The Bag of Words Torn Open:
Instance Retrieval goes Deep

AI Ukraine 2016
Kharkiv, Ukraine

James Pritts

Center for Machine Perception
Czech Technical University in Prague

Who are we?

Ondřej Chum
Associate Professor

Giorgos Tolias
Post-Doctoral candidate

Jiří Matas
Professor

Filip Radenović
PhD candidate

James Pritts
PhD candidate

Goals

- Introduce the Instance Retrieval Problem

- Compare two ways to **learn** an image encoding

  Bag-of-words (BoW) descriptor:

  ~1,000,000D vector

  Convolutional Neural Network (CNN) descriptor

  512D vector

- Demonstrate state-of-the-art retrieval performance

Part 1: The Instance Retrieval Task

Instance Retrieval Challenges

- Significant viewpoint and/or scale change
- Significant illumination change
- Severe occlusions
- Visually similar but different objects





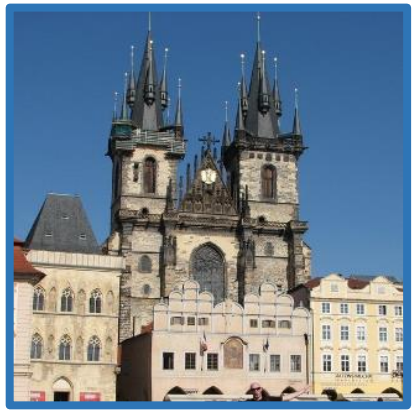
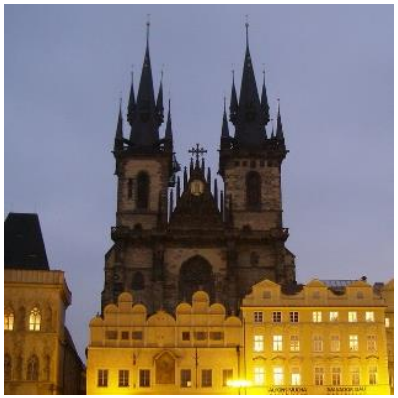
Instance Retrieval Challenges

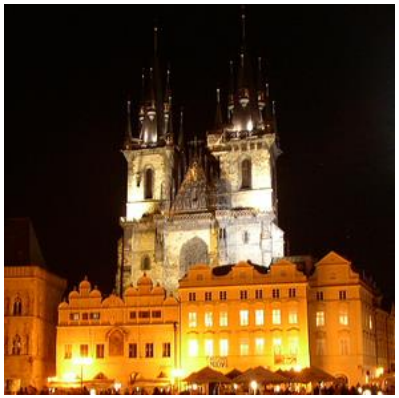
Significant viewpoint and/or scale change

Significant illumination change

Severe occlusions

Visually similar but different objects



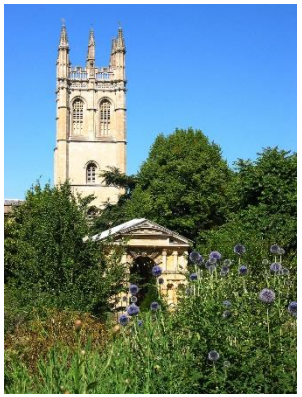
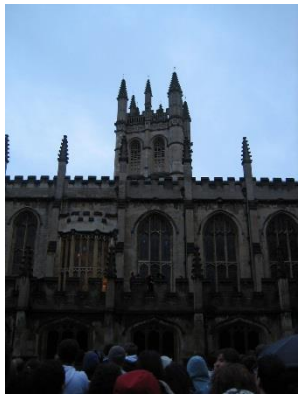
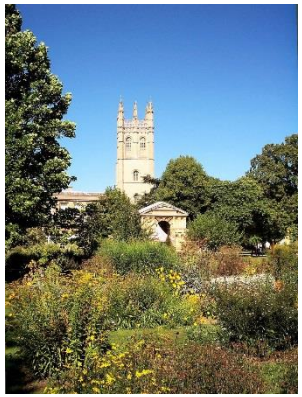
Instance Retrieval Challenges

Significant viewpoint and/or scale change

Significant illumination change

Severe occlusions

Visually similar but different objects


Instance Retrieval Challenges

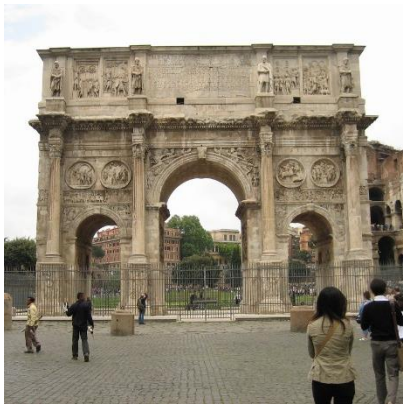
Significant viewpoint and/or scale change

Significant illumination change

Severe occlusions

Visually similar but different objects

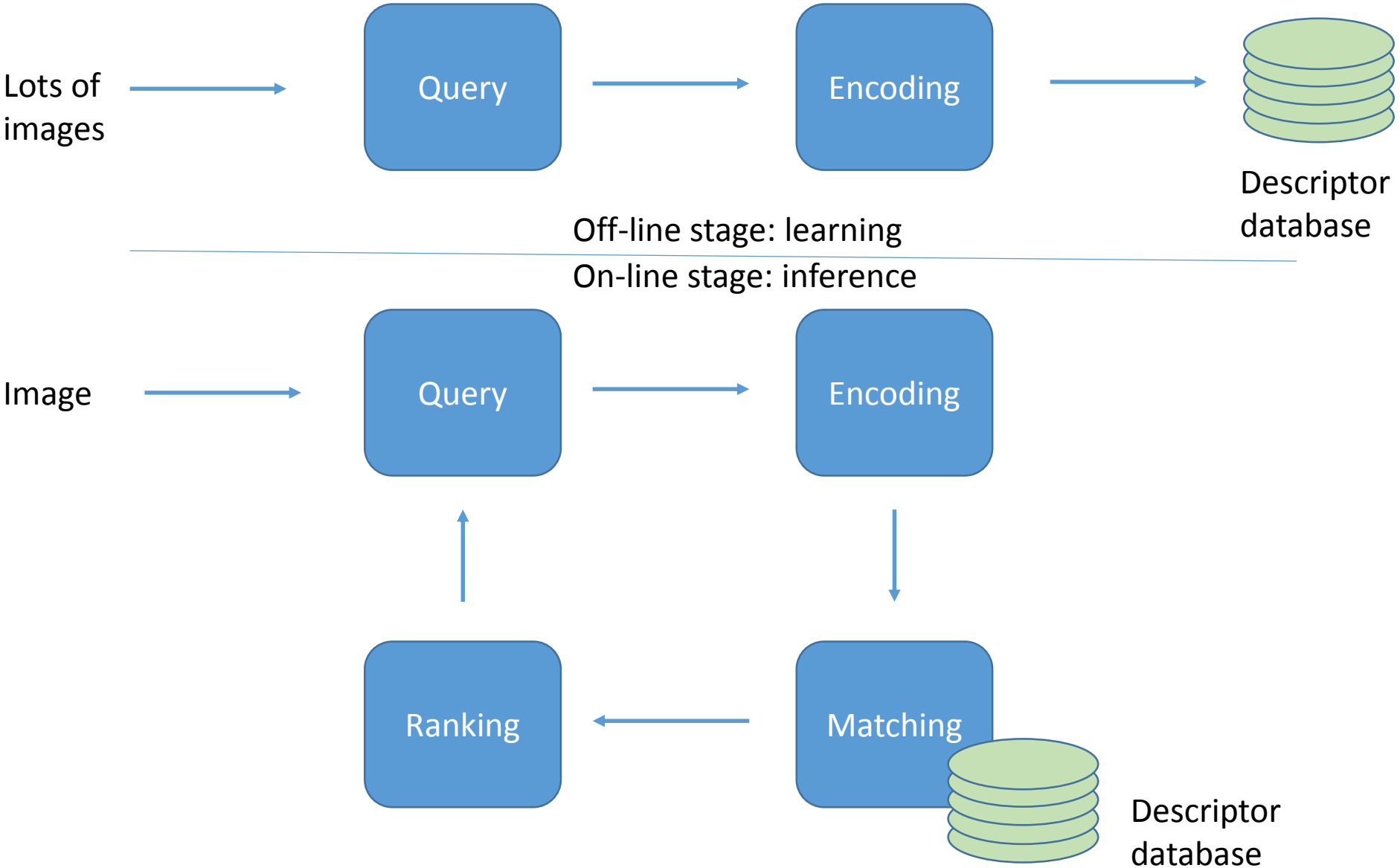


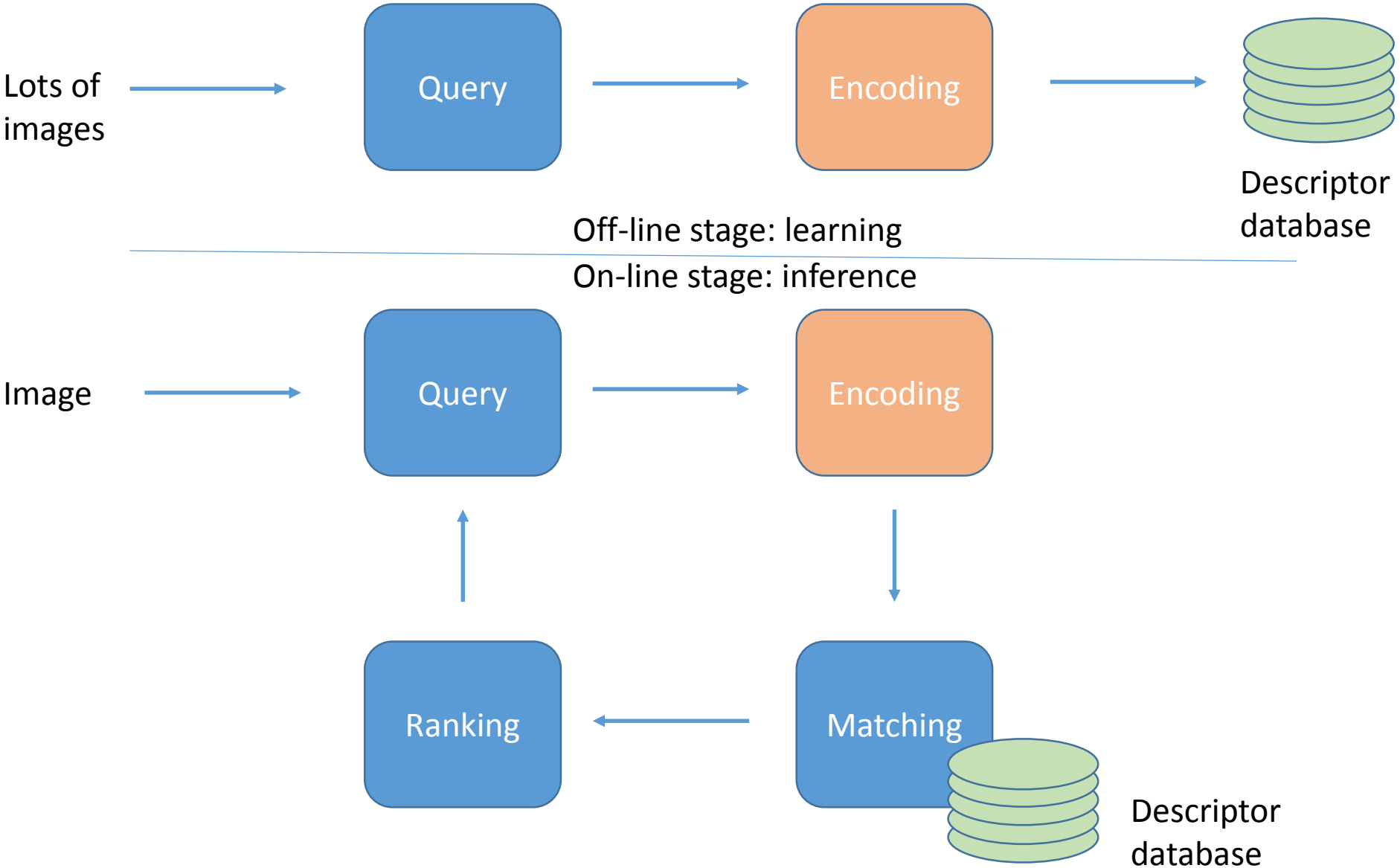
Instance Retrieval Demo

Click Here

Notional Instance Retrieval System

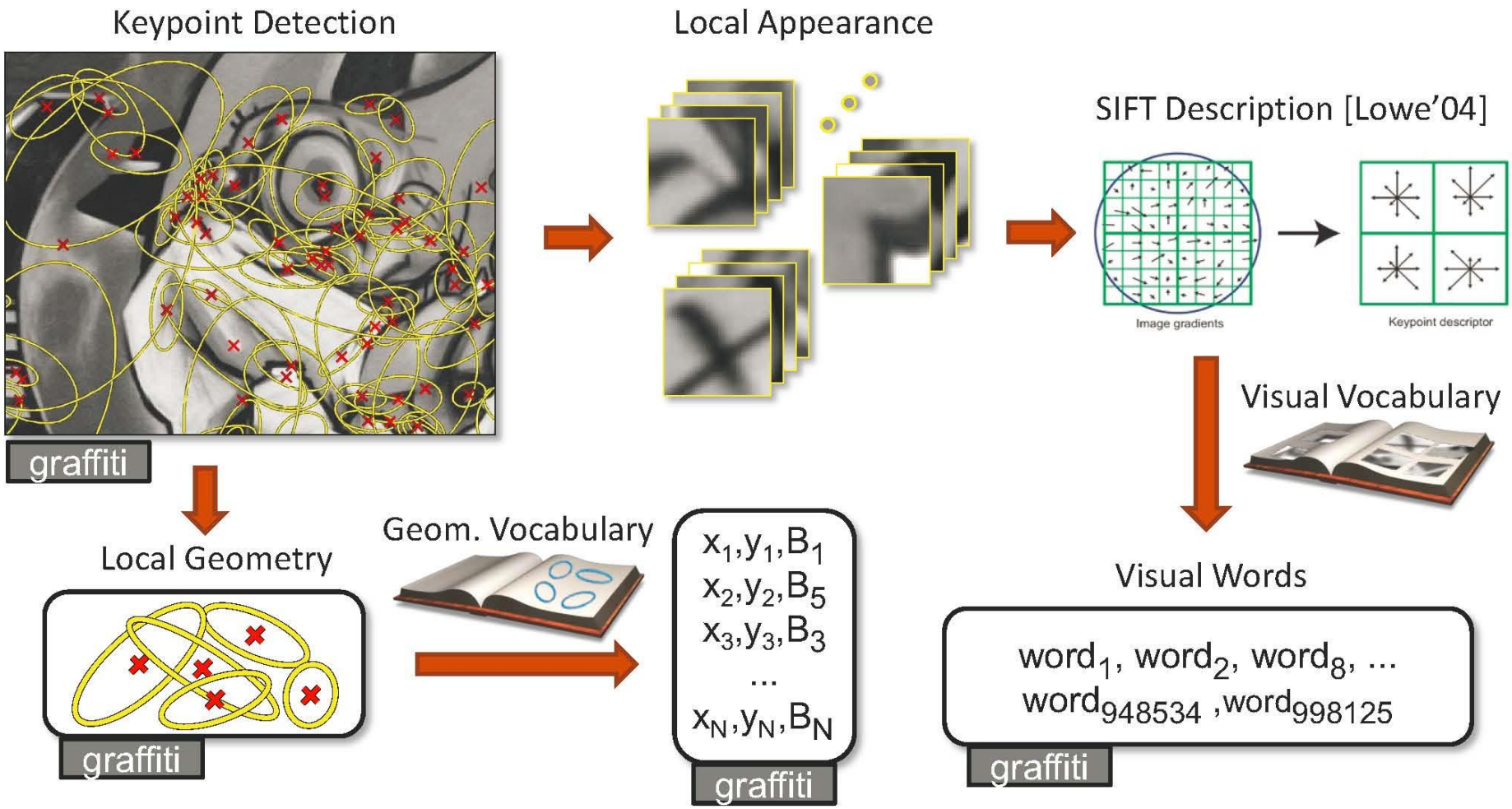


Notional Instance Retrieval System

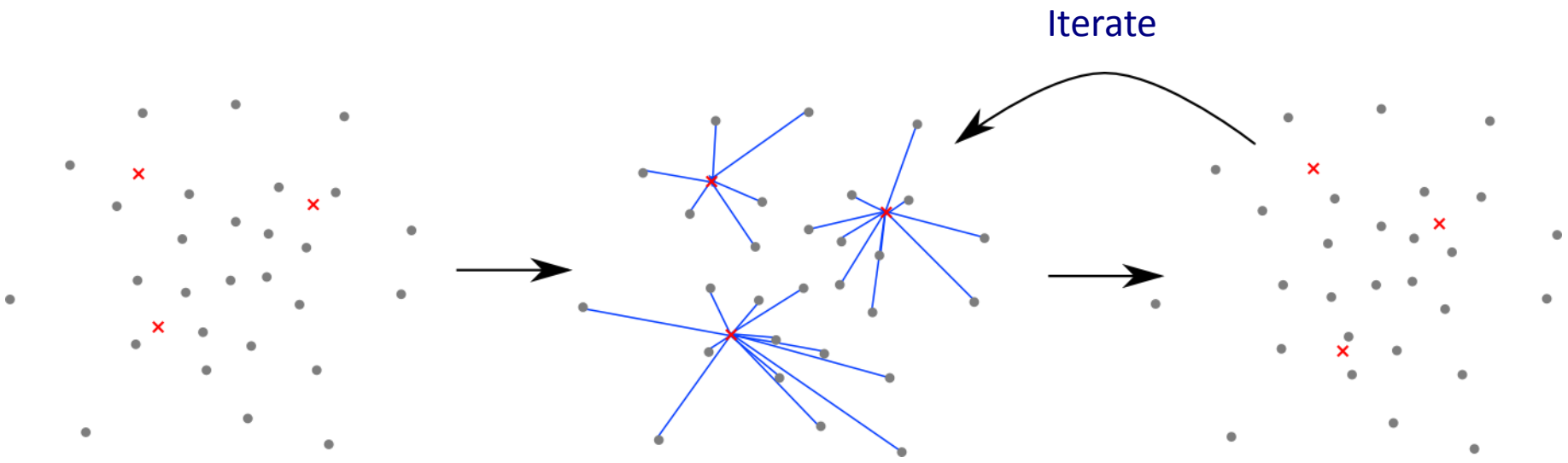


Part 2: The Bag of Words (BoW) representation

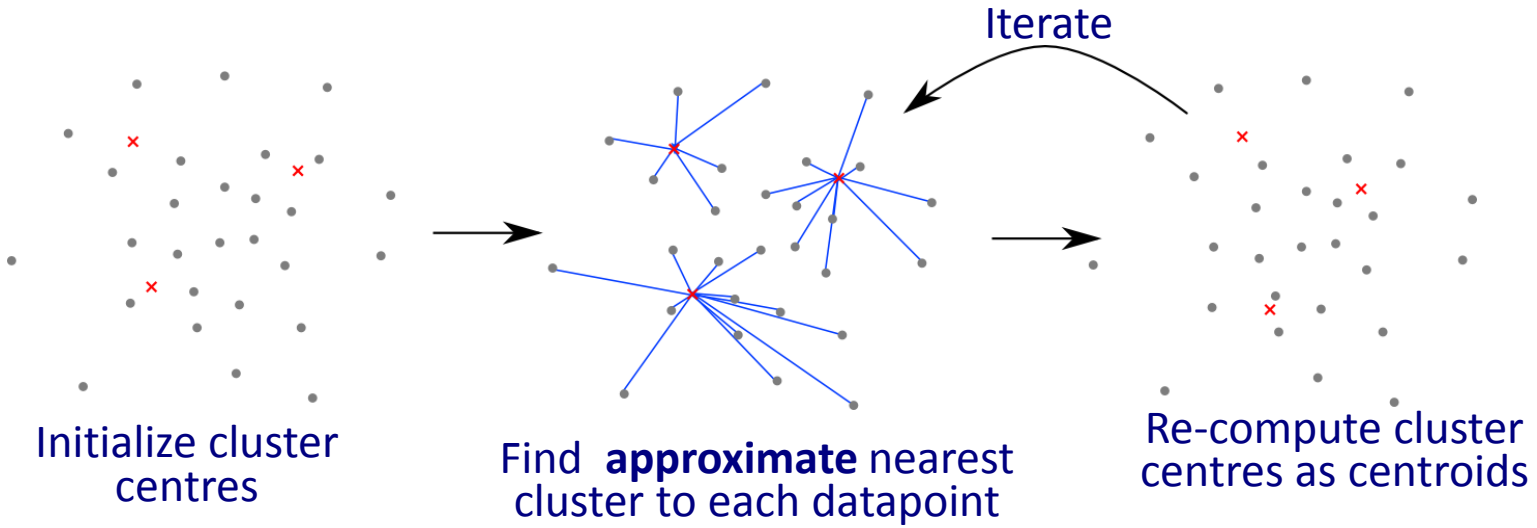
Bag of Words: Off-line stage



Quantization by K-Means



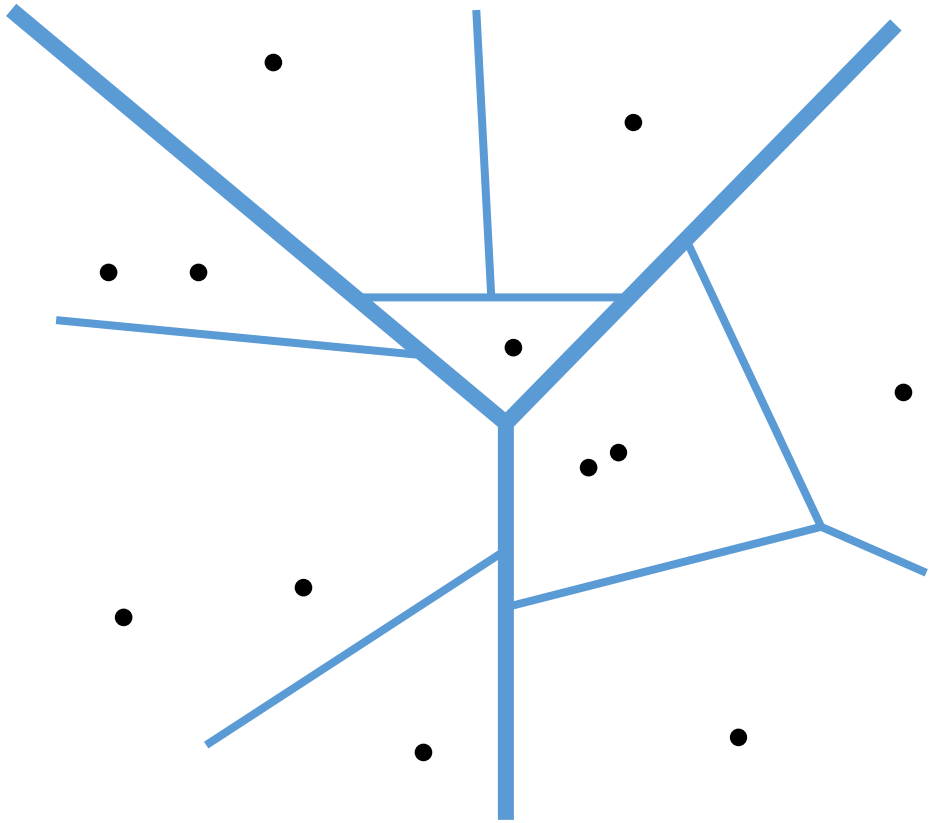
Quantization by Approximate K-Means



+ fast O(N log k)
+ reasonable quantization
- Can be inconsistent when ANN fails

Philbin, Chum, Isard, Sivic, and Zisserman – CVPR 2007
Object retrieval with large vocabularies and fast spatial matching

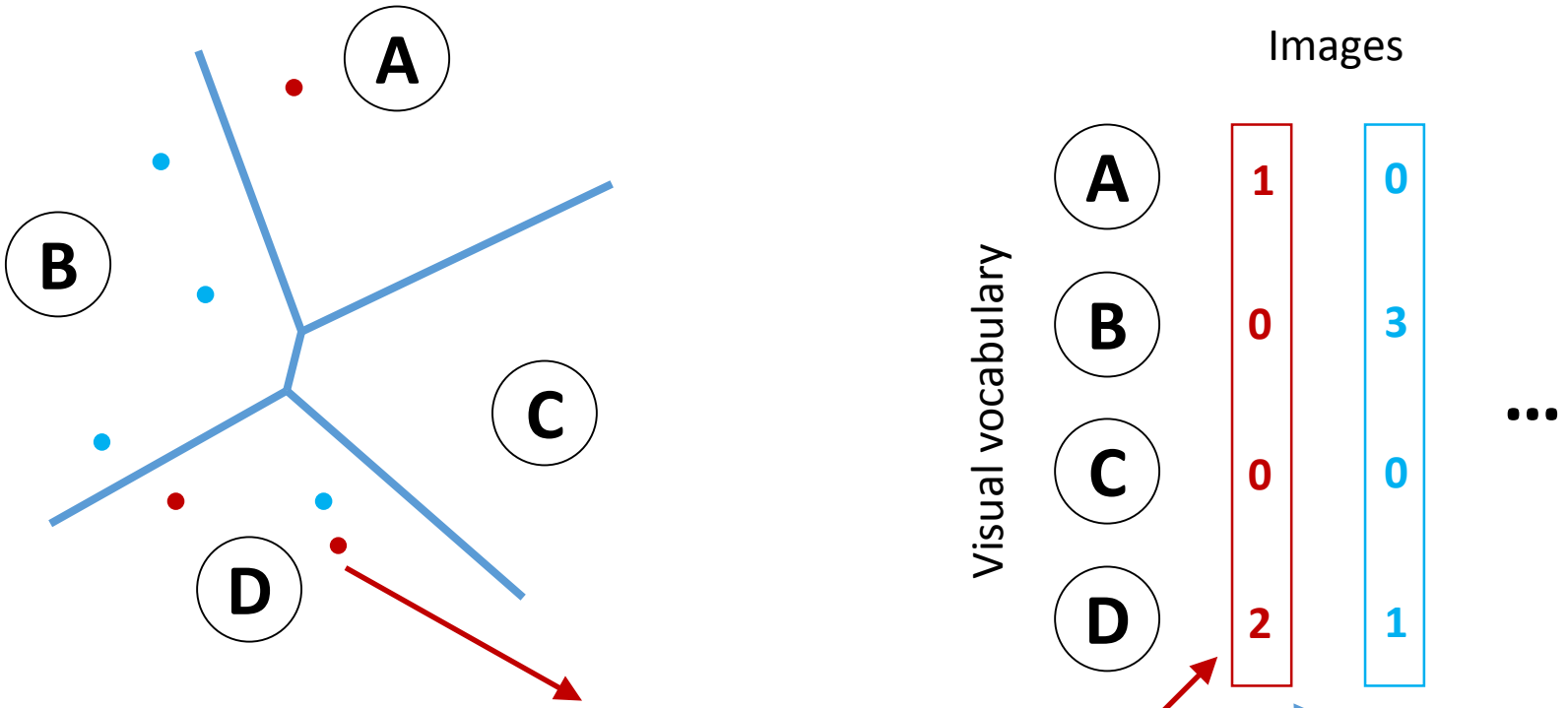
Quantization by Hierarchical K-means



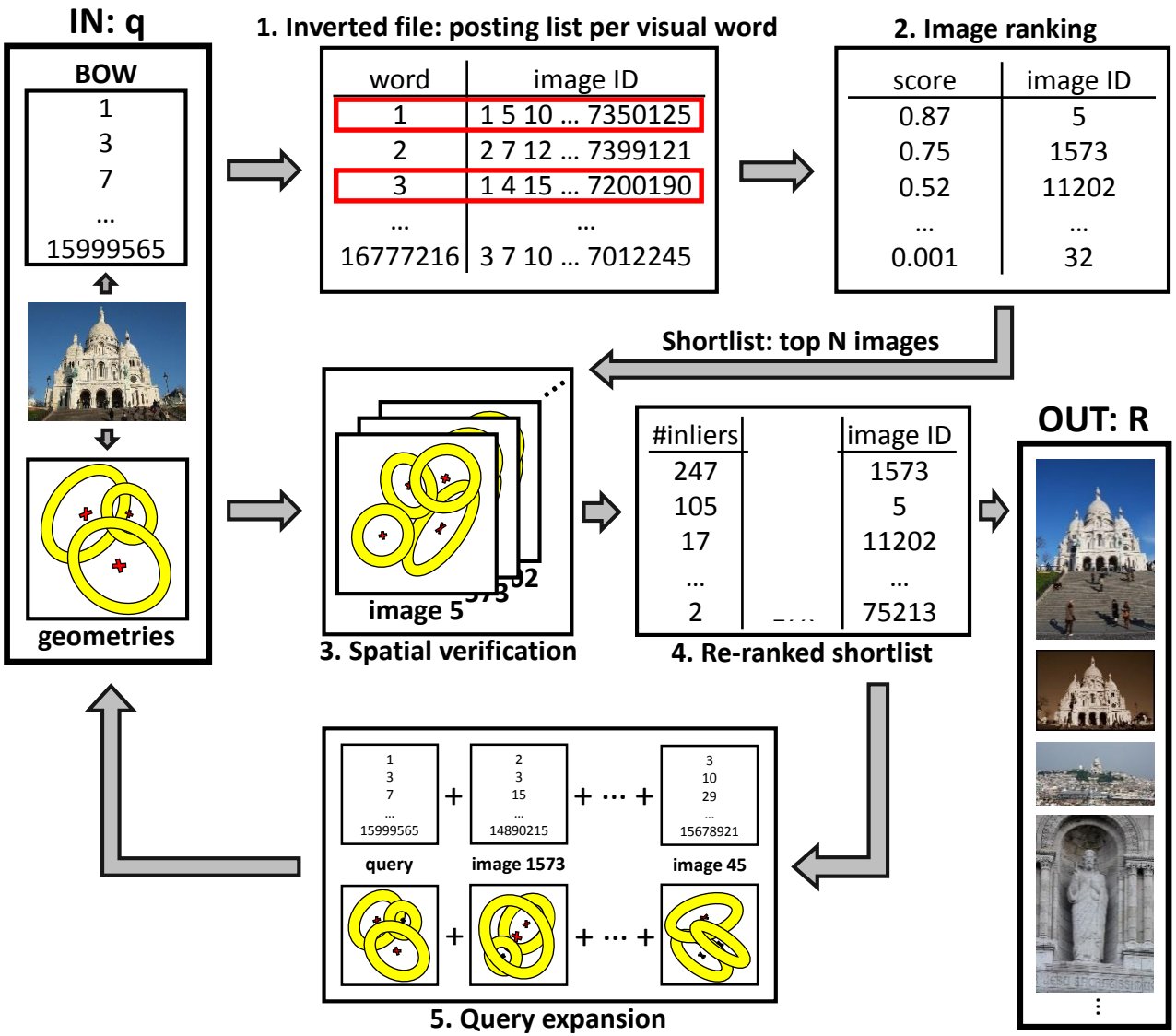
+ fast O(N log k)
+ incremental construction
- not so good quantization
- often imbalanced

Nistér & Stewénius: Scalable recognition with a vocabulary tree. CVPR 2006

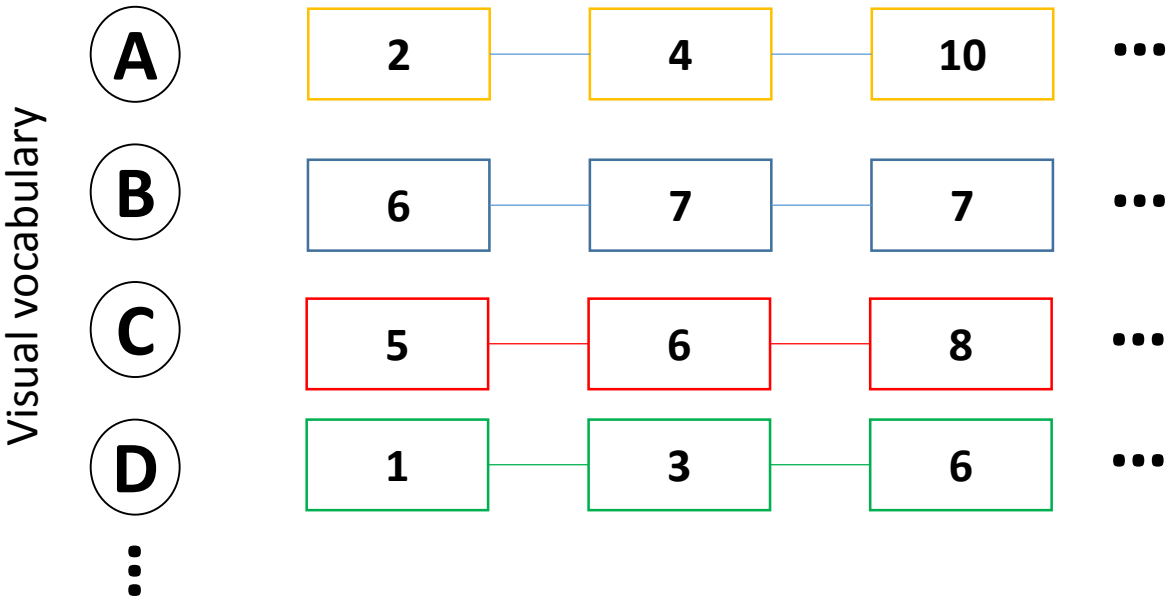
Bag-of-Words Image Representation



Bag of Words : On-line Stage

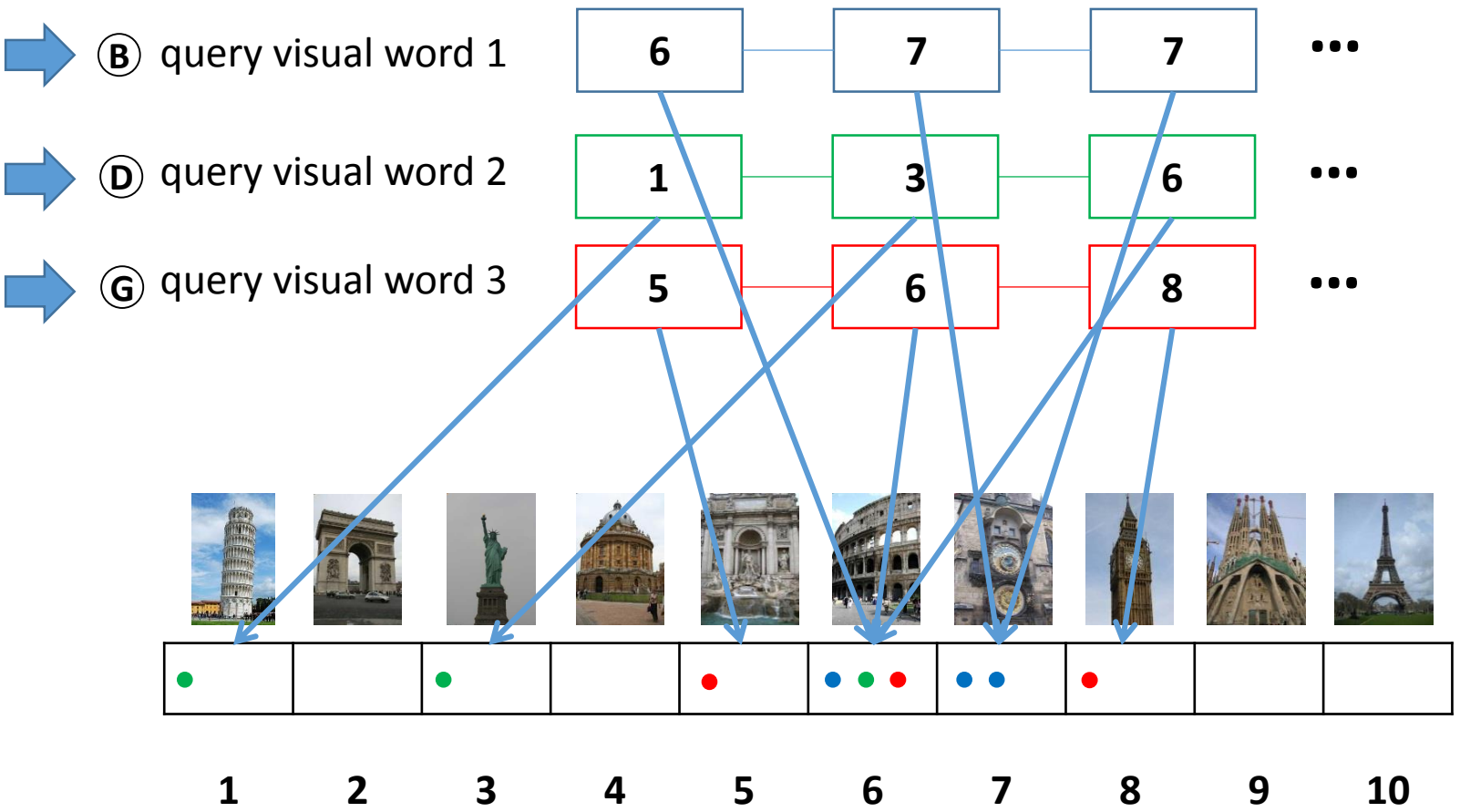


BoW and Inverted File



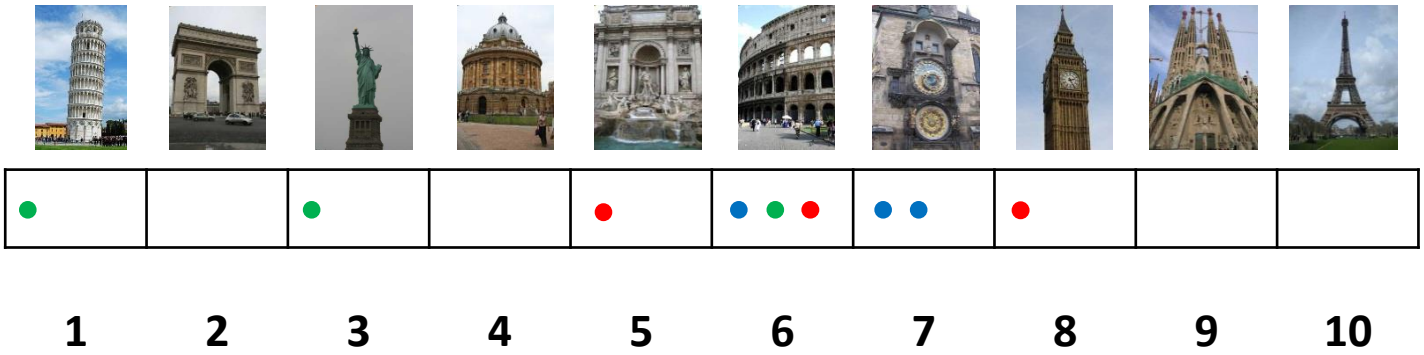
BoW and Inverted File

$$\text{score} = \frac{\mathbf{q}^\top \mathbf{x}}{||\mathbf{x}||}$$

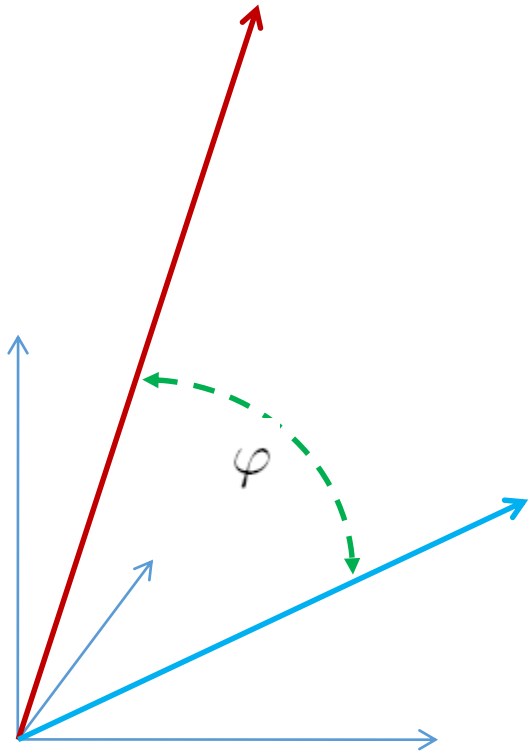


BoW and Inverted File

Efficient (fast)
Linear complexity (in # documents)
Can be interpreted as voting



Efficient Scoring



bag of words representation
(up to 1,000,000 D)

$$\cos\varphi = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} = \frac{1}{||\mathbf{x}||\,||\mathbf{y}||}\sum_{i=1}^{N} x_i y_i$$

$$\sum_{x_i\neq 0, y_i\neq 0} x_i y_i$$

Database

	A	B	C	D
α_1	1	0	0	2
α_2	0	2	0	1
α_3	1	0	0	0
⋮				

• Query α_q

A	0
B	3
C	0
D	1

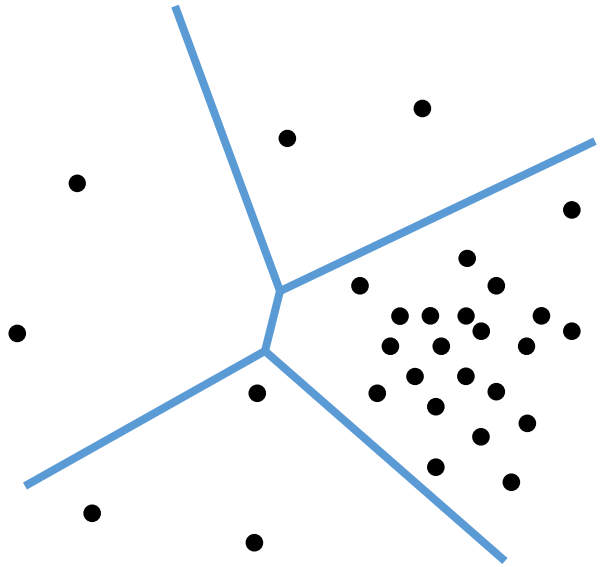
= Score

s_1
s_2
s_3
⋮

Word Weighting

Words (in text) common to many documents are less informative - ‘the’, ‘and’, ‘or’, ‘in’, …

$$idf_X = \log \frac{\text{\# documents}}{\text{\# docs containing } \textcircled{x}}$$

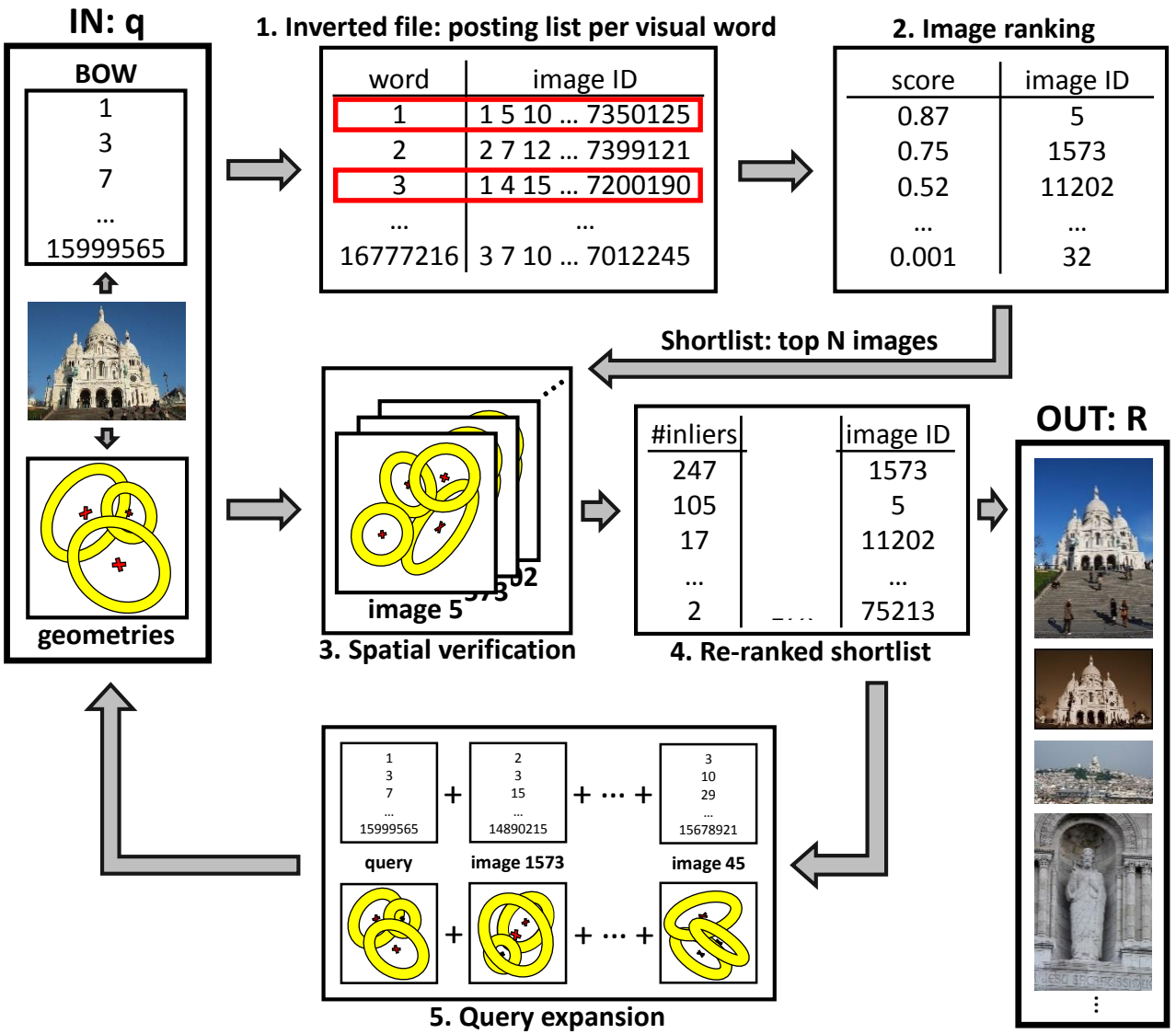


Images are represented by weighted histograms tf_X idf_X (rather than just a histogram of tf_X)

Words that are too frequent (virtually in every document) can be put on a stop list (ignored as if they were not in the document)

Baeza-Yates, Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999.

Bag of Words : On-line Stage



Query Expansion

Query image

Results

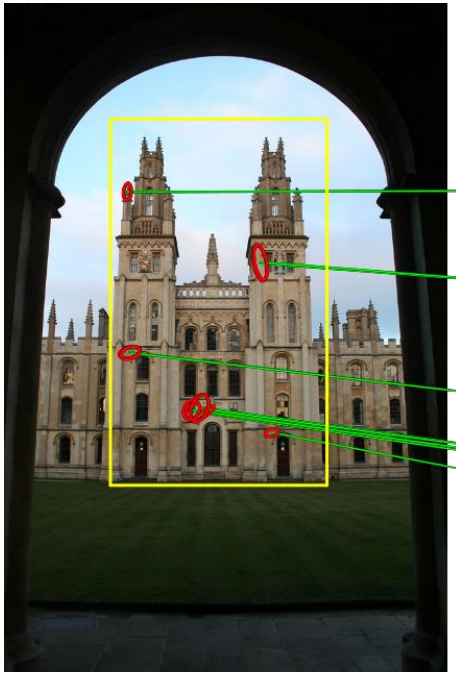
Spatial verification

New query

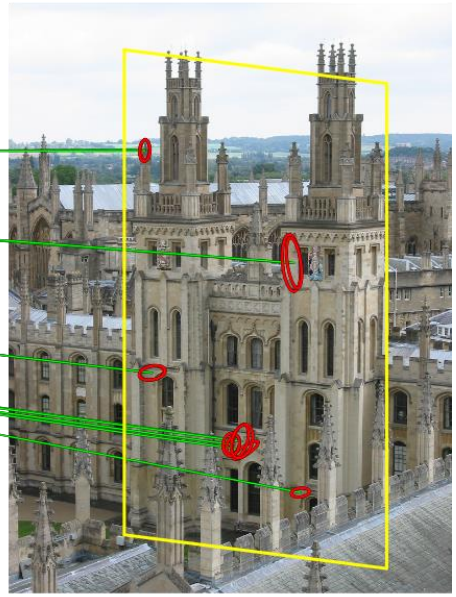
New results

Chum, Philbin, Sivic, Isard, Zisserman: Total Recall…, ICCV 2007

Query Expansion: Step by Step



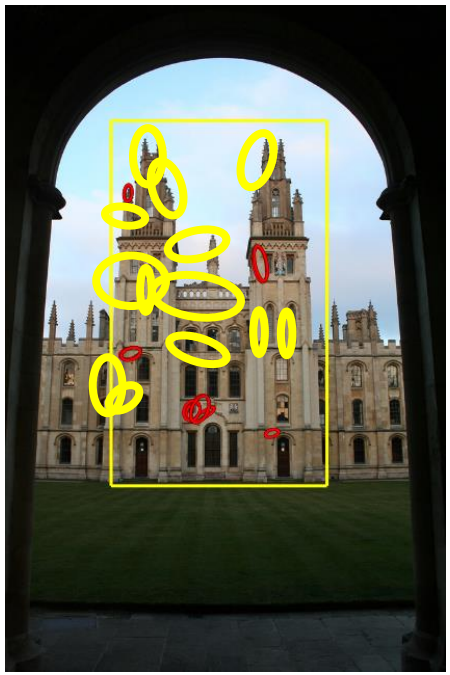
Query Image

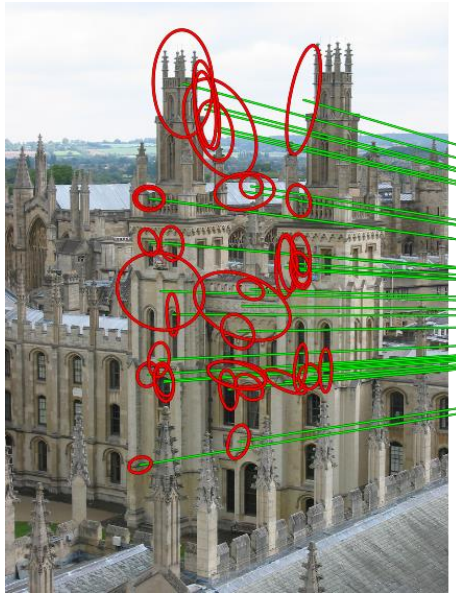


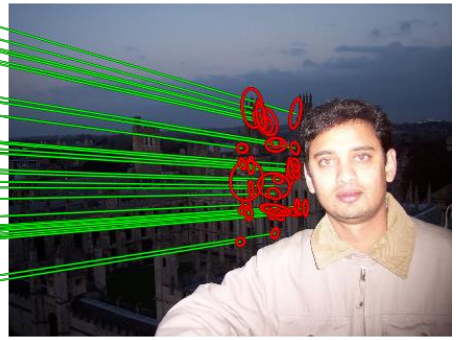
Retrieved image

Originally not retrieved

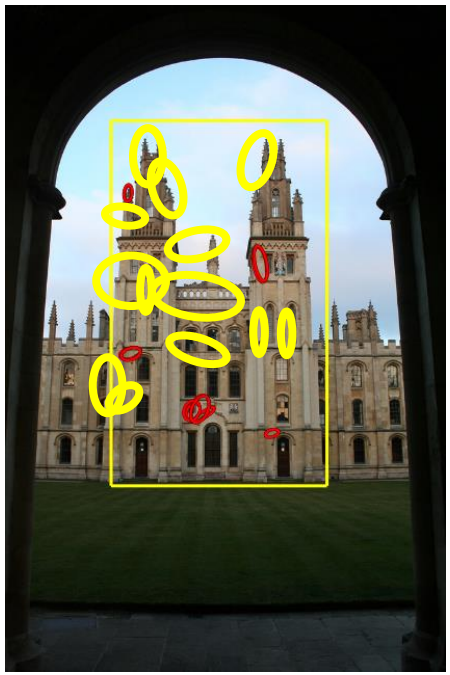
Query Expansion: Step by Step







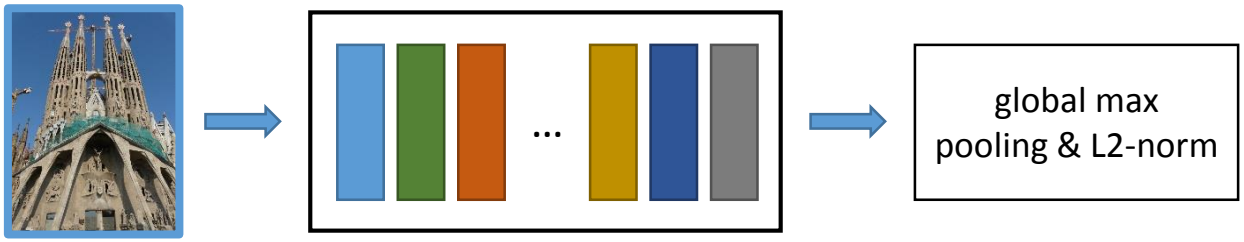
Query Expansion: Step by Step



The Bag of Words solution

Significant viewpoint scale change	covariant local features, invariant descriptors
Significant illumination change	color-normalized feature descriptors
Severe occlusions	locality of the features, geometric verification
Visually similar but different objects	Feature discriminability & geometric verification

** **Encoding is learned, but representation has many assumptions**

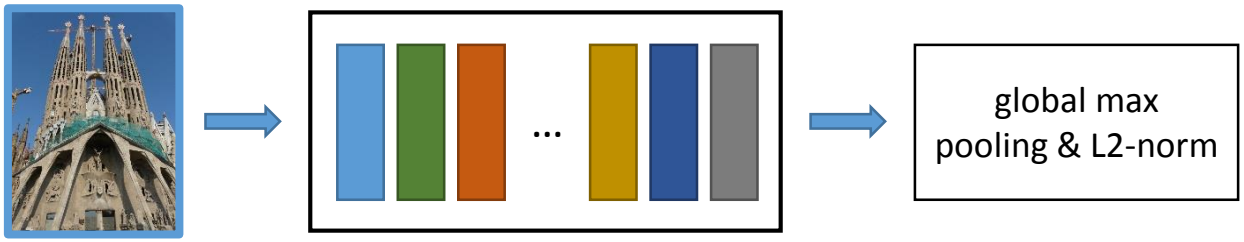


CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

Filip Radenović Giorgos Tolias Ondřej Chum

Center for Machine Perception, CTU in Prague

ECCV 2016



CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

Filip Radenović Giorgos Tolias Ondřej Chum

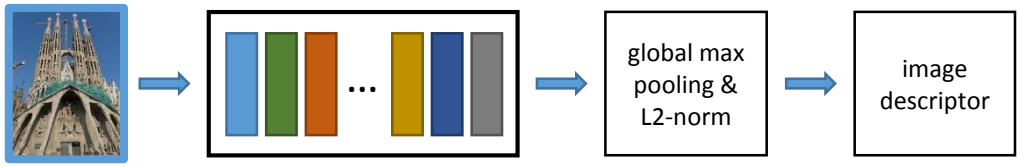
Center for Machine Perception, CTU in Prague

ECCV 2016

CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

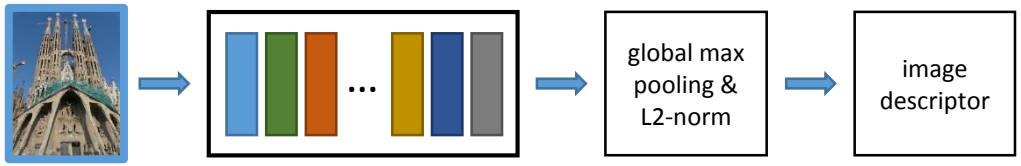
- compact image descriptors
- Nearest Neighbor search



CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

- compact image descriptors
- Nearest Neighbor search



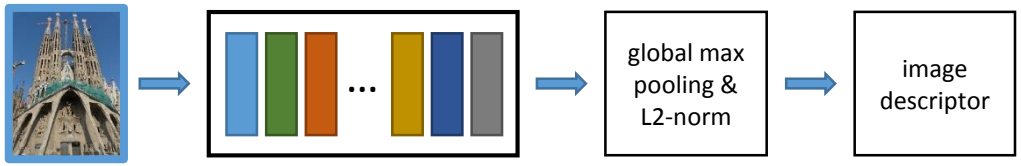
CNN Learning (Fine-Tuning)

- start with CNN trained for different but similar task (reasonable parameters)
- re-train with data relevant to your task

CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

- compact image descriptors
- Nearest Neighbor search

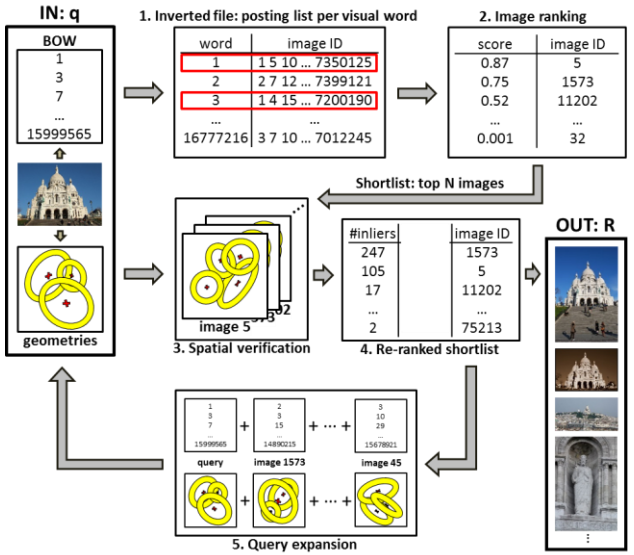


CNN Learning (Fine-Tuning)

- start with CNN trained for different but similar task (reasonable parameters)
- re-train with data relevant to your task

Bag of Words

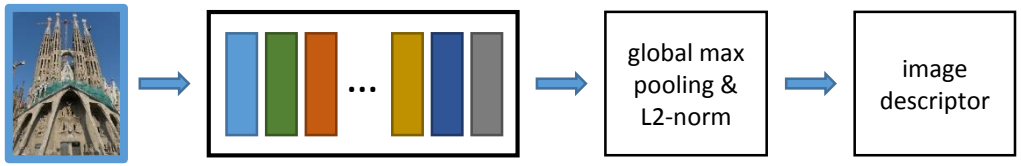
- state-of-the-art retrieval performance
- couples well with SfM



CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

- compact image descriptors
- Nearest Neighbor search



CNN Learning (Fine-Tuning)

- start with CNN trained for different but similar task (reasonable parameters)
- re-train with data relevant to your task

Bag of Words

- state-of-the-art retrieval performance
- couples well with SfM

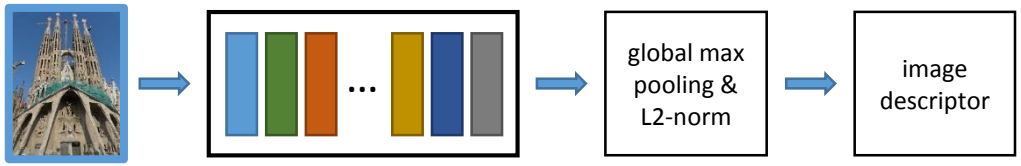
Unsupervised training data generation

- no human interaction

CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

CNN Image Retrieval

- compact image descriptors
- Nearest Neighbor search



CNN Learning (Fine-Tuning)

- start with CNN trained for different but similar task (reasonable parameters)
- re-train with data relevant to your task

Bag of Words

- state-of-the-art retrieval performance
- couples well with SfM

Unsupervised training data generation

- no human interaction

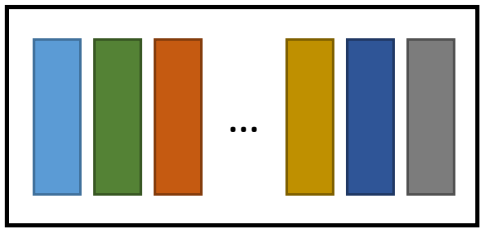
Hard Examples

“Lots of Training Examples”

Large Internet photo collection

Training

Image annotations



Convolutional Neural Network (CNN)

“Lots of Training Examples”

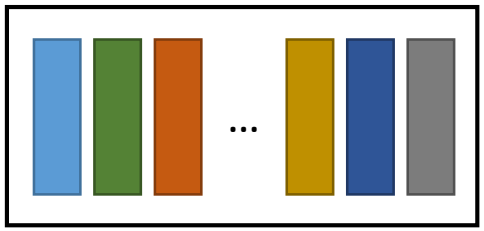
Manual cleaning of the training data done by Researchers

Very expensive $$$$

Large Internet photo collection

amazon mechanical turk™ Artificial Artificial Intelligence

Not accurate
Expensive $$



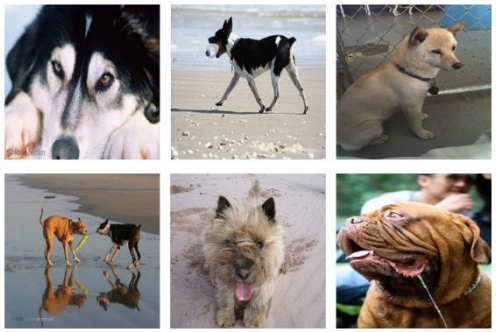
Convolutional Neural Network (CNN)

Automated extraction of training data

Very accurate
Free ~~$~~

Off-the-shelf CNN

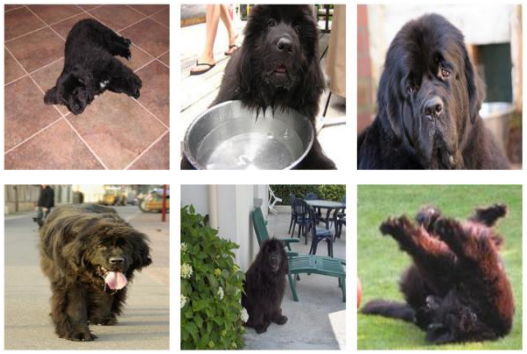
- Target application: classification
- Training dataset: ImageNet
- Architecture: AlexNet & VGG



Images from ImageNet.org

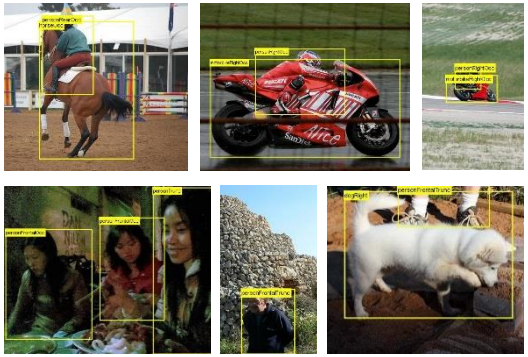
- Directly applicable to other tasks

Fine-grain classification



Images from ImageNet.org

Object detection



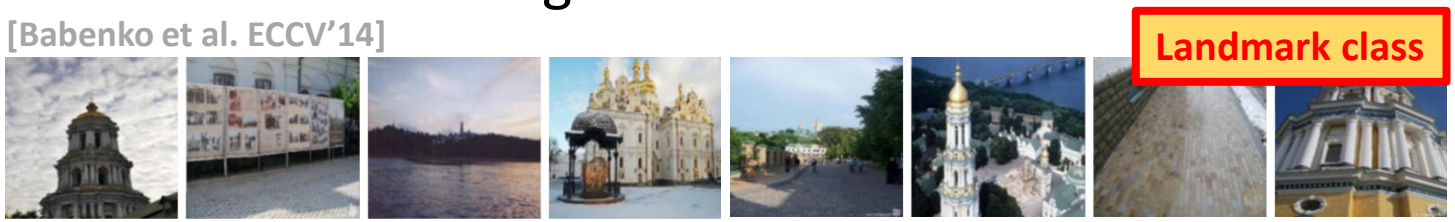
Images from PASCAL VOC 2012

Image retrieval

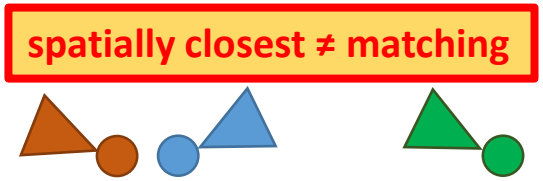
Annotations for CNN Image Retrieval

- CNN pre-trained for classification task used for retrieval
  [Gong et al. ECCV'14, Babenko et al. ICCV'15, Kalantidis et al. arXiv'15, Tolias et al. ICLR'16]

- Fine-tuned CNN using a dataset with landmark classes
  [Babenko et al. ECCV'14]



- NetVLAD: Weakly supervised fine-tuned CNN using GPS tags
  [Arandjelovic et al. CVPR'16]



- We propose: automatic annotations for CNN training

Retrieval and SfM



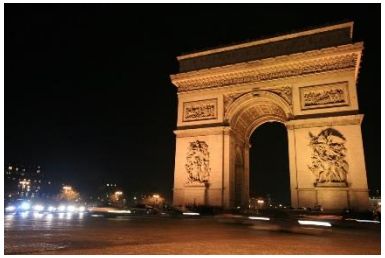










[Schonberger et al. CVPR'15]
[Radenovic et al. CVPR'16]

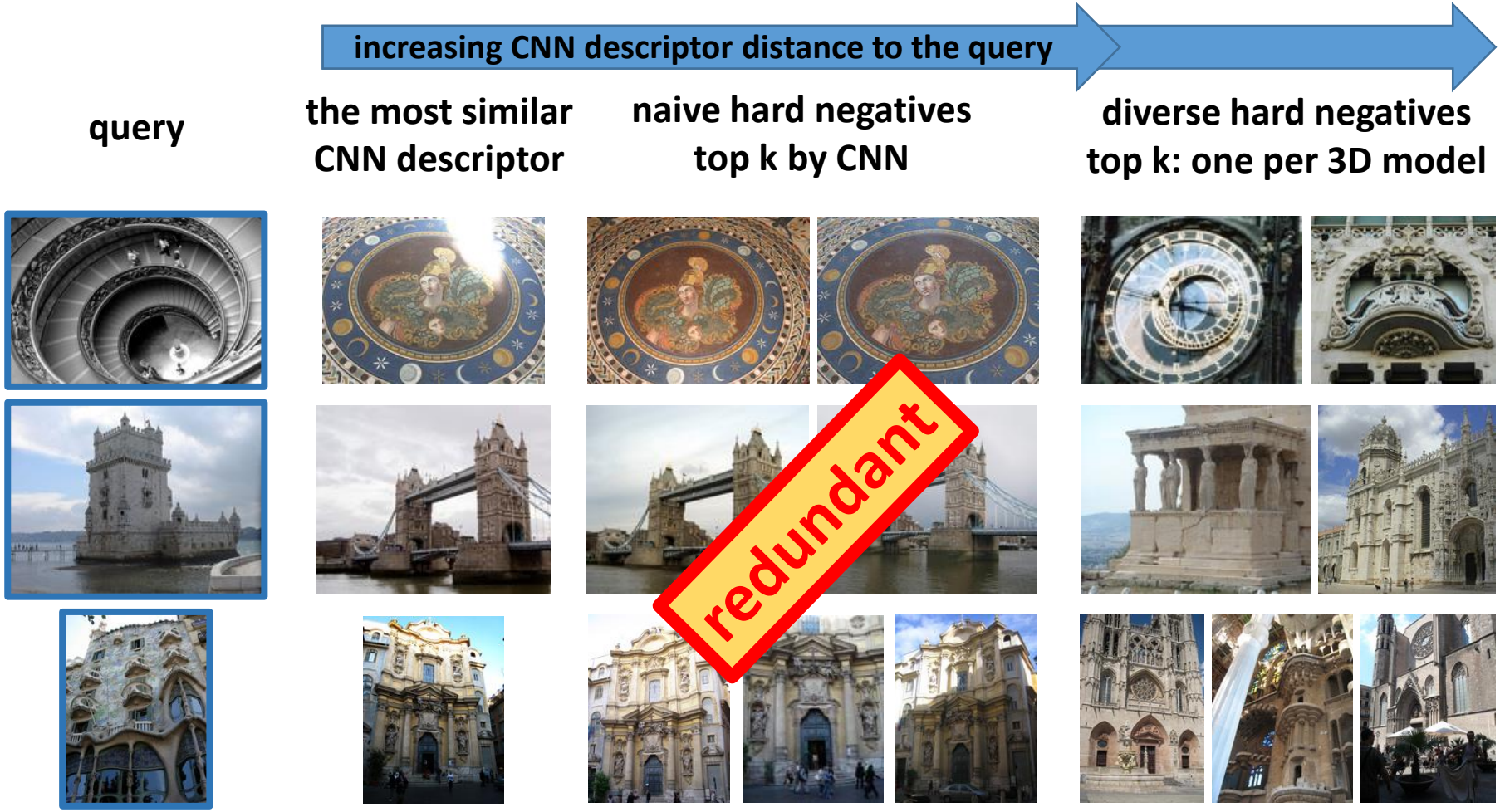
CNN learns from BoW – Training Data

Camera Orientation Known
Number of Inliers Known

[Schonberger et al. CVPR'15] 7.4M images → 713 training 3D models

Hard Negative Examples

Negative examples: images from different 3D models than the query
Hard negatives: closest negative examples to the query
Only hard negatives: as good as using all negatives, but faster



Hard Positive Examples

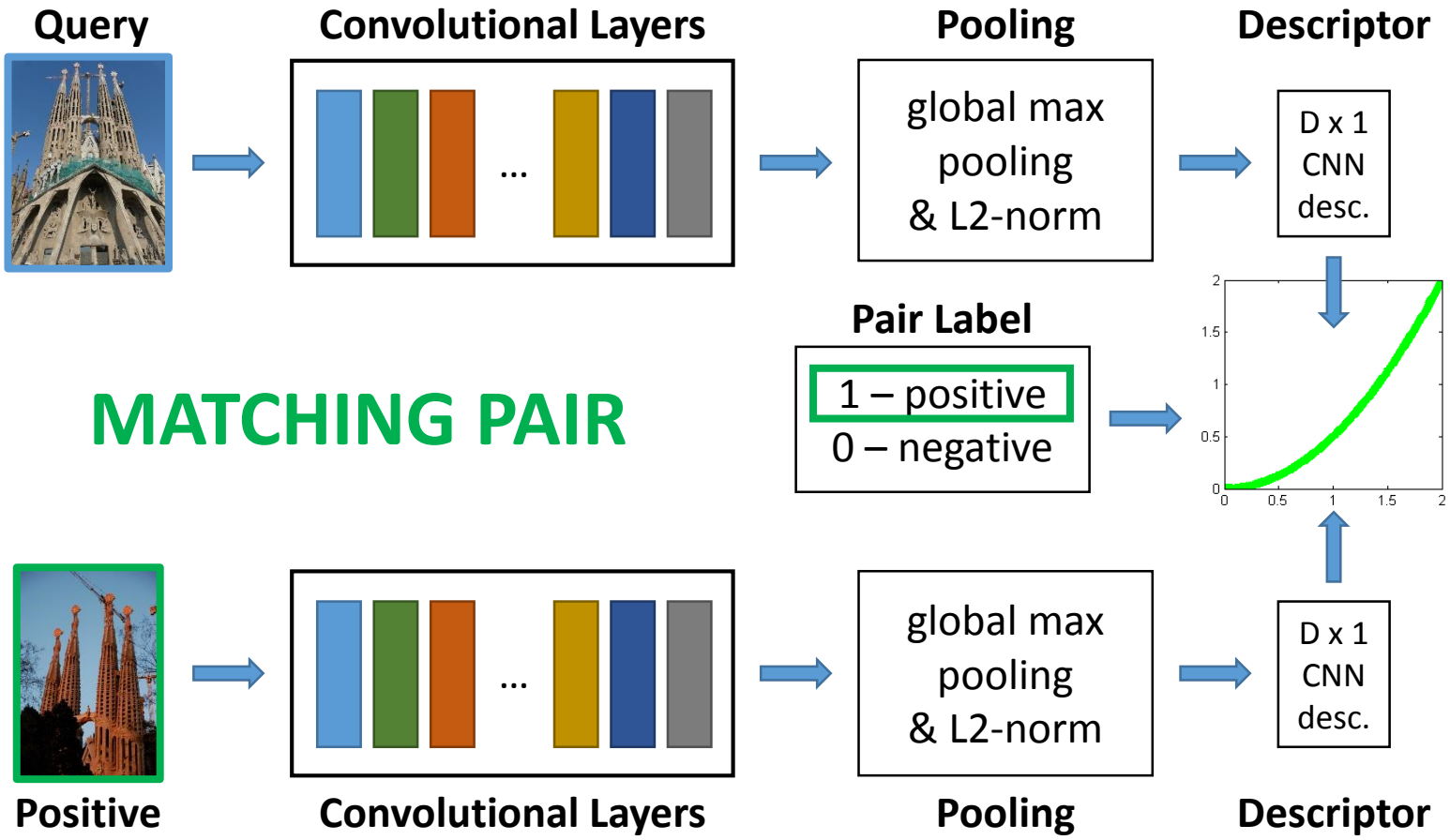
Positive examples: images from the same 3D model as the query
Hard positives: positive examples not close enough to the query

query | **top 1 by CNN** | **top 1 by BoW** | **random from top k by BoW**

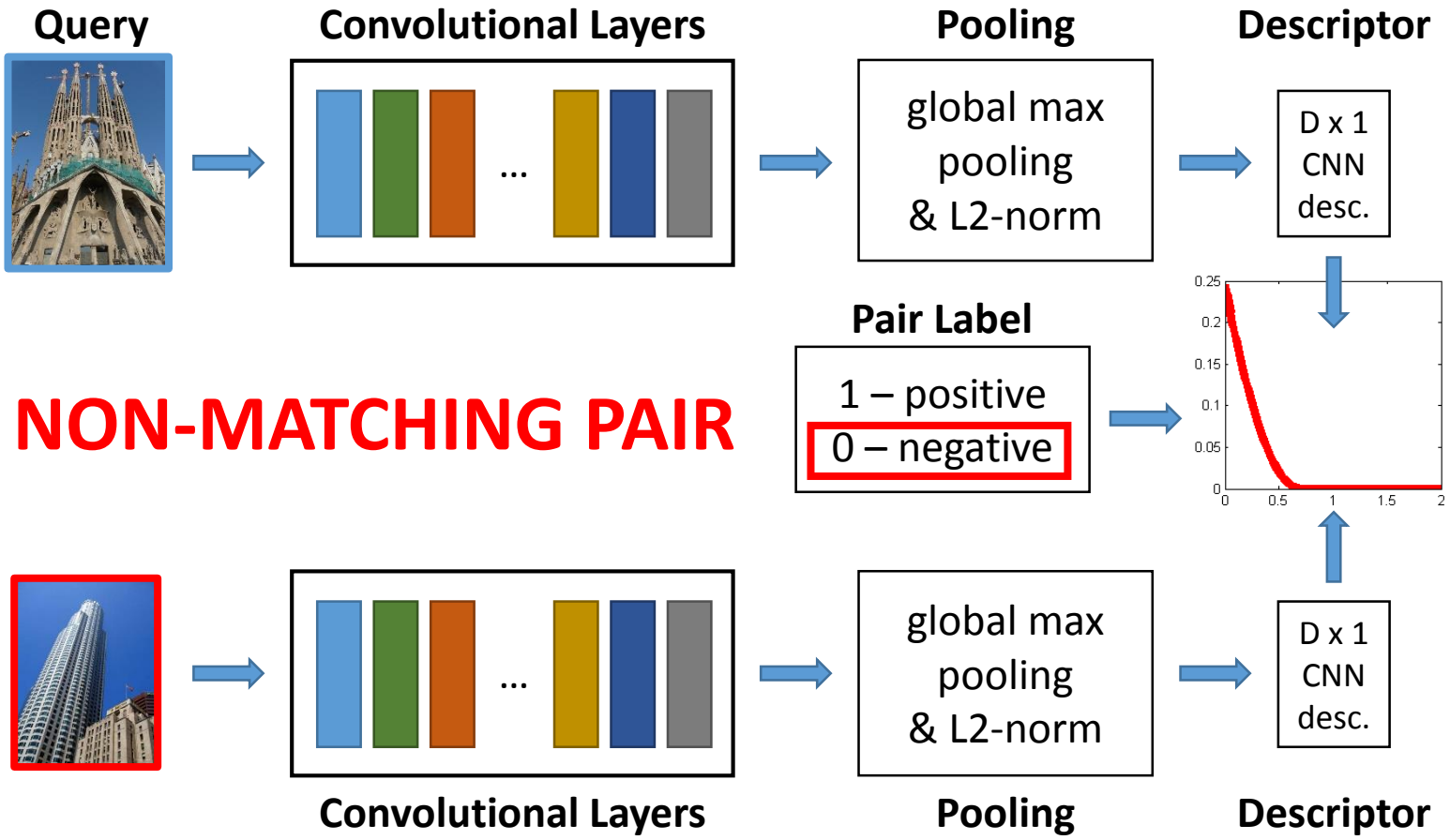
harder positives

used in NetVLAD

CNN Siamese Learning



CNN Siamese Learning

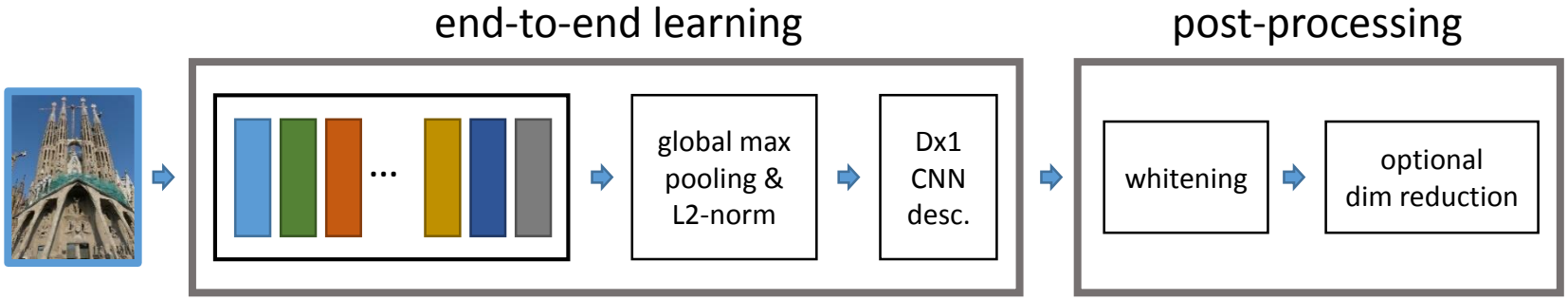


<u>Contrastive vs. Triplet loss</u>: Contrastive better with our data

Contrastive loss more strict, requires accurate training data

Triplet loss less sensitive to inaccurate annotation

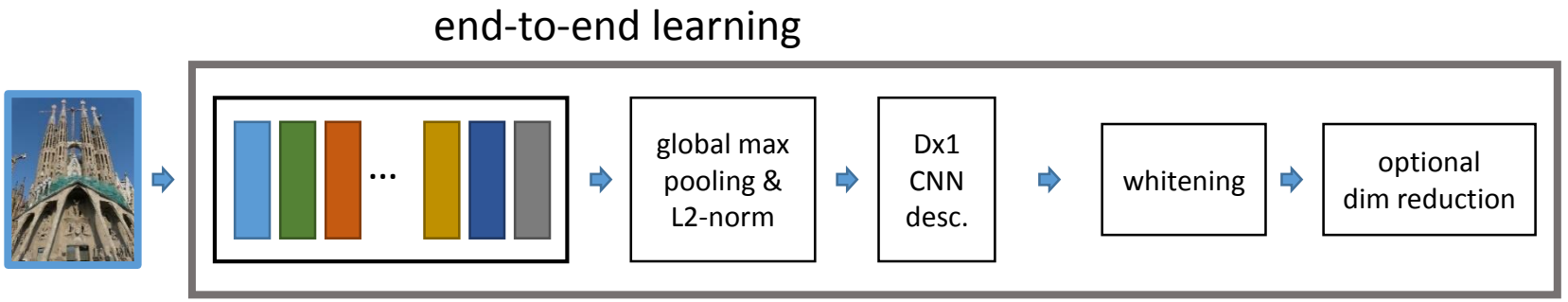
Whitening and dimensionality reduction



1. PCA_w – PCA of an independent set of descriptors
   [Babenko et al. ICCV'15, Tolias et al. ICLR'16]

2. L_w – We propose to learn whitening using labeled training data and linear discriminant projections
   [Mikolajczyk & Matas ICCV'07]

Whitening and dimensionality reduction

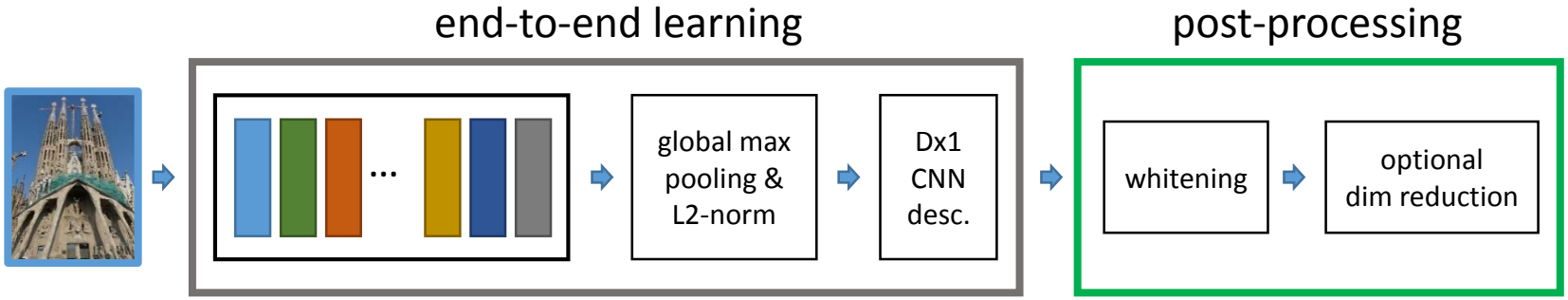


1. PCA_w – PCA of an independent set of descriptors
   [Babenko et al. ICCV'15, Tolias et al. ICLR'16]

2. L_w – We propose to learn whitening using labeled training data and linear discriminant projections
   [Mikolajczyk & Matas ICCV'07]

3. End-to-end Learning – Performs comparable or worse than L_w, while slowing down the convergence

Whitening and dimensionality reduction

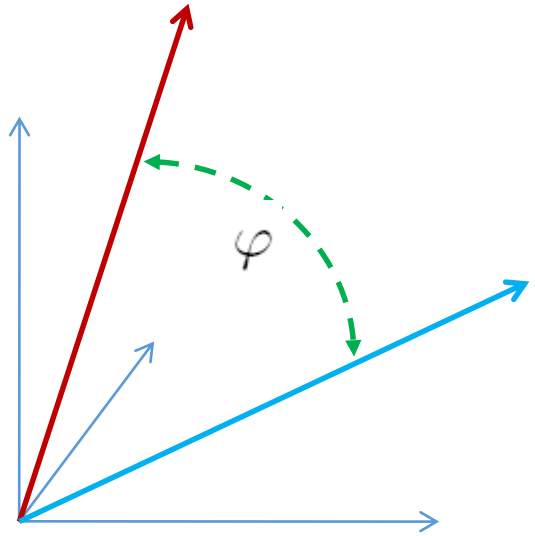


1. PCA_w – PCA of an independent set of descriptors
   [Babenko et al. ICCV'15, Tolias et al. ICLR'16]

2. L_w – We propose to learn whitening using labeled training data and linear discriminant projections
   [Mikolajczyk & Matas ICCV'07]

3. End-to-end Learning – Performs comparable or worse than L_w, while slowing down the convergence

Efficient Scoring and Ranking



CNN descriptor encoding
(512D)

$$\cos\varphi = \sum_{i=1}^{N} x_i y_i$$

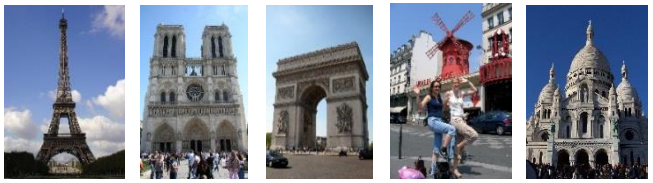
Nearest neighbors used on CNN descriptors

Can use any fast NN search, like ANN

Experiments – datasets

- **Oxford 5k dataset**
  [Philbin et al. CVPR'07]

- **Paris 6k dataset**
  [Philbin et al. CVPR'08]



- **Holidays dataset**
  [Jegou et al. ECCV'10]

- **100k distractor dataset**
  [Philbin et al. CVPR'07]

Training 3D models do not contain any landmark from these datasets

- **Protocol:** mean Average Precision (mAP)

Experiments – Learning (AlexNet)

- Careful choice of **positive** and **negative** training images makes a difference

Our learned whitening

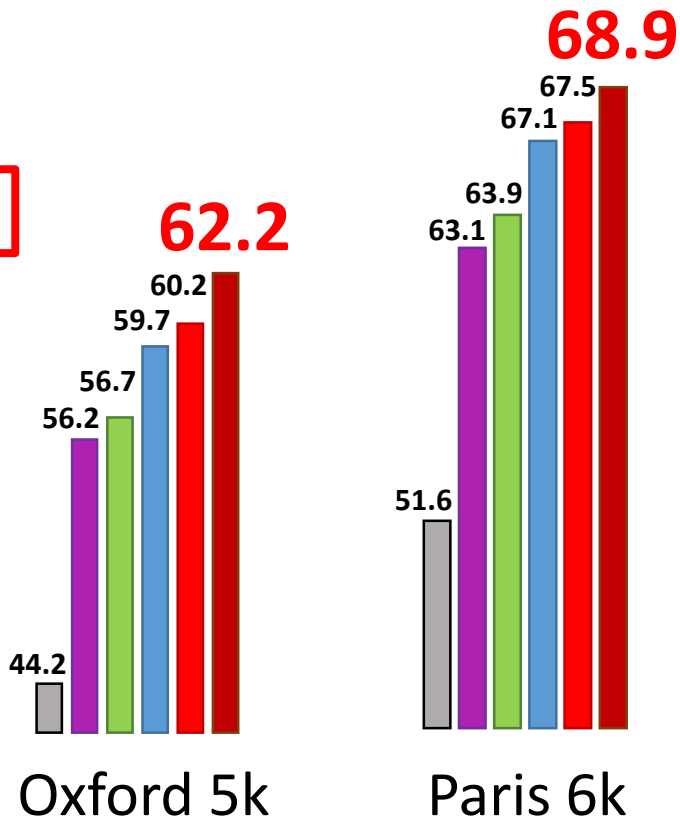
random(top k BoW) + top 1 / model CNN

top 1 BoW + top 1 / model CNN

top 1 CNN + top 1 / model CNN

top 1 CNN + top k CNN

Off-the-shelf



Experiments – Over-fitting and Generalization

- We added Oxford and Paris landmarks as 3D models and repeated fine-tuning

Only +0.3 mAP on average over all testing datasets

State-of-the-art

NetVLAD 256D

vs.

Our CNN 32D

Concurrent work:
[Gordo et al. ECCV'16]

Method		D	Oxf5k		Oxf105k		Par6k		Par106k		Hol	Hol
			$\text{Crop}_{\mathcal{I}}$	$\text{Crop}_{\mathcal{X}}$	$\text{Crop}_{\mathcal{I}}$	$\text{Crop}_{\mathcal{X}}$	$\text{Crop}_{\mathcal{I}}$	$\text{Crop}_{\mathcal{X}}$	$\text{Crop}_{\mathcal{I}}$	$\text{Crop}_{\mathcal{X}}$		101k
Compact representations												
mVoc/BoW [11]		128	48.8	–	41.4	–	–	–	–	–	65.6	–
Neural codes† [14]	(**fA**)	128	–	**55.7**	–	**52.3**	–	–	–	–	**78.9**	–
MAC‡	(**V**)	128	53.5	**55.7**	43.8	45.6	69.5	**70.6**	53.4	**55.4**	72.6	**56.7**
CroW [24]	(**V**)	128	**59.2**	–	**51.6**	–	**74.6**	–	**63.2**	–	–	–
⋆ MAC	(**fV**)	128	**75.8**	**76.8**	**68.6**	**70.8**	77.6	78.8	68.0	69.0	73.2	58.8
⋆ R-MAC	(**fV**)	128	72.5	76.7	64.3	69.7	**78.5**	**80.3**	**69.3**	**71.2**	**79.3**	**65.2**
MAC‡	(**V**)	256	54.7	56.9	45.6	47.8	71.5	72.4	55.7	**57.3**	76.5	**61.3**
SPoC [23]	(**V**)	256	–	53.1	–	**50.1**	–	–	–	–	80.2	–
R-MAC [25]	(**A**)	256	56.1	–	47.0	–	72.9	–	60.1	–	–	–
CroW [24]	(**V**)	256	**65.4**	–	**59.3**	–	**77.9**	–	**67.8**	–	83.1	–
NetVlad [35]	(**V**)	256				–	–	67.7	–	–	**86.0**	–
NetVlad [35]	(**fV**)	25				–	–	**73.5**	–	–	84.3	–
⋆ MAC	(**fA**)	256	6			58.0	68.9	72.2	54.7	58.5	76.2	63.8
⋆ R-MAC	(**fA**)	256	62.5	68.9	53.2	61.2	74.4	76.6	61.8	64.8	81.5	**70.8**
⋆ MAC	(**fV**)	256	**77.4**	**78.2**	**70.7**	**72.6**	80.8	81.9	72.2	73.4	77.3	62.9
⋆ R-MAC	(**fV**)	256	74.9	**78.2**	67.5	72.1	**82.3**	**83.5**	**74.1**	**75.6**	81.4	69.4
MAC‡	(**V**)	512	56.4	**58.3**	47.8	**49.2**	72.3	**72.6**	58.0	**59.1**	76.7	**62.7**
R-MAC [25]	(**V**)	512	66.9	–	61.6	–	**83.0**	–	**75.7**	–	–	–
CroW [24]	(**V**)	512	**68.2**	–	**63.2**	–	79.6	–	71.0	–	84.9	–
⋆ MAC	(**fV**)	512	**79.7**	80.0	**73.9**	**75.1**	82.4	82.9	74.6	75.3	79.5	67.0
⋆ R-MAC	(**fV**)	512	77.0	**80.1**	69.2	74.1	**83.8**	**85.0**	**76.4**	**77.9**	82.5	**71.5**
Extreme short codes												
Neural codes† [14]	(**fA**)	16	–	**41.8**	–	**35.4**	–	–	–	–	**60.9**	–
⋆ MAC	(**fV**)	16	**56.2**	**57.4**	**45.5**	**47.6**	57.3	62.9	43.4	48.5	51.3	25.6
⋆ R-MAC	(**fV**)	16	46.9	52.1	37.9	41.6	**58.8**	**63.2**	**45.6**	**49.6**	54.4	**31.7**
Neural codes† [14]	(**fA**)	32				**46.7**	–	–	–	–	**72.9**	–
⋆ MAC	(**fV**)	32				**59.5**	**63.9**	**69.5**	51.6	**56.3**	62.4	41.8
⋆ R-MAC	(**fV**)	32				55.1	**63.9**	67.4	**52.7**	55.8	68.0	**49.6**
Re-ranking (R) and query expansion (QE)												
BoW(1M)+QE [6]		–	82.7	–	76.7	–	80.5	–	71.0	–	–	–
BoW(16M)+QE [50]		–	84.9	–	79.5	–	82.4	–	77.3	–	–	–
HQE(65k) [8]		–	**88.0**	–	**84.0**	–	82.8	–	–	–	–	–
R-MAC+R+QE [25]	(**V**)	512	77.3	–	73.2	–	**86.5**	–	**79.8**	–	–	–
CroW+QE [24]	(**V**)	512	72.2	–	67.8	–	85.5	–	79.7	–	–	–
⋆ MAC+R+QE	(**fV**)	512	85.0	**85.4**	81.8	**82.3**	**86.5**	**87.0**	78.8	79.6	–	–
⋆ R-MAC+R+QE	(**fV**)	512	82.9	84.5	77.9	80.4	85.6	86.4	78.3	**79.7**	–	–

63.5

69.2

Teacher vs. Student

Method	Oxf5k	Oxf105k	Par6k	Par106k
BoW(16M)+R+QE	**84.9**	**79.5**	**82.4**	**77.3**
CNN(512D)	79.7	73.9	**82.4**	74.6
CNN(512D)+R+QE	**85.0**	**81.8**	**86.5**	**78.8**

Our CNN with re-ranking (R) and query expansion(QE) surpasses its teacher on all datasets!!!

Teacher vs. Student

query

BoW

top 10 (correct | incorrect)

first **incorrect** at rank 127

CNN

Teacher vs. Student

CNN descriptors

Significant viewpoint scale change	lots of training data
Significant illumination change	lots of training data
Severe occlusions	lots of training data
Visually similar but different objects	lots of training data

CNN descriptors

Significant viewpoint scale change	lots of training data
Significant illumination change	lots of training data
Severe occlusions	lots of training data
Visually similar but different objects	lots of training data

versus

Bag of Words

Significant viewpoint scale change	covariant local features, invariant descriptors
Significant illumination change	color-normalized feature descriptors
Severe occlusions	locality of the features, geometric verification
Visually similar but different objects	Feature discriminability & geometric verification

CNN descriptor learning

- Proposed a method to generate the necessary "lots of training examples" without any human interaction
- Strong supervision for hard negative, hard positive mining, and supervised whitening
- Data and trained networks available at: cmp.felk.cvut.cz/~radenfil/projects/siamac.html
- For more details about the paper visit **Poster O-1A-01**

So Is the Bag-of-Words REALLY torn?

Click Here

So is the Bag-of-Words REALLY torn?

Not yet, but don’t mess with tape ;)

Questions?

- Thanks for your attention

- Interested students should ask about our PhD program

Center for Machine Perception

Czech Technical University in Prague

http://cmp.felk.cvut.cz

Contact Jiri Matas or Ondrej Chum