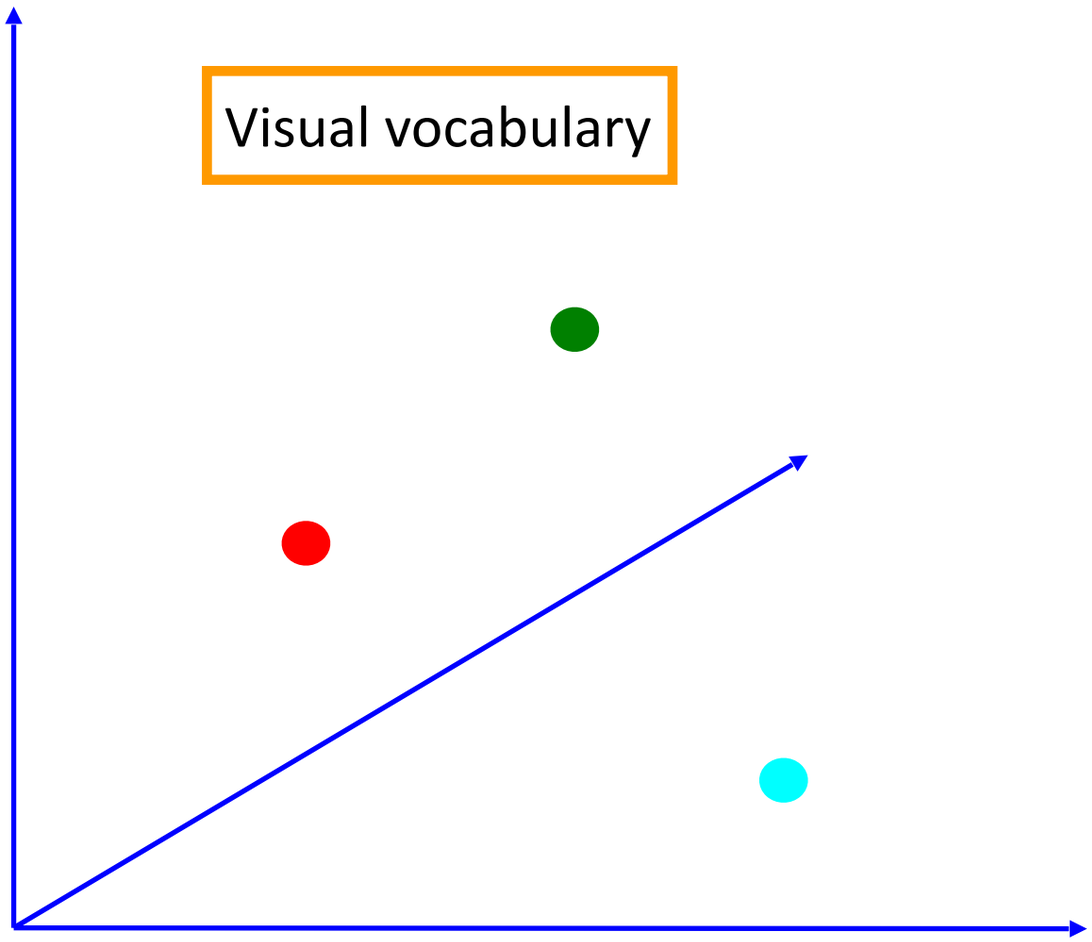






Clustering




Visual vocabulary


Clustering

K-means Clustering

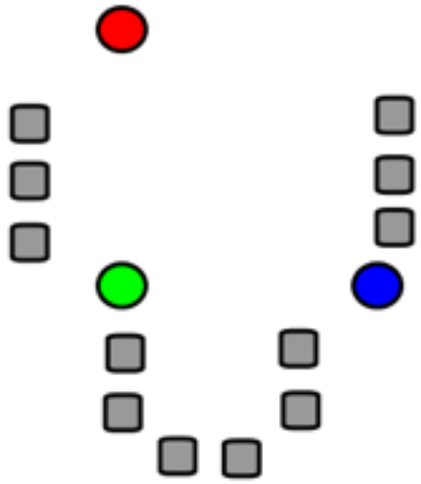
Given k:

1.Select initial centroids at random.

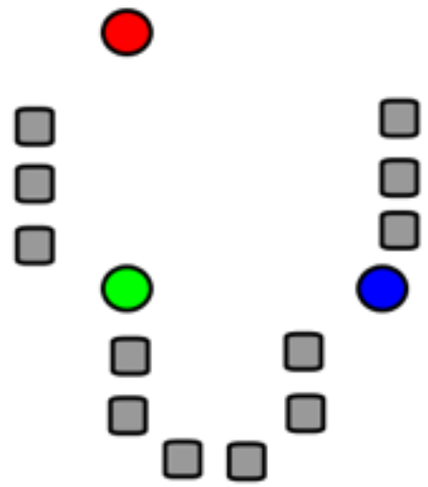
2.Assign each object to the cluster with the nearest centroid.

3.Compute each centroid as the mean of the objects assigned to it.

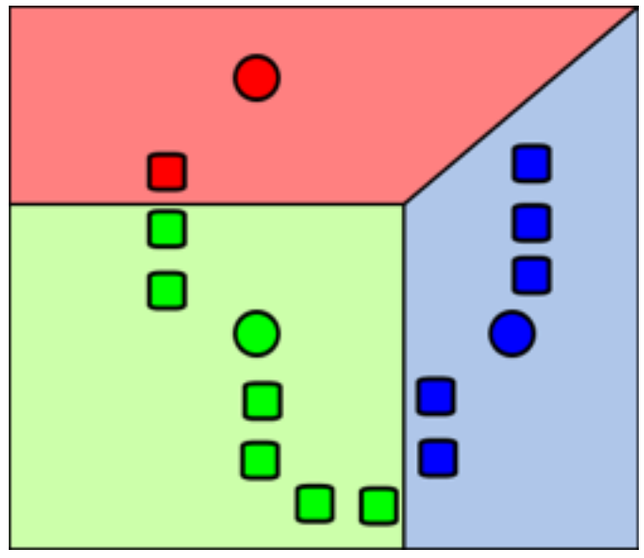
4.Repeat previous 2 steps until no change.



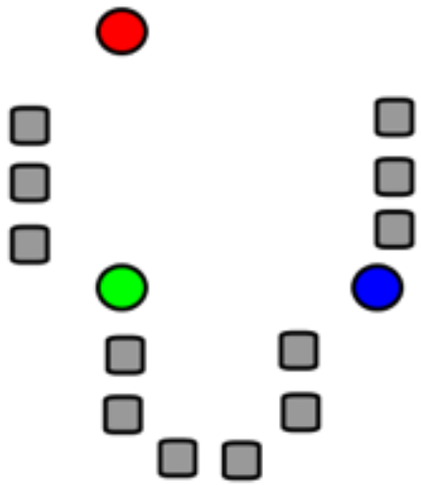
1. Select initial centroids at random



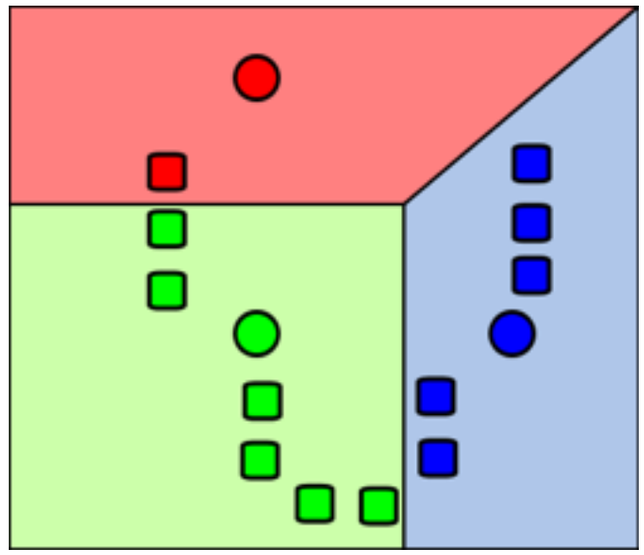
1. Select initial centroids at random



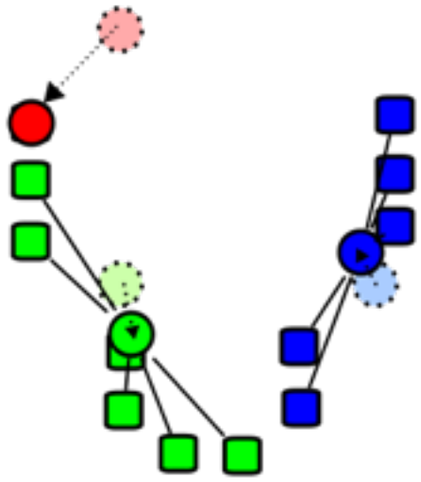
2. Assign each object to the cluster with the nearest centroid.



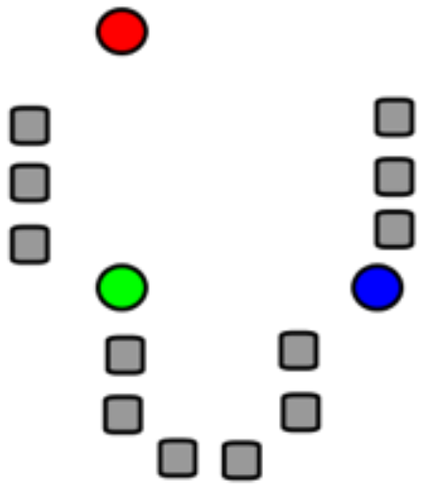
1. Select initial centroids at random



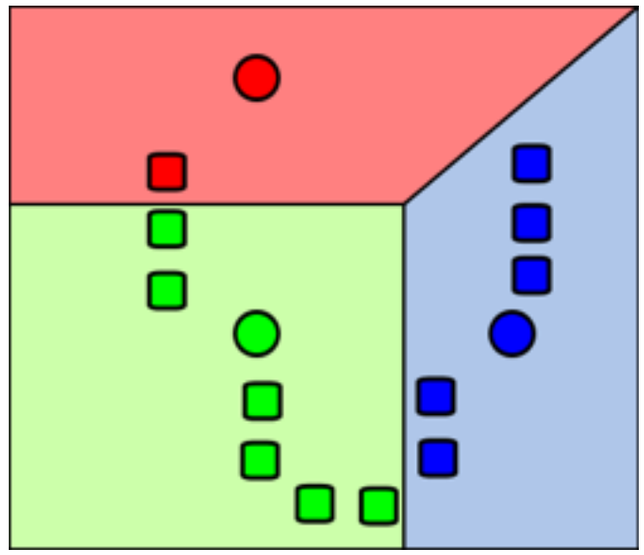
2. Assign each object to the cluster with the nearest centroid.



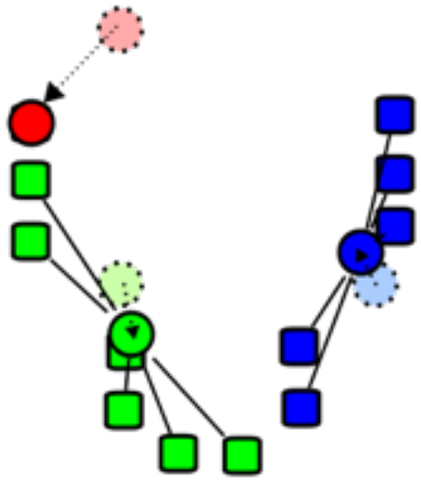
3. Compute each centroid as the mean of the objects assigned to it (go to 2)



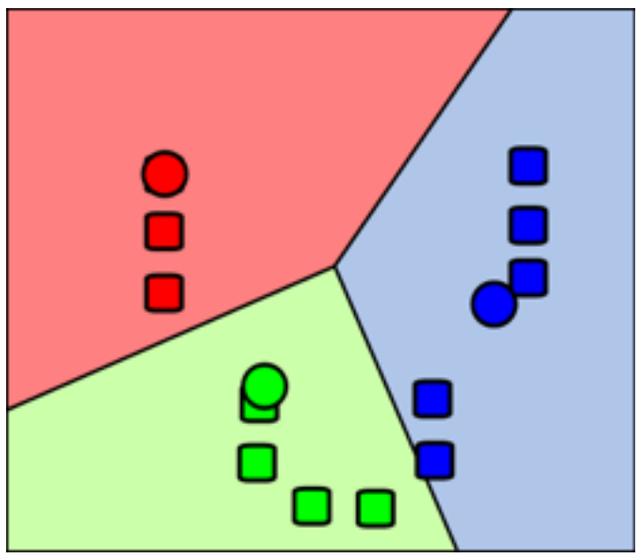
1. Select initial centroids at random



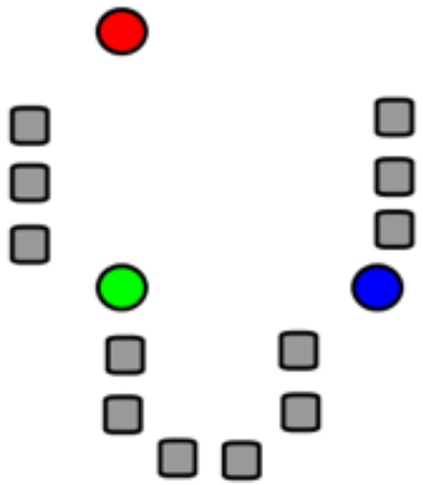
2. Assign each object to the cluster with the nearest centroid.



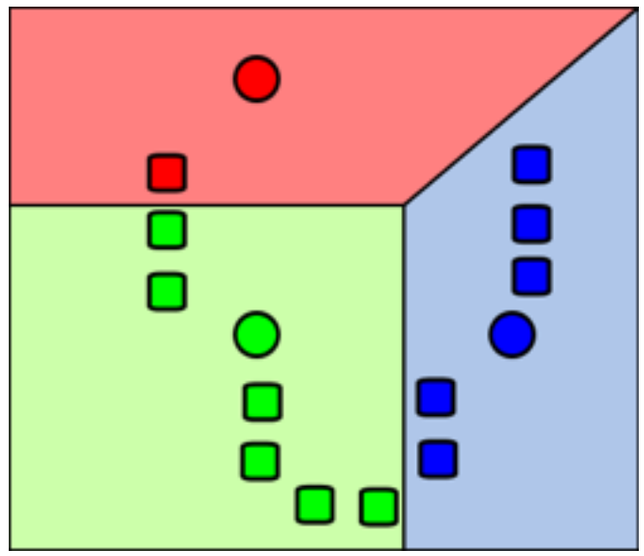
3. Compute each centroid as the mean of the objects assigned to it (go to 2)



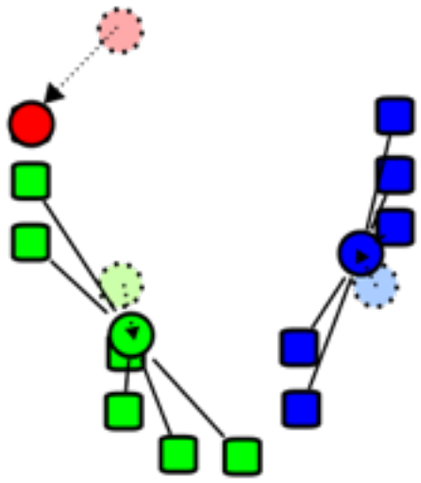
2. Assign each object to the cluster with the nearest centroid.



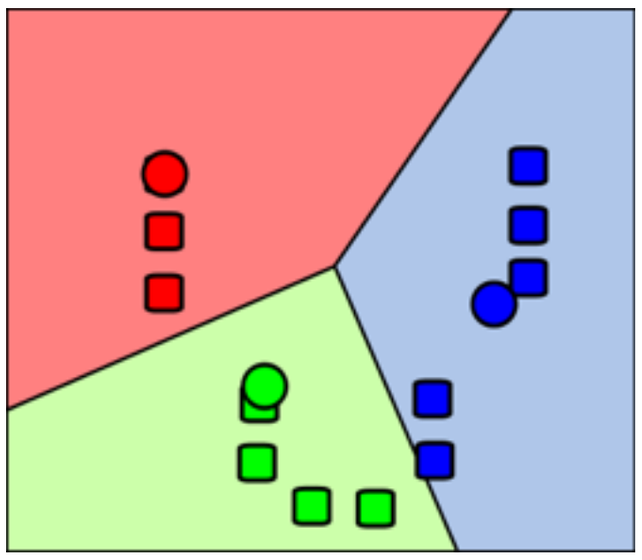
1. Select initial centroids at random



2. Assign each object to the cluster with the nearest centroid.



3. Compute each centroid as the mean of the objects assigned to it (go to 2)



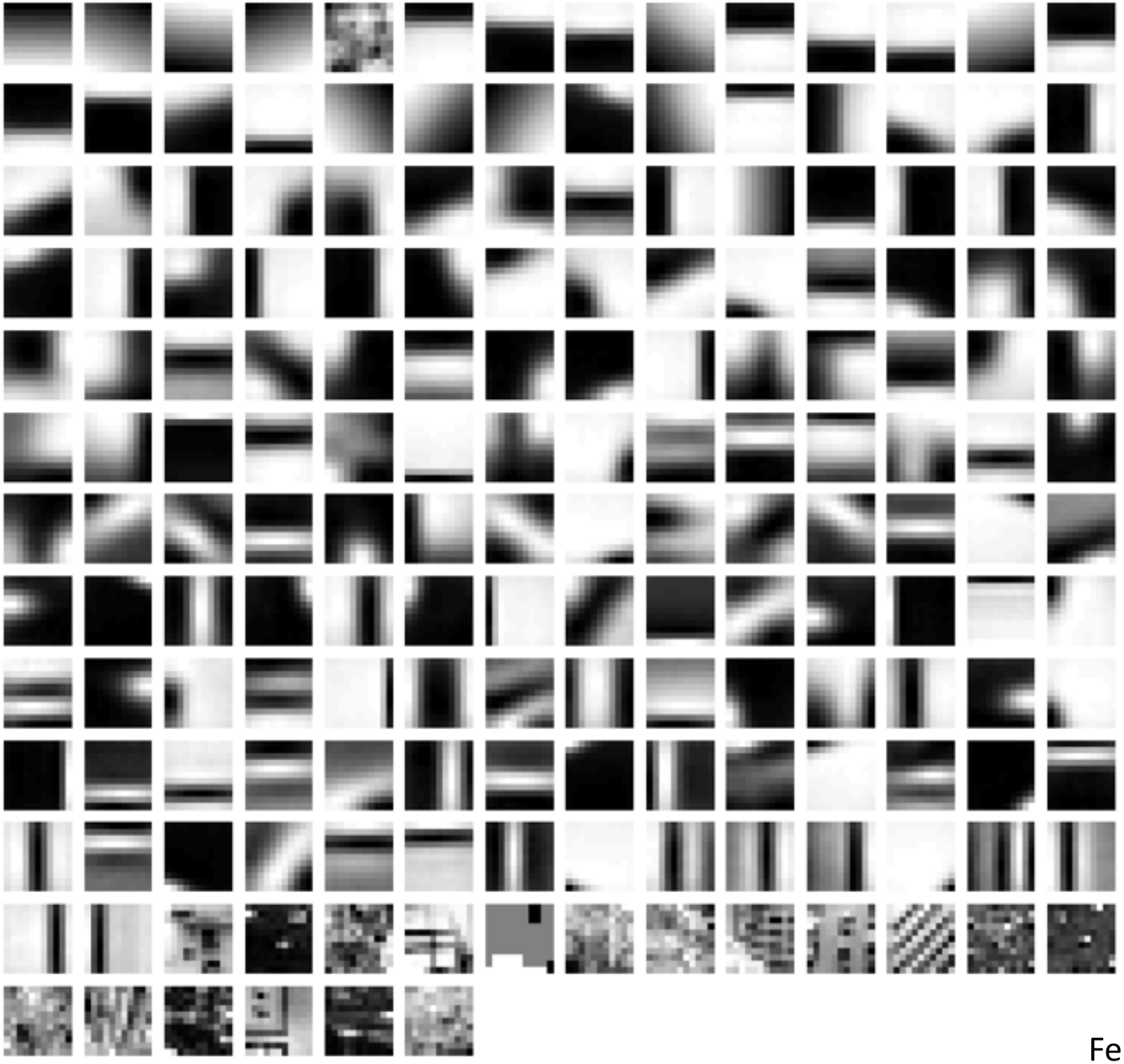
2. Assign each object to the cluster with the nearest centroid.

Repeat previous 2 steps until no change

*From what **data** should I learn the code book?*

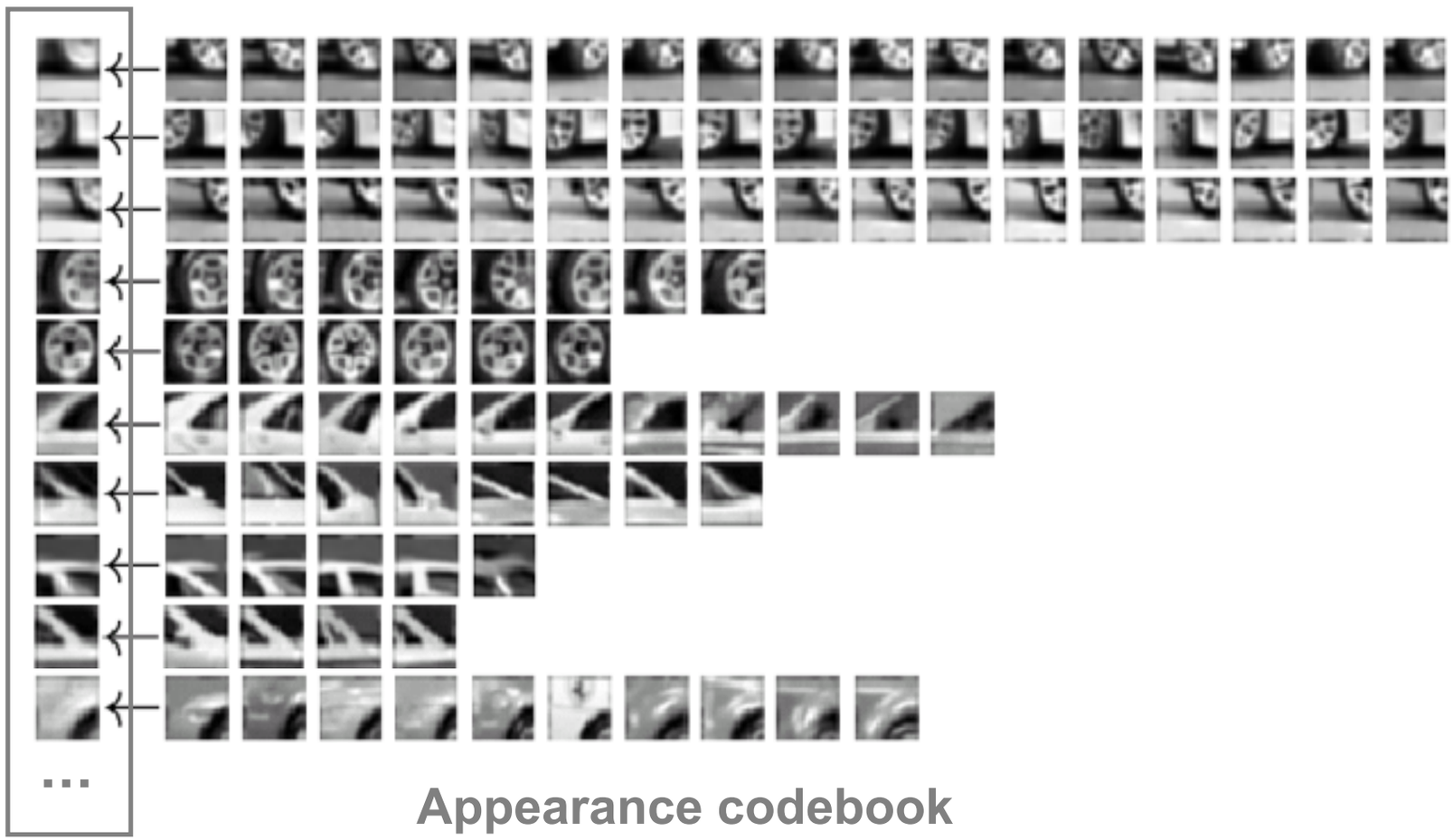
- Codebook can be learned on separate training set
- Provided the training set is sufficiently representative, the codebook will be “universal”

Example visual vocabulary



Fei-Fei et al. 2005

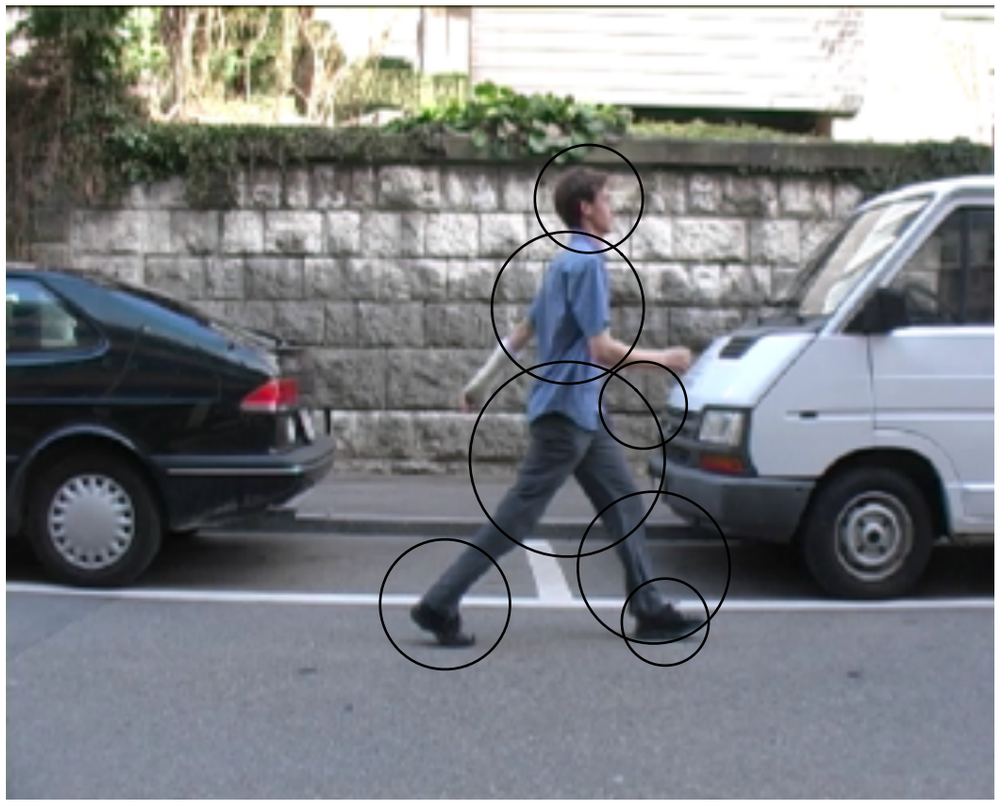
Example codebook

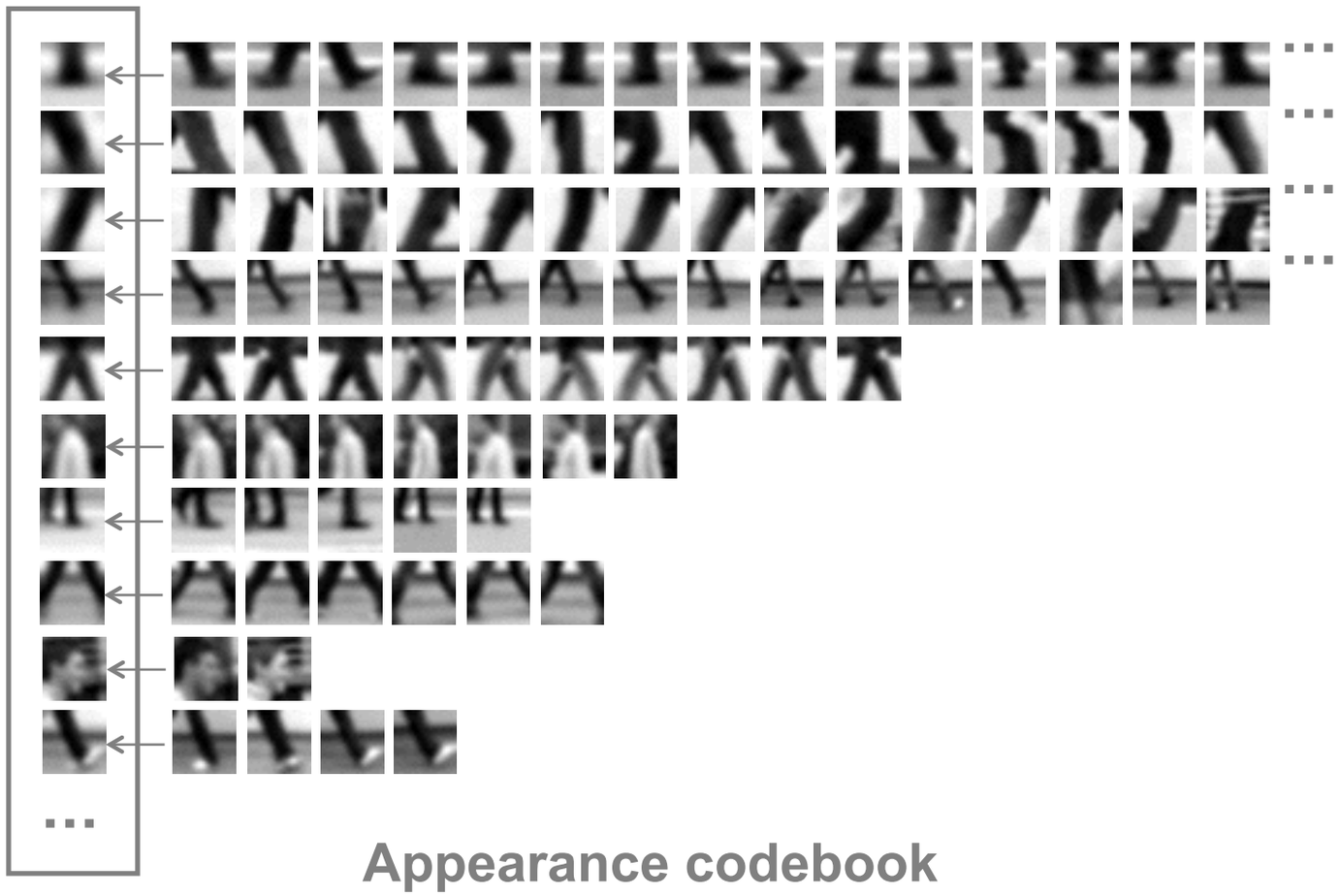


Appearance codebook

Source: B. Leibe

Another codebook





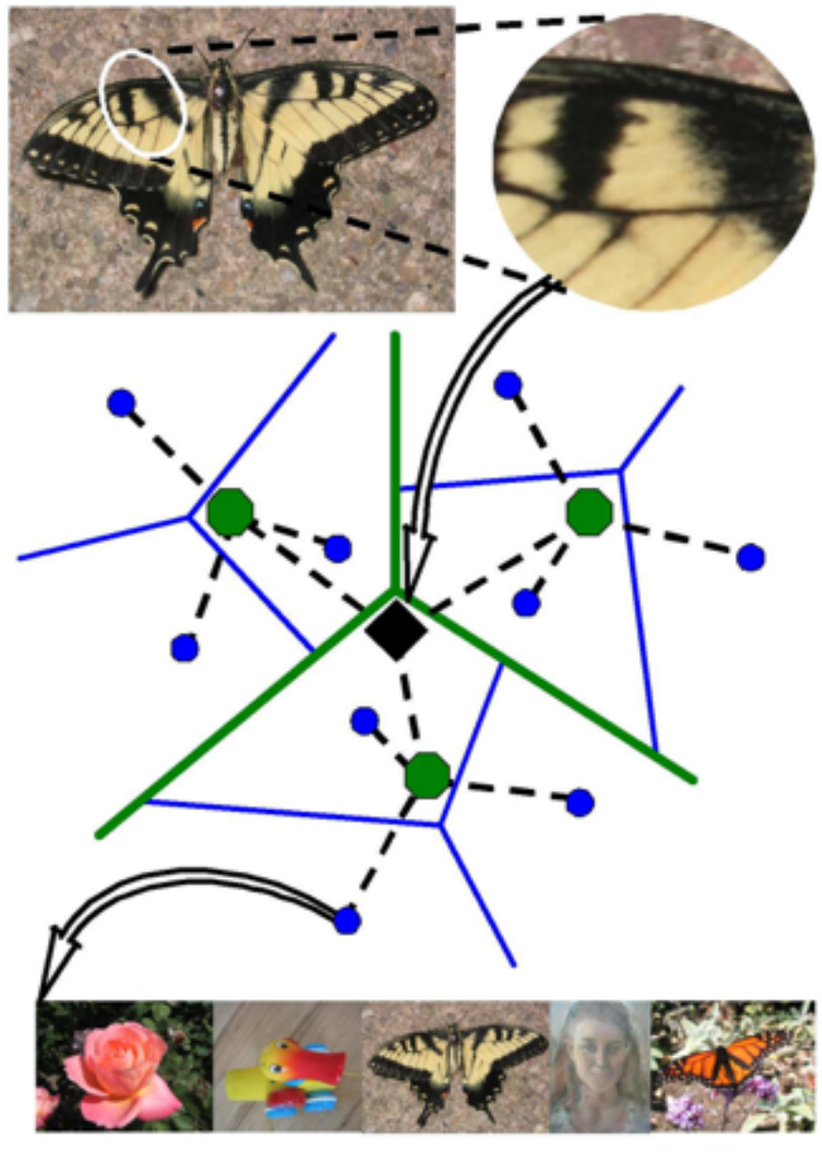
Appearance codebook

Source: B. Leibe

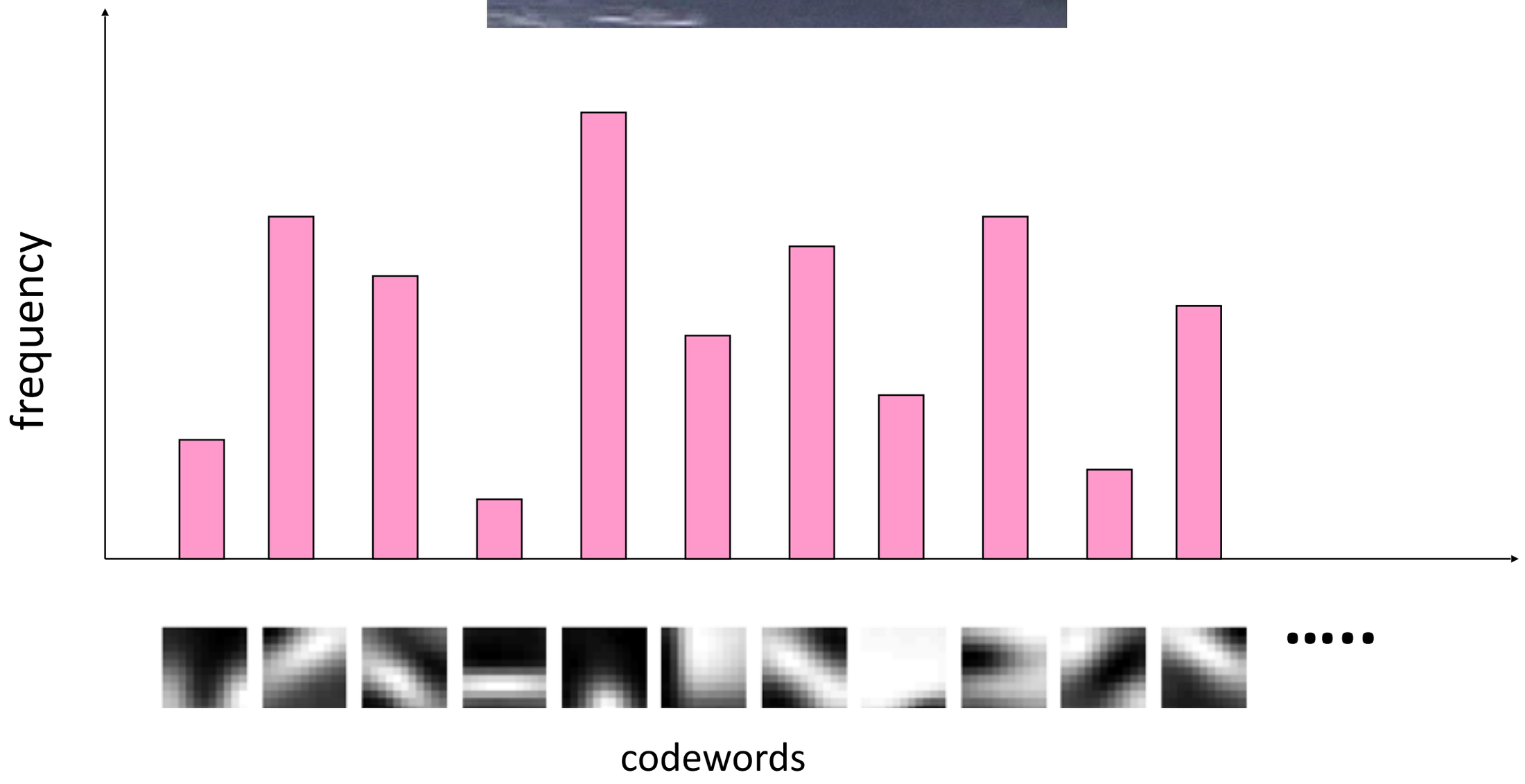
Visual vocabularies: Issues

- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large: quantization artifacts, overfitting

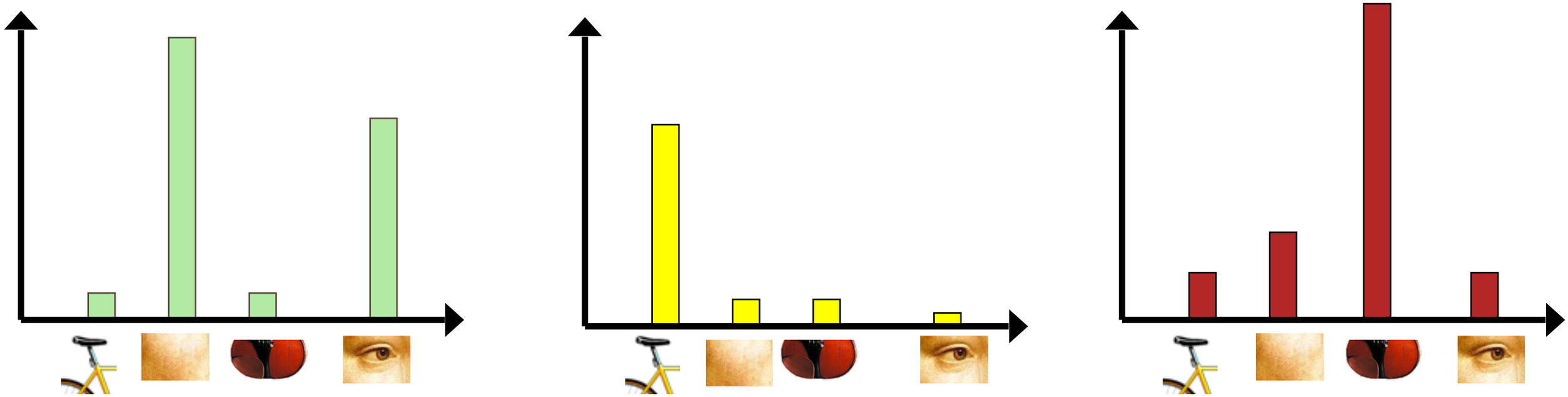
- Computational efficiency
  - Vocabulary trees
    (Nister & Stewenius, 2006)



Histogram


frequency
codewords

Classification

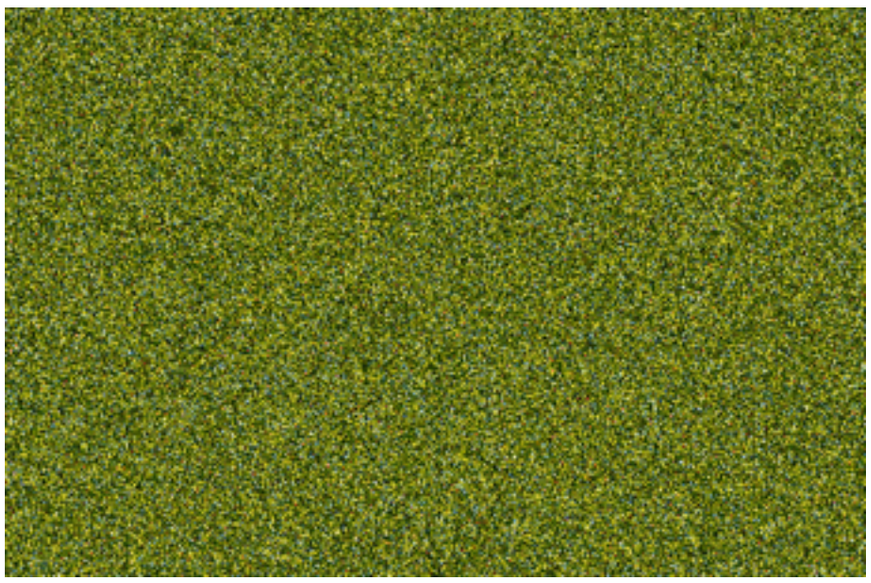


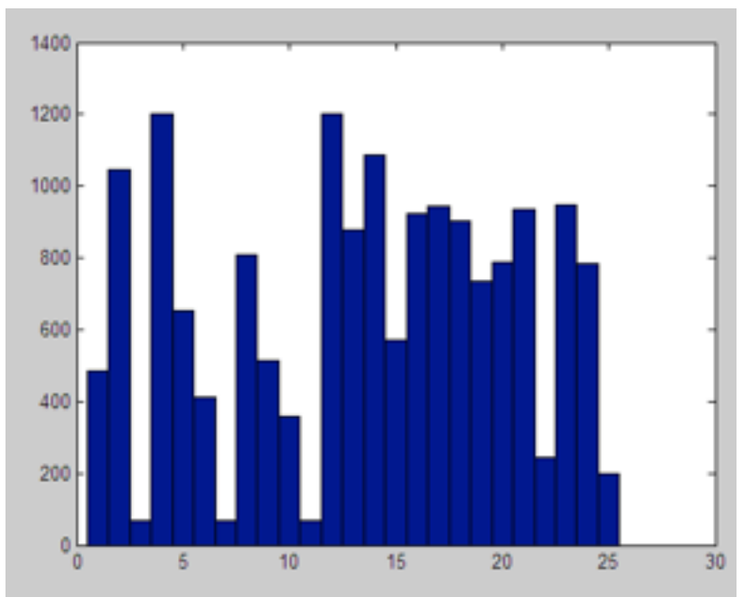
Given the bag-of-features representations of images from different classes, learn a classifier using machine learning

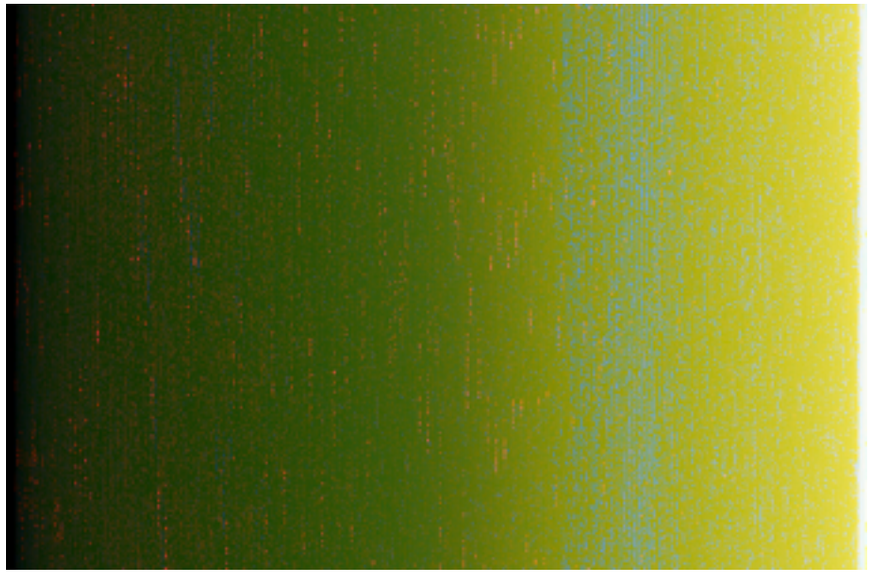
(more on this soon)

Extension to bag-of-words models

All of these images have the same color histogram!

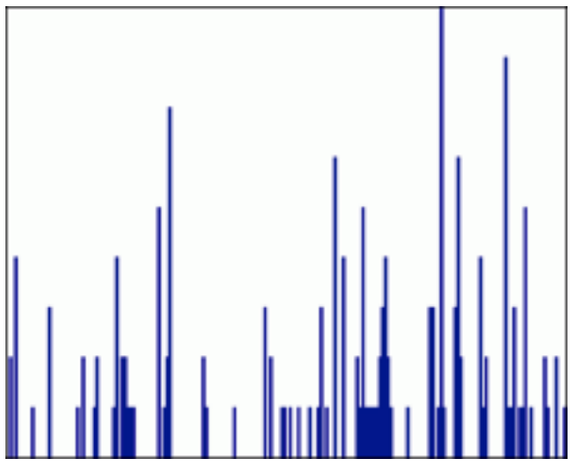






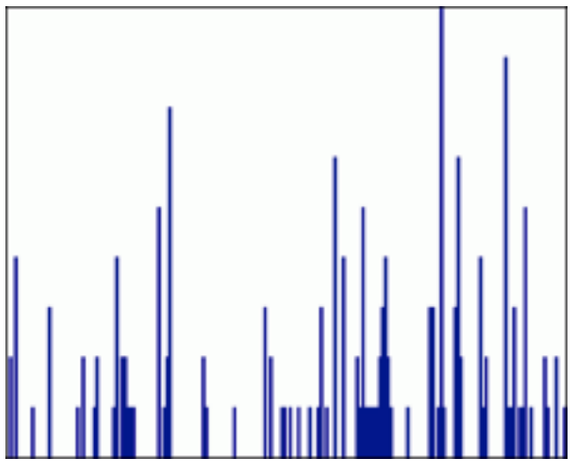
How can we encode the spatial layout?

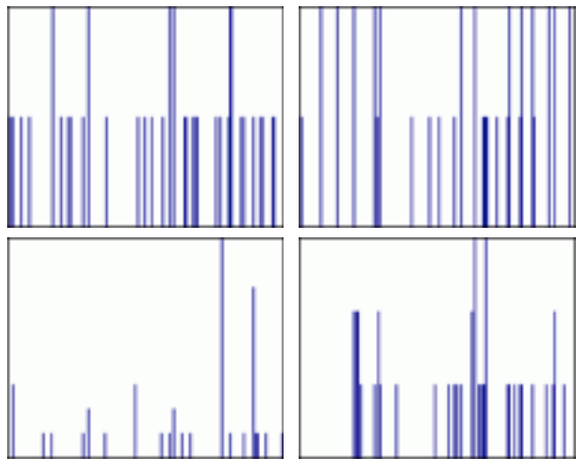
Spatial Pyramid representation



Lazebnik, Schmid & Ponce (CVPR 2006)

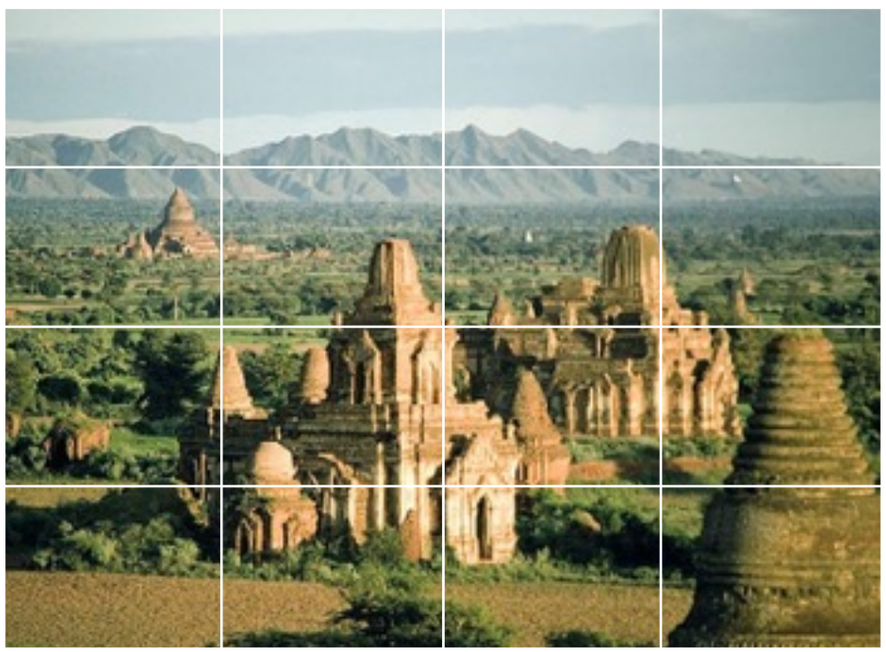
Spatial Pyramid representation

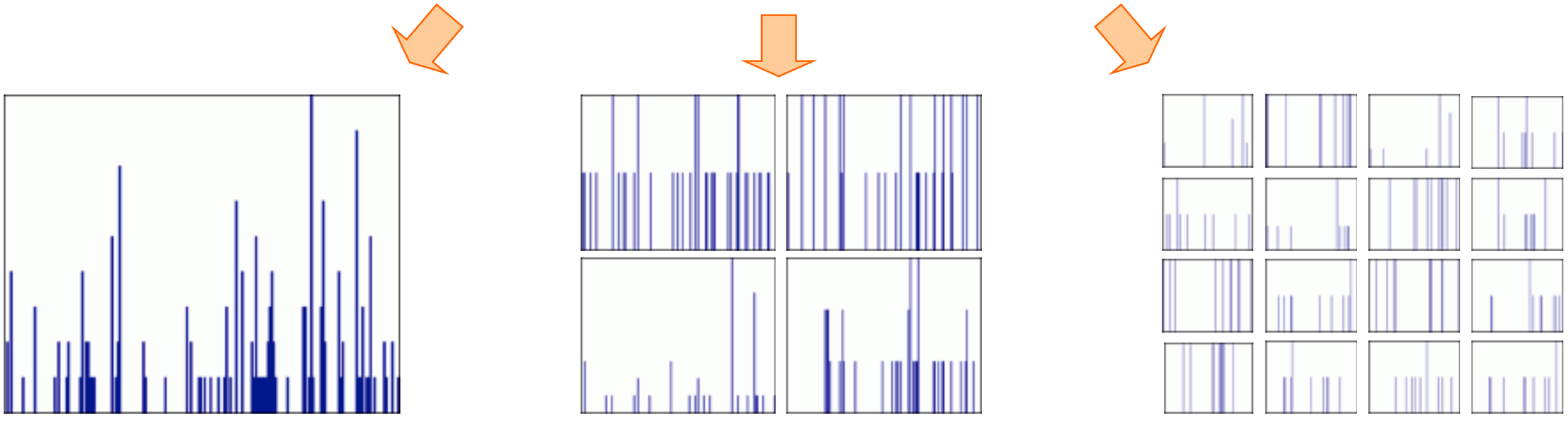




Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial Pyramid representation





Lazebnik, Schmid & Ponce (CVPR 2006)