Machine Perception

Lecture 6: Localization and SLAM (Part II)

Piotr Skrzypczyński
Institute of Robotics and Machine Intelligence
Poznań University of Technology

Lecture outline

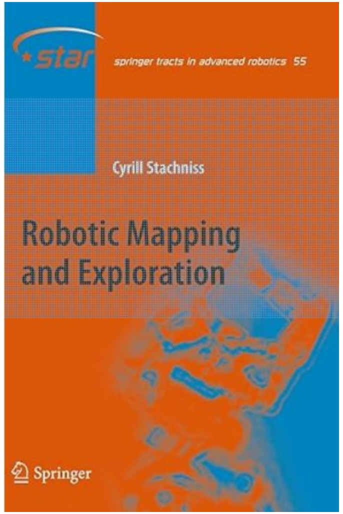
- Visual Place Recognition - definition, variants, role in SLAM.
- Approaches to Visual Place Recognition .
- Bag of Visual Words.
- Deep learning based approach - NetVLAD.
- Learning based place recognition in robotics.

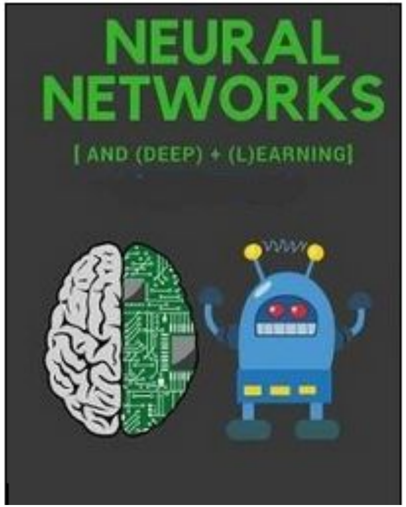
Literature

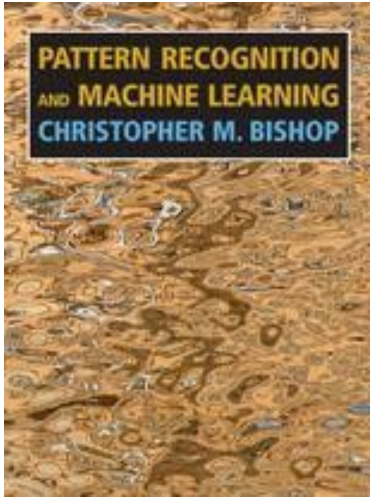
1. C. Stachniss, *Robotic Mapping and Exploration*, Springer Verlag, 2009.

2. M. Nielsen, *Neural Networks and Deep Learning* http://neuralnetworksanddeeplearning.com/

3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag Berlin, Heidelberg, 2006.







Place recognition

- Robotics : has the robot been to this place before ? Which images were taken around the same location ?
- Image retrieval : have I seen this image before? Which images in my database look similar to it ?

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**

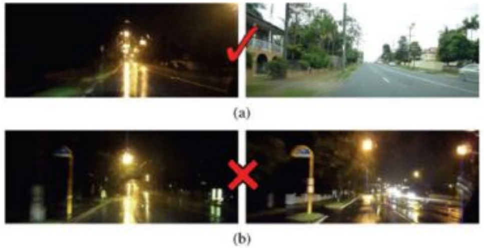


Fig. 1. Visual place recognition systems must be able to (a) successfully match very perceptually different images while (b) also rejecting incorrect matches between aliased image pairs of different places.

1. S. Lowry et al., "Visual Place Recognition: A Survey", IEEE Transactions on Robotics, vol. 32, no. 1, pp. 1-19, 2016

Place recognition

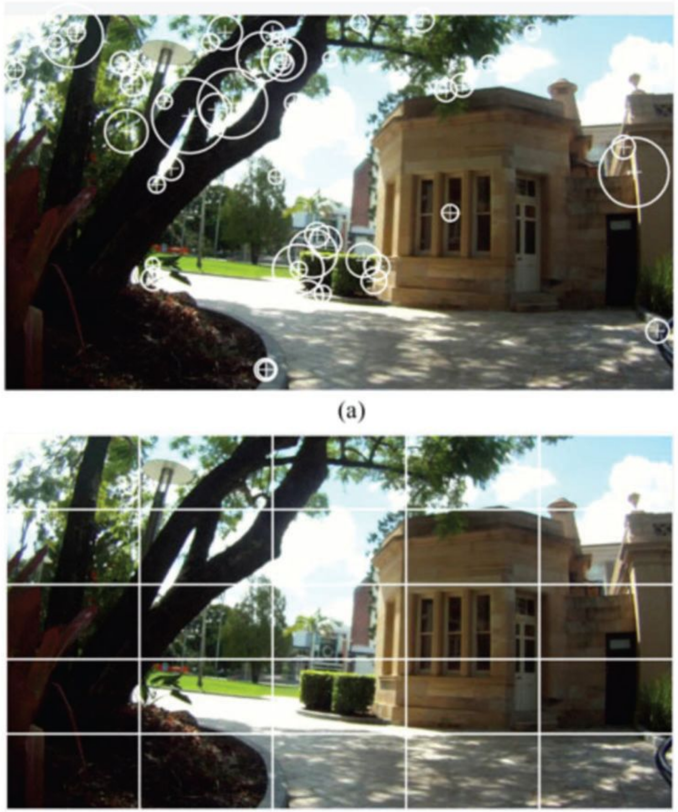
- **Visual Place Recognition**
  - goal : query an image in a database of N images
  - complexity: NM^2 feature comparisons (assumes each image has M features)

- **Appearance changes:** illumination, weather conditions, dynamic objects (people, cars,…), Viewpoint changes
- **Perceptual aliasing:** two different places may look similar (building, roads, …)

Place recognition

- **Approaches**

  - Local descriptors
  - Global descriptors
  - Learning-based methods
- Use analogies from text retrieval:
- Visual Words
- Vocabulary of Visual Words
- “Bag of Words” (BoW) approach



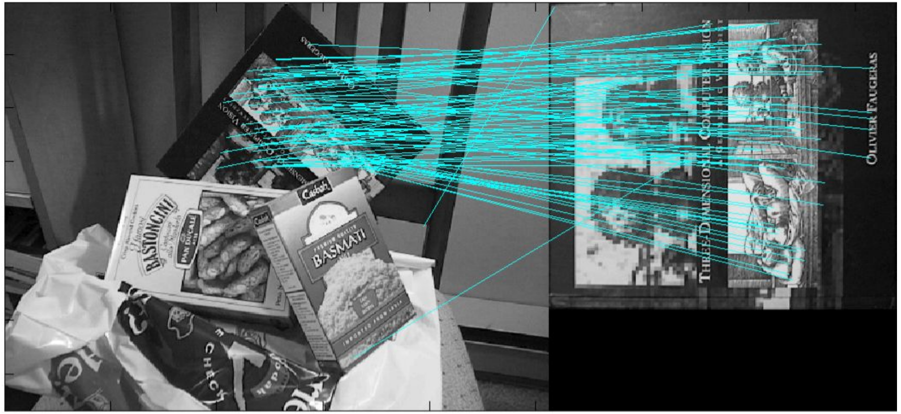
(a)

(b)

Place recognition

- **Approaches**

  - Local descriptors
  - Global descriptors
  - Learning-based methods
- Use analogies from text retrieval:
- Visual Words
- Vocabulary of Visual Words
- “Bag of Words” (BoW) approach



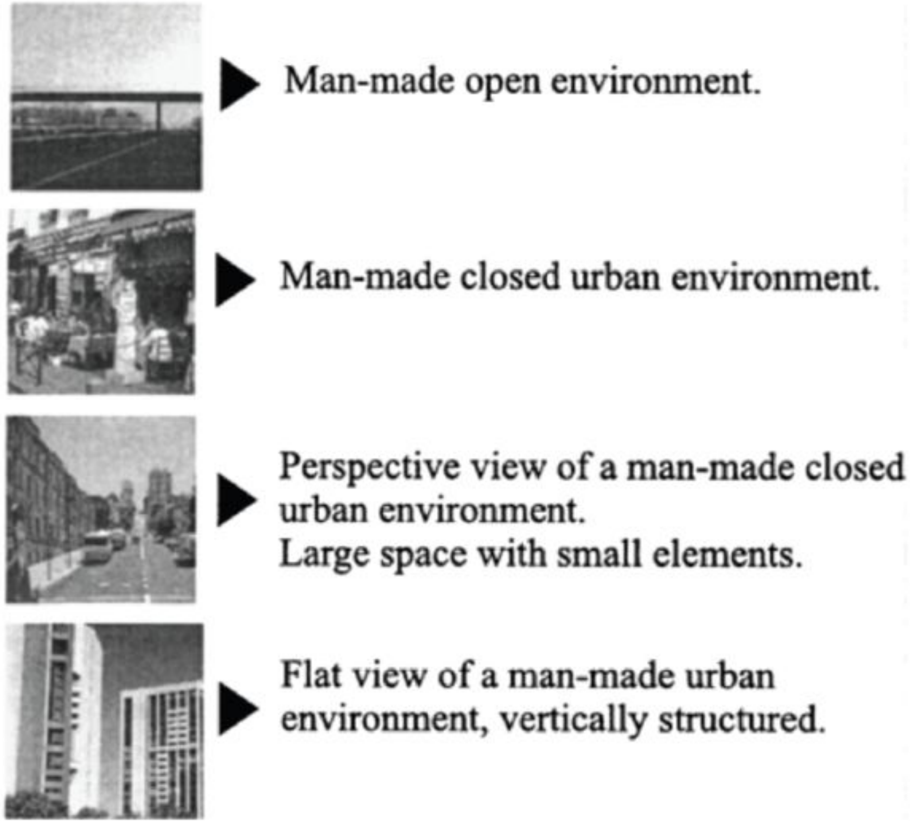
Global descriptors

Early approaches:

- color histograms
- principal component analysis
- •other statistics on edges,corners, and color patches

GIST descriptor:

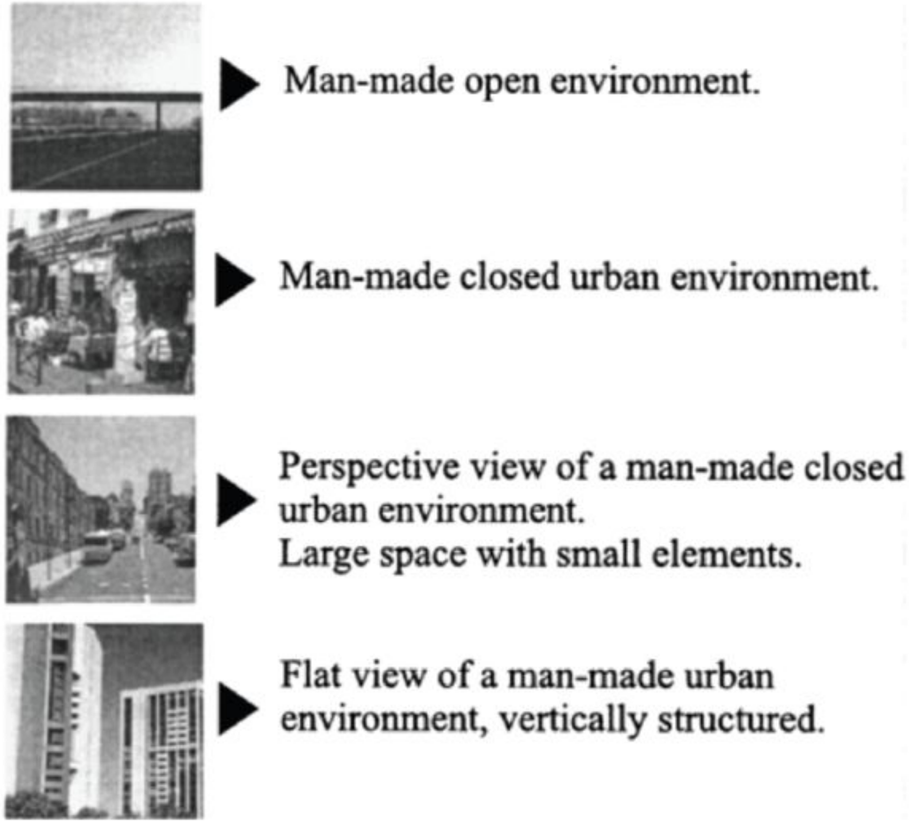
- image is filtered at different orientations and different frequencies to extract information from the image
- results are averaged to generate a compact vector that represents the “gist” of a scene



Global vs. local descriptors

Global descriptors:
- better at handling lighting conditions and seasonal variation
- more sensitive to viewpoint changes

Local descriptors:
- allow estimating feature (andcamera) geometry
- sensitive to lighting conditions and seasonal variations



Local descriptors: Bag of Visual Words

- Based on text retrieval and summarization methods

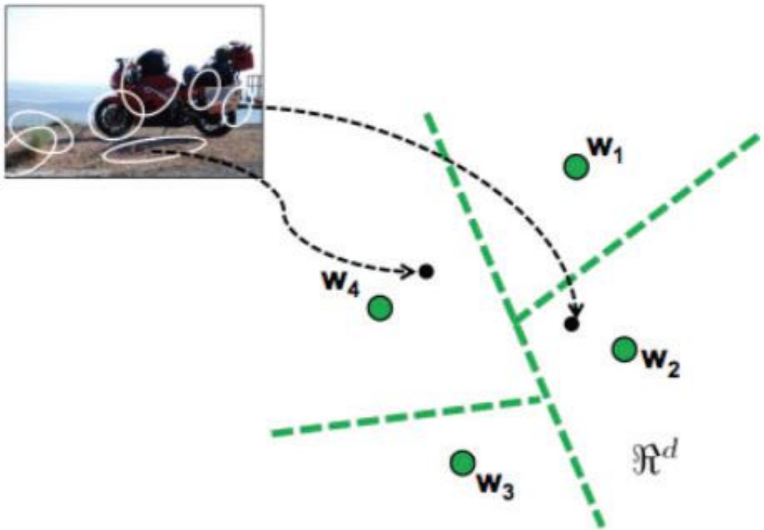
  1) Extract features and descriptors in image

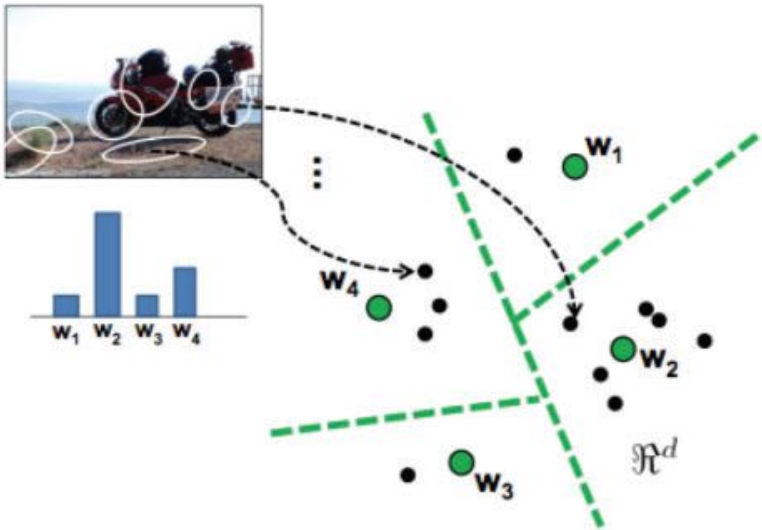
  2) Discretize feature space (clustering)

  3) Store the frequency of the features for each image

  Each cluster is a "visual word"

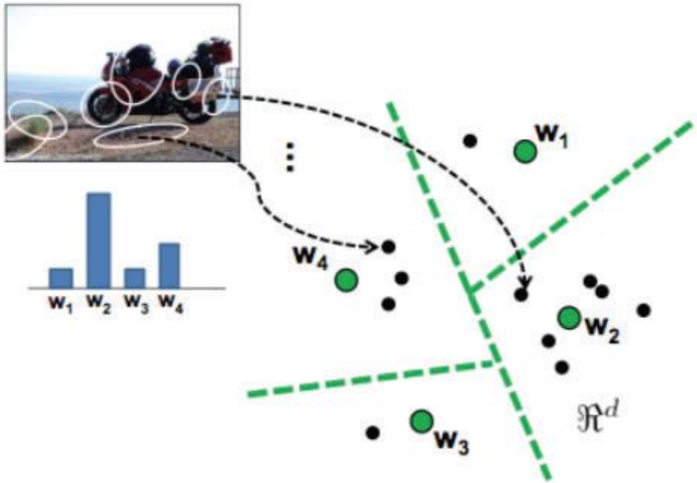
J. Sivic, A. Zisserman. Video Google: A text retrieval approach to object matching in videos. ICCV, 2003

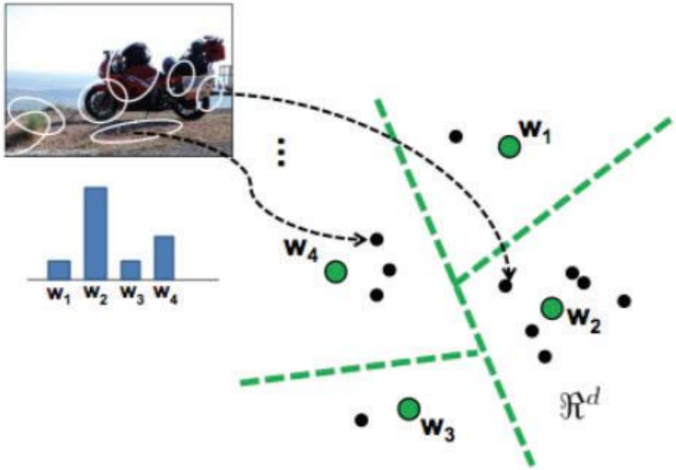




Local descriptors: Bag of Visual Words

- Two images are compared based on the corresponding histogram (Hamming distances, other metrics, …)
- Faster version: vocabulary tree
- Alternative: VLAD (Vector of Locally Aggregated Descriptors)

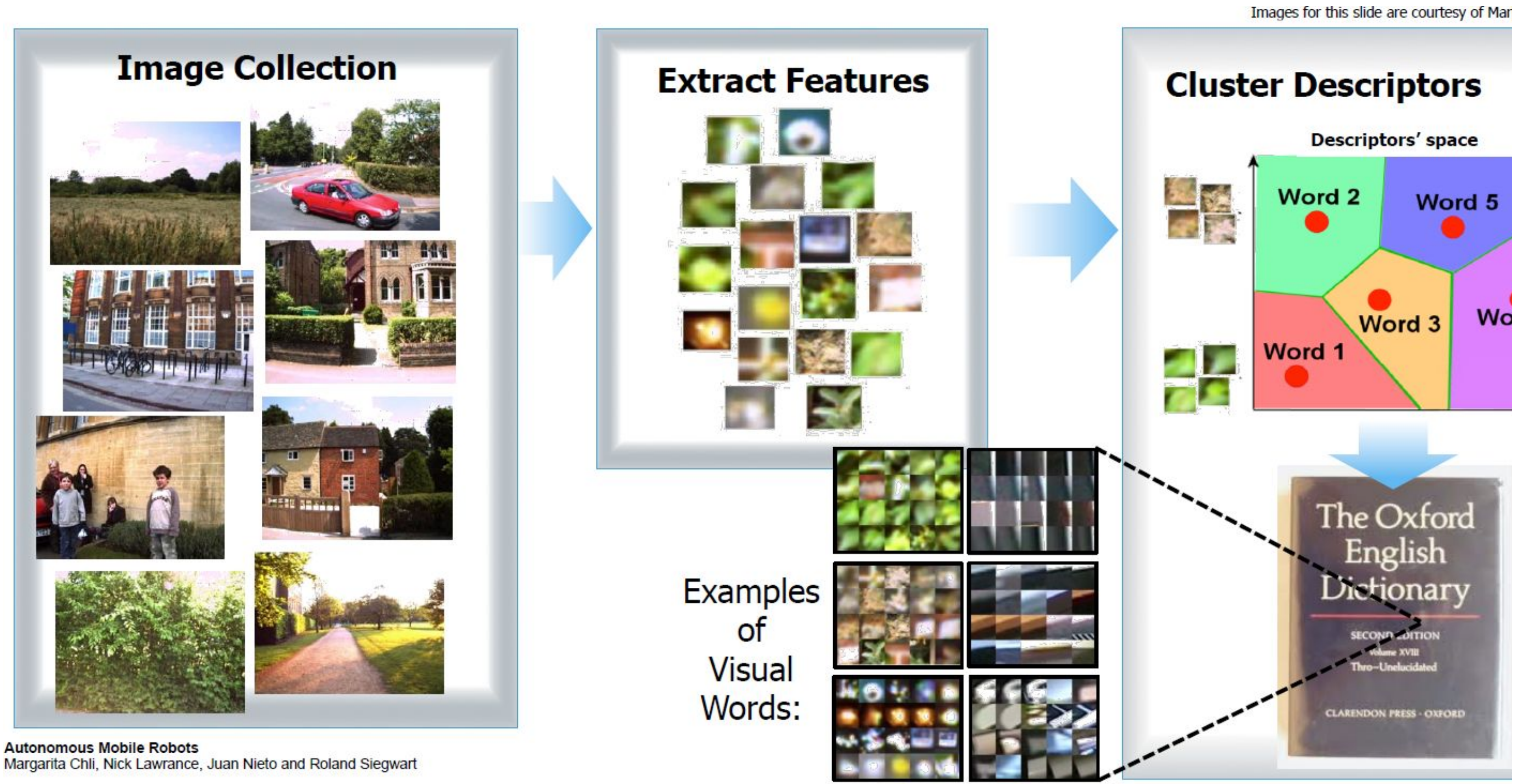




Place recognition (Bag of Visual Words)

- Collect a large enough dataset that is representative of all possible images that are relevant to your application (e.g., for automotive place recognition, you may want to collect million of street images sampled around the world)
- Extract features and descriptors from each image and map them into the descriptor space (e.g., for SIFT, 128 dimensional descriptor space)
- Cluster the descriptor space into *K* clusters
- The centroid of each cluster is a visual word.
- This is computed by taking the arithmetic average of all the descriptors within the same cluster, e.g., for SIFT, each cluster contains SIFT features that are very similar to each other;
- The visual word then is the average all the SIFT descriptors in that cluster

Place recognition (Bag of Visual Words)



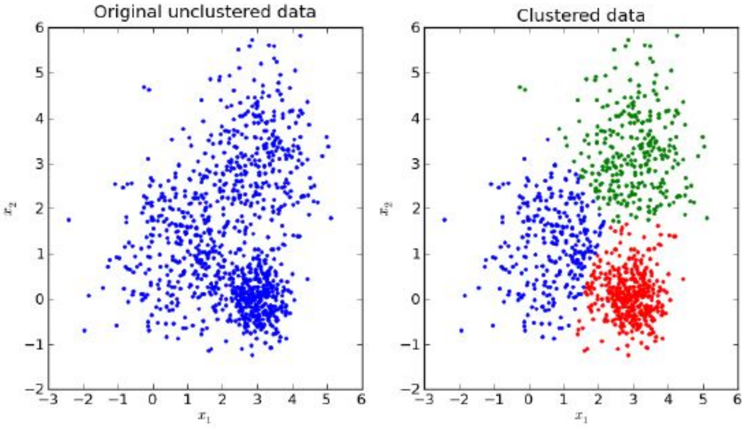
Autonomous Mobile Robots
Margarita Chli, Nick Lawrance, Juan Nieto and Roland Siegwart

Place recognition (Bag of Visual Words)

- The **k-means** clustering algorithm partitions n data point into k clusters in which each data point $\boldsymbol{x}$ belongs to the cluster $\boldsymbol{S}_i$ with center $\boldsymbol{m}_i$
- It minimizes the sum of squared Euclidean distances between points $\boldsymbol{x}\boldsymbol{x}$and their nearest cluster centers $\boldsymbol{m}_i$
- **Algorithm:**

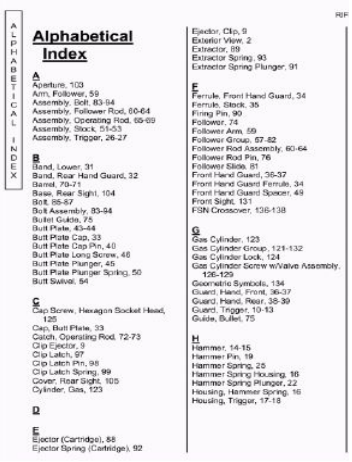
1. Randomly initialize k cluster centers
2. Iterate until convergence:
3. Assign each data point $\boldsymbol{x}_j$ to the nearest center $\boldsymbol{m}_i$
4. Recompute each cluster center as the mean of all points assigned to it

$$D(X, M) = \sum_{i=1}^{k} \sum_{x \in S_i} (x - m_i)^2$$



Place recognition (Bag of Visual Words)

- The **Image Vocabulary** is a data structure that lists all extracted visual words
- Each visual word is assigned a unique identifier (an integer number)
- Each word in the image vocabulary points to a list of images (from the entire image database) in which that word appears
- If the database grows, the vocabulary is updated accordingly



ALPHABETICAL INDEX

RIF

Alphabetical Index

A
Aperture, 103
Arm, Follower, 59
Assembly, Bolt, 83-94
Assembly, Follower Rod, 60-64
Assembly, Operating Rod, 65-69
Assembly, Stock, 51-53
Assembly, Trigger, 26-27

B
Band, Lower, 31
Band, Rear Hand Guard, 32
Barrel, 70-71
Base, Rear Sight, 104
Bolt, 85-87
Bolt Assembly, 83-94
Bullet Guide, 75
Butt Plate, 43-44
Butt Plate Cap, 33
Butt Plate Cap Pin, 40
Butt Plate Long Screw, 46
Butt Plate Plunger, 45
Butt Plate Plunger Spring, 50
Butt Swivel, 54

C
Cap Screw, Hexagon Socket Head, 125
Cap, Butt Plate, 33
Catch, Operating Rod, 72-73
Clip Ejector, 9
Clip Latch, 97
Clip Latch Pin, 98
Clip Latch Spring, 99
Cover, Rear Sight, 105
Cylinder, Gas, 123

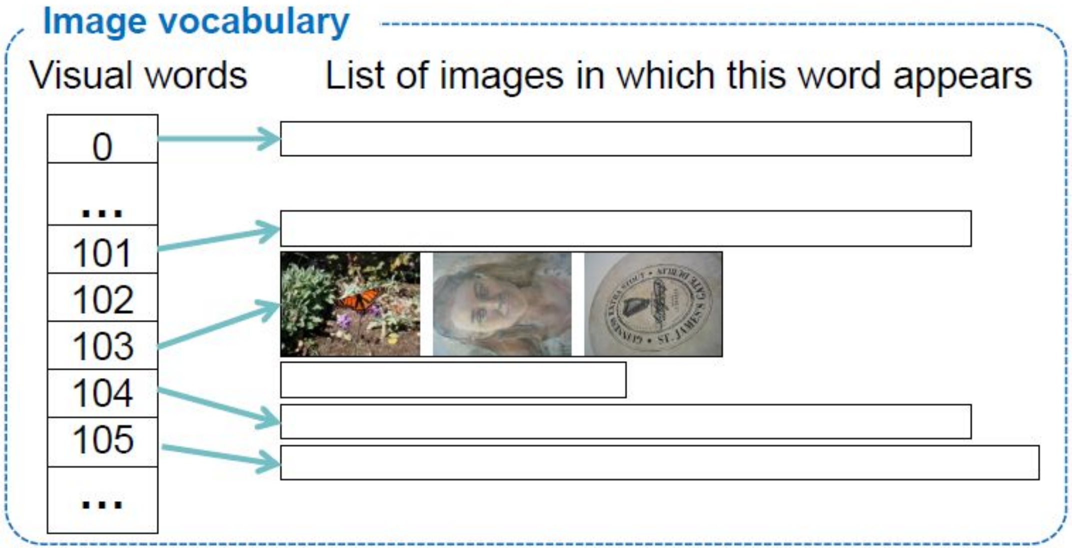
D

E
Ejector (Cartridge), 88
Ejector Spring (Cartridge), 92
Ejector, Clip, 9
Exterior View, 2
Extractor, 89
Extractor Spring, 93
Extractor Spring Plunger, 91

F
Ferrule, Front Hand Guard, 34
Ferrule, Stock, 35
Firing Pin, 90
Follower, 74
Follower Arm, 59
Follower Group, 57-82
Follower Rod Assembly, 60-64
Follower Rod Pin, 76
Follower Slide, 81
Front Hand Guard, 36-37
Front Hand Guard Ferrule, 34
Front Hand Guard Spacer, 49
Front Sight, 131
FSN Crossover, 136-138

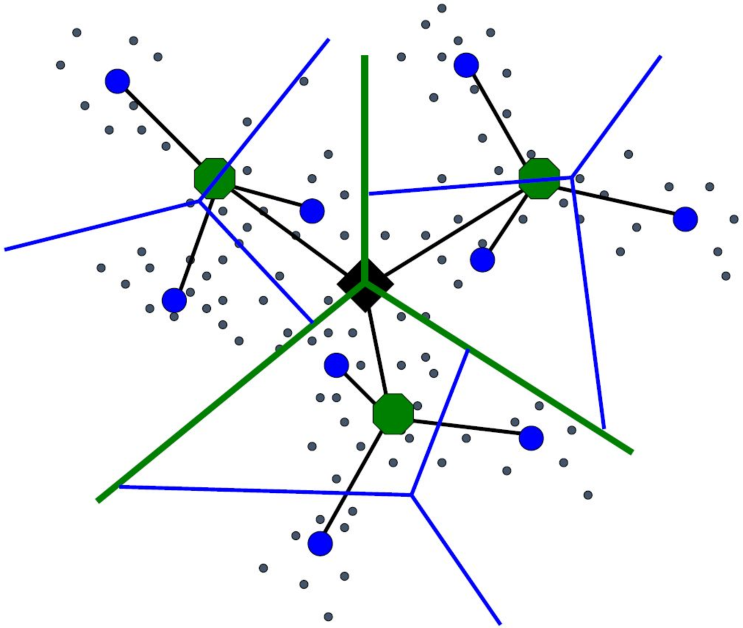
G
Gas Cylinder, 123
Gas Cylinder Group, 121-132
Gas Cylinder Lock, 124
Gas Cylinder Screw w/Valve Assembly, 126-129
Geometric Symbols, 134
Guard, Hand, Front, 36-37
Guard, Hand, Rear, 38-39
Guard, Trigger, 10-13
Guide, Bullet, 75

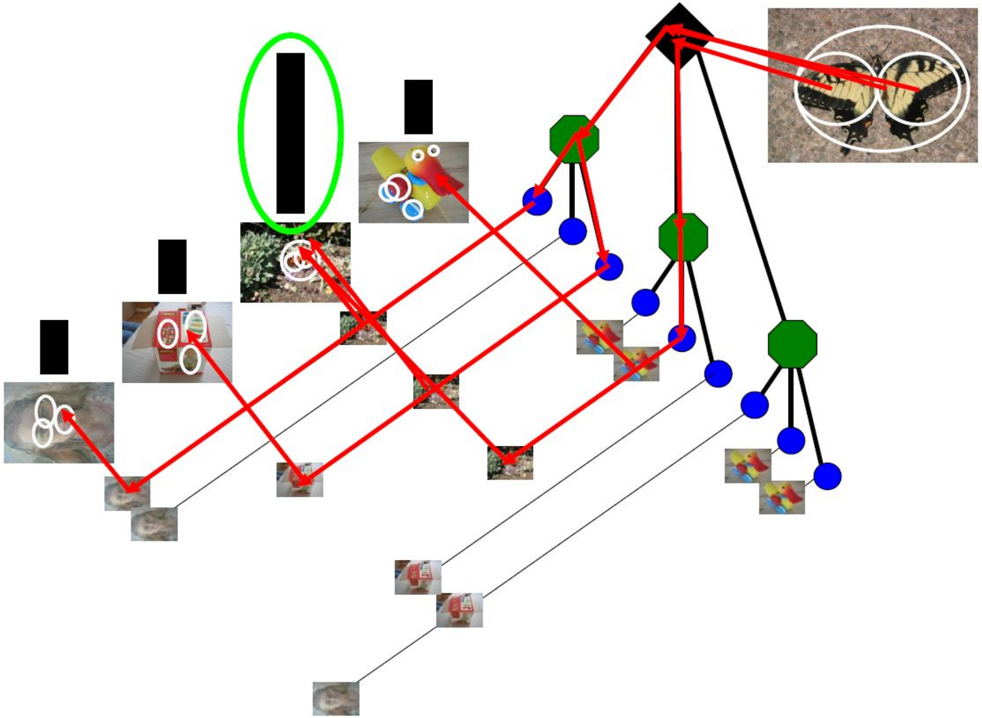
H
Hammer, 14-15
Hammer Pin, 19
Hammer Spring, 25
Hammer Spring Housing, 16
Hammer Spring Plunger, 22
Housing, Hammer Spring, 16
Housing, Trigger, 17-18



Place recognition (Bag of Visual Words)

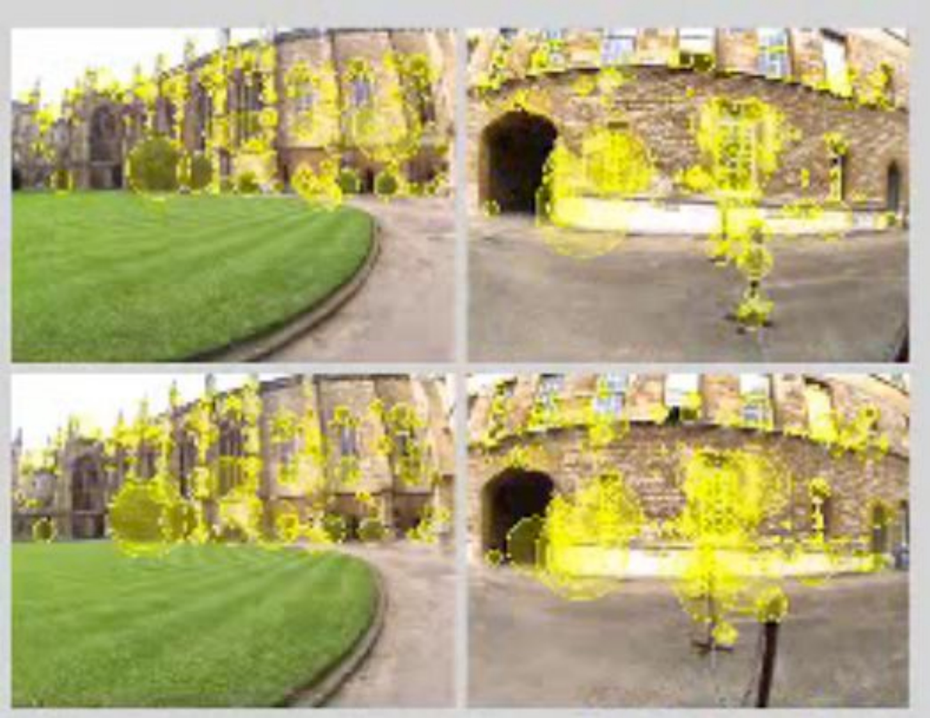
- Within the vocabulary, each visual word points to a list of images where that word occurs.
- During retrieval, each feature contributes to update the voting array.
- The image with most votes is returned.





Place recognition (FAB-MAP)

- Place recognition for robot localization using stereo images
- Build the visual vocabulary using SURF features
- Probabilistic model of the world:
- World = a set of discrete places
- Place = a set of consecutive images
- At a new frame, compute: P(being at a known place), P(being at a new place)
- Captures the dependencies of visual
- words to distinguish the most
- characteristic structure of each
- scene (using the Chow-Liu tree)



Place recognition (FAB-MAP)

robots.ox.ac.uk/~mjc/appearance_based_results.htm

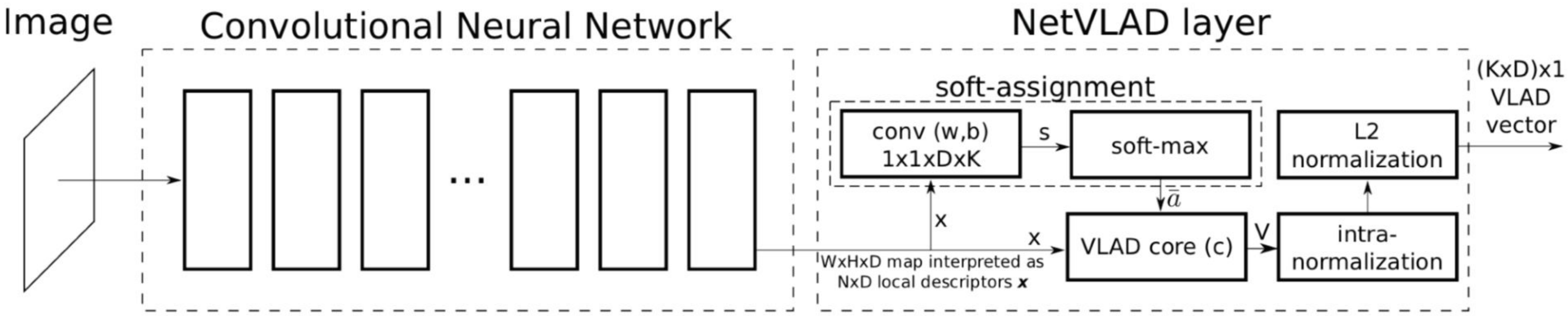
(a) p=0.9989 (b) p=0.999 (c) p=0.999

Place recognition (NetVLAD)

- Earlier approaches using AlexNet or similar and use layers activations as descriptors
- **NetVLAD:**
  - CNN-based approach
  - Trained on the task of place recognition

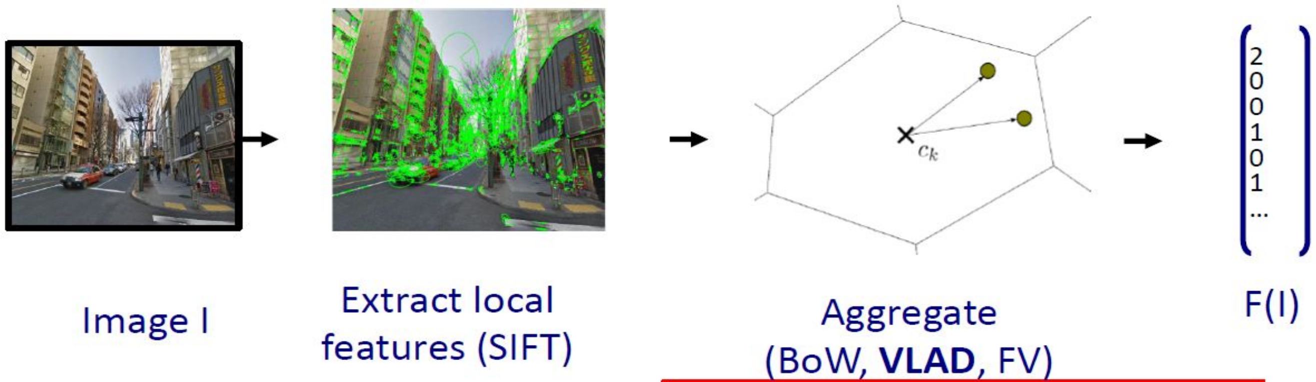
How to get labeled data ?

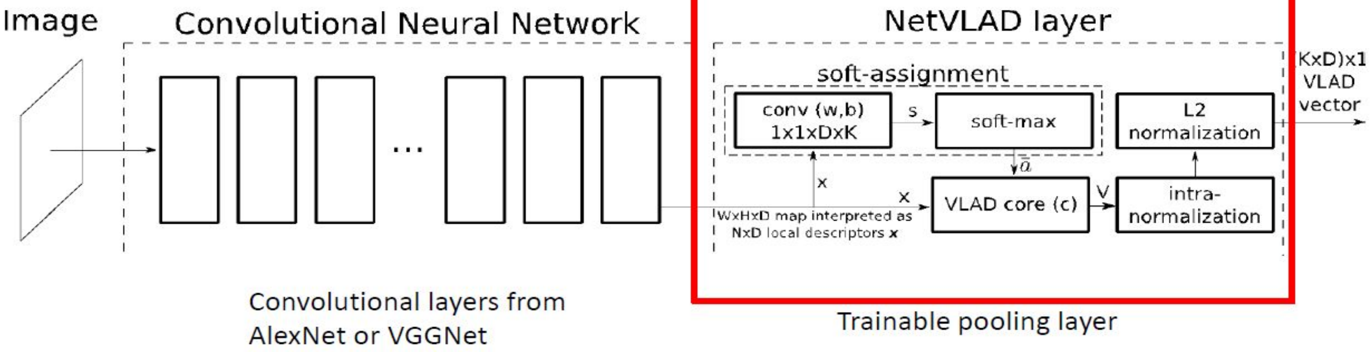
- a large dataset of panoramic images from the Google StreetView
- positions based on their (noisy) GPS
- seasonal variations
- illumination changes



Place recognition (NetVLAD)

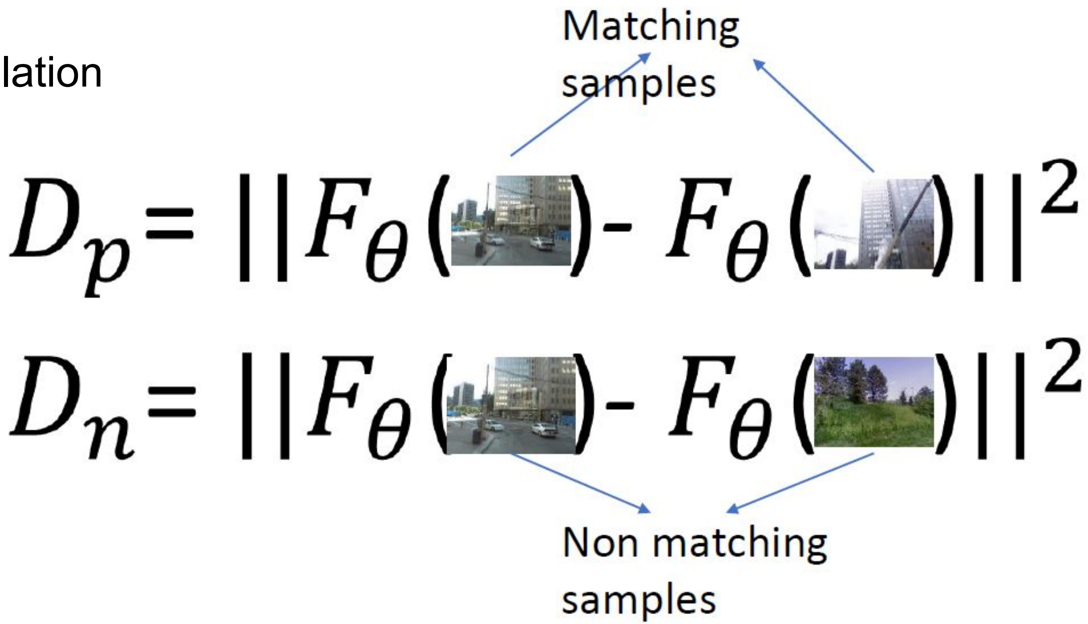
- Mimic the classical pipeline with deep learning

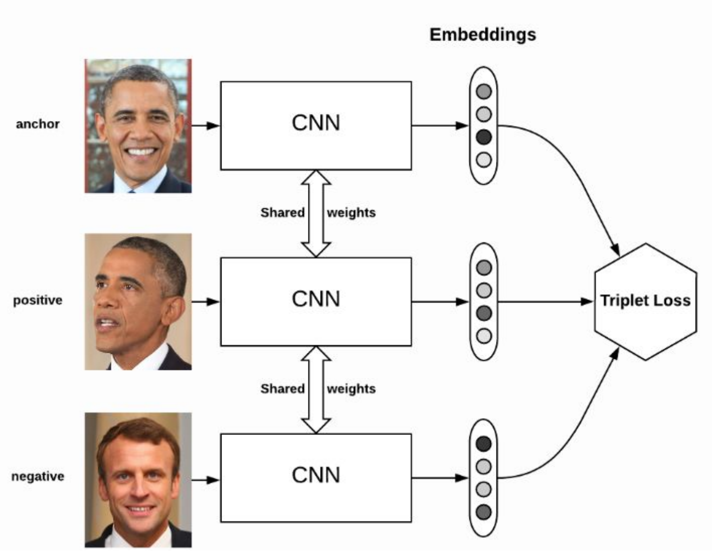




Place recognition (NetVLAD)

- NetVlad loss - triplet loss formulation





$$L_\theta = \sum_{samples} \max(D_{p(\theta)} + \underbrace{m}_{\text{margin}} - D_{n(\theta)}, 0)$$

Place recognition (NetVLAD)

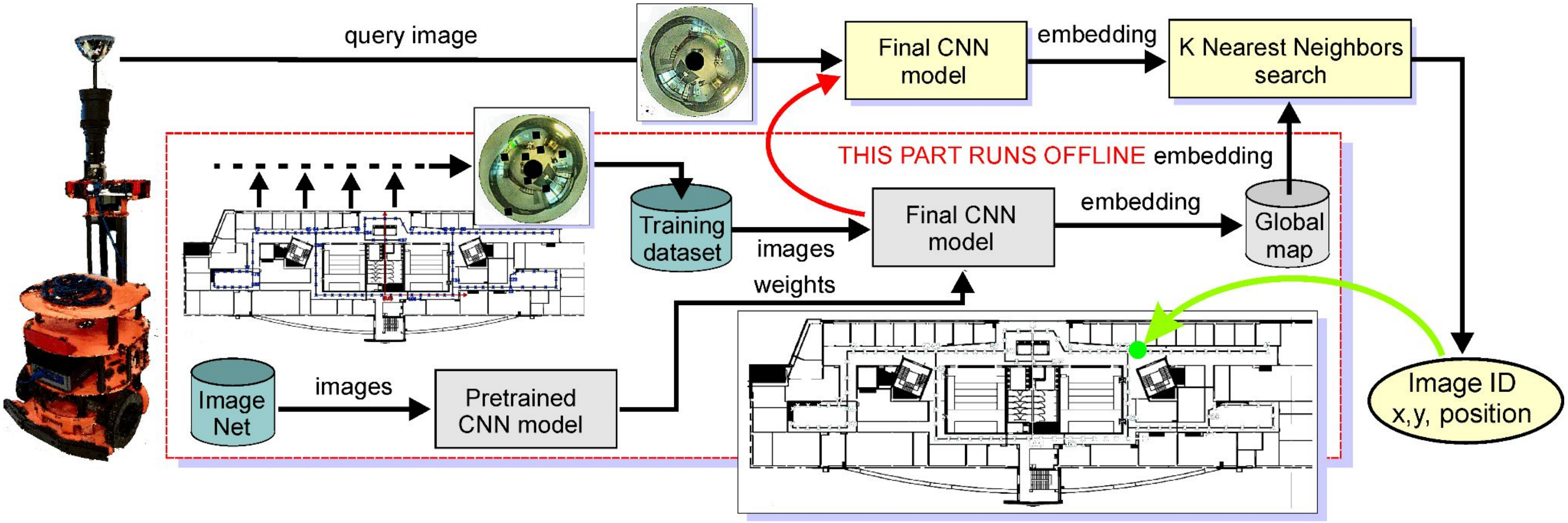
- Code, dataset and trained network available online:
  http://www.di.ens.fr/willow/research/netvlad/

Query

Top result

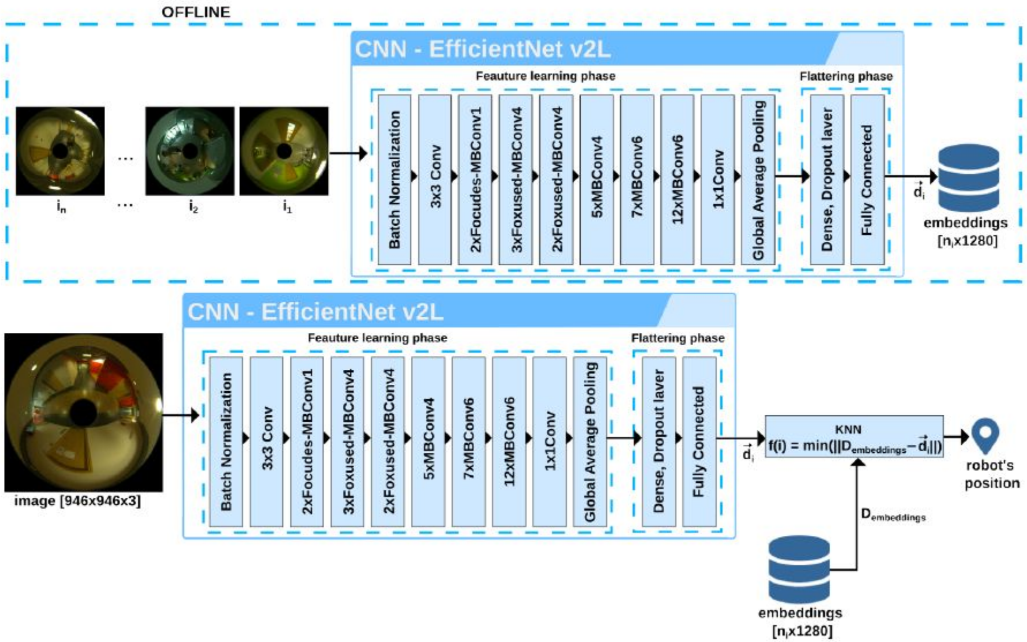
Place recognition with omnidirectional images

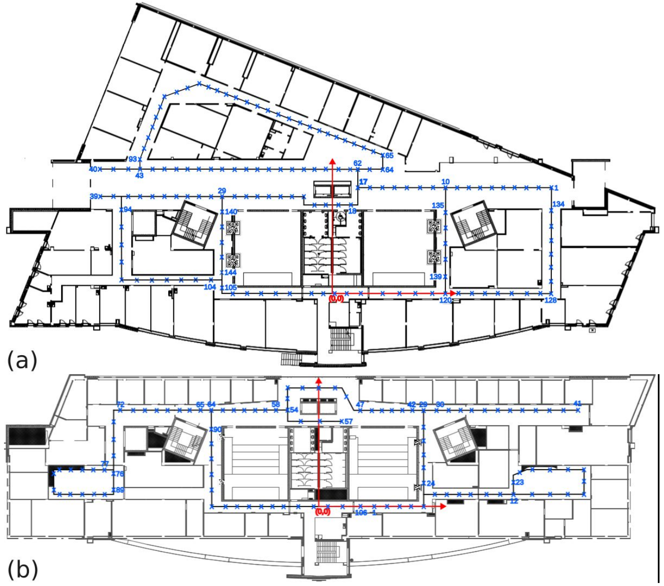
- Appearance-based localization: visual place recognition with omnidirectional images obtained from a catadioptric camera



Place recognition with omnidirectional images

- Diagram of the CNN-based image description blocks that produce embeddings used as global descriptors in the localization system. The global map is built from n_i images converted to embedding vectors d_i that are stored in the map (global descriptors).





Outcome of the lecture

- A brief review of the Visual Place recognition approaches.
- More detailed presentation of the Bag of Visual Words idea and FAB-MAP as a localisation system that uses the BoW concept.
- NetVLAD as an example of end-to-end trainable place recognition system..
- Example of a simple place recognition system that uses embeddings. .

IRIM

POZNAN UNIVERSITY OF TECHNOLOGY