Large Scale Image Retrieval

Jiří Matas

Center for Machine Perception,Department of Cybernetics,
Faculty of Electrical Engineering,
Czech Technical University in Prague

work done as collaboration of :
Ondrej Chum, Andrej Mikulik, Filip Radenovic,
Michal Perdoch James Pritts, Dmytro Mishkin,
Johannes Schönberger, Jan-Michael Frahm, Jan Cech

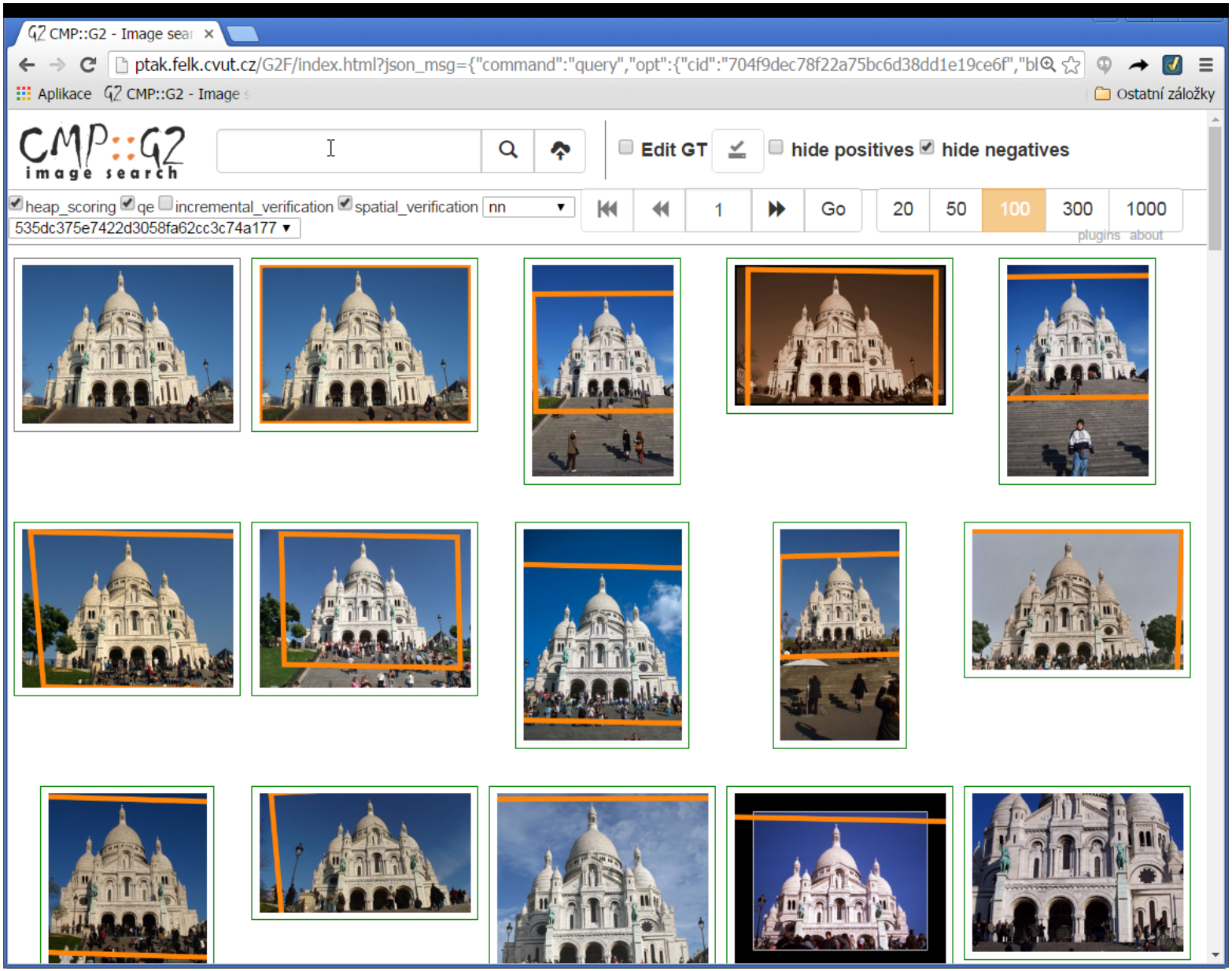
Supported by the Czech Science Foundation Project GACR P103/12/G084

Outline

1. The standard retrieval problem:
   - *The visually most similar* - *All visually similar*
2. The standard method - Bag of Words with refinements: query expansion, context, geometric verification

3. Beyond similarity retrieval – obtaining new information
   - What is this? Where is this?
   - What is interesting here?
   - Where should I look?
   - Show me different appearances of the same scene
4. Application in 3D reconstruction.
5. The challenges in visual object retrieval.
6. Conclusions.

Standard Image Retrieval



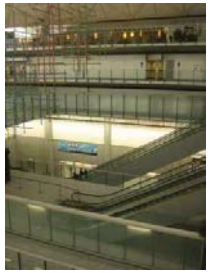
CMP Image Retrieval 2.0 Live Demo

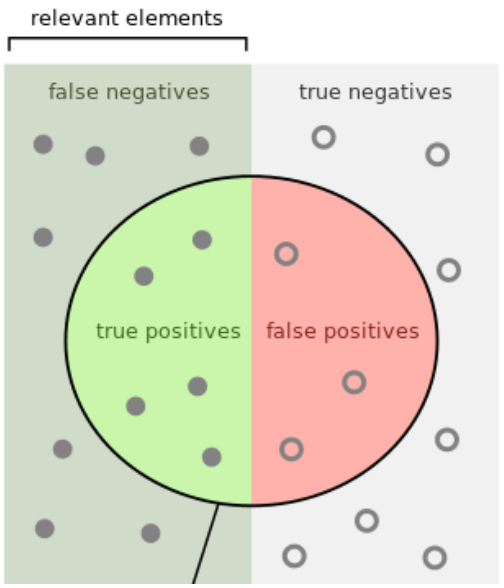
Standard Image Retrieval Evaluation

Query:

- 10 database images
- 5 relevant images

Ranking:



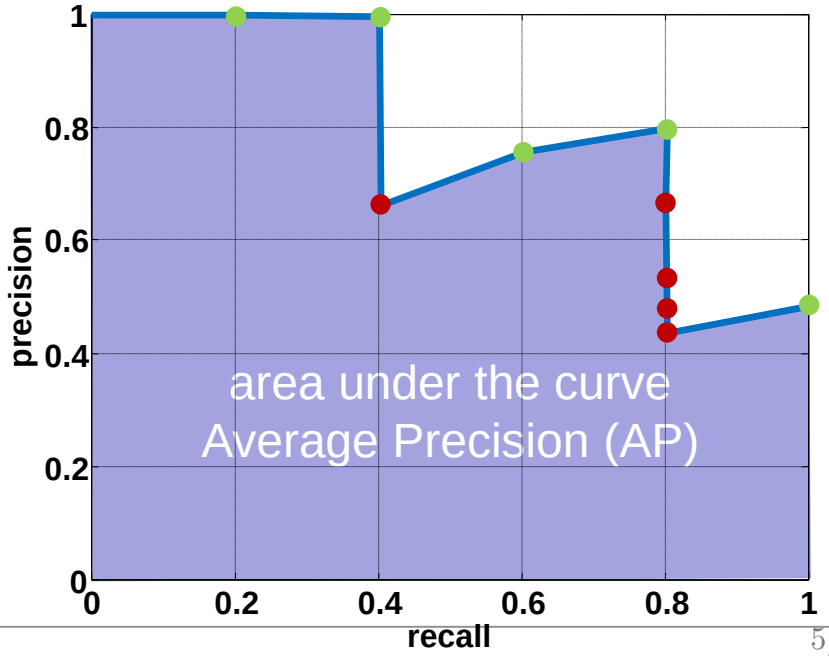


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$



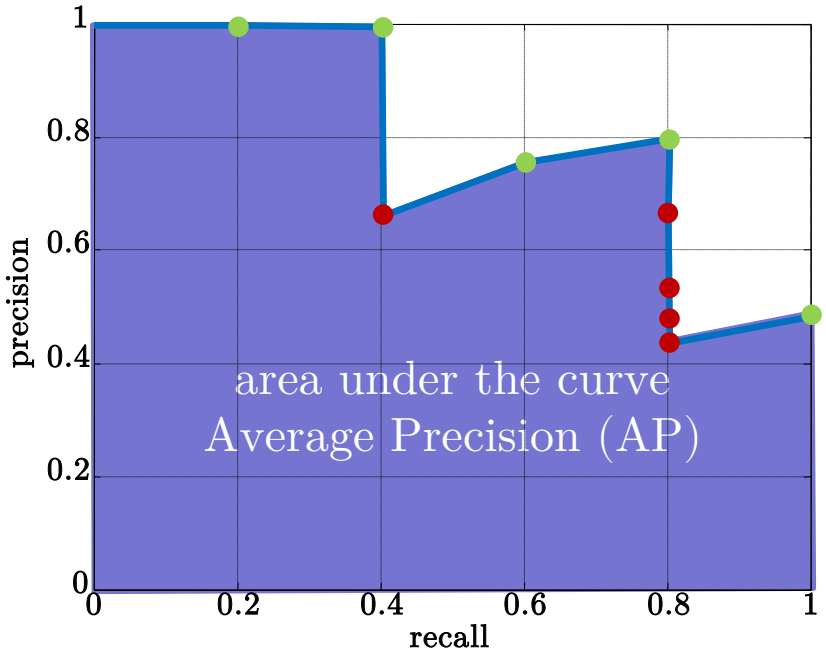
Standard Image Retrieval - Evaluation

1. recall and precision

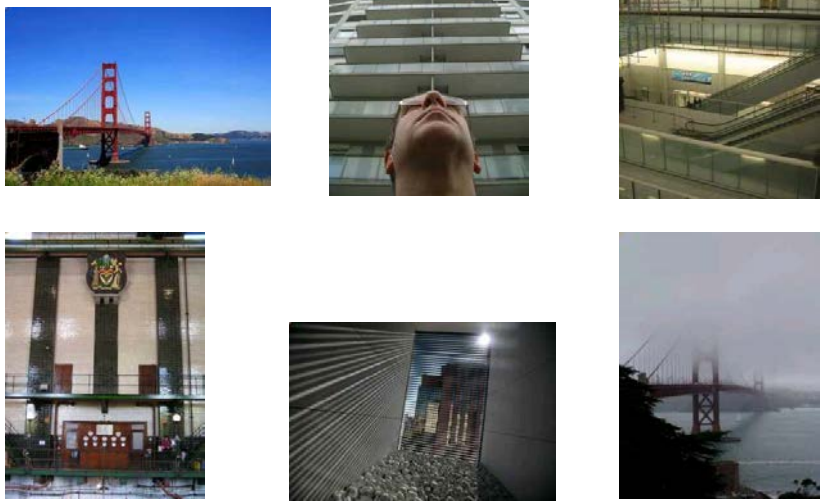
Query

Database size: 10 images
Relevant (total): 5 images

precision = #relevant / #returned
recall = #relevant / #total relevant

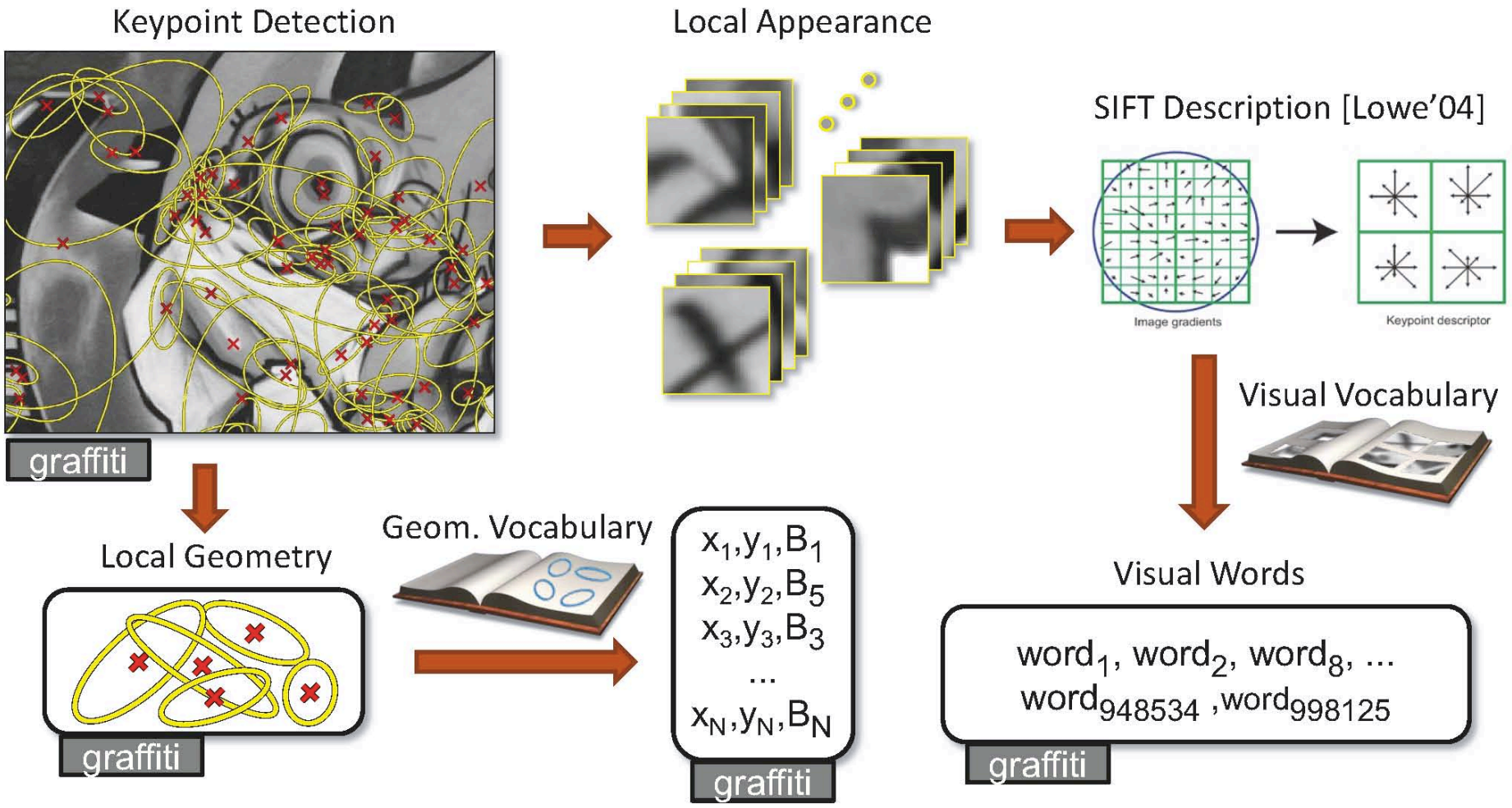


Results (ordered):

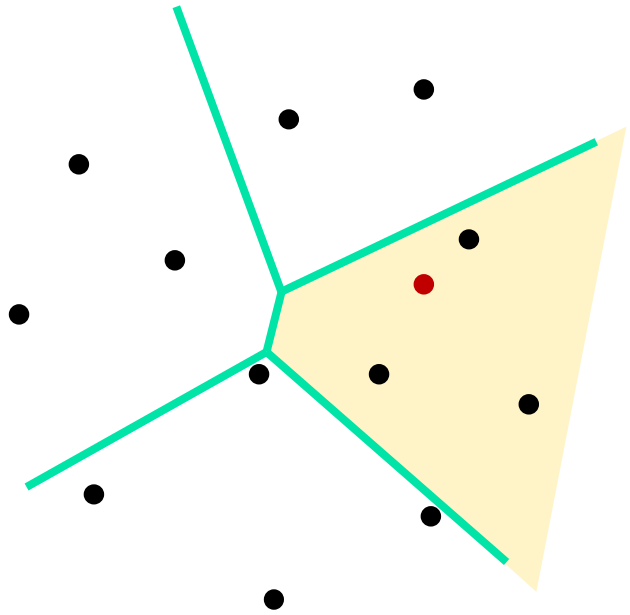


Bag-of-Words (BoW): Off-line Stage

Bag-of-Words is the standard image representation for specific object retrieval



Building a Visual Vocabulary



Feature distance
0 : features in the same cell
∞ : features in different cells

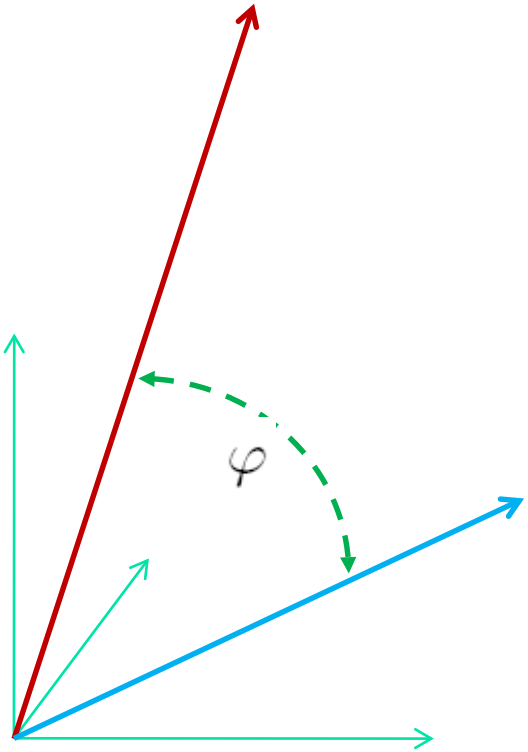
+ most of the features are not considered (infinitely distant)
+ near-by descriptors accessible instantly – storing a list of features for each cell

Partition the feature space

(k – means clustering)

- extremely time consuming (k=10^7, |dataset|>10^{10})
- approximate k-means used (**FLANN**)
- L_2 most common (for computational convenience), other options L_1, Hellinger,

Efficient Scoring



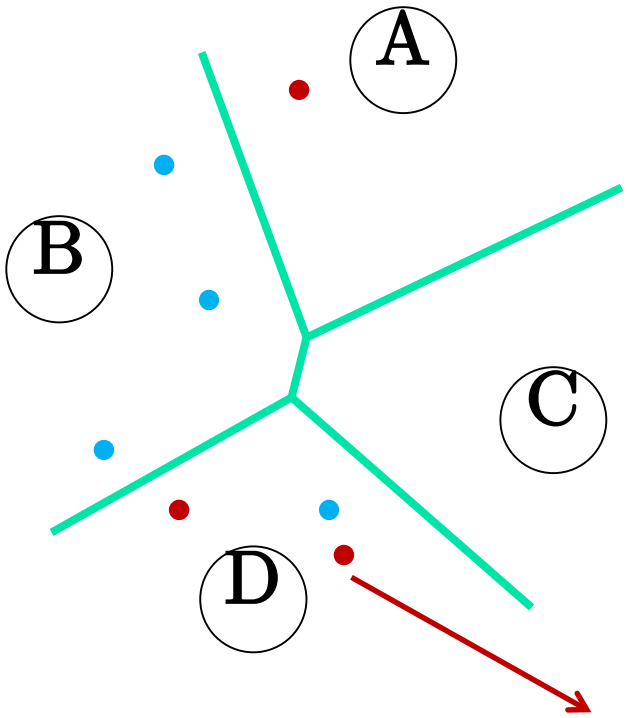
bag of words representation
(up to 1,000,000 D)

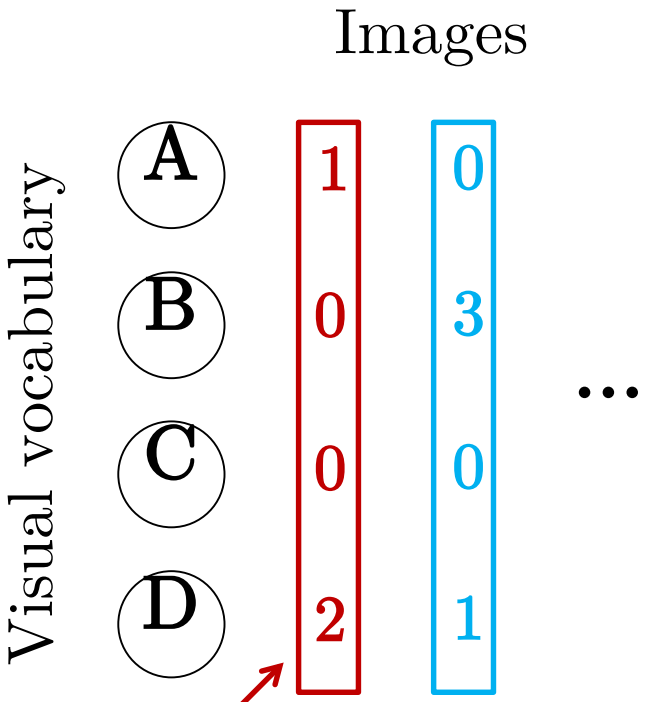
$$\cos\varphi = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} = \frac{1}{||\mathbf{x}||\,||\mathbf{y}||}\sum_{i=1}^{N} x_i y_i$$

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

	Database					Query		Score
	A	B	C	D				
α_1	1	0	0	2		A 0		s_1
α_2	0	2	0	1	• α_q	B 3	=	s_2
α_3	1	0	0	0		C 0		s_3
	⋮					D 1		⋮

Bags of Words – Image Representation





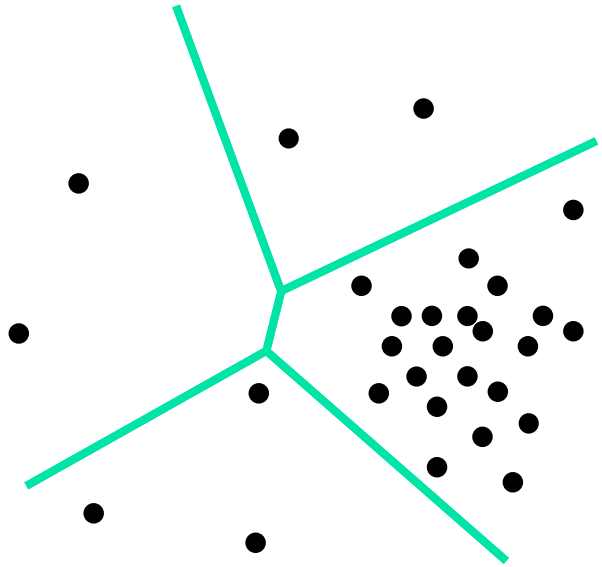
Term-frequency (tf) – visual word D is twice in the image

Images are represented by sparse vector / histogram of visual words present in them

Bag of Words: Weighting words

- Words (in text) common to many documents are less informative – 'the', 'and', 'or', 'in', ...
- 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)
- Images are represented by weighted histograms tf_X idf_X rather than just a histogram of tf_X)

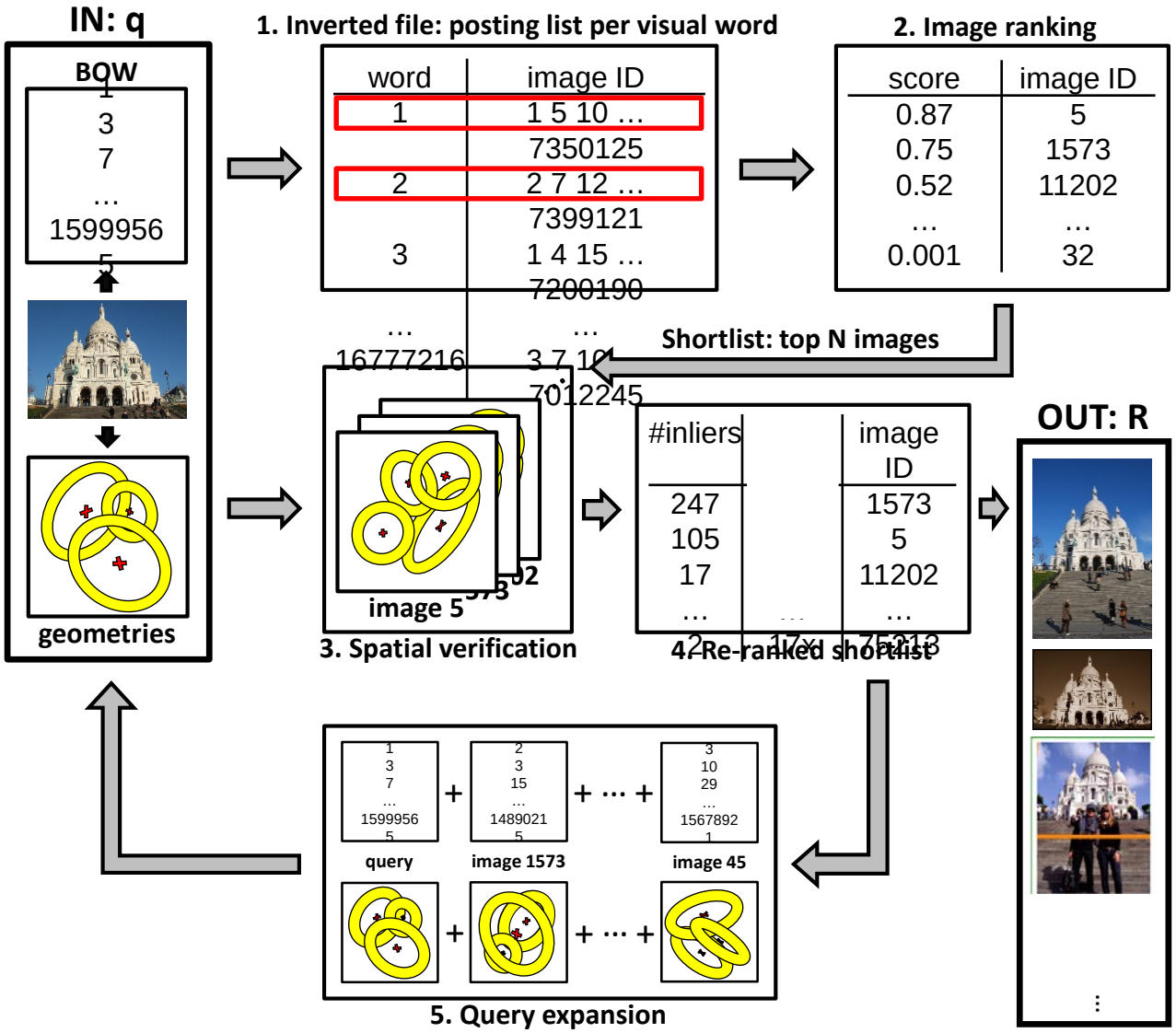
$$idf_X = \log \frac{\#\ \text{documents}}{\#\ \text{docs containing}\ x}$$



features from all documents

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

Bag of Words: Retrieval



Spatial Verification

Re-rank top ranked images (removing false positives)

- RANSAC

NOTE: Standard BoW score ranking performed without geometric information

IMPORTANT: Geometric verification crucial for query expansion

Sivic, Zisserman: Video Google, ICCV 2003

Philbin, Chum, Isard, Sivic, Zisserman: Object retrieval with large vocabularies and fast spatial matching, CVPR'07

Spatial Verification and Query Expansion

Query image

Results

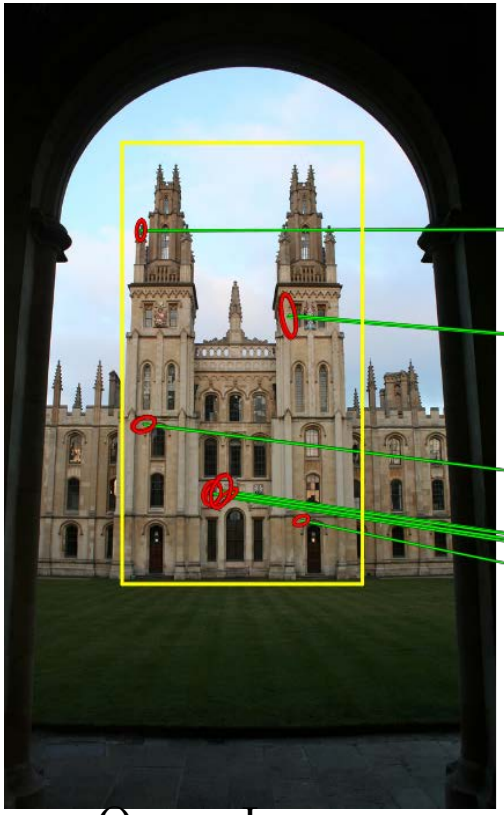
Spatial verification

New query

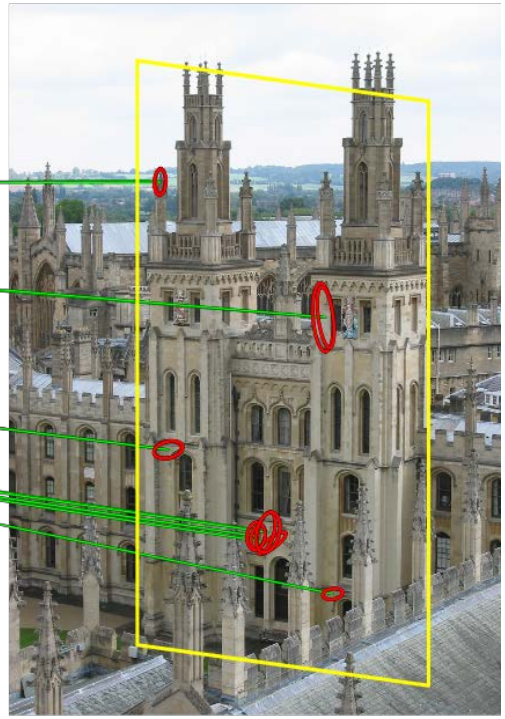
New results

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

Query Expansion Step by Step



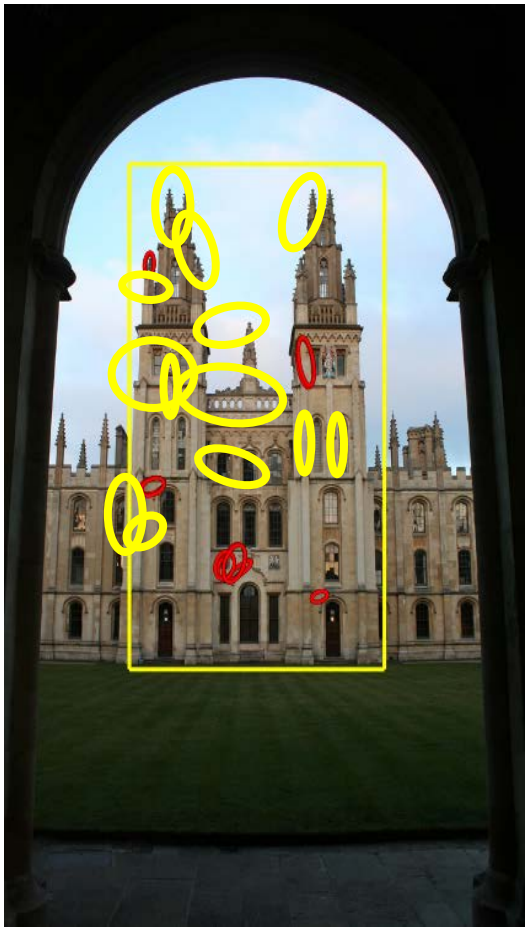
Query Image

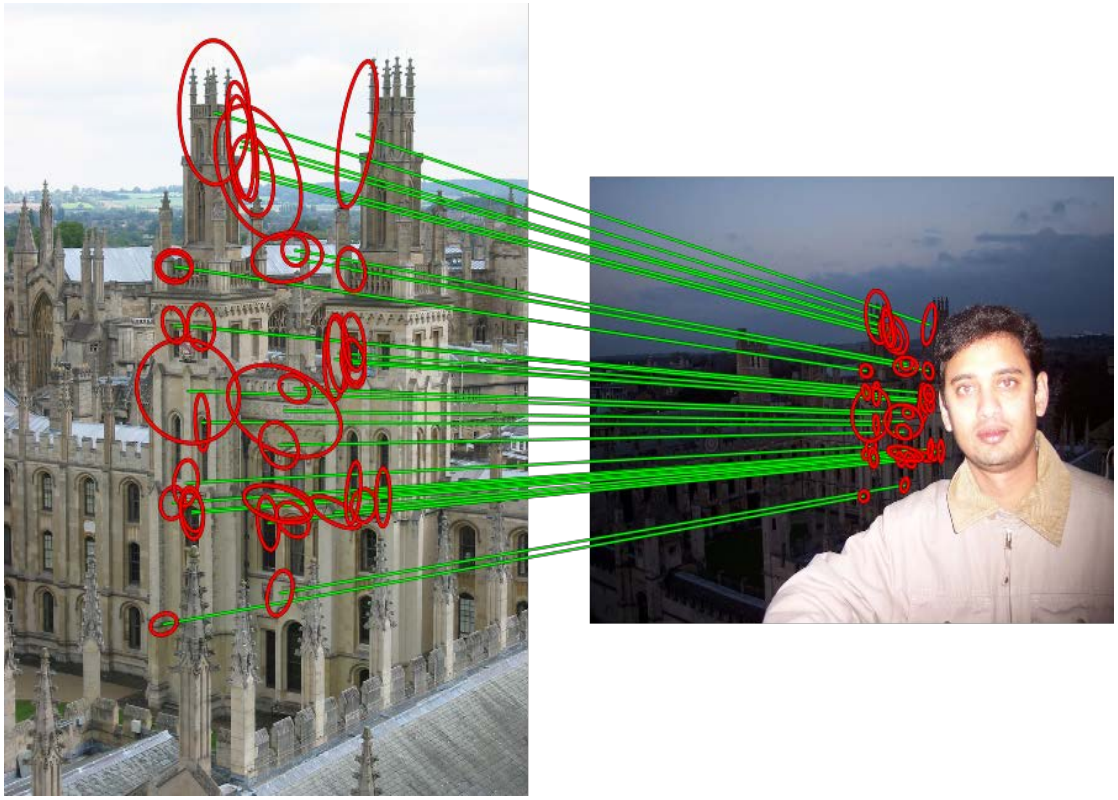


Retrieved image

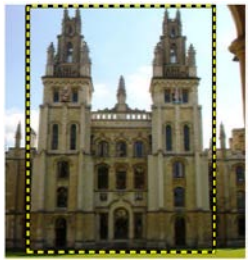
Originally not retrieved

Query Expansion Step by Step



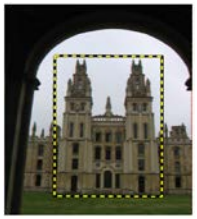
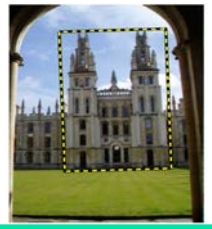


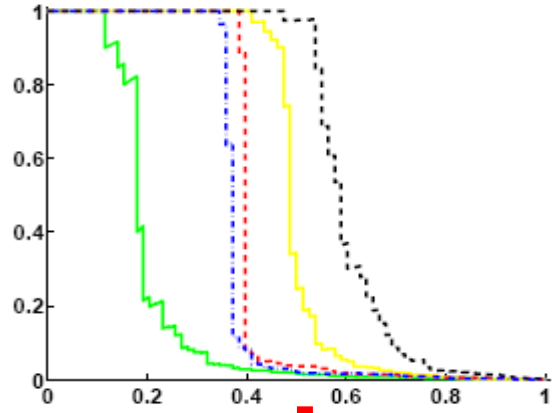
Query Expansion Results



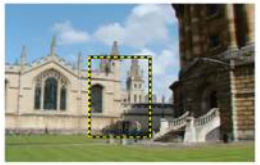
Query image

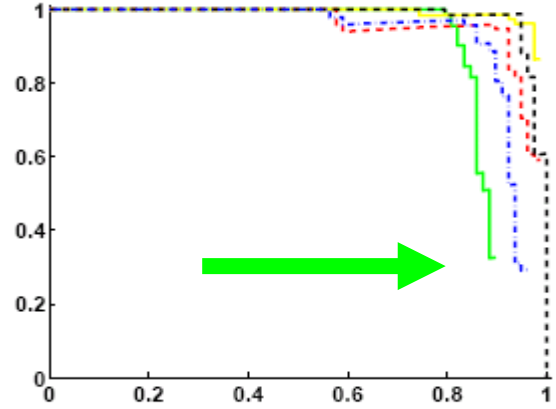
Original results (good)



Expanded results (better)





Is this what we want?

- Visually most similar
  Results identical to query for large datasets

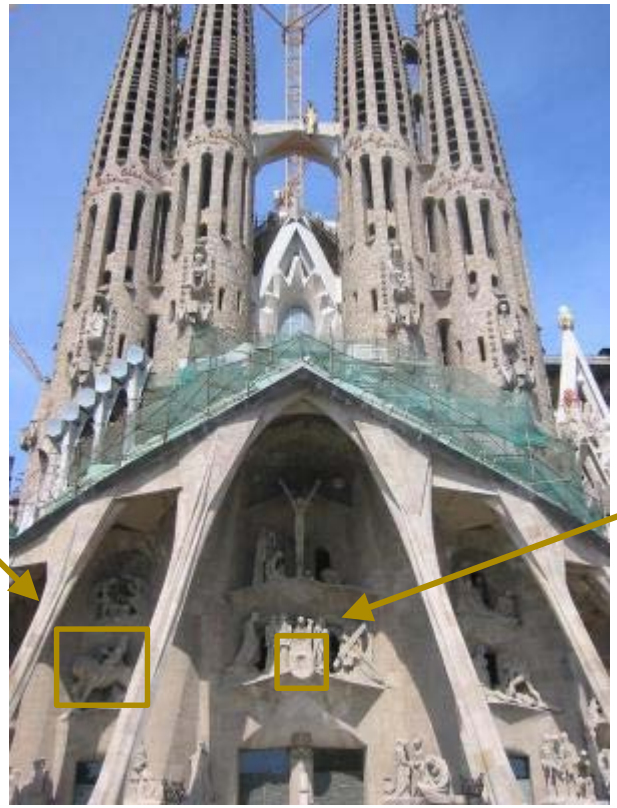
- All visually similar
  Output of varying length
  Ground truth hard to obtain

- Users won't look at tens of near-duplicate images!

Beyond Similarity Retrieval

New Retrieval Problems

What is this?

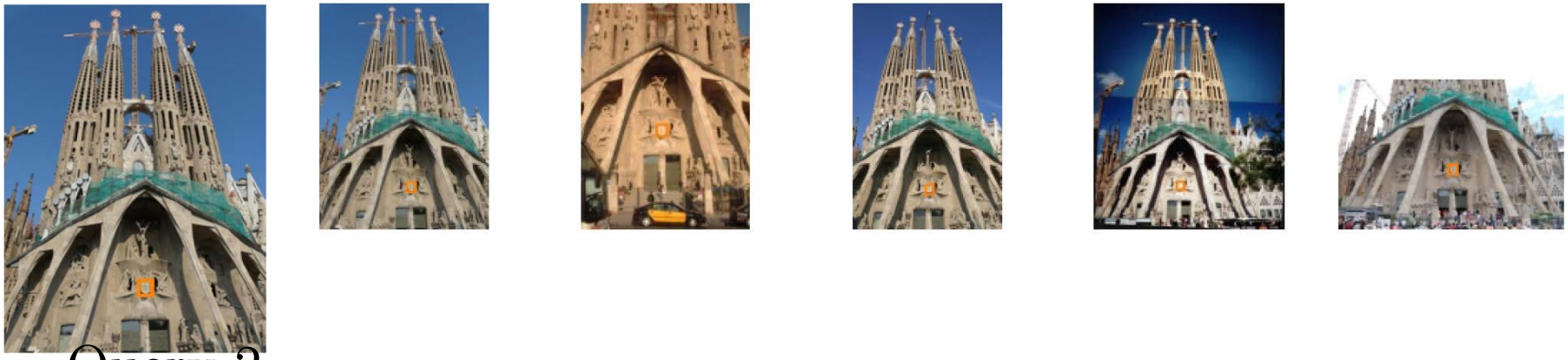


... and what is that?

Let's **zoom-in!**

Standard Retrieval (for browsing, ...

Query 1



Query 2

Standard retrieval: similar results for both queries, no new information

Standard Retrieval and Details

query

rank:

1

2

32

64

65

EASY

query

rank:

1

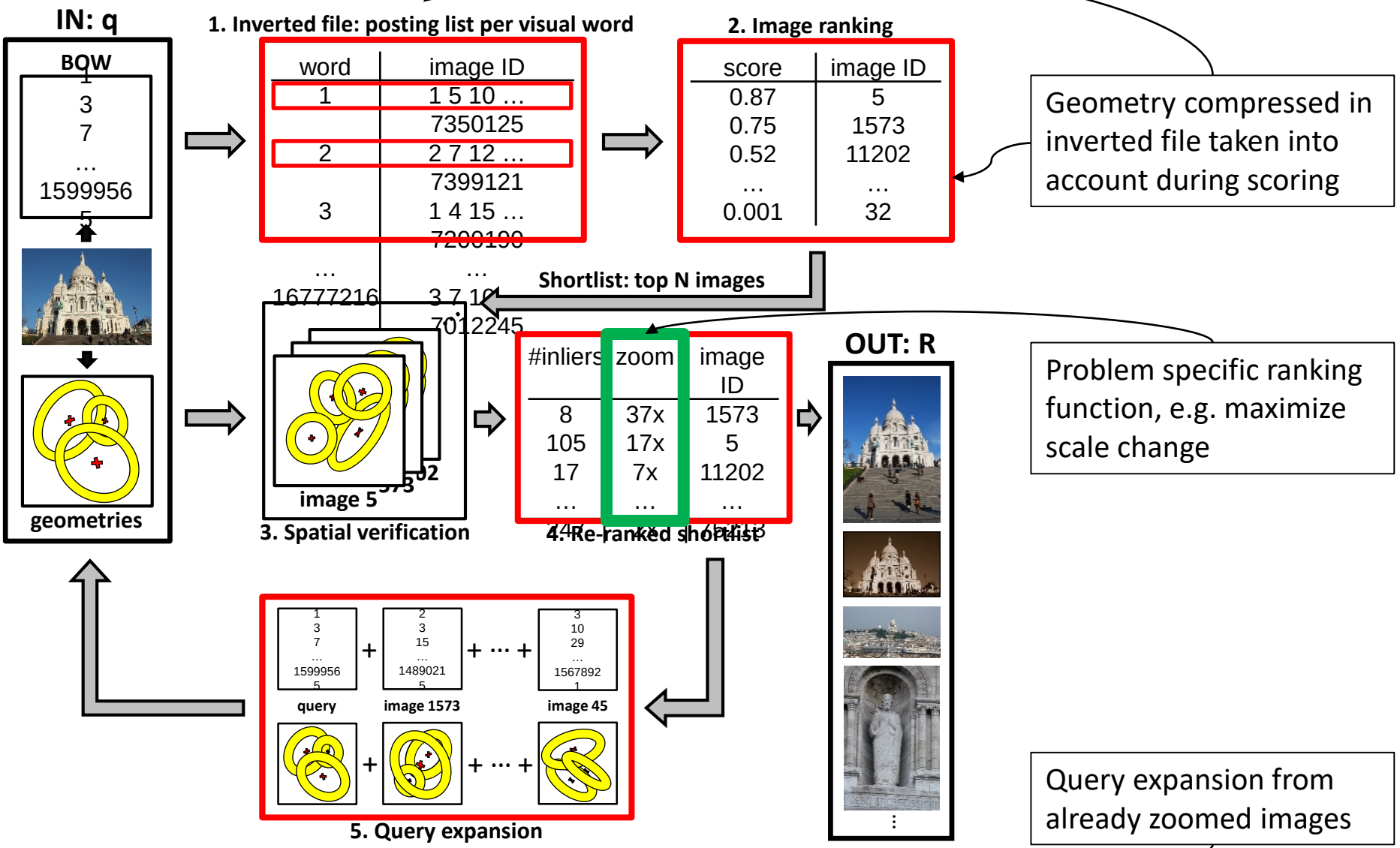
2048

16384

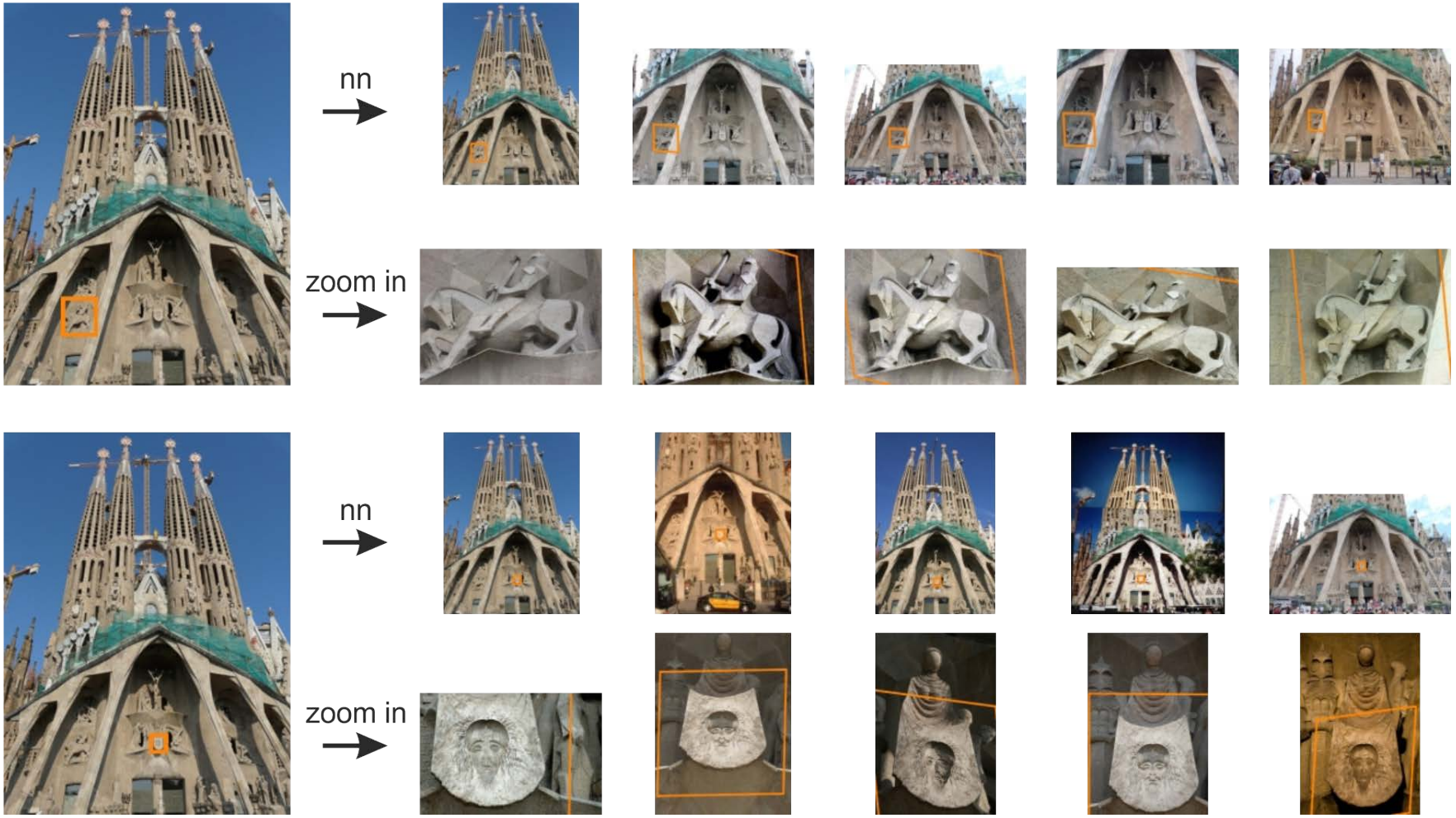
81368

DIFFICULT

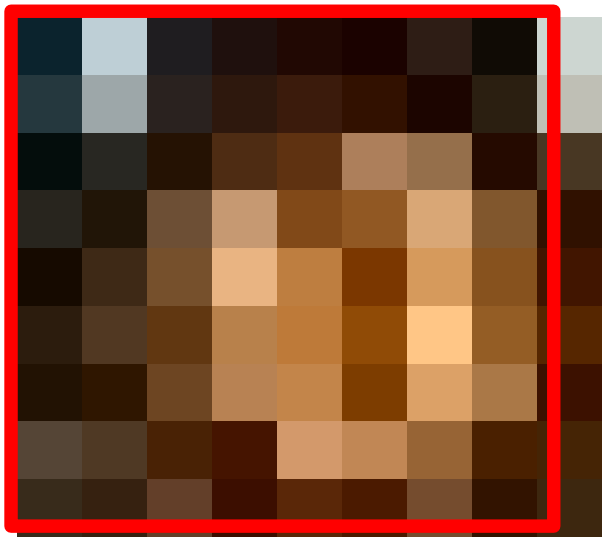
Zoom-in: On-line Stage



“What is this?” examples

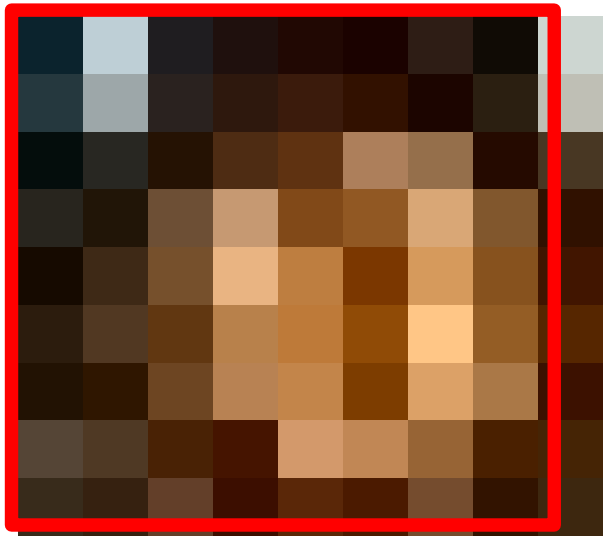


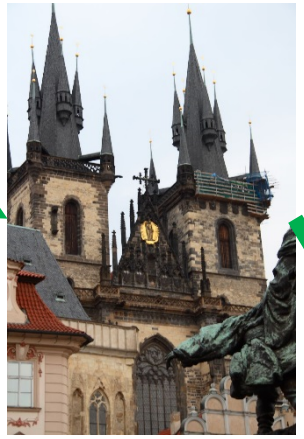
Zoom-in: Example



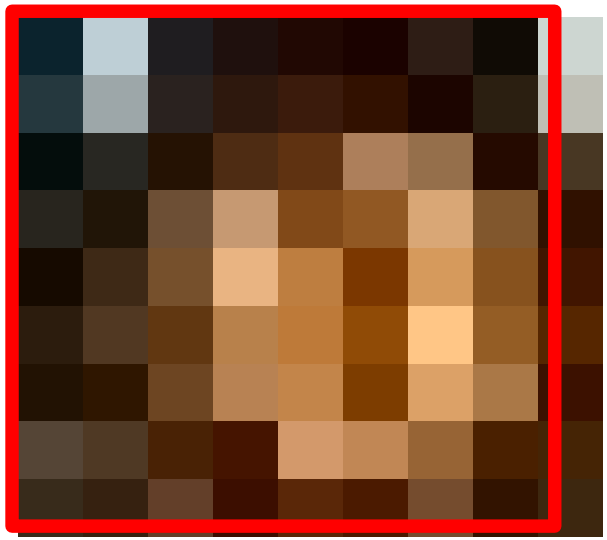
?

Zoom-in: Query Expansion

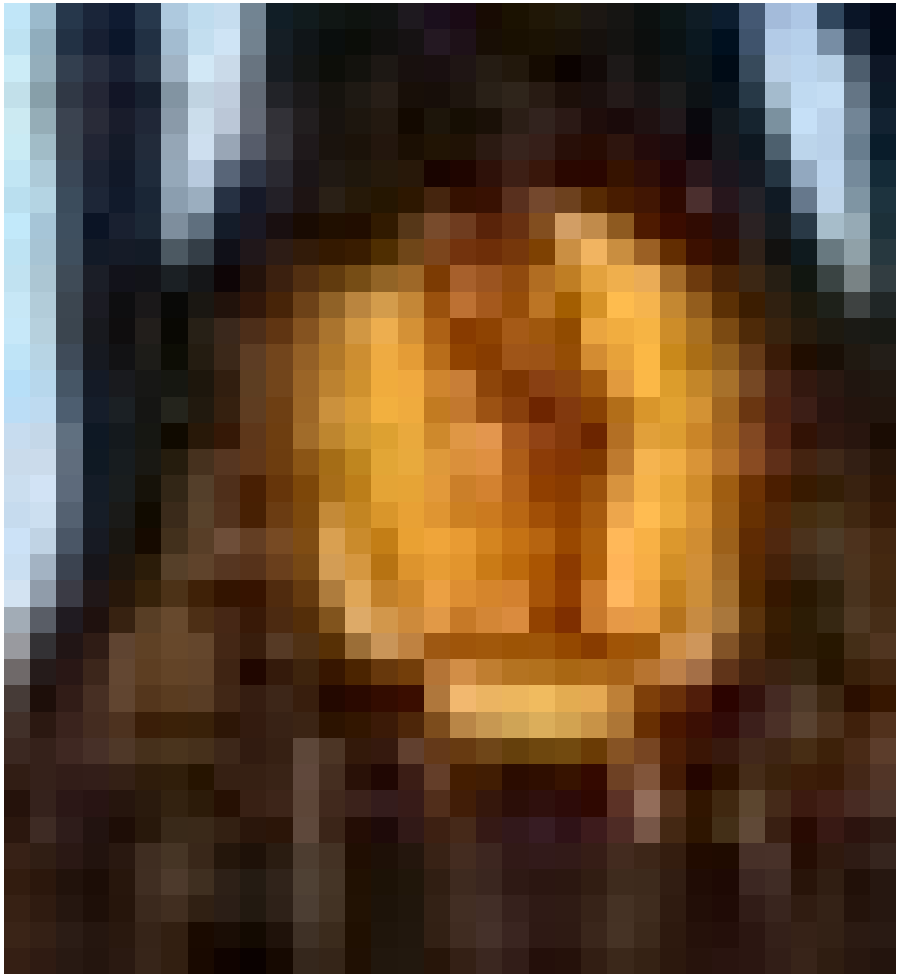


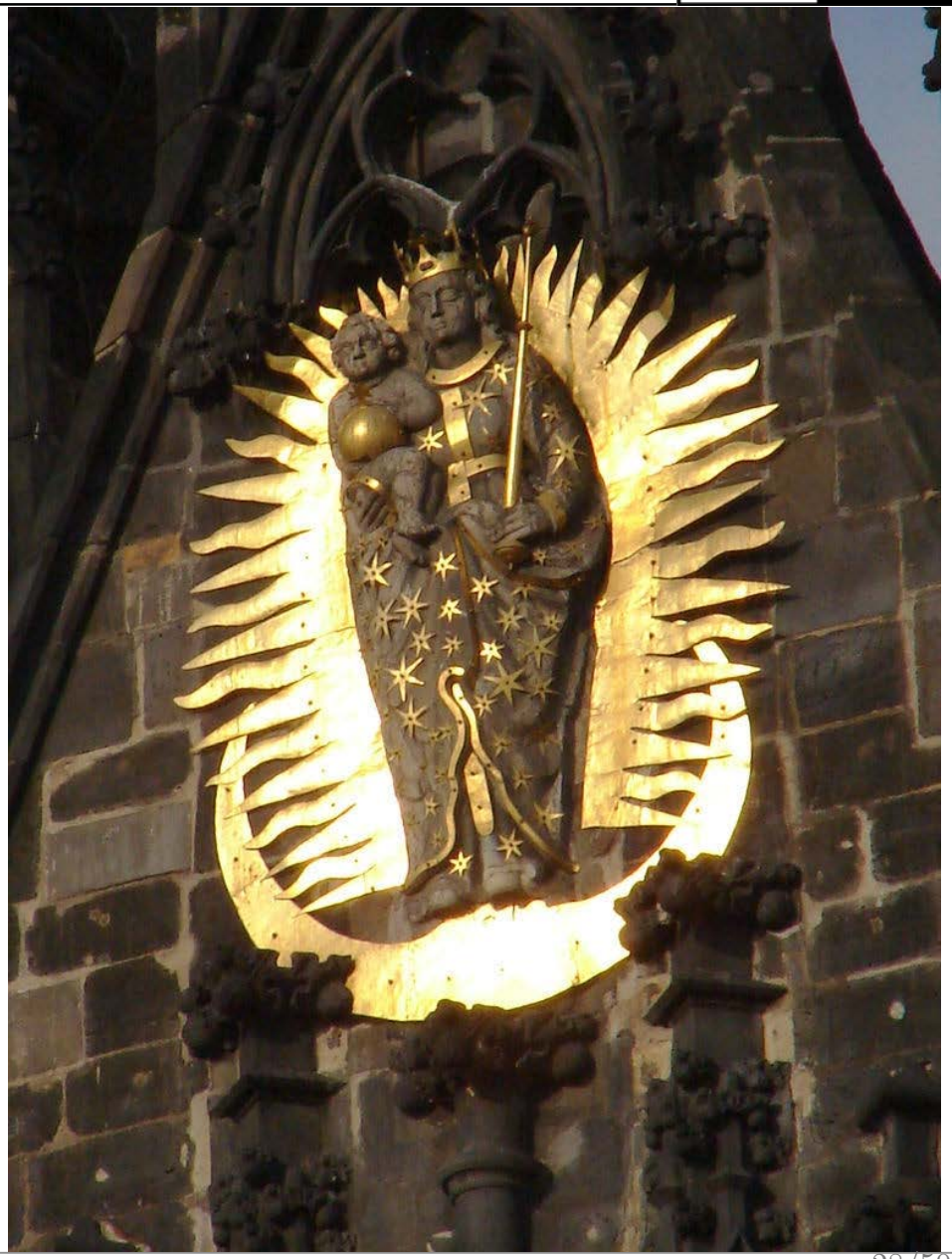


Zoom-in: Example

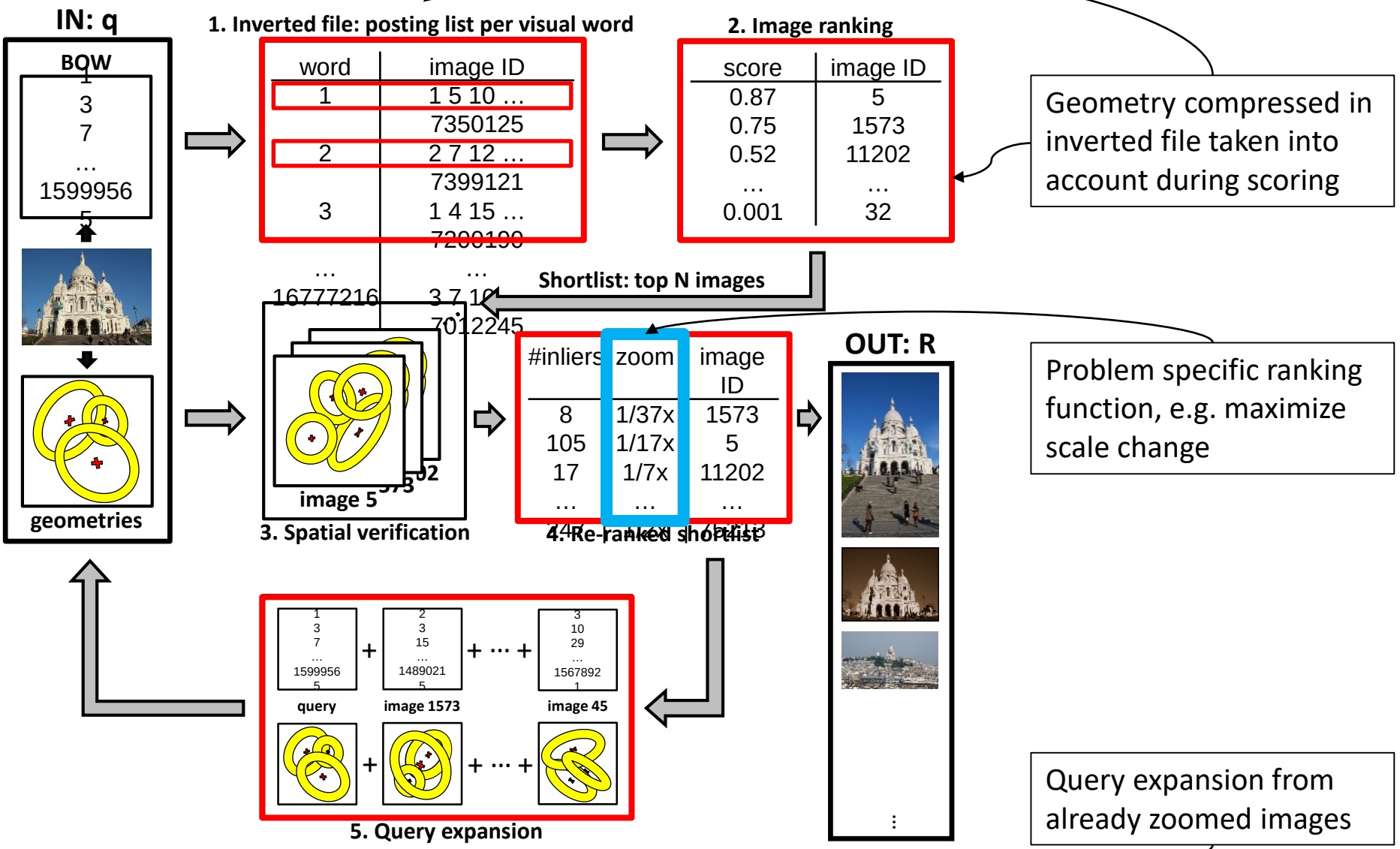


Zoom in examples

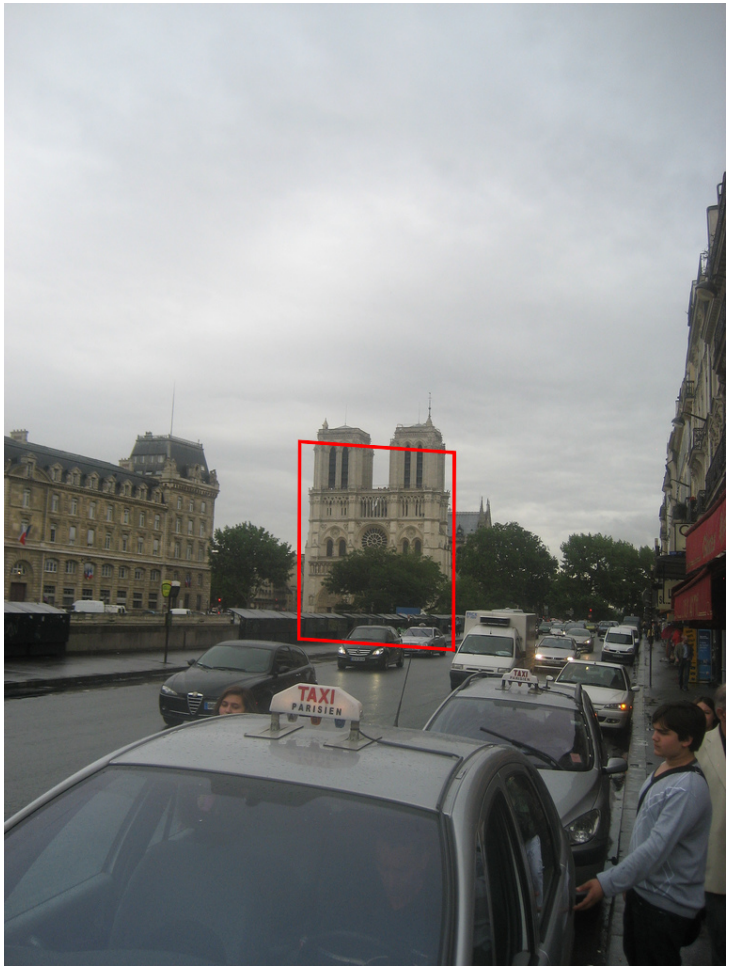




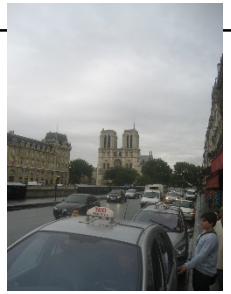
Zoom-out: On-line Stage



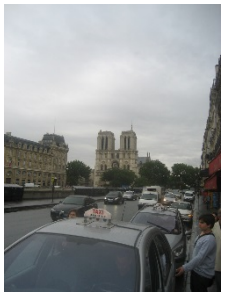
Zoom-out: Iterate

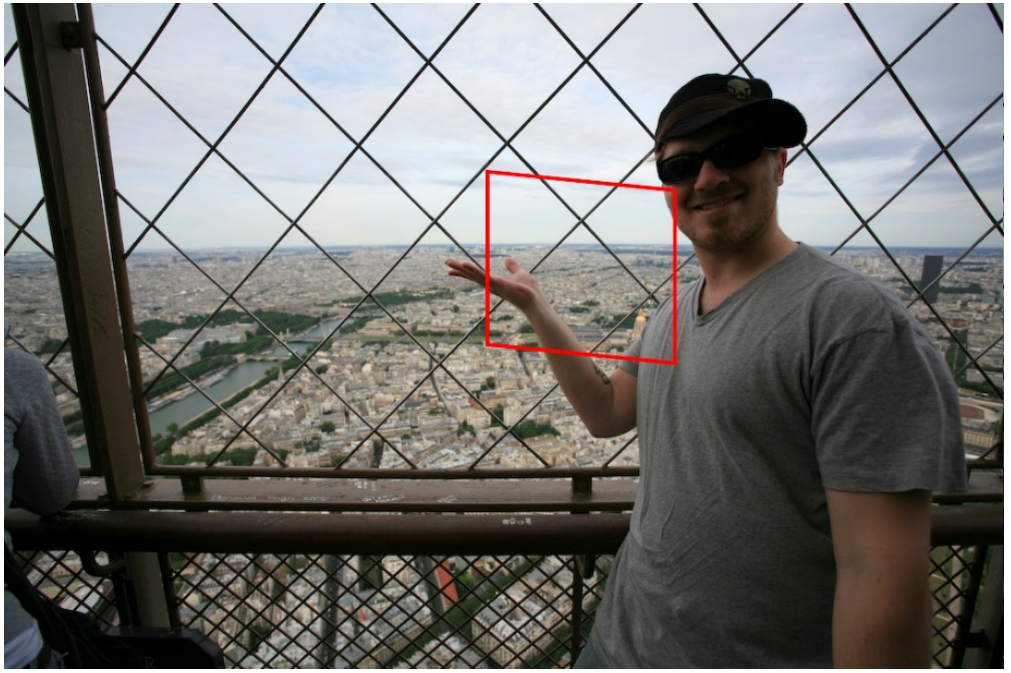


Zoom-out: Iterate



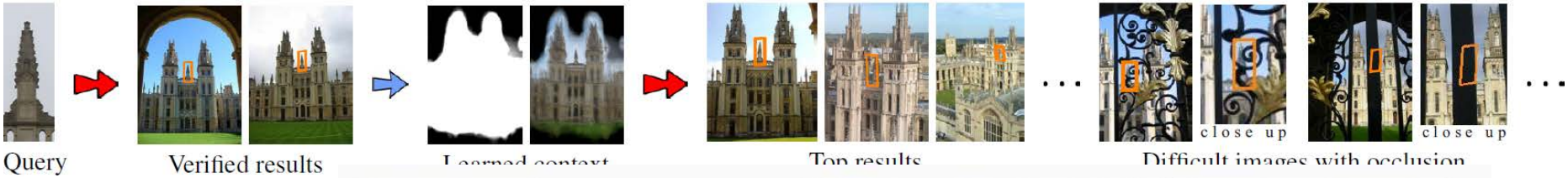
Zoom-out: Iterate





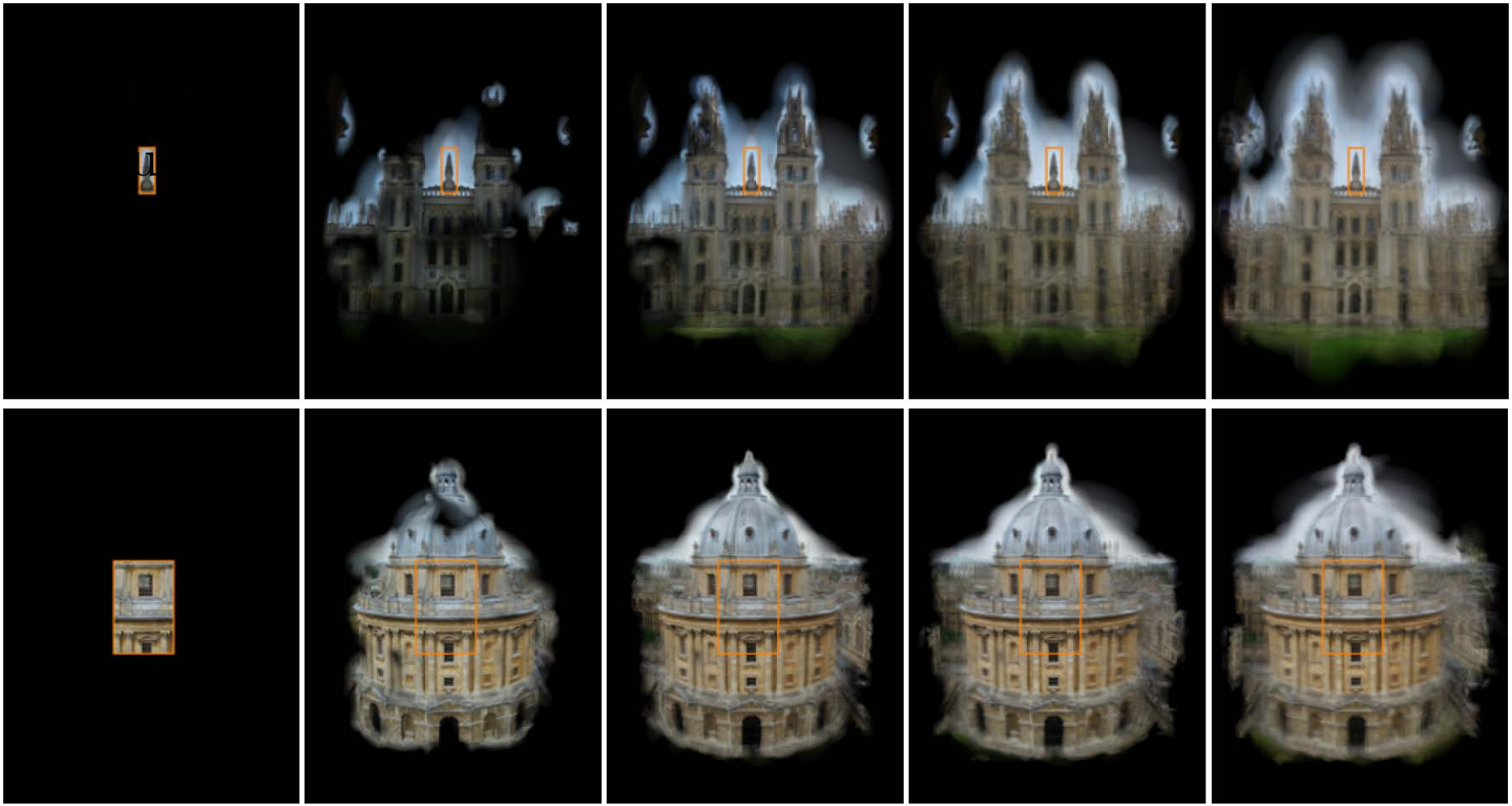
Context expansion

- the model of the object is grown beyond the boundaries of the initial query,
- a feature added into the model that is not inside the context is inactive until confirmed by feature(s) from another image with the same visual word and similar geometry.
- Once a feature is confirmed, it adds the neighbourhood around its center to the context.



Learning the Context

Feature patches back-projected into the context from spatially verified images.

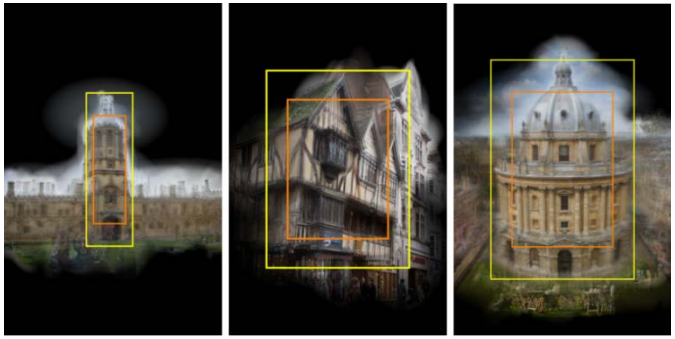


2 5 10 20

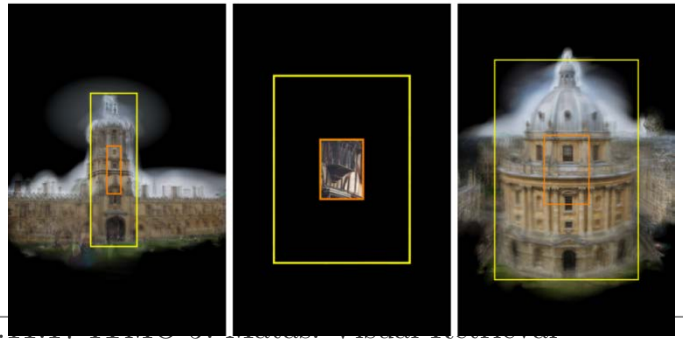
How Much Do We Need to See?

Oxford landmarks – 3 queries
100%, 50%, and 10% of the query bounding box

Context learned from the full bounding box

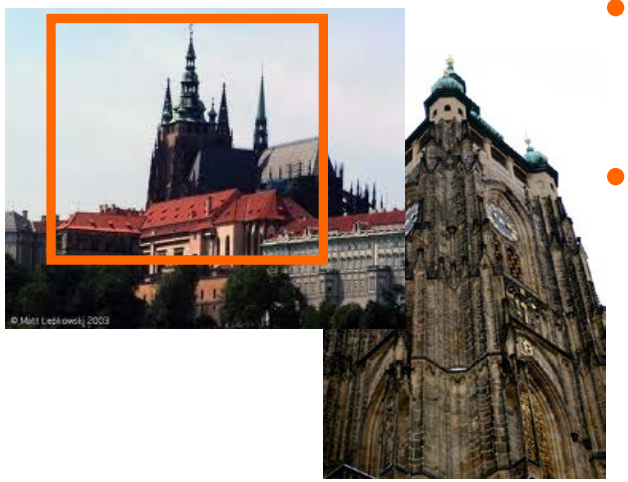


Context learned from 50% of the bounding box

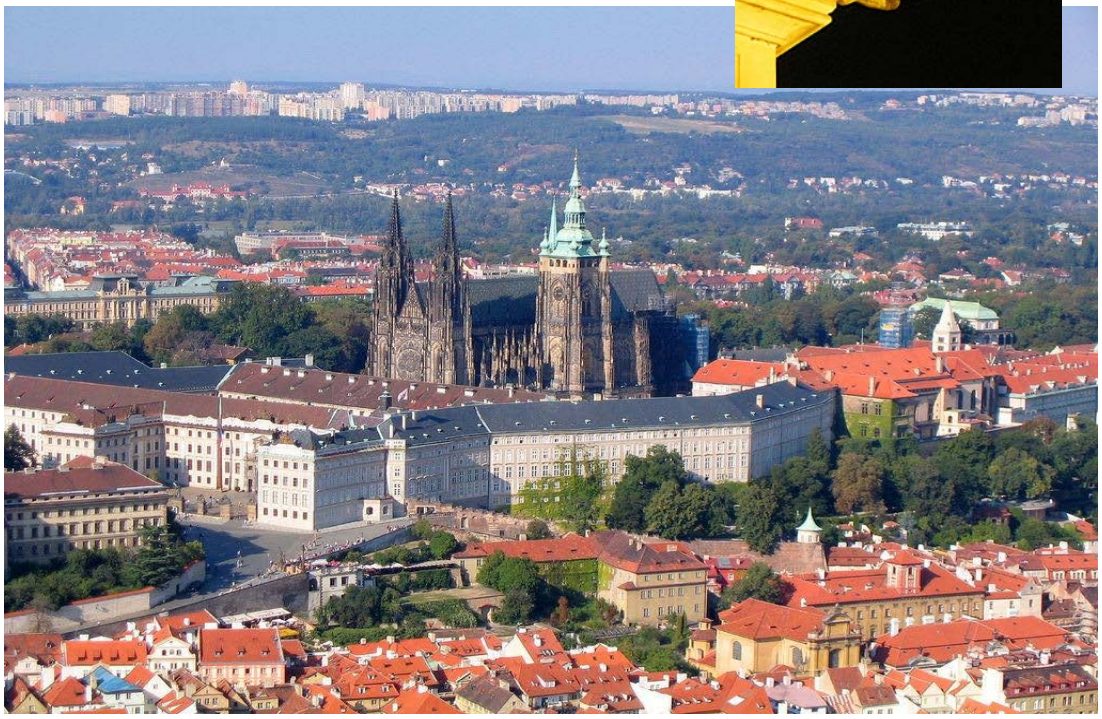


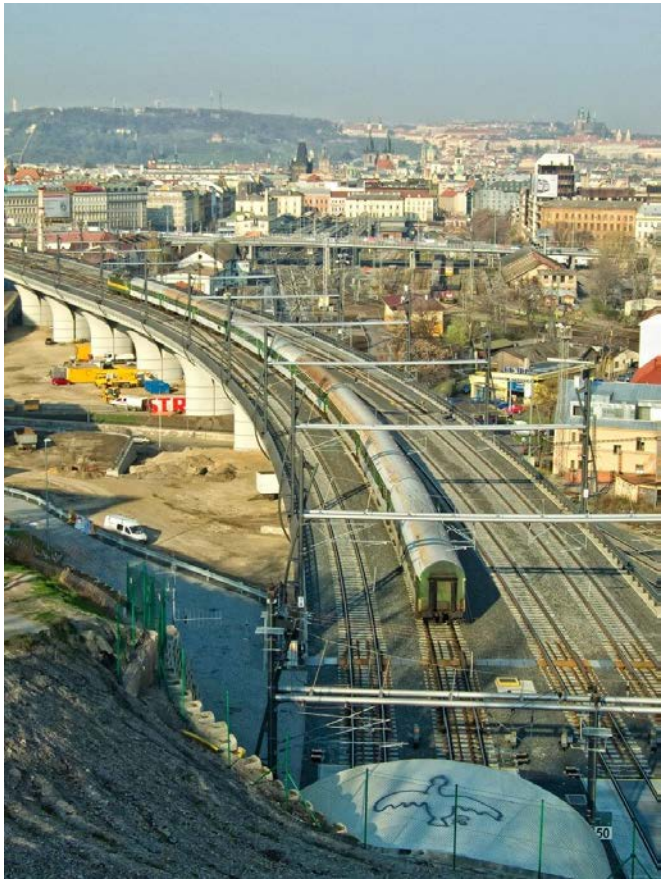
Context learned from 10% of the bounding box

Retrieval with large viewpoint change



- Image retrieval is not an efficient execution of two view matching.
- Significant part is about finding paths, *sequences of matches*



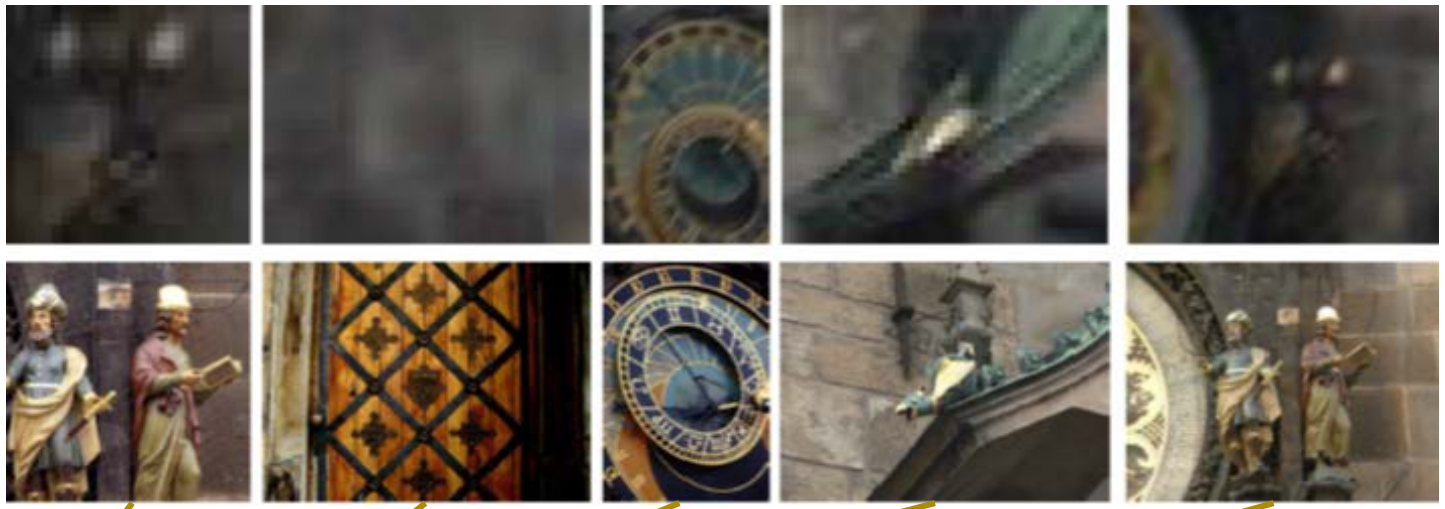


What is interesting here?

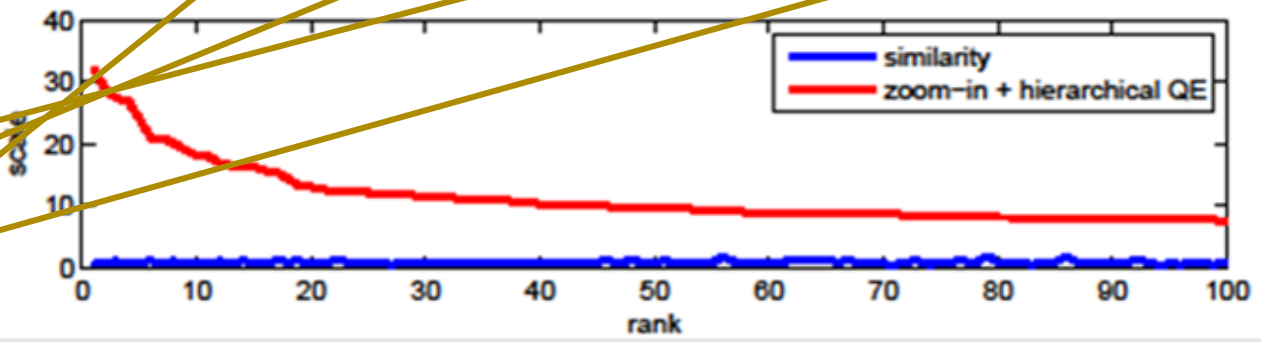
What should you not miss?

Highest Resolution Transform

Given a query and a dataset, for every pixel in the query image:
Find the database image with the maximum resolution depicting the pixel



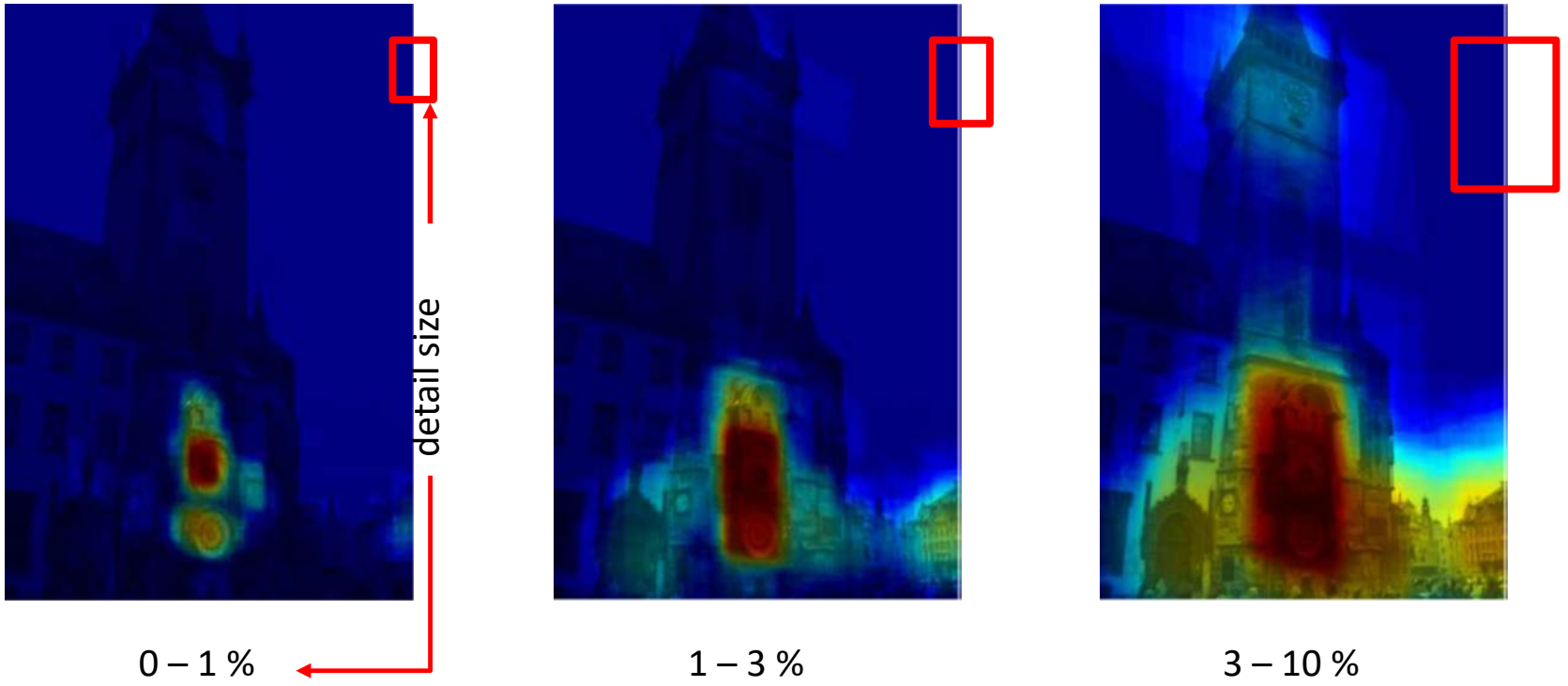
37.3x 27.0x 22.8x 21.9x 21.6x



Mikulík, Radenović, Chum, Matas: **Efficient Image Detail Mining,** ACCV 2014

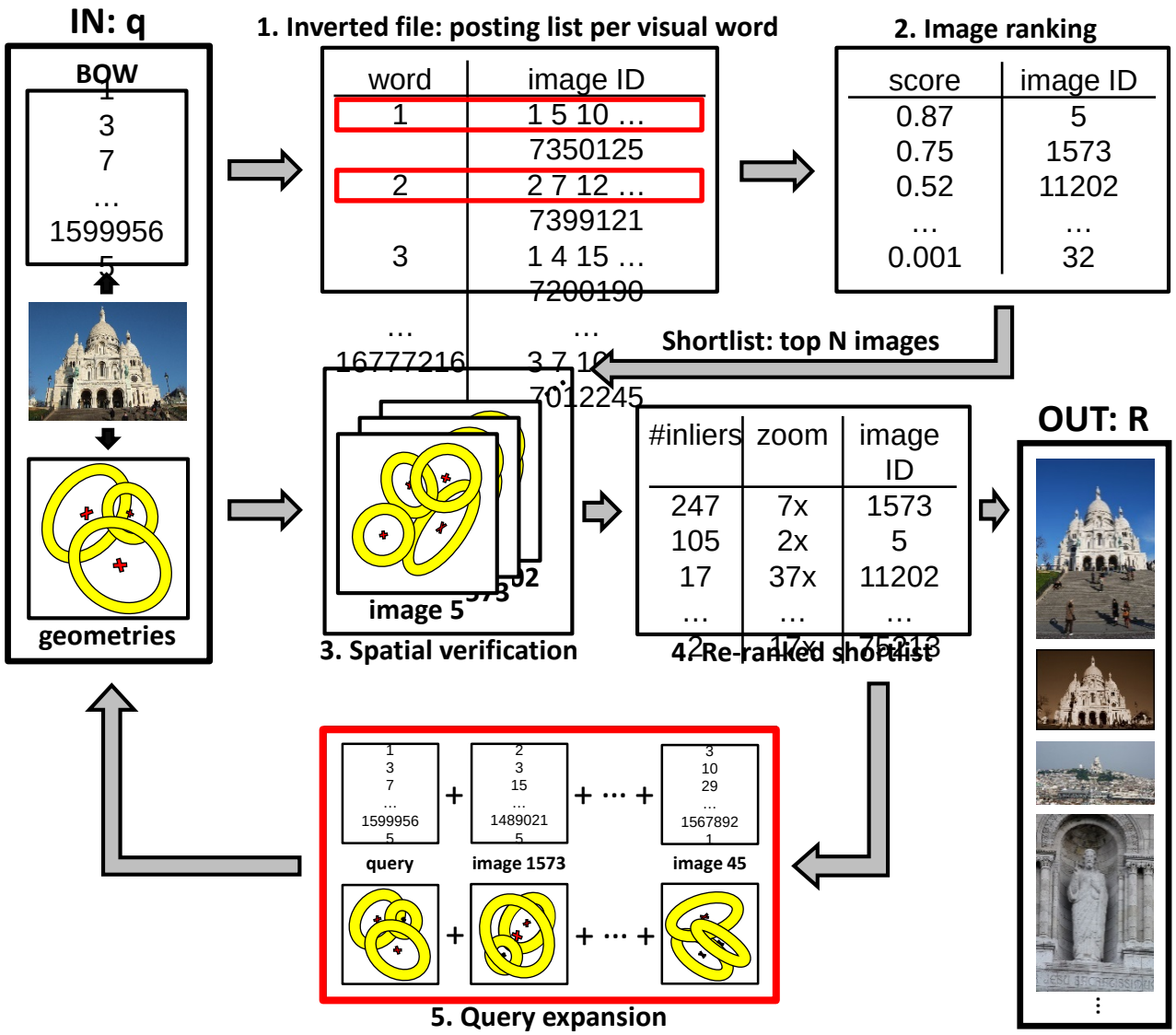
What most people find interesting?

Most commonly photographed parts
Given a query and a dataset, for every pixel in the query image:
Find the frequency with which it is photographed in detail



Mikulík, Radenović, Chum, Matas: **Efficient Image Detail Mining,** ACCV 2014

All Details: On-line Stage



All Details: Hierarchical Query Expansion

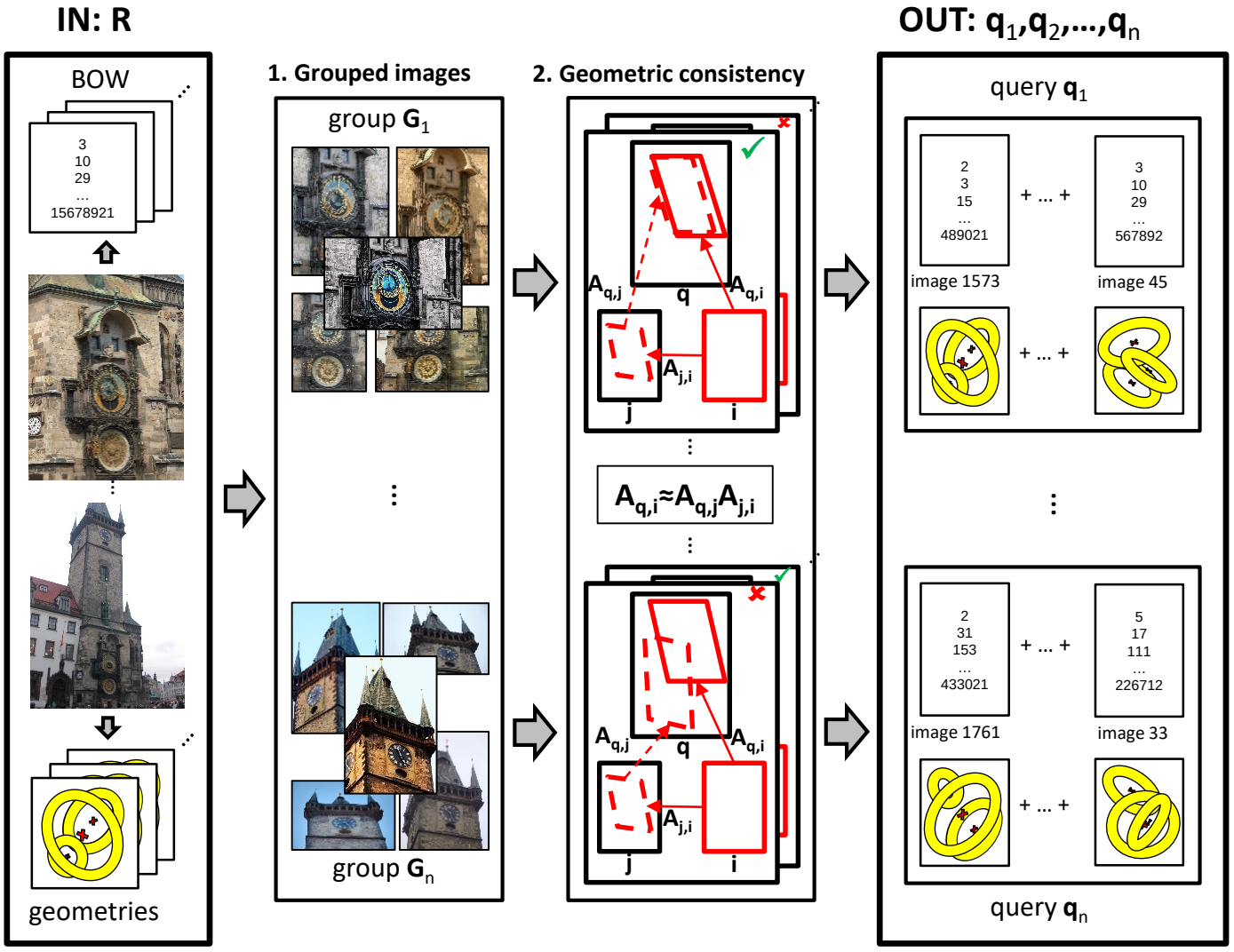


Image Retrieval for 3D Reconstruction

SfM - 3D Reconstruction

- Few thousand images

  Exhaustive matching of all image pairs

  [Snavely, Seitz, Szeliski: **Photo tourism**, SIGGRAPH 2006]

  \+ High level of details reconstructed

  \- Infeasible for larger photo collections

- Few million images

  Matching images through standard image retrieval

  [Heinly, Schonberger, Dunn, Frahm:
  Reconstructing the World in Six Days, CVPR 2015]

  \+ Efficient and scalable image matching

  \- Details not reconstructed

Retrieval for 3D Reconstruction

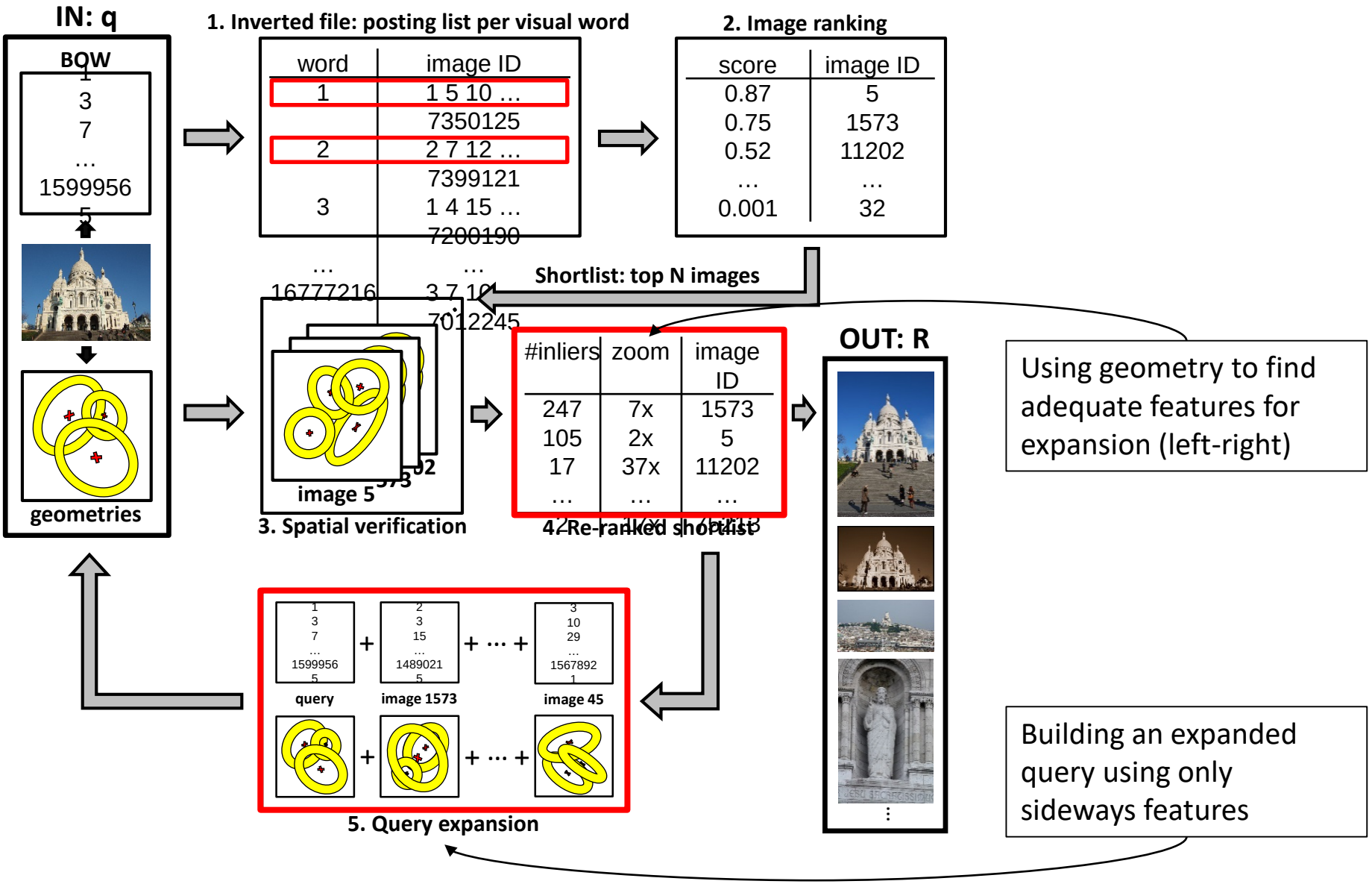
- Visually most similar search
  - Many near duplicates
  - Details lost
- Zoom-in and details search
  - Details retrieved
  - Transition images to match the details
- Zoom-out search
  - Viewpoint change
  - More context
- Sideways crawl
  - Significant viewpoint change
  - More context

Schoenberger, Radenović, Chum, Frahm:
From Single Image Query to Detailed 3D Reconstruction, CVPR 2015

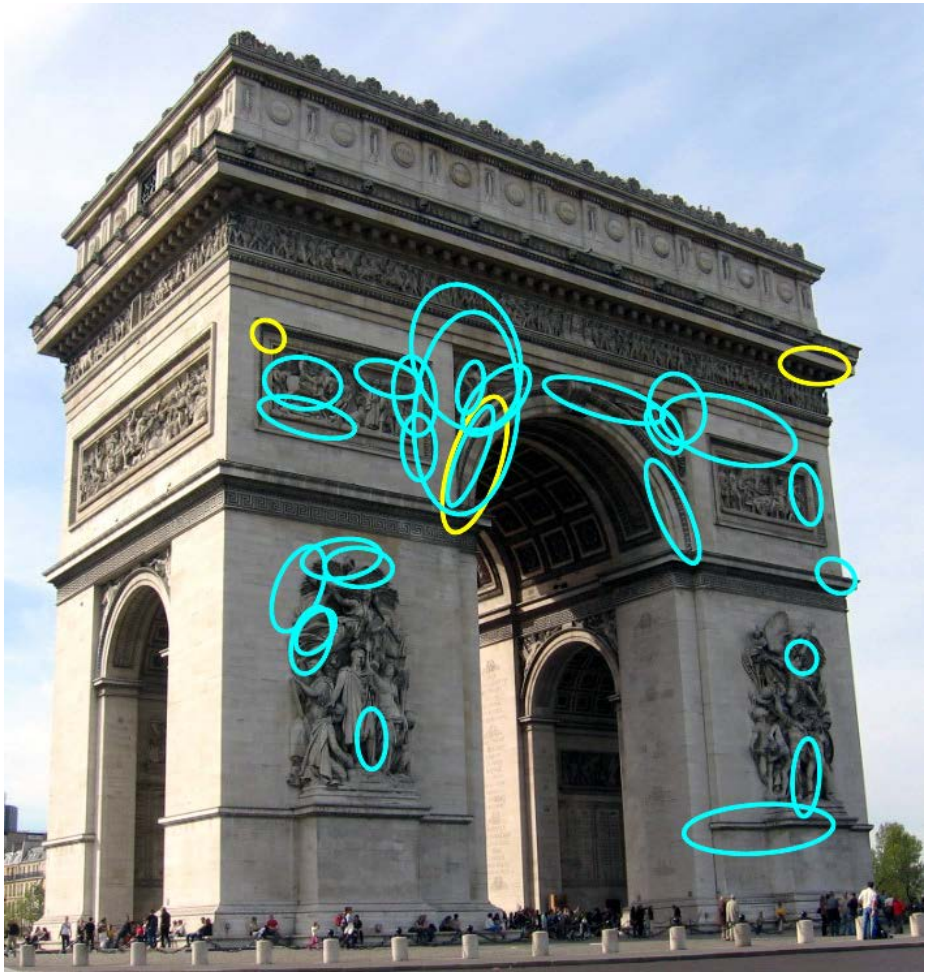
Sideways image crawl

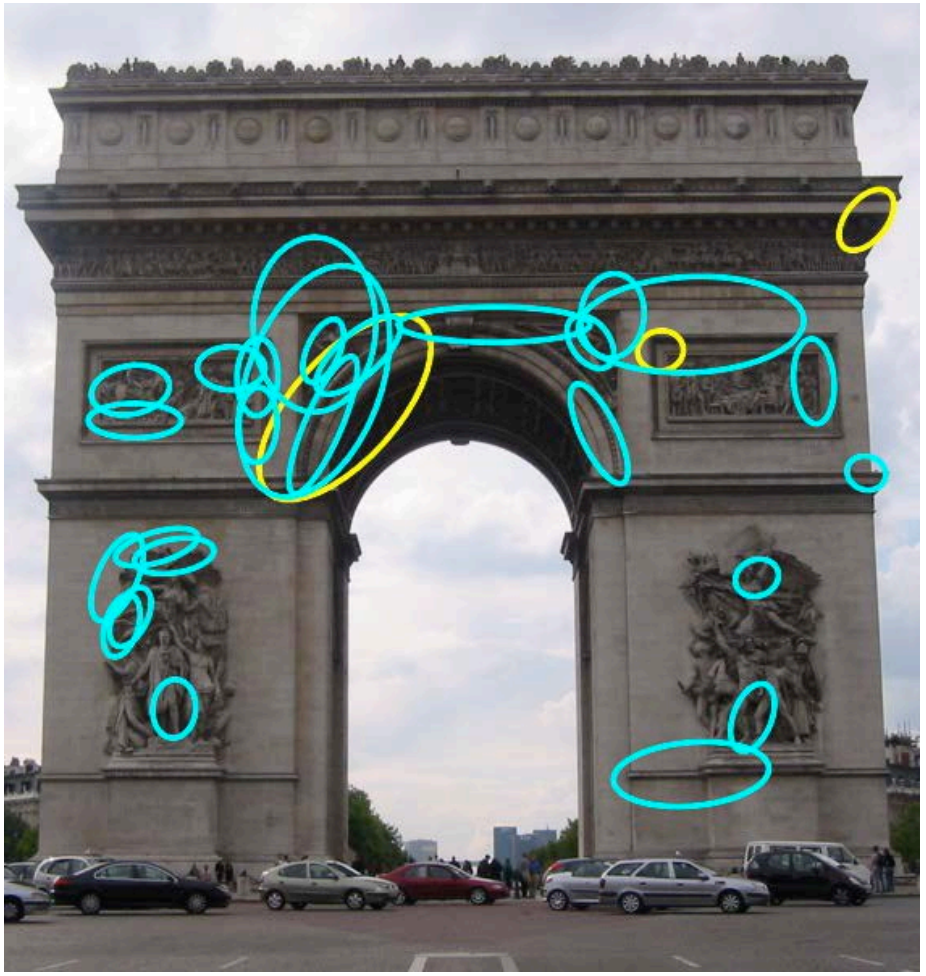
Schoenberger, Radenović, Chum, Frahm: **From Single Image Query to Detailed 3D Reconstruction,** CVPR 2015

Sideways crawl: On-line Stage

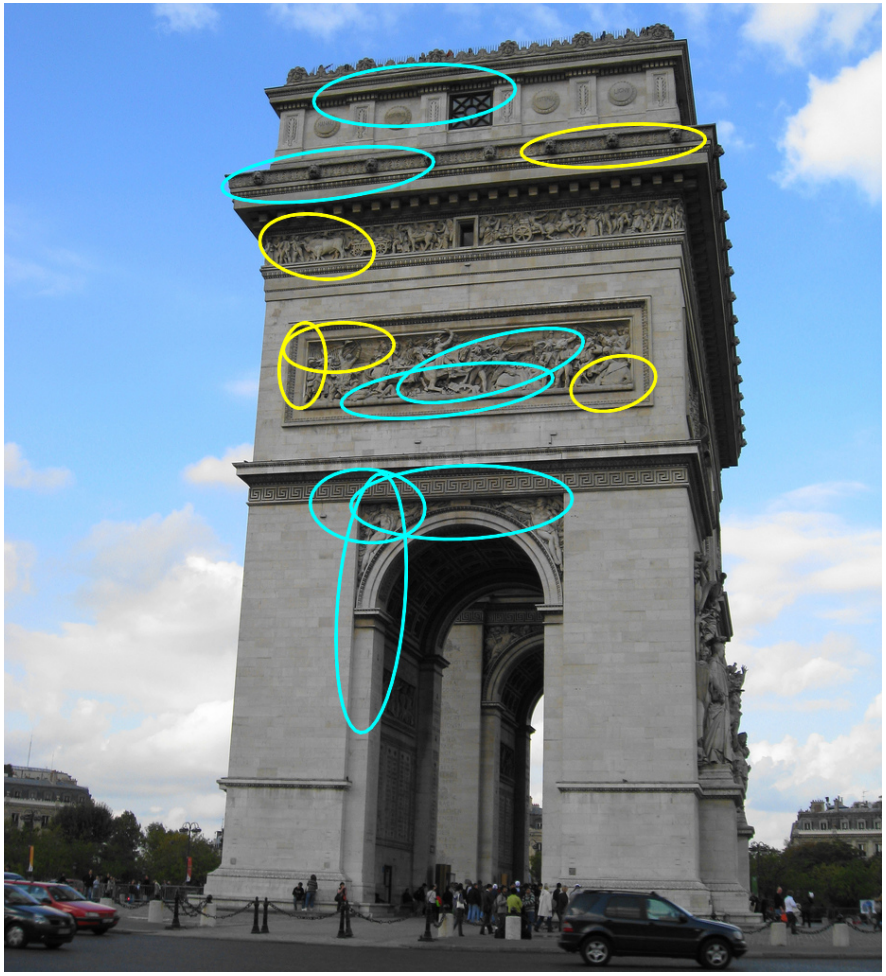


Sideways Left: Step by Step





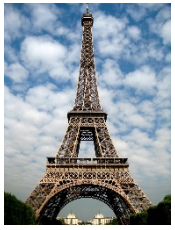
Sideways Left: Step by Step

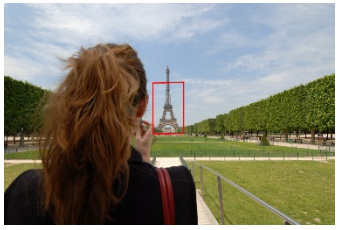
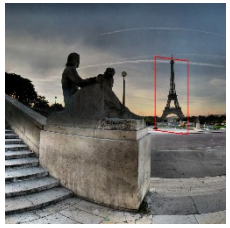


Summary

Visually most similar

Zoom-in / details



Zoom-out

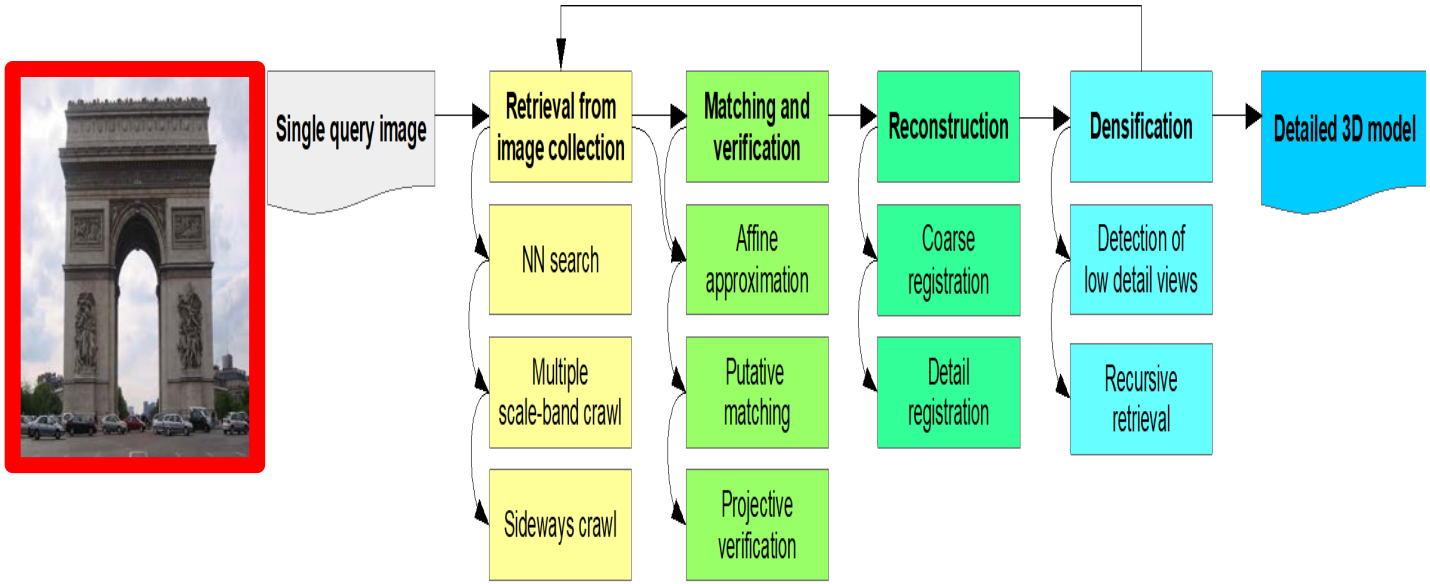
Sideways right

Application of Zoom in, out, ...
From Single Image Query to Detailed 3D Reconstruction

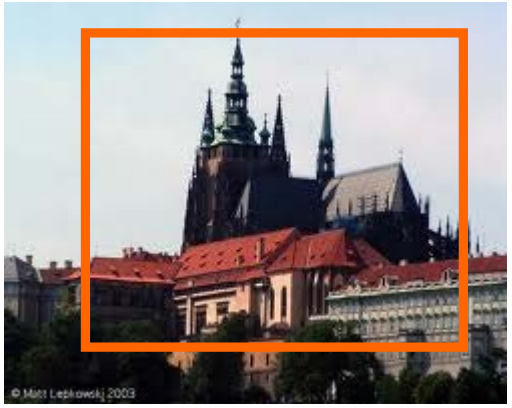
Johannes L. Schönberger[1], Filip Radenović[2], Ondrej Chum[2], Jan-Michael Frahm[1]

- Input: Single query image
  7.4 million images downloaded from the Internet

- Output: Detailed 3D model, Video, YouTube
- Recursive retrieval and Structure-from-Motion for detail reconstruction and model extension



Retrieval with illumination change



Illumination constraints

Video, YouTube CVPR 2016 Spotlight

Radenović F., Schönberger J. L., Ji D., Frahm J., Chum O., Matas J.
From Dusk till Dawn: Modeling in the Dark, CVPR 2016

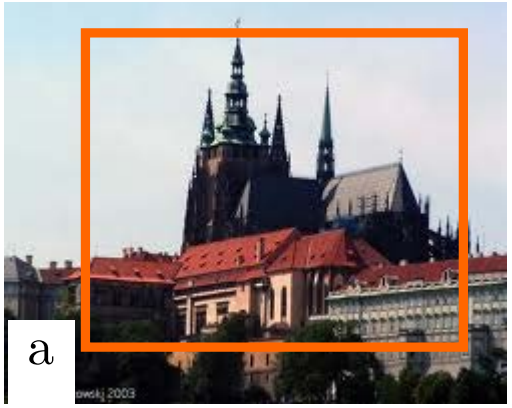
Conclusions

- For a certain class of objects and problems, image retrieval is mature -- reliable, fast, ...

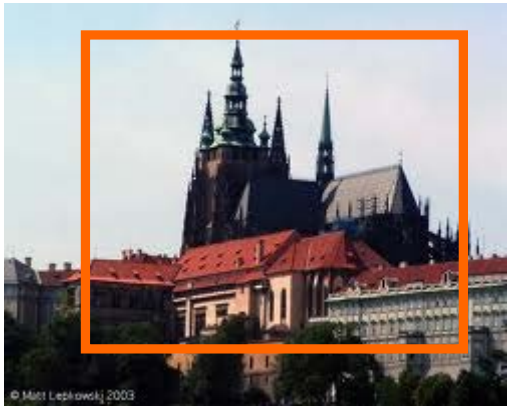
- There is more than the “specific most-similar object retrieval”, i.e. the nearest neighbor problem. E.g. Find :
  - the most geometrically different, yet the same scene
  - spots in the scene that attract attention

  - a representative collection of appearances of the scene – find the database image most different to already retrieved
  - images of the scene, most similar in illumination, viewpoint, arrangement) to *another image,* a secondary query

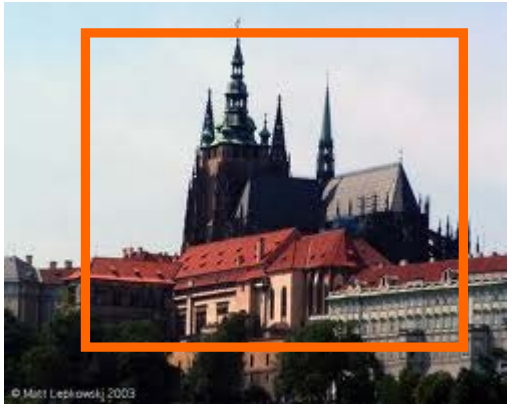
Retrieval with occlusion

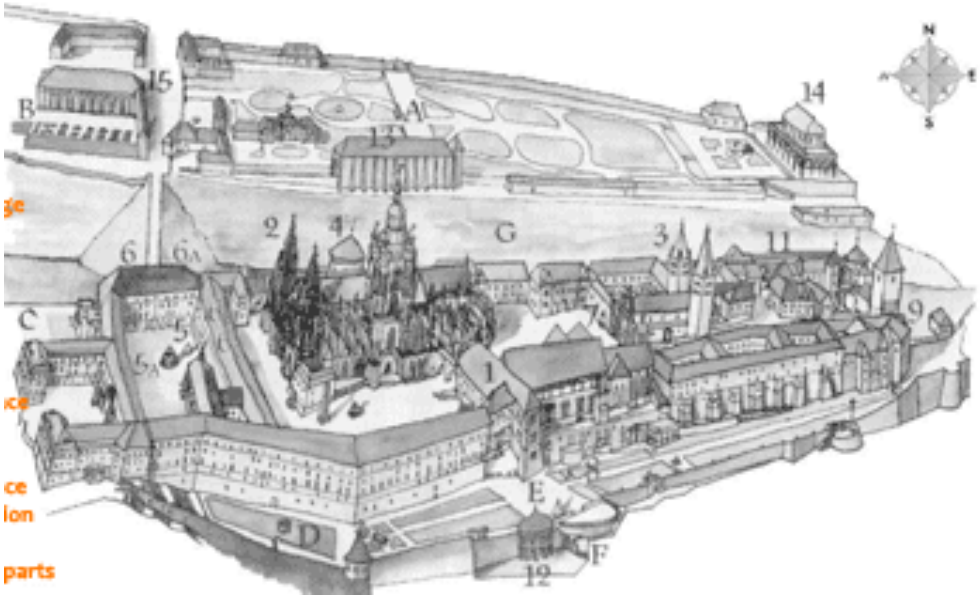


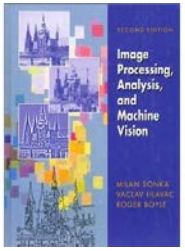
Retrieval with large time difference

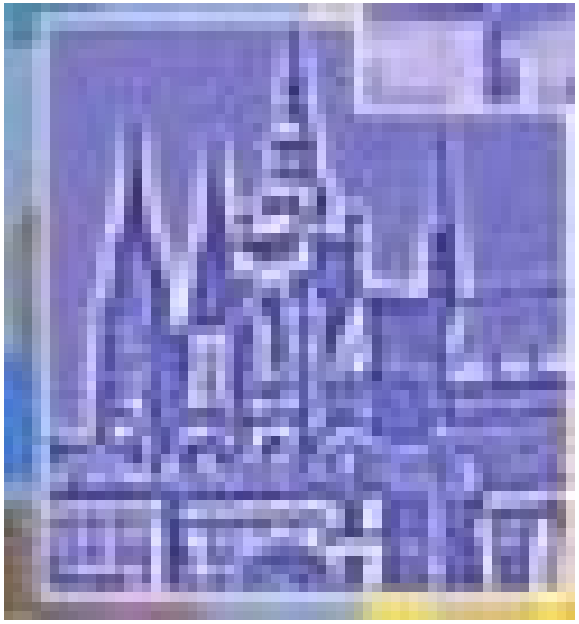


Retrieving different modalities









Retrieving non-overlapping images

Retrieving *sequences* helps in complex cases

Thank you!

Questions?