image retrieval

Computer Vision Methods 2025

CTU in Prague

Giorgos Tolias

slide credits also to: Jiri Matas, Ondrej Chum, Andrej Mikulik, Filip Radenovic, Michal Perdoch, James Pritts, Dmytro Mishkin, Jan Cech

outline

- task formulation
- bag-of-words (BoW) model
- BoW generic point-of-view → other approaches
- fast spatial (geometric) matching
- special retrieval tasks

the task

image retrieval - example

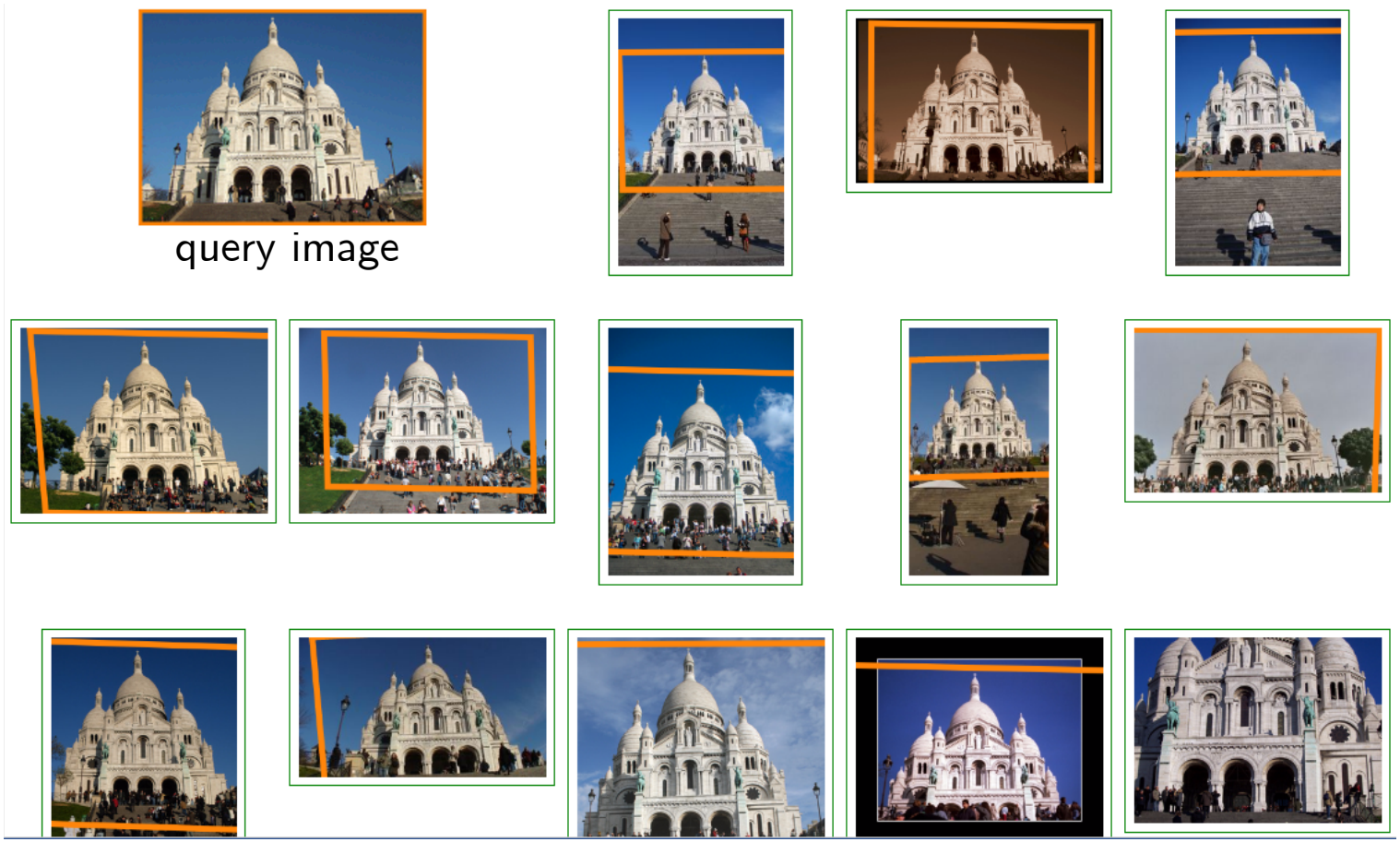


image retrieval

for image copy detection

image retrieval

for image copy detection

image retrieval

for image geo-localization

image retrieval




for image geo-localization

image retrieval

for image geo-localization

image retrieval

for image geo-localization

image retrieval

History [edit]

Originally a family business, the pub was founded in 1499,[1] and therefore celebrated its 500th anniversary in 1999. It has been referred to as the oldest brewery in Prague. In 1762 the

for landmark recognition

image retrieval

for e-commerce

Figure from Li et al. 2016

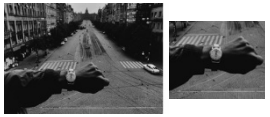
image retrieval

for logo infringement identification

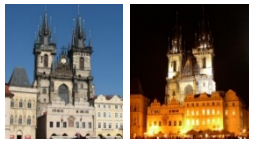
image retrieval - task

- given a query image and an image collection (database)

- retrieve database images according to their relevance to the query
  - rank all images
  - retrieve most relevant
- relevance defined in different ways, according to the task



identical photo



identical object

infringement

- retrieval approaches characterized by performance
  - quality: ability to retrieve relevant images first
  - speed: small query time
  - memory footprint: low memory requirements

  - scalability: applicability to large image collections

image retrieval - task

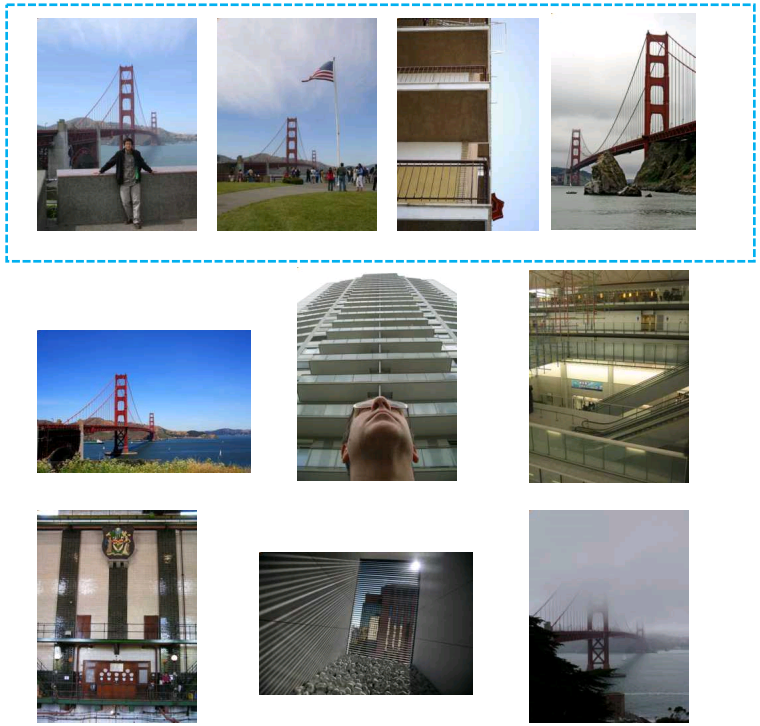
- retrieval is a ranking task for a given query "q"
- relevance is defined for pairs (q,a), "a" is a database image
- degree of relevance s(q,a)
  - estimate it to perform retrieval
  - need for ground-truth
- difficult to obtain continuous degree of relevance in ground-truth
  - binary approximation
    - 1: positive / matching /relevant pair
    - 0: negative / non-matching / non-relevant pair

image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



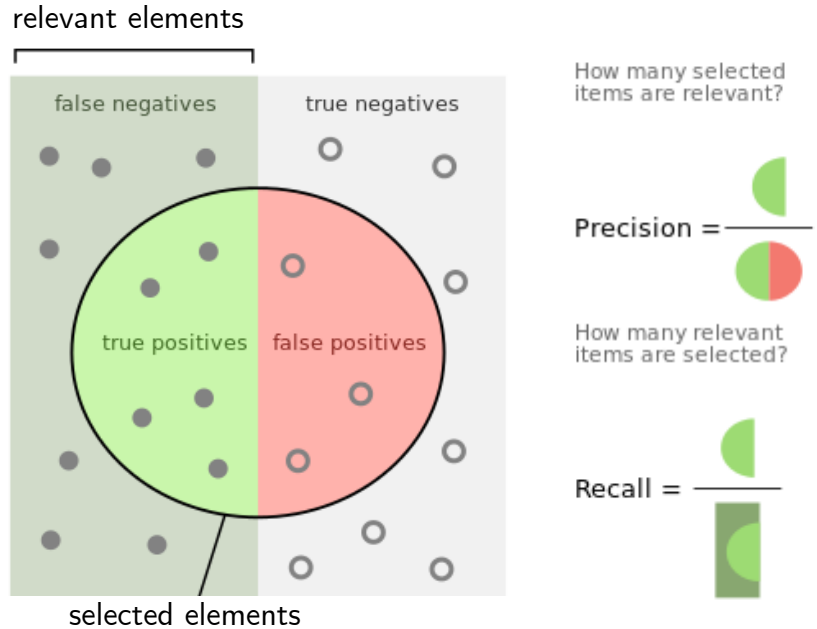


image retrieval - evaluation

query:

ranking:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

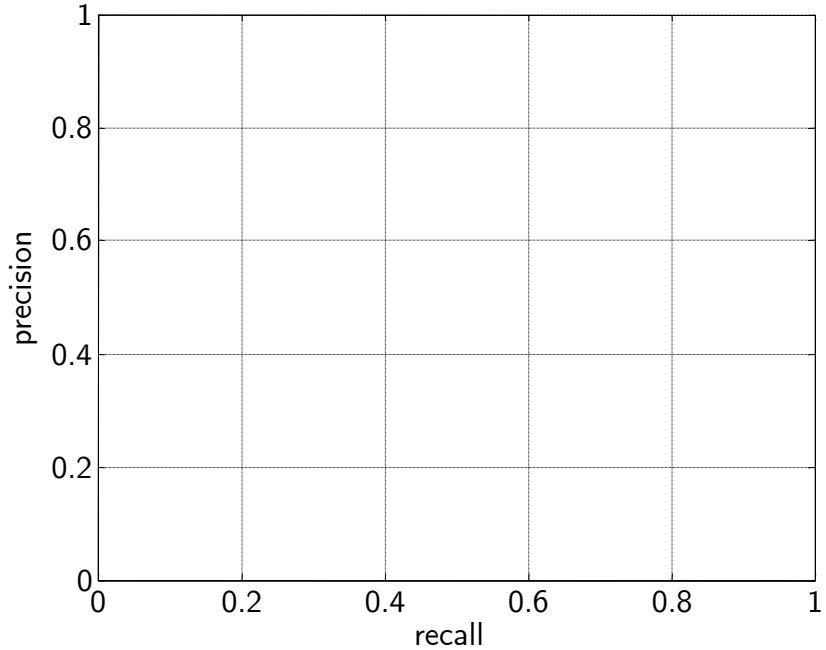
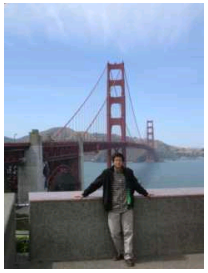


image retrieval - evaluation

query:

ranking:



- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

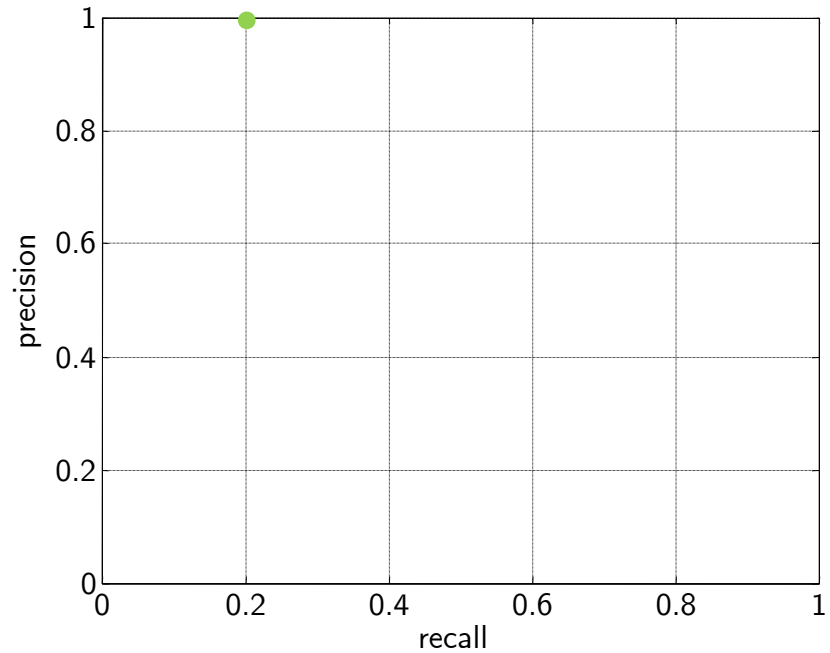
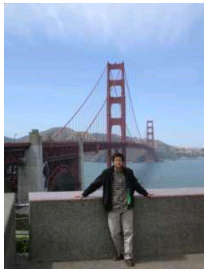


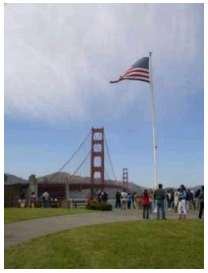
image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:





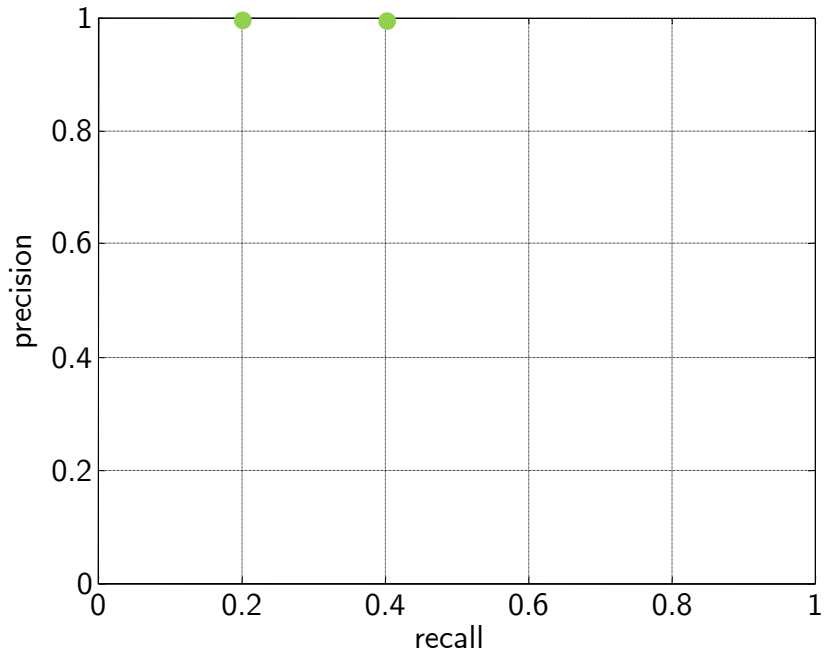
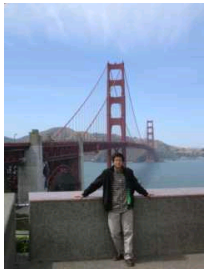
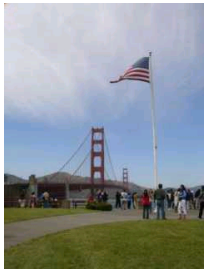


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:

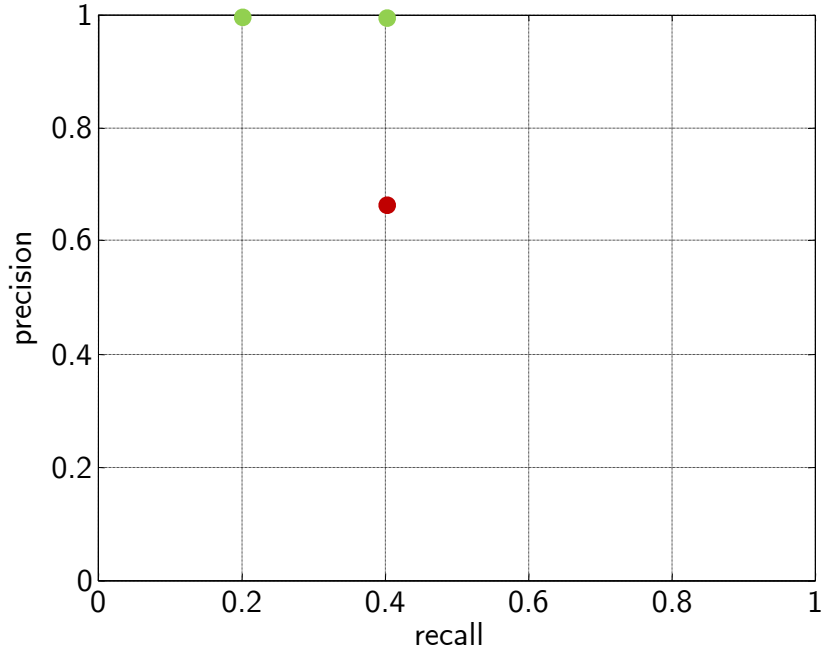
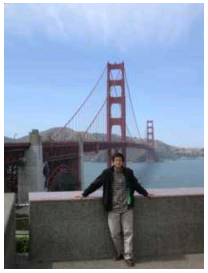
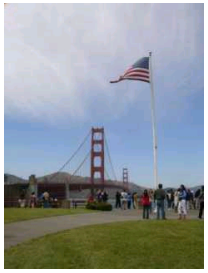
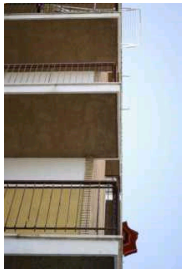
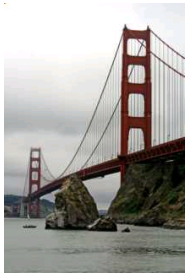



image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:

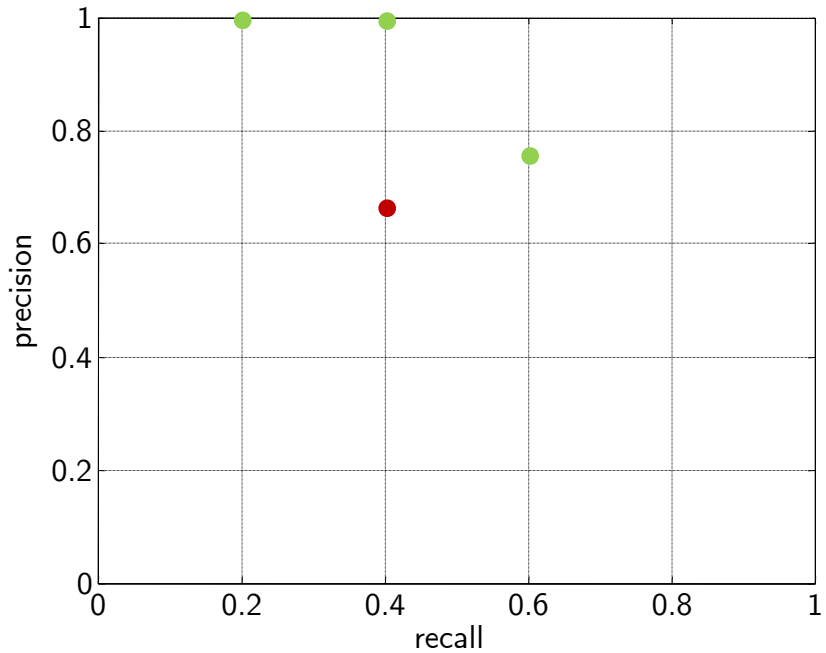
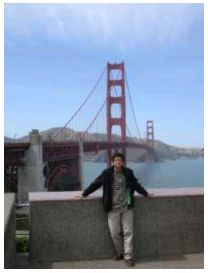
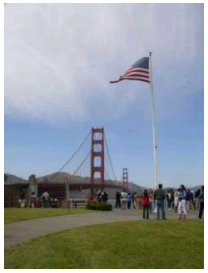
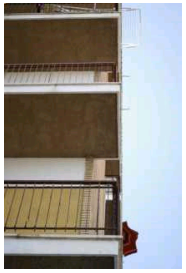
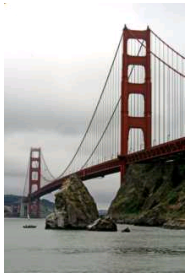


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:

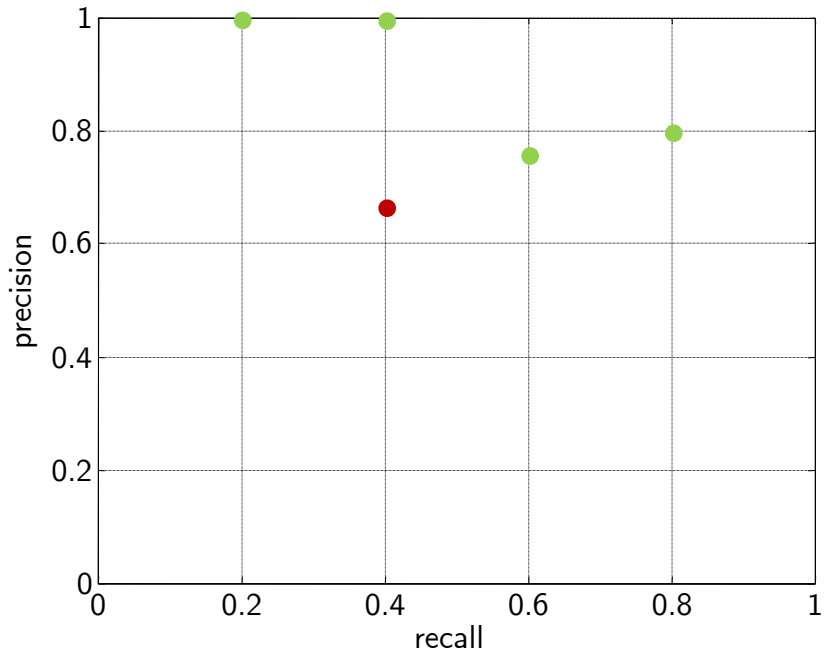
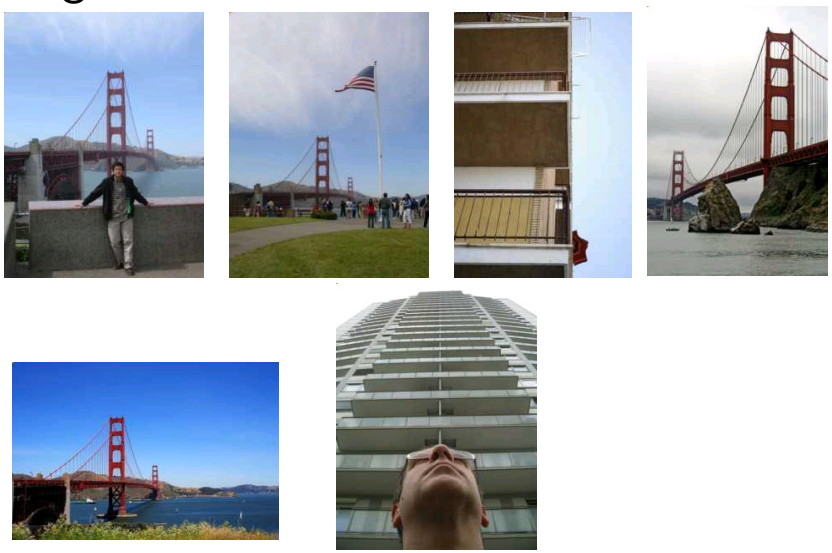



image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



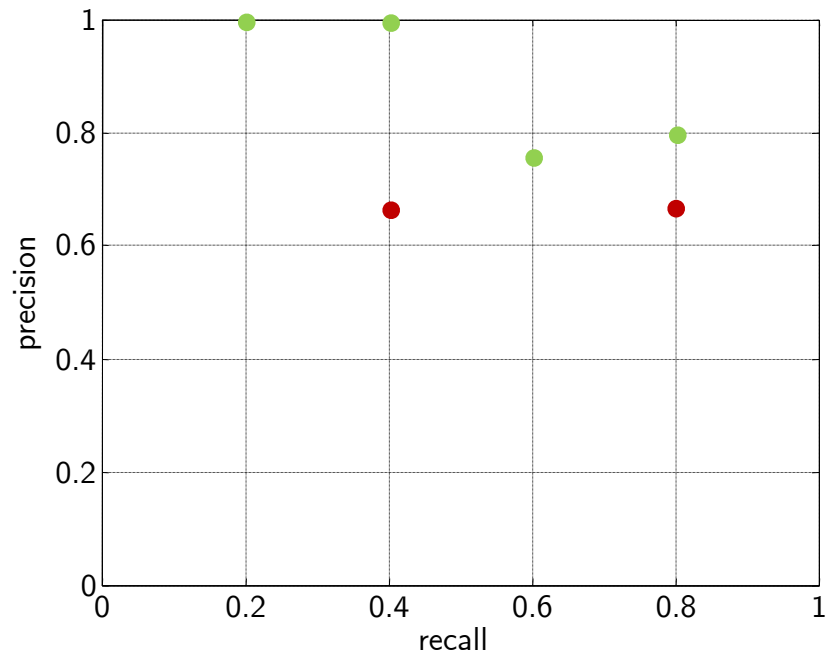
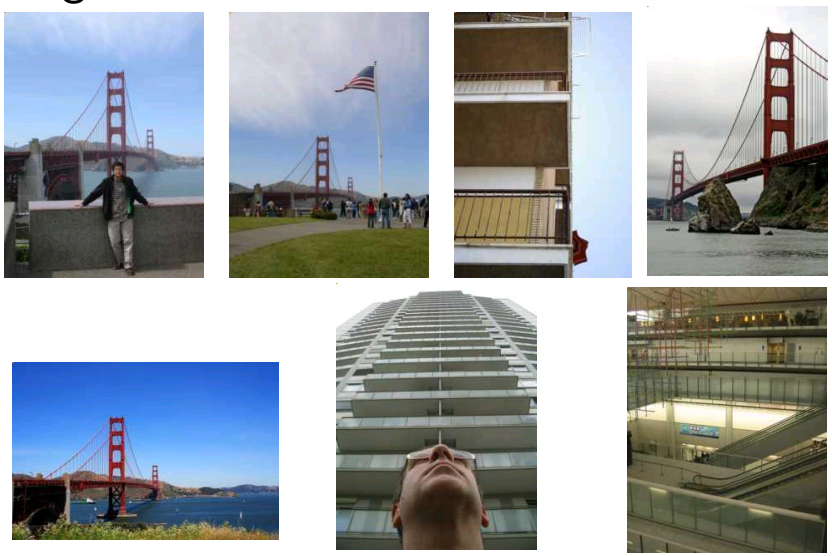


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



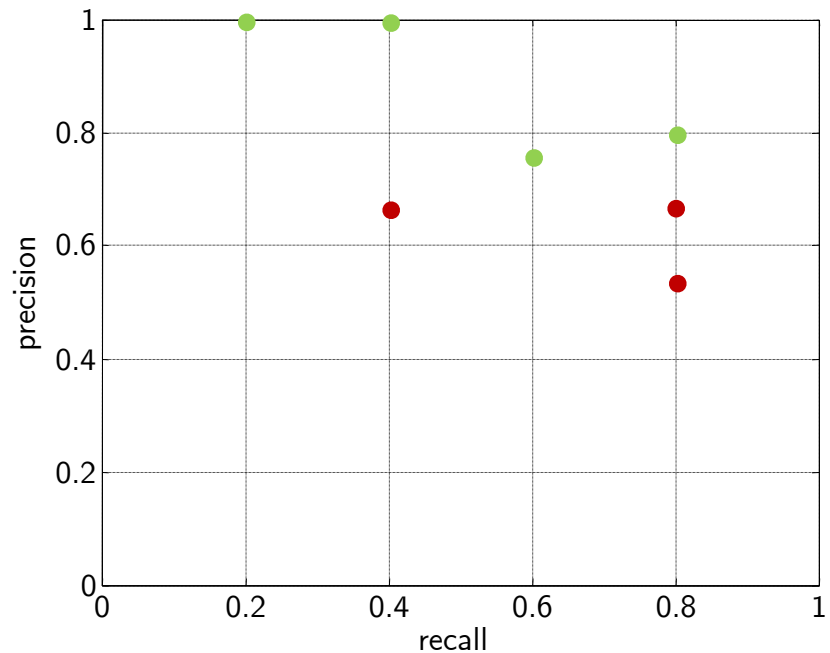
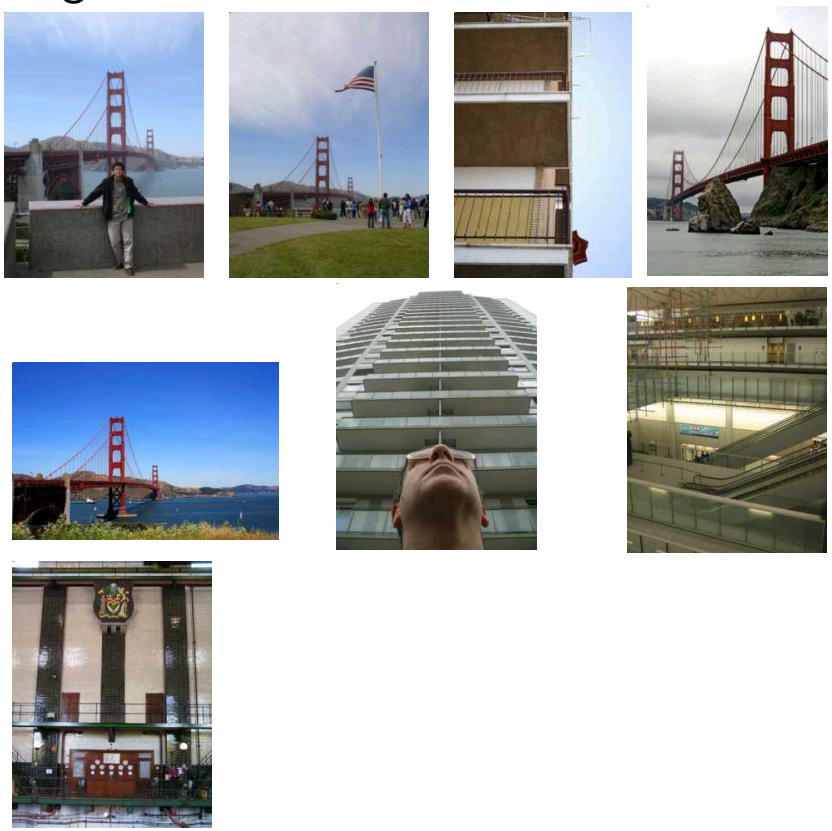


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



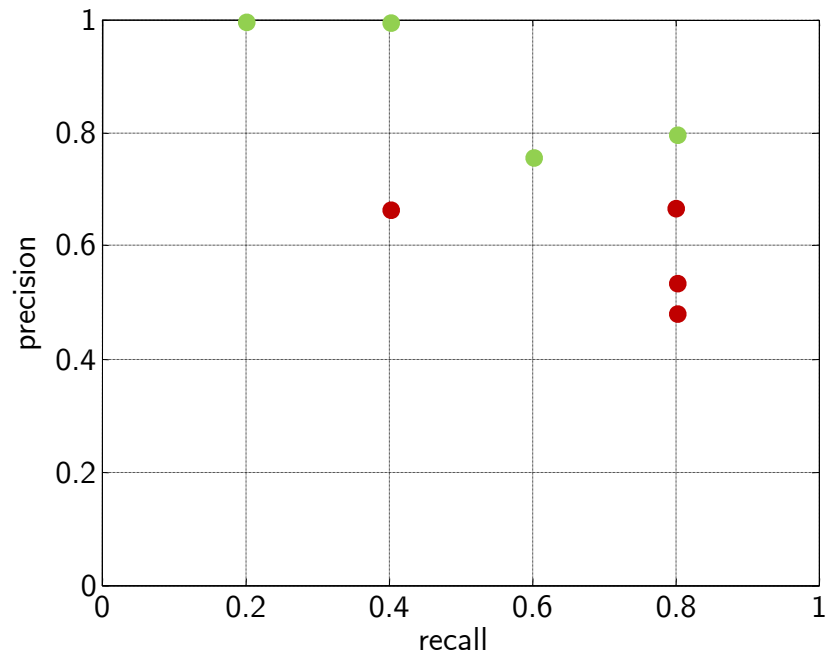
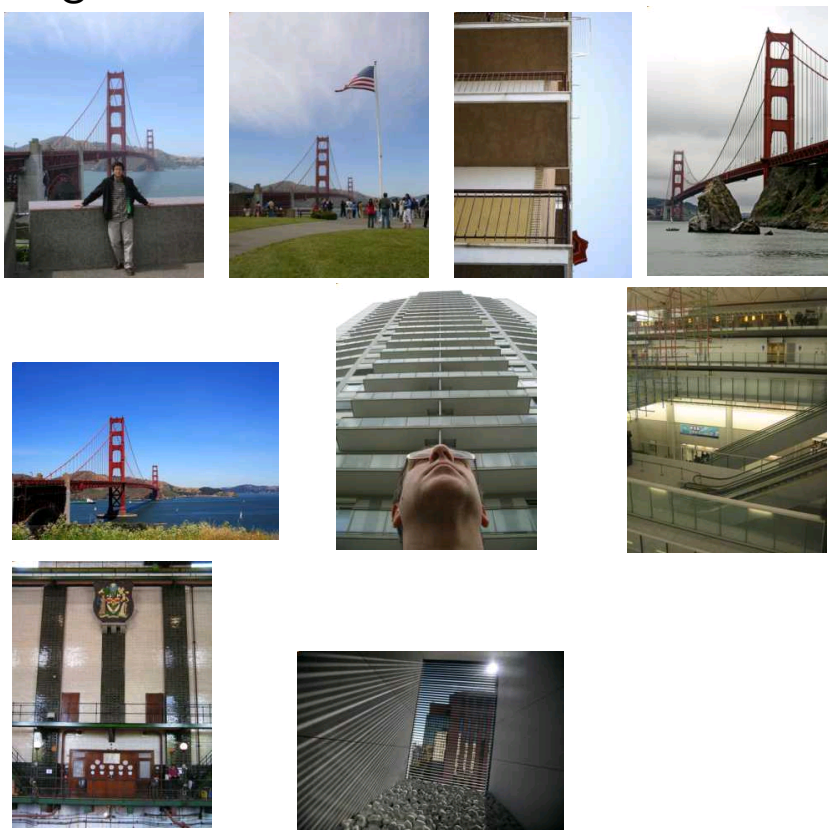


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



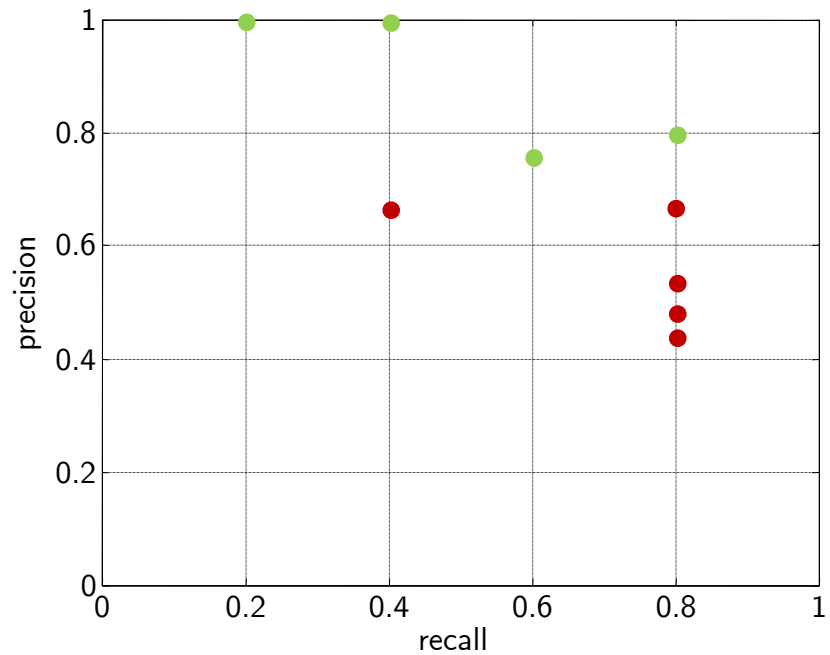
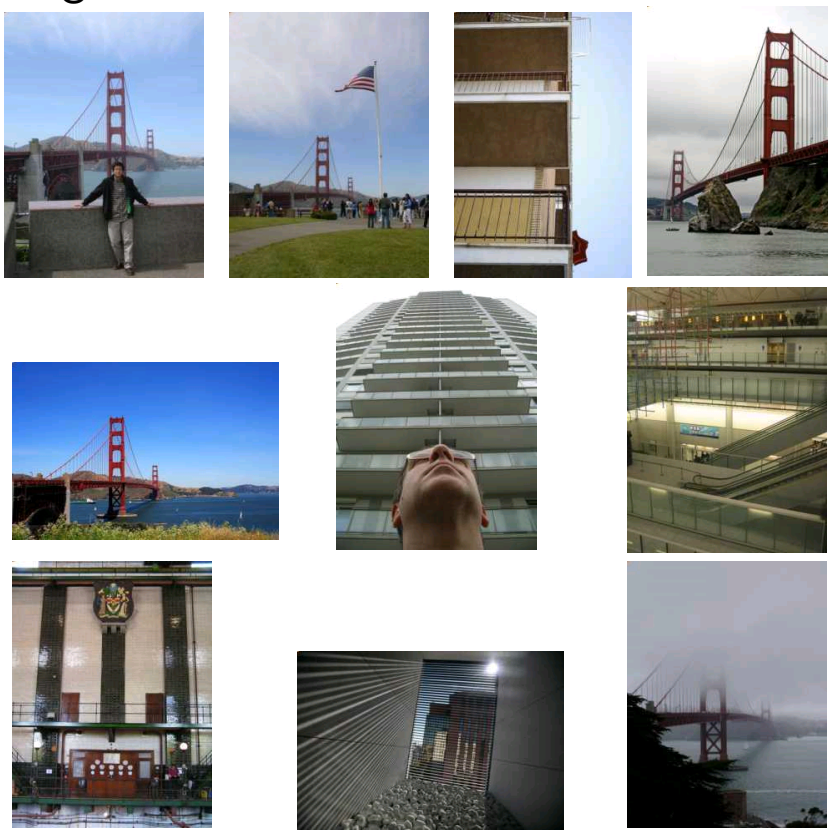


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



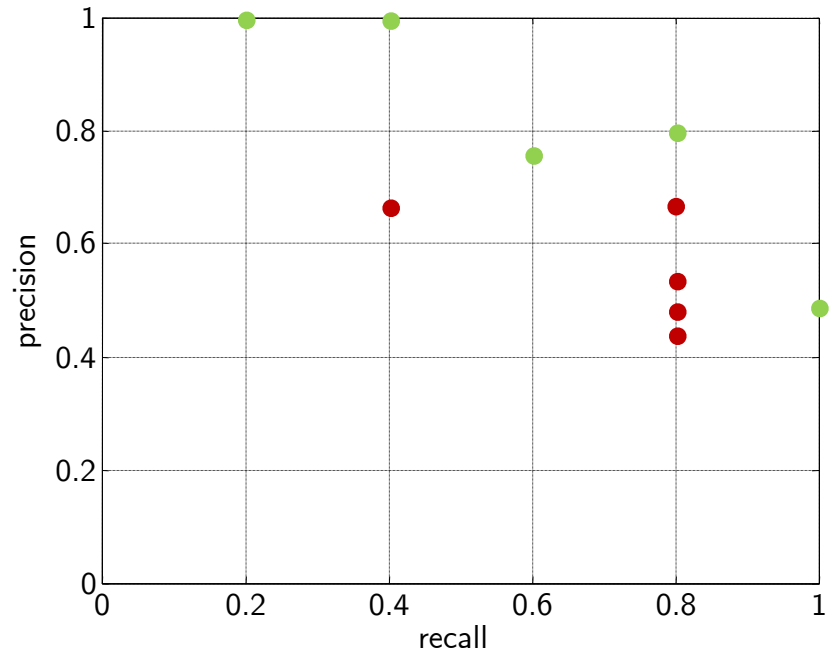
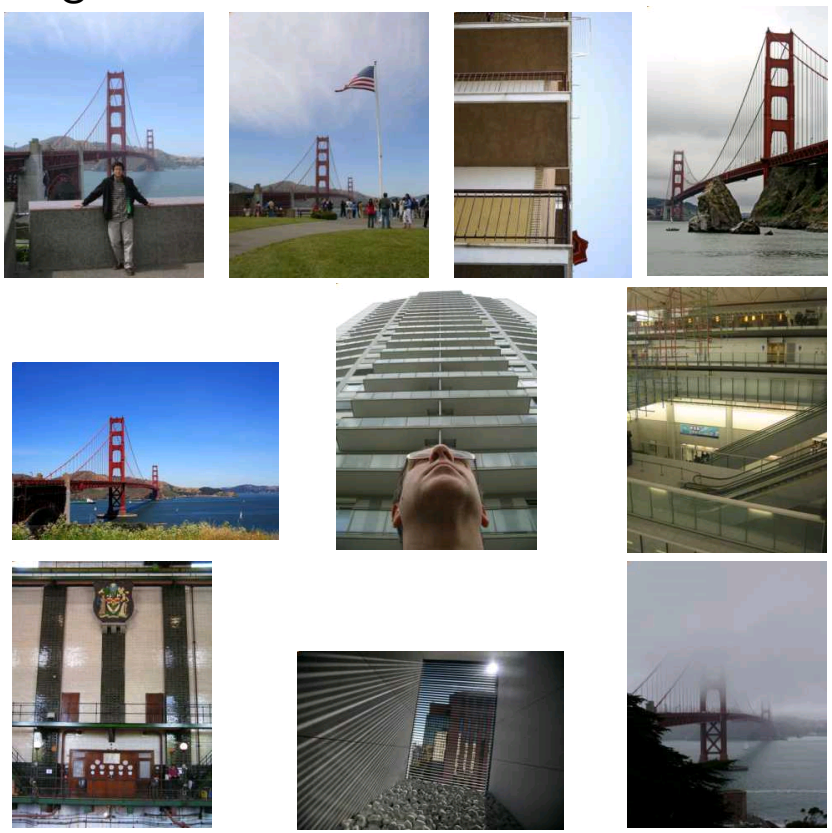


image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:



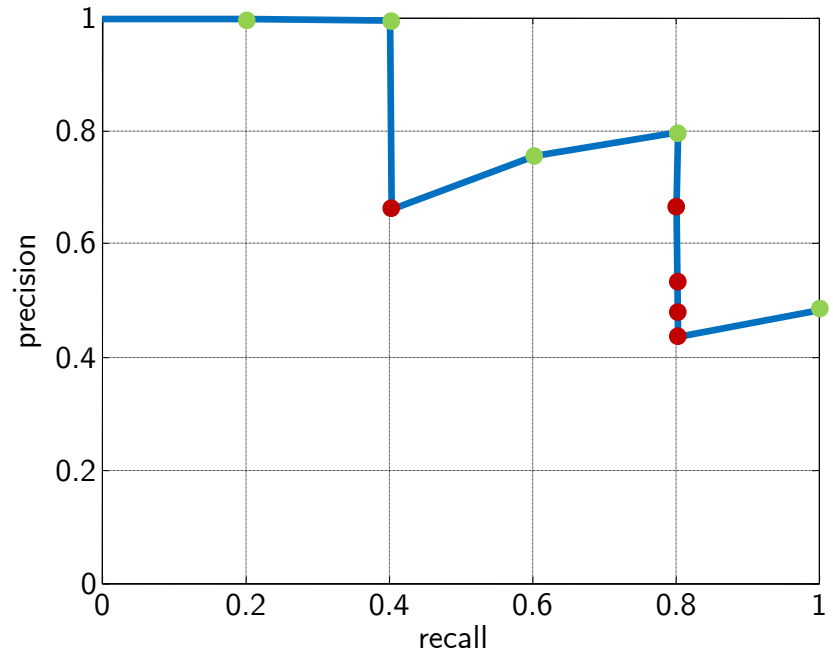
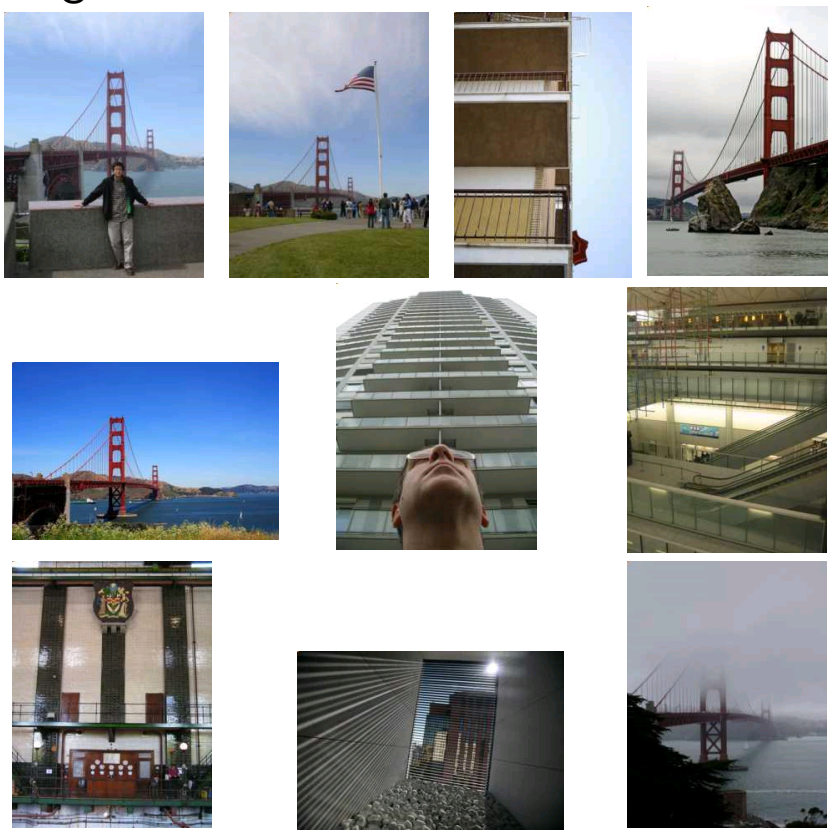


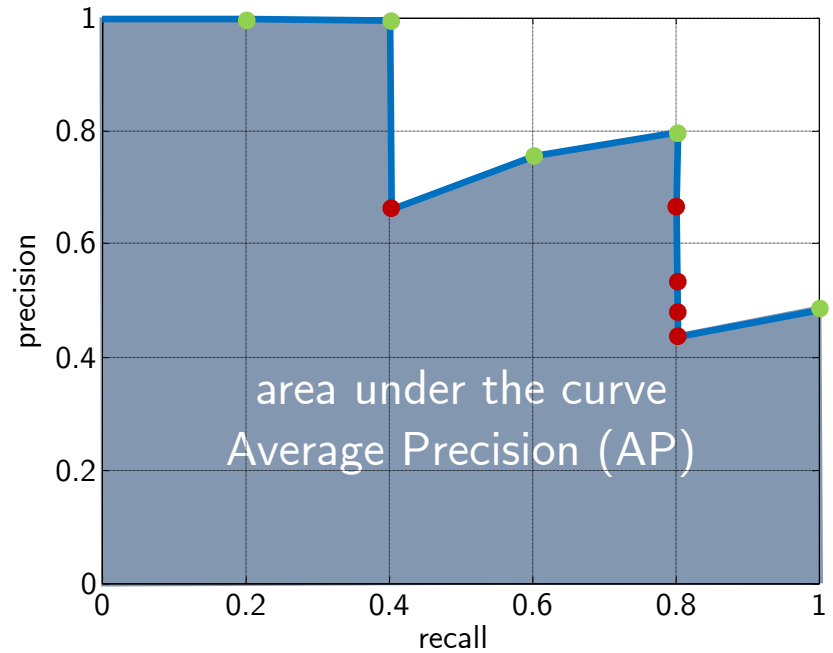
image retrieval - evaluation

query:

- 10 database images
- 5 relevant images, according to binary pairwise labels
- select a "window" of top-k images
- measure precision@k and recall@k
- what should k be? should we use only one window?

ranking:

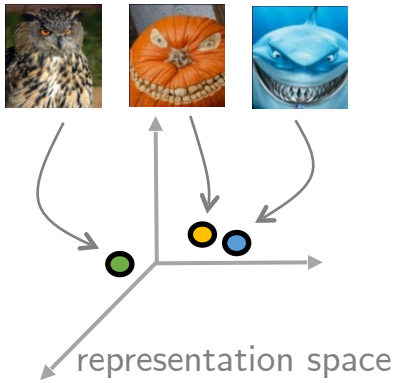


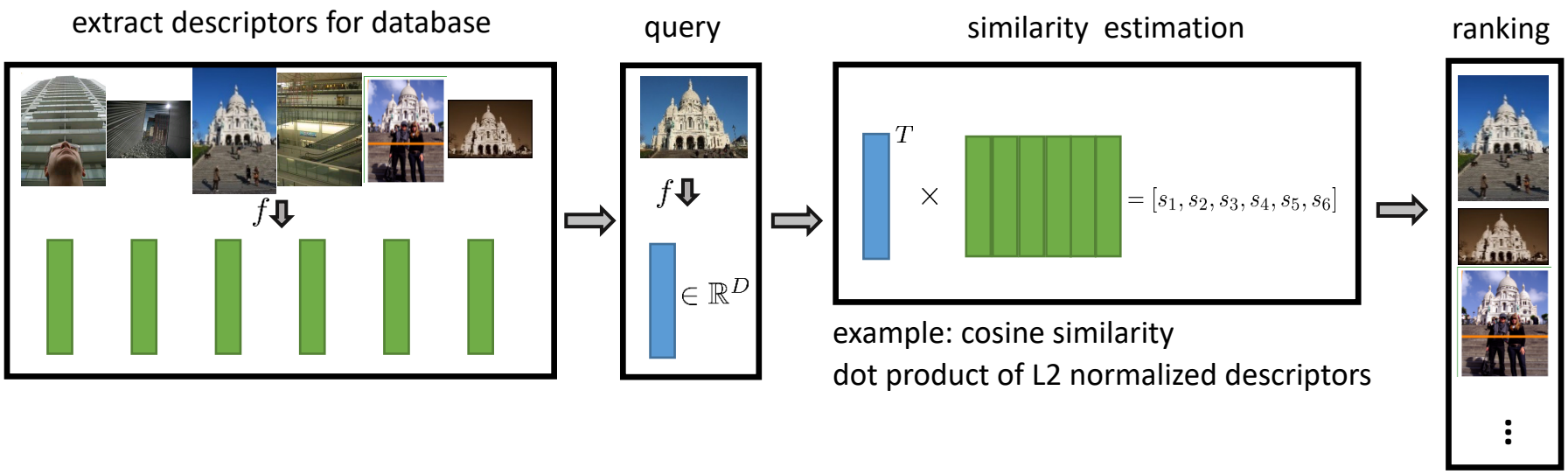


visual representation

visual representation – vector space

- map images to a high dimensional representation space
- image descriptor $X = f(\text{  }) \in \mathbb{R}^D$
- function f is hand-crafted or learned
  - so that proximity reflects relevance
- retrieval is performed by nearest neighbor search



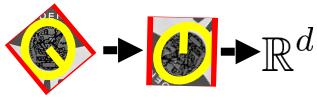


- how to obtain (global) image descriptors?
  - let's revisit local features & descriptors first...

visual representation – local descriptors

$\mathcal{X} = g($  $)$

$\mathcal{X} = \{\mathbf{x}\},\ \mathbf{x} \in \mathbb{R}^d$



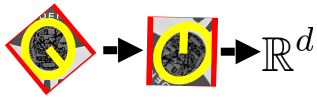
$X = h(g($  $)) = h(\mathcal{X}) \in \mathbb{R}^D$

visual representation – local descriptors

- represent an image by a set of local descriptors

$\mathcal{X} = g($  $)$

$\mathcal{X} = \{\mathbf{x}\},\ \mathbf{x} \in \mathbb{R}^d$



- obtain global descriptor as a function of a local descriptor set

$X = h(g($  $)) = h(\mathcal{X}) \in \mathbb{R}^D$

visual representation – local descriptors

- represent an image by a set of local descriptors

$\mathcal{X} = g($  $)$

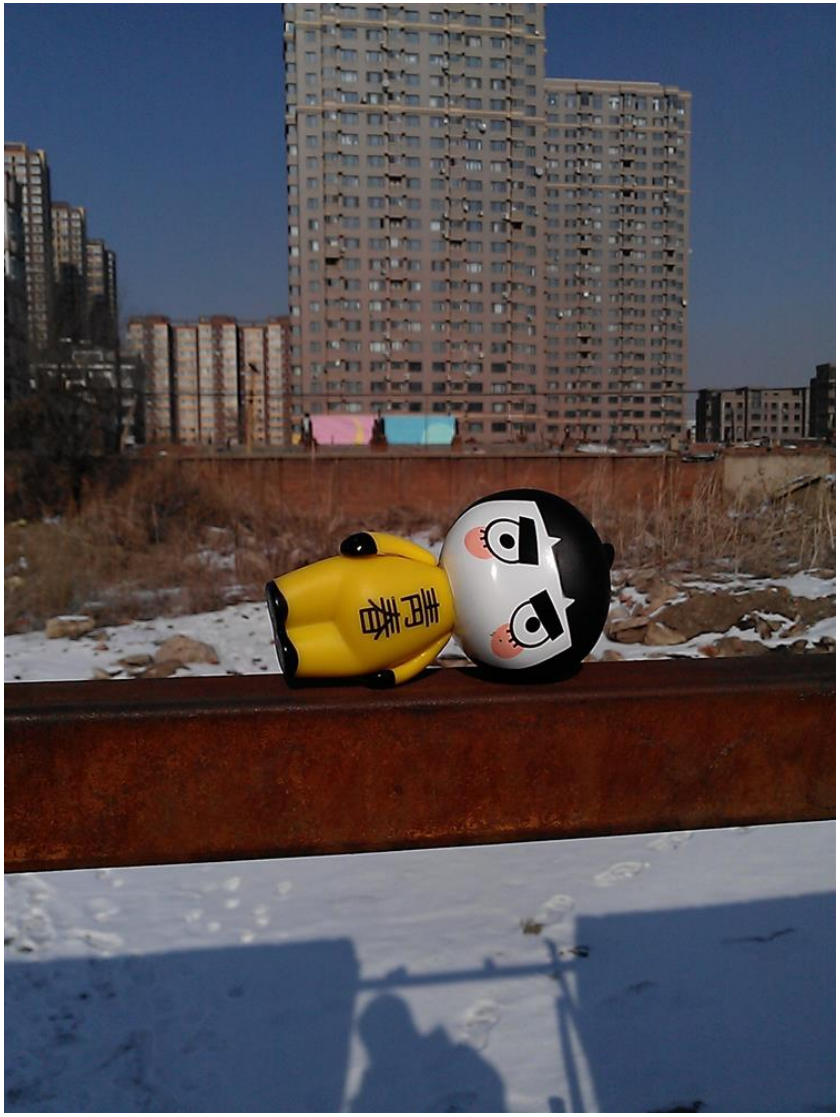
$\mathcal{X} = \{\mathbf{x}\},\ \mathbf{x} \in \mathbb{R}^d$

- obtain global descriptor as a function of a local descriptor set

$X = h(g($  $)) = h(\mathcal{X}) \in \mathbb{R}^D$

- inherit their invariance properties
- robustness to occlusions and background clutter

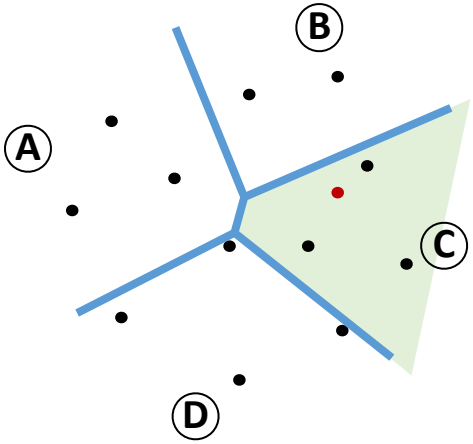
local features



bag-of-words
BoW

bag-of-words - representation

- inspired by natural language processing
  - text document represented by the set of words in it
- quantize the space of local descriptors
  - eg. with k-means clustering
  - each cluster is a visual word
- visual codebook: set of visual words (centroids)



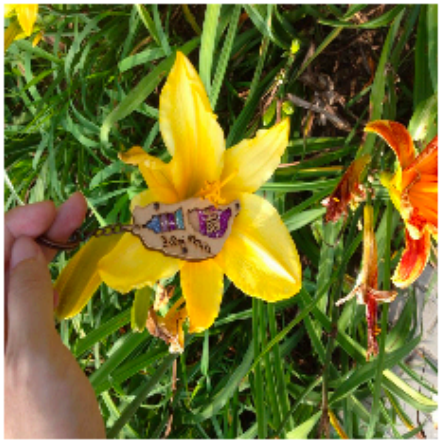
[Sivic & Zisserman. ICCV'03]

[Csurka et al. ECCVW'04]

features (patches) in the same cluster

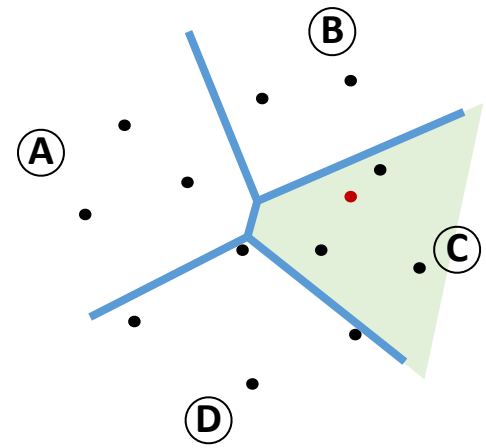

CANADA

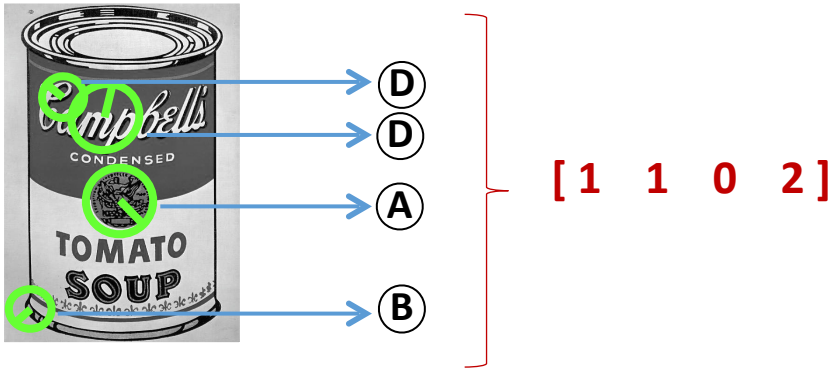

LOVE



bag-of-words - representation

- inspired by natural language processing
  - text document represented by the set of words in it
- quantize the space of local descriptors
  - eg. with k-means clustering
  - each cell (cluster) is a visual word
- visual codebook: set of visual words (centroids)
- assign local descriptors to visual words
  - according to the nearest centroid
- histogram of visual word occurrences (term-frequency) per image

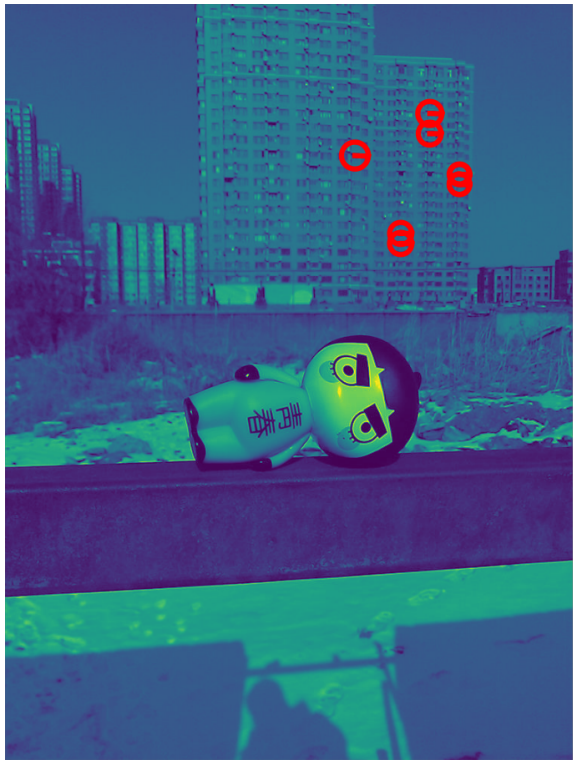




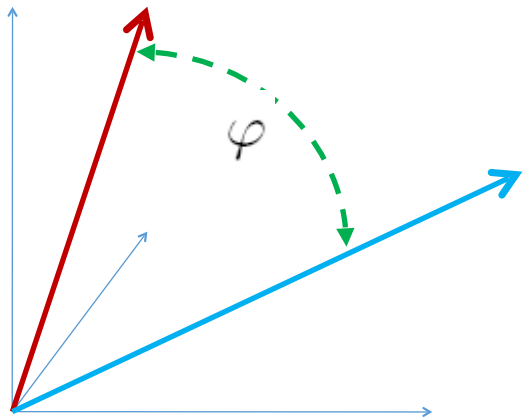
[Sivic & Zisserman. ICCV'03]

[Csurka et al. ECCVW'04]

local features with the same visual word



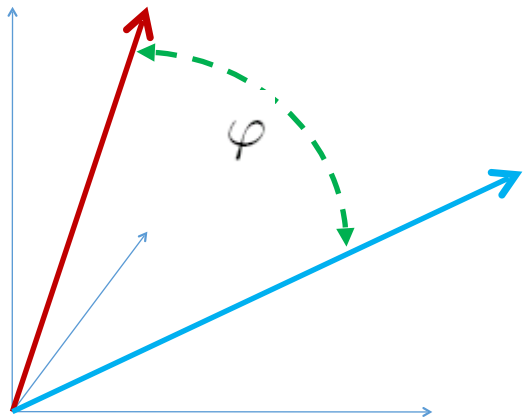
bag-of-words - similarity



bag of words representation

$$\cos\varphi = \frac{X^\top Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

bag-of-words - similarity

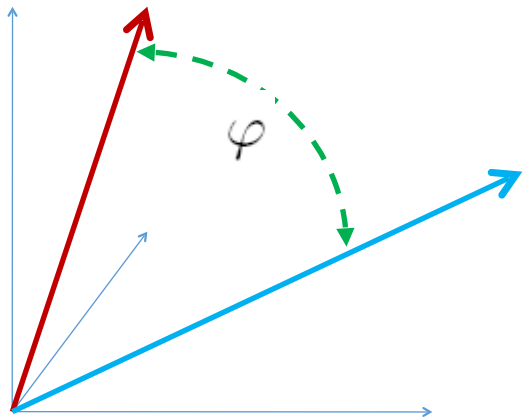


bag of words representation

$$\cos\varphi = \frac{X^\top Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database				query		score
	A	B	C	D			
Y_1	1	0	0	2	A	0	s_1
• X_q					B	3	=
					C	0	
					D	1	

bag-of-words - similarity

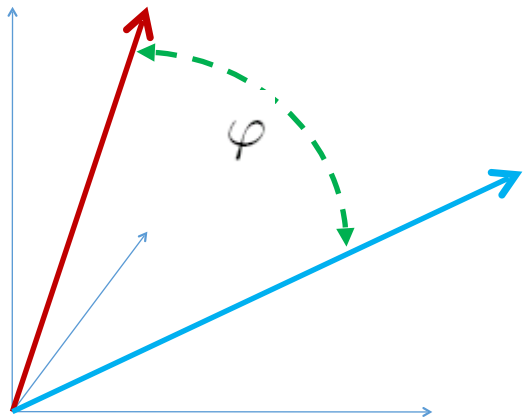


bag of words representation

$$\cos\varphi = \frac{X^T Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database	query	score
	A B C D		
Y_1	(1 0 0 2)	A 0	s_1
• X_q		B 3	=
		C 0	
		D 1	

bag-of-words - similarity

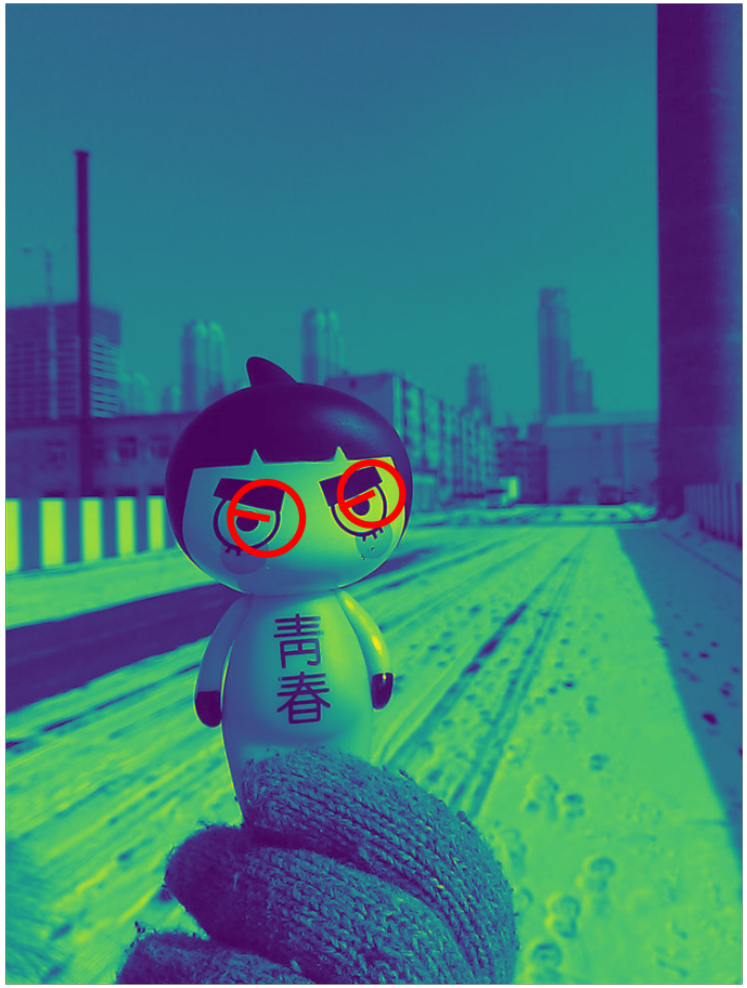


bag of words representation

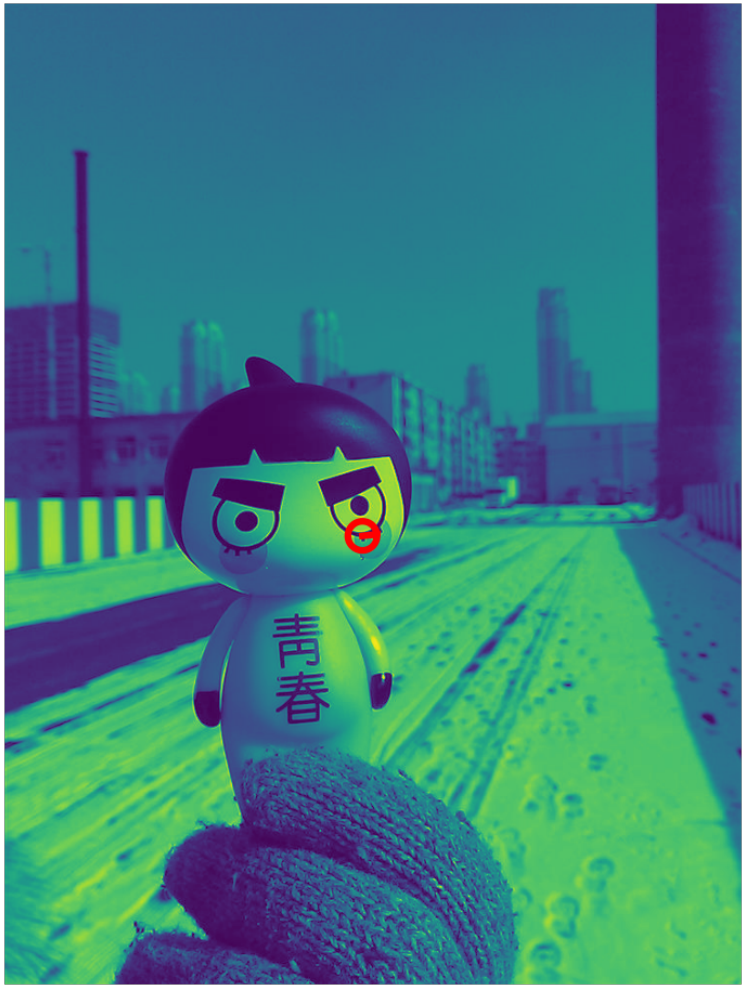
$$\cos\varphi = \frac{X^T Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database					query	score
	A	B	C	D			
Y_1	1	0	0	2		A 0	s_1
Y_2	0	2	0	1	$\bullet\ X_q$	B 3 =	s_2
Y_3	1	0	0	0		C 0	s_3
	⋮					D 1	⋮

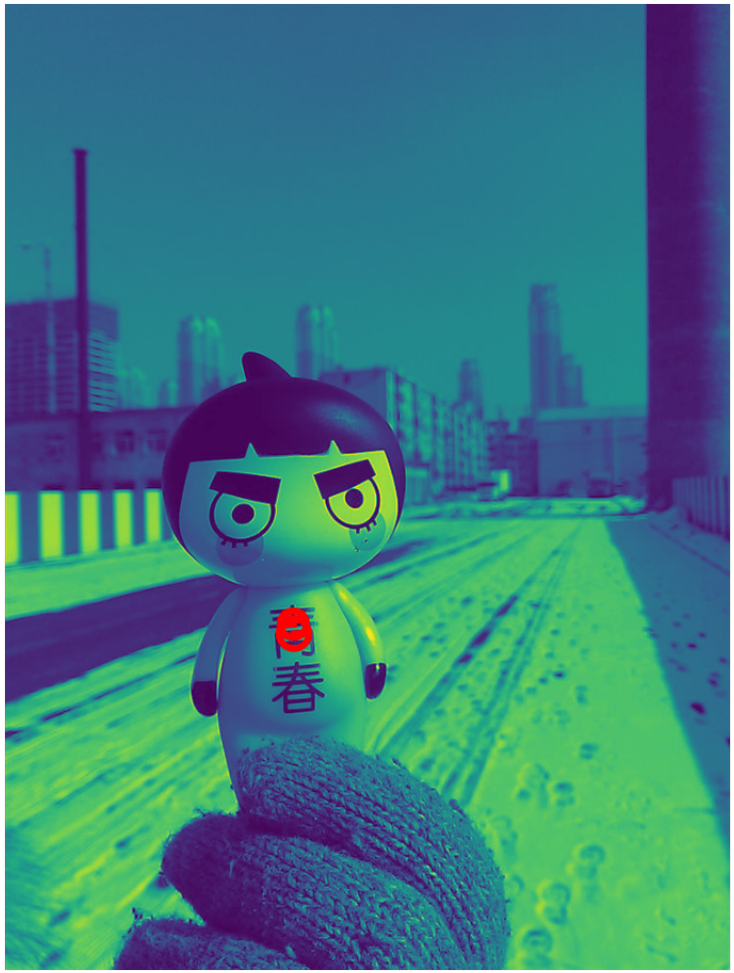
local features with the same visual word

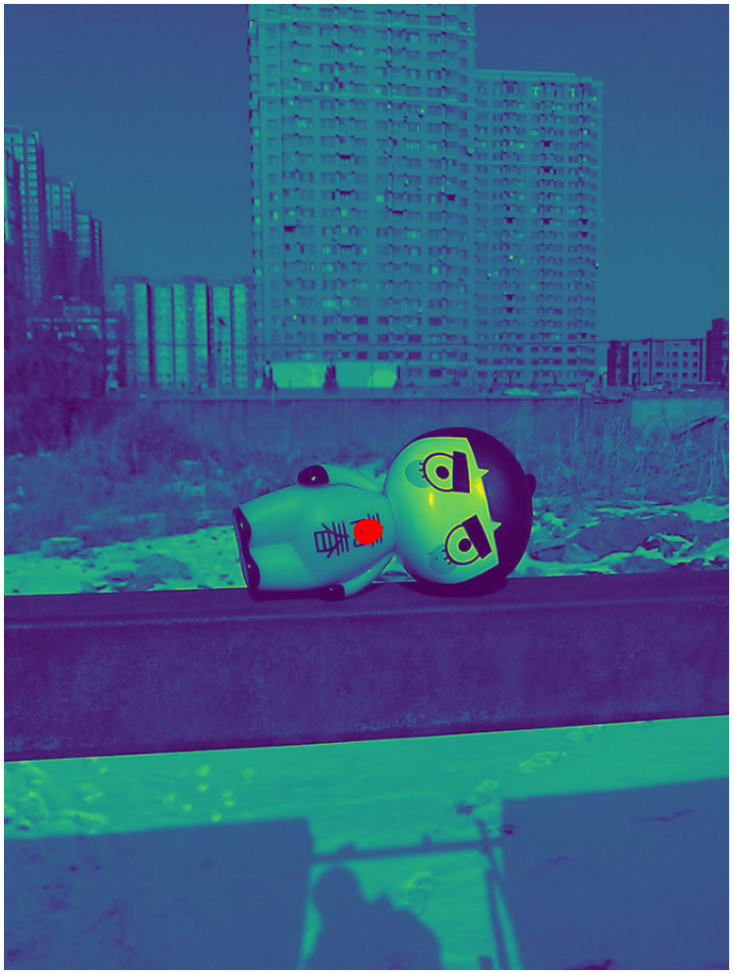


local features with the same visual word

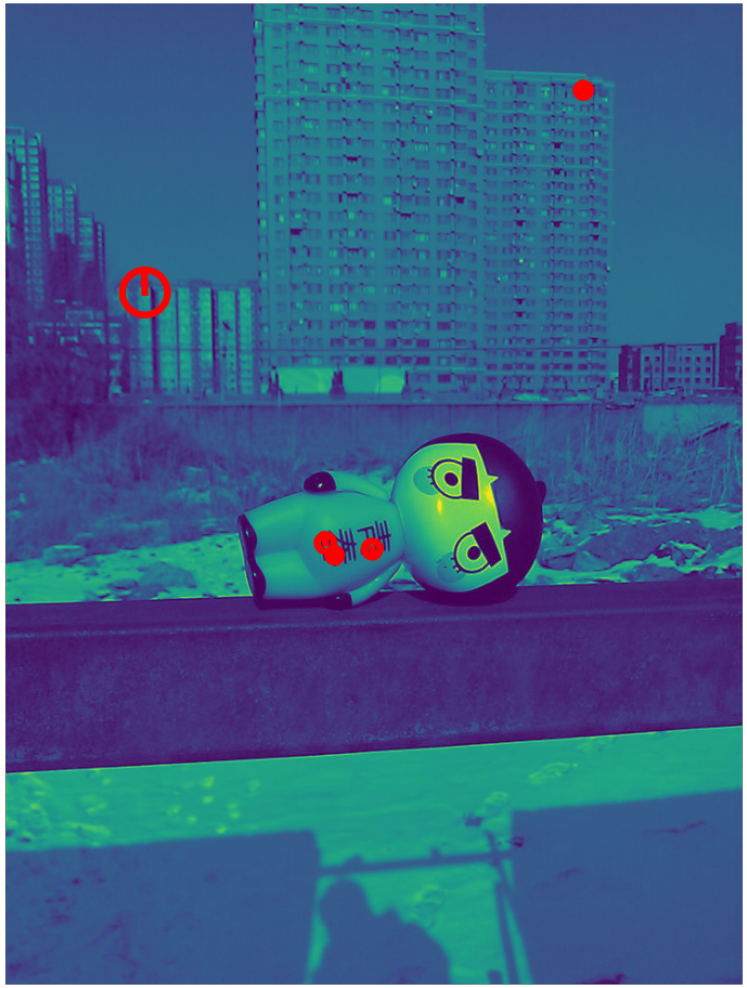


local features with the same visual word

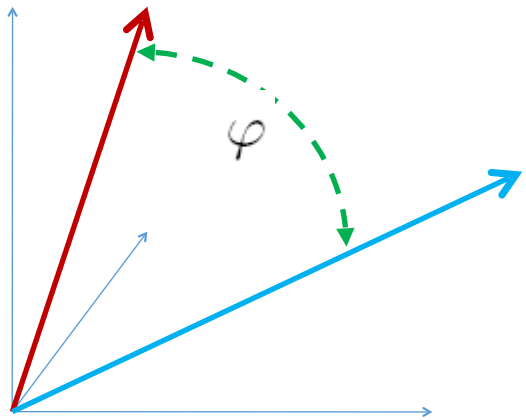




local features with the same visual word



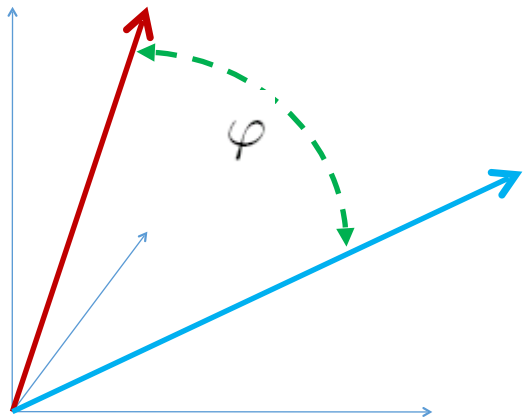
bag-of-words - similarity



bag of words representation

$$\cos\varphi = \frac{X^\top Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

bag-of-words - similarity

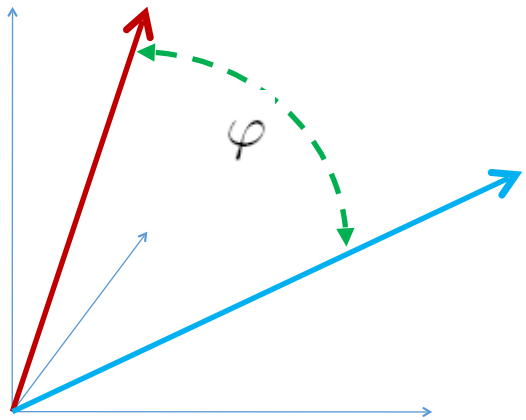


bag of words representation

$$\cos\varphi = \frac{X^\top Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database	query		score
	A B C D			
Y_1	(1 0 0 2)	A 0		s_1
		• X_q B 3	=	
		C 0		
		D 1		

bag-of-words - similarity

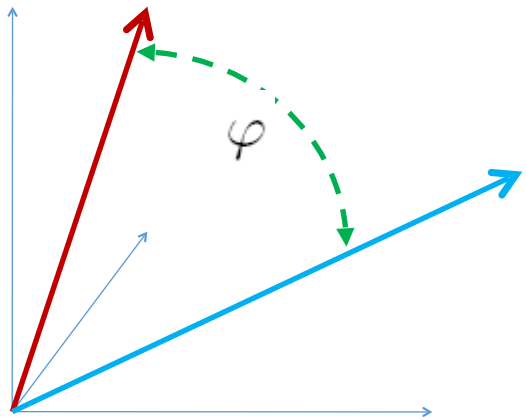


bag of words representation

$$\cos\varphi = \frac{X^T Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database	query	score
	A B C D		
Y_1	(1 0 0 2)	A 0	s_1
• X_q		B 3	=
		C 0	
		D 1	

bag-of-words - similarity

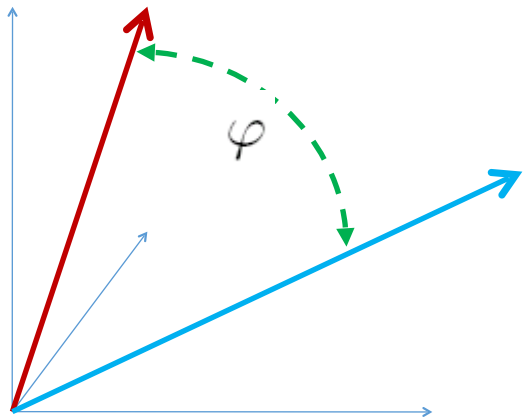


bag of words representation

$$\cos\varphi = \frac{X^T Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database					query	score
	A	B	C	D			
Y_1	1	0	0	2		A 0	s_1
Y_2	0	2	0	1	$\bullet\ X_q$	B 3 =	s_2
Y_3	1	0	0	0		C 0	s_3
	⋮					D 1	⋮

bag-of-words - similarity

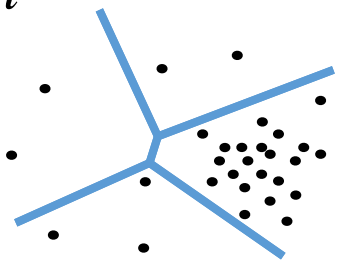


bag of words representation

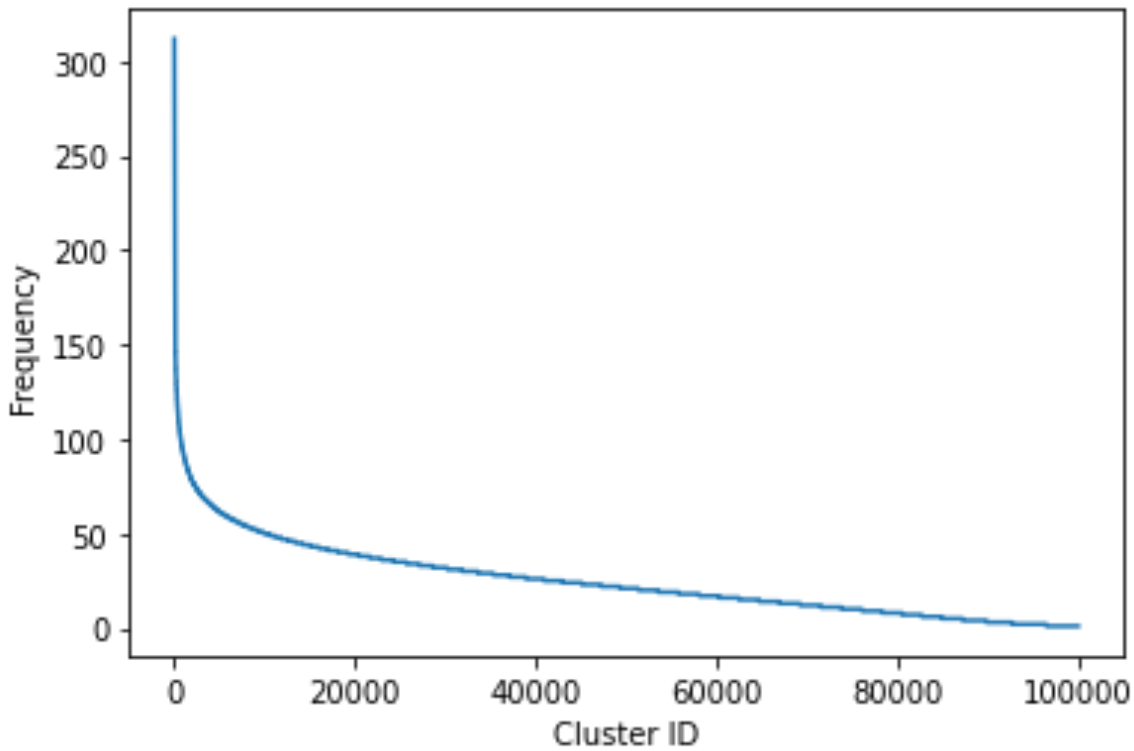
$$\cos\varphi = \frac{X^T Y}{||X||\ ||Y||} = \frac{1}{||X||}\frac{1}{||Y||}\sum_{i=1}^{K} X(i)Y(i)$$

	database					query	score
	A	B	C	D			
Y_1	1	0	0	2		A 0	s_1
Y_2	0	2	0	1	$\bullet\ X_q$	B 3 =	s_2
Y_3	1	0	0	0		C 0	s_3
	⋮					D 1	⋮

- inverse-document-frequency (idf):
  - multiply dimension i with $w_i = \log \frac{\#\text{ images}}{\#\text{ images with word } i}$
  - down-weighs highly populated visual words



cluster sizes



cluster size = 312

cluster size = 92

cluster size = 92

bag-of-words – retrieval

query

1
3
7
...
35678

visual words

inverted file

visual word	image ID
1	1 5 10 ... 7350125
2	2 7 12 ... 7399121
3	1 4 15 ... 7200190
...	...
40000	3 7 10 ... 7012245

similarity estimation & sorting

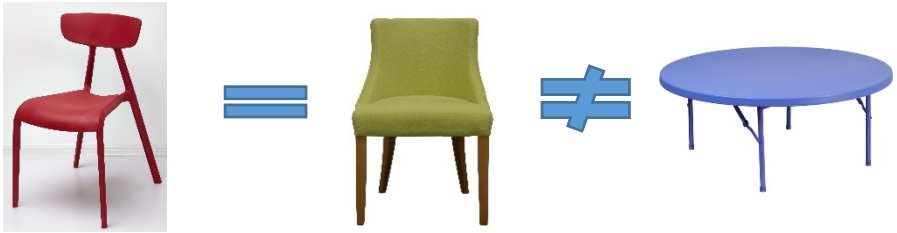
score	image ID
0.87	5
0.75	1573
0.52	11202
...	...
0.001	32

retrieval result

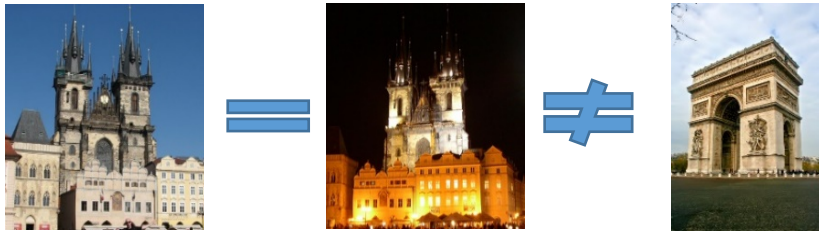
- efficient computation by skipping zero elements
- inverted-file structure
  - image-id list per visual word
    - contains images with that word
  - query time: visit only lists of words in the query

impact of codebook size

- small (eg. 128)
  - dense representation: compatible with approximate NN search
  - memory: #images x codebook-size x descriptor-element-footprint
  - appropriate for category-level matching

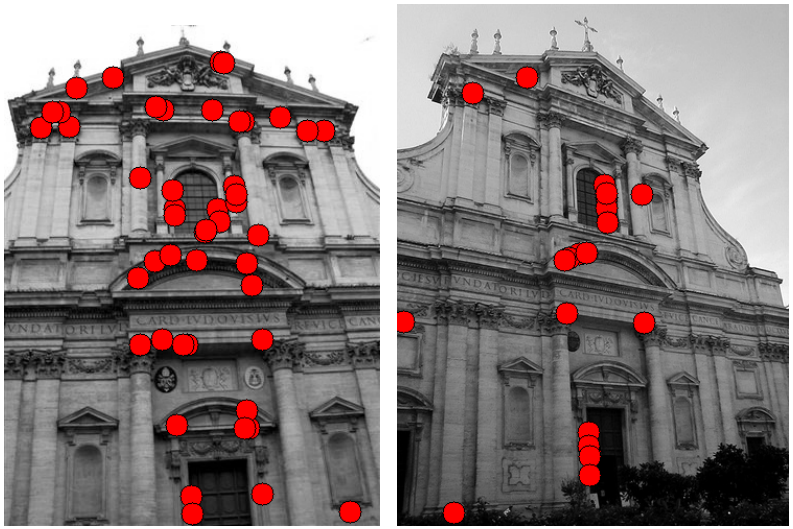


- very large (eg. 10^6)
  - sparse representation: use of inverted-file
  - memory: total-number-of-features x image-id-footprint
  - appropriate for instance-level matching

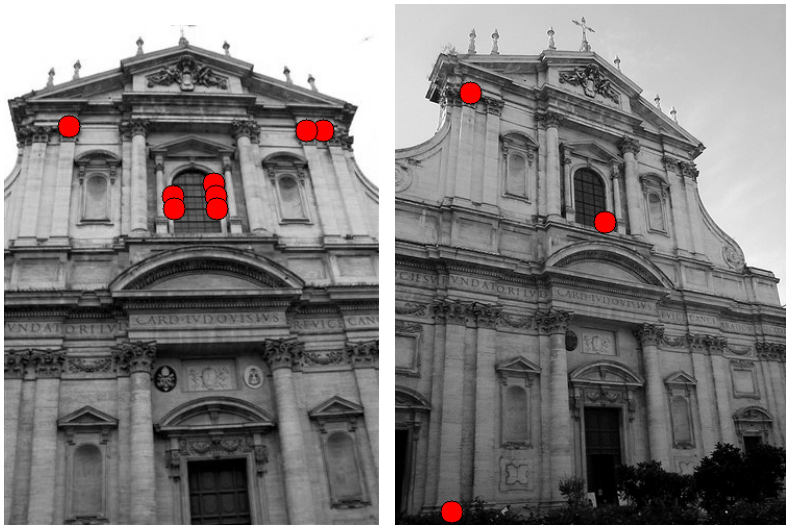


bag-of-words - example

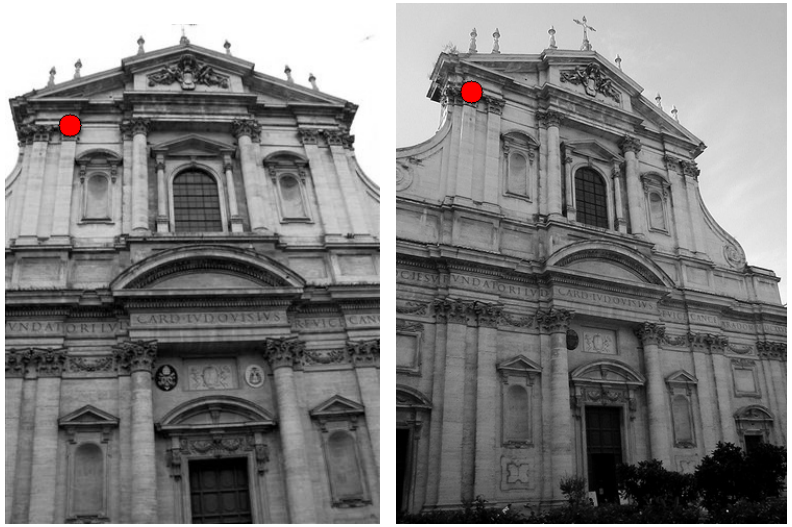
- local descriptors mapped to the same visual word for different codebook size



64 visual words



10^3 visual words



10^6 visual words

BoW & other approaches

from a different point of view

bag-of-words

$$S(\mathcal{X},\mathcal{Y}) \propto \left(\sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})\right)^\top \sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{y})$$

- local embedding: $\phi(\mathbf{x}) = [0,\ldots,1,\ldots,0]^\top,\ \mathbf{x}$ assigned to word $q(\mathbf{x})$

$q(\mathbf{x})$-th element

D [0 0 0 1]
D [0 0 0 1]
A [1 0 0 0]
B [0 1 0 0]

[1 1 0 2]

[1 0 1 1]

[0 0 0 1] D
[1 0 0 0] A
[0 0 1 0] C

bag-of-words

image similarity → $$\boxed{S(\mathcal{X},\mathcal{Y})} \propto \left(\sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})\right)^\top \sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{y})$$

- local embedding: $\phi(\mathbf{x}) = [0,\ldots,1,\ldots,0]^\top$, $\mathbf{x}$ assigned to word $q(\mathbf{x})$

$q(\mathbf{x})$-th element

D [0 0 0 1]
D [0 0 0 1]
A [1 0 0 0]
B [0 1 0 0]

[1 1 0 2]

[1 0 1 1]

[0 0 0 1] D
[1 0 0 0] A
[0 0 1 0] C

bag-of-words

image similarity → $\boxed{S(\mathcal{X},\mathcal{Y})} \propto \left(\sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})\right)^\top \boxed{\sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{y})}$ global descriptor

- local embedding: $\phi(\mathbf{x}) = [0,\ldots,\underset{q(\mathbf{x})\text{-th element}}{1},\ldots,0]^\top,\ \mathbf{x}$ assigned to word $q(\mathbf{x})$

D [0 0 0 1]
D [0 0 0 1]
A [1 0 0 0]
B [0 1 0 0]

[1 1 0 2]

[1 0 1 1]

[0 0 0 1] D
[1 0 0 0] A
[0 0 1 0] C

bag-of-words

image similarity

$$\boxed{S(\mathcal{X},\mathcal{Y})} \propto \left(\sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})\right)^\top \boxed{\sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{y})}$$

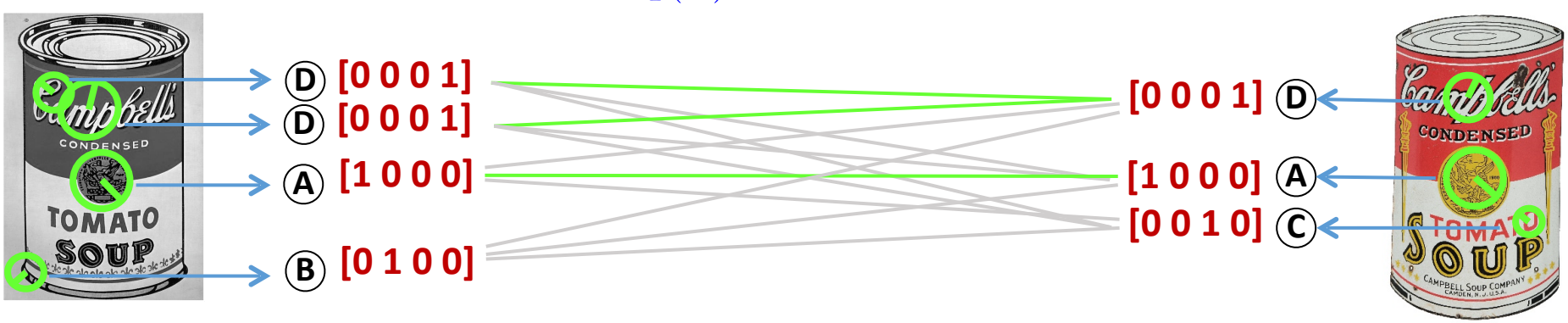
global descriptor

$$= \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x})^\top \phi(\mathbf{y})$$

voting approach

- local embedding: $\phi(\mathbf{x}) = [0,\ldots,1,\ldots,0]^\top$, $\mathbf{x}$ assigned to word $q(\mathbf{x})$

$q(\mathbf{x})$-th element



bag-of-words

image similarity

$$\boxed{S(\mathcal{X},\mathcal{Y})} \propto \left(\sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})\right)^\top \boxed{\sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{y})}$$

global descriptor

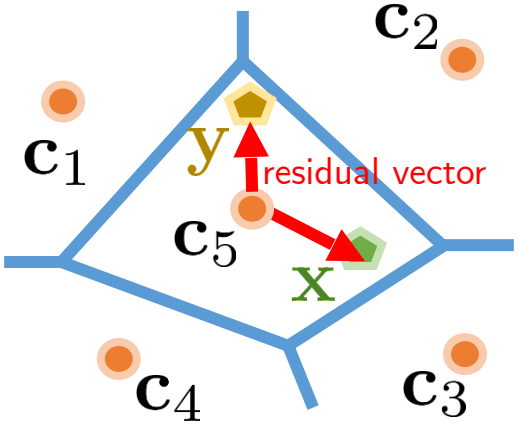
$$= \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x})^\top \phi(\mathbf{y})$$

voting approach

- local embedding: $\phi(\mathbf{x}) = [0,\ldots,\underbrace{1}_{q(\mathbf{x})\text{-th element}},\ldots,0]^\top,\ \mathbf{x}$ assigned to word $q(\mathbf{x})$
- same result achieved in two ways:
  - voting approach: accumulate local similarity of all pairs
  - dot product of two vectors (efficient)
- accumulate 0 if different visual word, 1 (or idf) if same visual word
  - what about more fine-grained votes?

vector of locally aggregated descriptors (VLAD)

- fine-grained votes equal to
  - 0, if $q(\mathbf{x}) \neq q(\mathbf{y})$ (different visual words)
  - $r(\mathbf{x})^\top r(\mathbf{y})$, if $q(\mathbf{x}) = q(\mathbf{y})$ (same visual word)
- residual vector $r(\mathbf{x}) = \mathbf{x} - \mathbf{c}_{q(\mathbf{x})}$
- local embedding $\phi(\mathbf{x}) = [\mathbf{0}^\top, \ldots, \underbrace{r(\mathbf{x})^\top}_{q(\mathbf{x})\text{-th sub-vector}}, \ldots, \mathbf{0}^\top]^\top$



[Jegou et al. CVPR'10]

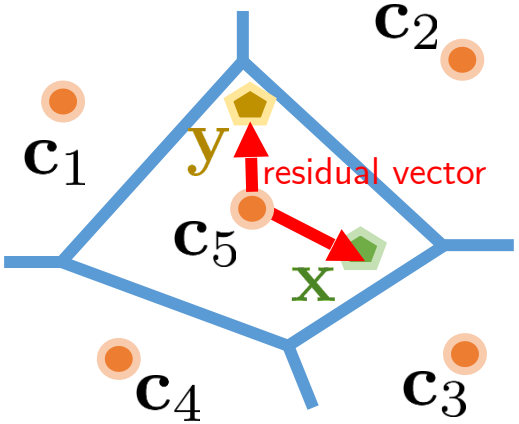
vector of locally aggregated descriptors (VLAD)

- voting approach
  - intuitive but inefficient

$$S(\mathcal{X},\mathcal{Y}) \propto \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x})^\top \phi(\mathbf{y})$$

- global descriptor similarity
  - efficient via dot product (dK dimensions)

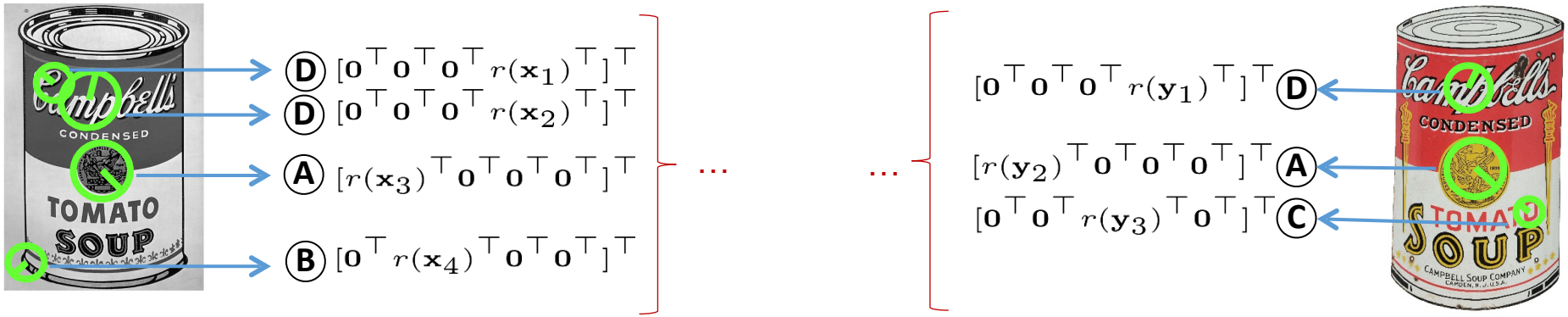
$$S(\mathcal{X},\mathcal{Y}) \propto \left(\sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) \right)^\top \sum_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{y})$$



[Jegou et al. CVPR'10]

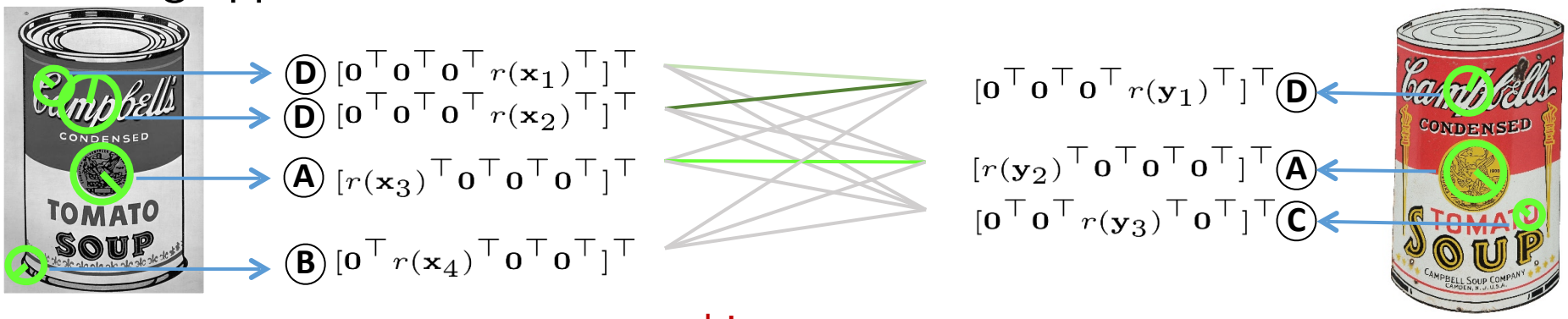
vector of locally aggregated descriptors (VLAD)

- global descriptor similarity



- sum to obtain global descriptors
- dot product for similarity

- voting approach



- non-binary votes
- according to descriptor similarity

VLAD and related approaches

- VLAD variations
  - use L2 normalized residuals [Delhumeau et al. ACMMM'13]
    - only angles matter
  - L2 normalization of each d-dimensional VLAD sub-vector
    - handles features from repeating patterns
    - [Arandjelovic & Zisserman, CVPR'13]

- Fisher vectors [Perronnin et al. CVPR'10]
  - Gaussian-Mixture-Model instead of k-means
  - different derivation: gradient log-likelihood of local descriptors
  - soft-assign to all visual words
  - residual vectors normalized (divided) by cluster variance

selective match kernel (SMK)

- use a local similarity function to estimate votes

$$S(\mathcal{X},\mathcal{Y}) \propto \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} \hat{s}(\phi(\mathbf{x})^\top \boxed{\phi(\mathbf{y})})$$

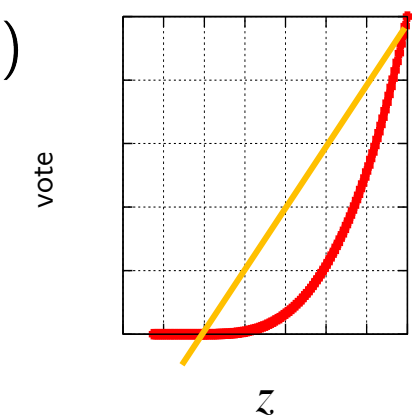
residual vector

- equivalently iterate over common visual words $\mathcal{W}$ between $\mathcal{X}$ and $\mathcal{Y}$

$$S(\mathcal{X},\mathcal{Y}) \propto \sum_{k\in\mathcal{W}} \sum_{\mathbf{x}\in\mathcal{X}_k} \sum_{\mathbf{y}\in\mathcal{Y}_k} s(\phi(\mathbf{x})^\top \phi(\mathbf{y})), \text{ where } \mathcal{X}_k = \{\mathbf{x} : q(\mathbf{x}) = k\}$$

- non-linear local similarity function (more selective votes)

$$s(z) = \begin{cases} z^\alpha, & \text{if } z > t \\ 0, & \text{otherwise} \end{cases}$$



- only seen as a voting approach
  - no equivalence to global descriptor. why?

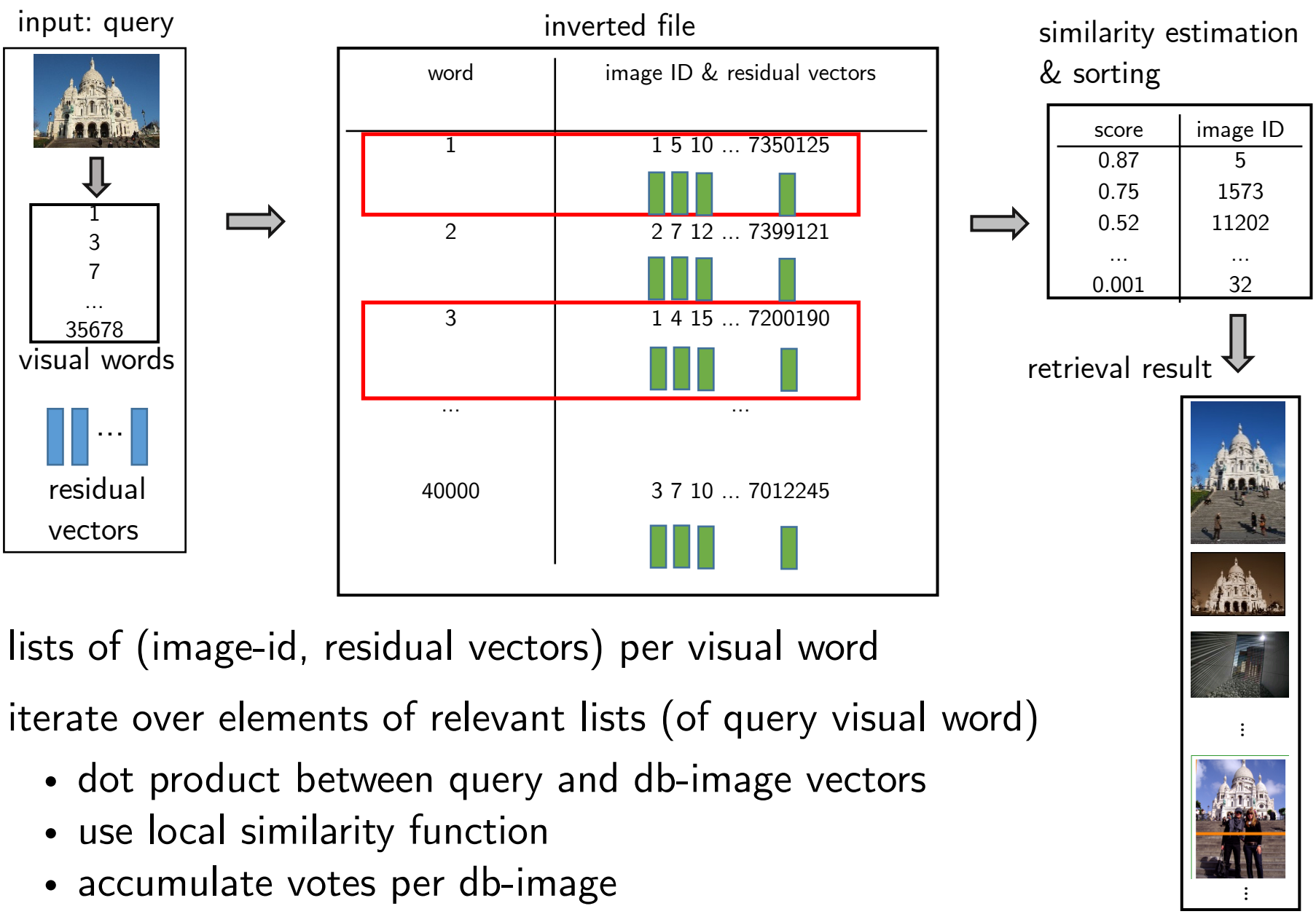
[Tolias et al. ICCV'13]

impact of the local similarity function

- numerical example:

- exponent $\alpha = 1$
  - [0.5 0.5 0.5 0.5 0.5] → 2.5
  - [0.7 0.7 0.7 0.2 0.2] → 2.5
- exponent $\alpha = 2$
  - [0.5 0.5 0.5 0.5 0.5] → 1.25
  - [0.7 0.7 0.7 0.2 0.2] → 1.55
- exponent $\alpha = 3$
  - [0.5 0.5 0.5 0.5 0.5] → 0.62
  - [0.7 0.7 0.7 0.2 0.2] → 1.05

SMK – retrieval



- lists of (image-id, residual vectors) per visual word
- iterate over elements of relevant lists (of query visual word)
  - dot product between query and db-image vectors
  - use local similarity function
  - accumulate votes per db-image

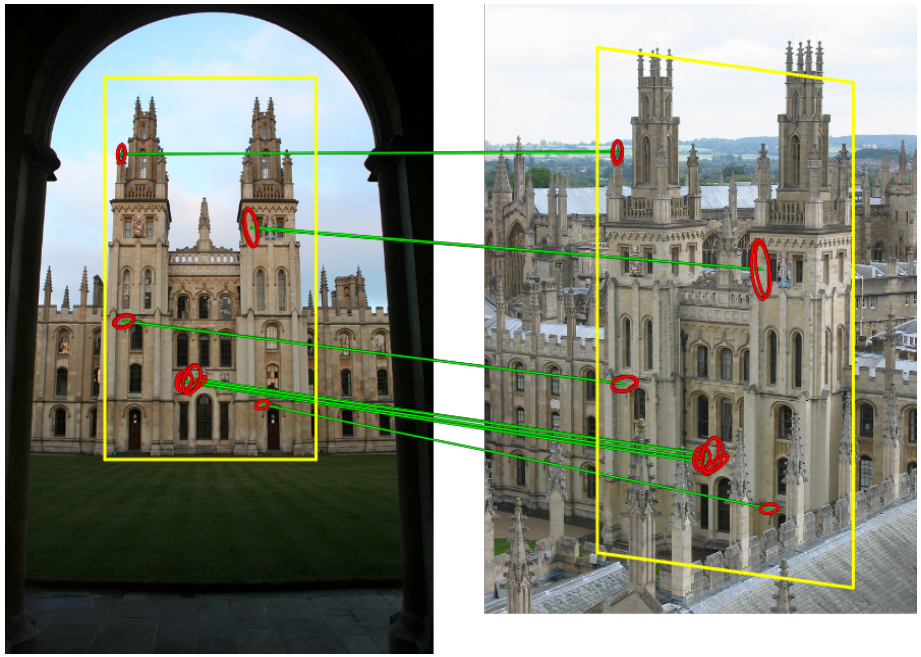
settings for identical object retrieval

- BoW
  - very large codebook (eg. 10^6)
  - compatible with inverted-file
  - lower memory requirements than SMK but less precise
  - provides tentative correspondences
- VLAD
  - small codebook (eg. 128)
  - dimensionality reduction
  - small memory footprint
  - efficient search
    - compatible with approximate nearest neighbor search methods
- SMK
  - large codebook (eg. 65k)
  - compatible with inverted-file
  - memory demanding and slower search than BoW but better performance
  - provides tentative correspondences

fast spatial matching

spatial matching and verification

- RANSAC-like approach to find inlier correspondences
  - find the geometric transformation with the largest number of inliers
  - use the number of inliers as image-to-image similarity
- 1st stage of retrieval: rank all images with BoW
- 2nd stage: re-rank top-ranked images according to the number of inliers
  - efficiency is important → method inspired by RANSAC but sped-up

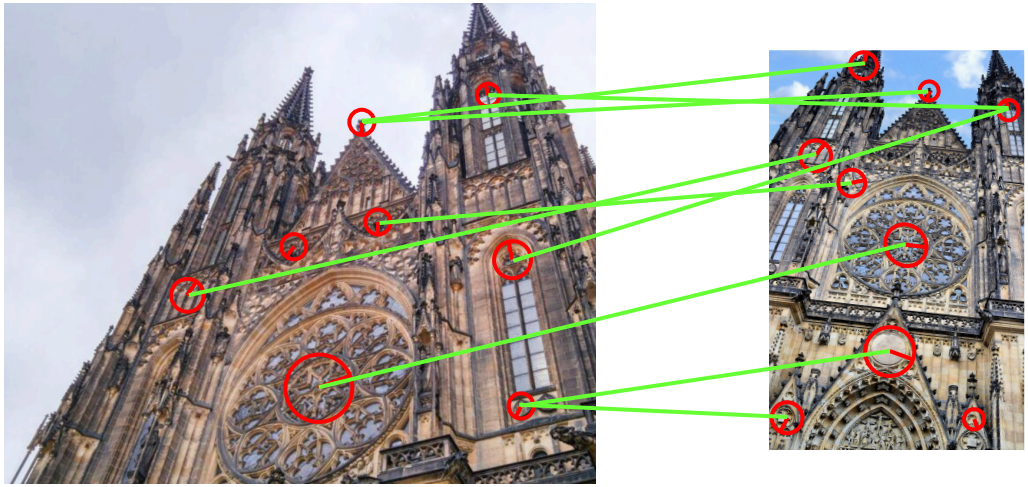


query top retrieved

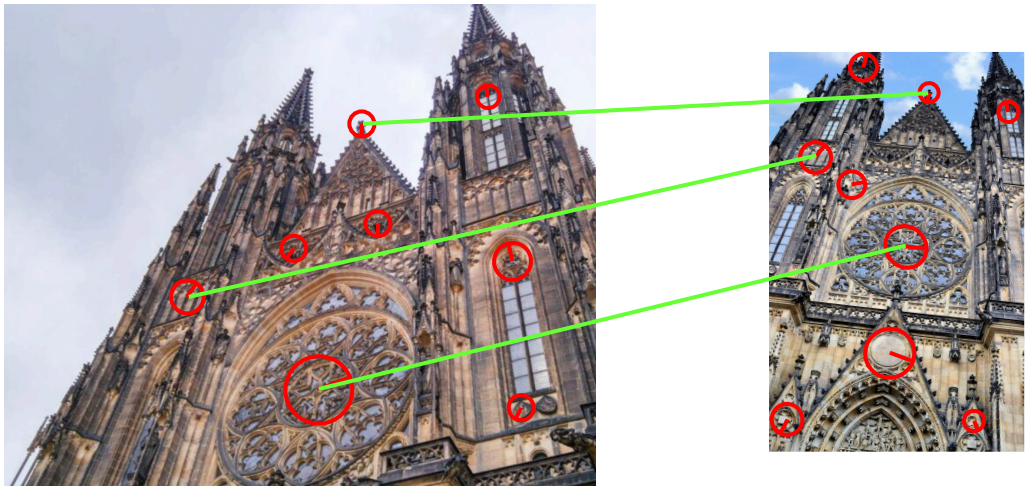
[Philbin et al. CVPR'07] [Perdoch et al. CVPR'09]

spatial verification: input - output

input:
tentative correspondences
based on visual words



out:
inlier correspondences
also geometric transformation
but that's not what we are after



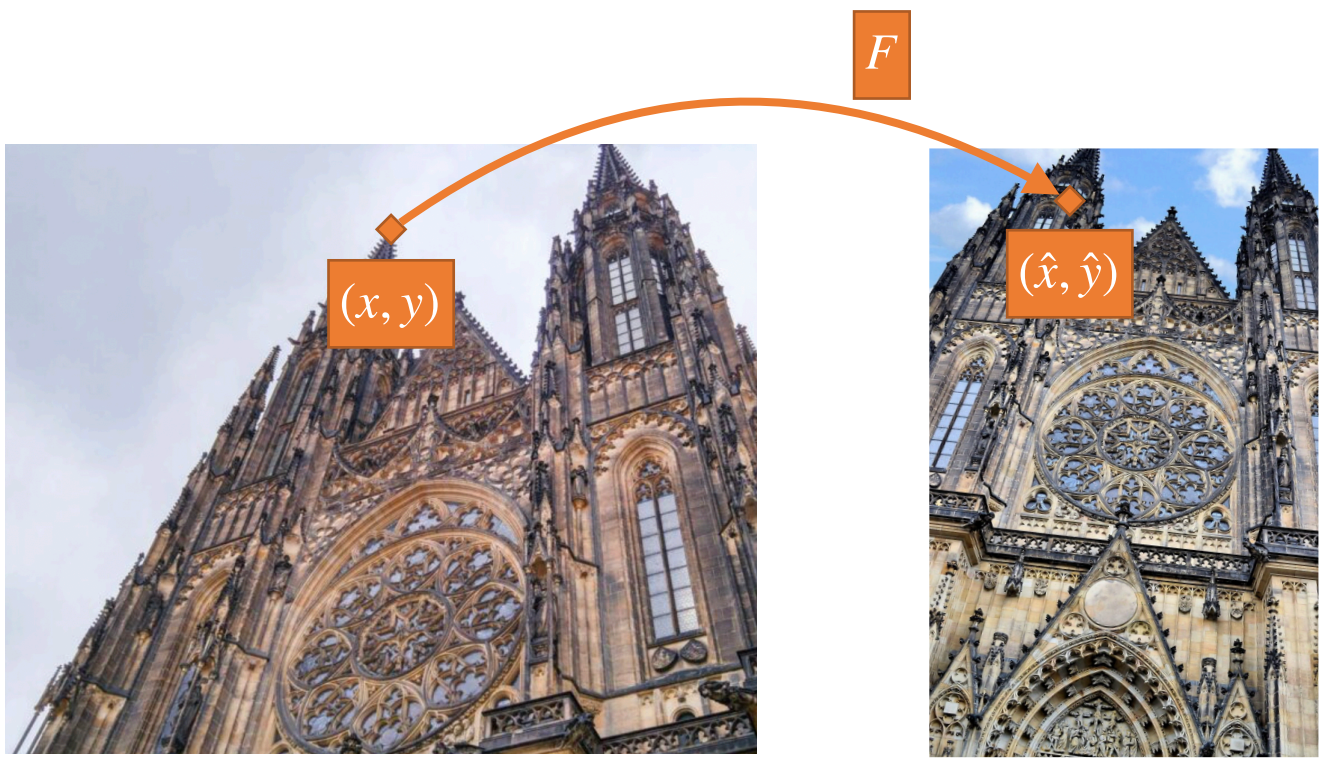
[Philbin et al. CVPR'07] [Perdoch et al. CVPR'09]

preliminaries: 2D image-to-image transformations

- 2D transformations as linear transformations in homogeneous coordinates

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = F \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- homogeneous coordinates: $[x, y, 1]^\top$
- represent a 2D transformation as a 3x3 matrix F



preliminaries: 2D transformations

- translation in 2D cartesian coordinates

$$\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} + \begin{bmatrix} x \\ y \end{bmatrix}$$

- translation in homogenous coordinates

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- rotation in homogenous coordinates

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- rotation & translation in homogenous coordinates

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & \Delta x \\ sin\theta & cos\theta & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

preliminaries: 2D transformations

- Euclidean/rigid transformation (3 DoF) - preserves distances and angles

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & \Delta x \\ sin\theta & cos\theta & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- similarity transformation (4 DoF) - preserves angles, not distances

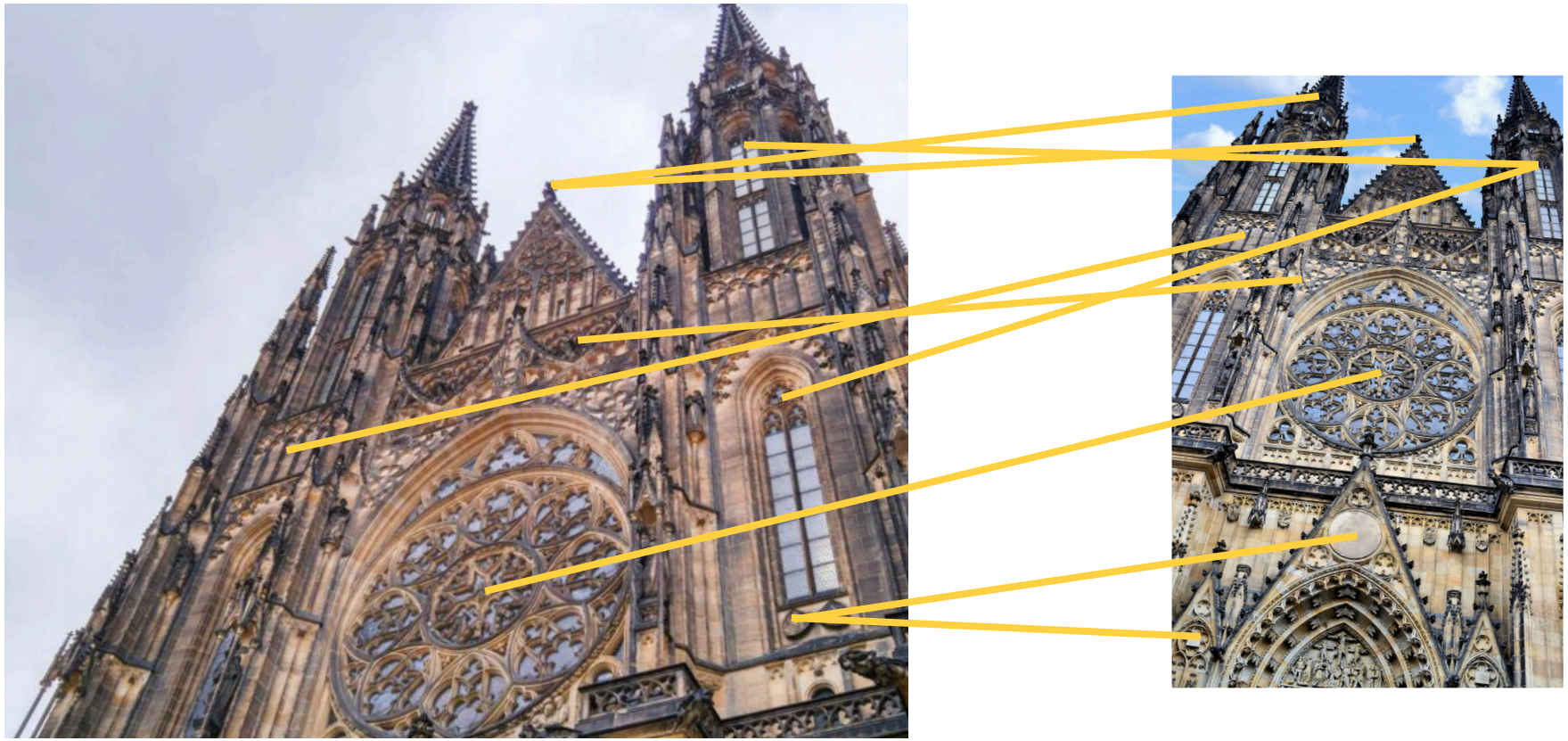
$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = \begin{bmatrix} s\ cos\theta & -s\ sin\theta & \Delta x \\ s\ sin\theta & s\ cos\theta & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- affine transformation (6 DoF) - preserves parallel lines, not angles/distances

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \Delta x \\ a_{21} & a_{22} & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

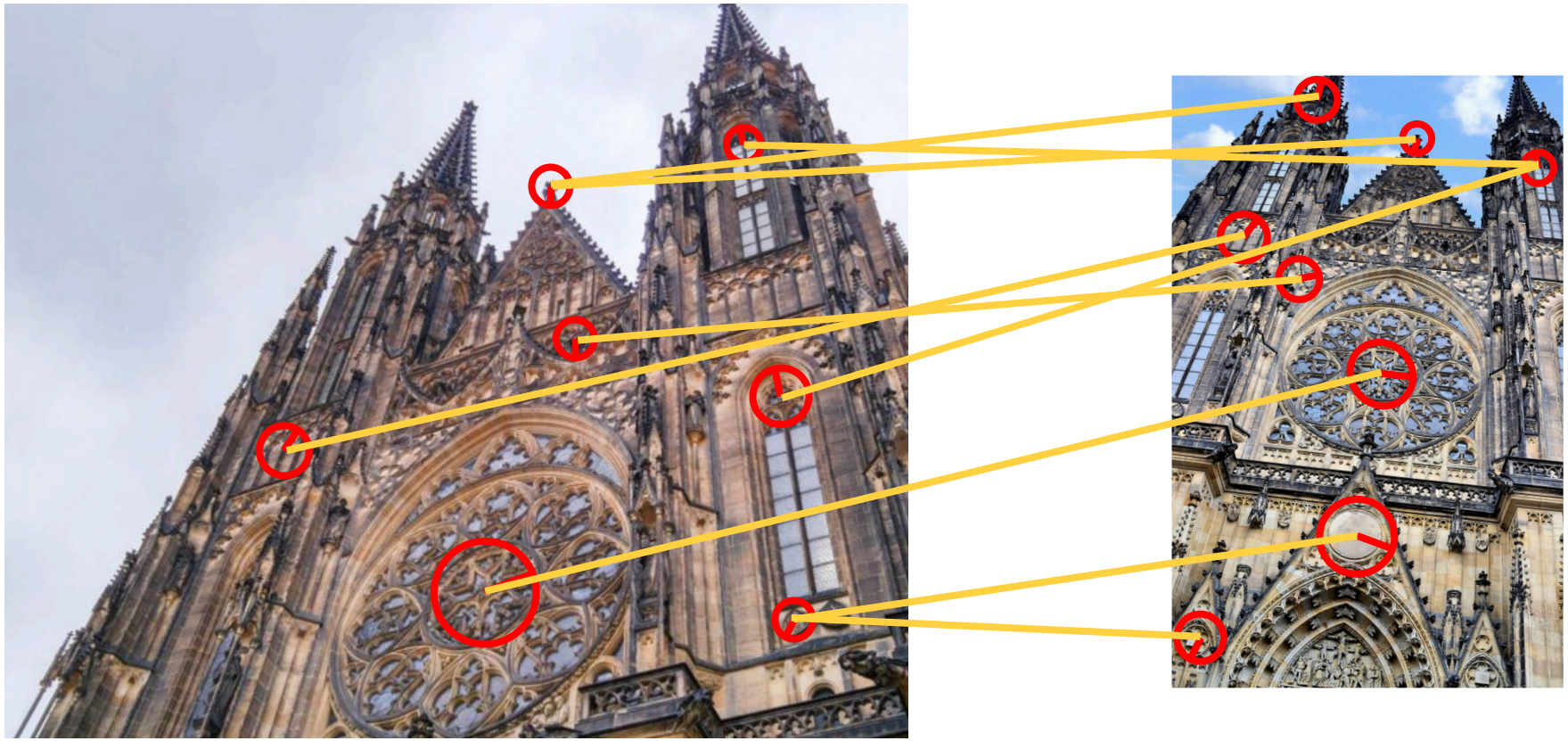
point correspondences vs local feature correspondences

- point correspondences: $(x, y) \leftrightarrow (x', y')$
- local feature correspondences: $(A, x, y) \leftrightarrow (A', x', y')$
  - A represents the shape of the local feature
  - eg. scale and rotation, affine shape



point correspondences vs local feature correspondences

- point correspondences: $(x, y) \leftrightarrow (x', y')$
- local feature correspondences: $(A, x, y) \leftrightarrow (A', x', y')$
  - A represents the shape of the local feature
  - eg. scale and rotation, affine shape

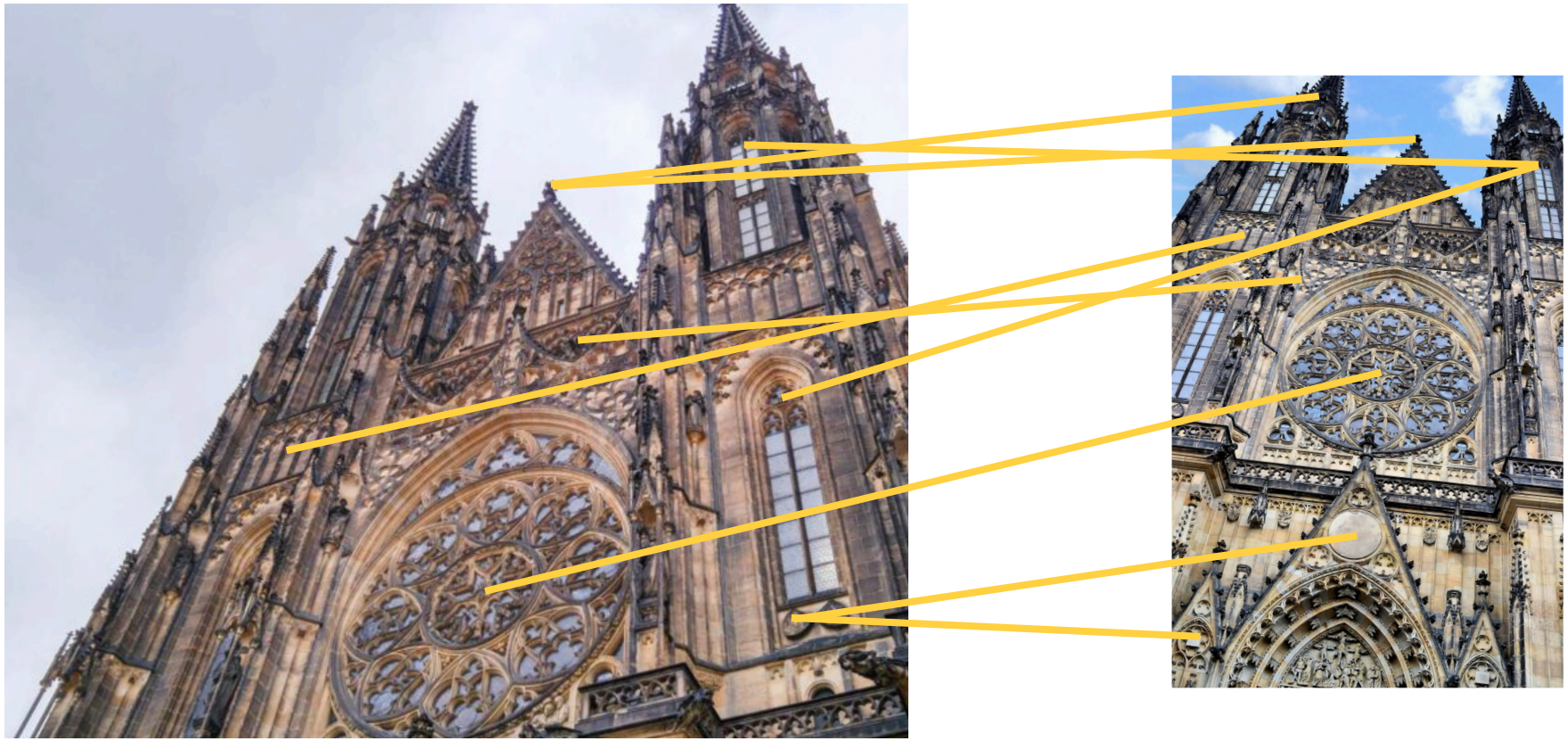


transformation hypothesis

- sample the minimum number of correspondences required to estimate the parameters of the geometric transformation

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = F \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

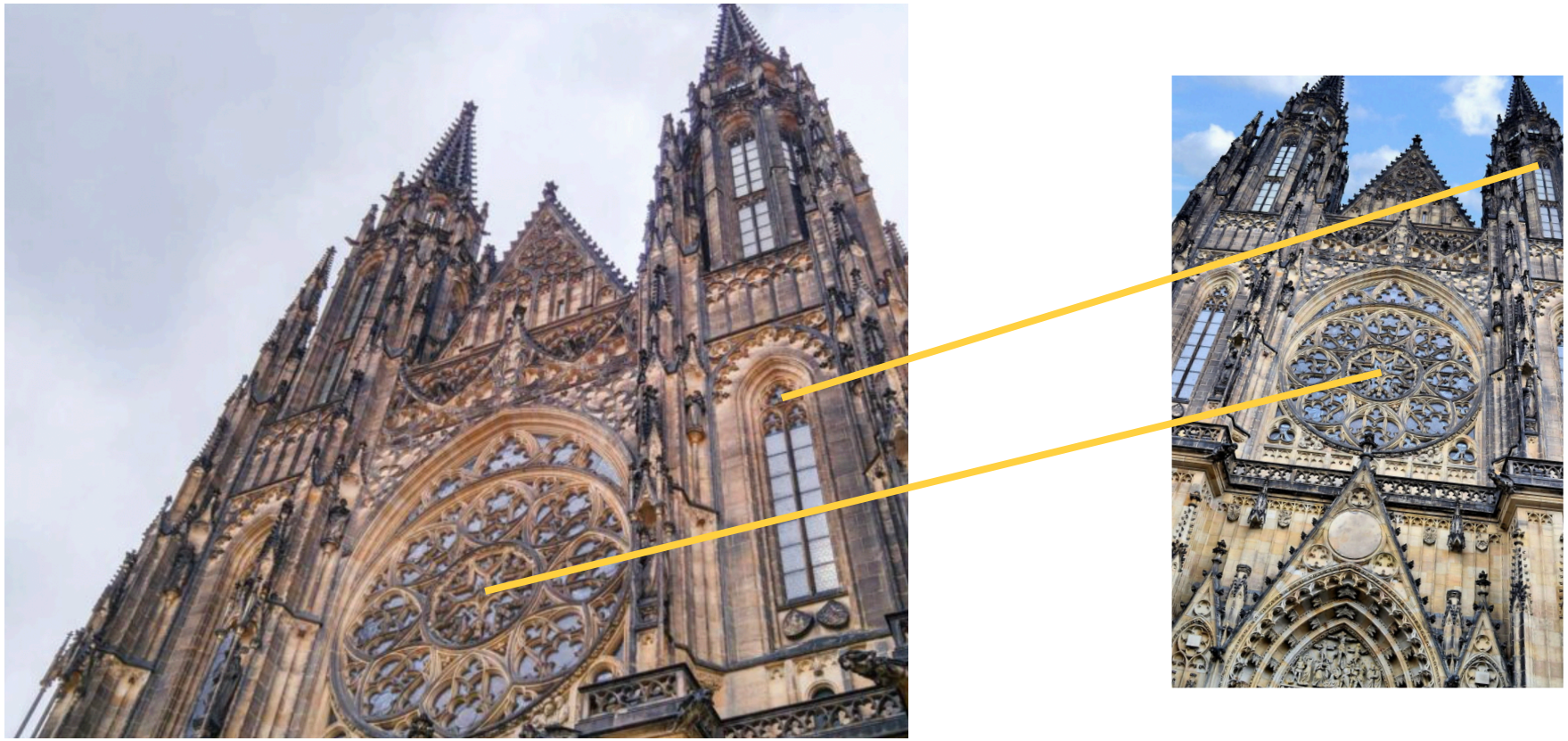
- one correspondece → two equations
  - similarity transformation (4 DoF) → 2 correspondences
  - affine transformation (6 DoF) → 3 correspondences



transformation hypothesis

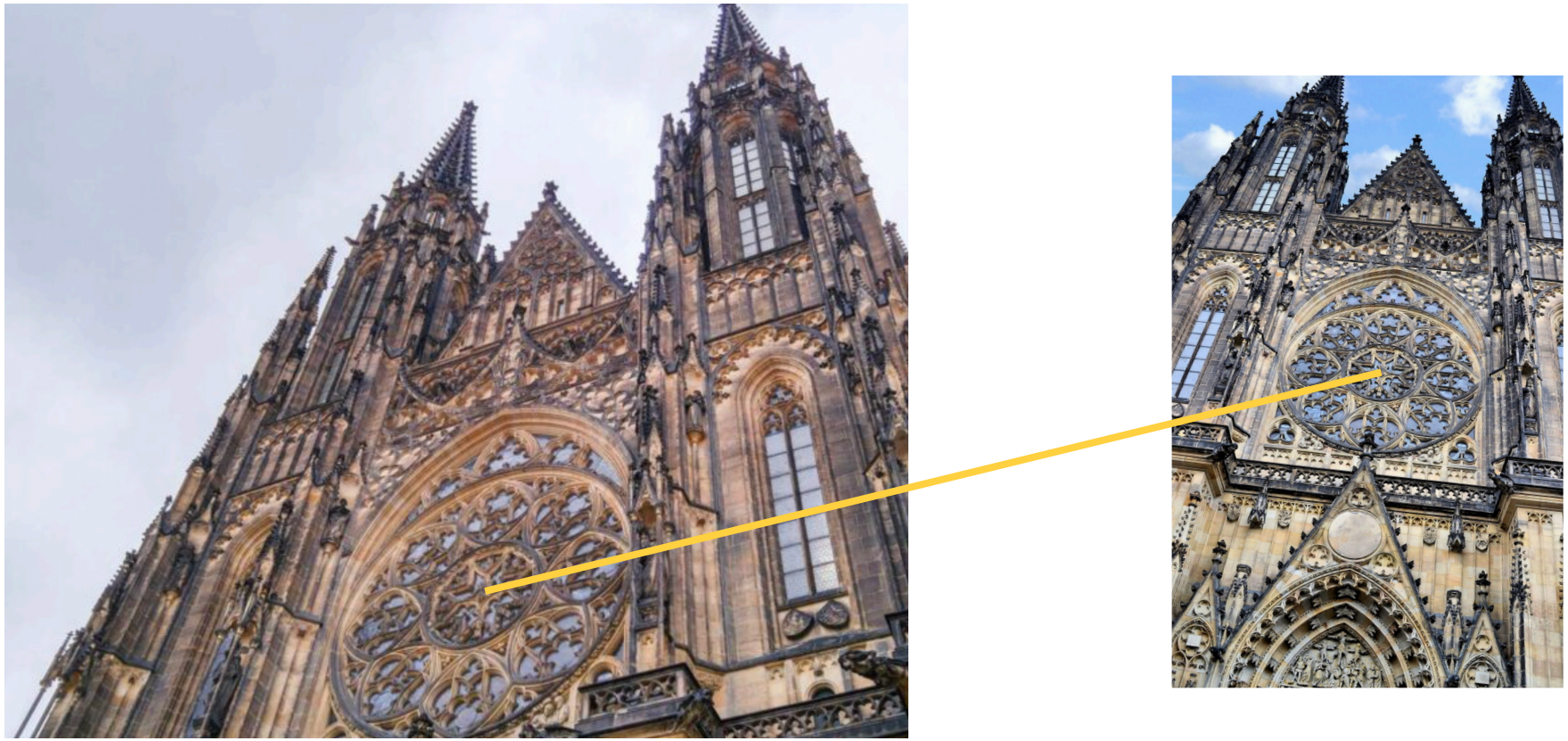
- sample the minimum number of correspondences required to estimate the parameters of the geometric transformation
- one correspondece → two equations
  - similarity transformation (4 DoF) → 2 correspondences
  - affine transformation (6 DoF) → 3 correspondences

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ 1 \end{bmatrix} = F \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$



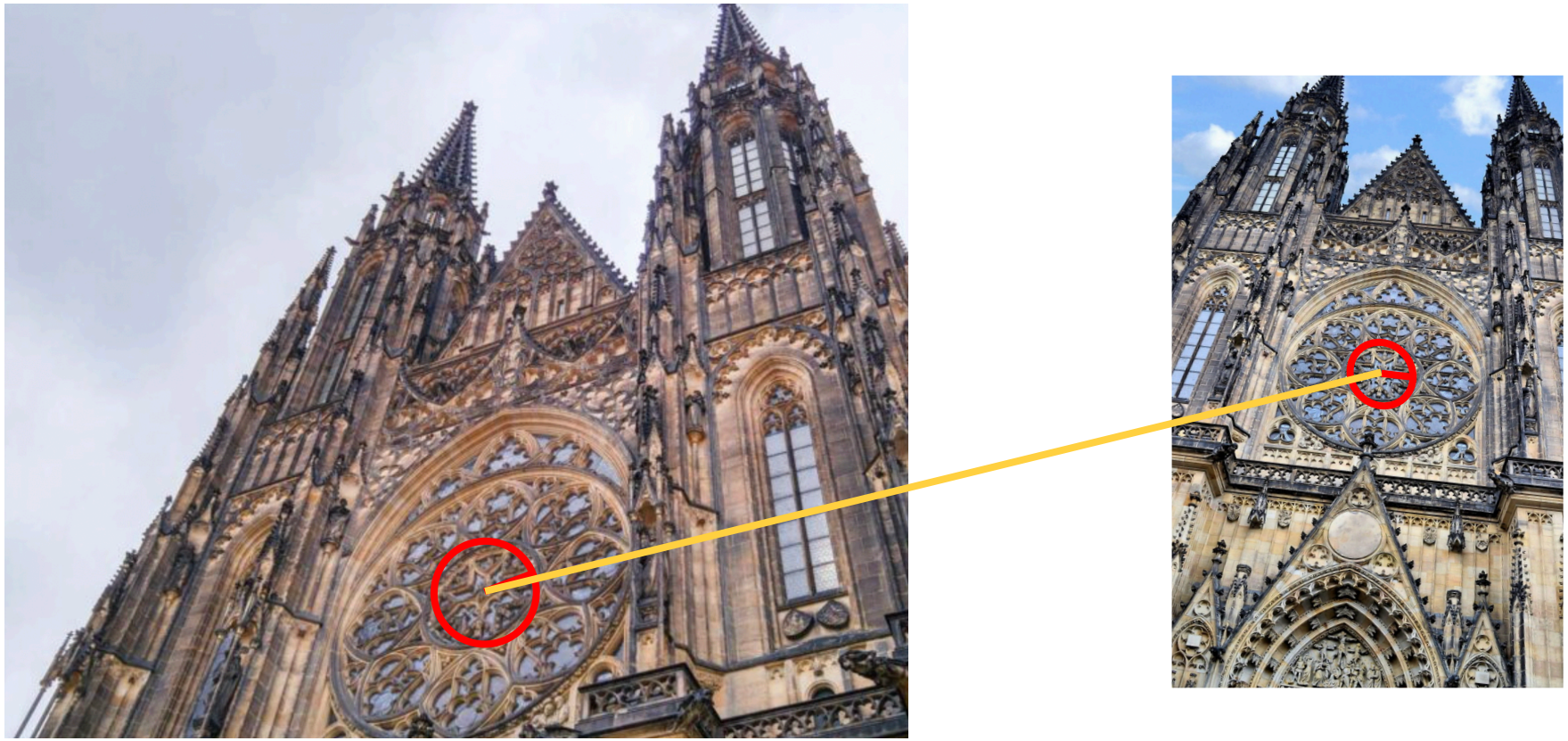
transformation hypothesis from a single correspondence

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



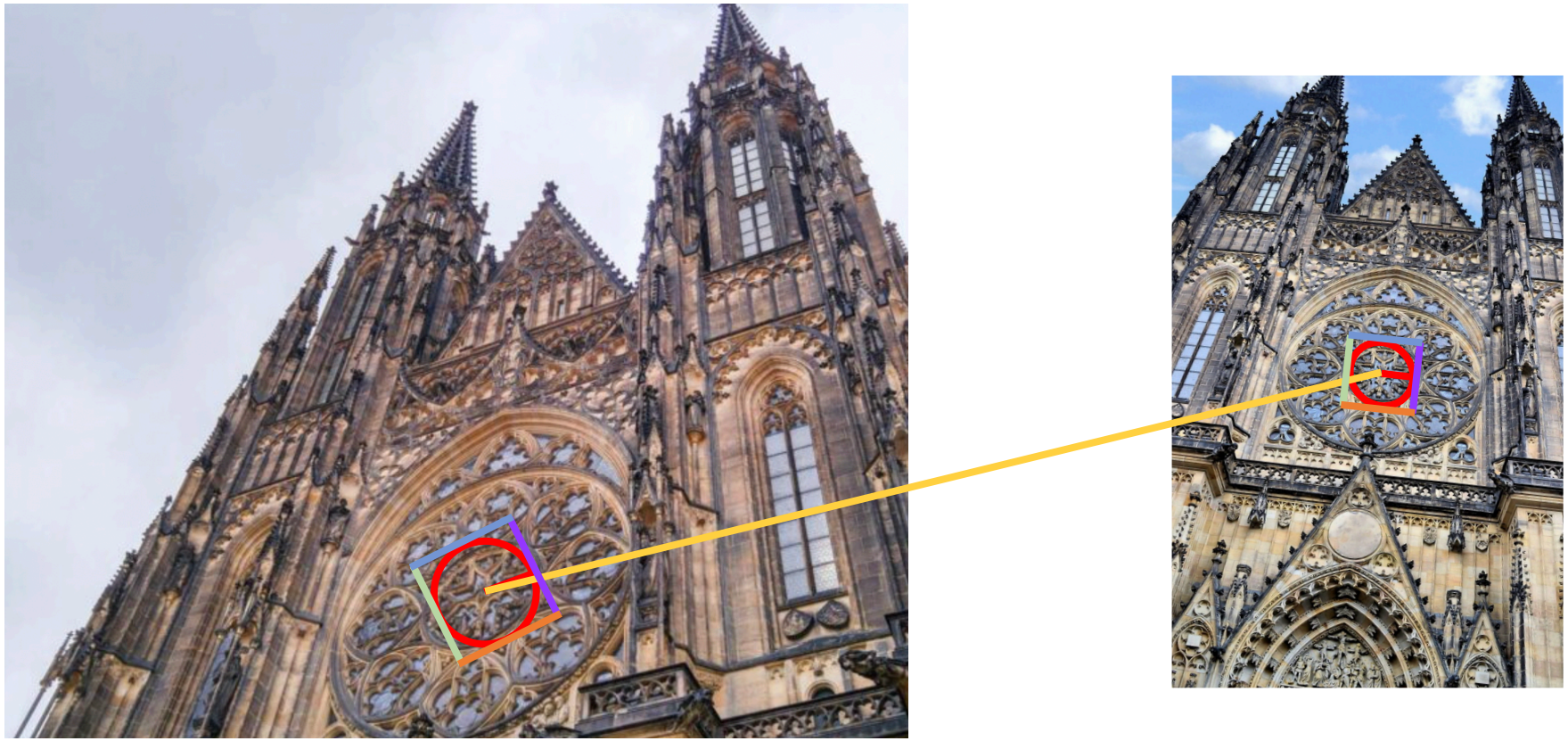
transformation hypothesis from a single correspondence

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



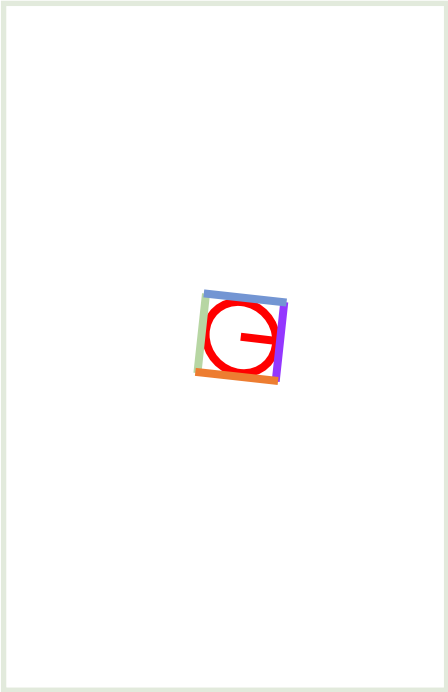
transformation hypothesis from a single correspondence

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



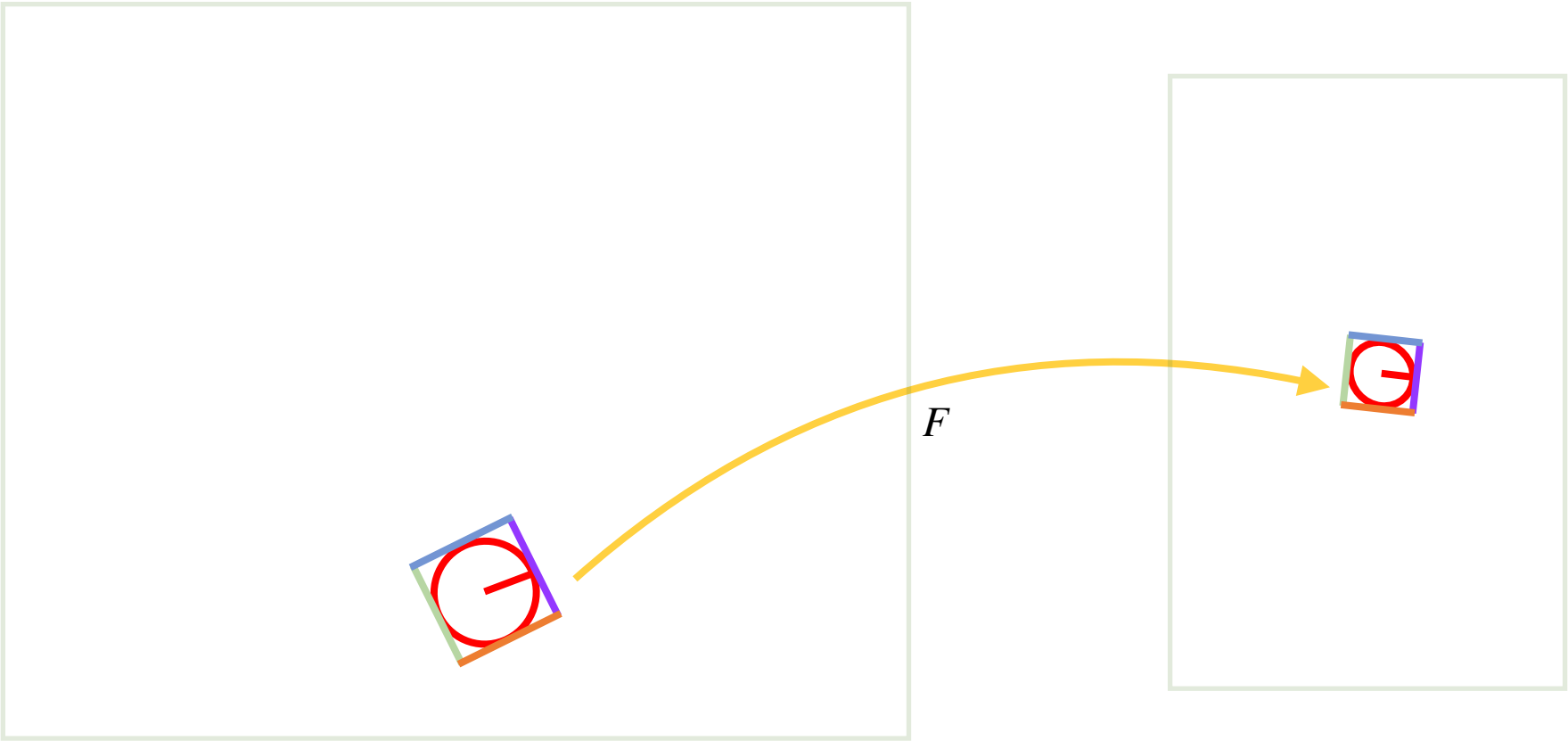
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



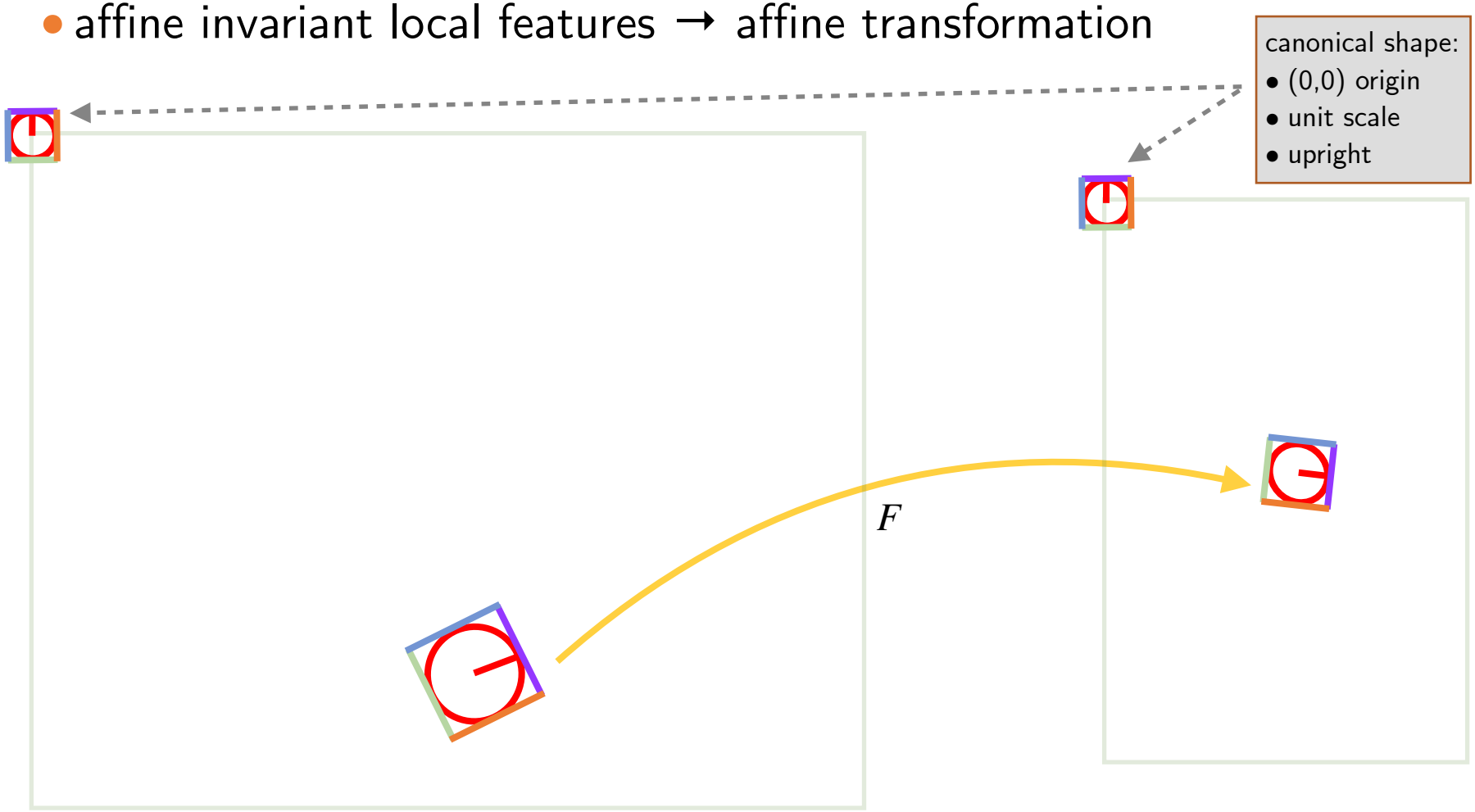
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



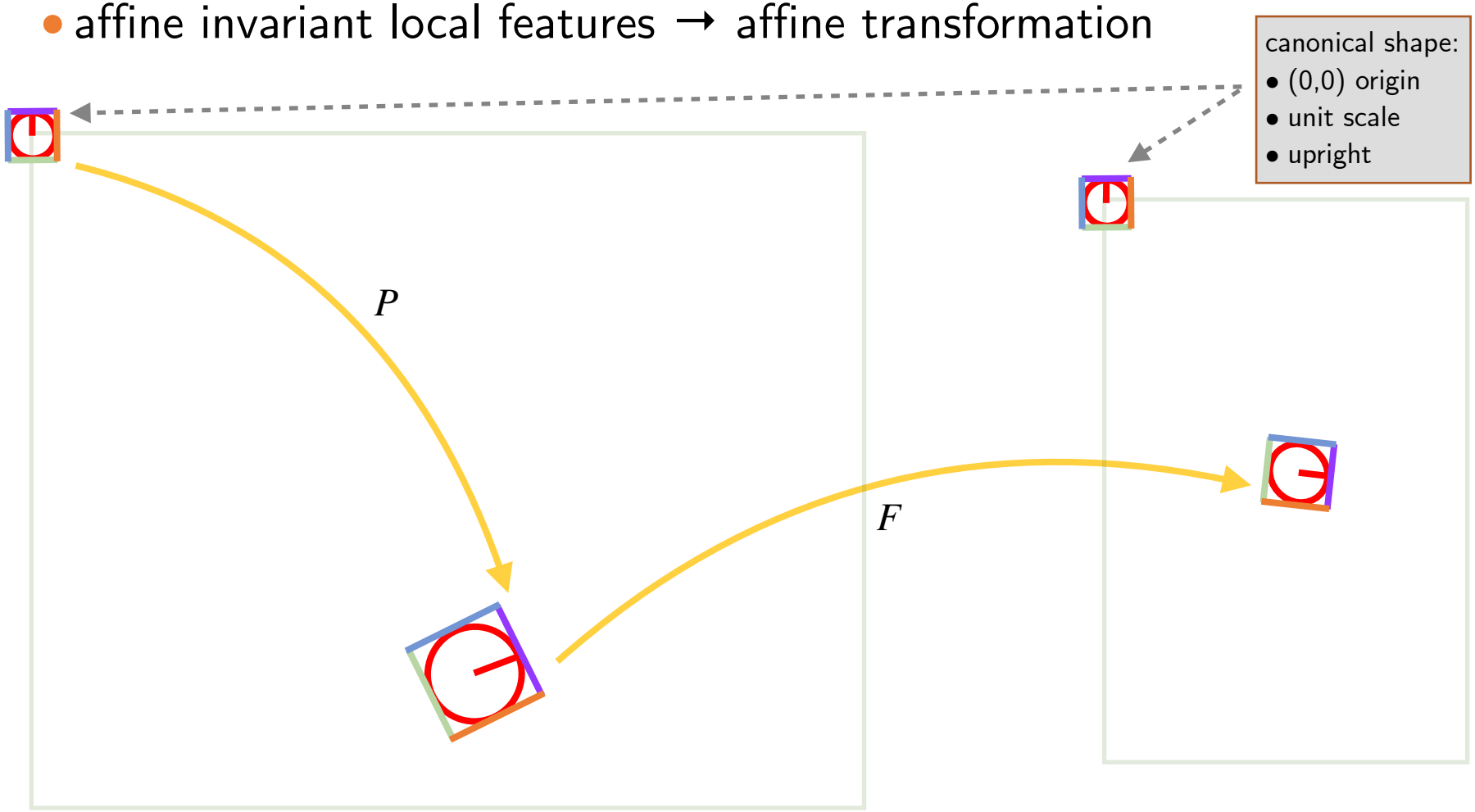
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



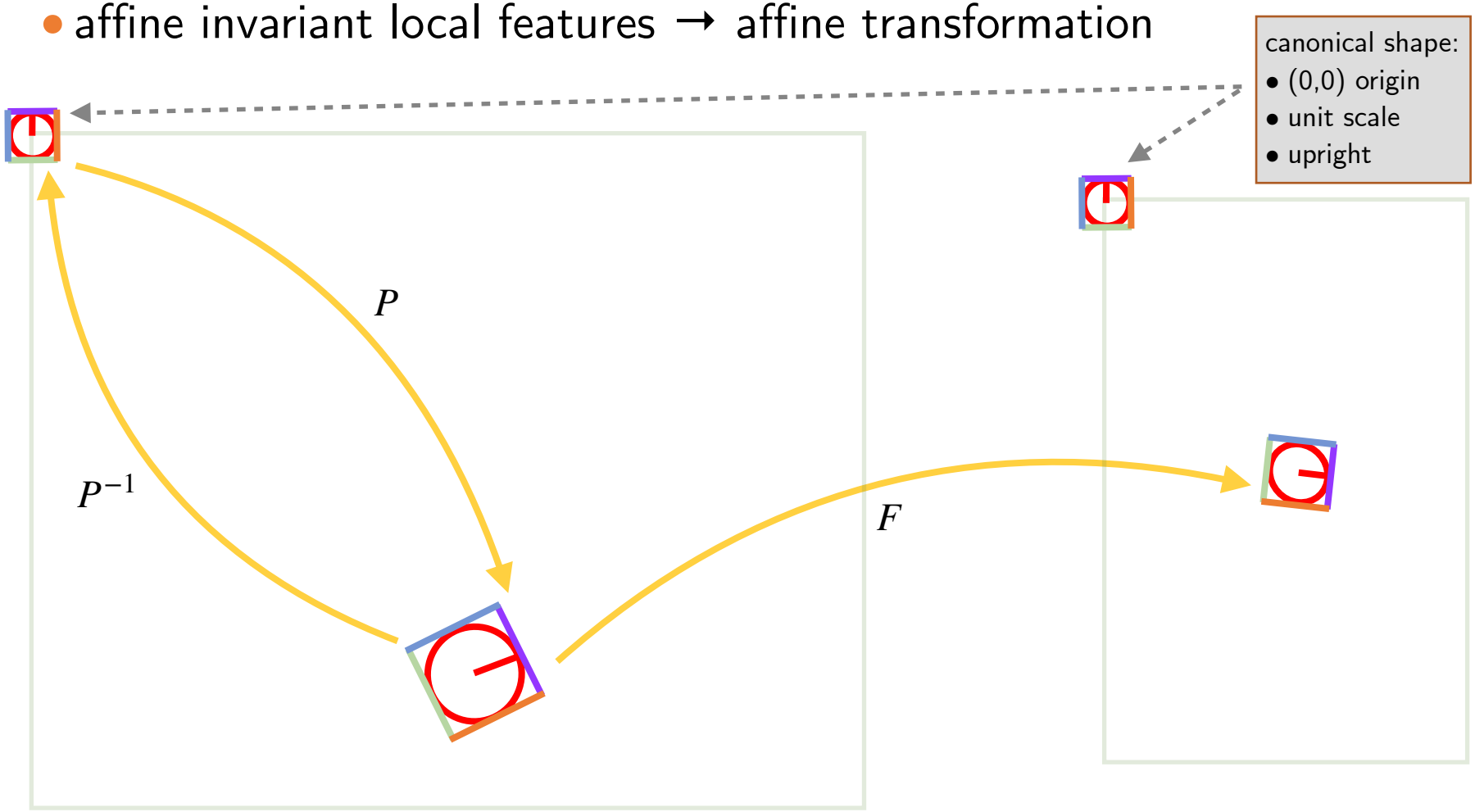
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



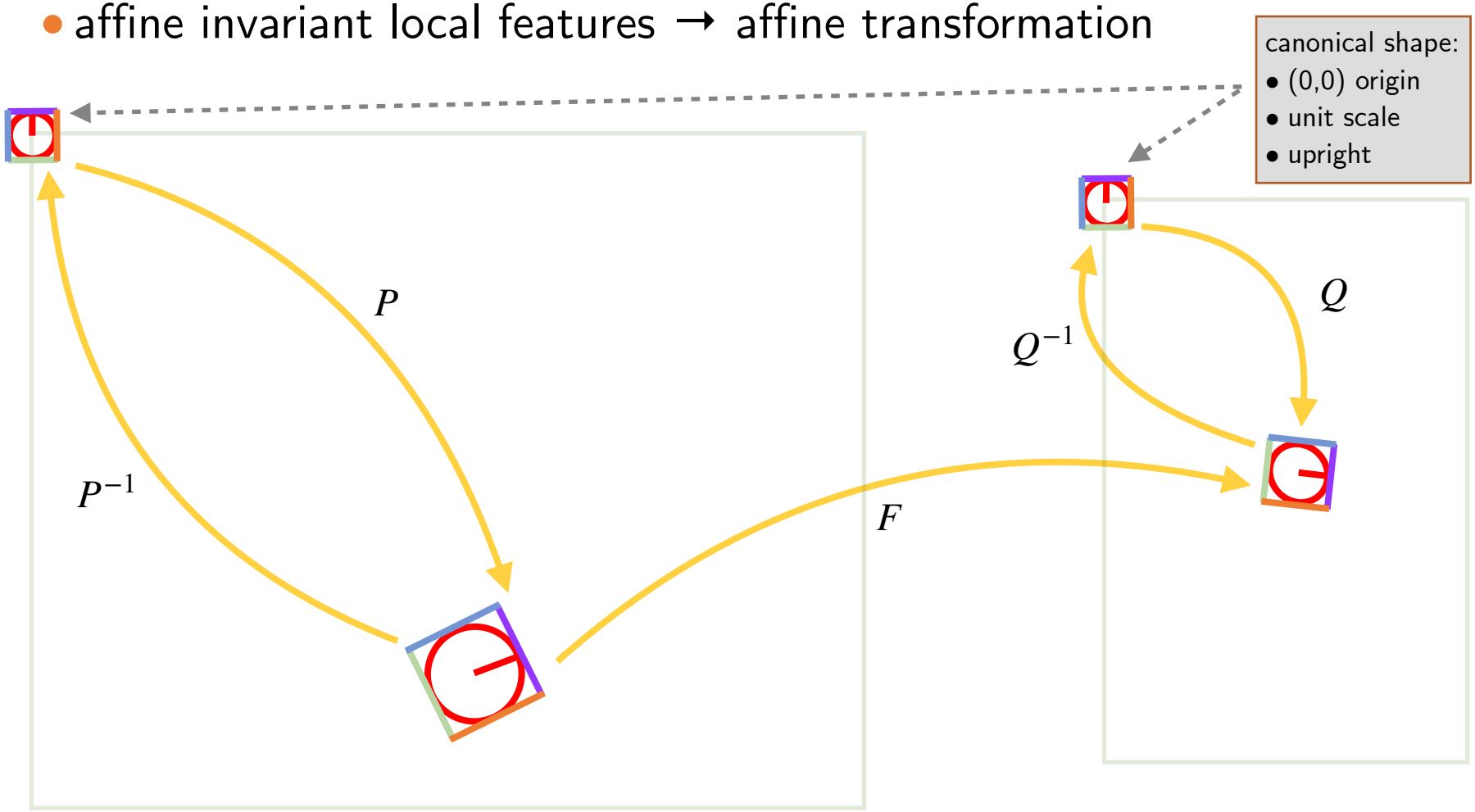
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



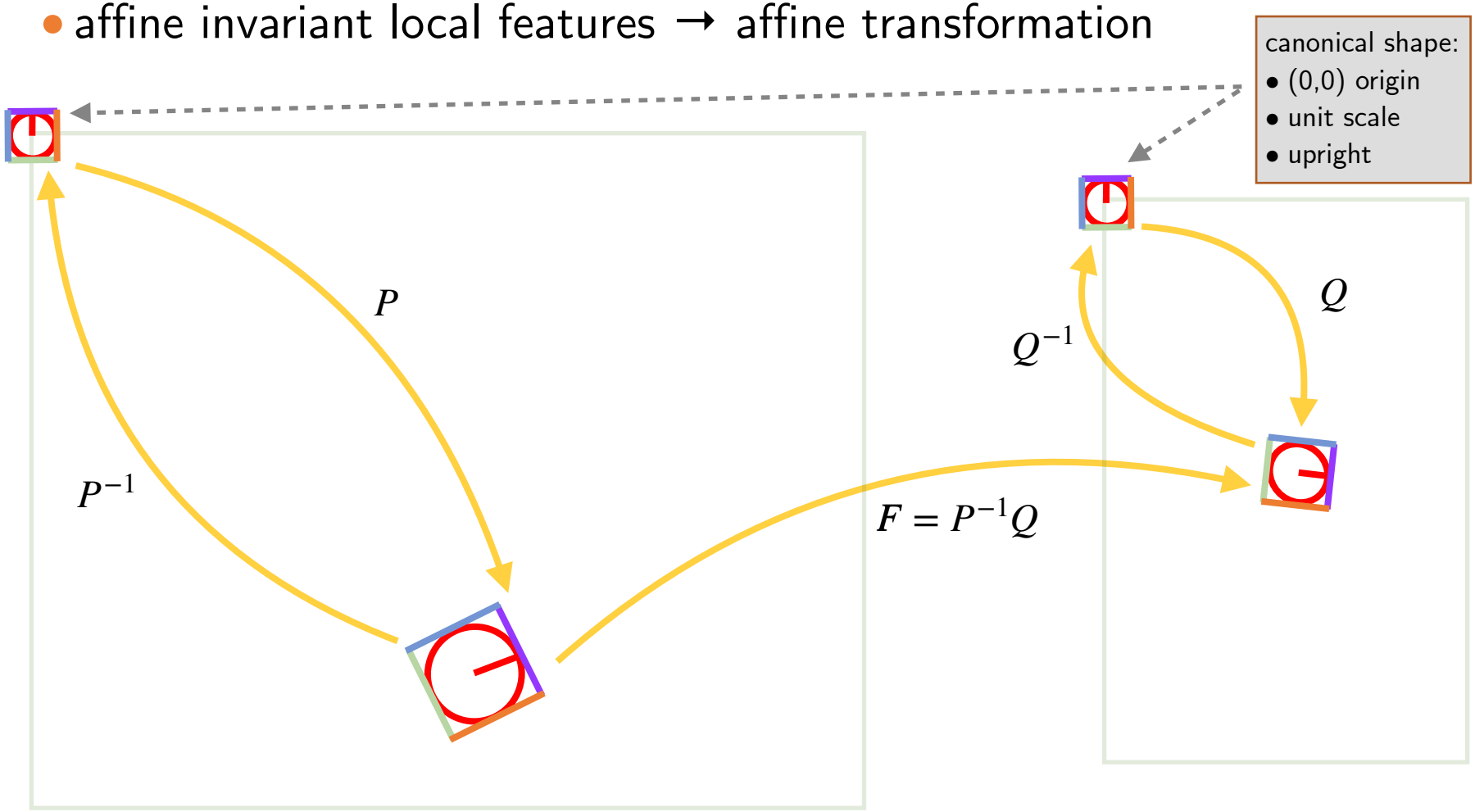
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



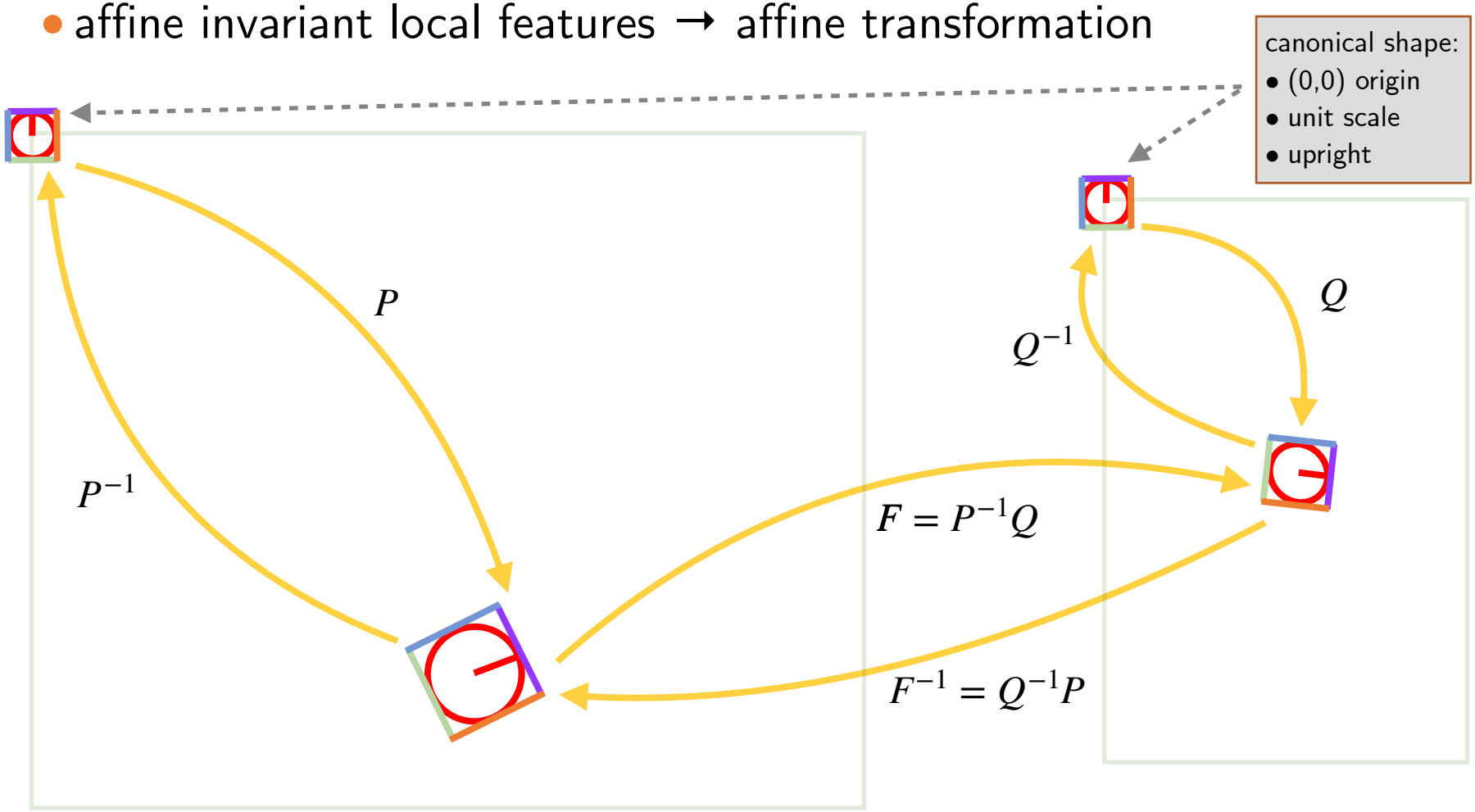
spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



spatial verification

- estimate the transformation hypothesis from a single correspondence
- scale and rotation invariant local features → similarity transformation
- affine invariant local features → affine transformation



conventional RANSAC

- correspondence $c_i = (x_i, y_i) \leftrightarrow (x_i', y_i')$
- correspondence set $C = \{c_1, c_2, \ldots c_n\}$

conventional RANSAC with point correspondences

- for j in 1…T # max number of hypotheses
  - sample a minimal set C_m from C # size depends on DoFs
  - estimate F from C_m # inlier free sample for good estimate
  - for i in 1…n # loop over all correspondences
    - $\begin{bmatrix} \hat{x}_i \\ \hat{y}_i \\ 1 \end{bmatrix} = F \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$ # transform points from an image to the other
    - is_inlier = error($[\hat{x}_i, \hat{y}_i], [x_i', y_i']$) $< \tau$ # error between projected and given point
    - …… # inlier count, track F with most inliers, early termination, …

fast spatial matching

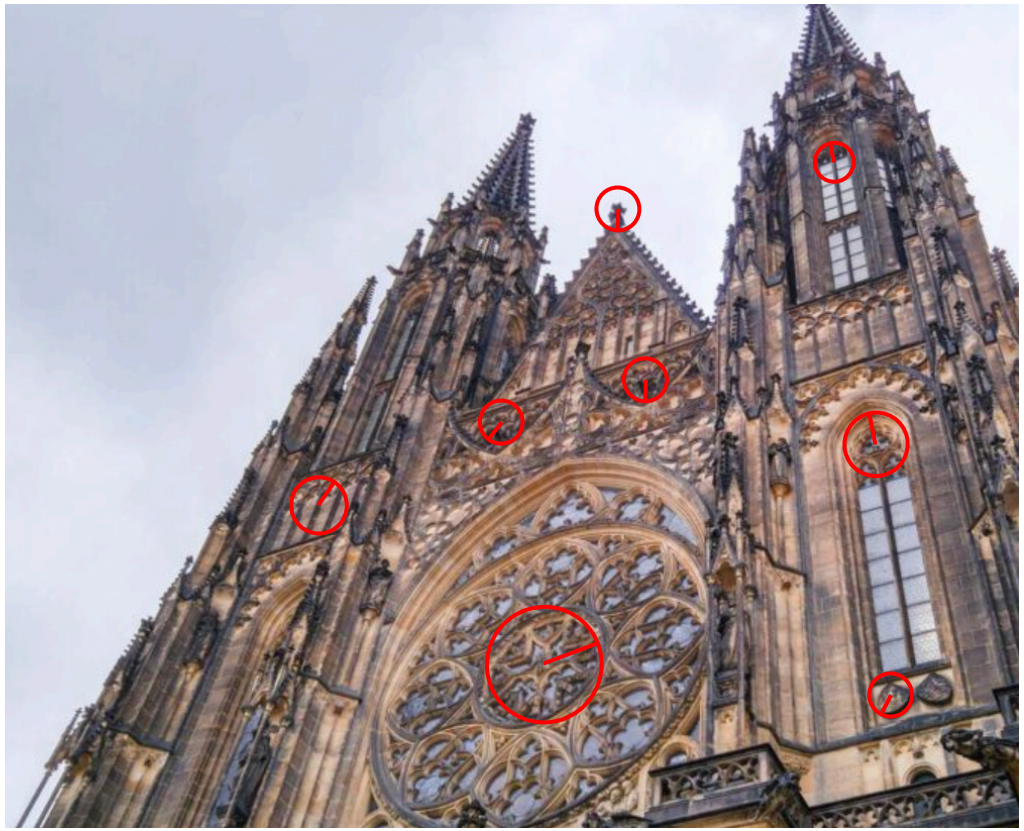
- correspondence $c_i = (A_i, x_i, y_i) \leftrightarrow (A'_i, x'_i, y'_i)$
- correspondence set $C = \{c_1, c_2, \ldots c_n\}$

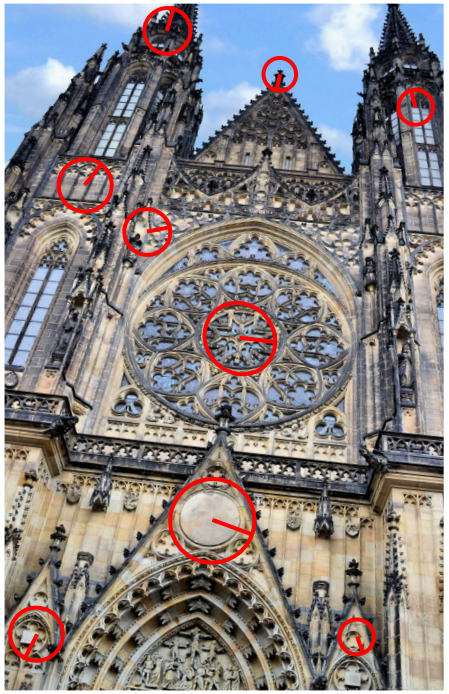
fast spatial matching with local feature correspondences

- for j in 1…n # loop over all correspondences
  - estimate F from c_j # single correspondence hypothesis
  - for i in 1…n # loop over all correspondences
    - $\begin{bmatrix} \hat{x}_i \\ \hat{y}_i \\ 1 \end{bmatrix} = F \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$ # transform points from an image to the other
    - is_inlier = error($[\hat{x}_i, \hat{y}_i], [x'_i, y'_i]$) $< \tau$ # error between projected and corresponding point
    - …… # inlier count, track F with most inliers, early termination, …

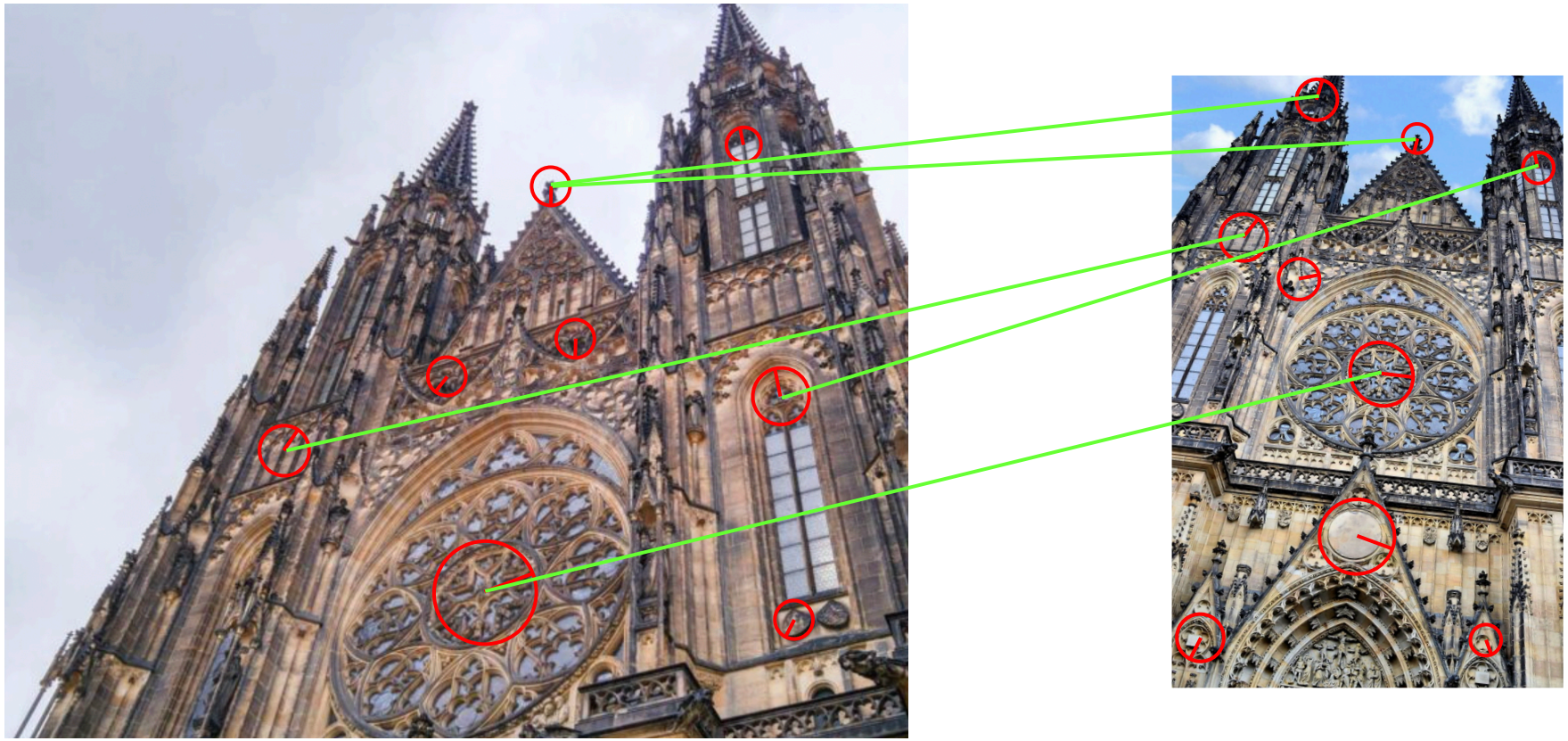
no randomness

transformation hypothesis verification - inlier counting

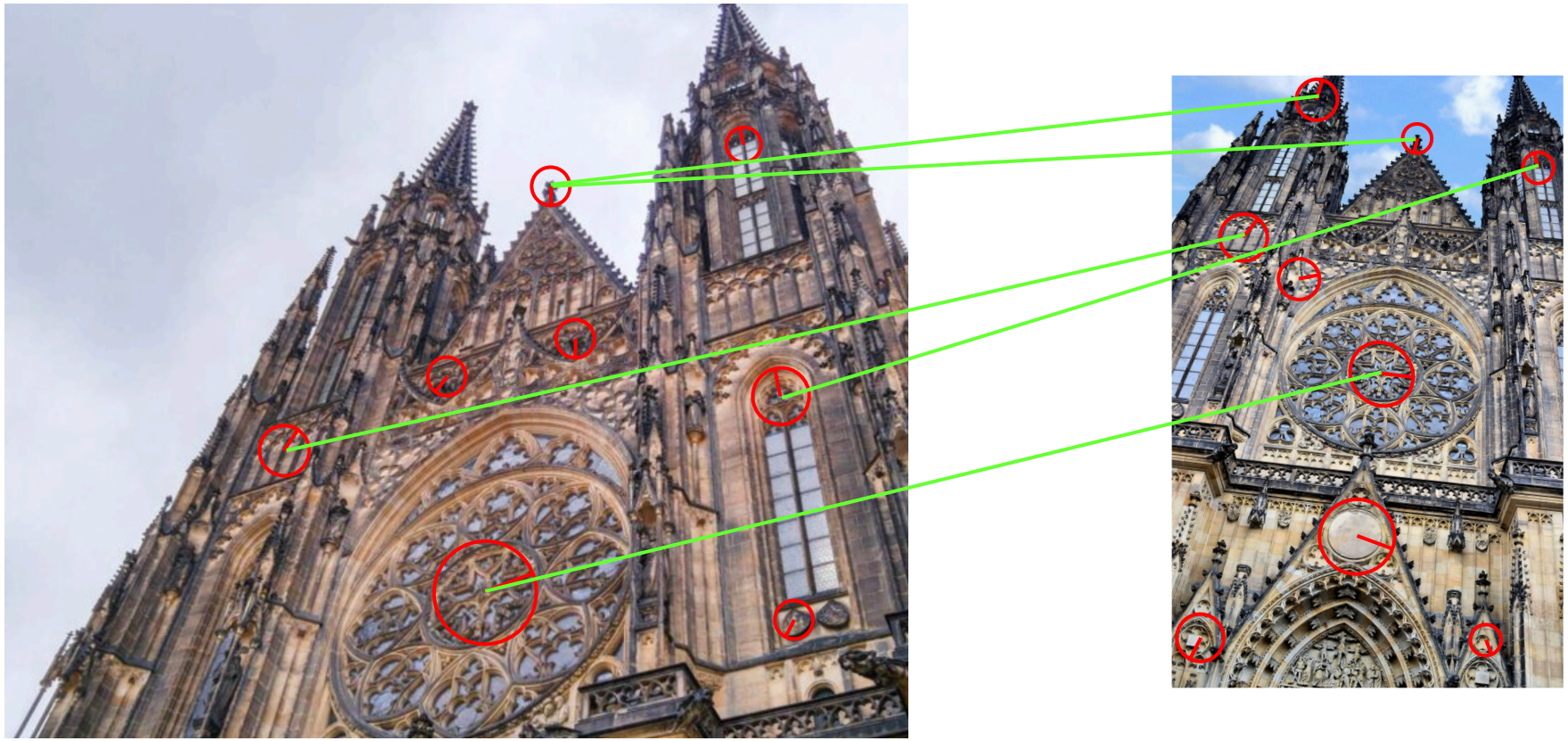




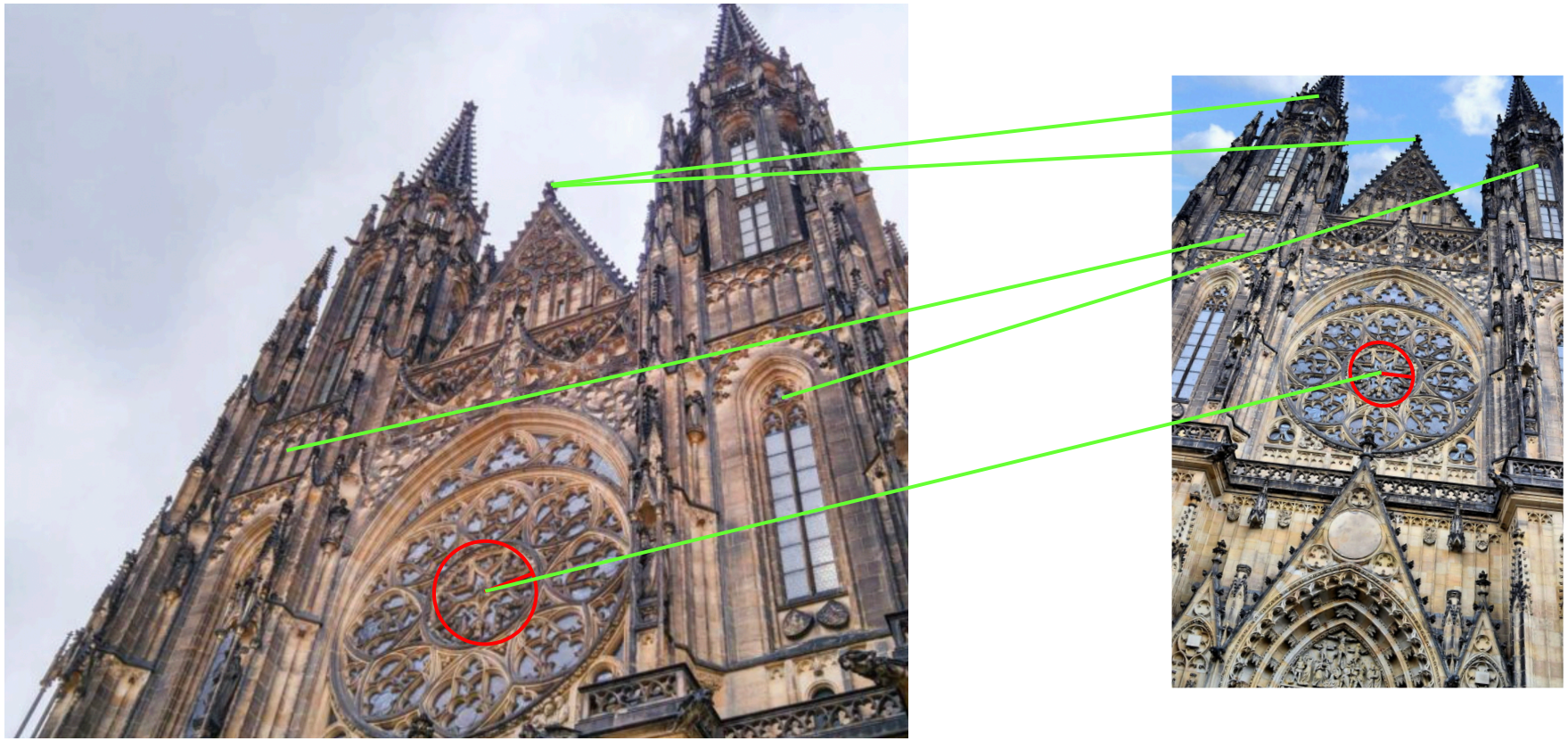
transformation hypothesis verification - inlier counting



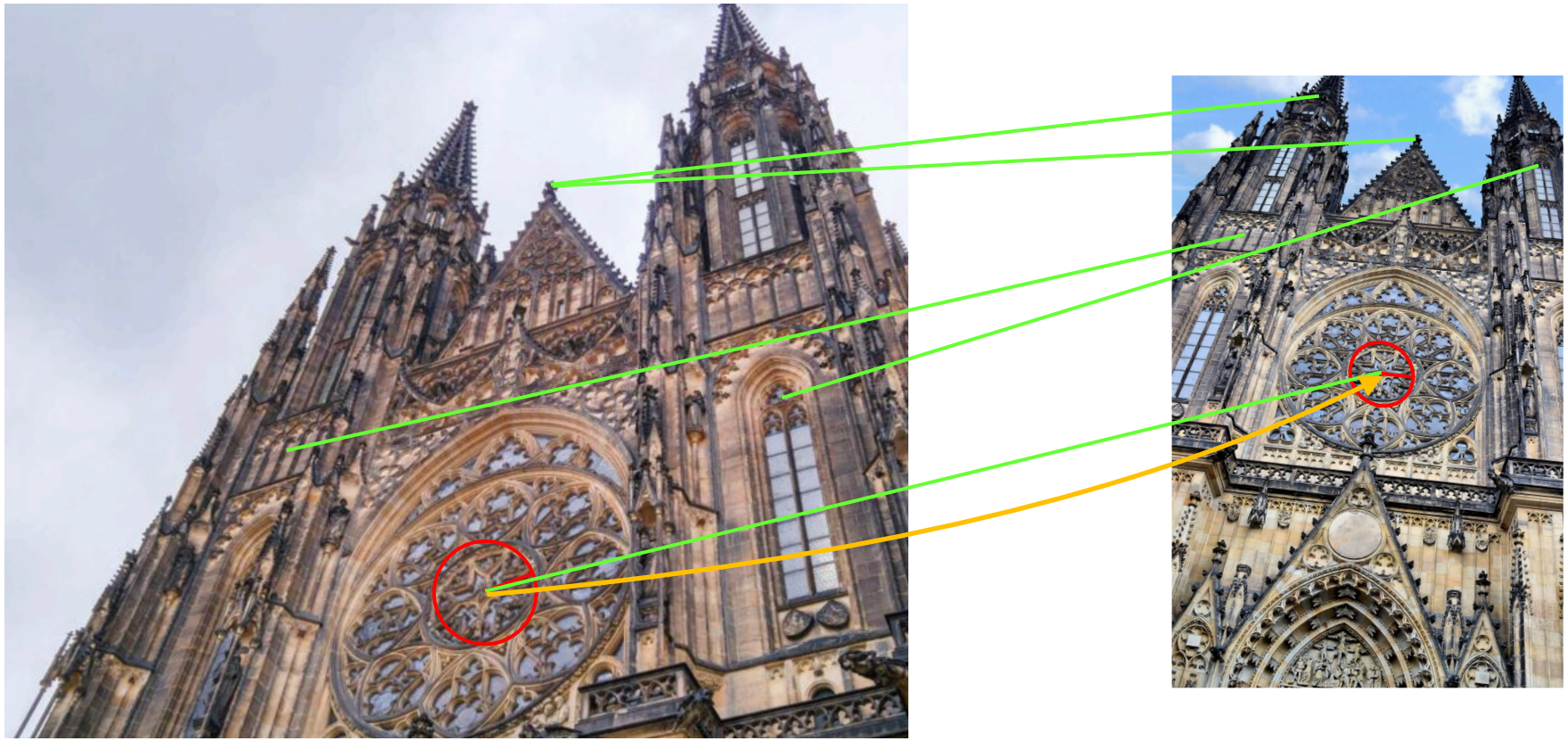
transformation hypothesis verification - inlier counting



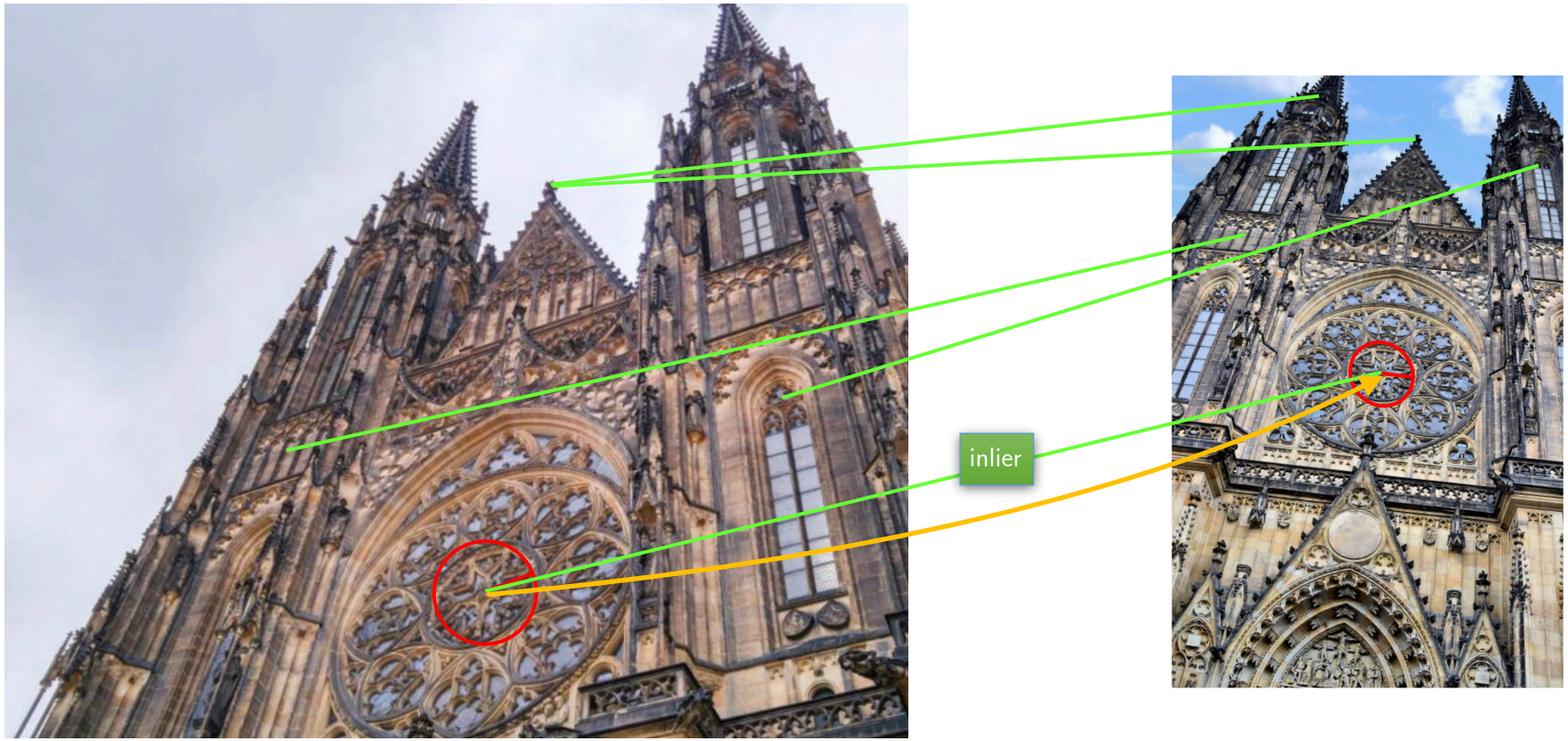
transformation hypothesis verification - inlier counting



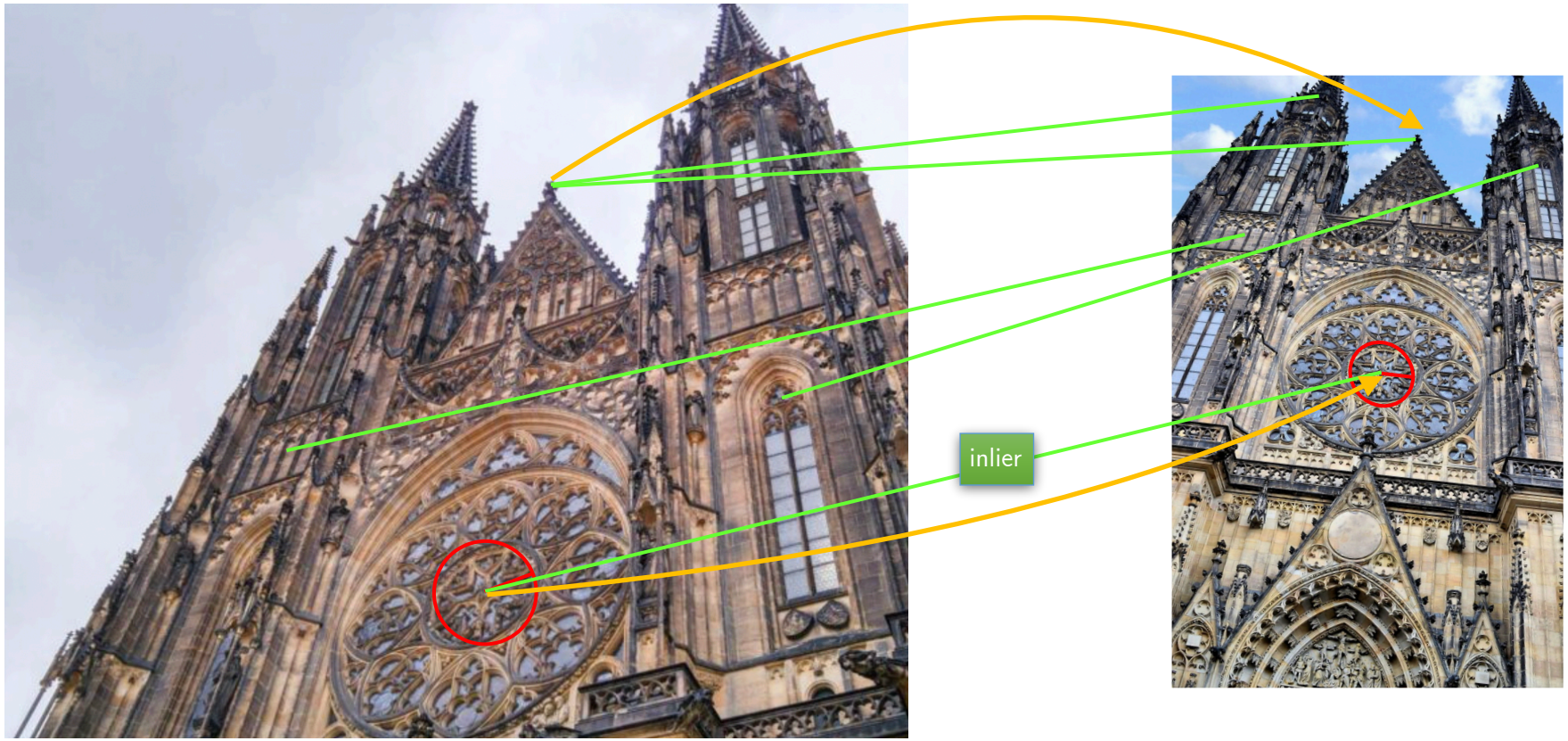
transformation hypothesis verification - inlier counting



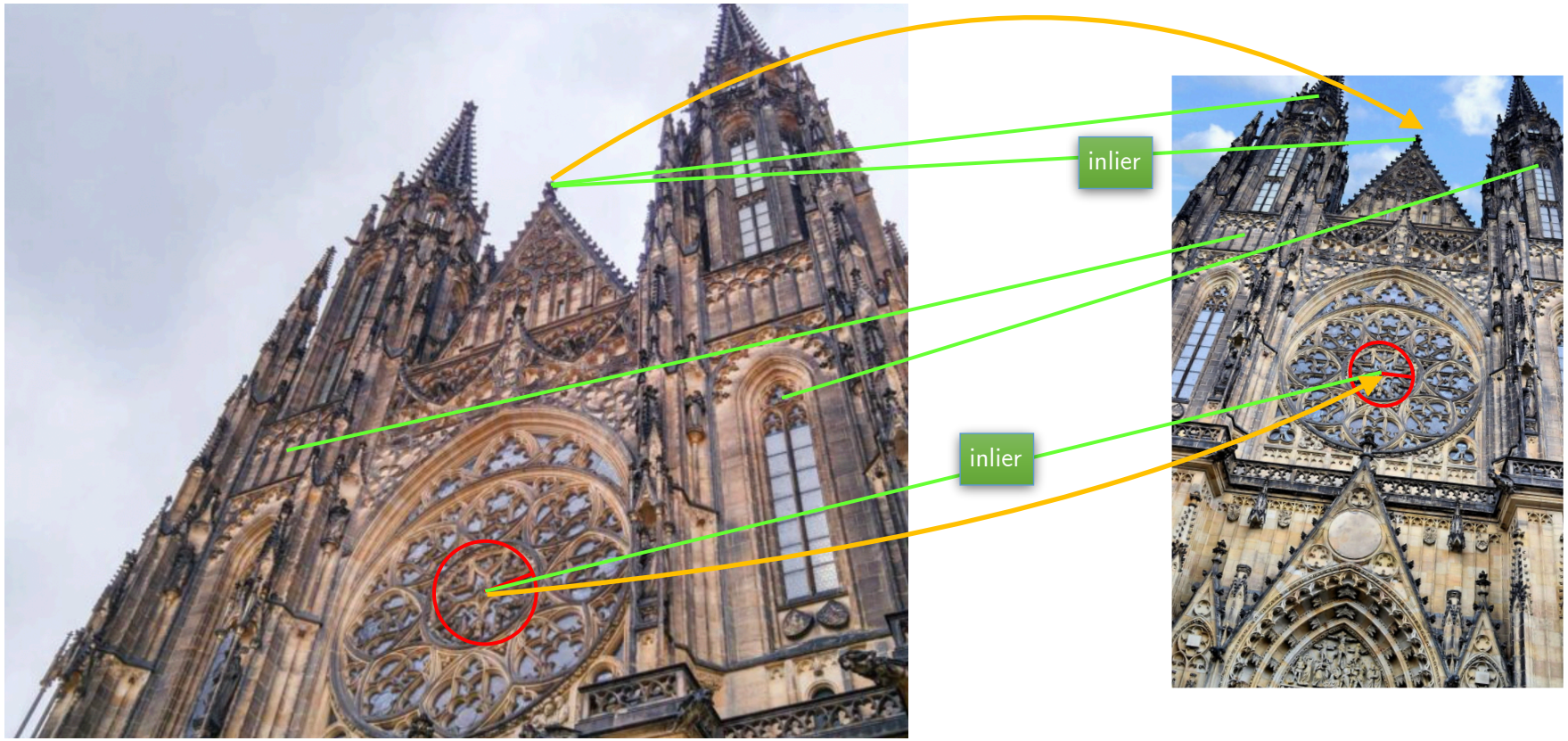
transformation hypothesis verification - inlier counting



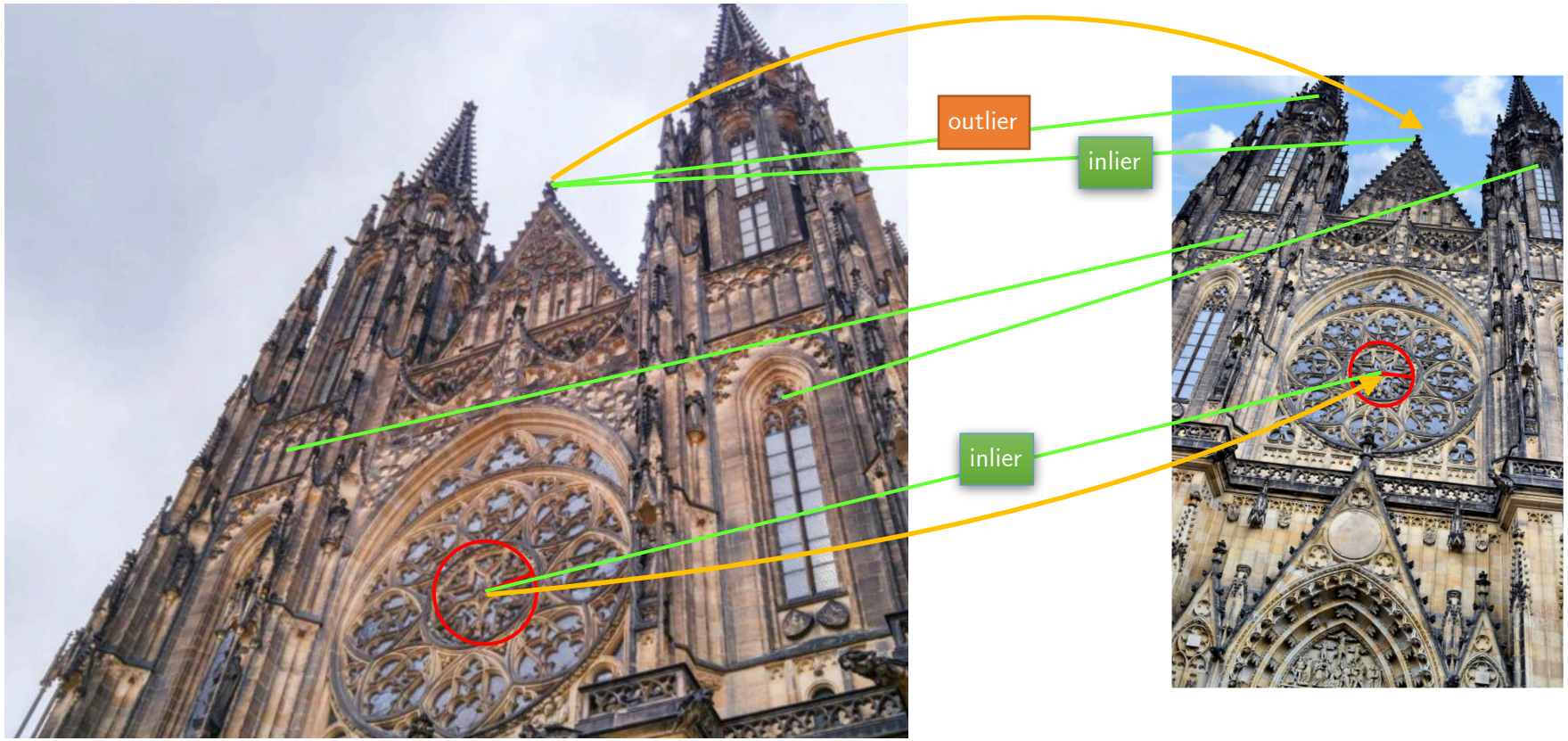
transformation hypothesis verification - inlier counting



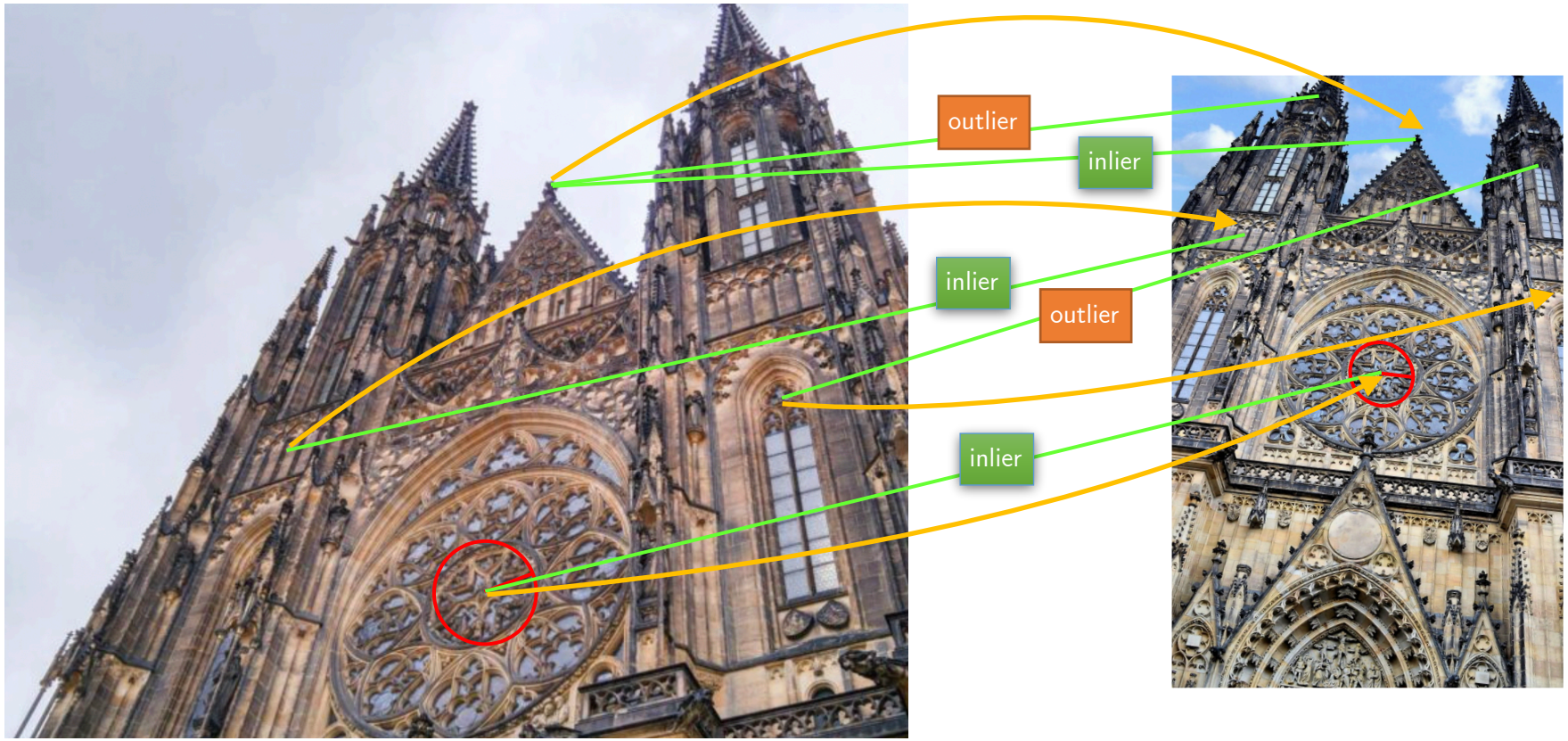
transformation hypothesis verification - inlier counting



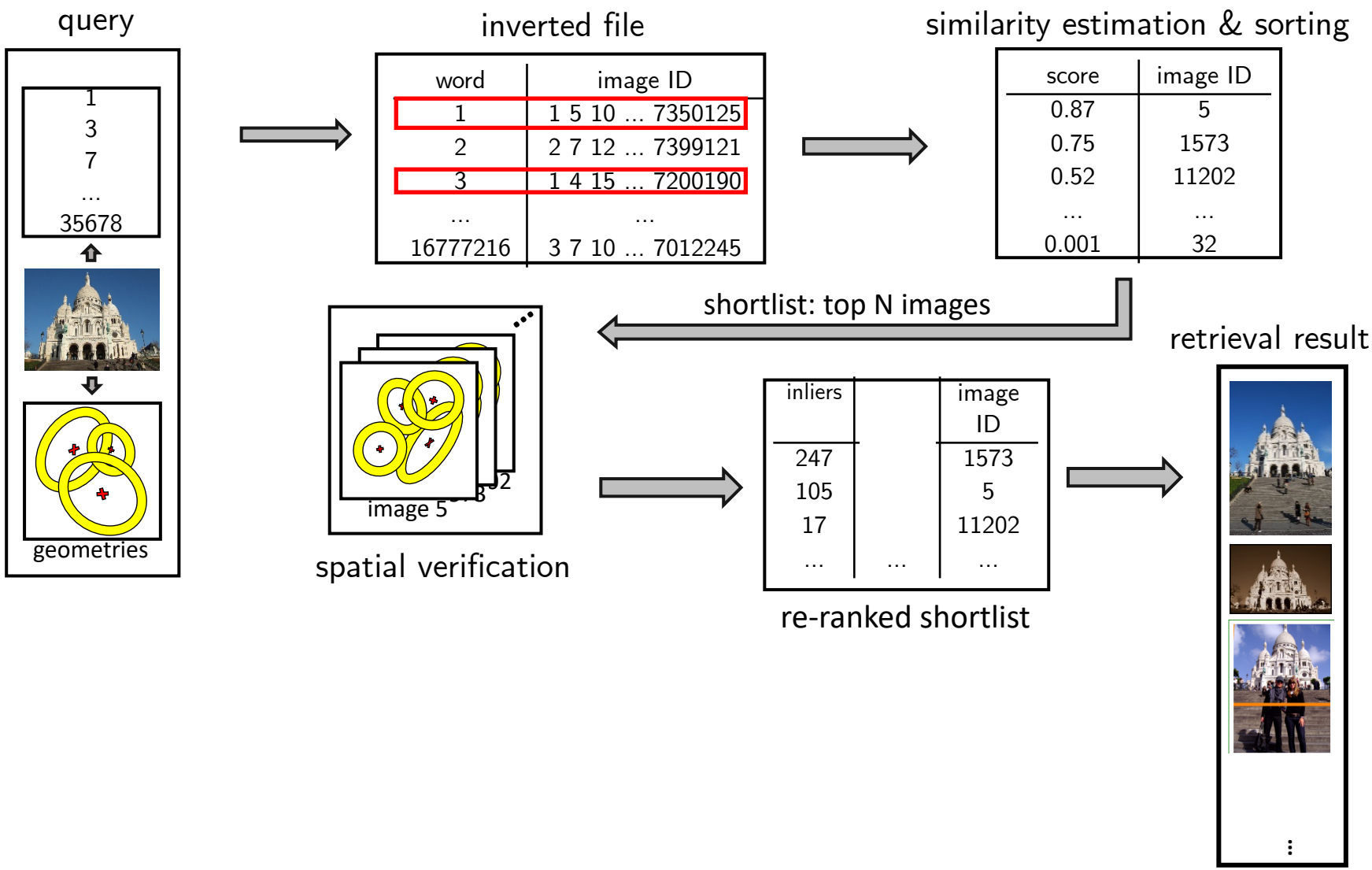
transformation hypothesis verification - inlier counting



transformation hypothesis verification - inlier counting



spatial verification - retrieval



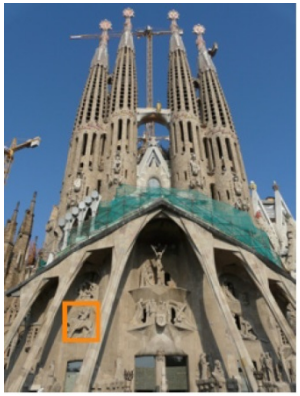
[Philbin et al. CVPR'07]

[Perdoch et al. CVPR'09]

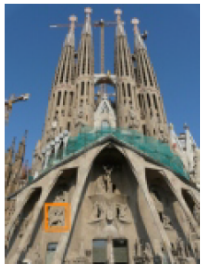
special retrieval tasks

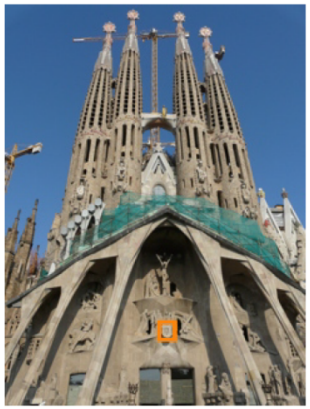
beyond visual similarity retrieval

- most similar images: near identical
  - is this what we want?

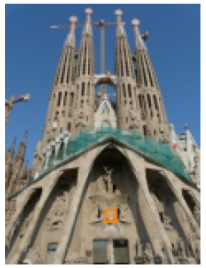


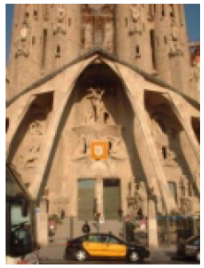
query 1

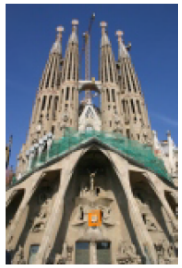




query 2

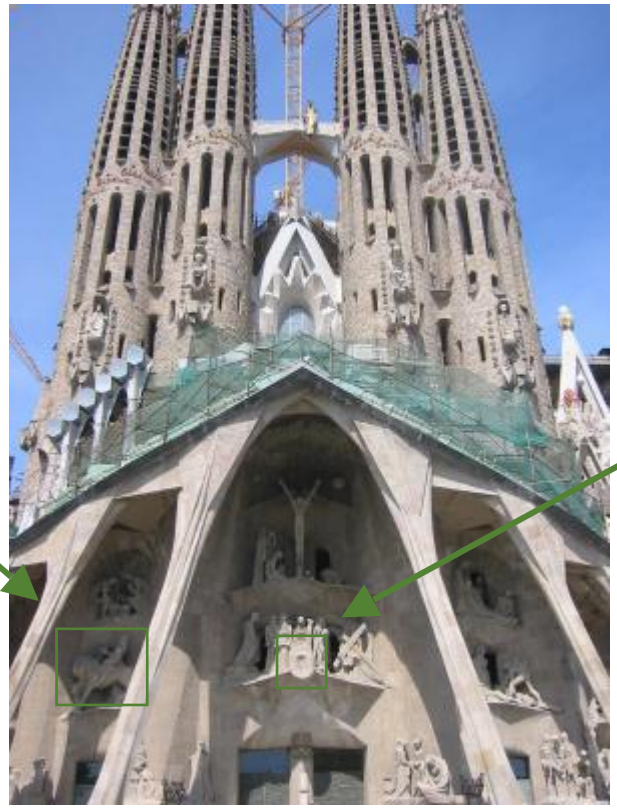






beyond visual similarity retrieval

what is this?



... and what is that?

let's zoom-in!

image retrieval for zoom-in

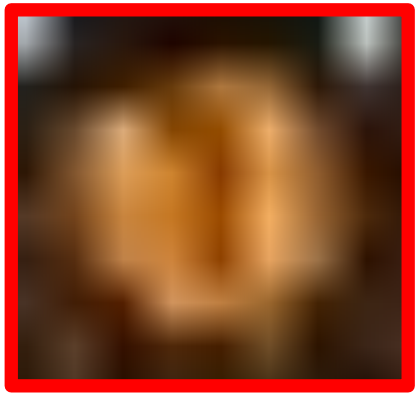


image retrieval for zoom-in

image retrieval for zoom-in

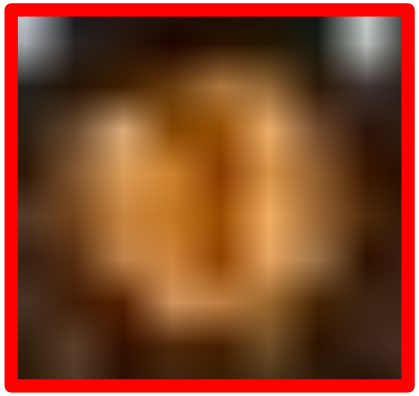


image retrieval for zoom-in

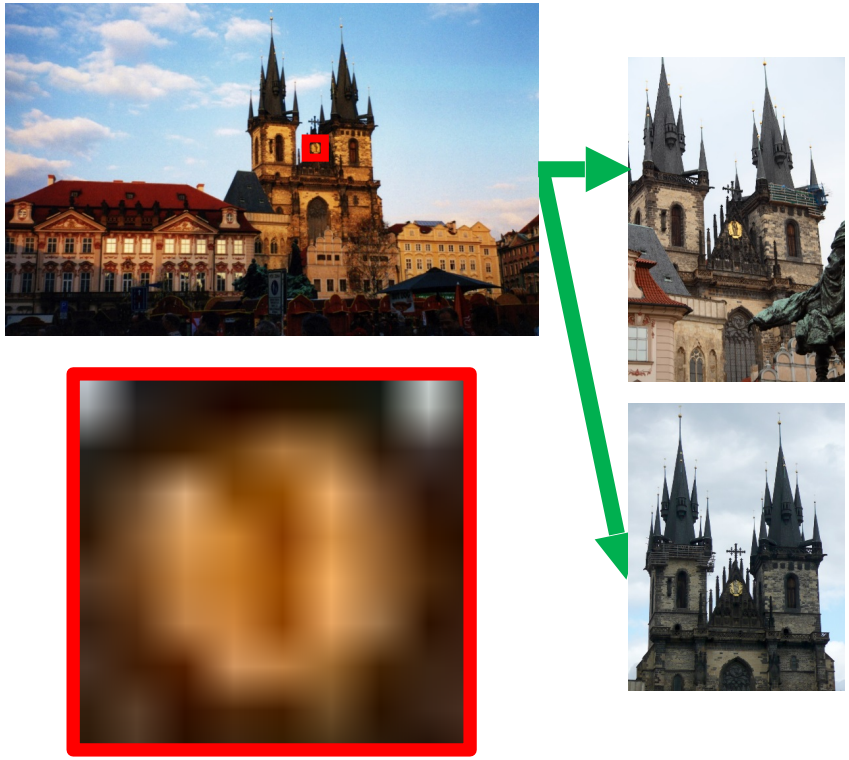


image retrieval for zoom-in

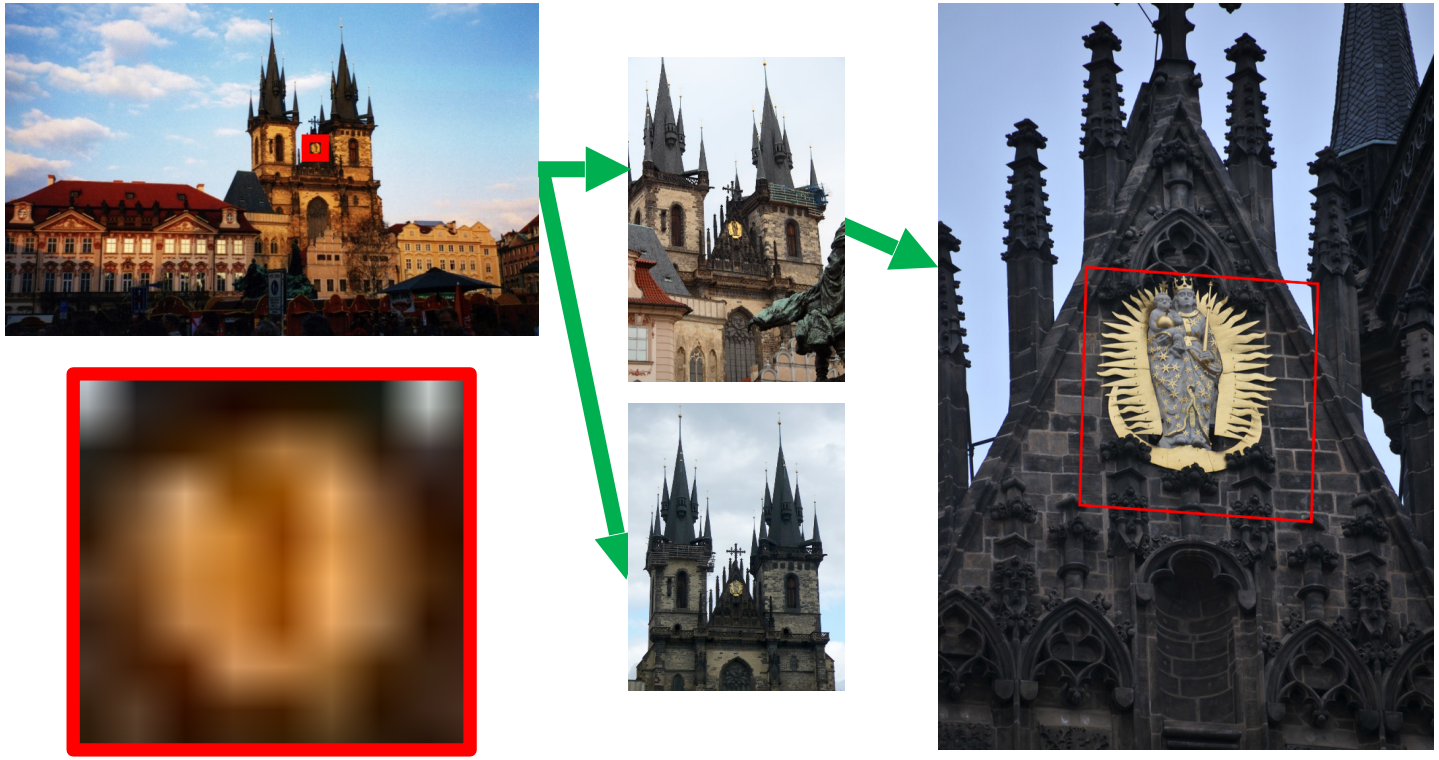


image retrieval for zoom-in

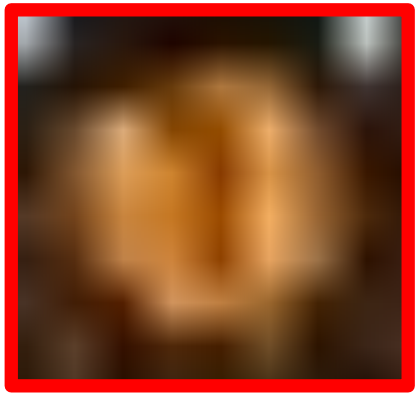


image retrieval for zoom-in

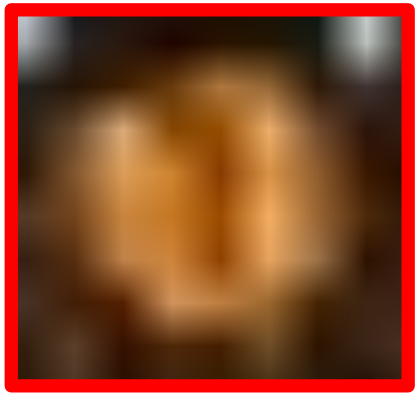
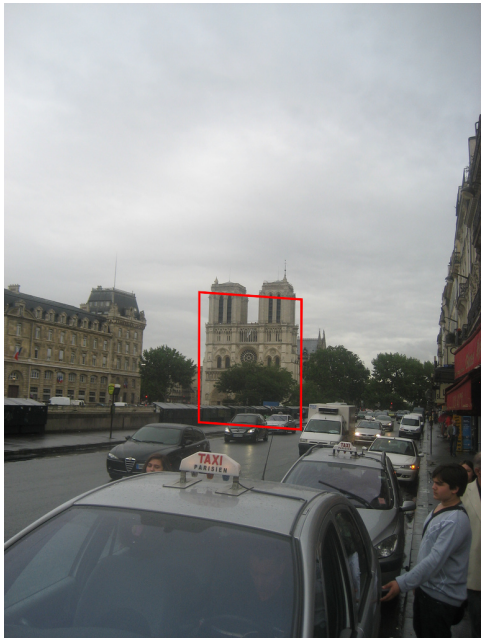
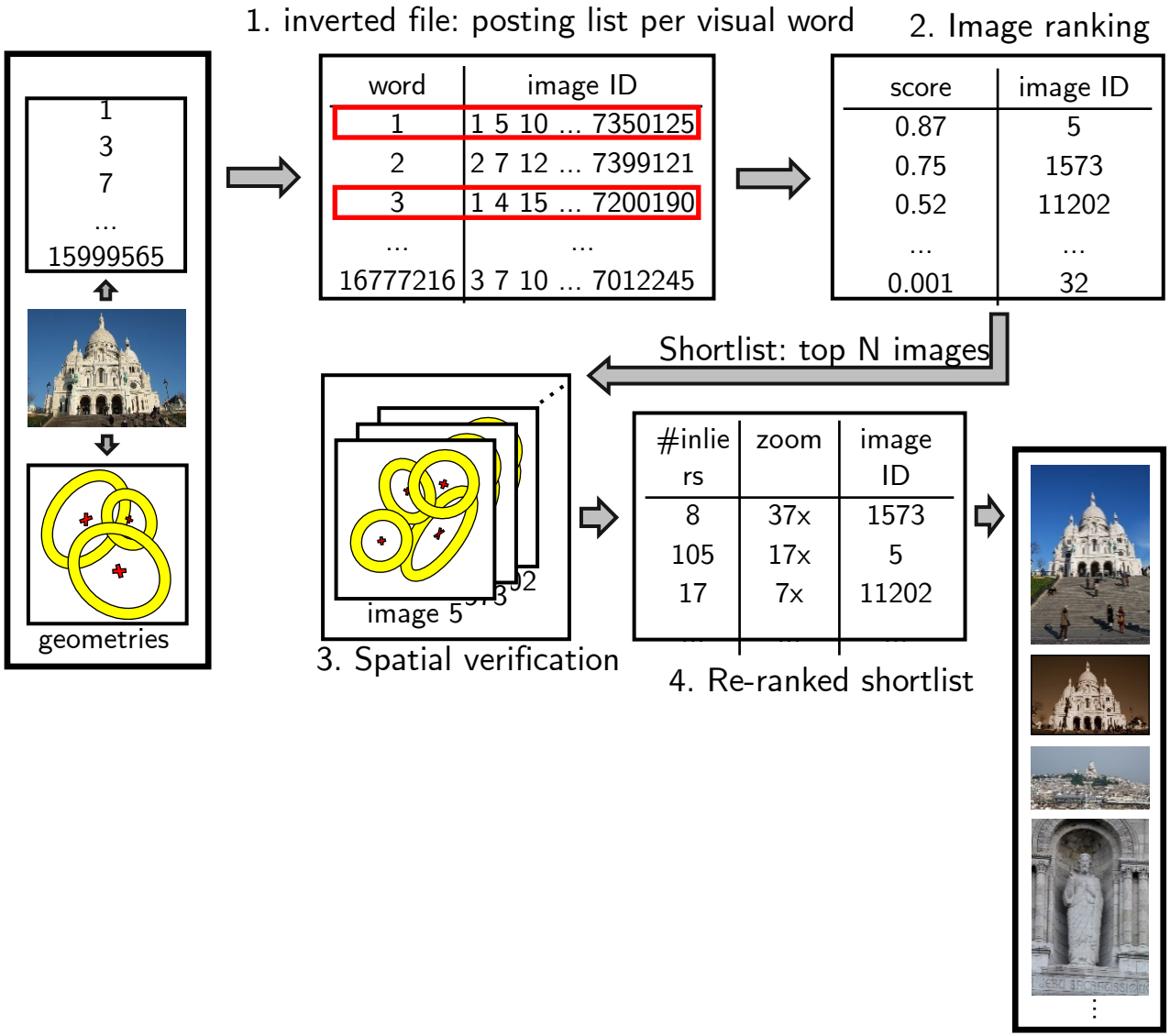


image retrieval for zoom-out

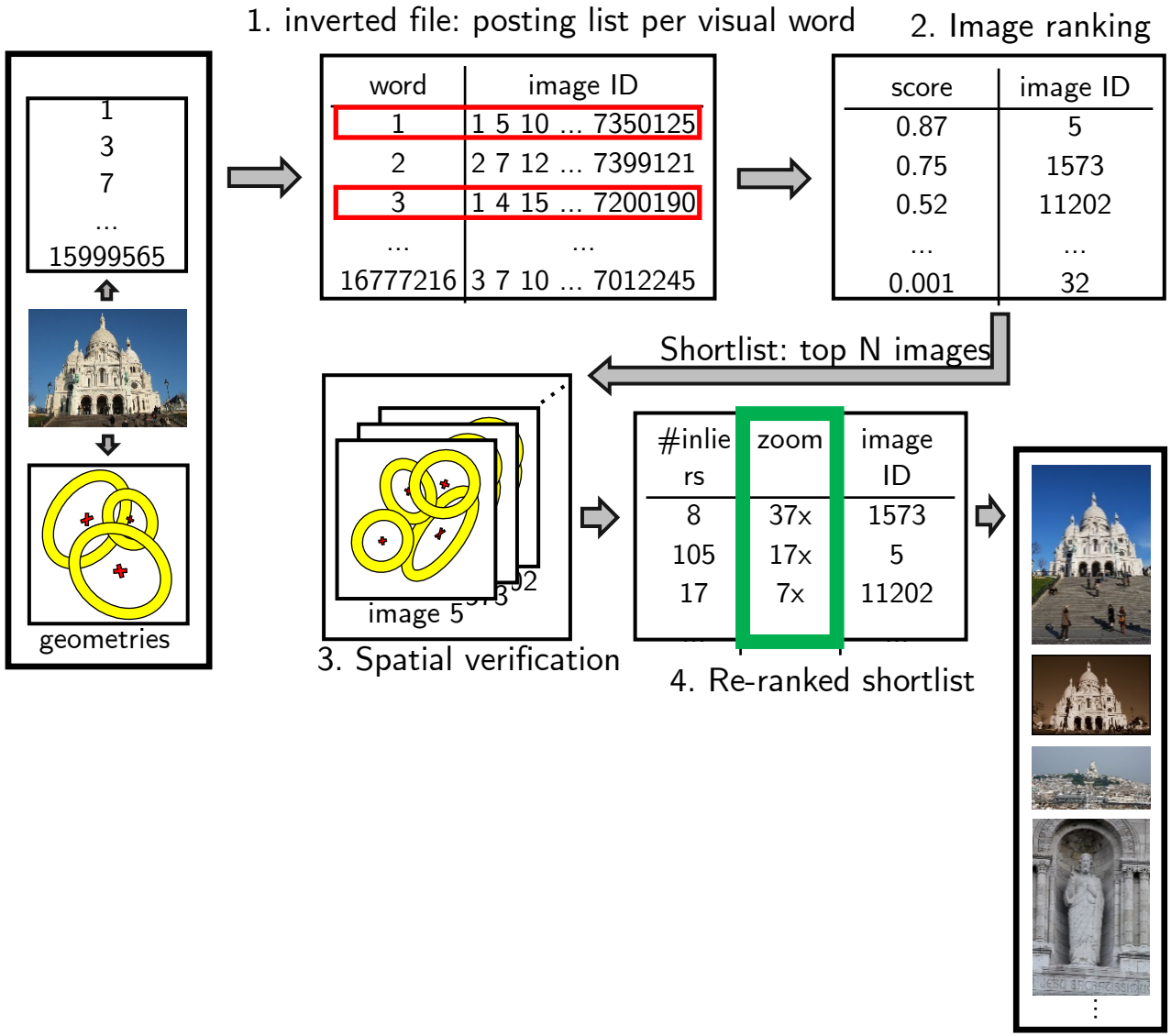


zoom in/out



[Mikulik et al. ACCV'14]

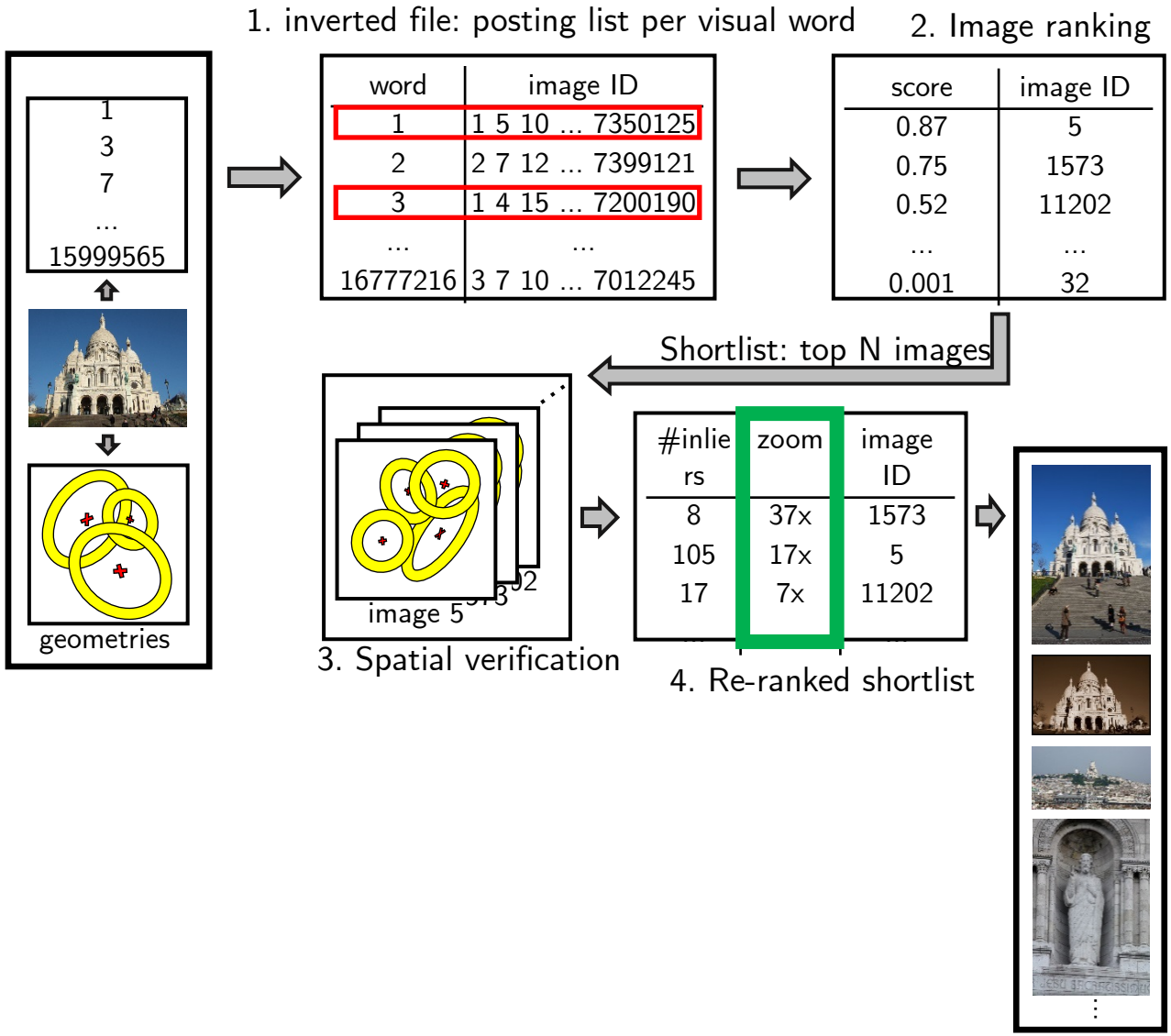
zoom in/out



problem specific ranking function, e.g. maximize scale change

[Mikulik et al. ACCV'14]

zoom in/out



geometry compressed in inverted file taken into account during scoring

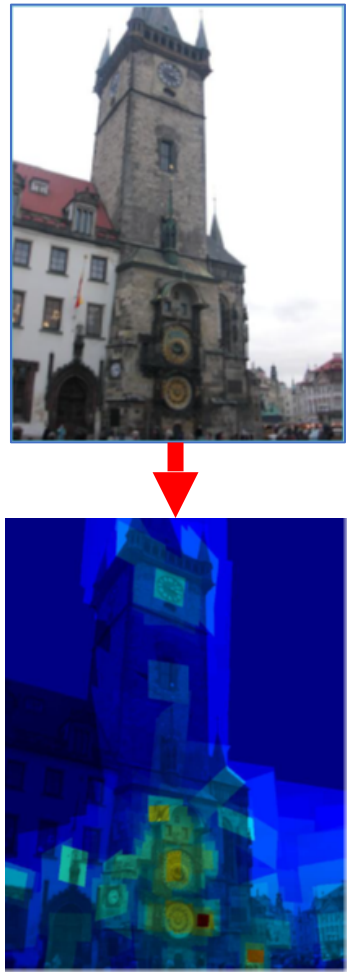
problem specific ranking function, e.g. maximize scale change

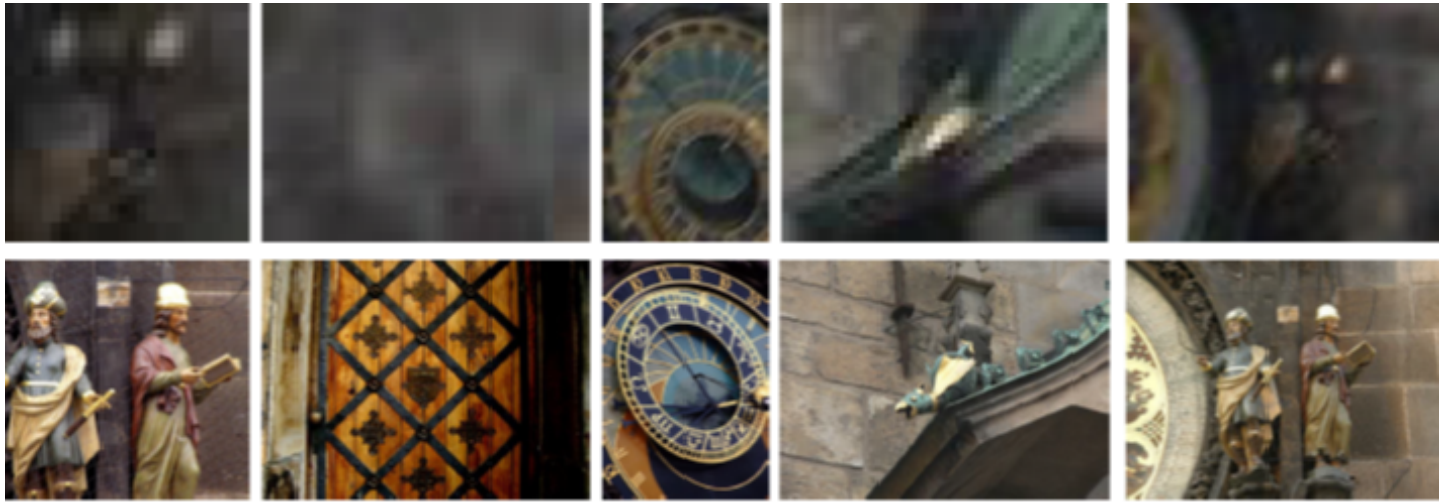
[Mikulik et al. ACCV'14]

what should you not miss?

highest resolution transform

- given a query, 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

[Mikulik et al. ACCV'14]

retrieval for 3D reconstruction

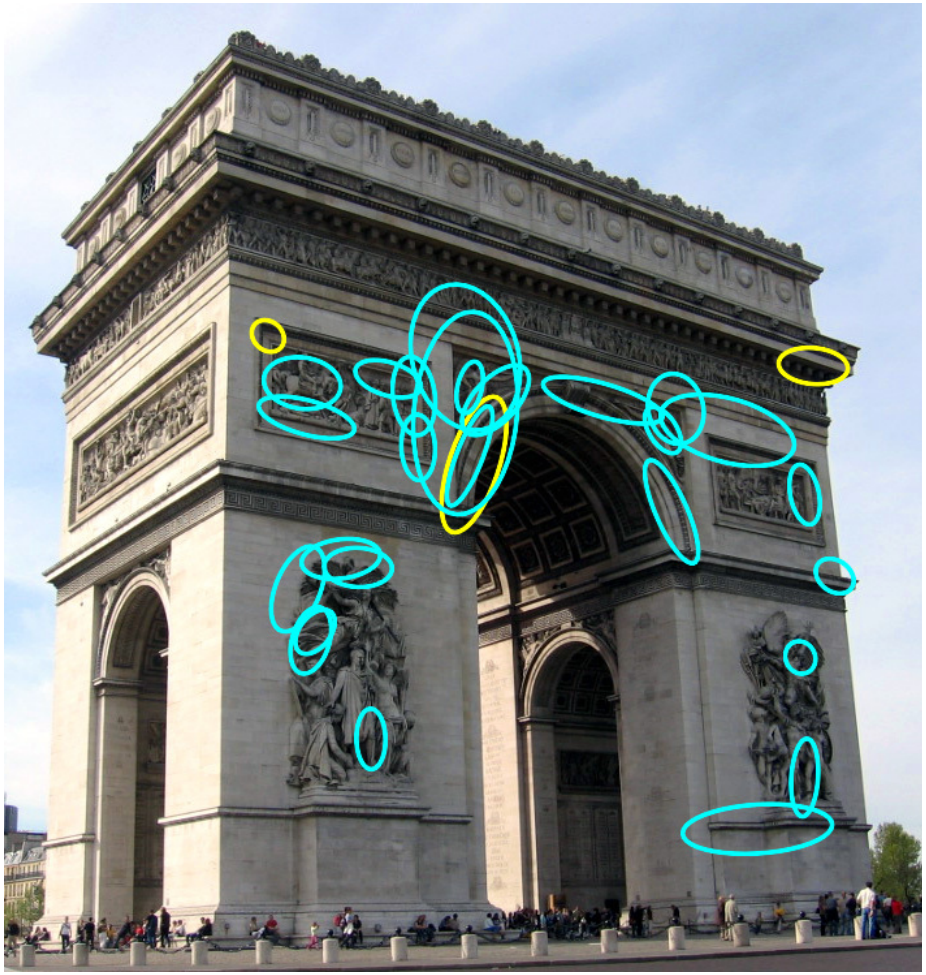
- visually most similar search
  - many near duplicates
  - details lost
- zoom-in and details search
  - details retrieved
  - transition images to match the details
- zoom-out search
  - viewpoint change
  - more context
- sideways crawl
  - significant viewpoint change
  - more context

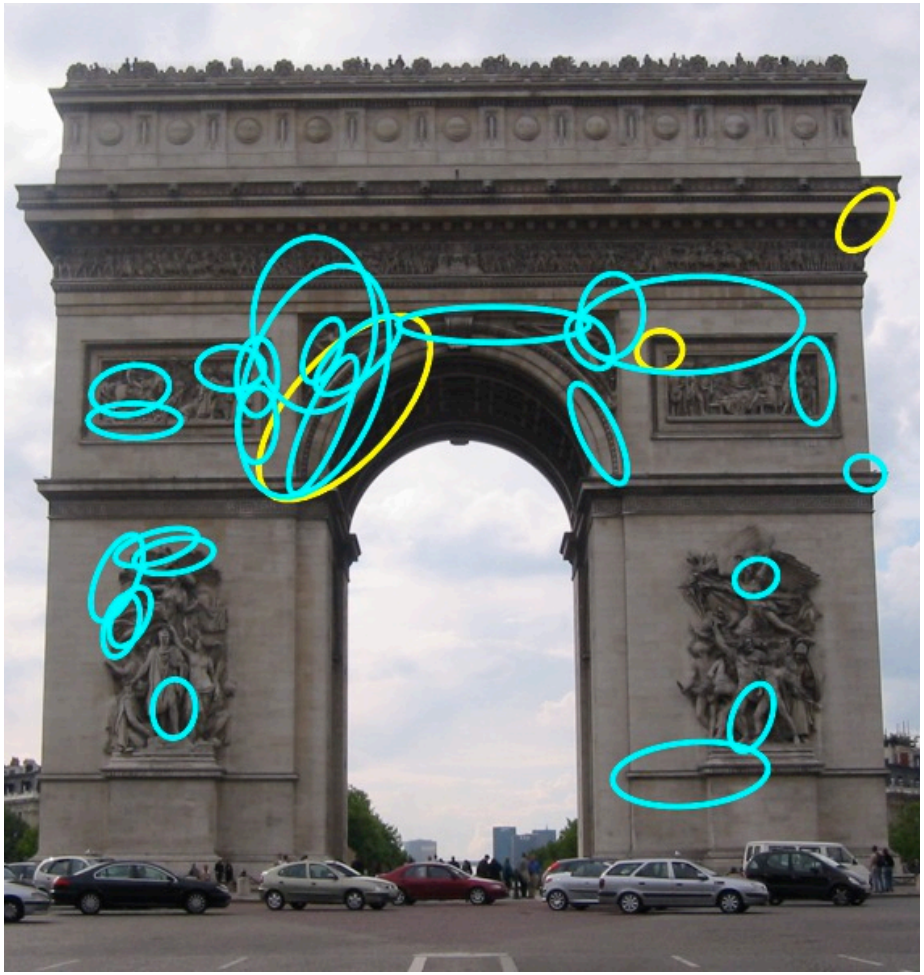
[Schonberger et al. CVPR'15]

sideways image crawl

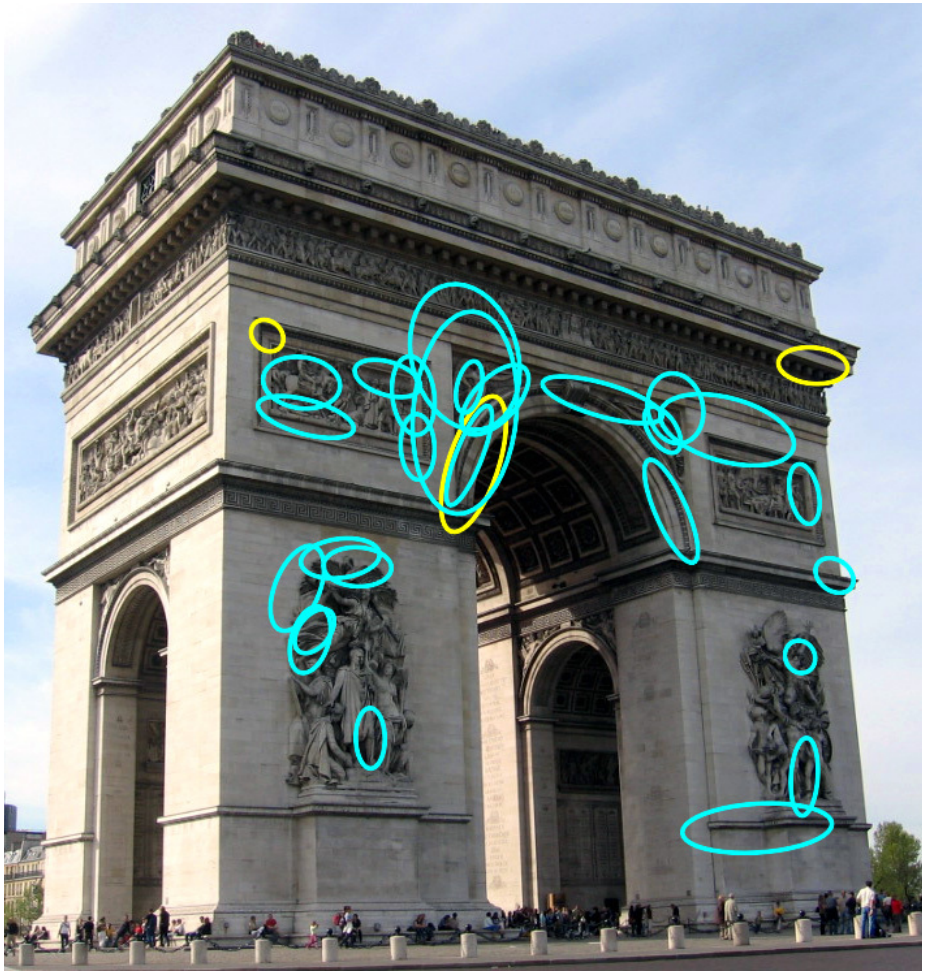
sideways left: step by step

sideways left: step by step

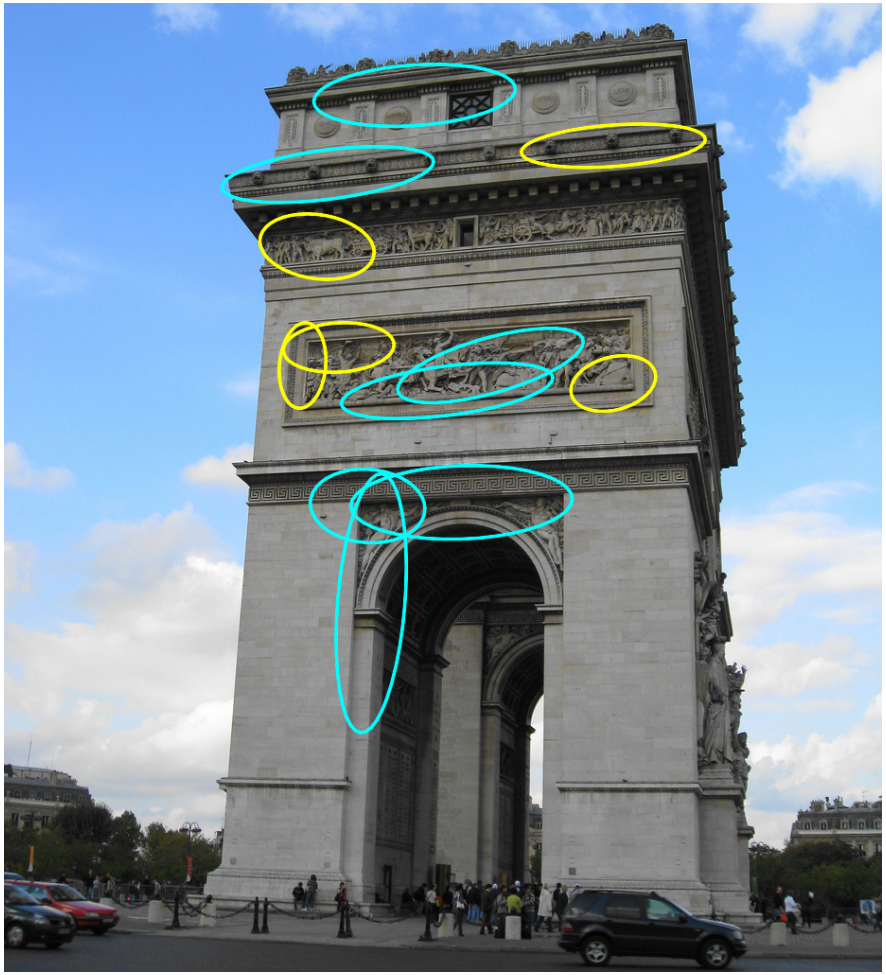




sideways left: step by step



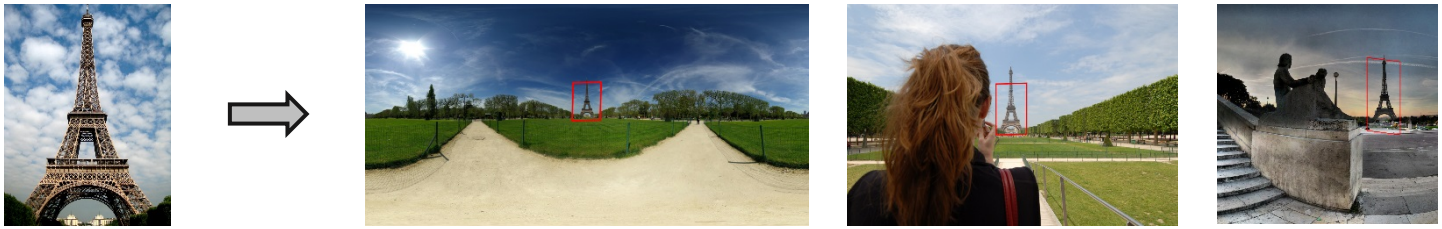
sideways left: step by step



summary

visually most similar

zoom-in / details



zoom-out

sideways right

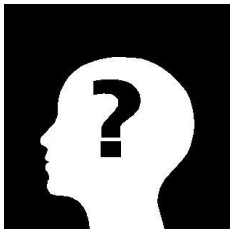
3D reconstruction – from a single image

[Schonberger et al. CVPR'15]

3D reconstruction – from a single image

[Schonberger et al. CVPR'15]

questions




640 x 331 · jpeg
小料：关晓彤吴磊、陈立农、秦岚...
lieqinews.com
Comunista
mural.uv.es
MAKI Sushi Roll & Wings La Reform...
www.facebook.com
シロフクコーヒー - Home | Facebook
Facebook
blogspot.com
Change Password Svg Png Icon Fre...
www.onlinewebfonts.com
Free other icon File Page 184 - New...
www.newdesignfile.com
point d interrogation photo cerveau ...
mytourdumonde.net
YellTale*A place for fun & fine articl...
blogspot.com
来年はApple SiliconなMacBook Pro...
www.gizmodo.jp
Punto Interrogativo In Testa Illustra...
Dreamstime
El origen de los nombres de las em...
zonaj.net
Lockee, le générateur de cadenas n...
portaileduc.net
0:15
$6
Burger King 2 for $6 Whopper Deal ...
iSpot.tv

bibliography

- BoW
  - J Sivic, A Zisserman, Video Google: A text retrieval approach to object matching in videos, ICCV 2003
  - G Csurka, C Dance, L Fan, J Willamowski, C Bray, Visual categorization with bags of keypoints, ECCVW 2004
- VLAD
  - H Jegou, M Douze, C Schmid, P Pérez, Aggregating local descriptors into a compact image representation, CVPR 2010
  - J Delhumeau, PH Gosselin, H Jégou, P Pérez, Revisiting the VLAD image representation, ACMMM 2013
- Fisher vectors
  - F Perronnin, Y Liu, J Sánchez, H Poirier, Large-scale image retrieval with compressed fisher vectors CVPR 2010
- SMK / Hamming Embedding
  - G Tolias, Y Avrithis, H Jégou, To aggregate or not to aggregate: Selective match kernels for image search ICCV 2013
  - H Jegou, M Douze, C Schmid, Hamming embedding and weak geometric consistency for large scale image search, CVPR 2010
- spatial verification
  - J Philbin, O Chum, M Isard, J Sivic, A Zisserman, Object retrieval with large vocabularies and fast spatial matching, CVPR 2007
  - M Perdoch, O Chum, J Matas, Efficient representation of local geometry for large scale object retrieval, CVPR 2009
- other retrieval tasks
  - A Mikulık, F Radenovic, O Chum, J Matas, Efficient Image Detail Mining, ACCV 2014
  - JL Schonberger, F Radenovic, O Chum, JM Frahm, From single image query to detailed 3D reconstruction, CVPR 2015