

[Vondrick, IJCV'16]



[Chen, ECML'18]

16.485: VNAV - Visual Navigation for Autonomous Vehicles

Luca Carlone

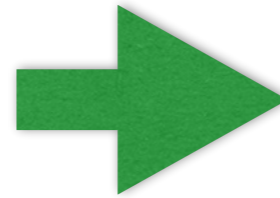
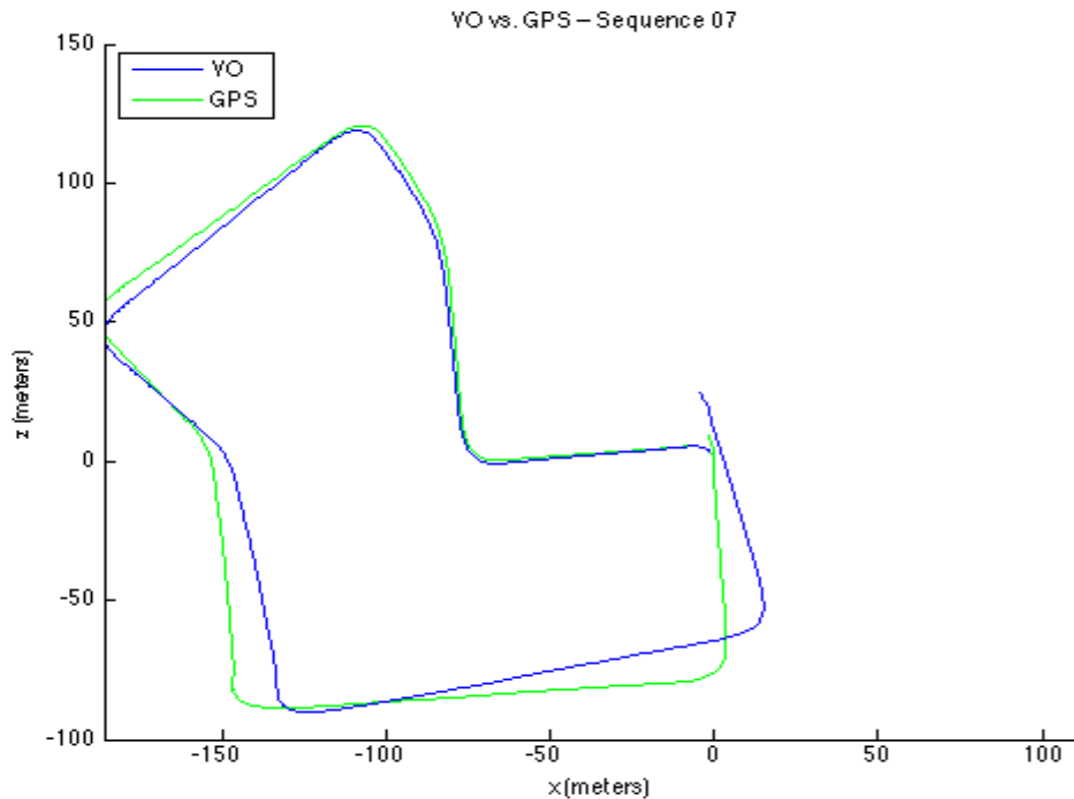
Lecture 22: Place Recognition
and Object Detection



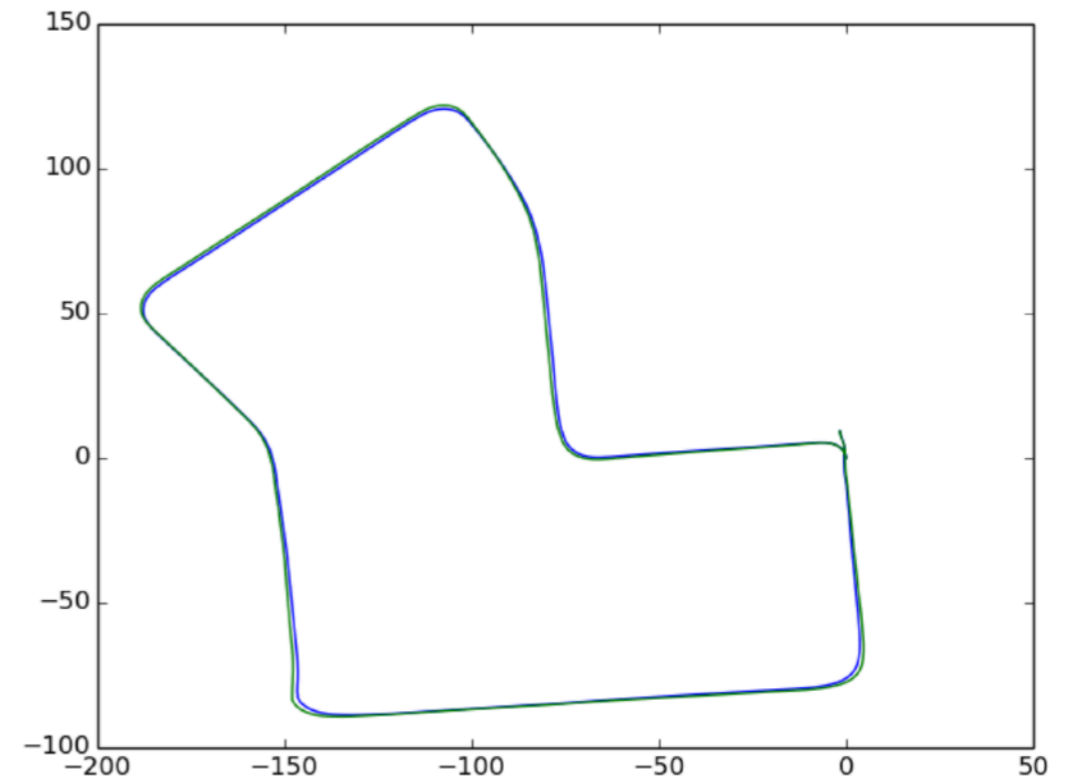
based on slides by Kasra Khosoussi

Next Week: SLAM

Visual odometry



SLAM



SLAM requires:

- place recognition => loop closure detection and / or
- Object detection => landmark detection

Today

- **Place recognition - Bag of Words**
- **Object detection / recognition**

Visual Place Recognition: A Survey

Stephanie Lowry, Niko Sünderhauf, Paul Newman, *Fellow, IEEE*, John J. Leonard, *Fellow, IEEE*, David Cox, Peter Corke, *Fellow, IEEE*, and Michael J. Milford, *Member, IEEE*

Abstract—Visual place recognition is a challenging problem due to the vast range of ways in which the appearance of real-world places can vary. In recent years, improvements in visual sensing capabilities, an ever-increasing focus on long-term mobile robot autonomy, and the ability to draw on state-of-the-art research in other disciplines—particularly recognition in computer vision and animal navigation in neuroscience—have all contributed to significant advances in visual place recognition systems. This paper presents a survey of the visual place recognition research landscape. We start by introducing the concepts behind place recognition—the role of place recognition in the animal kingdom, how a “place” is defined in a robotics context, and the major components of a place recognition system. Long-term robot operations have revealed that changing appearance can be a significant factor in visual place recognition failure; therefore, we discuss how place recognition solutions can implicitly or explicitly account for appearance change within the environment. Finally, we close with a discussion on the future of

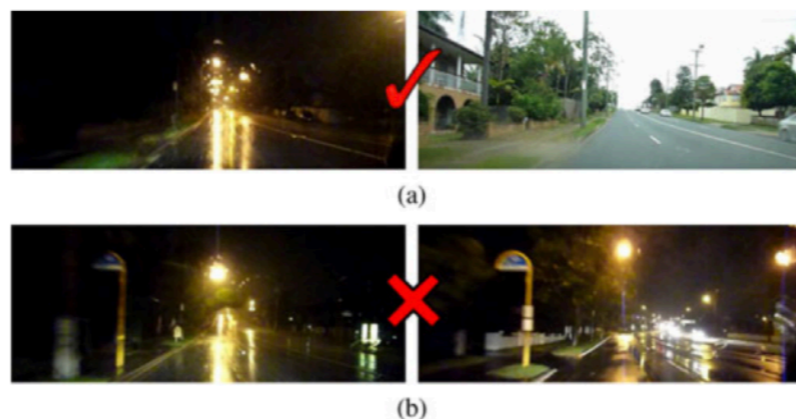


Fig. 1. Visual place recognition systems must be able to (a) successfully match very perceptually different images while (b) also rejecting incorrect matches between aliased image pairs of different places.

+ a few more recent papers

Guess the Speaker - Speech #1

Vocabulary	Freq
America	3
new	21
knowledge	8
first	8
years	9
made	6
now	6
history	5
man	8
science	6
will	20
space	18
Iran	0
behind	5
moon	5
sanctions	0



President John F. Kennedy: 'We choose to go to the moon'

Credit: NASA

October 9, 2017 | 2:07 PM EDT

President John F. Kennedy gave a speech at Rice University in 1962 about the quest to put a man on the moon. "We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard," he said to a cheering crowd.

Guess the Speaker - Speech #2

Vocabulary	Freq (#2)	Freq (#1)
America	12	3
new	10	21
knowledge	0	8
first	2	8
years	0	9
made	5	6
now	3	6
history	4	5
man	0	8
science	0	6
will	40	20
space	0	18
Iran	11	0
behind	1	5
moon	0	5
sanctions	4	0

Are they similar? Not quite ...

For this particular vocabulary, the angle between the two vectors (histograms) is about 50 [deg]

Idea

Use the **distribution of a special set of words** to **efficiently** retrieve a query document (or find similar ones) from a **large** database

Bag of Words (Natural Language Processing)

Representation:

- ▶ Build a **vocabulary**
- ▶ Represent documents as distributions (histograms) over the vocabulary

BoW : document \mapsto histogram(document|vocabulary)

Document Retrieval:

- ▶ Store the BoW histogram for every document in a DB
- ▶ Represent the **query** document as a histogram
- ▶ Compare the query histogram with histograms of documents in DB
- ▶ Return the best (or best n) matches
- ▶ Verify potential matches

Bag of Visual Words

Representation:

- ▶ Build a **visual vocabulary**
- ▶ Represent **images** as distributions (histograms) over the vocabulary

BoVW : **image** \mapsto histogram(**image**|vocabulary)

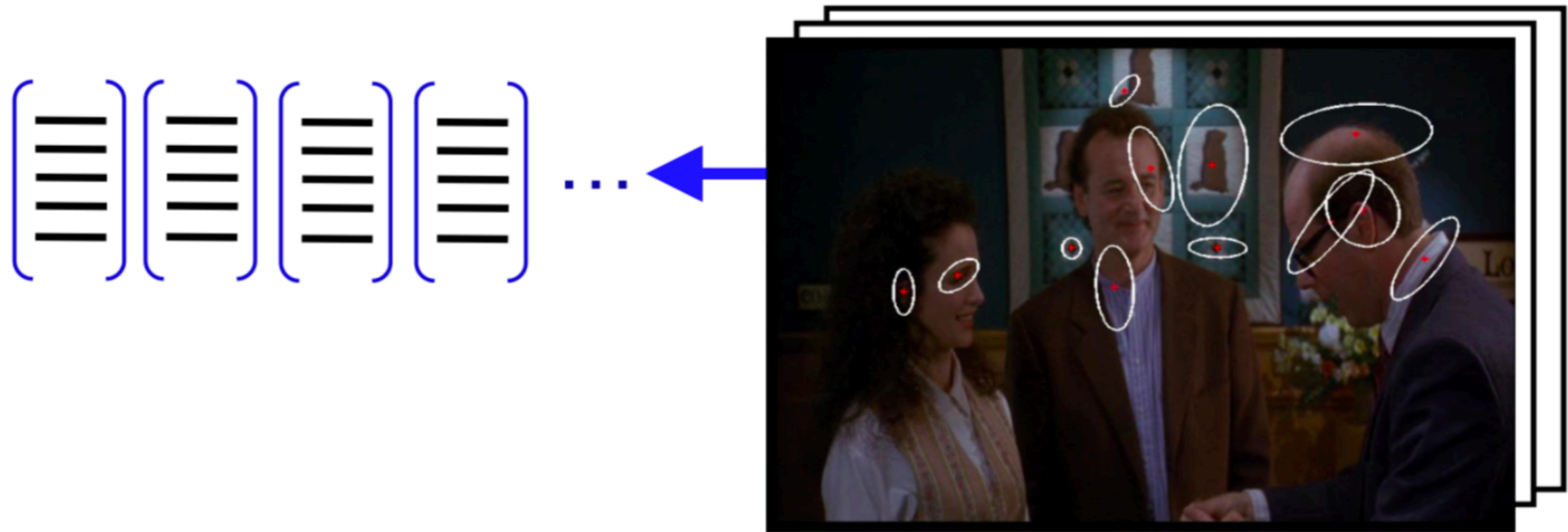
Image Retrieval:

- ▶ Store the BoW histogram for every **image** in a DB
- ▶ Represent the **query image** as a histogram
- ▶ Compare the query histogram with histograms of **images** in DB
- ▶ Return the best (or best n) matches
- ▶ Verify potential matches using **geometric/spatial** verification (**RANSAC**)

Build the Vocabulary

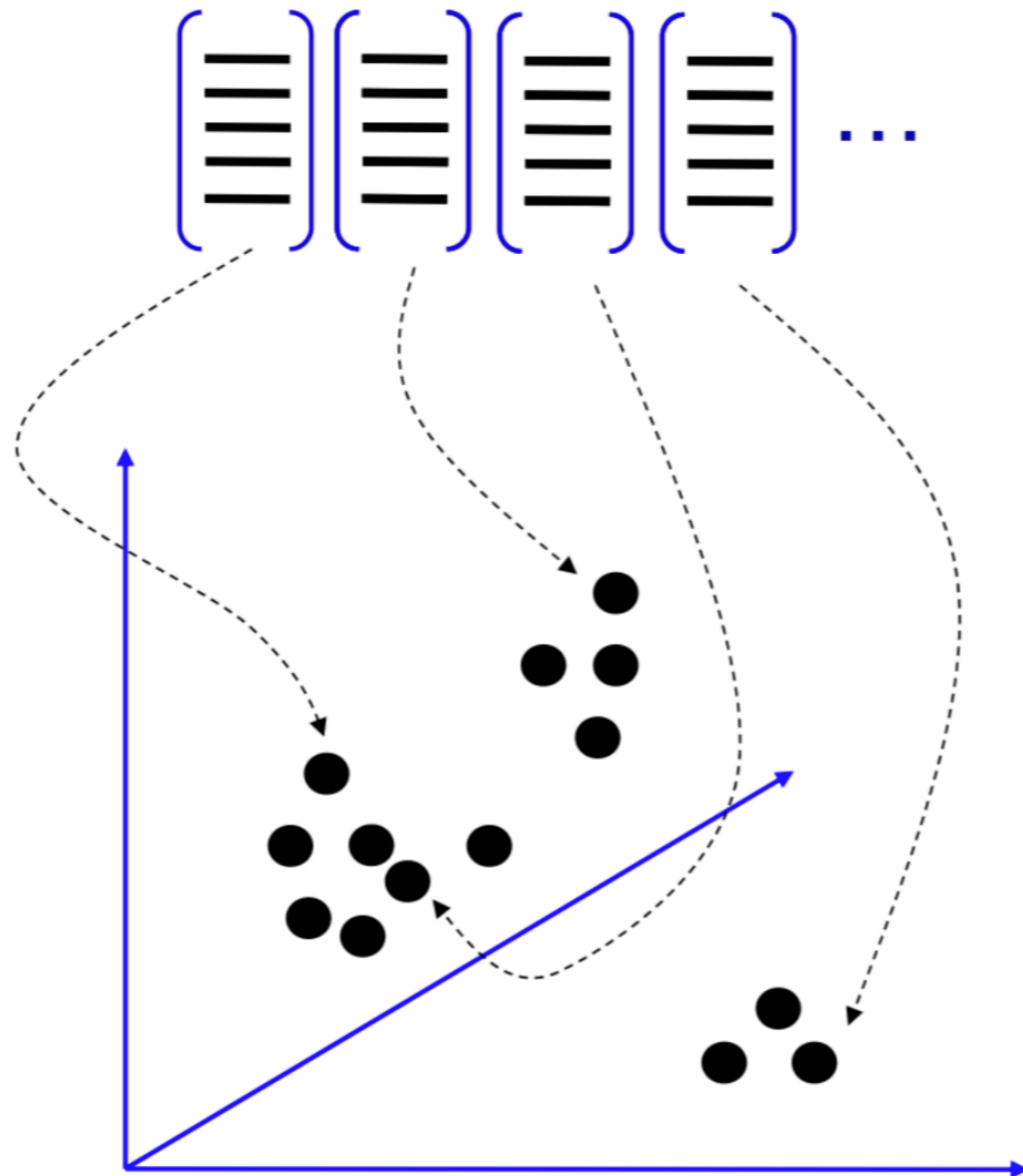
- ① Pick a set of images
- ② Extract keypoints and their descriptors from every image
 - ▶ Need to be fast, invariant to viewpoint variations, etc.
- ③ Cluster the descriptors into k clusters (using, e.g., *k-means*)
- ④ Pick the k cluster centers as your vocabulary

Extract Keypoints and Descriptors

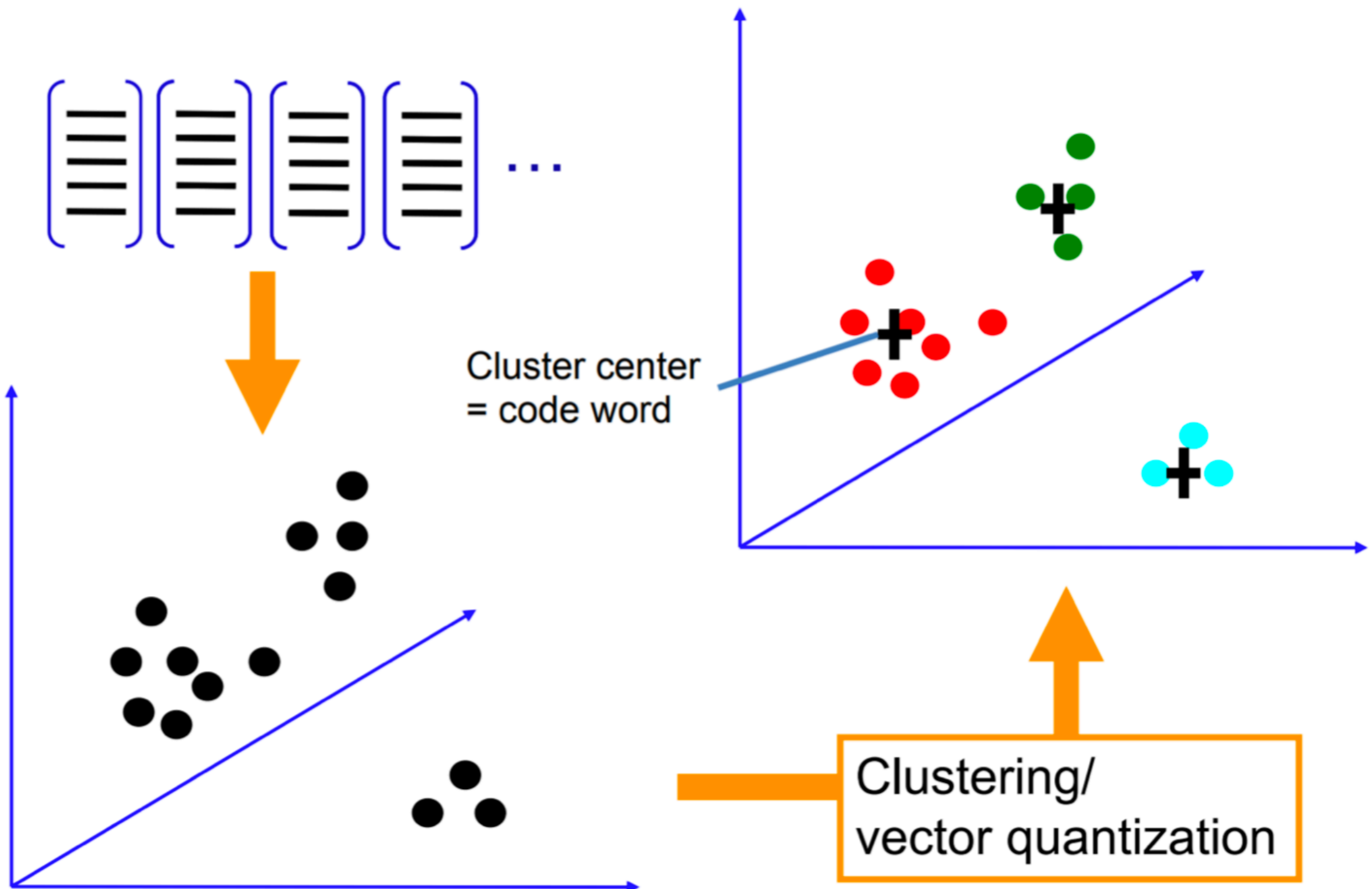


Credit: *Fei-Fei Li*

Descriptor Space



Cluster the the Descriptors to Build the Vocabulary



k-means Clustering

Find a *k*-partitioning (clustering)

$\{\mathcal{C}_i\}_{i=1}^k$ for \mathcal{X} by minimizing

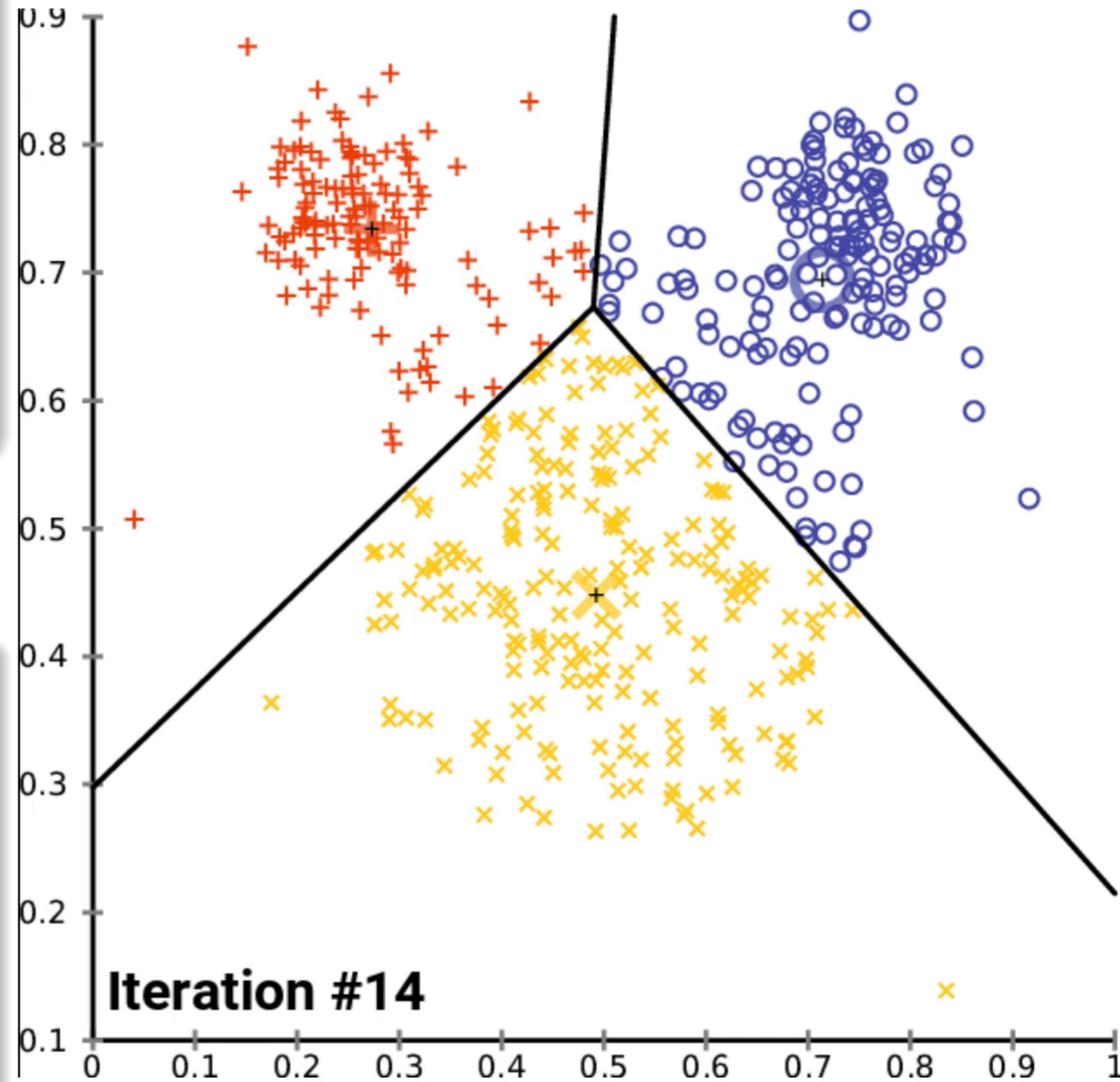
$$\sum_{i=1}^k \sum_{x \in \mathcal{C}_i} \|x - \mu_i\|^2$$

where μ_i is the mean of cluster \mathcal{C}_i

This is NP-hard - A simple idea:

Initialize cluster centers and, until convergence, alternate between

- 1 Associating points to nearest cluster centers
 \Leftrightarrow solve for \mathcal{C}_i 's given μ_i 's
- 2 Computing cluster centers given the associations
 \Leftrightarrow solve for μ_i 's given \mathcal{C}_i 's

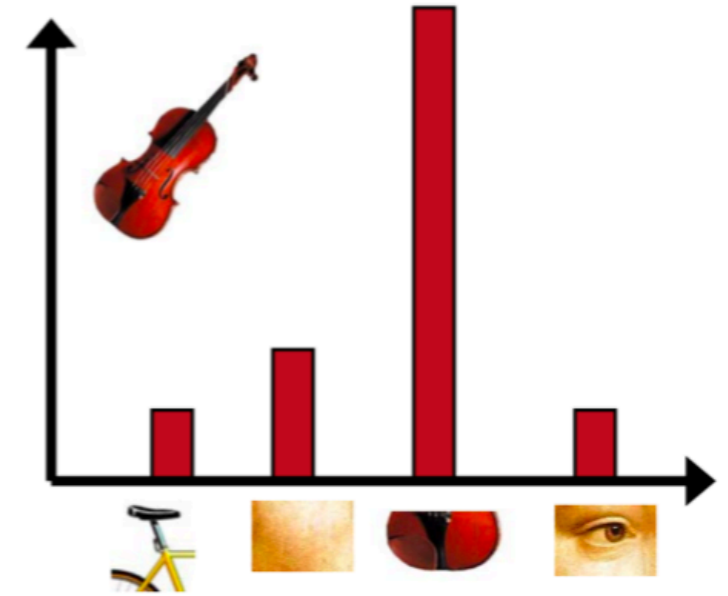
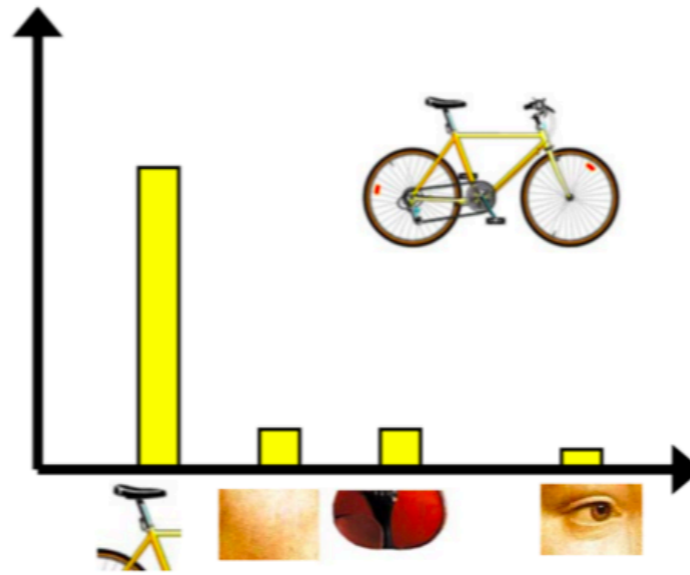
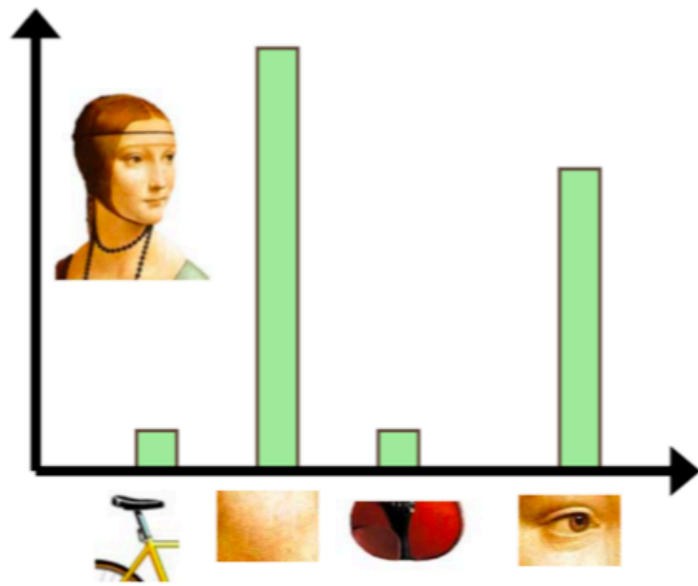


Representation



Credit: *Fei-Fei Li*

Representation



Search in DB for a Query Image via Inverted File Index

- ▶ Given a query image, we need to search the database for similar images
- ▶ The database is large - in SLAM, it's always growing!

- ▶ **Idea:** for each visual word, maintain a list of images that contain that word
- ▶ Given a query image:
 - ① Extract visual words (i.e., BoW representation)
 - ② Look up the inverted file index (DB) to find documents containing same words
 - ③ Sort candidates based on weighted distance/similarity between BoW vectors

Index

- A*-optimal experiment design, 387
- abstract form convex problem, 137
- active constraint, 128
- activity planning, 149, 195
- aff** (affine hull), 23
- affine
 - combination, 22
 - dimension, 23
 - function, 36
 - composition, 79, 95, 508, 642, 645
 - hull, 23
 - independence, 32
 - invariance, 486
 - analytic center, 449
- angle, 633
 - approximation, 448
 - constraint, 406
 - problem, 405, 408
- anisotropy**, 461
- approximate Newton method, 519
- approximation
 - Chebyshev, 6, 293
 - complex, 197
 - fitting angles, 448
 - ℓ_1 -norm, 193, 294
 - least-squares, 293
 - log-Chebyshev, 344
 - matrix norm, 194

TF-IDF Weights

- ▶ **Issue:** Relying on very common words could be misleading (e.g., “the”, “is”)
 - ▶ Both speeches contained many instances of “will” – any speech in the world would contain tons of these!
- ▶ On the other hand, unique/rare words are very informative
 - ▶ Not every presidential speech contains the word “moon”!
- ▶ **Solution:** For each word in the vocabulary, multiply its “term frequency” (TF) (i.e., histogram bar) by its “inverse document frequency” (IDF) in the (training) database

$$\text{IDF weight for word } i \triangleq \log \left(\frac{\# \text{ “documents”}}{\# \text{ “documents” that contain } i\text{th word}} \right)$$

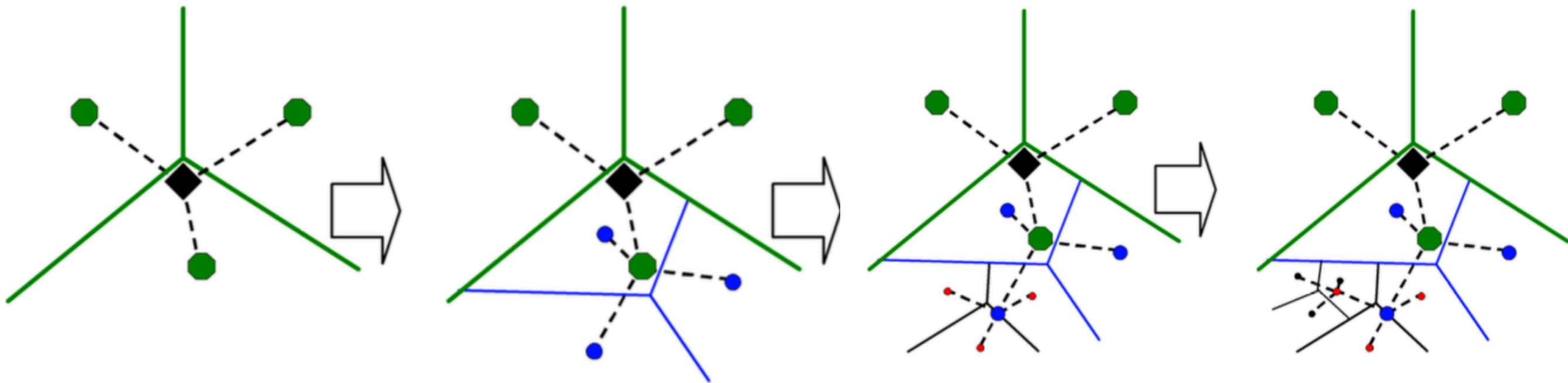
- ▶ Total weight for word $i = \text{TF}_i \times \text{IDF}_i$ (i.e., i th component of the BoW vector)
- ▶ Comparing two (very sparse) BoW vectors:

$$\text{dist}(v_{\text{query}}, v_{\text{DB}}) = \left\| \frac{v_{\text{query}}}{\|v_{\text{query}}\|} - \frac{v_{\text{DB}}}{\|v_{\text{DB}}\|} \right\|$$

- ▶ Many dist/similarity functions, norms (ℓ_1 and ℓ_2) and normalization schemes

Need Large Vocabularies: Vocabulary Tree

- ▶ Faster quantization (logarithmic time complexity in vocabulary size)
- ▶ Therefore can afford larger vocabularies
- ▶ Hierarchical clustering:



Credit: *Nister and Stewenius*

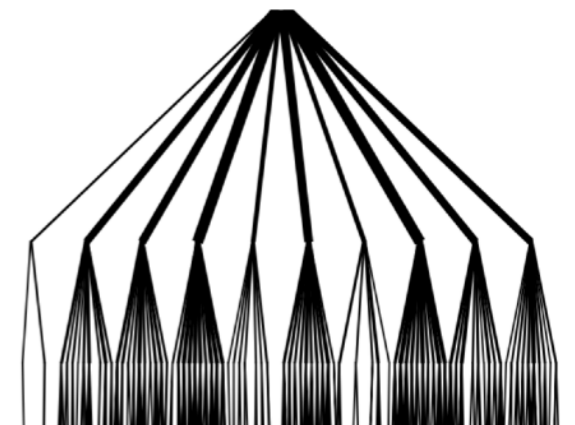
Scalable recognition with a vocabulary tree

D Nister, H Stewenius

2006 IEEE Computer Society Conference on Computer Vision and Pattern ...

4135

2006



BoW-based Loop-Closure Detection in Action

Bags of Binary Words for Fast Place Recognition in Image Sequences

667

2012

D Gálvez-López, JD Tardos

IEEE Transactions on Robotics 28 (5), 1188-1197

Bags of Binary Words for Fast Place Recognition in Image Sequences

Dorian Gálvez-López, Juan D. Tardós

*Robotics, Perception and Real Time Group
Departamento de Informática e Ingeniería de Sistemas
Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza, Spain*



Bags of Binary Words for Fast Place Recognition in Image Sequences

Dorian Gálvez-López, Juan D. Tardós

*Robotics, Perception and Real Time Group
Departamento de Informática e Ingeniería de Sistemas
Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza, Spain*