NYU TANDON SCHOOL OF ENGINEERING

AI4CE

Robot Vision

Visual Place Recognition

Dr. Chen Feng

cfeng@nyu.edu

ROB-UY 3203, Spring 2024

Overview

- Overview of Visual Place Recognition
- Bag of Visual Words (BoVW)
- VLAD

References

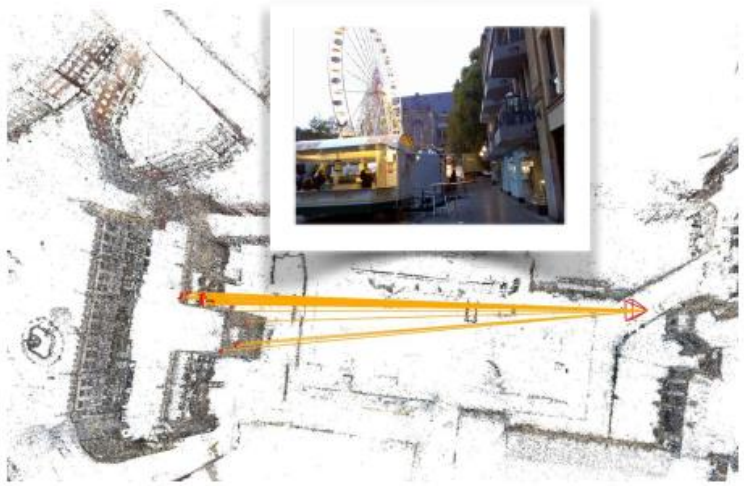
- Szeliski 2011
  - Section 14.4

- (BoVW)Jupyter notebook by Olga Vysotska: https://github.com/ovysotska/in_simple_english/bl ob/master/bag_of_visual_words.ipynb
- (VLAD) Jégou, H., Douze, M., Schmid, C. and Pérez, P., 2010, June. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3304-3311). IEEE.
- (NetVLAD) Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5297-5307).
- Kendall, A., Grimes, M. and Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 2938-2946).

Overview of Visual Place Recognition

- **We predict where the images are captured using database of**
  - Geo-tagged images
    - **Visual place recognition**

  - 3D points and descriptors
    - **Image based localization**



Slides from Sattler et. al., 2015

Keywords Related to Visual Place Recognition

Structure from Motion

Visual SLAM

Image-based localization

Visual place recognition

Image retrieval

Image classification

Slides from Sattler et. al., 2015

Keywords Related to Visual Place Recognition

Structure from Motion

Visual SLAM

Bundle adjustment

Tracking

Geometric verification (RANSAC)

Image-based localization

Loop detection & closure

Visual place recognition

ANN

Feature detection & description

Image description & indexing

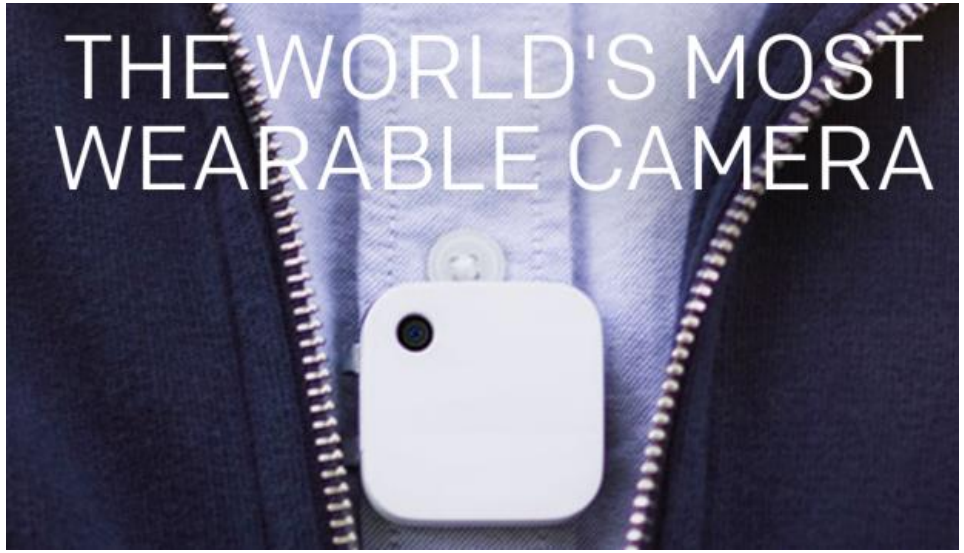
Image retrieval

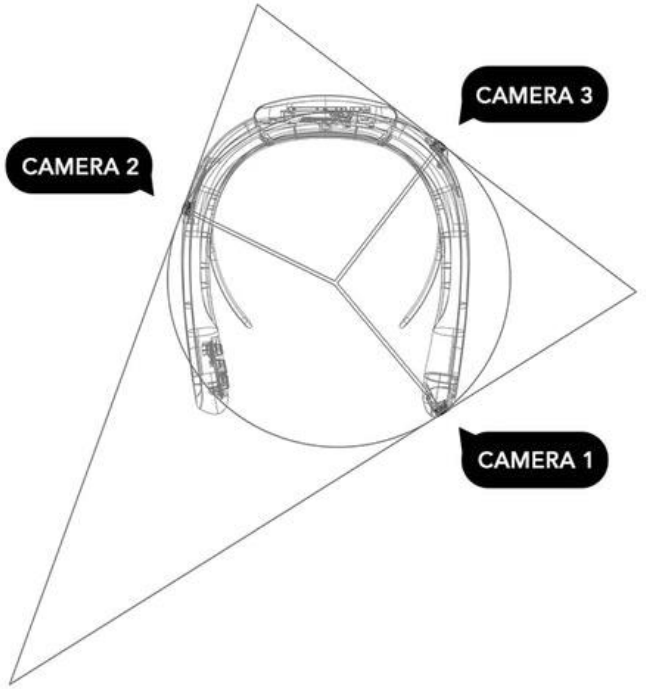
Image classification

Slides from Sattler et. al., 2015

Potential Application of Visual Place Recognition

- Accelerate incremental SfM and SLAM
  - Localization and navigation of robots and cars
- Integrate images to the real-world coordinates
- Acquiring images is becoming easier
  - e.g. Narrative Clip, FITT360





Slides from Sattler et. al., 2015

Street-level Visual Place Recognition

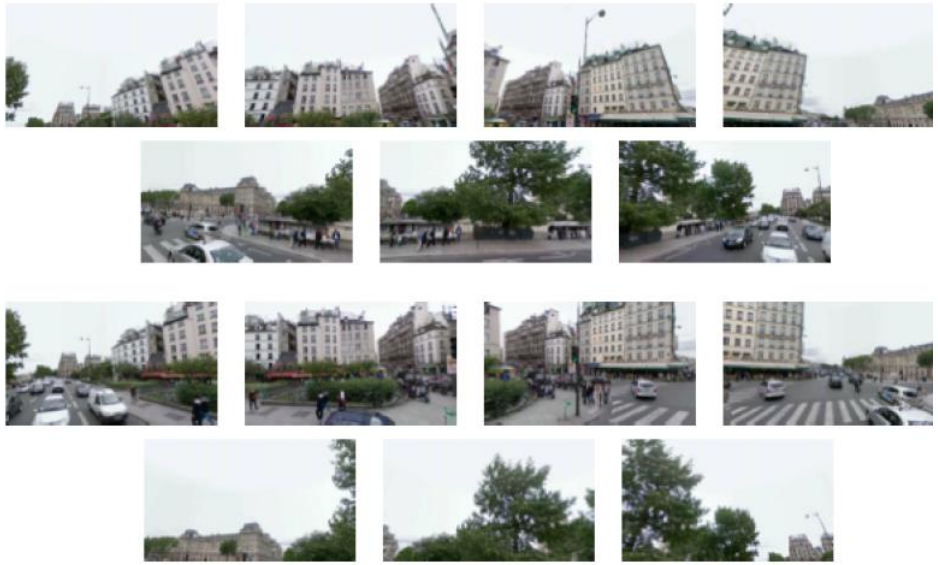
- Given a query image of a particular street or a building, we seek to find one, or at most a few, images in the geotagged database depicting the same place!

Slides from Sattler et. al., 2015

Database of Geo-Tagged Images

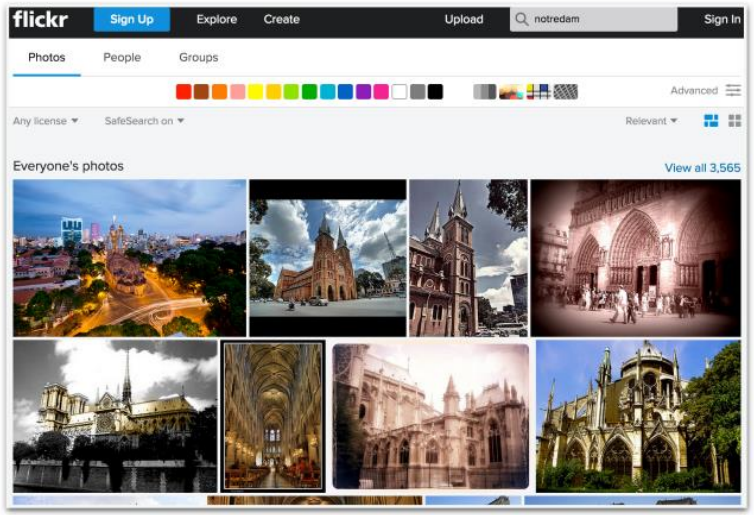
- Typical source of geo-tagged images:
  - Google Streetview (panorama)
  - Driving/walking with a ladybug camera

- We can use original panoramas or generate perspective cutouts



Slides from Sattler et. al., 2015

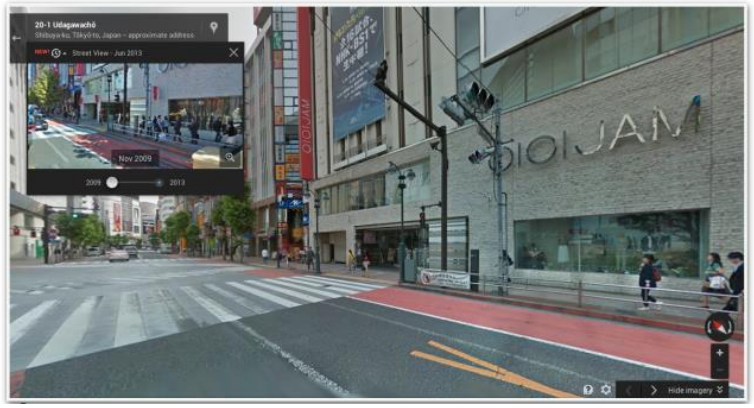
Database of Geo-Tagged Images



Landmarks

Internet Community Photos,
e.g. Flickr, Panoramio
+ Densely captured in different times!
(200M images on Flickr (2012))
- Not really depicting the place :)

See Part 2 for image based localization.
See also [Hays-CVPR'08, Hays-ICCV'09, Quack-CVIR'08].



Street-level

StreetView (panoramic) images
+ Almost all the streets on cities !
- A few images at the same locations
(different times available now!)

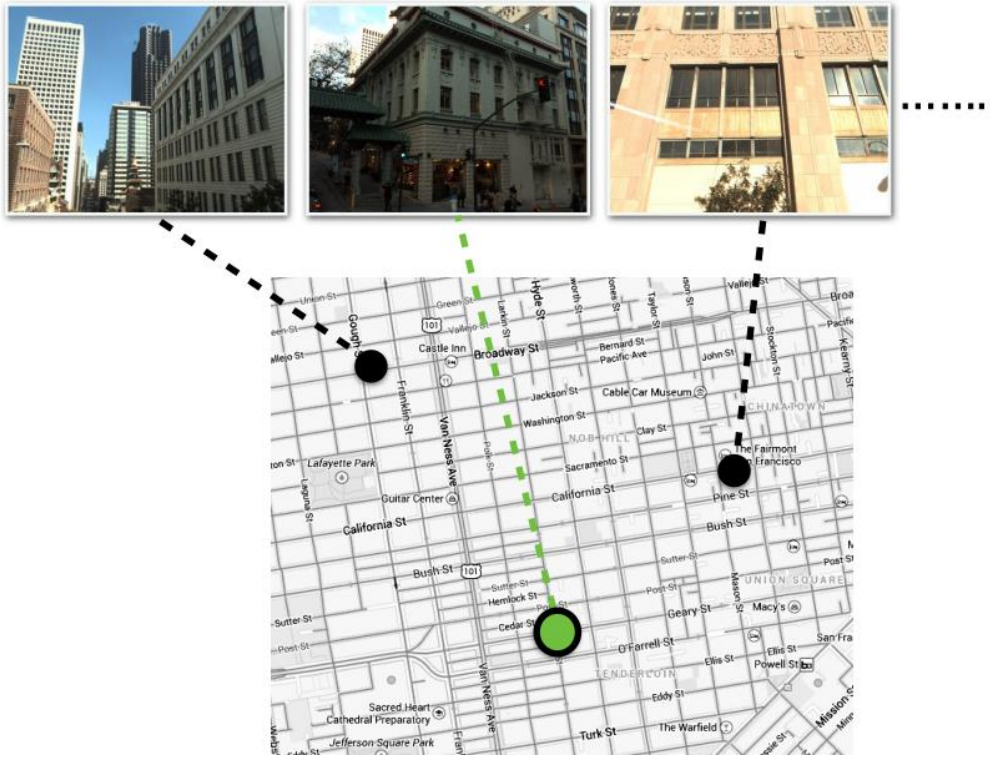
Slides from Sattler et. al., 2015

Visual place recognition by image retrieval

We are interested in "a single query testing image" and "a large image database".

Query image

Database of geo-tagged images



Slides from Sattler et. al., 2015

A standard image retrieval [Philbin-CVPR07]

Query image

Database images (millions, billions)

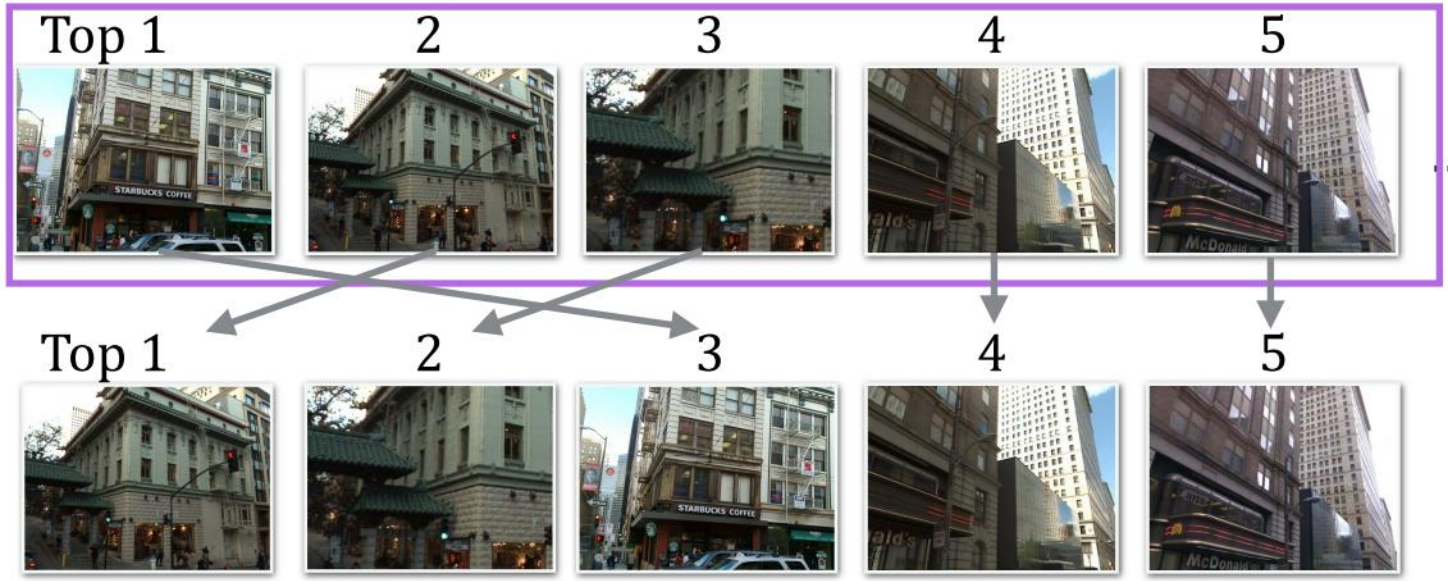
Slides from Sattler et. al., 2015

A standard image retrieval [Philbin-CVPR07]

Query image

Step1: Initial ranking/shortlisting

Step2: Re-ranking to improve the list



Step1 should be fast -> BoVW, VLAD, Fisher vectors

Step2 can be a bit more costly -> geometric verification

Slides from Sattler et. al., 2015

Visual Place Recognition Pipeline

Query image

Database images



.....

Offline

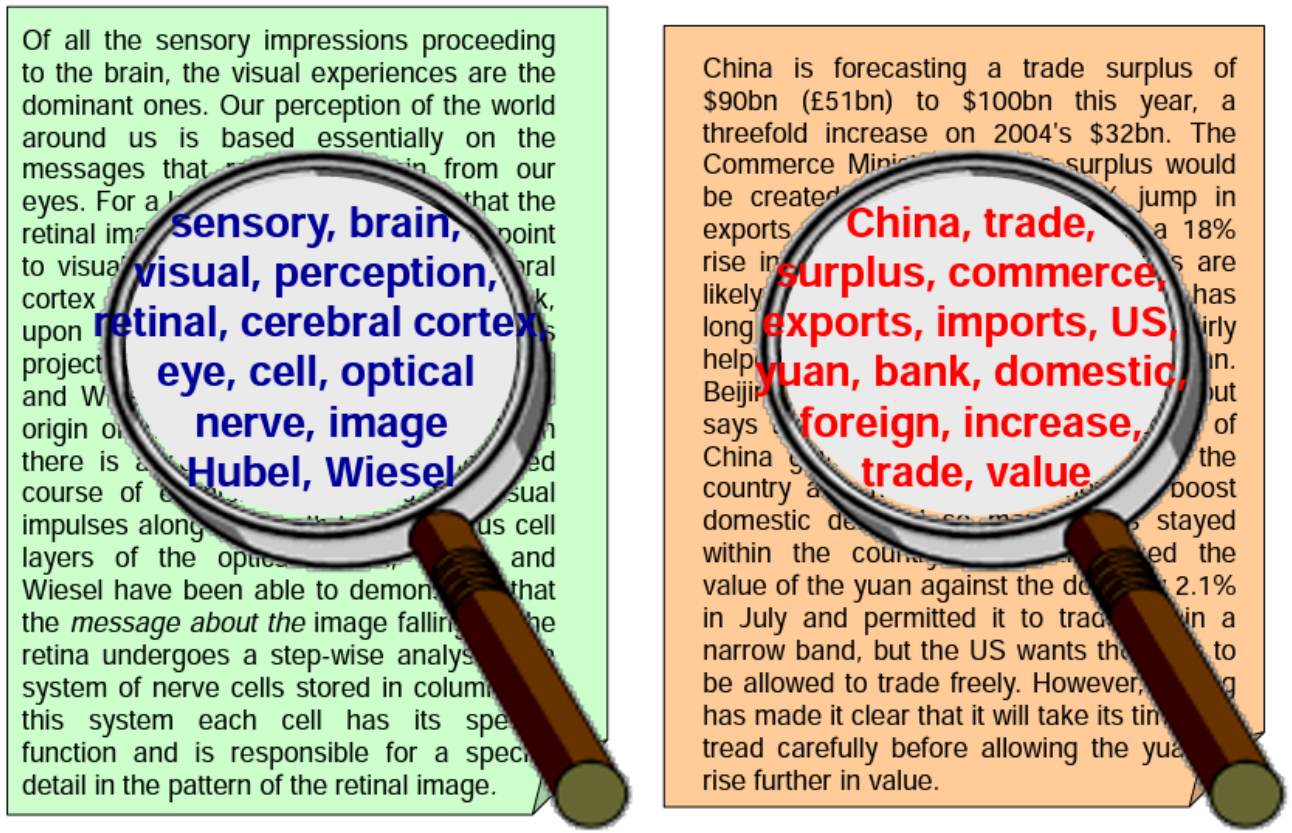
1. Feature detection & description
2. Training visual vocabulary
3. Image description

4. Feature detection & description
5. Image description

6. Initial ranking
7. Re-ranking with geometric verification

Slides from Sattler et. al., 2015

Bag-of-Visual-Words(BoVW)

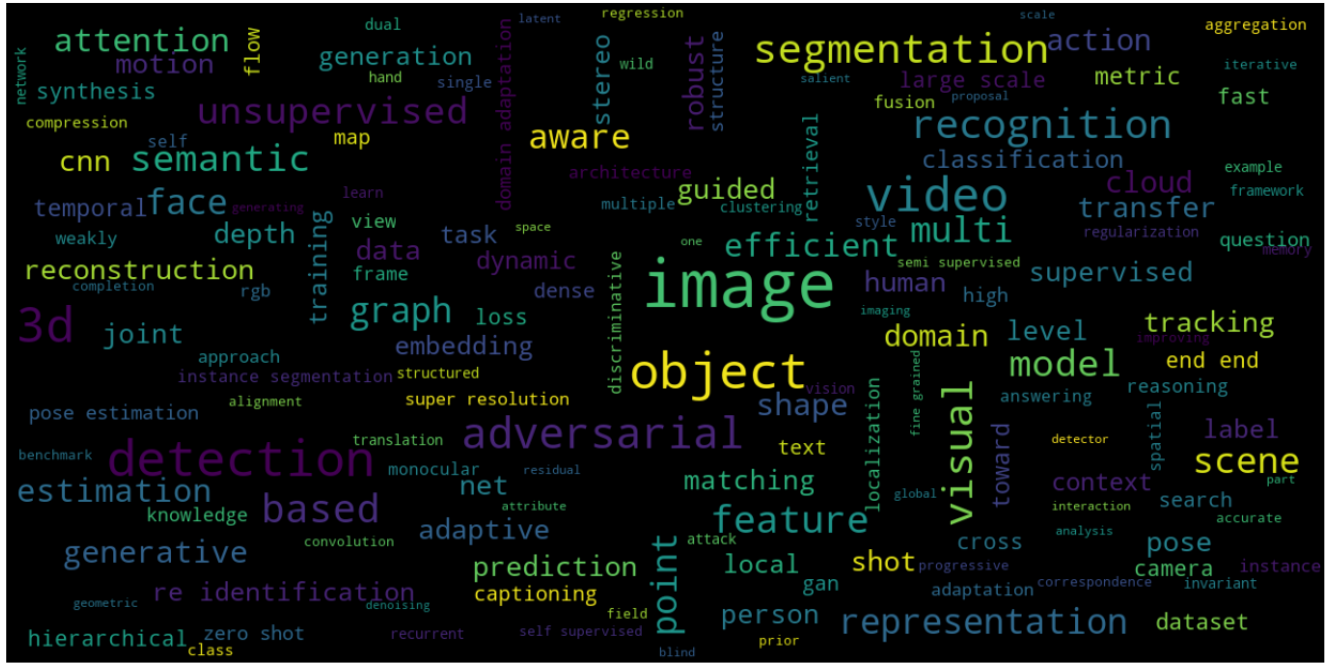


Analogy to Text Documents

Slides adapted from FeiFei Li

BoVW Origin: Bag-of-Words Model

- Orderless document representation: frequencies of words from a dictionary (Salton&McGill(1983))

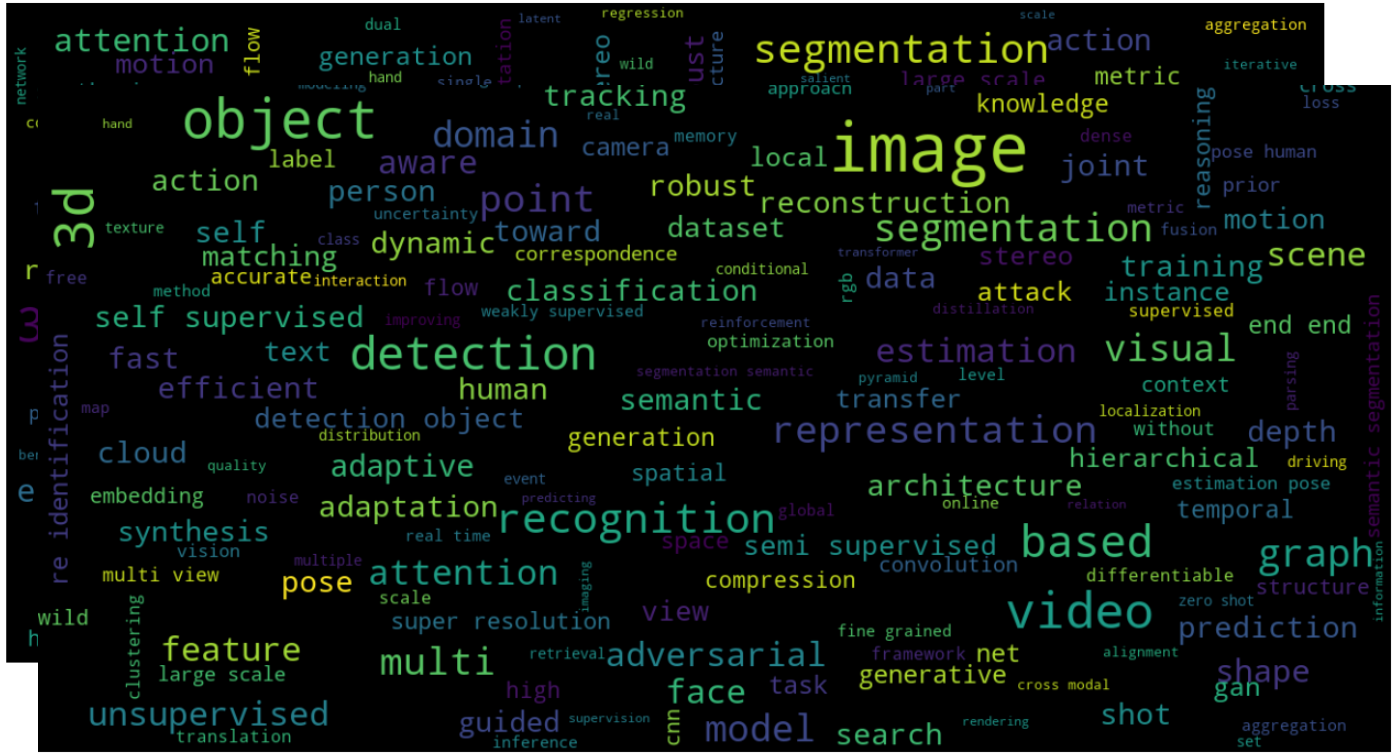


CVPR - 2019

https://github.com/hoya012/CVPR-2021-Paper-Statistics

BoVW Origin: Bag-of-Words Model

- Orderless document representation: frequencies of words from a dictionary (Salton&McGill(1983))

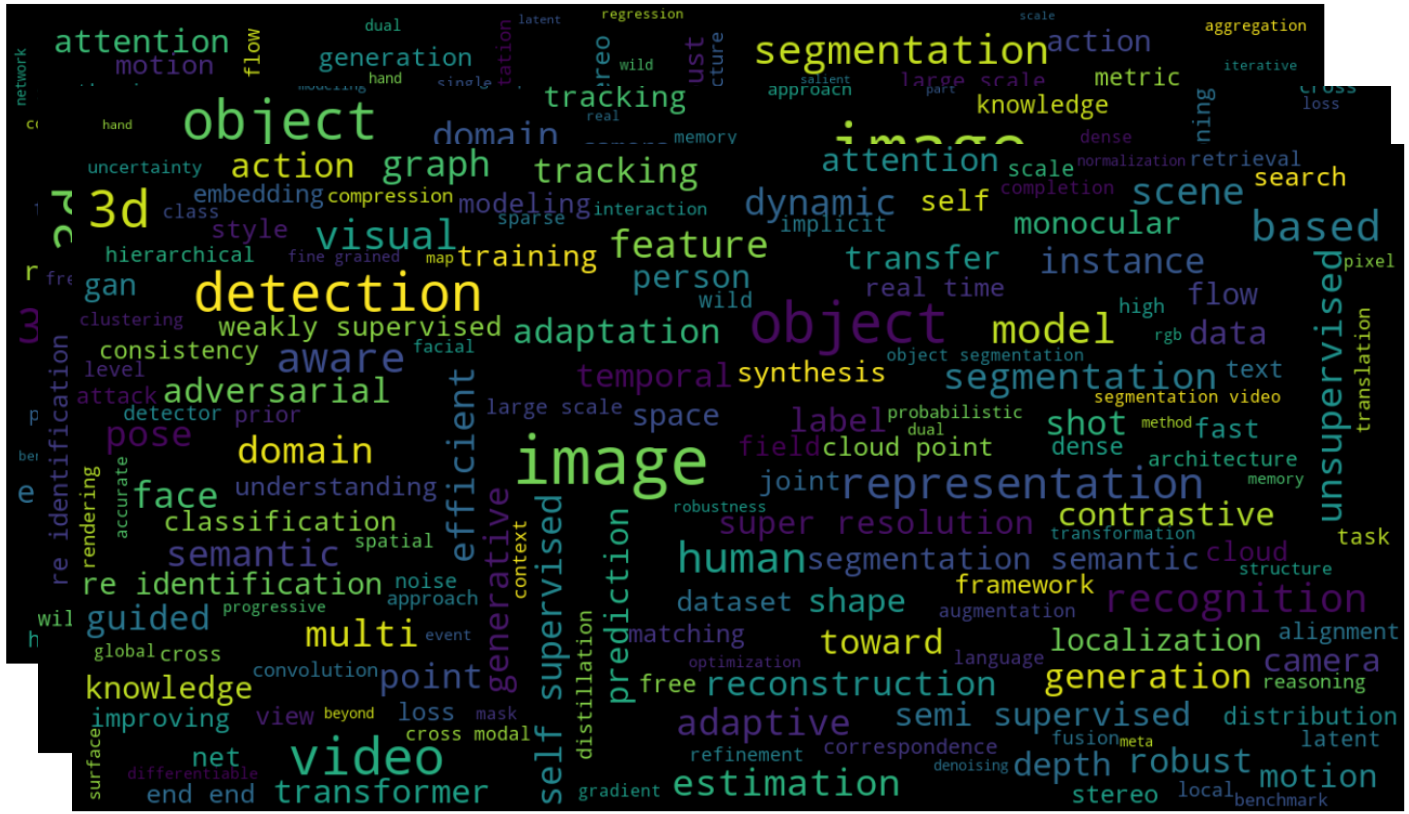


CVPR - 2020

https://github.com/hoya012/CVPR-2021-Paper-Statistics

BoVW Origin: Bag-of-Words Model

- Orderless document representation: frequencies of words from a dictionary (Salton&McGill(1983))



CVPR - 2021

https://github.com/hoya012/CVPR-2021-Paper-Statistics

BoVW Pipeline

1. Extract features (e.g., SIFT)
2. Learn “visual dictionary” (e.g., K-Means Clustering)
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”

Image

Bag of ”Visual Words”

Slides adapted from Cyrill Stachniss

BoVW Pipeline: Visual Overview

1. Extract features
2. Learn “visual dictionary” (Offline)
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”

Slides adapted from Cyrill Stachniss

BoVW Pipeline: Visual Overview

1. **Extract features**
2. Learn “visual dictionary” (Offline)
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”

Slides adapted from Cyrill Stachniss

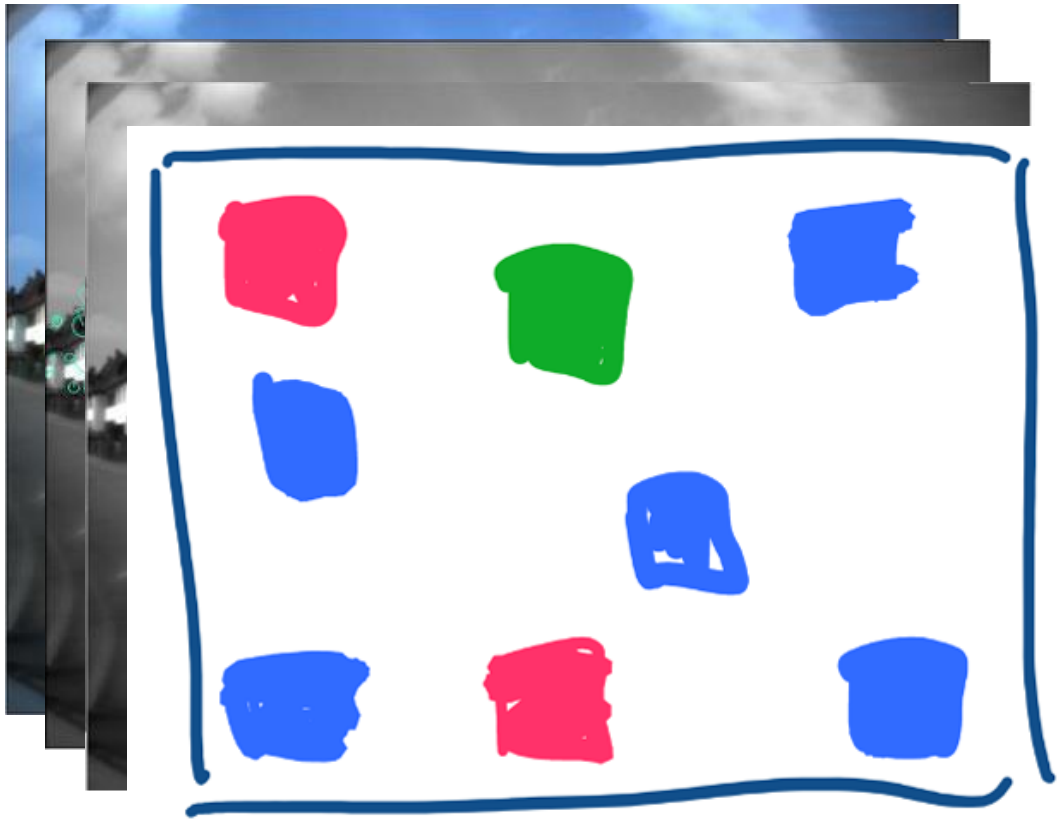
BoVW Pipeline: Visual Overview

1. Extract features
2. Learn “visual dictionary”
3. **Quantize features using visual vocabulary**
4. Represent images by frequencies of “visual words”

Slides adapted from Cyrill Stachniss

BoVW Pipeline: Visual Overview

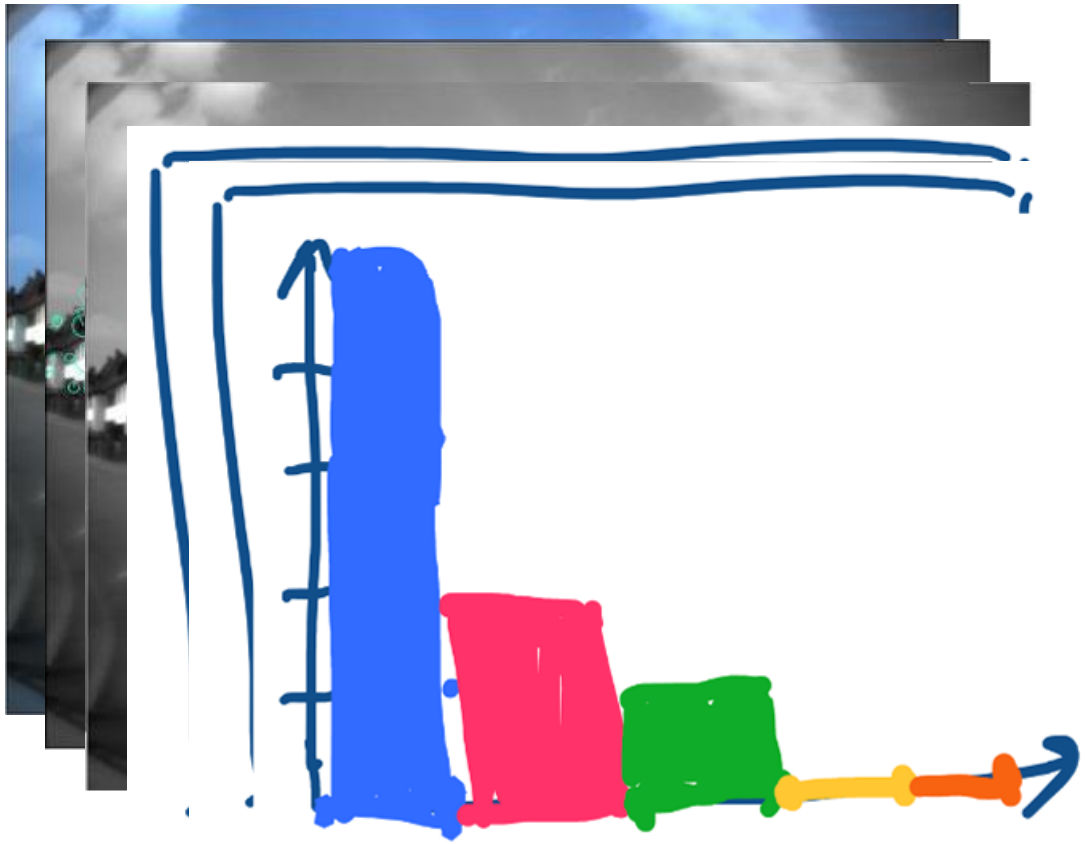
1. Extract features
2. Learn “visual dictionary”
3. **Quantize features using visual vocabulary**
4. Represent images by frequencies of “visual words”



Slides adapted from Cyrill Stachniss

BoVW Pipeline: Visual Overview

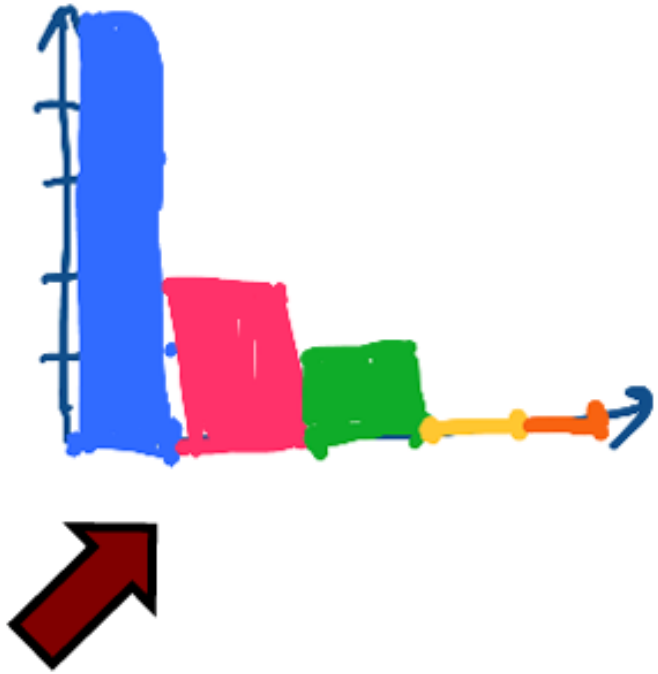
1. Extract features
2. Learn “visual dictionary”
3. Quantize features using visual vocabulary
4. **Represent images by frequencies of “visual words”**



Slides adapted from Cyrill Stachniss

BoVW: Dictionary

- A dictionary defines the list of words that are considered
- The dictionary defines the x-axes of all the word occurrence histograms
- The dictionary must remain fixed

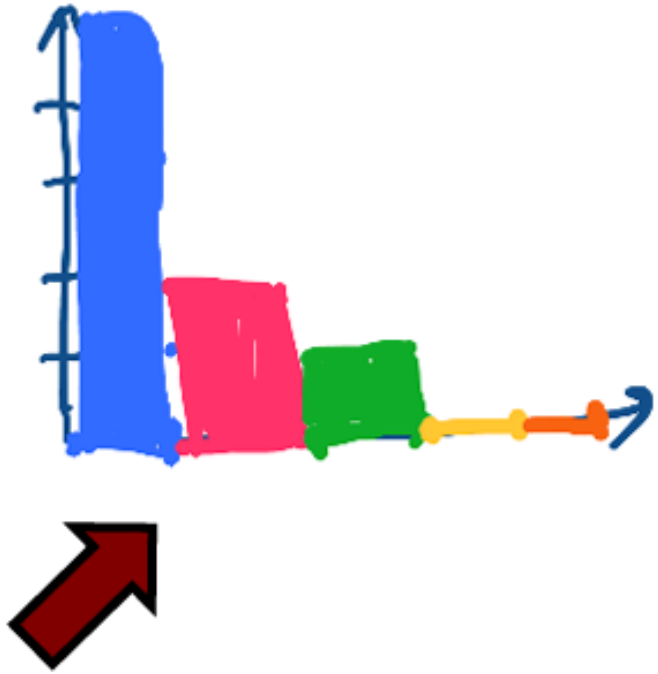


Slides adapted from Cyrill Stachniss

BoVW: Dictionary

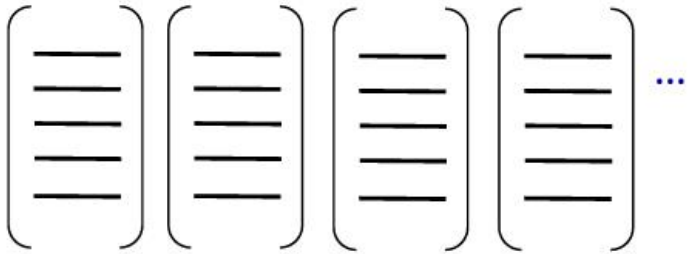
- A dictionary defines the list of words that are considered
- The dictionary defines the x-axes of all the word occurrence histograms
- The dictionary must remain fixed

The dictionary is typically learned from data. How can we do that?

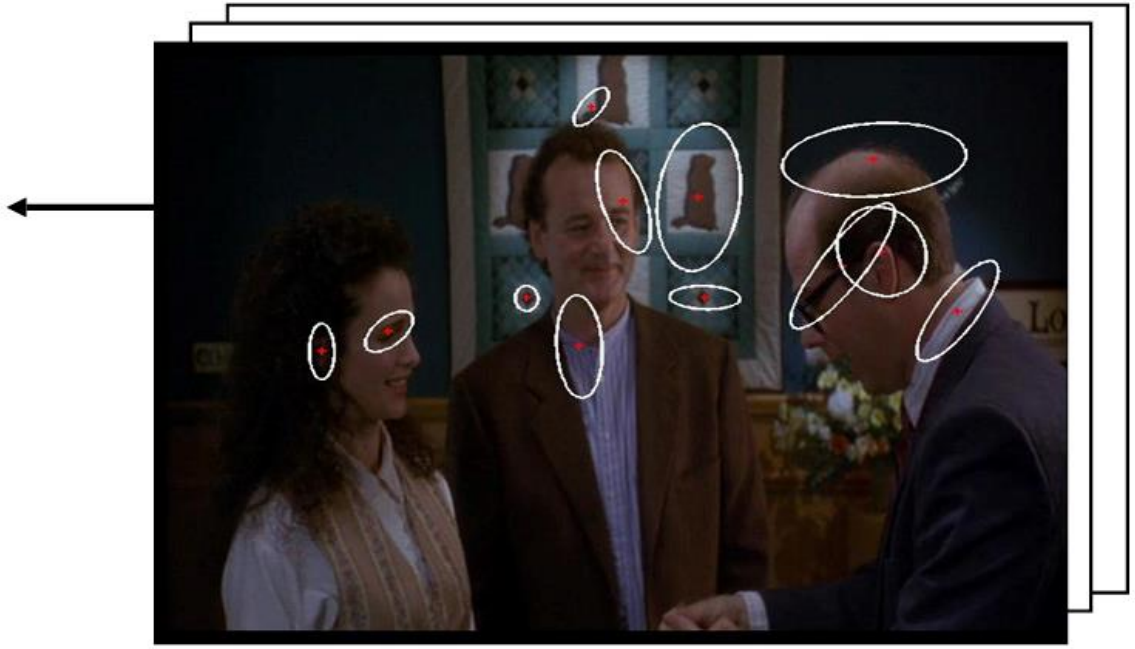


Slides adapted from Cyrill Stachniss

BoVW: Learn Visual Dictionary

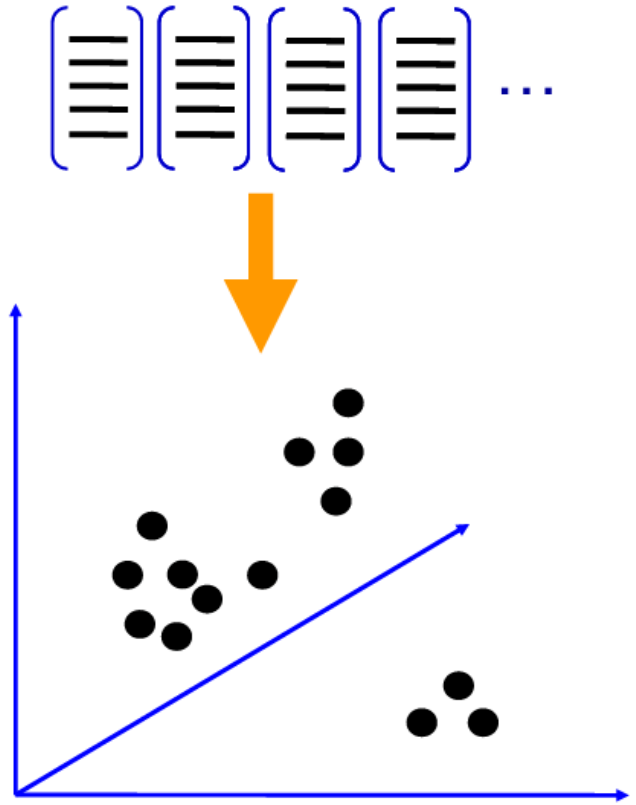


Compute SIFT Descriptor



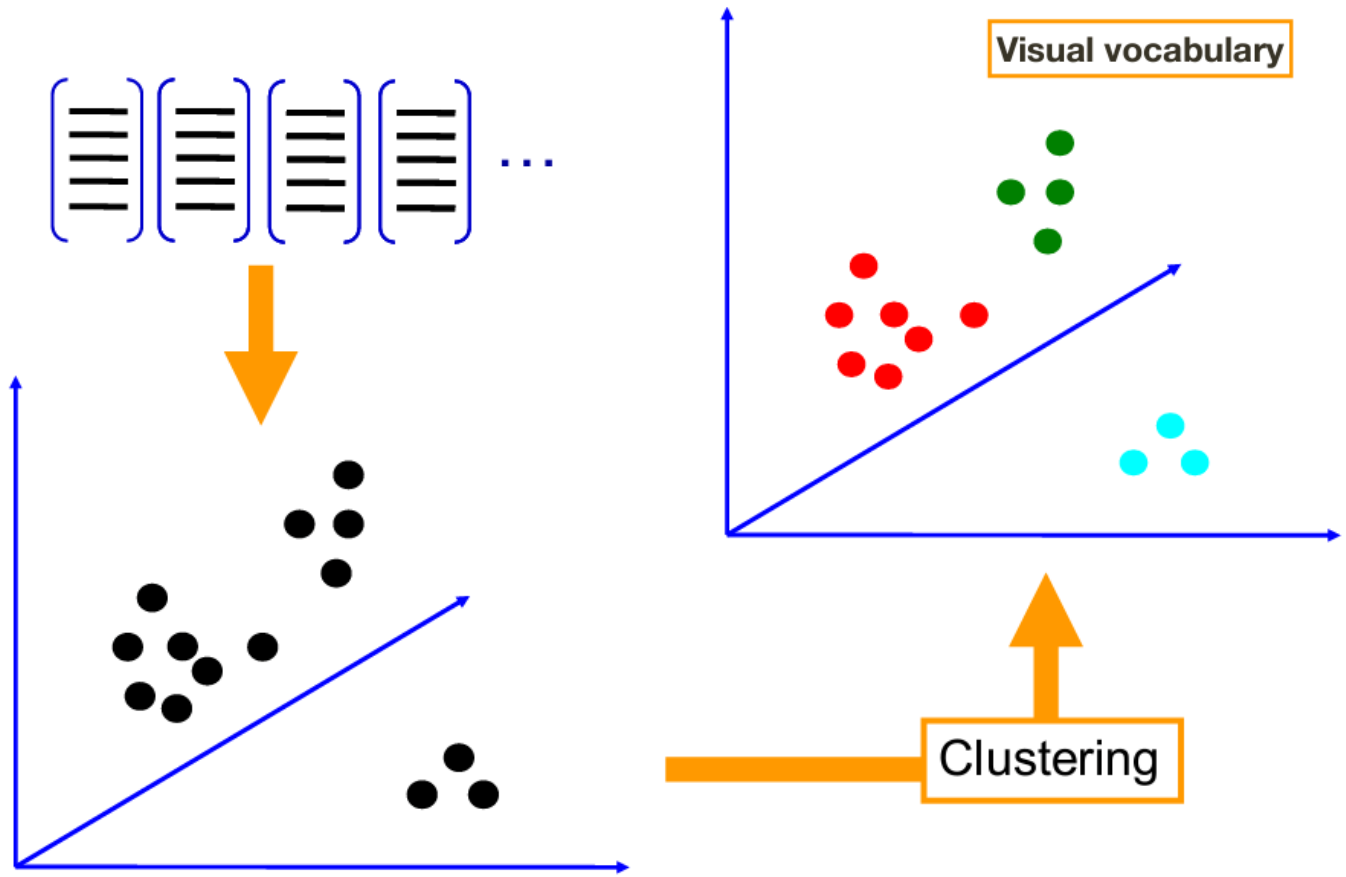
Slides adapted from Ioannis (Yannis) Gkioulekas (CMU)

BoVW: Learn Visual Dictionary



Slides adapted from Ioannis (Yannis) Gkioulekas (CMU)

BoVW: Learn Visual Dictionary



Slides adapted from Ioannis (Yannis) Gkioulekas (CMU)

BoVW: K-Means Cluster Recap

- Want to minimize sum of squared Euclidean distances between points x_i and their nearest cluster centers m_k

- Algorithm
  - Randomly initialize K cluster centers
  - Iterate until convergence:
    - Assign each data point to the nearest center
    - Recompute each cluster center as the mean of all points assigned to it

$$D(X, M) = \sum_{\text{cluster } k} \sum_{\substack{\text{point } i \text{ in} \\ \text{cluster } k}} (x_i - m_k)^2$$

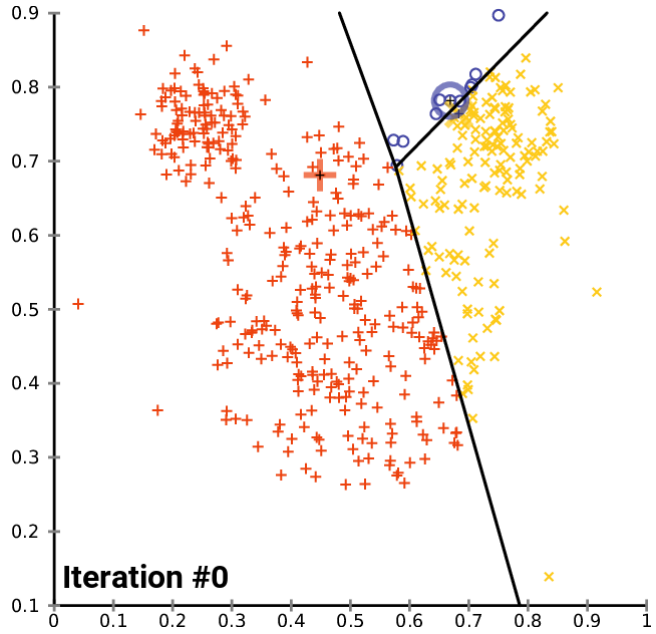


Image Credit: Wikipedia

Slides adapted from Ioannis (Yannis) Gkioulekas (CMU)

BoVW: K-Means Cluster Recap

- Want to minimize sum of squared Euclidean distances between points x_i and their nearest cluster centers m_k

- Algorithm
  - Randomly initialize K cluster centers
  - Iterate until convergence:
    - Assign each data point to the nearest center
    - Recompute each cluster center as the mean of all points assigned to it

$$D(X, M) = \sum_{\text{cluster } k} \sum_{\substack{\text{point } i \text{ in} \\ \text{cluster } k}} (x_i - m_k)^2$$

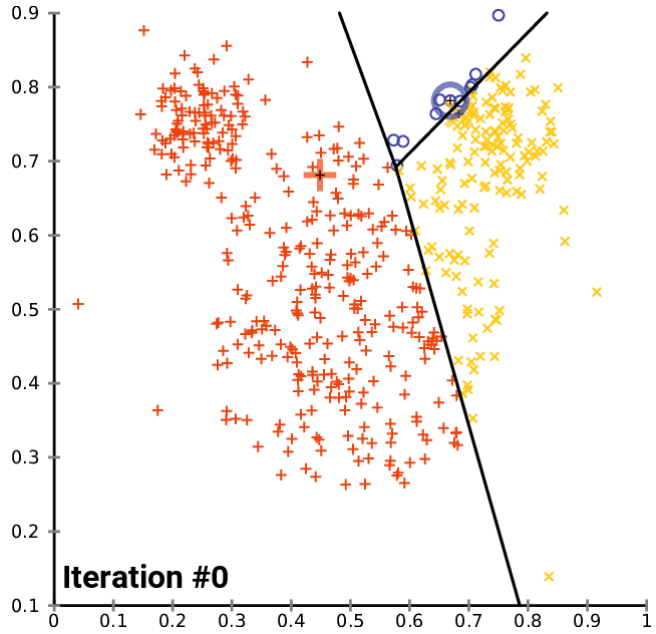
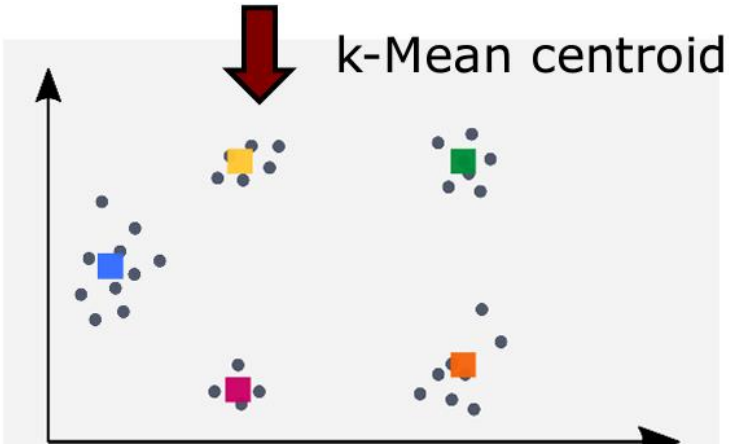


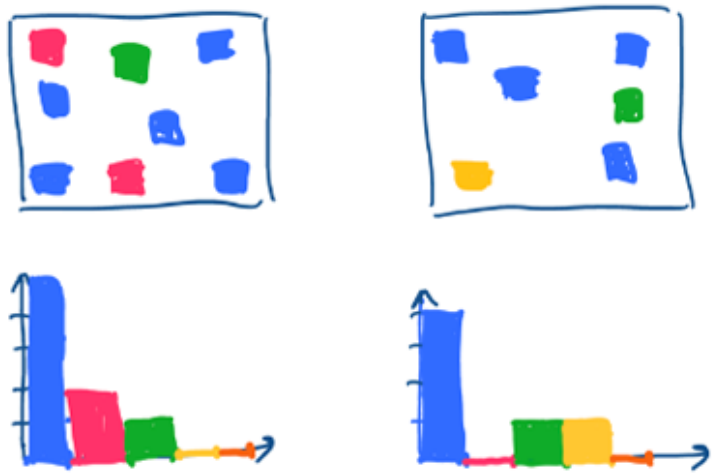
Image Credit: Wikipedia

We use k-means to compute the dictionary of visual words.

Slides adapted from Ioannis (Yannis) Gkioulekas (CMU)

BoVW: Represent Images as Visual Words

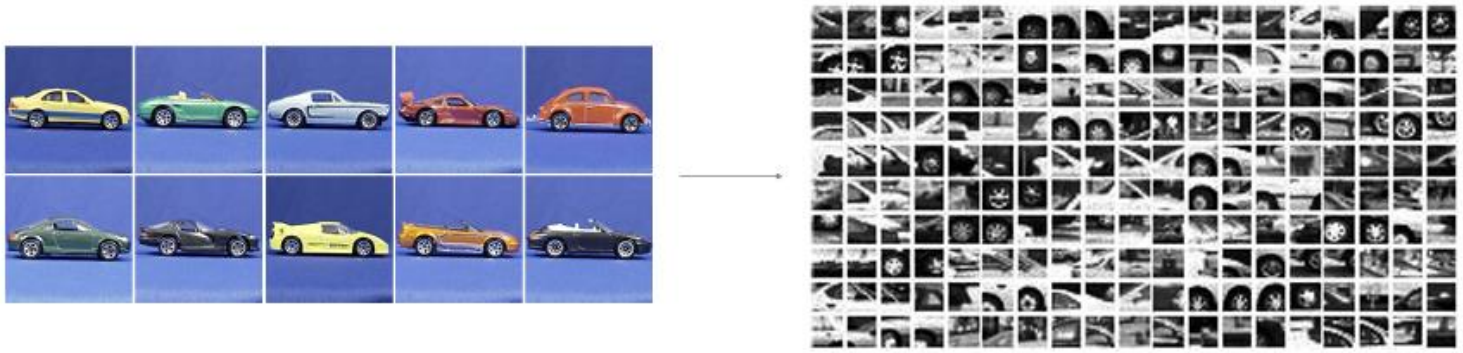


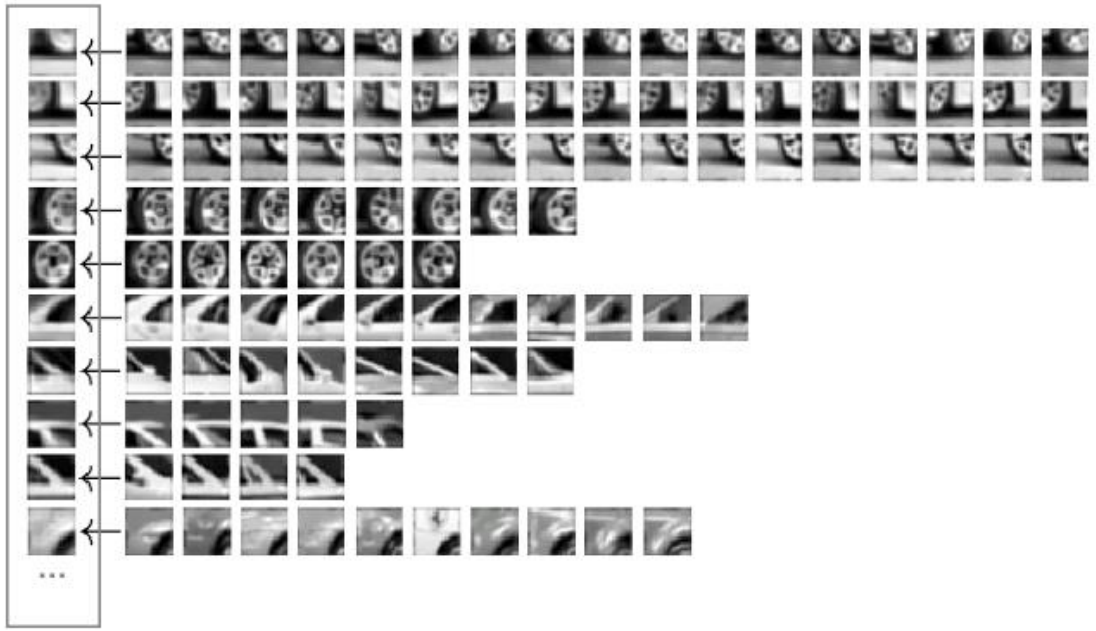


Every image turns into a histogram!

Slides adapted from Ioannis (Yannis) Gkioulekas (CMU)

BoVW: Visual Dictionary Example





Slides adapted from Ioannis (Yannis) Gkioulekas (CMU) Image Credit: B. Leibe

BoVW: Visual Dictionary Example

Slides adapted from Ioannis (Yannis) Gkioulekas (CMU) Image Credit: B. Leibe

BoVW: Summary

- Compact summary of the image content
- Largely invariant to viewpoint changes and deformations
- Ignores the spatial arrangement
- Unclear how to choose optimal size of the dictionary
  - Too small: Words not representative of all image regions
  - Too large: Over-fitting

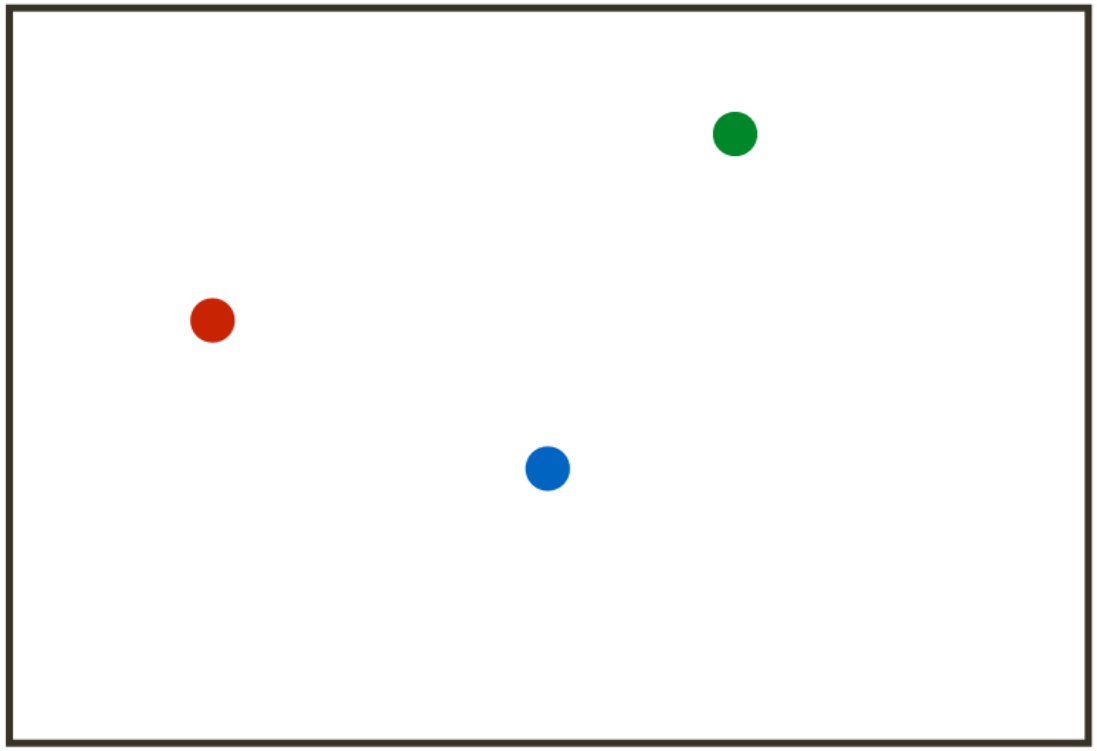
Slides adapted from Cyrill Stachniss

VLAD (Vector of Locally Aggregated Descriptors)

- There are more advanced ways to ‘count’ visual words than incrementing its histogram bin
- For example, it might be useful to describe how local descriptors are quantized to their visual words
- In the VLAD representation, instead of incrementing the histogram bin by one, we increment it by the residual vector $x - c(x)$

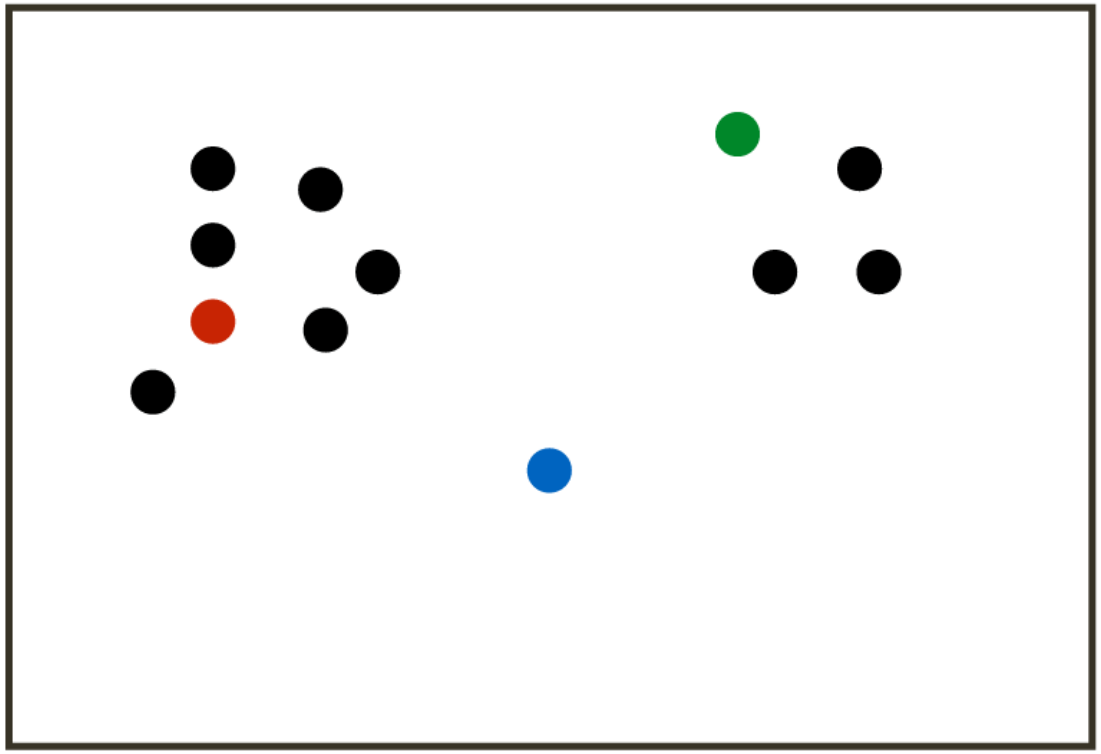
Slides adopted from Ioannis (Yannis) Gkioulekas (CMU)

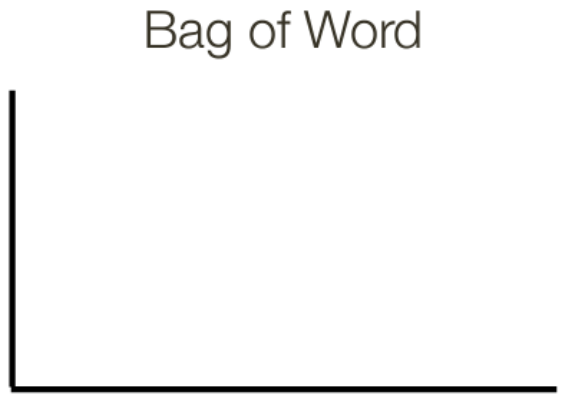
VLAD: Example



Slides adopted from Ioannis (Yannis) Gkioulekas (CMU)

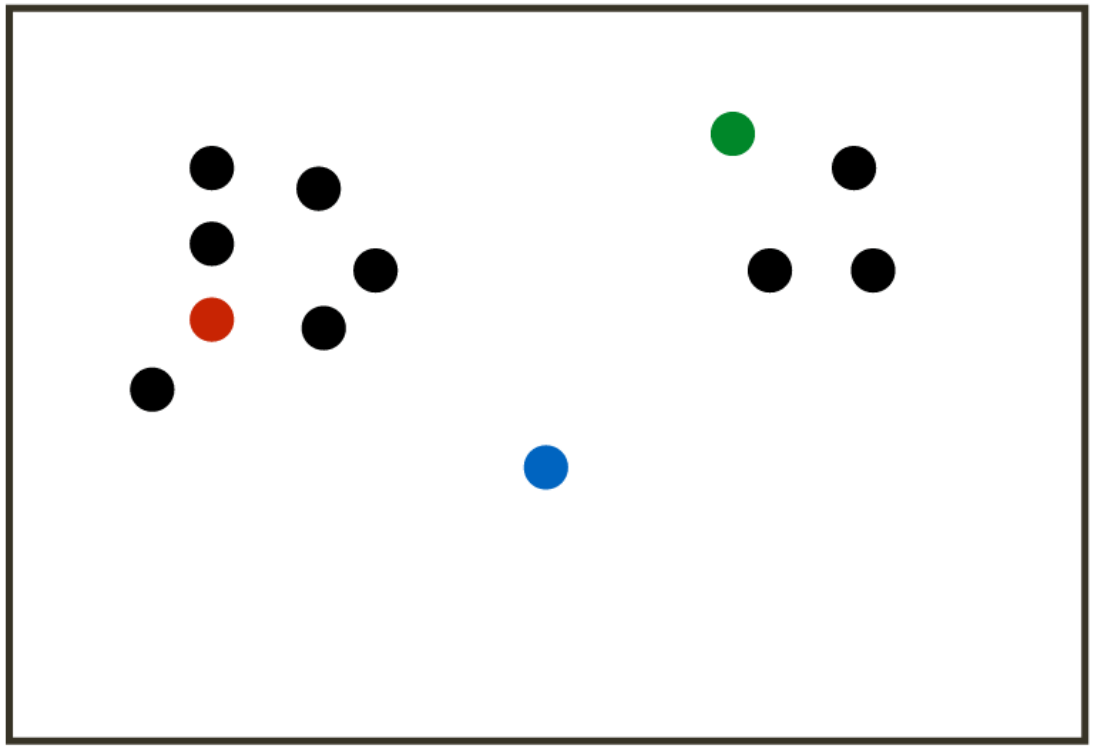
VLAD: Example

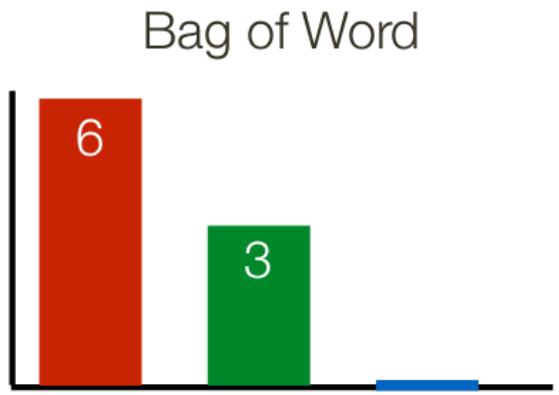




Slides adopted from Ioannis (Yannis) Gkioulekas (CMU)

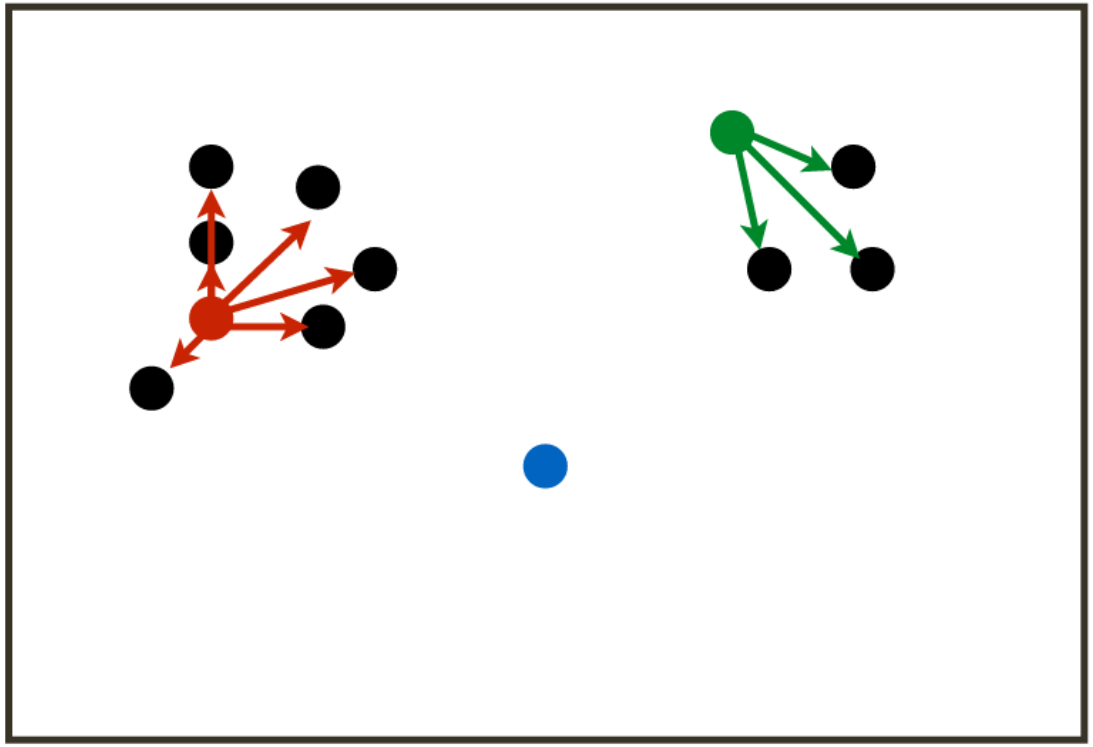
VLAD: Example

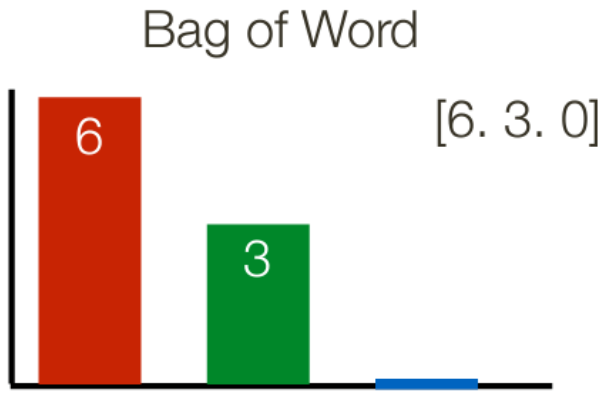




Slides adopted from Ioannis (Yannis) Gkioulekas (CMU)

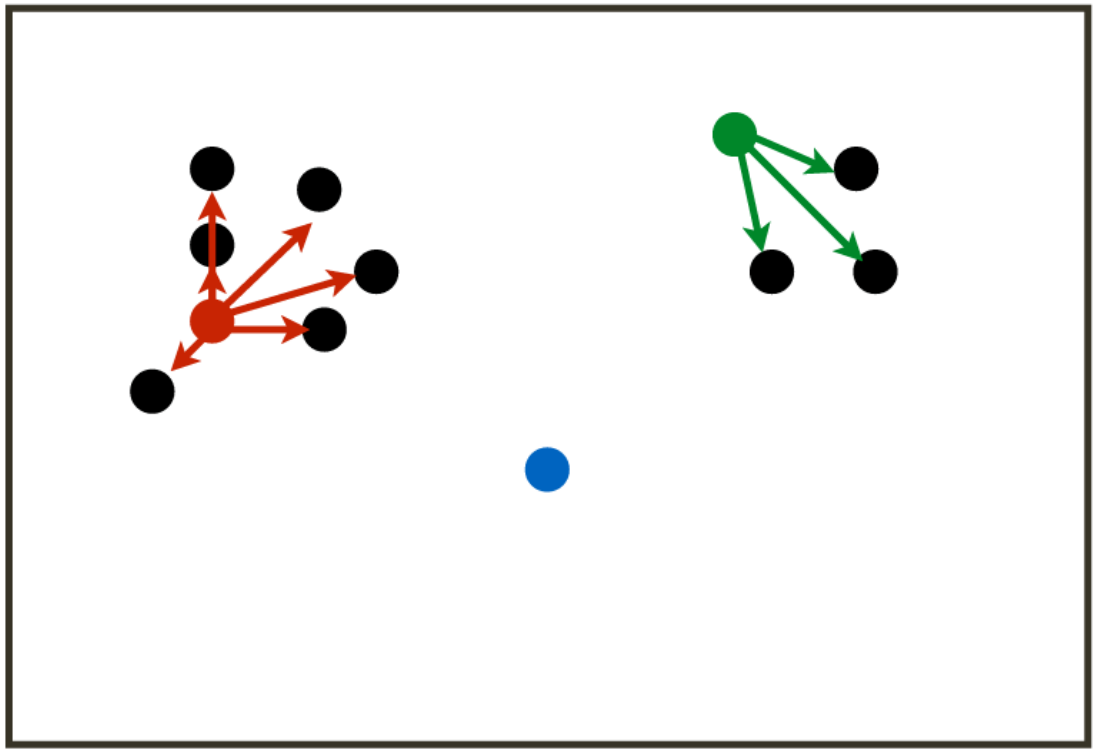
VLAD: Example

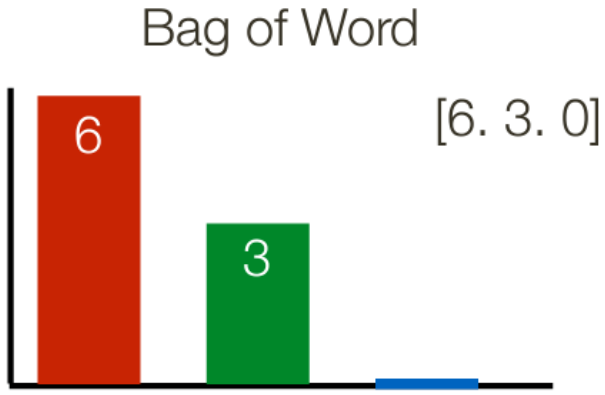




Slides adopted from Ioannis (Yannis) Gkioulekas (CMU)

VLAD: Example





Slides adopted from Ioannis (Yannis) Gkioulekas (CMU)

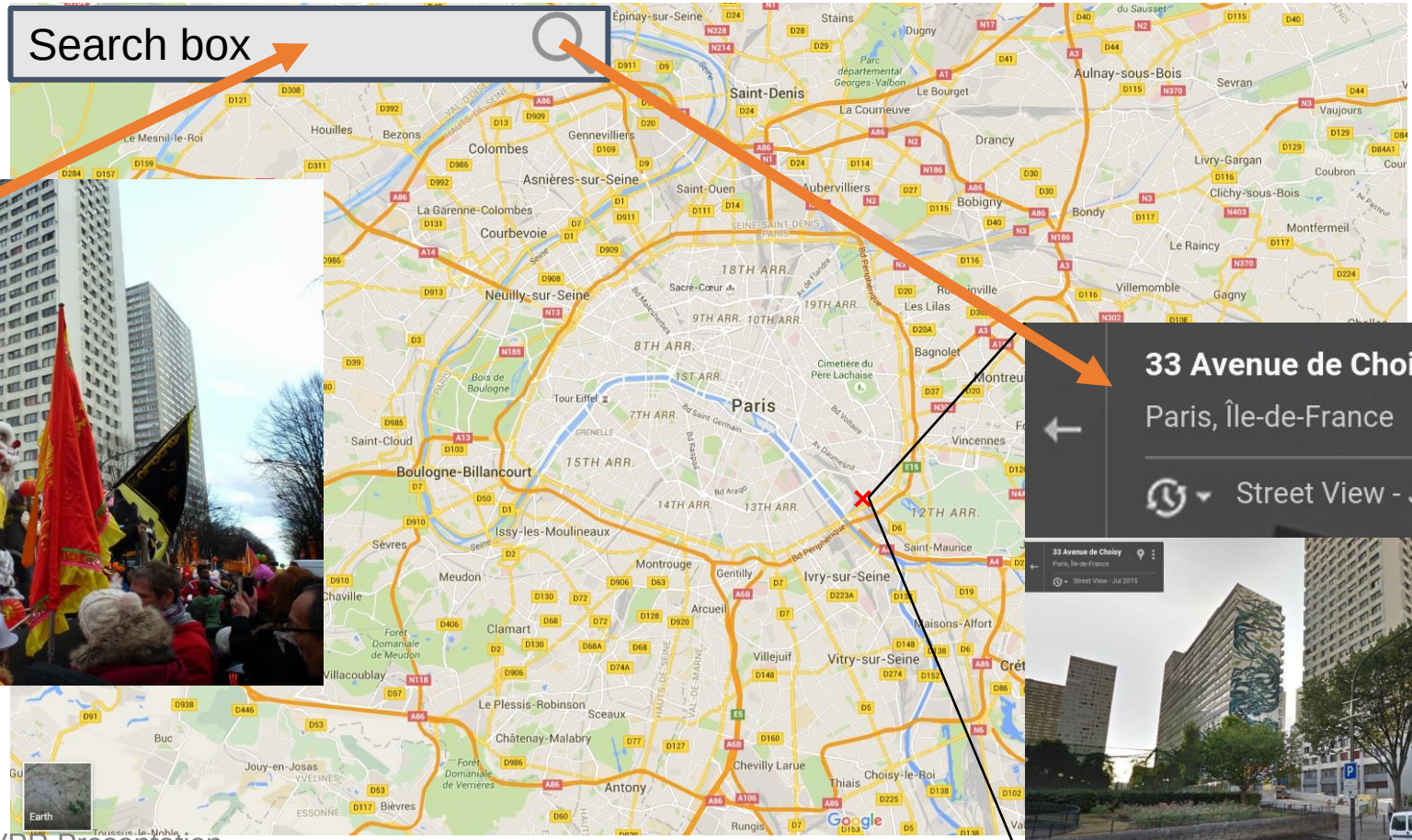
BoVW	VLAD (FV)
Sparse (with a large vocab.)	Dense
▸ Inverted file indexing	▸ ANN, product quantization
▸ Size per image ~= #features x 8 byte (4 bytes for the index and 4 bytes for the weight) E.g. 4000 x 8 = 32K byte	▸ Size per image ~= #centroids x desc dim. x 4 byte (4 bytes (single precision) per dimension} E.g. 256 x 128 x 4 = 65K byte
+ Can provide matches	+ No extra memory requirement to use densely detected features

Slides from Sattler et. al., 2015

NetVLAD: CNN architecture for weakly supervised place recognition

- Goal: Visual Place Recognition

Query Image



Slides adopted from NetVLAD CVPR Presentation

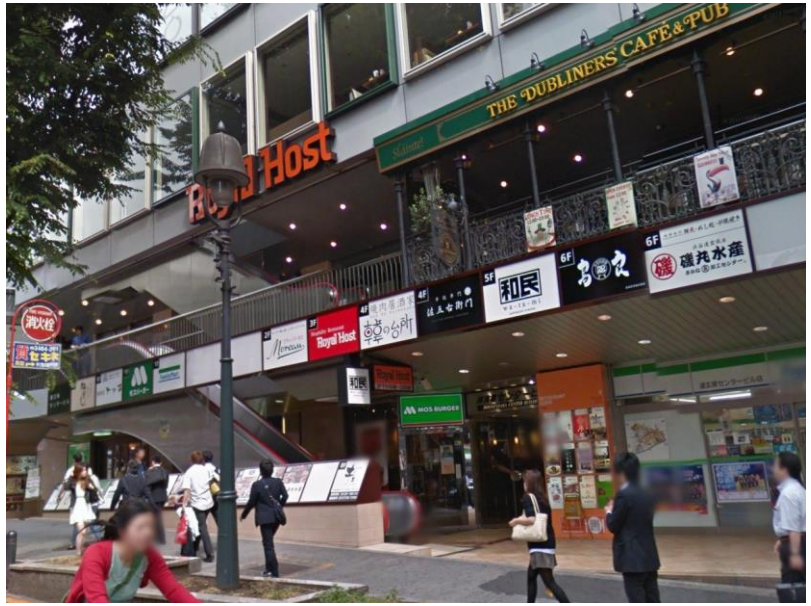
NetVLAD: Difficulty in VPR

- **Lighting changes: Different time of day / year**
- Changes in camera viewpoint
- Occluders and ambiguous objects: People, cars, trees, pavement...
- Big data: World-scale localization

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Difficulty in VPR

- Lighting changes: Different time of day / year
- **Changes in camera viewpoint**
- Occluders and ambiguous objects: People, cars, trees, pavement...
- Big data: World-scale localization



Slides adopted from NetVLAD CVPR Presentation

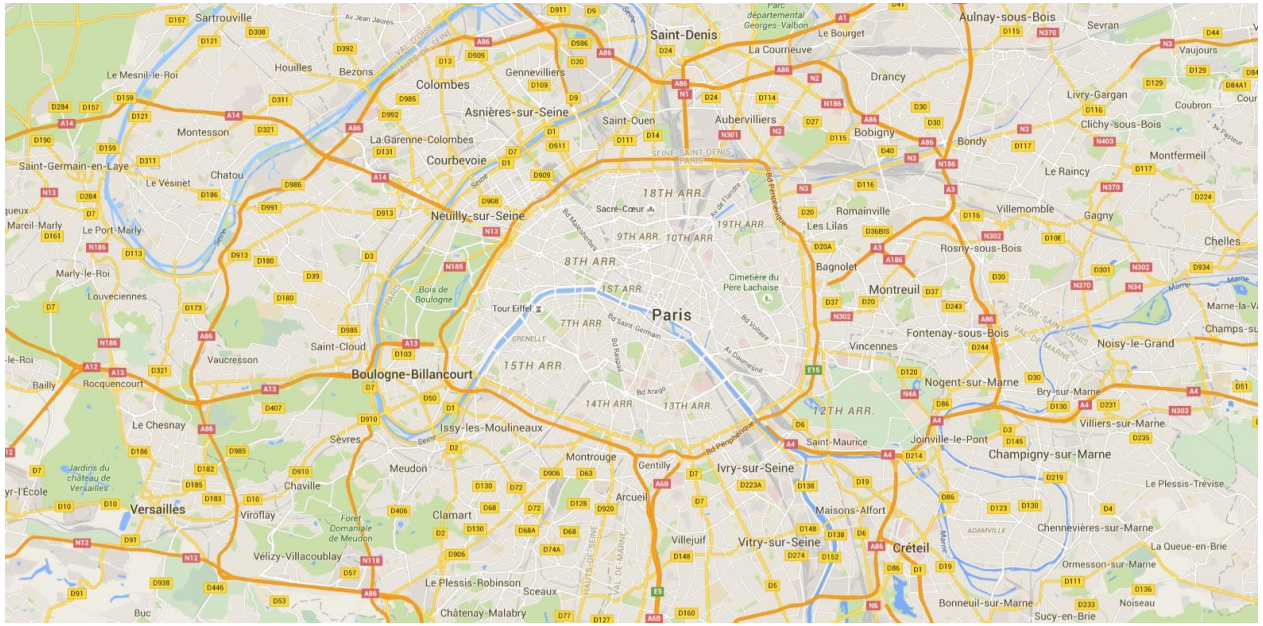
NetVLAD: Difficulty in VPR

- Lighting changes: Different time of day / year
- Changes in camera viewpoint
- **Occluders and ambiguous objects: People, cars, trees, pavement...**
- Big data: World-scale localization

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Difficulty in VPR

- Lighting changes: Different time of day / year
- Changes in camera viewpoint
- Occluders and ambiguous objects: People, cars, trees, pavement...
- **Big data: World-scale localization**



Slides adopted from NetVLAD CVPR Presentation

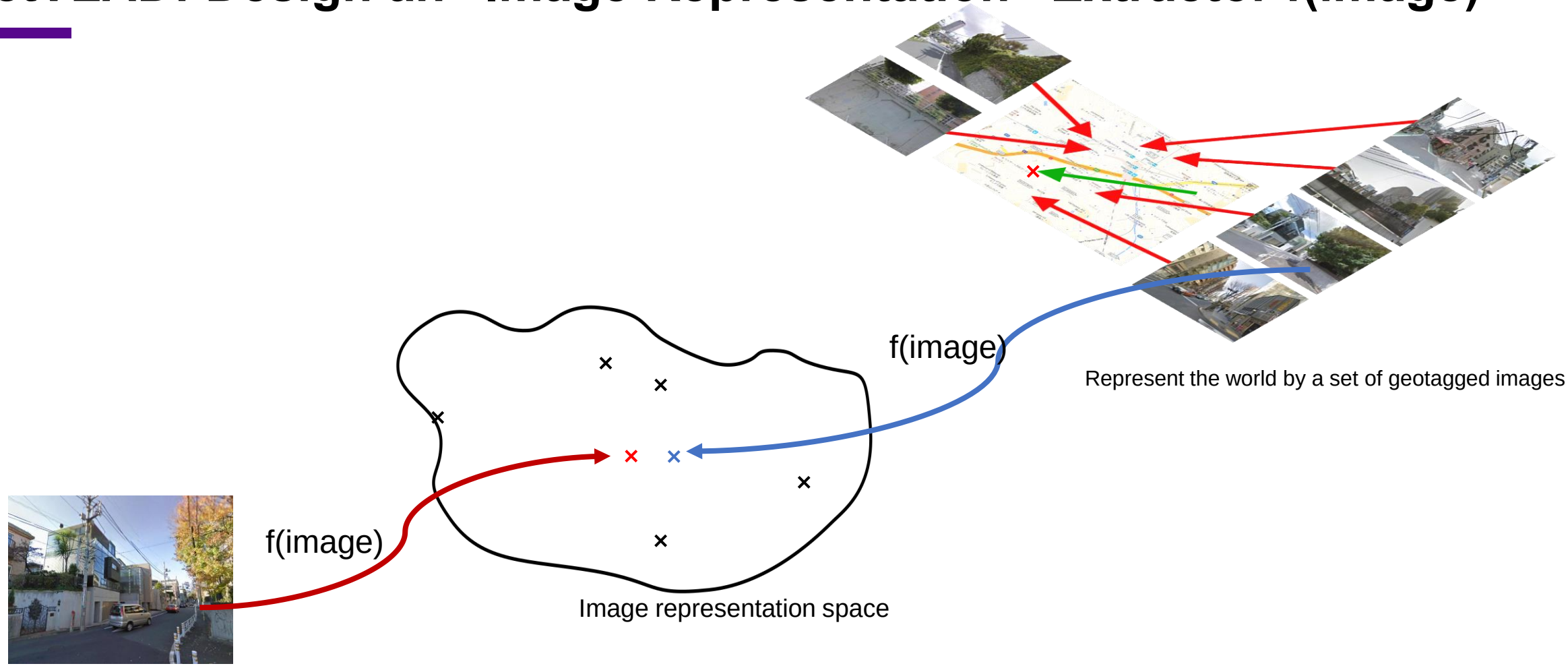
NetVLAD: Design an “Image Representation” Extractor f(image)

Represent the world by a set of geotagged images

Query

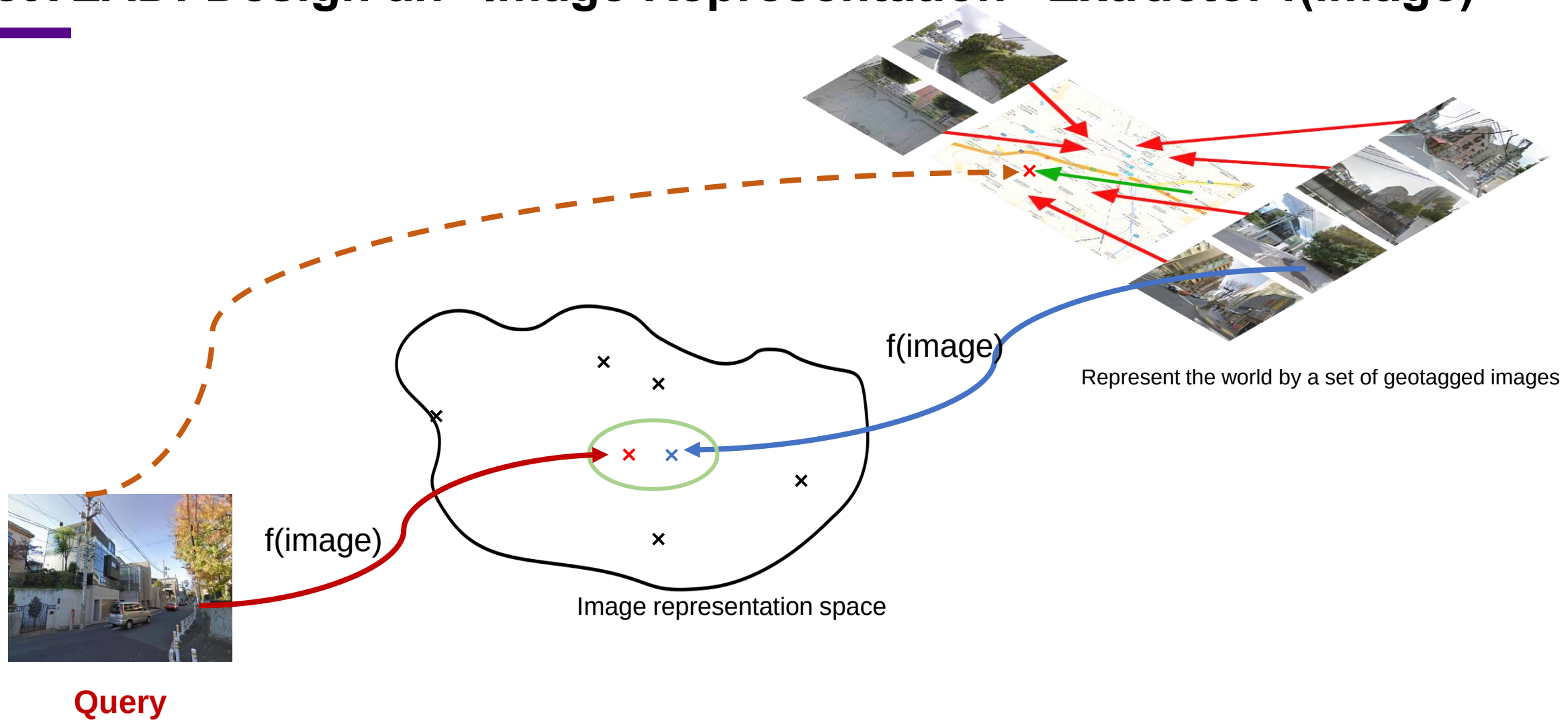
Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Design an “Image Representation” Extractor f(image)



Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Design an "Image Representation" Extractor f(image)



Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Can we apply CNNs to place recognition?

Questions:

1. (Model) What is a good CNN architecture?
2. (Data) How to get the lots of annotated training data?
3. (Loss) What is the appropriate loss for end-to-end training?

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Can we apply CNNs to place recognition?

Questions:

1. **(Model) What is a good CNN architecture?**
2. (Data) How to get the lots of annotated training data?
3. (Loss) What is the appropriate loss for end-to-end training?

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Model

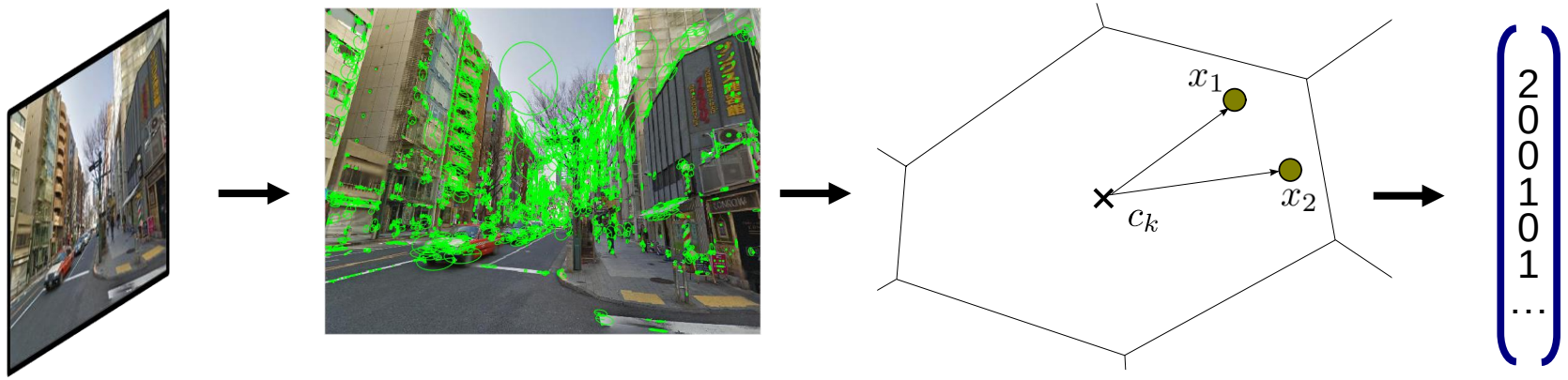


Image I

Extract local features (SIFT)

Aggregate (BoW, **VLAD**, FV)

f(I)

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Model

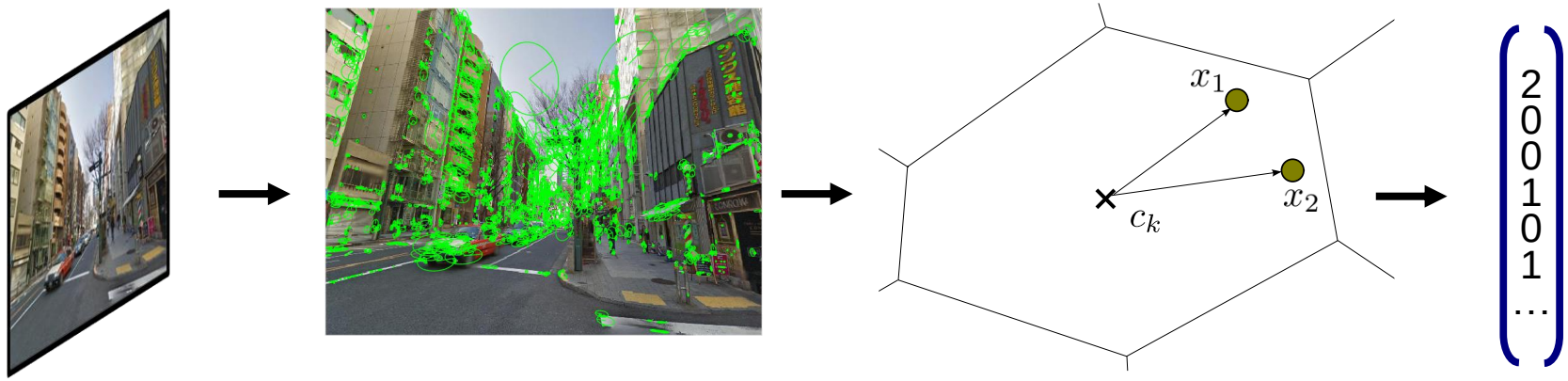
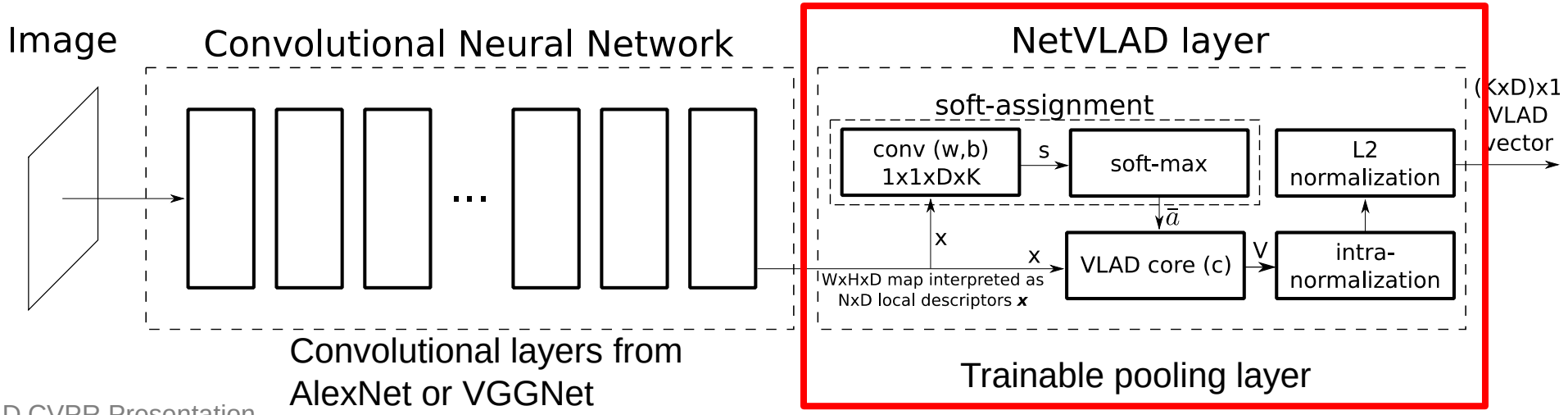


Image I → Extract local features (SIFT) → Aggregate (BoW, **VLAD**, FV) → f(I)



But make it trainable end-to-end

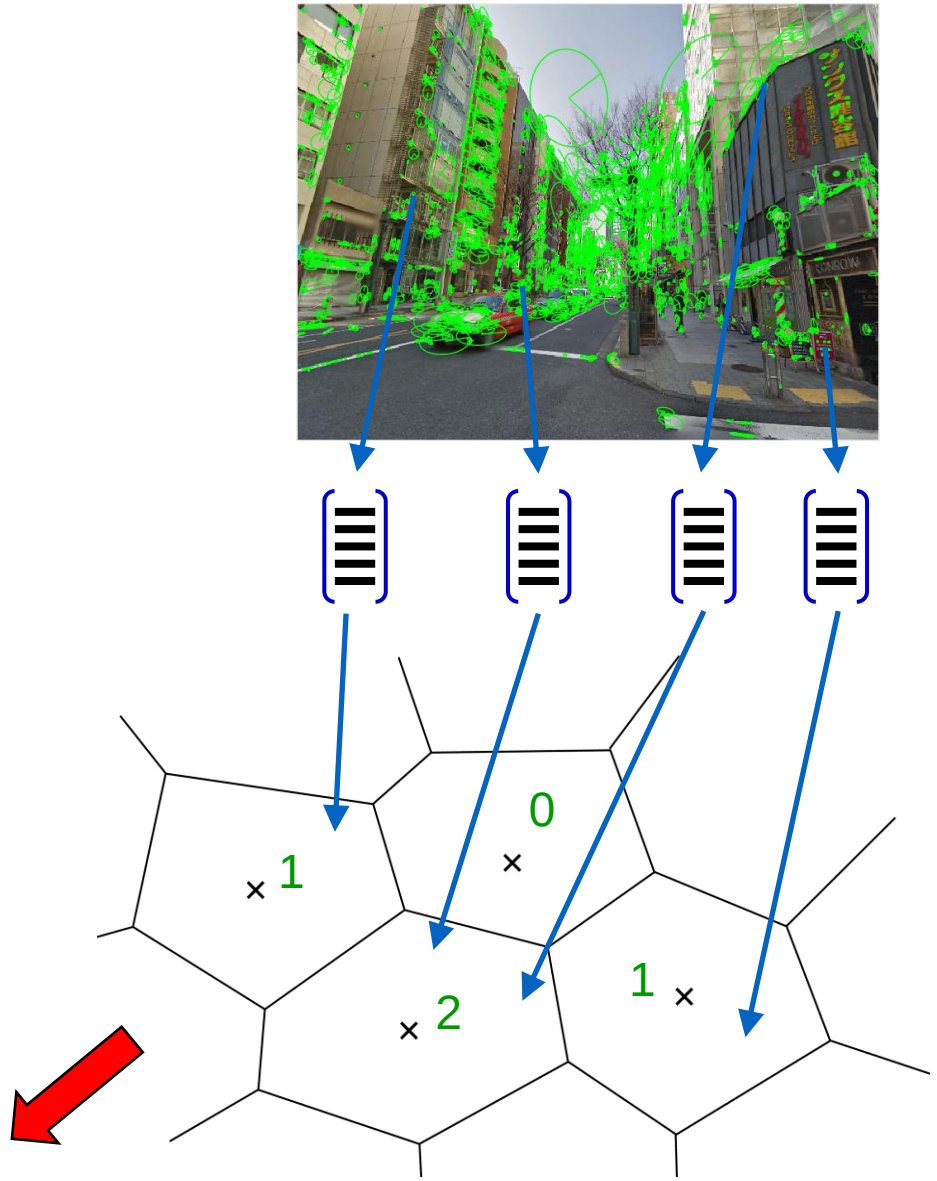
Slides adopted from NetVLAD CVPR Presentation

NetVLAD: BoVW Recap

0/1 assignment of descriptor *i* to cluster *k*

$$B(k) = \sum_{i=1}^{N} a_k(x_i)$$

Sum over all N descriptors in the image



B = [1, 0, 2, 1, …]

Slides adopted from NetVLAD CVPR Presentation

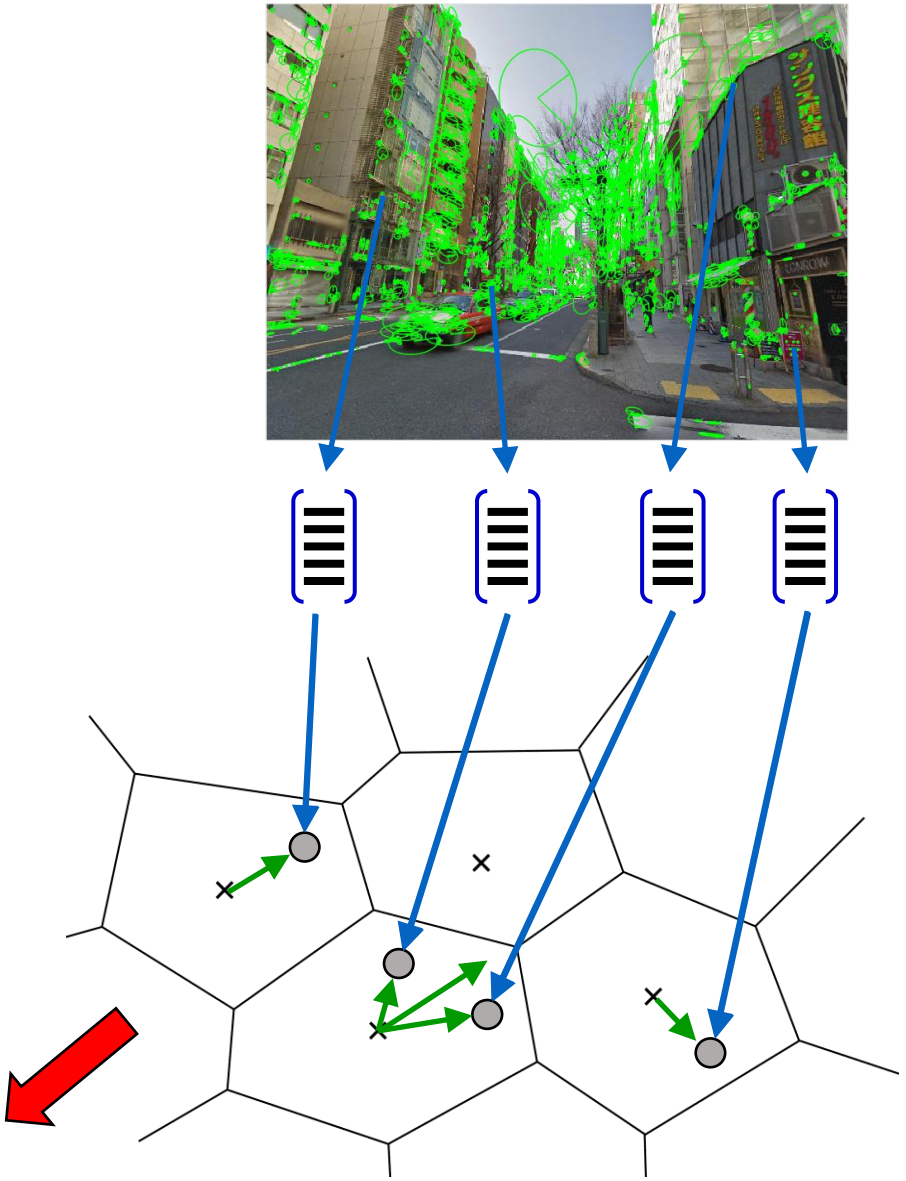
NetVLAD: VLAD Recap

0/1 assignment of descriptor *i* to cluster *k*

$$V(:,k) = \sum_{i=1}^{N} a_k(x_i)(x_i - c_k)$$

Residual vector

Sum over all N descriptors in the image



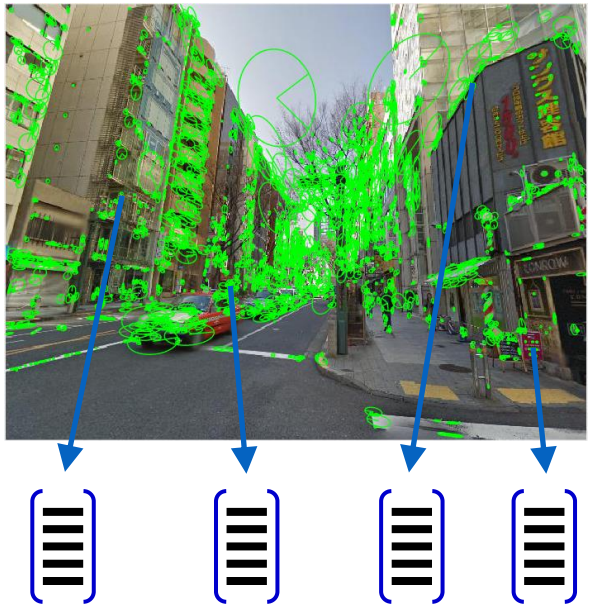
Slides adopted from NetVLAD CVPR Presentation

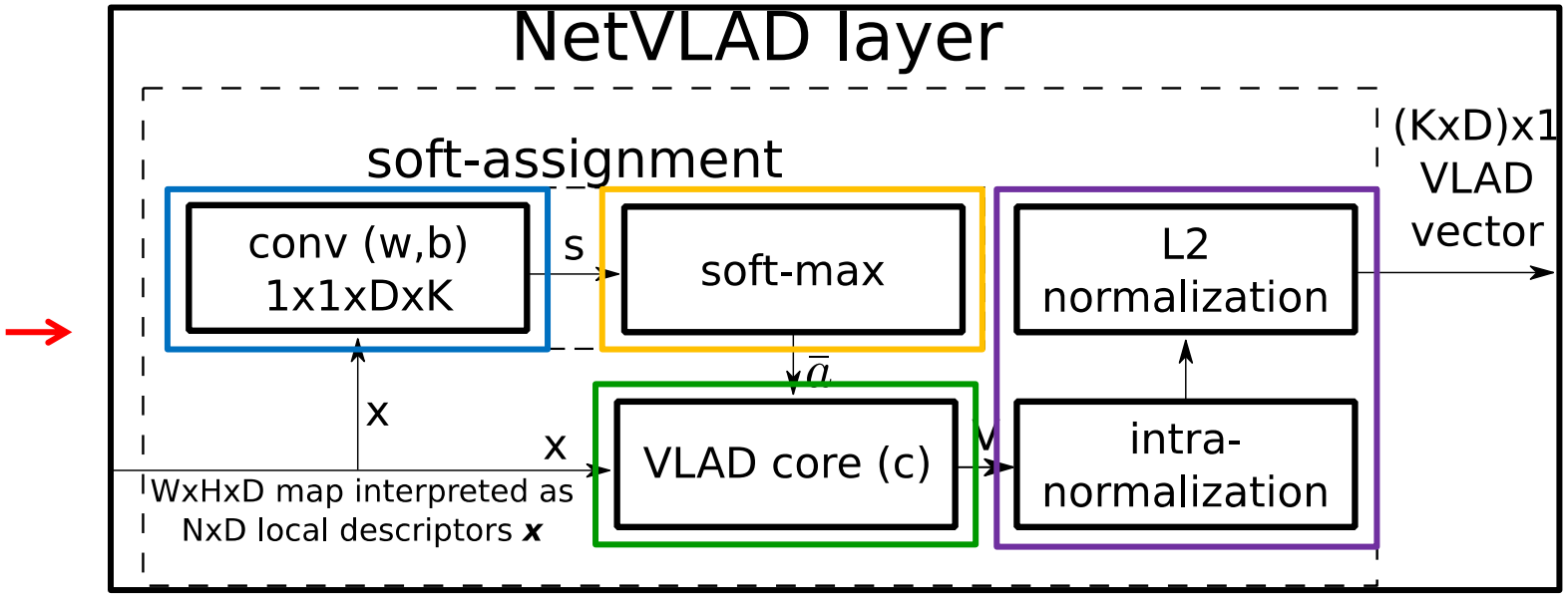
NetVLAD: NetVLAD Modification

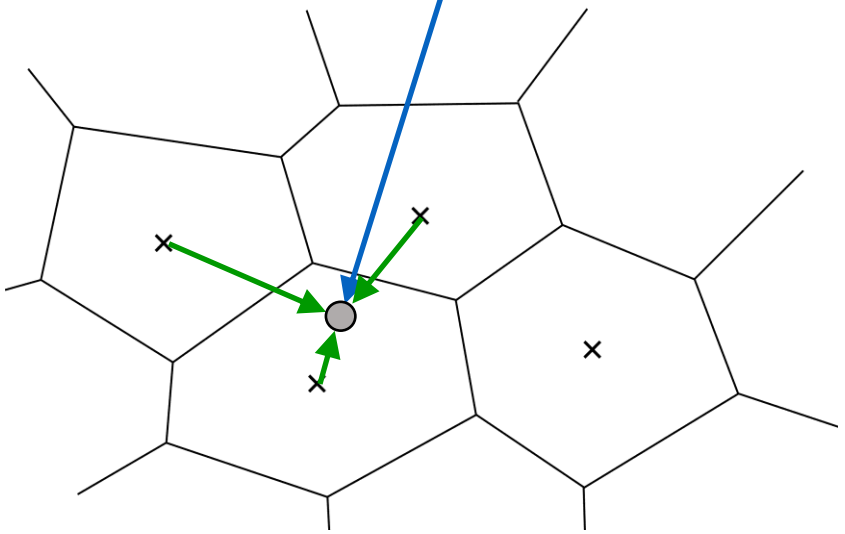
Soft-max → $\bar{a}$

$$V(:,k) = \sum_{i=1}^{N} \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'}^T x_i + b_{k'}}} (x_i - c_k)$$

Residual vector







NetVLAD: Can we apply CNNs to place recognition?

Questions:

1. (Model) What is a good CNN architecture?
2. **(Data) How to get the lots of annotated training data?**
3. (Loss) What is the appropriate loss for end-to-end training?

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Google Street View Time Machine

same locations at different times and seasons

Slides adopted from NetVLAD CVPR Presentation

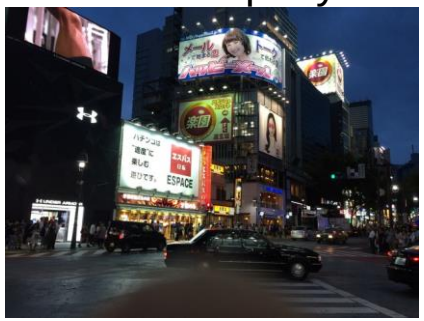
NetVLAD: Google Street View Time Machine

Tokyo 24/7 [Torii et al. 15]

- Database: 76k images from Street View
- Queries: 315 images from mobile phone cameras

Day query

Sunset query



Night query

Database image

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Can we apply CNNs to place recognition?

Questions:

1. (Model) What is a good CNN architecture?
2. (Data) How to get the lots of annotated training data?
3. **(Loss) What is the appropriate loss for end-to-end training?**

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Weakly Supervised Ranking Loss

$$L_\theta = \sum_j l\left(\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_j^q)\right)$$

hinge loss

margin

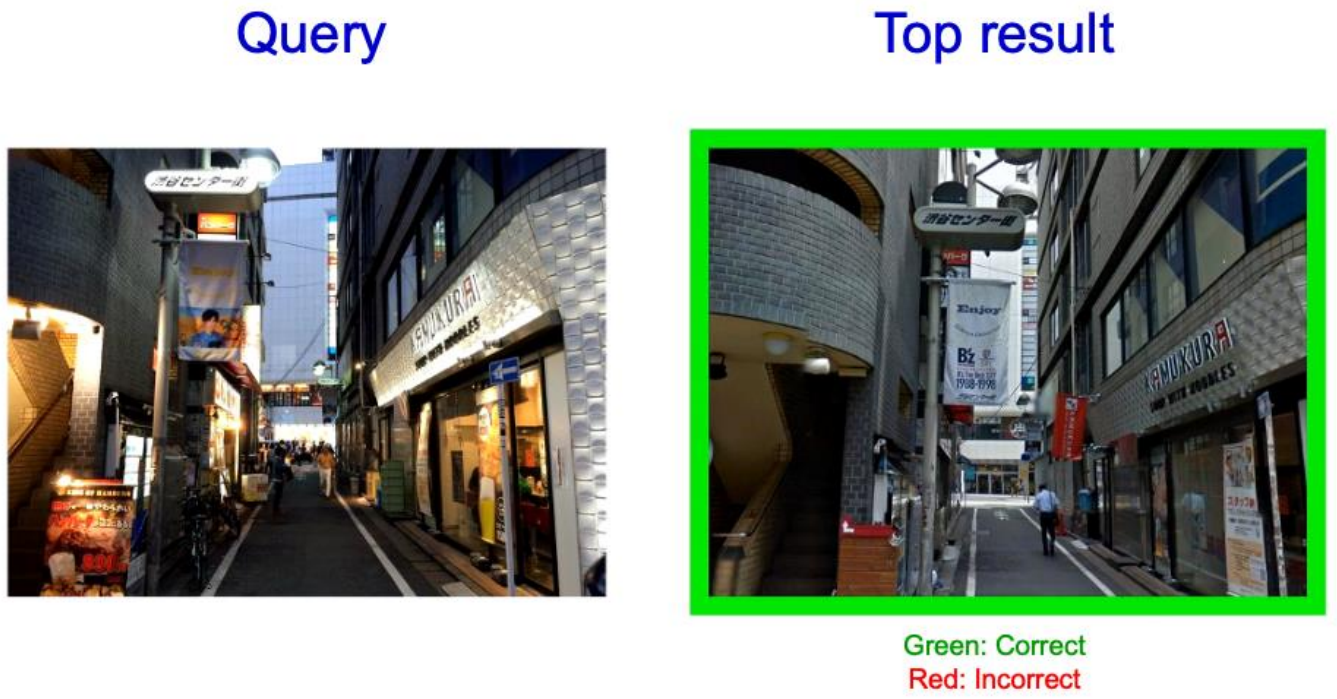
Sum over negatives

Distance to the best potential positive

Distance to the negative

Slides adopted from NetVLAD CVPR Presentation

NetVLAD: Results



Code and trained networks available online:

http://www.di.ens.fr/willow/research/netvlad/ (the link is in the paper)

Slides adopted from NetVLAD CVPR Presentation

Summary: Visual Place Recognition Pipeline

Query image

Database images



.....

Offline

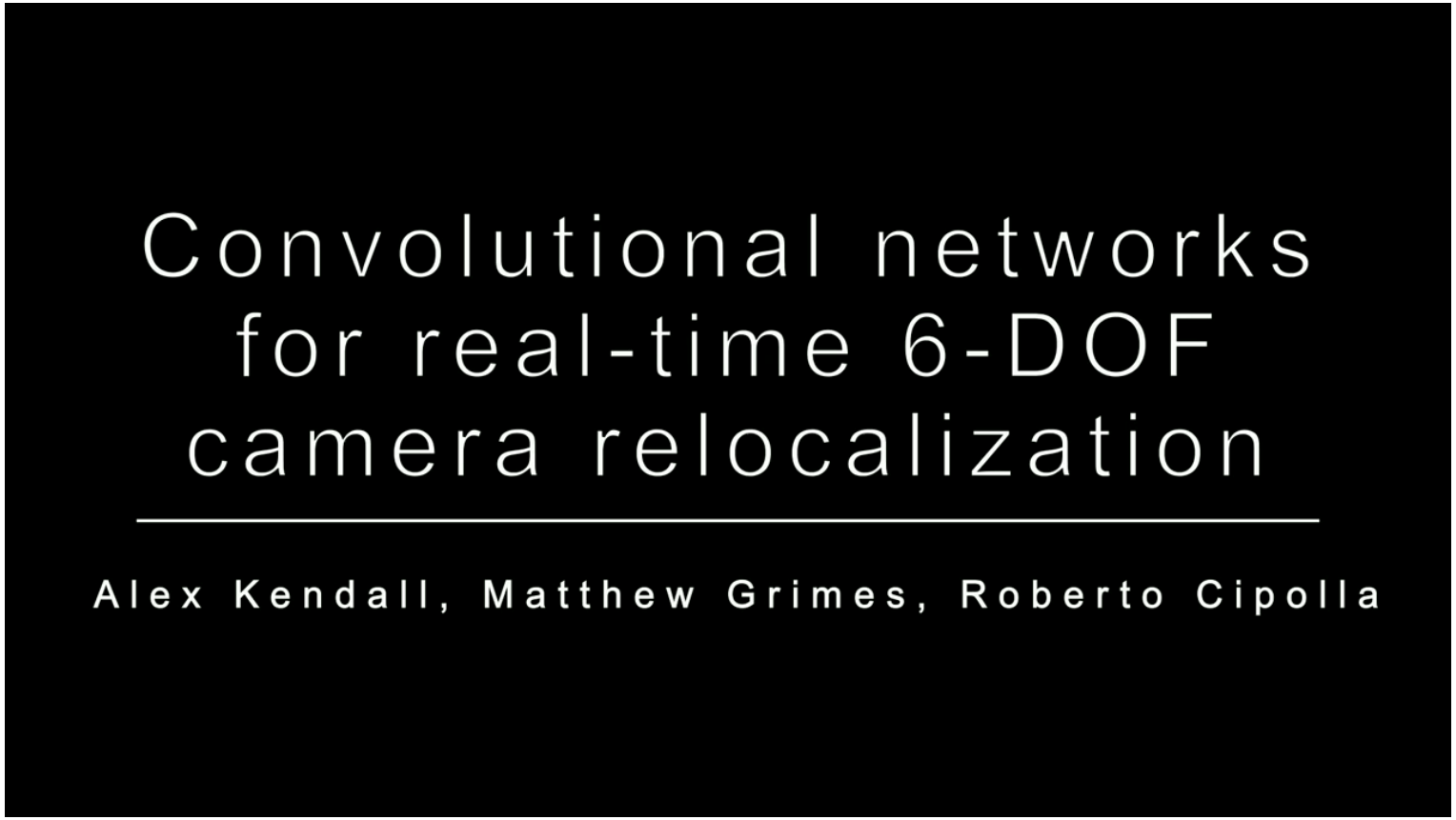
1. Feature detection & description
2. Training visual vocabulary
3. Image description

4. Feature detection & description
5. Image description

6. Initial ranking
7. Re-ranking with geometric verification

Slides from Sattler et. al., 2015

Posenet: Convolutional networks for real-time 6-DOF camera relocalization



https://www.youtube.com/watch?v=u0MVbL_RyPU

Slides adopted from Kendall at all, 2016

References

- Szeliski 2022
  - Section 6.3

- Forsyth & Ponce 2011
  - Section 15.1, Chapter 17

- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

- Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." arXiv preprint arXiv:1506.01497 (2015).
- Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.