Large Scale Image Retrieval and Mining

Ondra Chum

Introduction

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Visual Recognition Group

Computer Vision, Machine Learning, Recognition, Robotics, Medical

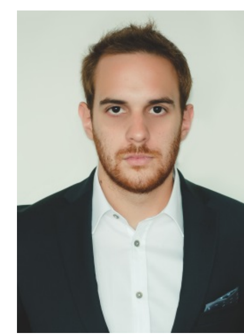
Image retrieval, Classification, Geometry, Robust model fitting

Giorgos Tolias
(Grece)

Javier Aldana
(Mexico)

Arun Mukundan
(India)

James Pritts
(USA)



Filip Radenović
(Montenegro)

Outline

- Image and specific object retrieval
- Clustering, min-Hash
- Geometry in image retrieval
- Beyond visual nearest neighbour search
- Retrieval for 3D
- Retrieval with CNN
- Advertisement

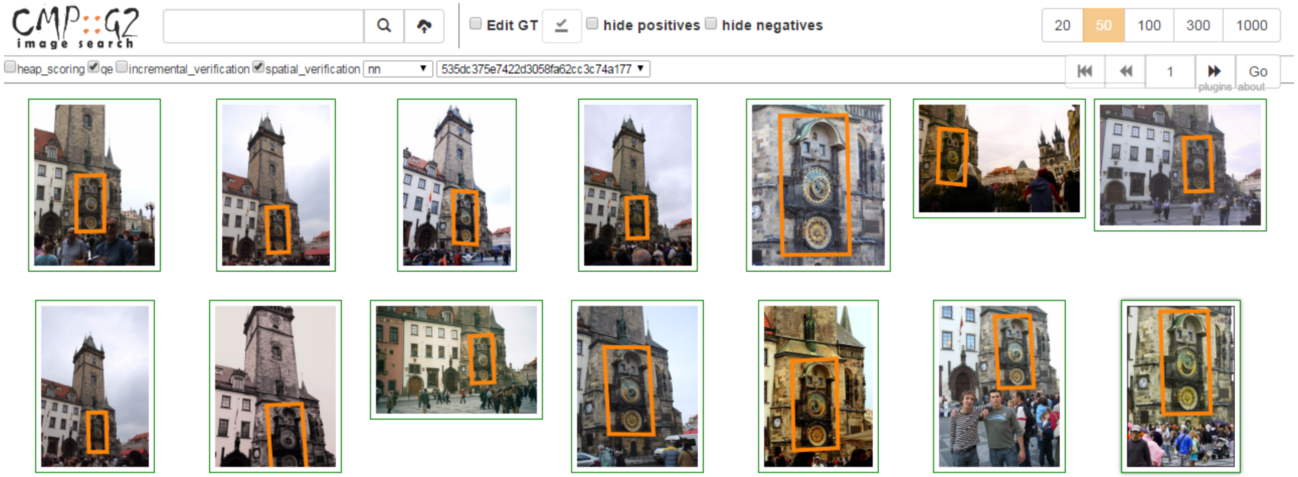


IMAGE RETRIEVAL

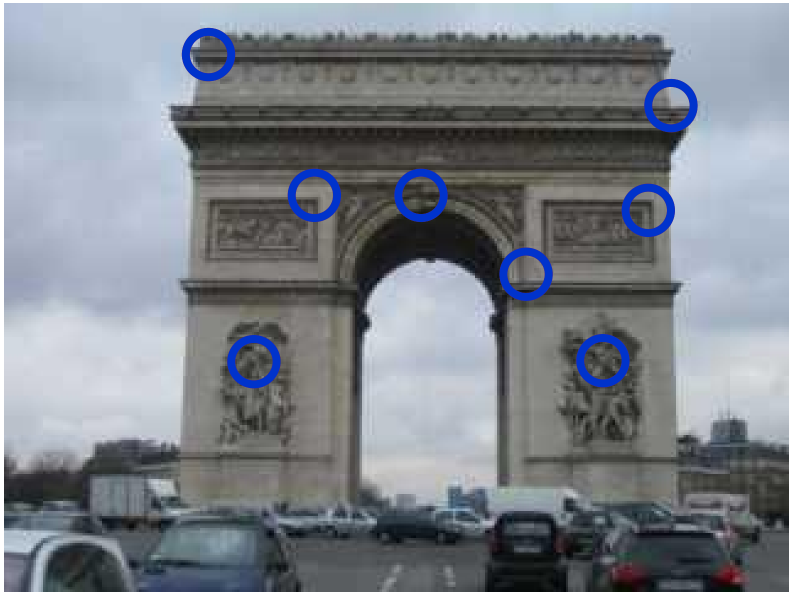
Video Google

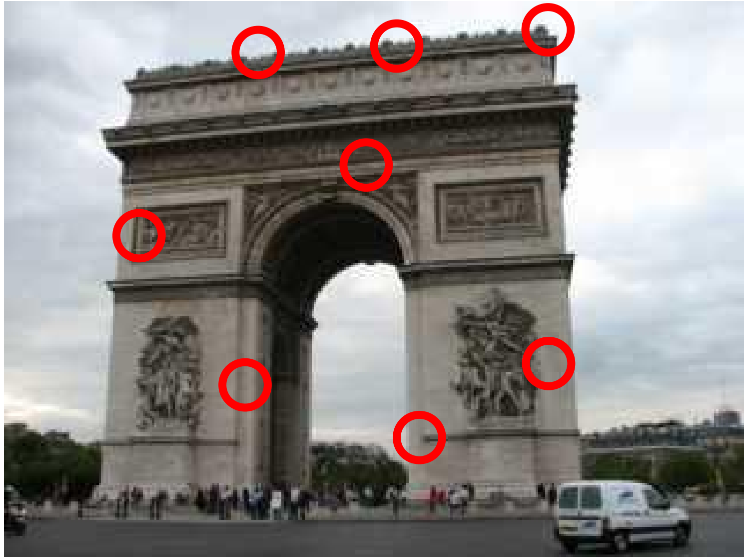
- Feature detection and description
- Vector quantization
- Bag of Words representation
- Scoring
- Verification

Sivic & Zisserman – ICCV 2003
Video Google: A Text Retrieval Approach to Object Matching in Videos

Local Features

aka feature points, key points, anchor points, distinguished regions, …

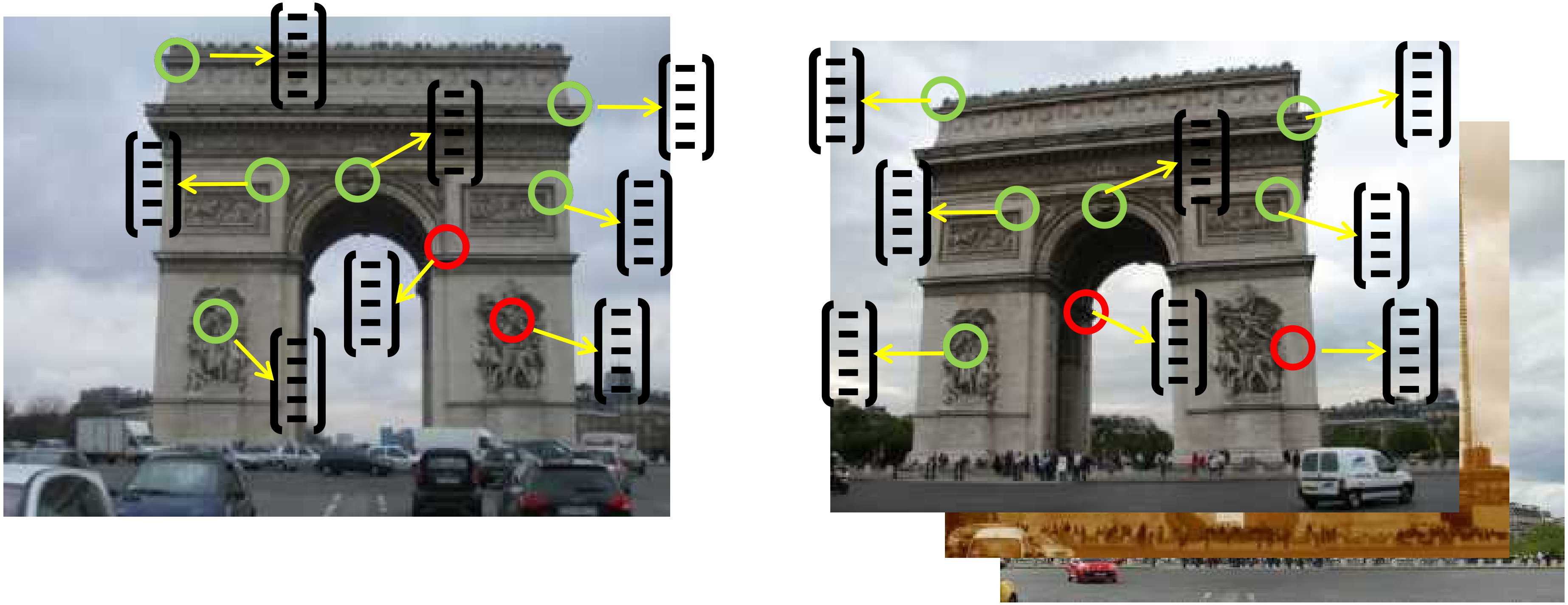




- Detect features in images independently, local = robust to occlusions
- Repeatable features

Local Features

aka feature points, key points, anchor points, distinguished regions, …



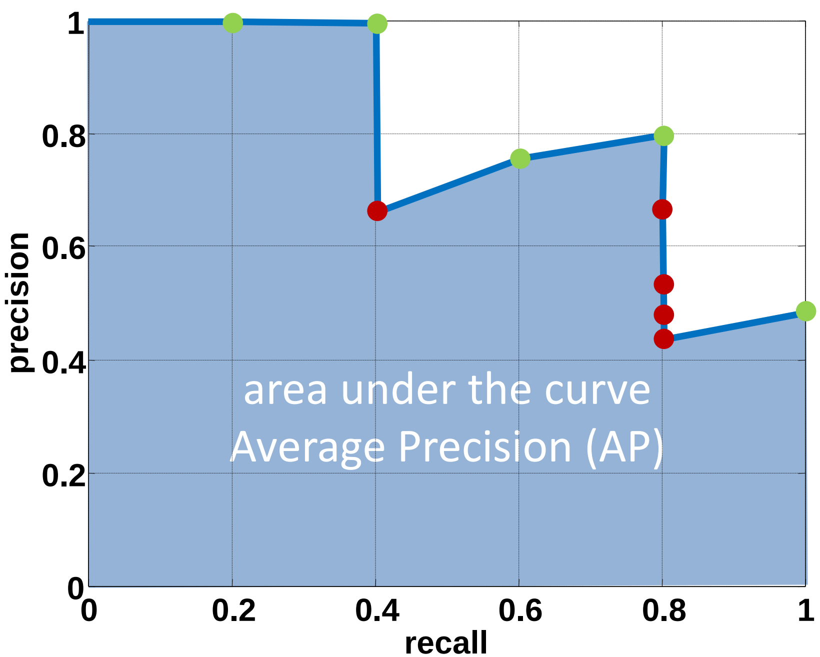
- Detect features in images independently, local = robust to occlusions
- Repeatable features
- Feature descriptor: patch to a vector
- Similar features have similar descriptors – nearest neighbour search
- Retrieval – matching millions of images at the same time

Retrieval Quality

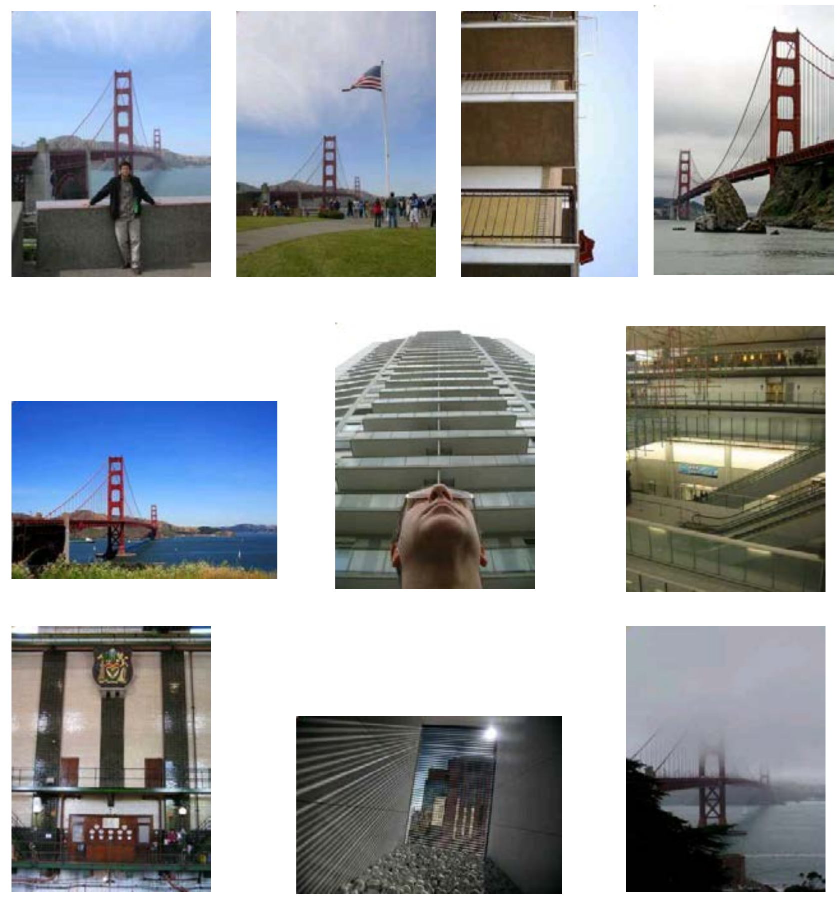
Query

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

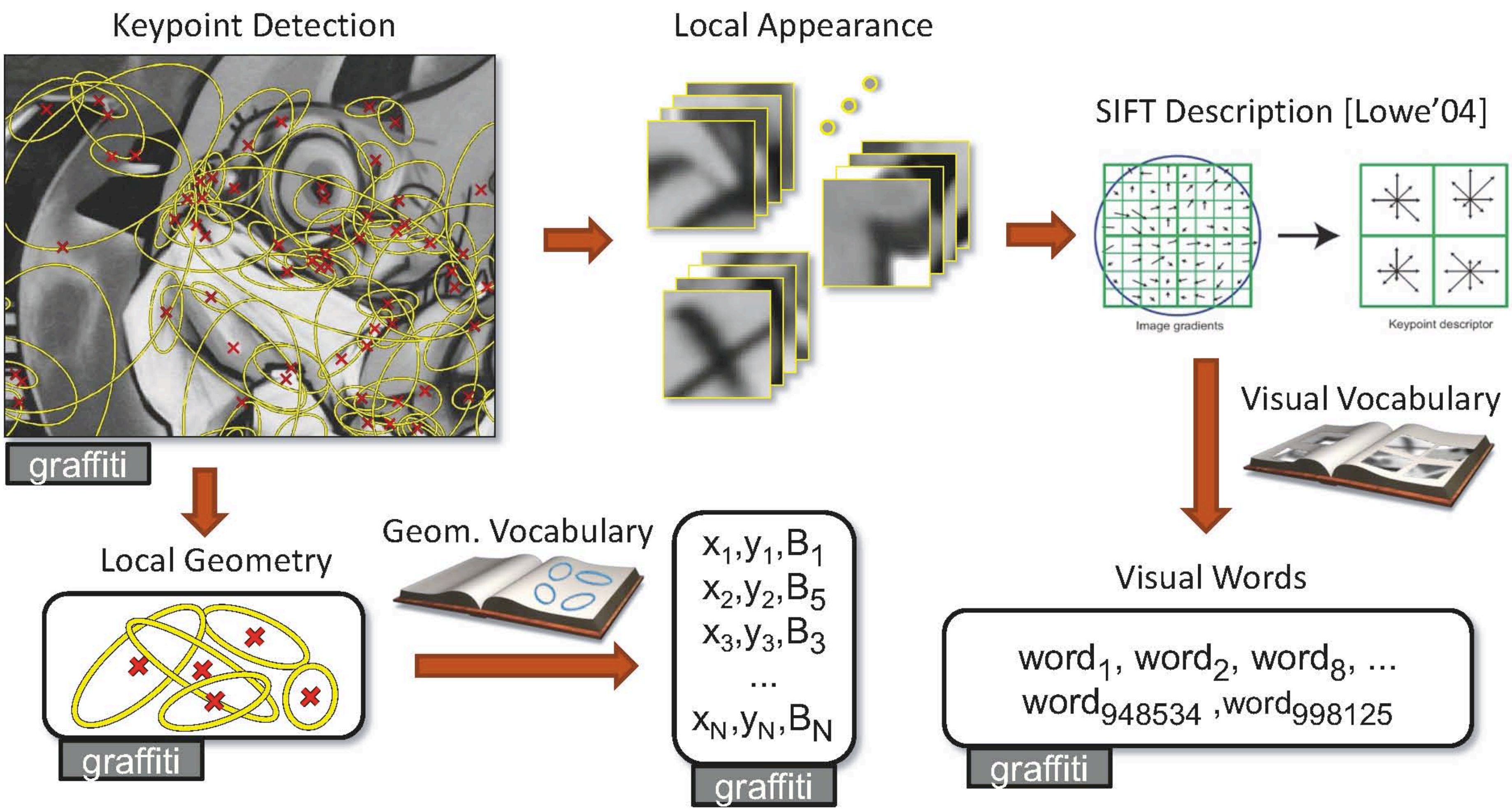
precision = #relevant / #returned
recall = #relevant / #total relevant



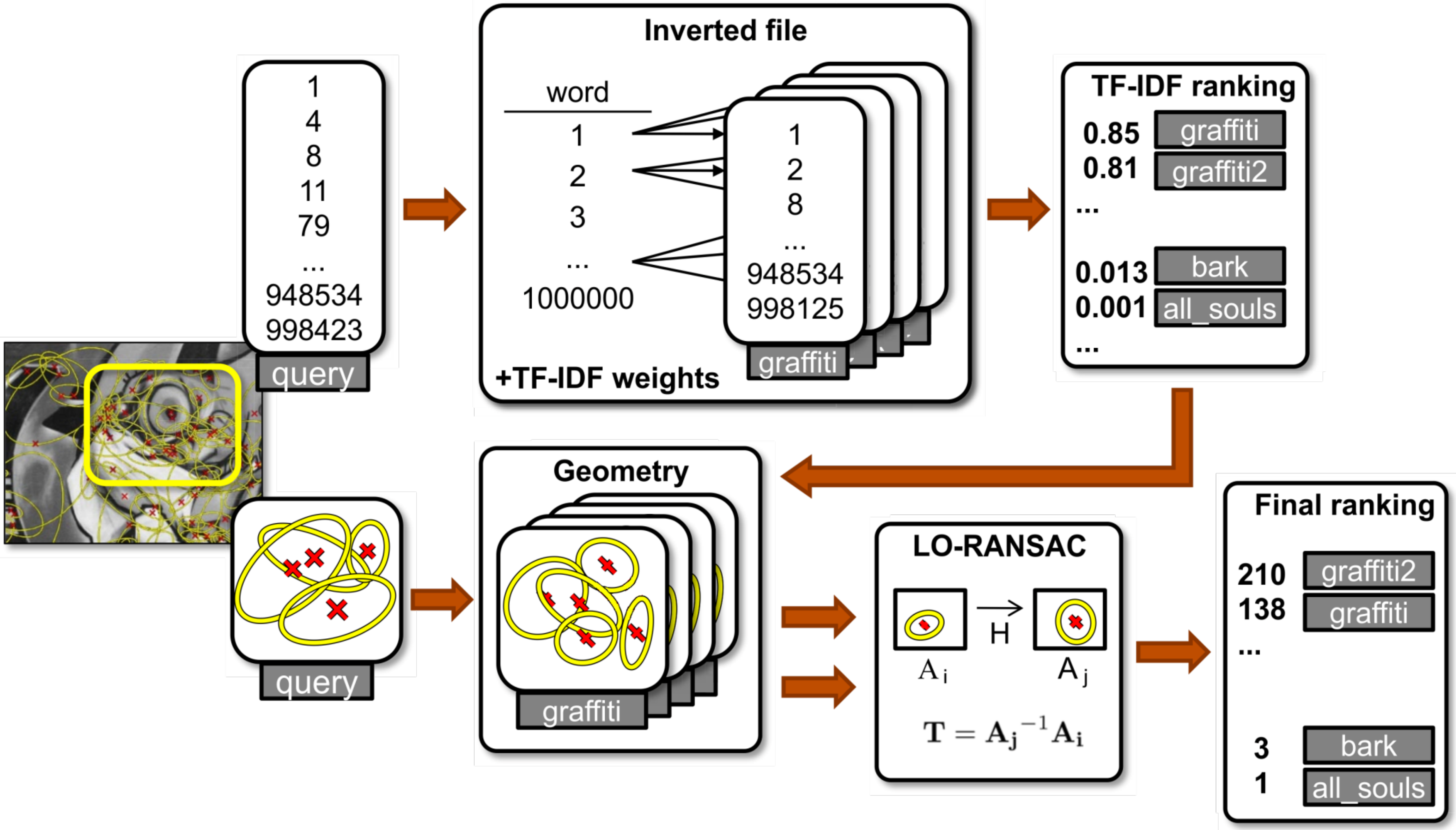
Results (ordered):



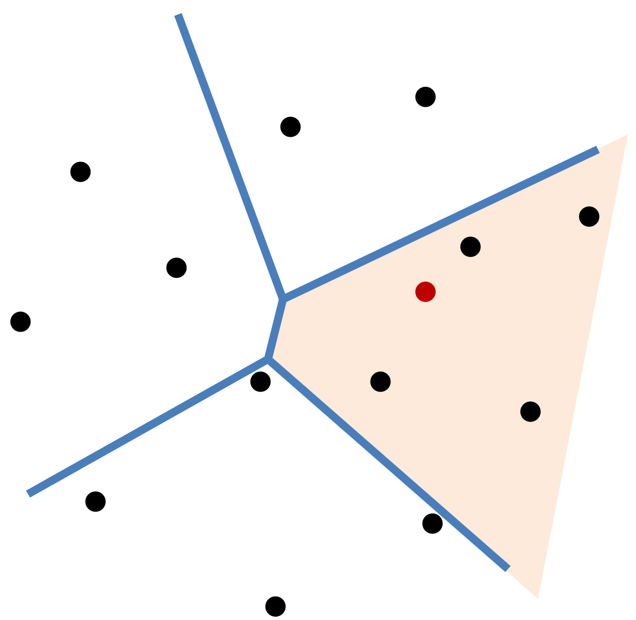
Bag-of-Words (BoW): Off-line Stage



Bag-of-Words : On-line Stage



Feature Distance Approximation

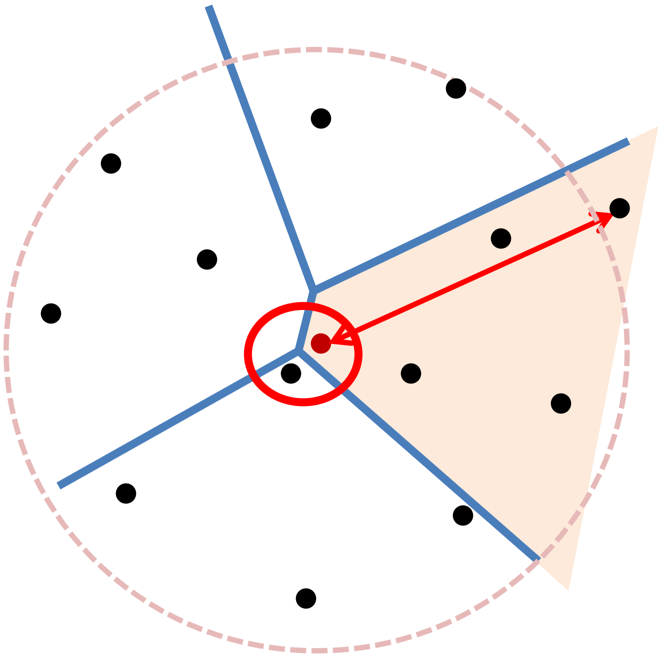


Partition the feature space
(k – means clustering)

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

Feature Distance Approximation



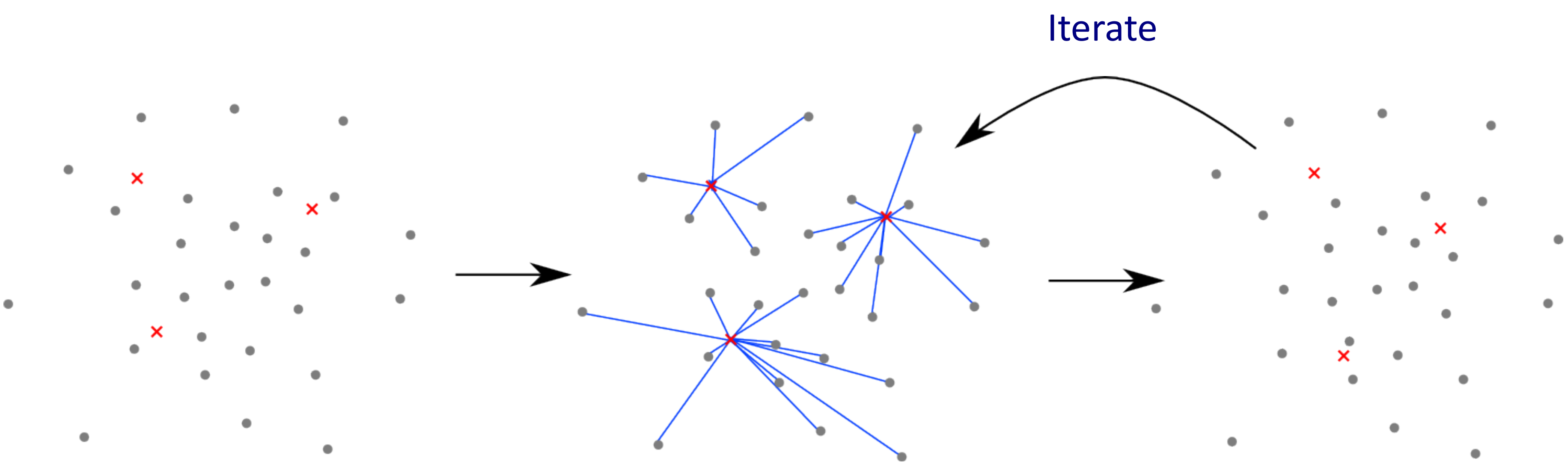
Feature distance

0 : features in the same cell

∞ : features in different cells

- quantization effects
- large (even unbounded) cells

Vector Quantization via k-Means

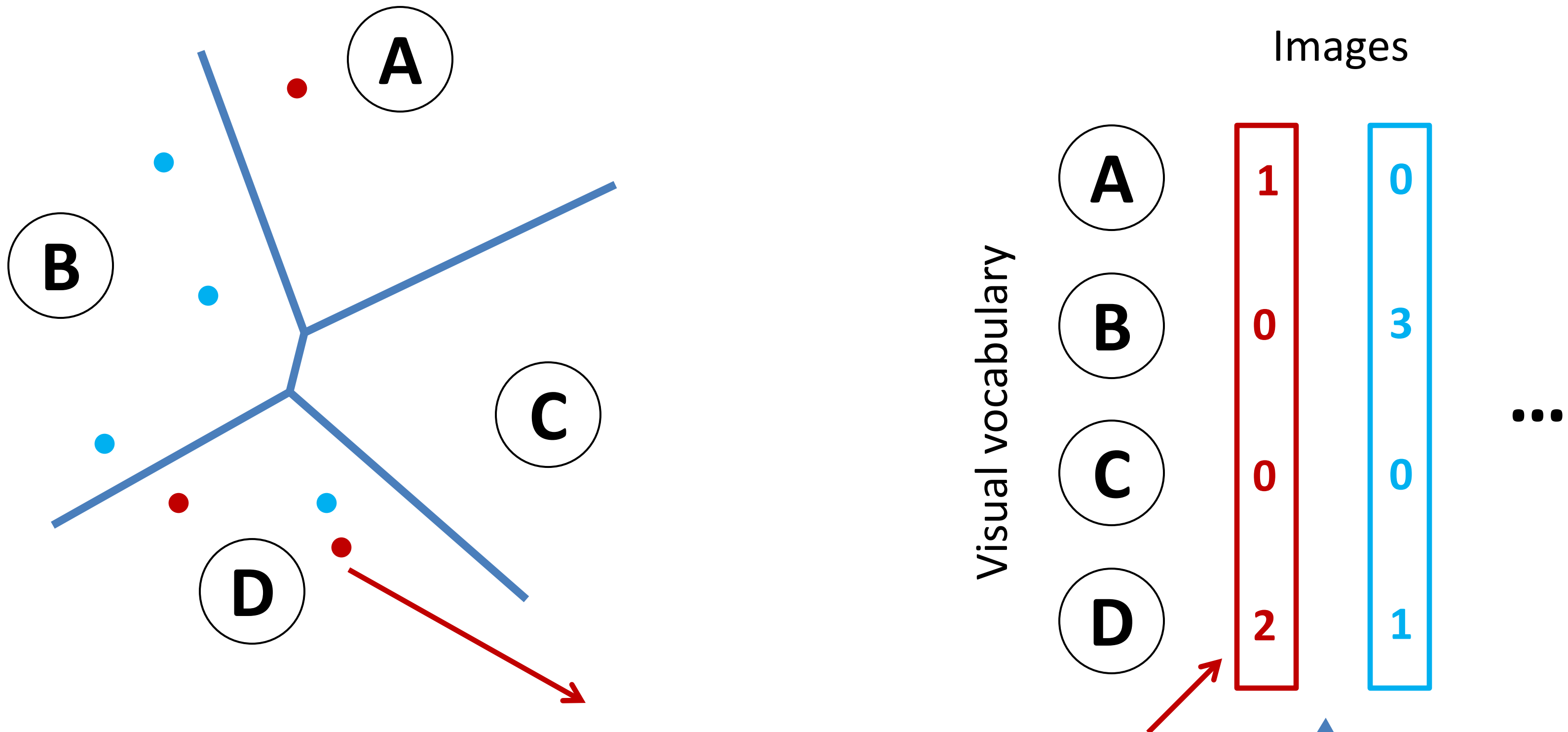


Initialize cluster centres

Find nearest cluster to each datapoint (**slow**) O(N k)

Re-compute cluster centres as centroids

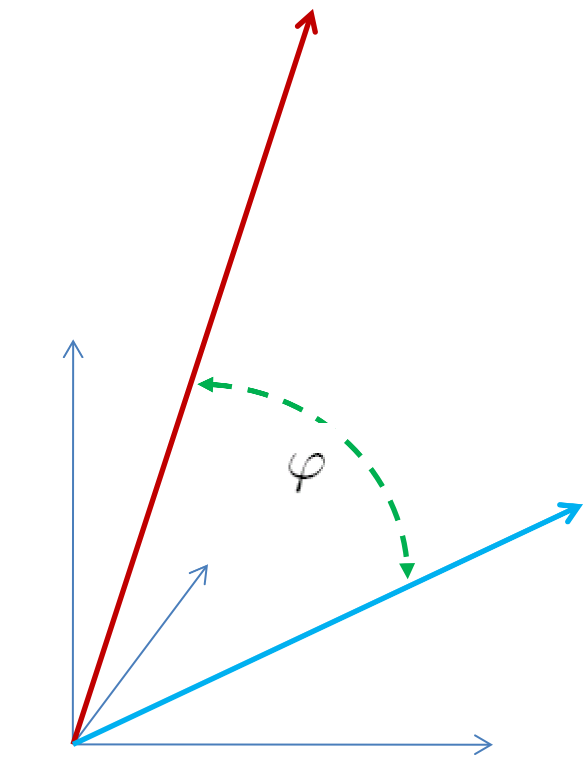
Bags of Words Image Representation



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

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

Efficient Scoring



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

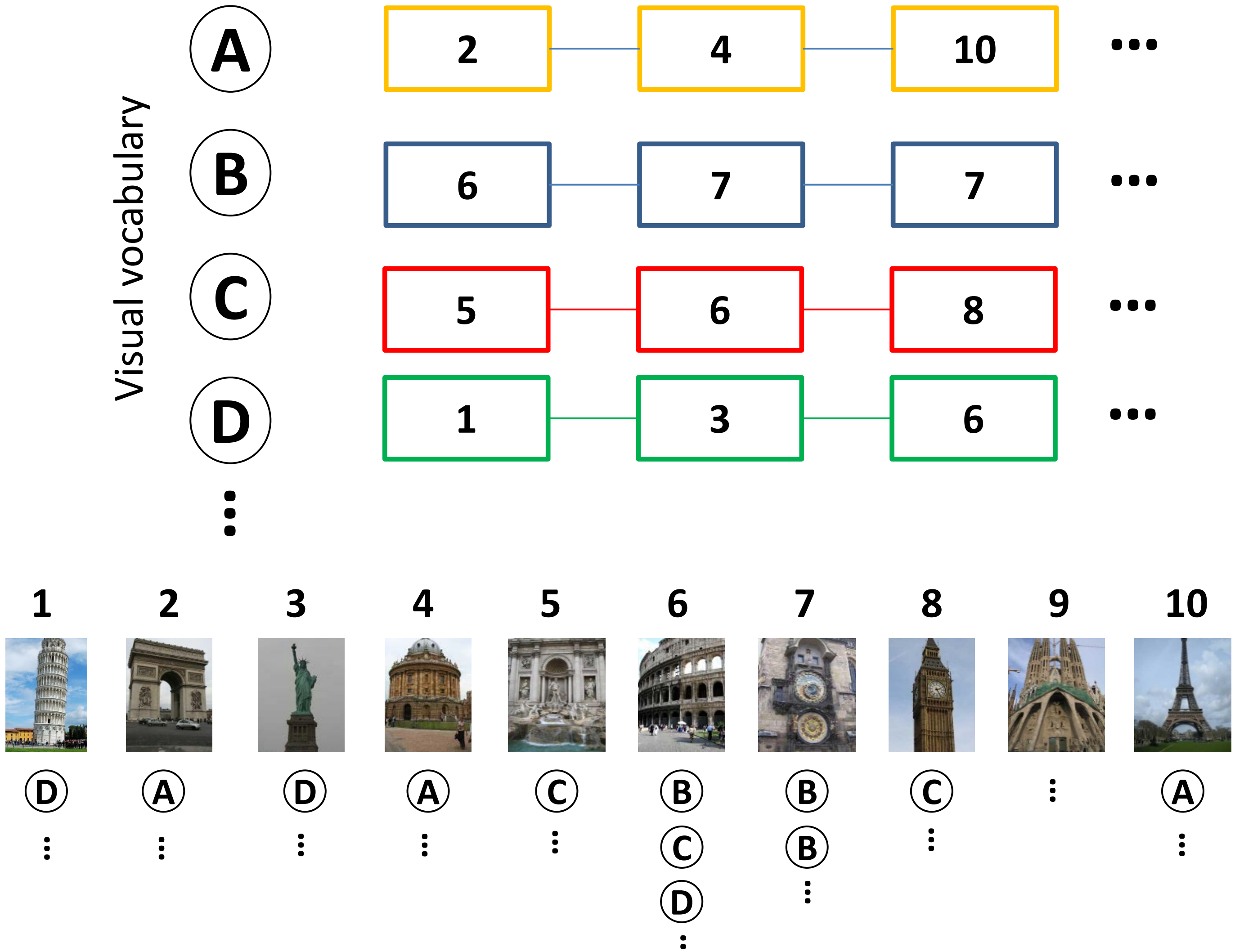
$$\cos\varphi = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||\;||\mathbf{y}||} = \frac{1}{||\mathbf{x}||\;||\mathbf{y}||}\sum_{i=1}^{N} x_i y_i$$

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

Database | Query | Score

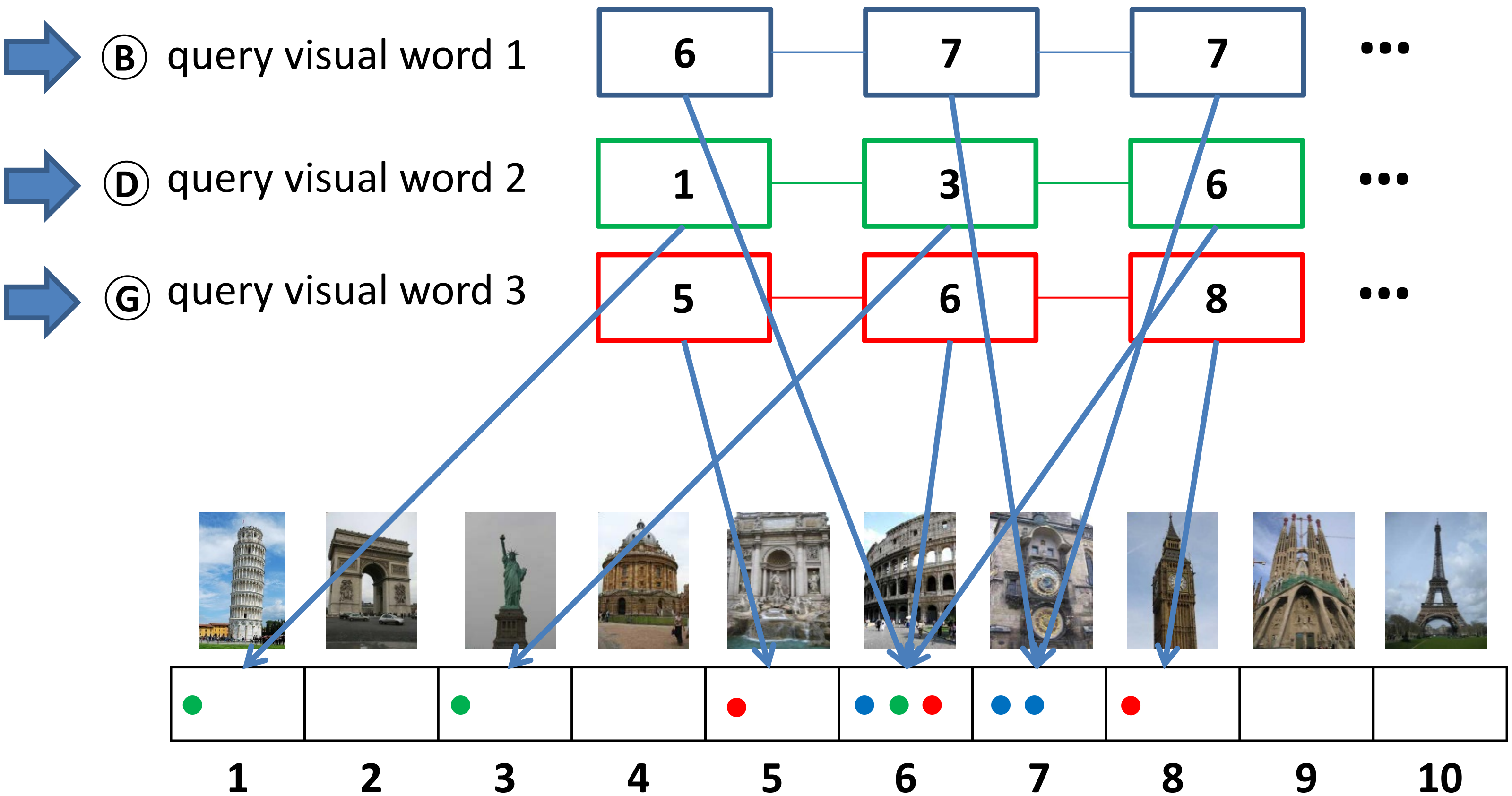
$$\begin{array}{c|cccc} & A & B & C & D \\ \hline \alpha_1 & (1 & 0 & 0 & 2) \\ \alpha_2 & (0 & 2 & 0 & 1) \\ \alpha_3 & (1 & 0 & 0 & 0) \\ & & \vdots & & \end{array} \;\bullet\; \alpha_q \begin{array}{cc} A & 0 \\ B & 3 \\ C & 0 \\ D & 1 \end{array} \;=\; \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ \vdots \end{array}$$

BoW and Inverted File



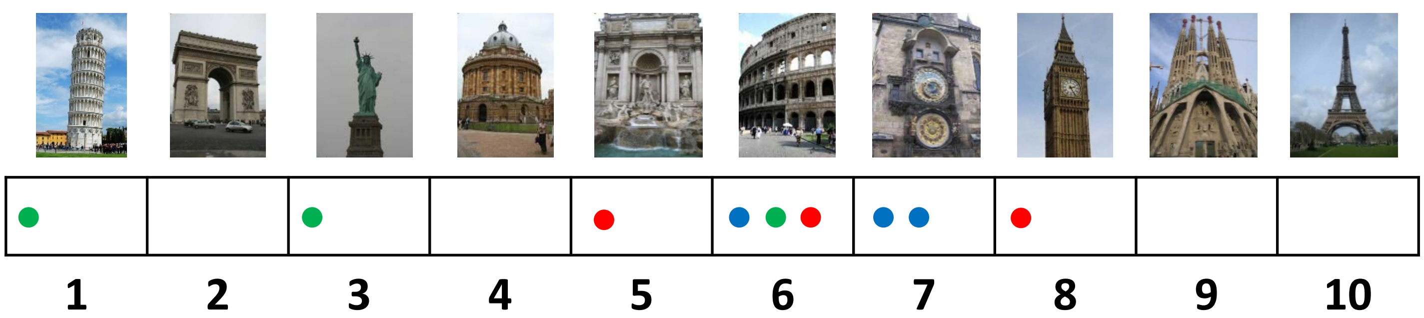
BoW and Inverted File

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



BoW and Inverted File

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



Geometric Re-ranking

1. Perform ranking without geometric information
   - BoW
   - VLAD
   - Fischer vectors
   - CNN descriptors
2. Re-rank top ranked images (removing false positives)
   - RANSAC

Sivic, Zisserman: Video Google, ICCV 2003

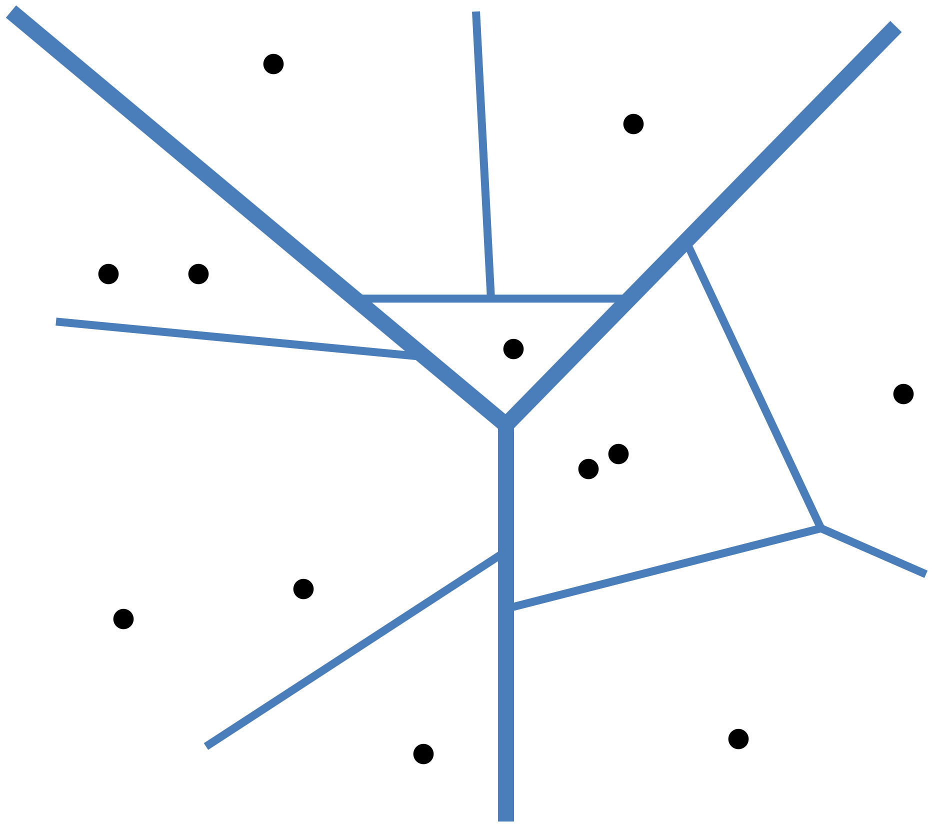
Philbin, Chum, Isard, Sivic, Zisserman: Object retrieval with large vocabularies and fast spatial matching, CVPR’07

Visual Words and Vector Quantization

Vector Quantization

- k-means
- Fixed quantization [Tuytelaars and Schmid ICCV 2007]
- Agglomerative [Leibe, Mikolajczyk and Schiele BMVC 2006]
- Hierarchical k-means
- Approximate k-means
- Hamming embedding
- Learning fine vocabularies

Hierarchical k-means



\+ fast O(N log k)

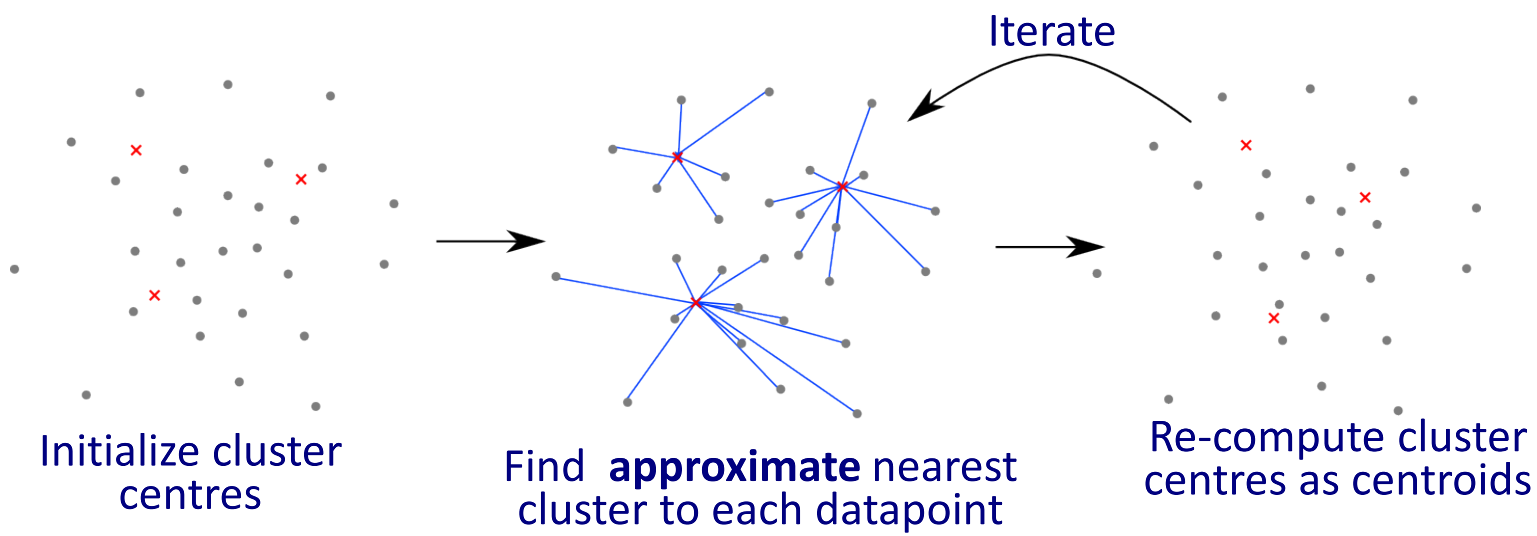
\+ incremental construction

\- not so good quantization

\- often imbalanced

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

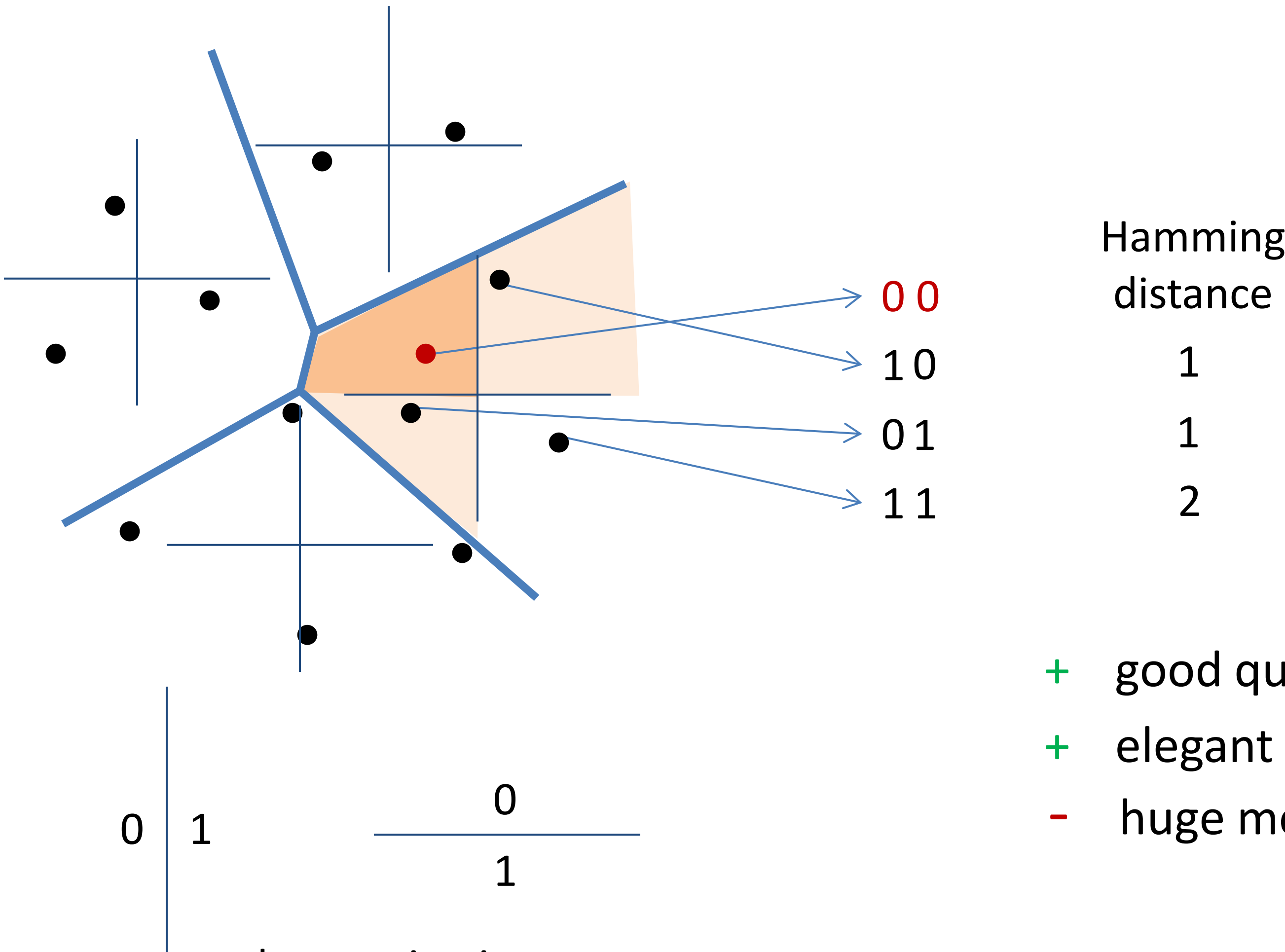
Approximate k-means



- \+ 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

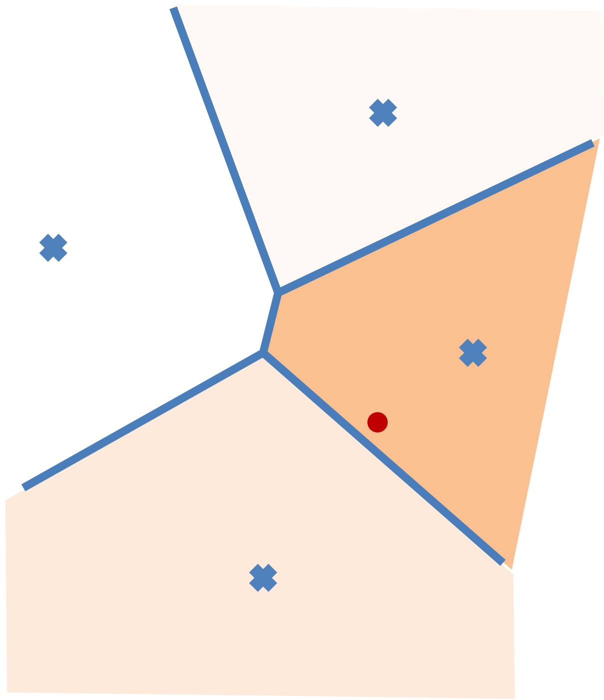
Hamming Embedding



- \+ good quantization
- \+ elegant idea
- \- huge memory footprint

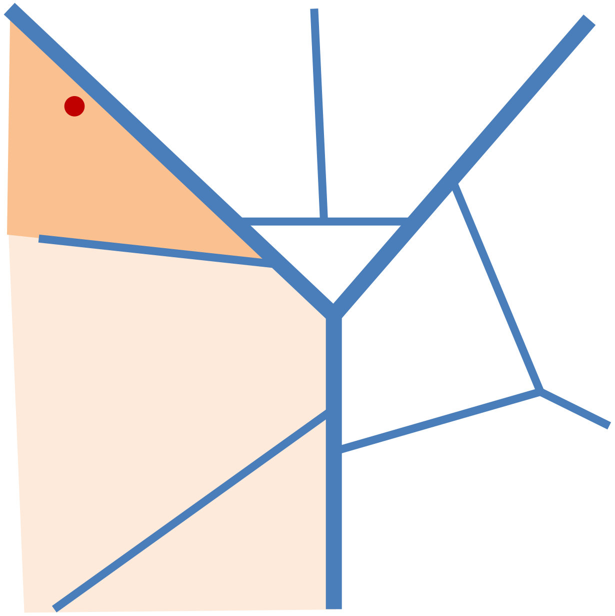
Jegou, Douze, and Schmid – ECCV 2008
Hamming embedding and weak geometric consistency for large scale image search

Soft Assignment



(Approximate) k-means
- database side
- query side

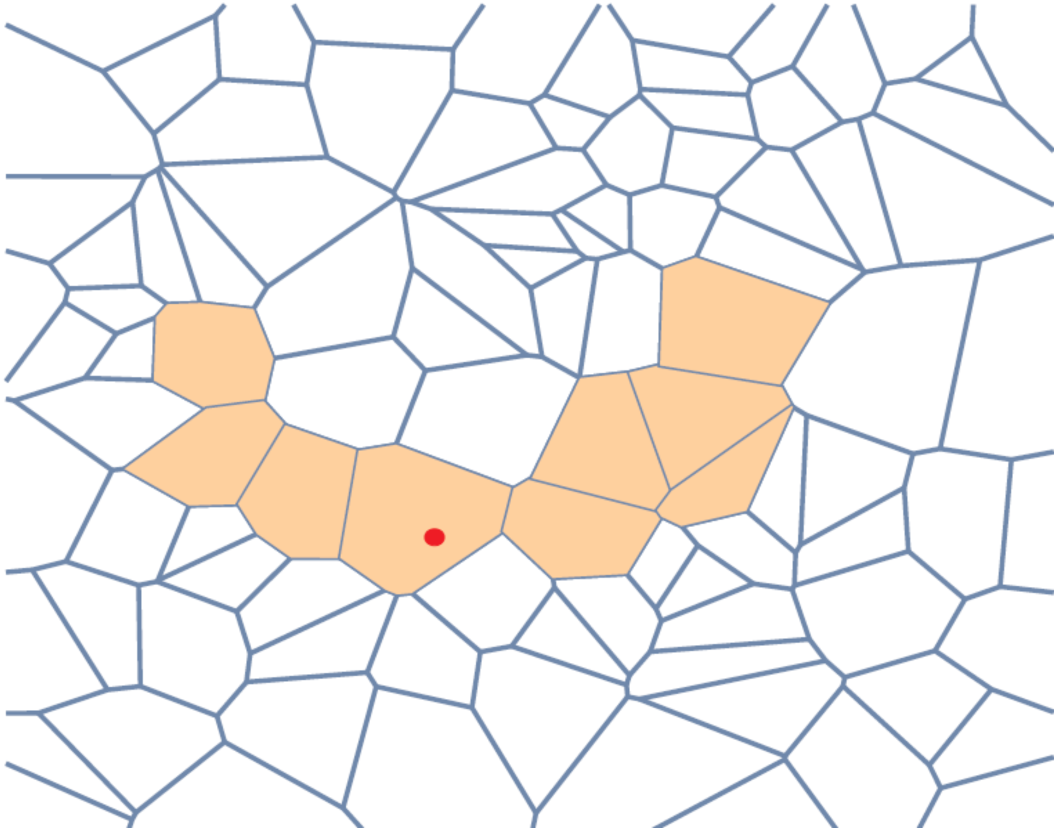
Philbin, Chum, Isard, Sivic, and Zisserman – CVPR 2008
Lost in Quantization



Hierarchical k-means

Nistér & Stewénius – CVPR 2006 Scalable recognition with a vocabulary tree

Learning Fine Vocabularies



Fine vocabulary (16 million visual words)
Using wide-baseline stereo matches on 6 million images to learn what is similar

Mikulik, Perdoch, Chum, and Matas: Learinig a Fine Vocabulary, ECCV 2010

min-Hash

min-Hash

min-Hash is an efficient representation of a set A_i

$$m(\mathcal{A}_i, f) = \arg\min_{X \in \mathcal{A}_i} f(X)$$

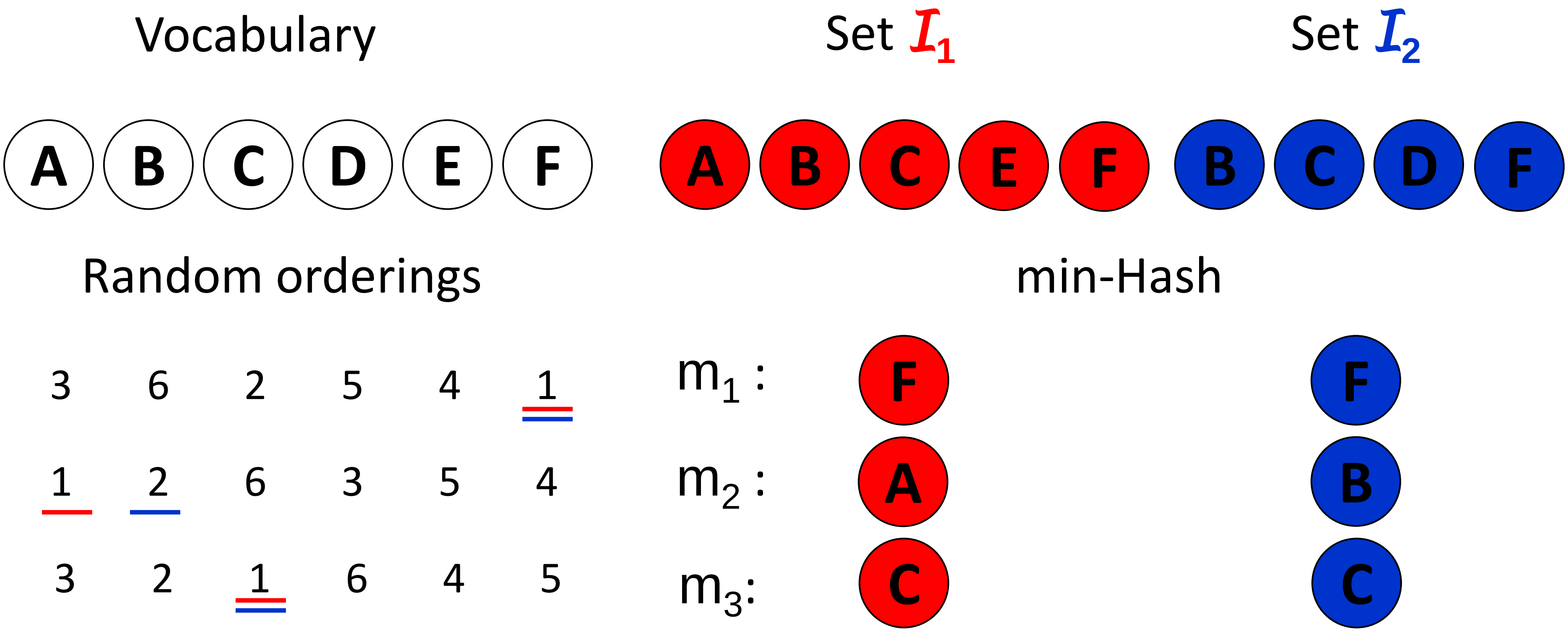
set of visual words   hash function   visual word

min-Hash is a locality sensitive hashing (LSH) function m that selects an element (visual word) $m(A_i)$ from each set A_i of visual words detected in image *i* so that

$$P\{m(\mathcal{I}_1) == m(\mathcal{I}_2)\} = \frac{|\mathcal{I}_1 \cap \mathcal{I}_2|}{|\mathcal{I}_1 \cup \mathcal{I}_2|}$$

Image similarity $\quad \mathrm{sim}(\mathcal{I}_1, \mathcal{I}_2) = \frac{|\mathcal{I}_1 \cap \mathcal{I}_2|}{|\mathcal{I}_1 \cup \mathcal{I}_2|}$

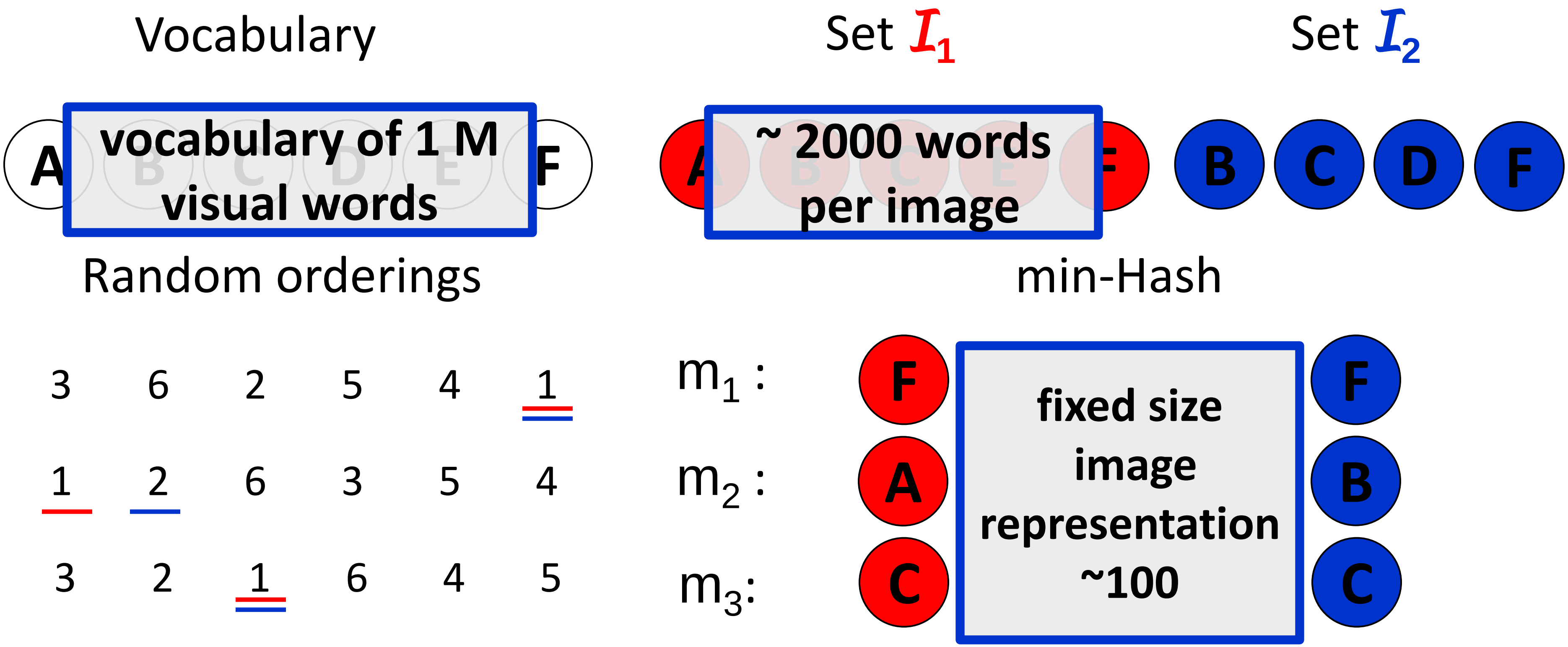
min-Hash



sim (I_1, I_2) = 1/2

Estimated similarity of I_1 and I_2 from 3 min-Hashes = 2/3

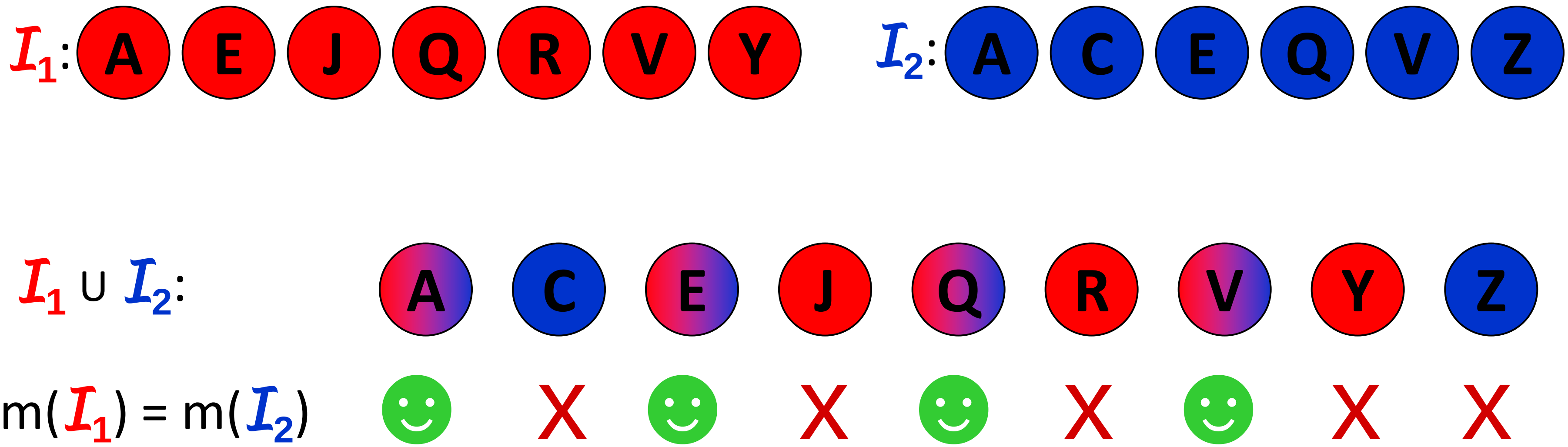
min-Hash

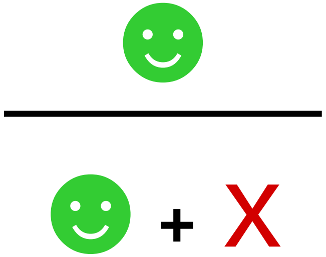


sim (I_1, I_2) = 1/2

Estimated similarity of I_1 and I_2 from 3 min-Hashes = 2/3

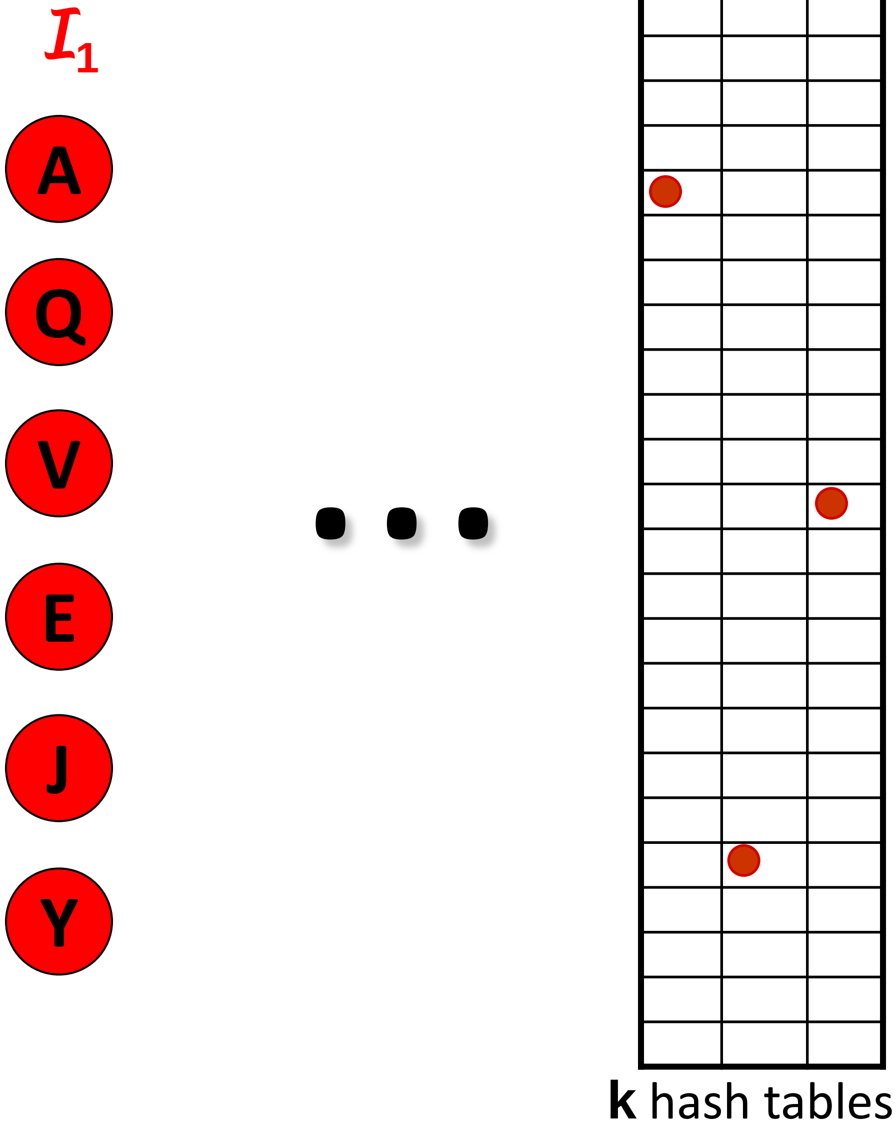
Set Overlap and min-Hash



$P(m(\mathcal{I}_1) = m(\mathcal{I}_2)) =$  $= \dfrac{|\mathcal{I}_1 \cap \mathcal{I}_2|}{|\mathcal{I}_1 \cup \mathcal{I}_2|}$

min-Hash

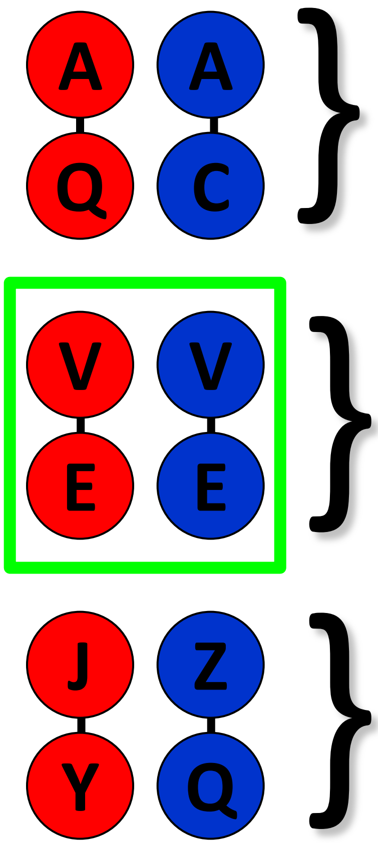
a sketch = **s**-tuple of min-Hashes

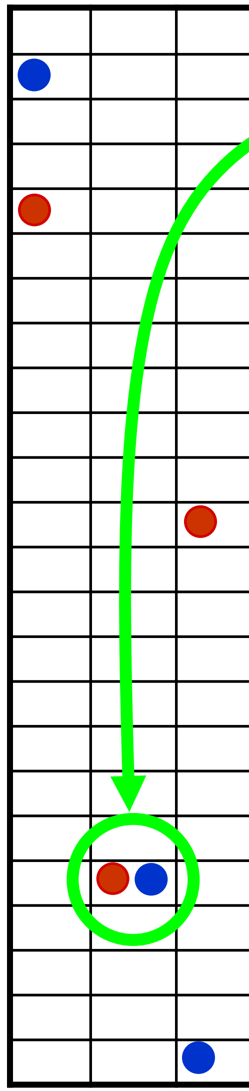


min-Hash

a sketch = **s**-tuple of min-Hashes

$\mathcal{I}_1$ $\mathcal{I}_2$





k hash tables

Sketch collision

collision:
all s min-Hashes must agree

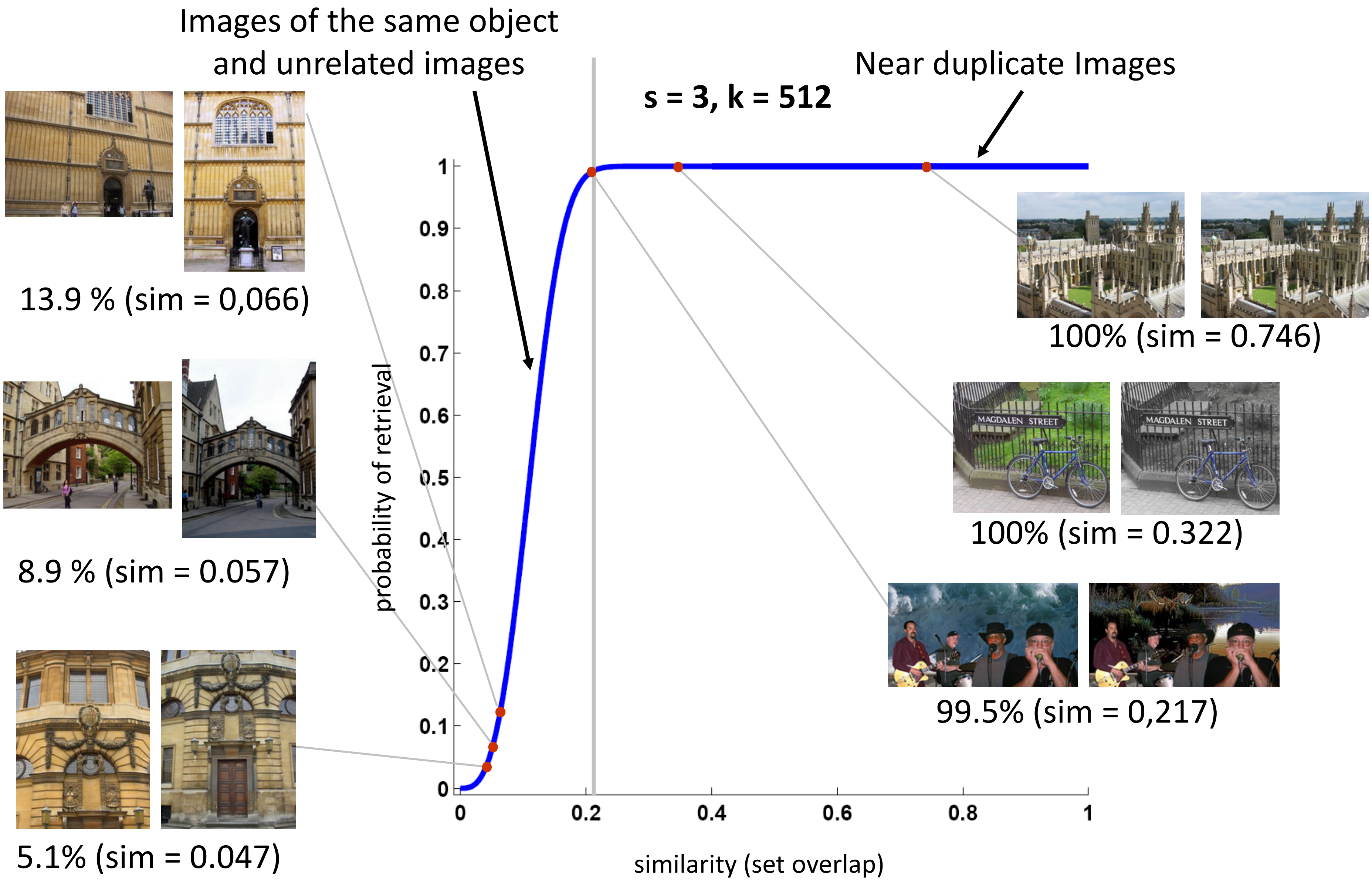
P{collision} = $\mathrm{sim}(\mathcal{I}_1,\mathcal{I}_2)^s$

retrieval:

1. generate k sketches
2. at least one of k sketches must collide

P{retrieval} =
$1 - (1 - \mathrm{sim}(\mathcal{I}_1, \mathcal{I}_2)^s)^k$

Probability of Retrieving an Image Pair



Weighted min-Hash

For hash function (set overlap similarity) $f_j(X_w) = x \quad x \sim \mathrm{Un}(1,0)$

all words X_w have the same chance to be a min-Hash

For hash function $\boxed{f_j(X_w) = \frac{-\log x}{d_w} \qquad x \sim \mathrm{Un}(1,0)}$

the probability of X_w being a min-Hash is proportional to d_w

$I_1 \cup I_2$: A C E J Q R V Y Z

d_A d_C d_E d_J d_Q d_R d_V d_Y d_Z

$$P(m(\mathcal{A}) = m(\mathcal{B})) = \frac{\sum_{X_w \in \mathcal{A} \cap \mathcal{B}} d_w}{\sum_{X_w \in \mathcal{A} \cup \mathcal{B}} d_w}$$

Chum, Philbin, Zisserman: Near Duplicate Image Detection: min-Hash and tf-idf Weighting, BMVC 2008

Image Clustering via min-Hash

Image Clusters as Connected Components

Standard Approach (using image retrieval):

Quadratic method in the size of database D -- $O(D^2)$
the multiplicative constant at the quadratic term ~ 1 – quadratic even for small D

1. Take each image in turn
2. Use a image retrieval system to retrieve related images
3. Compute connected components of the graph

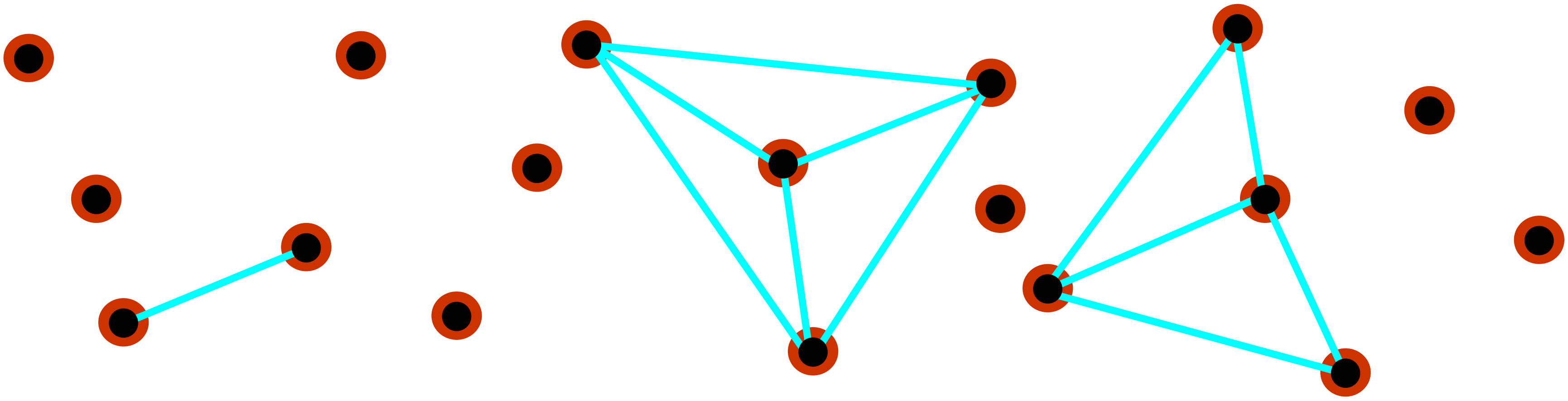
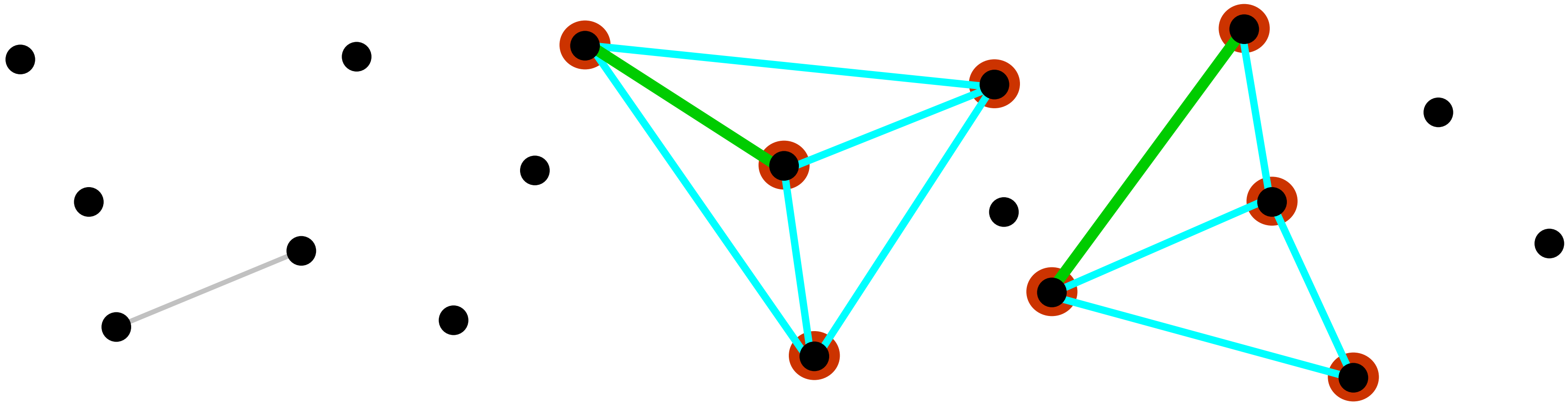


Image Clusters as Connected Components

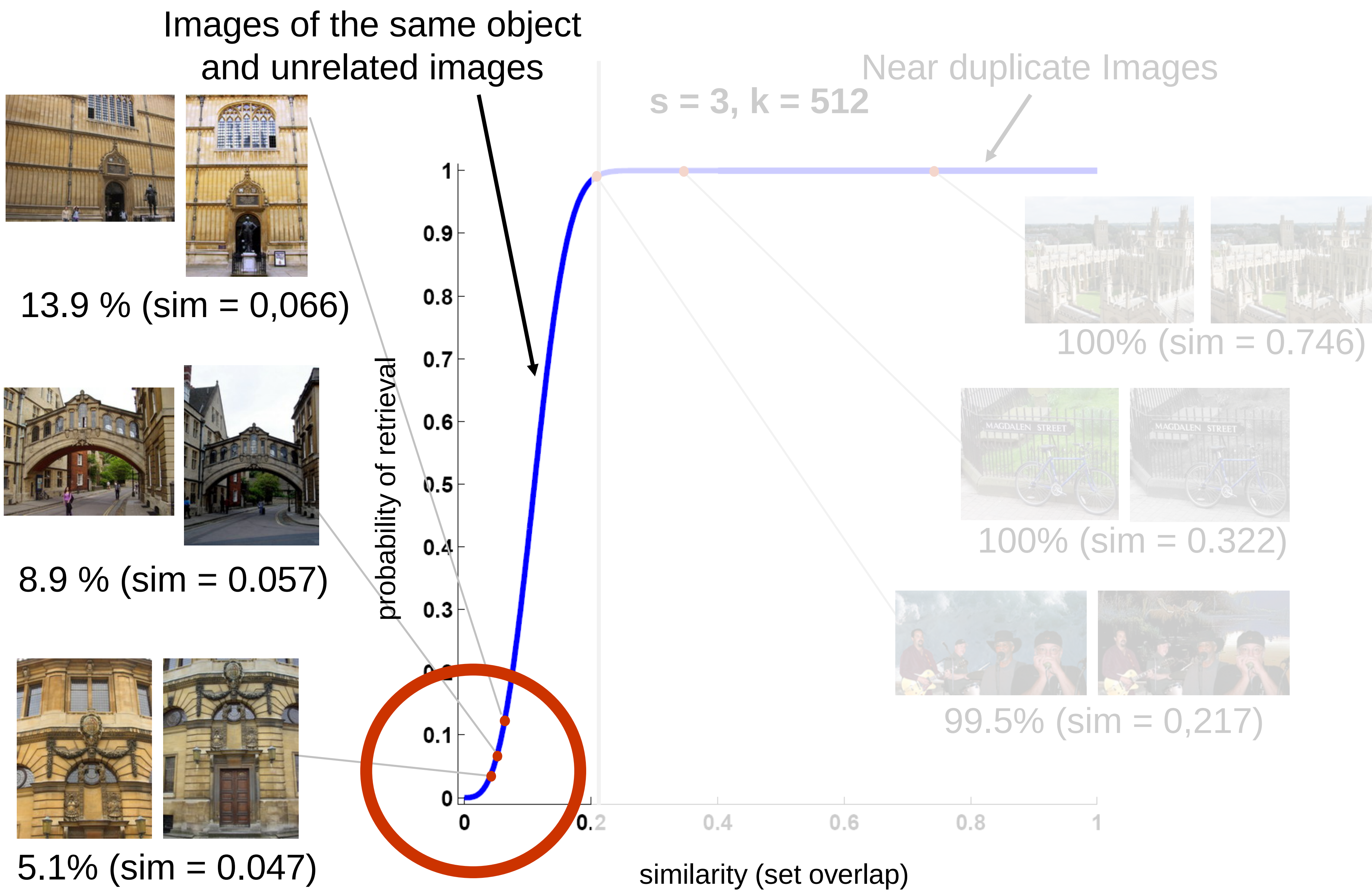
Proposed method:

1. Seed Generation – hashing (fast, low recall)
   characterize images by pseudo-random numbers stored in a hash table
   time complexity equal to the sum of second moments of Poisson random variable -- linear for database size D up to 2^{50}

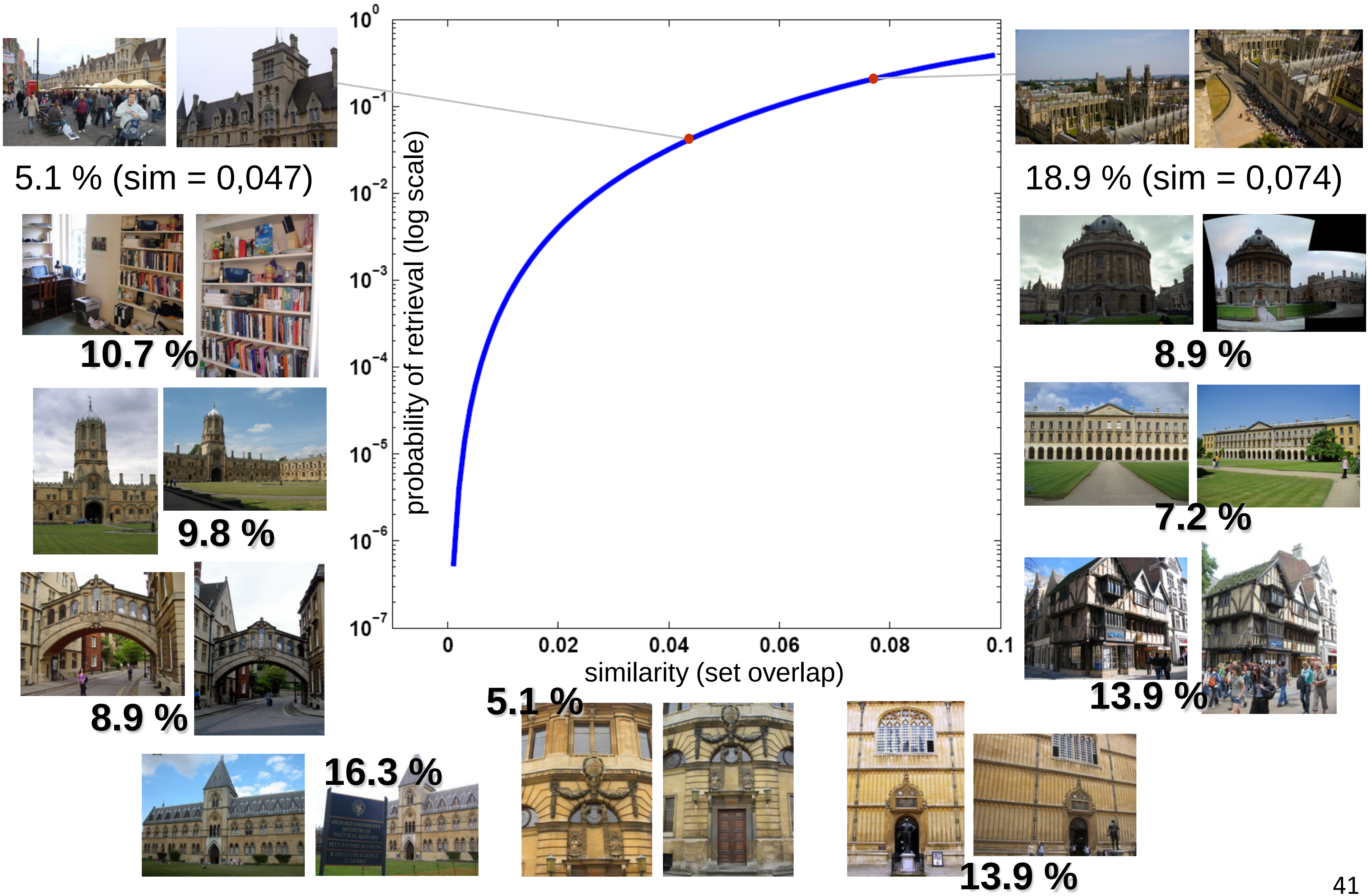
2. Seed Growing – retrieval (thorough – high recall)
   complete the clusters only for cluster members c << D, complexity *O(cD)*



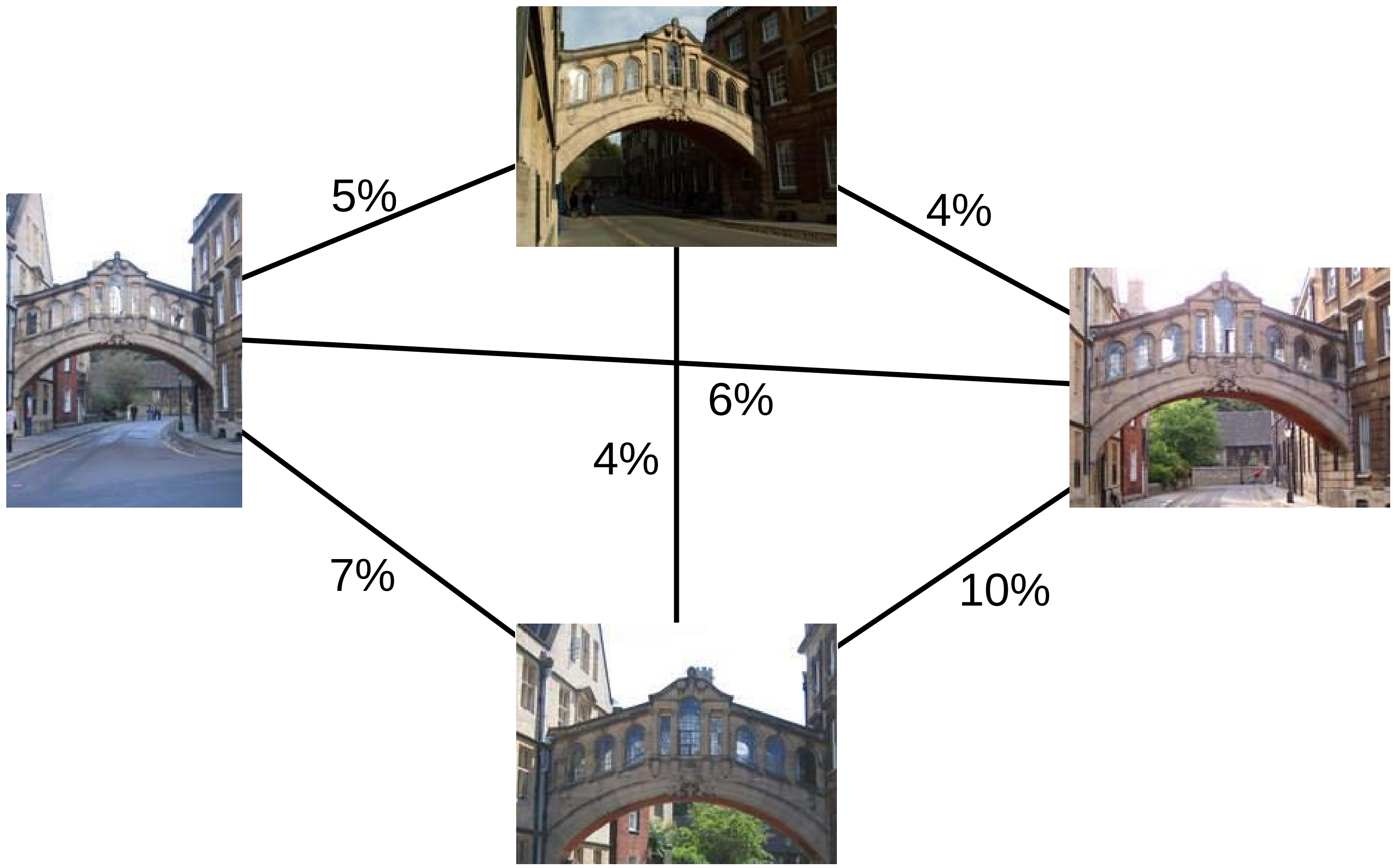
Probability of Retrieving an Image Pair



Spatially Related Images

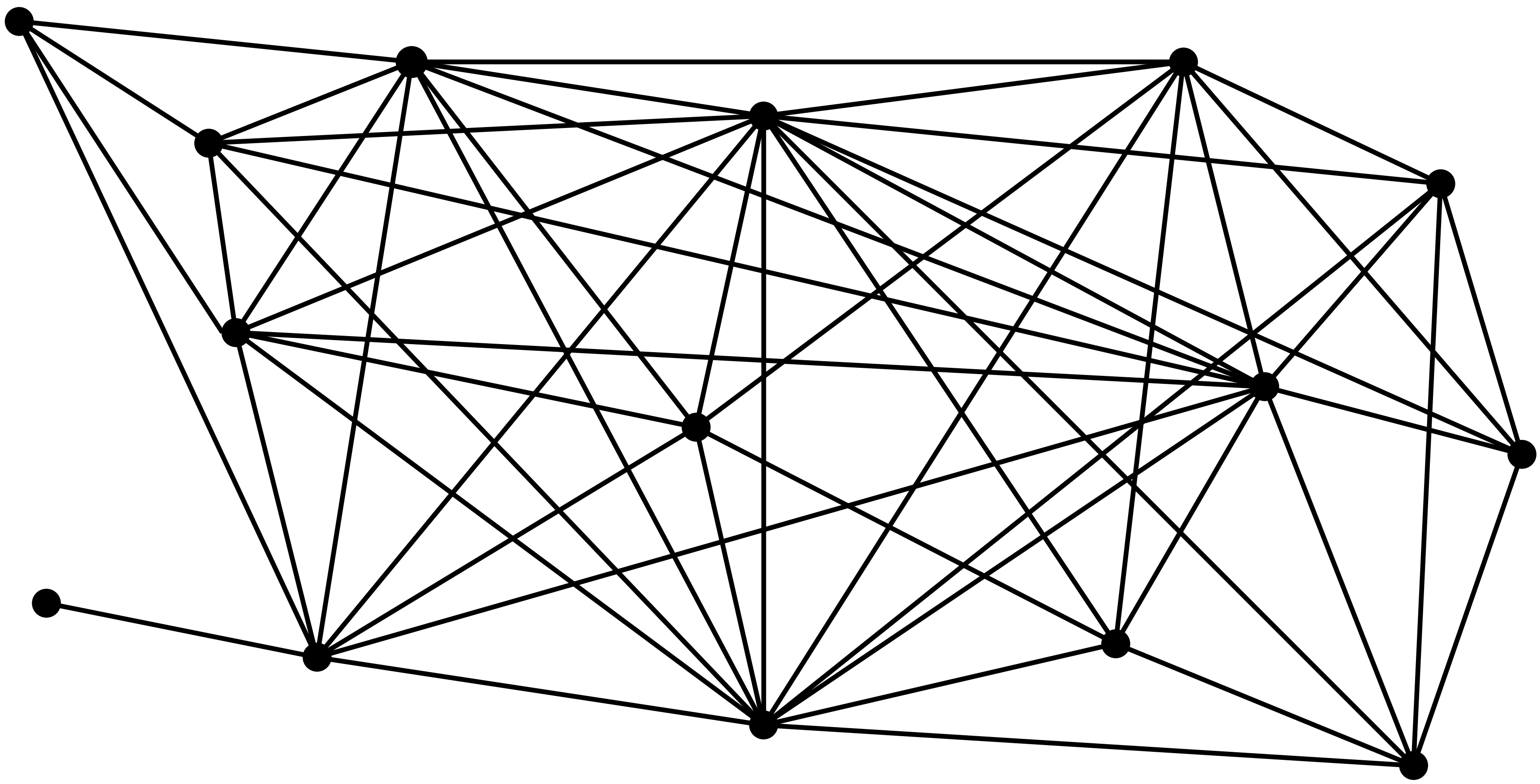


Seed Generation



P (no seed) = 68.88 %

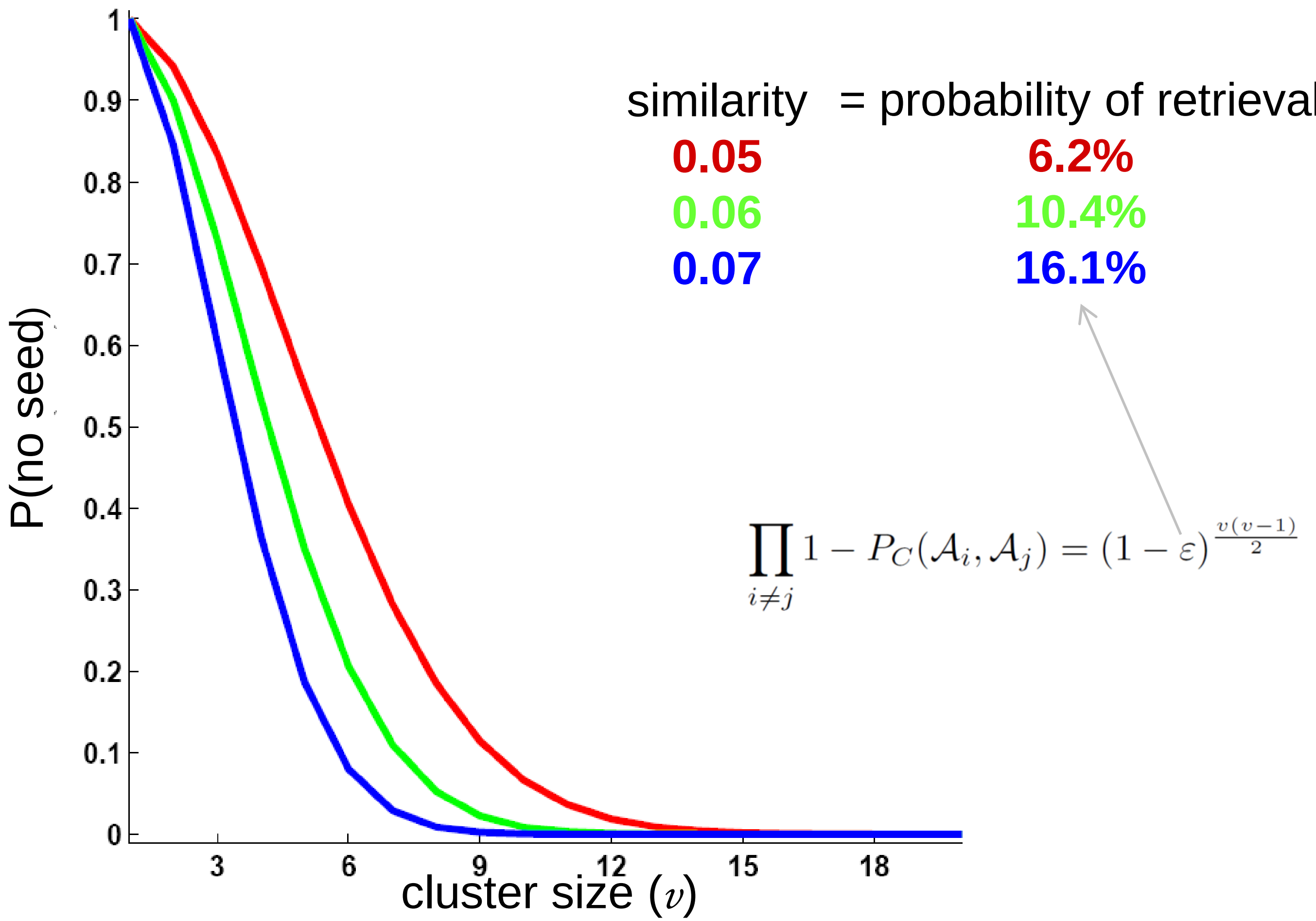
Seed Generation



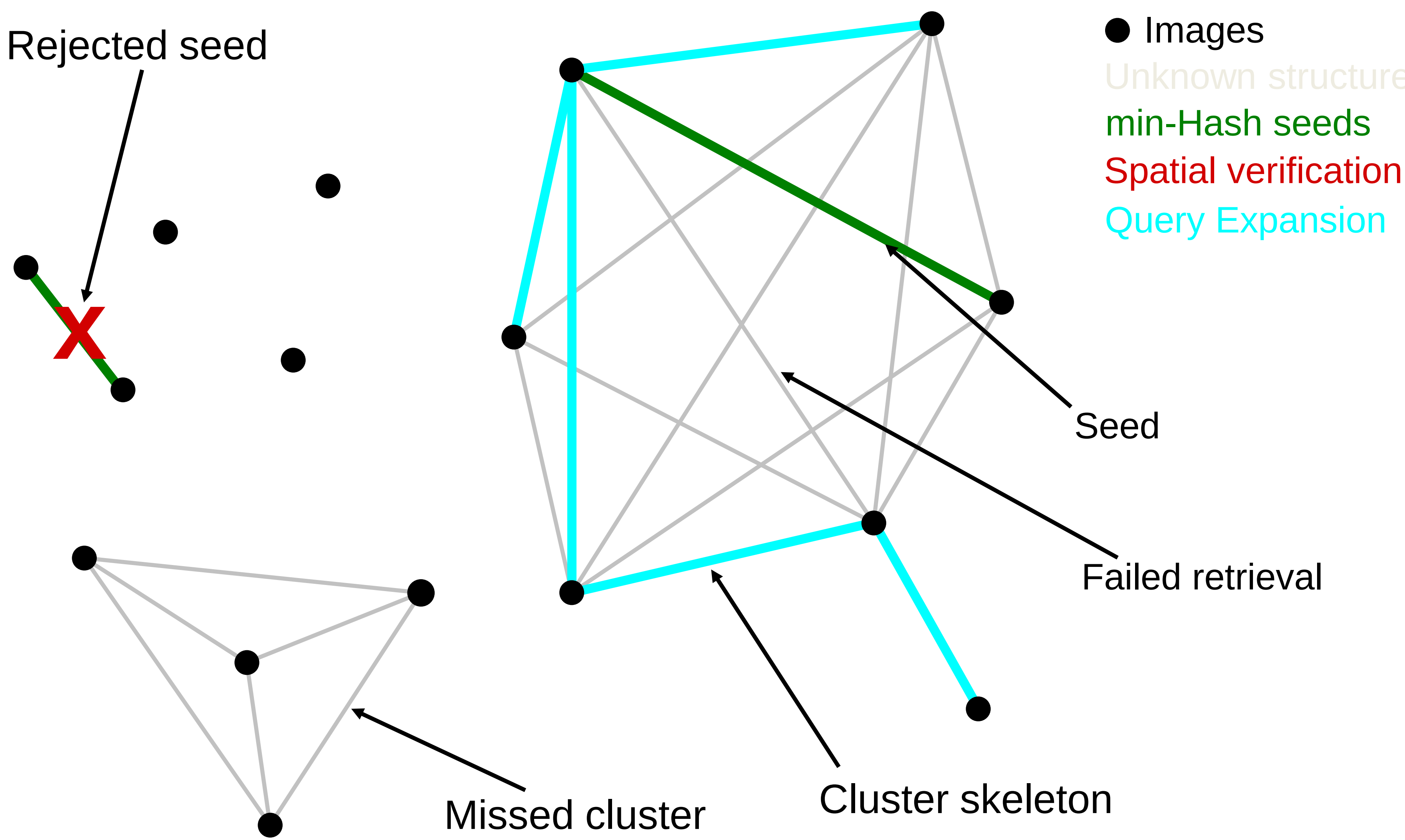
P (no seed) = 1.94 %

At Least One Seed in Cluster

Estimate of the probability of failure plot against the size of the cluster
assumption used in this plot: all images in the cluster are related

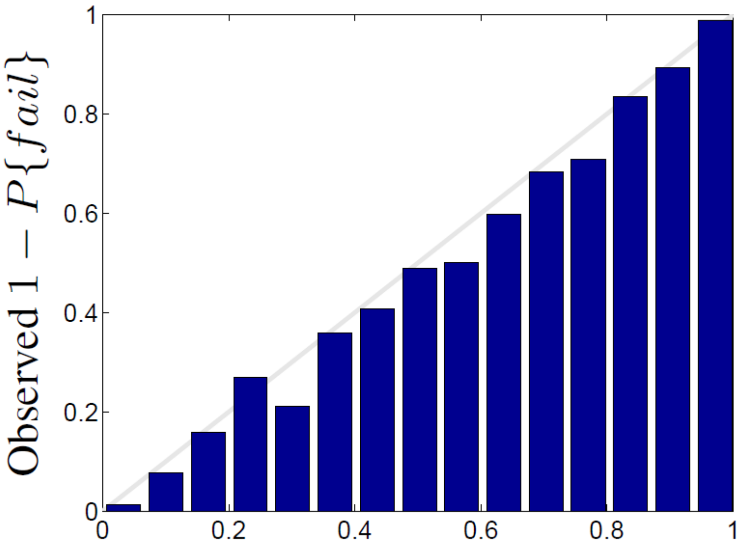


Summary of the Method



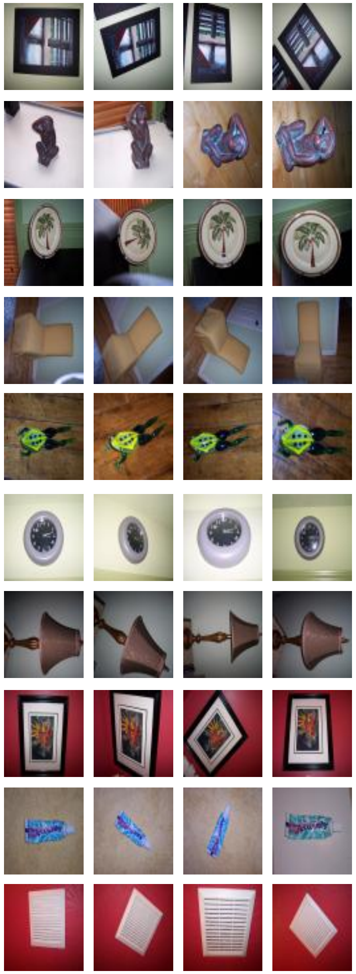
UKY Dataset

Cluster of 4 images = 6 image pairs
Are the probabilities of retrieval (close to) independent?



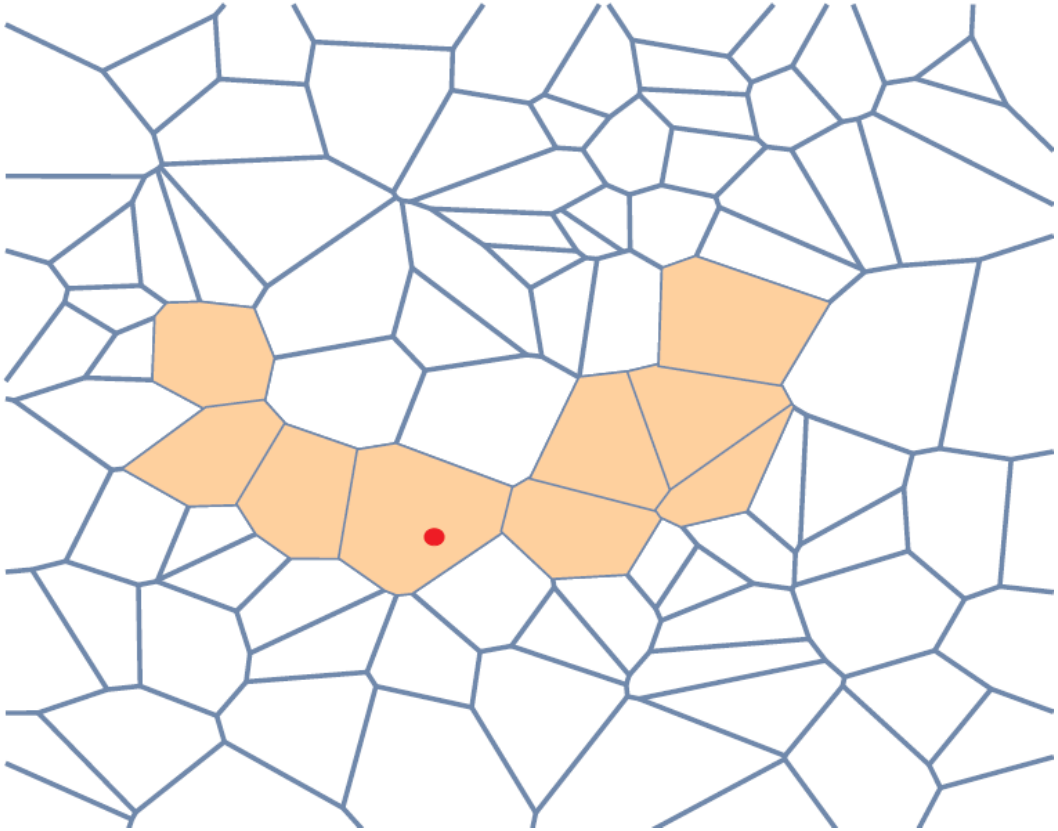
$$1 - (1 - \varepsilon)^6$$

computed exactly from the images descriptors



Application

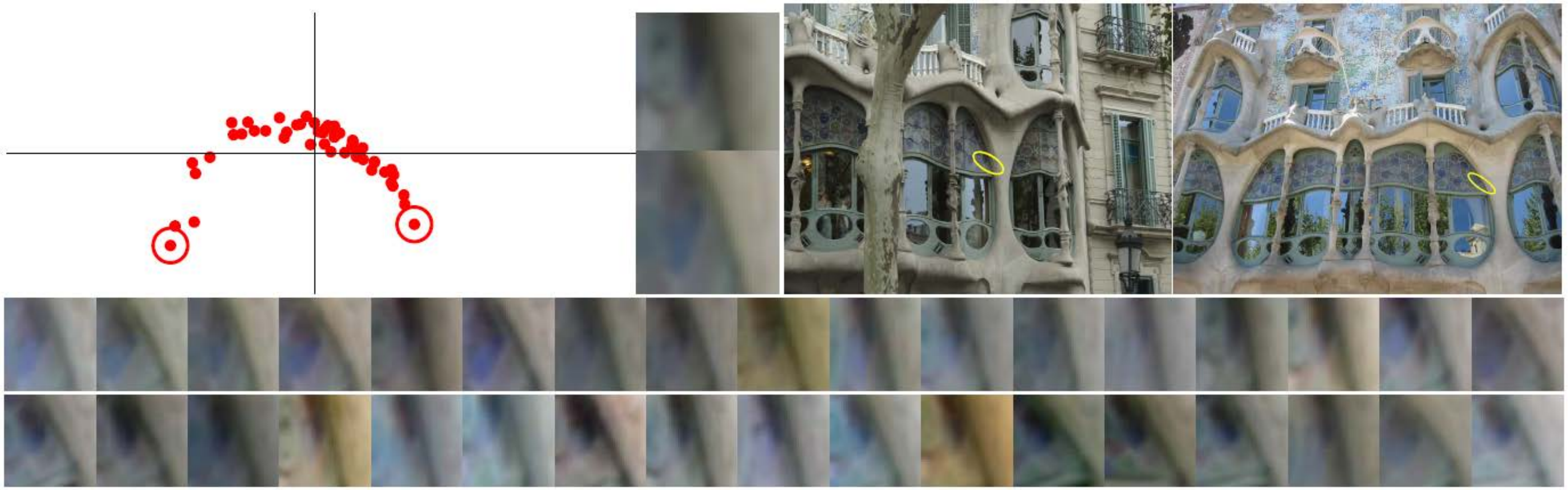
Learning Fine Vocabularies



Fine vocabulary (16 million visual words)
Using wide-baseline stereo matches on 6 million images to learn what is similar

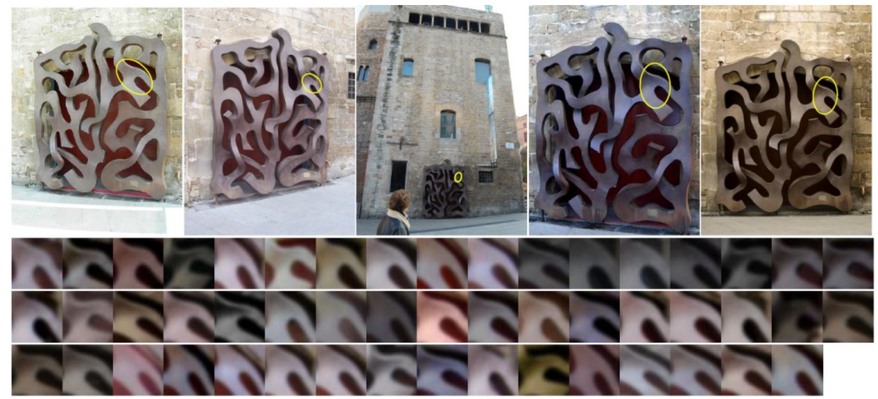
Mikulik, Perdoch, Chum, and Matas: Learinig a Fine Vocabulary, ECCV 2010

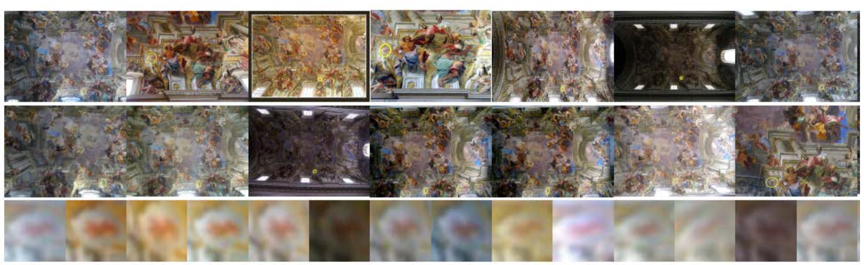
Appearance Variance of a Single Feature



Mikulik, Perdoch, Chum, Matas: Learning Vocabularies over a Fine Quantization, IJCV 2012

- over 5 million images
- almost 20k clusters of 750k images (visual word based)
- 733k successfully matched in WBS matching (raw descriptor based)
- over 111 M feature tracks established (12.3 M with 6+ features)
- 564 M features in the tracks (319.5 M in tracks of 6+ features)





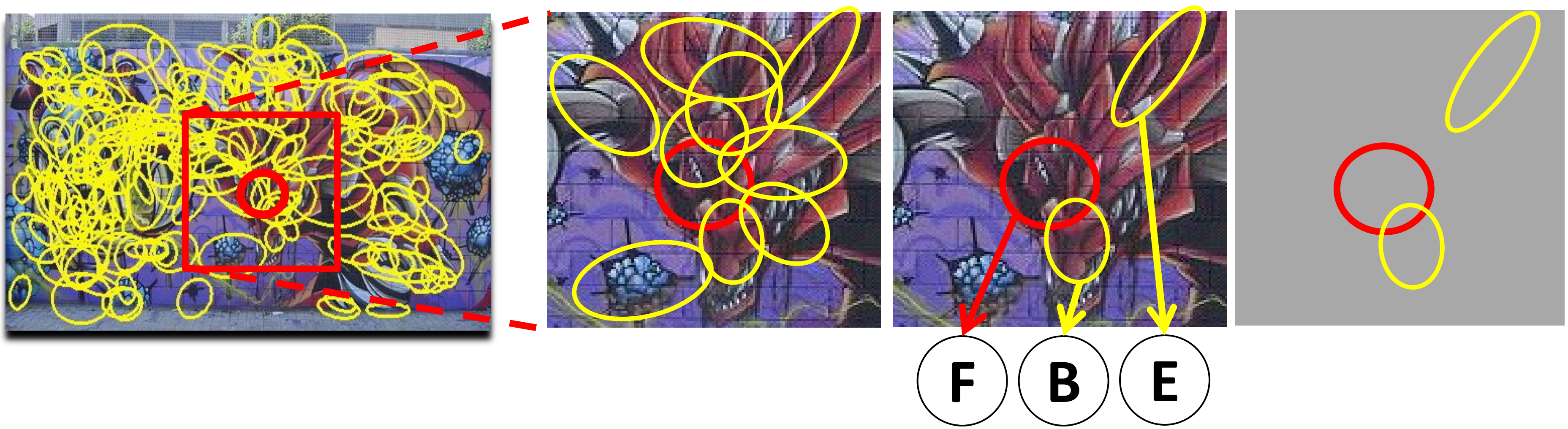
http://cmp.felk.cvut.cz/~qqmikula/publications/ijcv2012/index.html

Geometric min-Hash

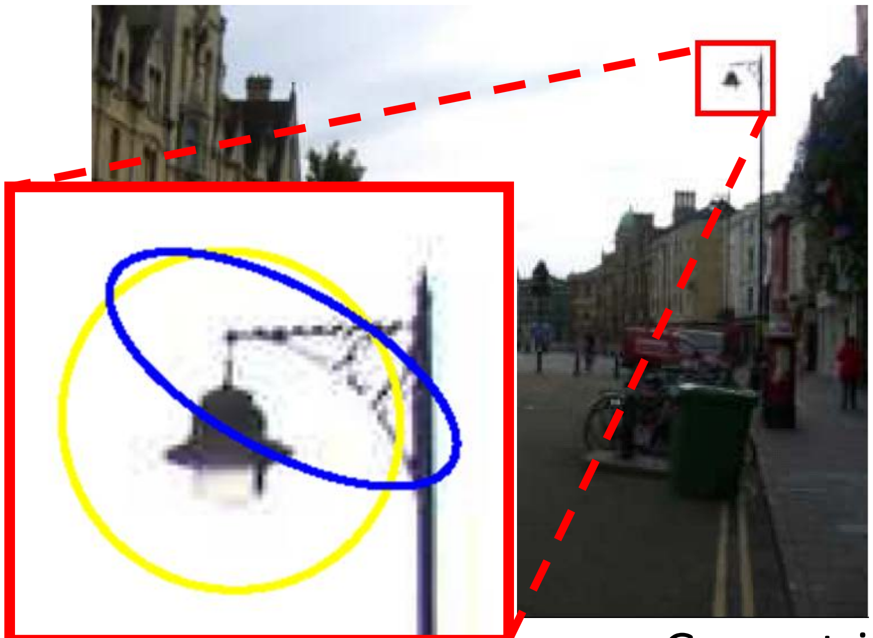
Geometric min-Hash algorithm

1. Keep features with unique visual word in the image
2. Obtain the “central feature” by min-Hash
3. Select scale and spatial neighbourhood of the central feature
4. Select secondary min-Hash(es) from the neighbourhood
5. Relative pose of the sketch features is a geometric invariant (as in geometric hashing)

Sketch of GmH: s-tuple of visual words + geometric invariant



Object Discovery



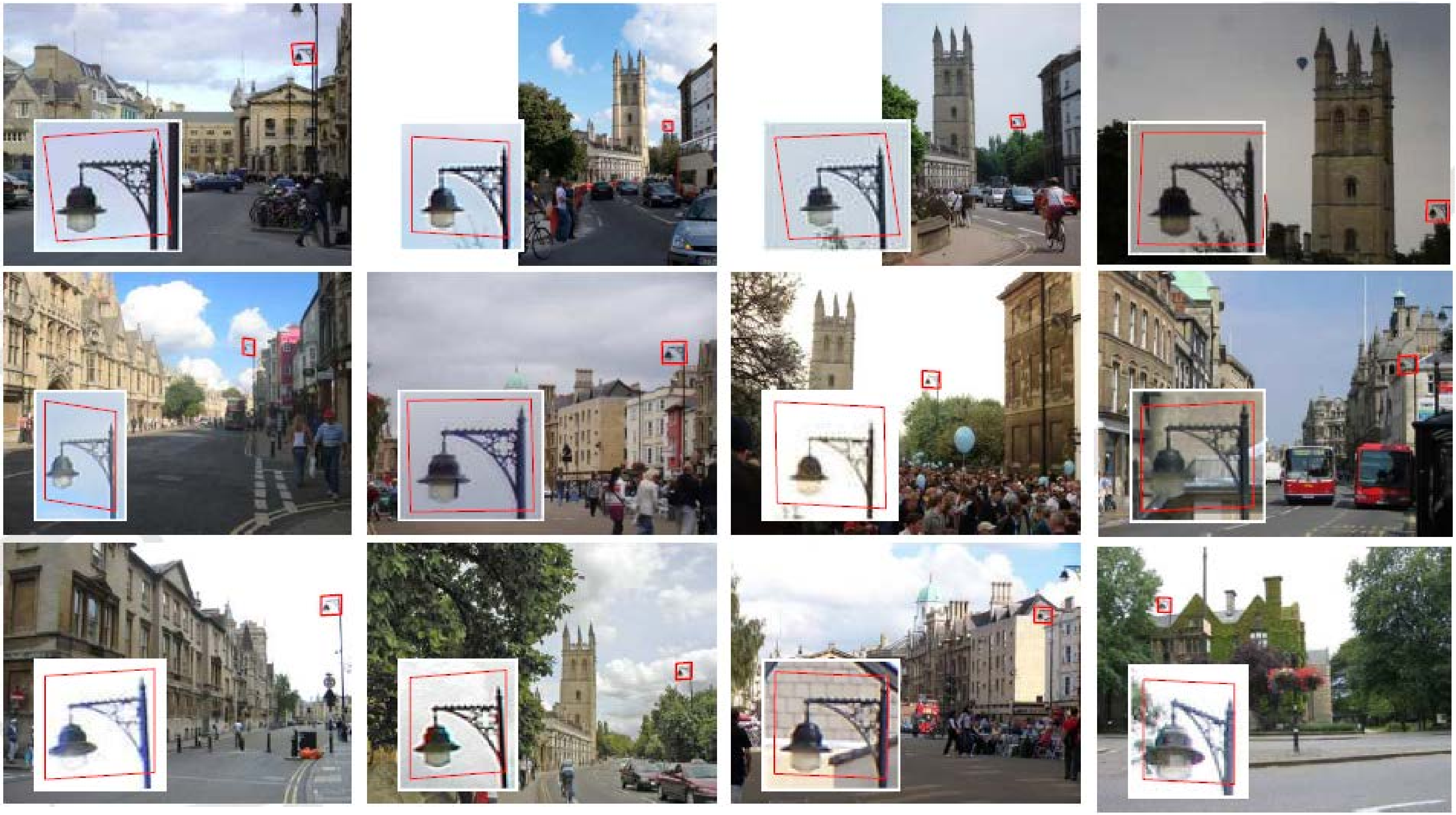
Geometric min-Hash
sketch collision
s = 2, k = 256

Verification by co-segmentation critical for small objects

[Cech, Matas, Perdoch CVPR 08], code available on WWW
[Ferrari, Tuytelaars,Van Gool, ECCV 2004]

Object Discovery

Other instances of the discovered object by (sub)image retrieval



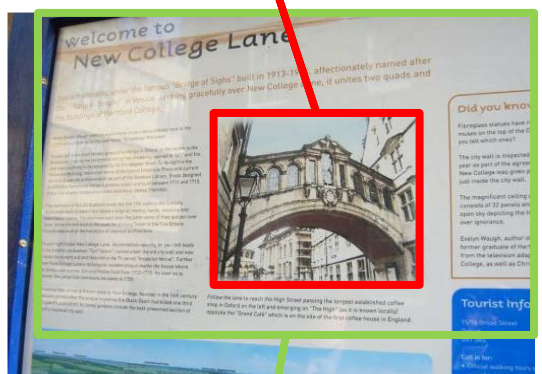
Unsupervised Discovery of Co-occurrence in Sparse High Dimensional Data

Over-counting

71 features

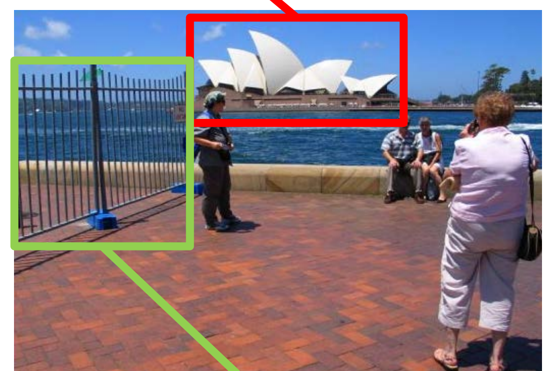
3700 features

929 features



1960 features

155 features



474 features

Chum and Matas:
Unsupervised Discovery of Co-occurrence in Sparse High Dimensional Data, CVPR 2010

Independence Assumption Violation

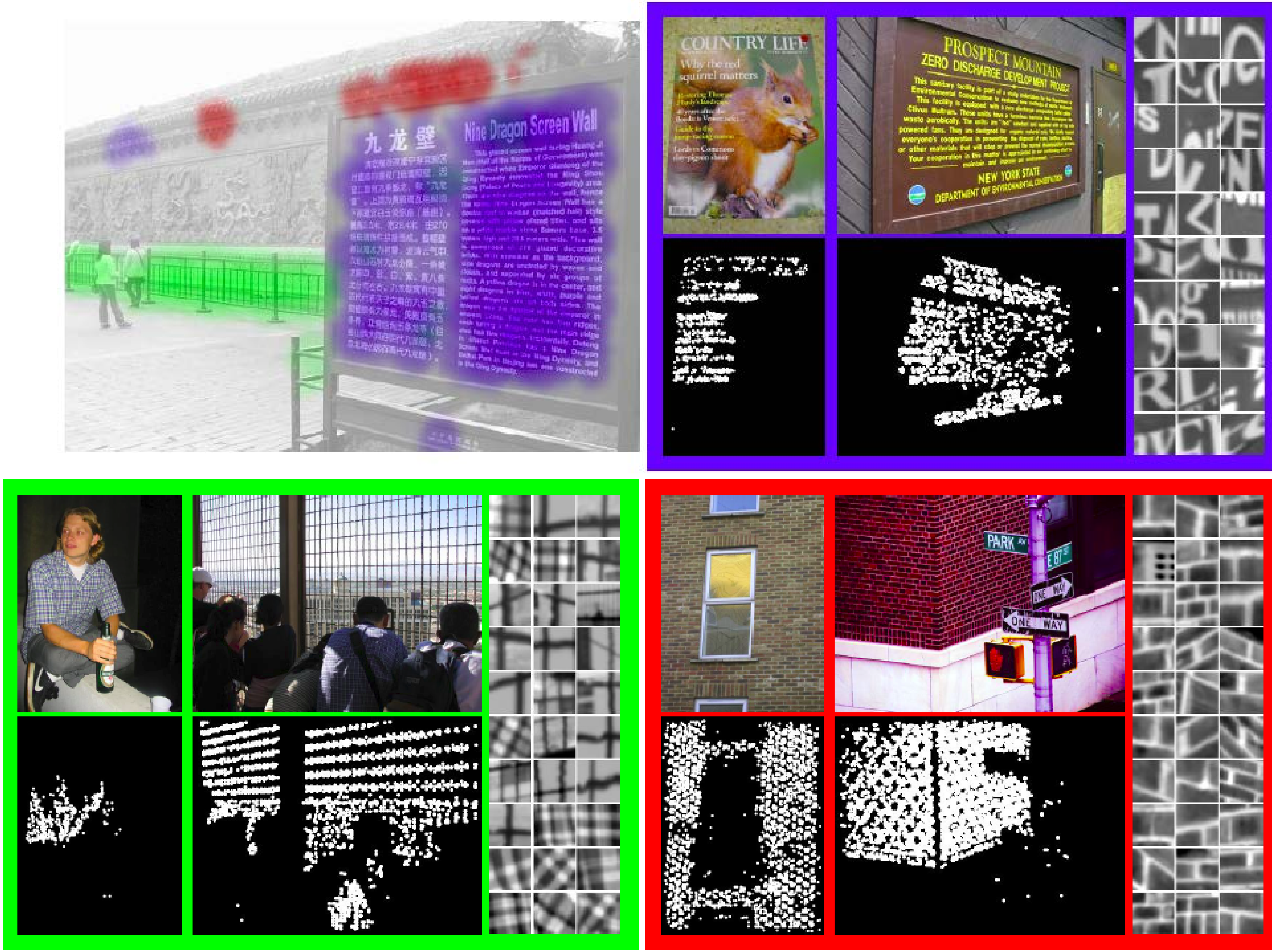
Query

Results (water)

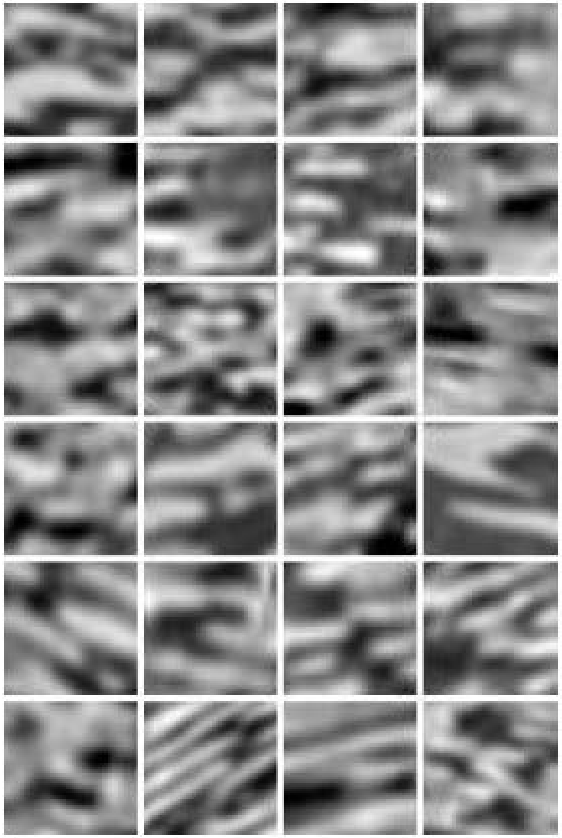
Results (Stockholm town hall)

- Over-counting of dependent observations
- Detect co-occurring visual words
  - Interchange the role of images and visual words
  - Use min-Hash to obtain sets of co-occurring visual words
- Down-weight / eliminate co-occurring features

Examples of Co-occurring Features



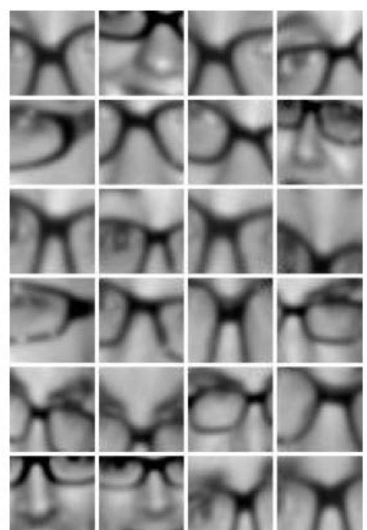
More Examples



More Examples





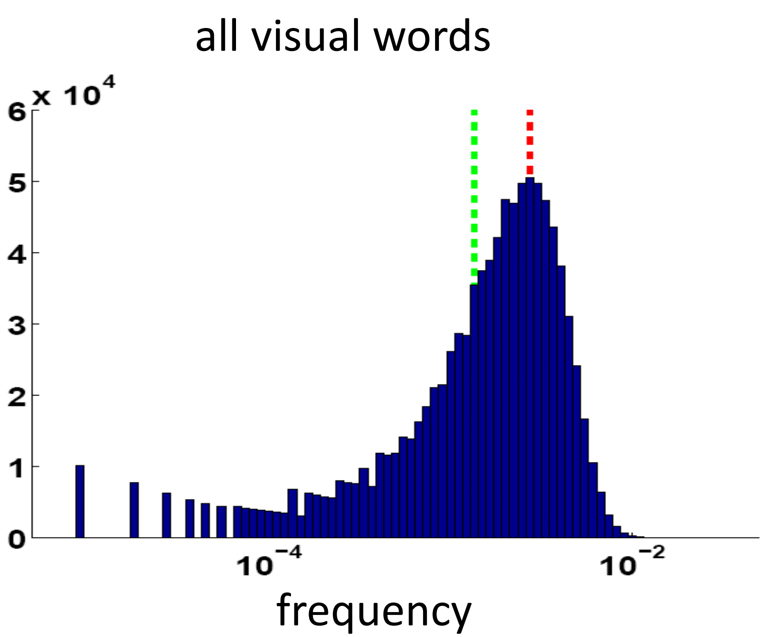



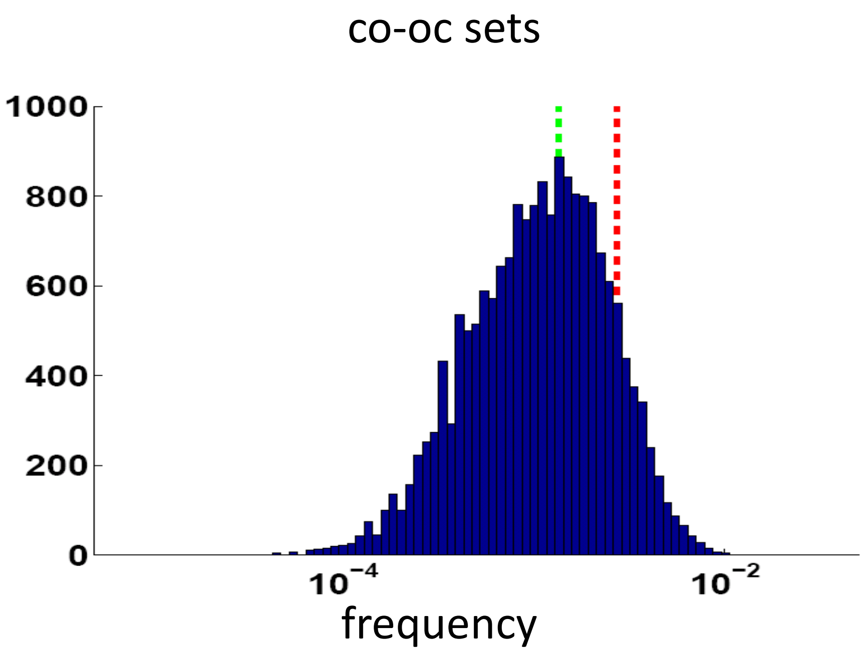
Flickr images

=

lots of faces

Visual Word Frequency





co-occurring visual words do not have to be frequent

greedy algorithms (such as a-Priori) fail

GEOMETRY IN IMAGE RETRIEVAL

Robust Estimation: Hough vs. RANSAC

Voting:

- discretized parameter space
- votes for parameters consistent with the measurements
- more votes higher support

\+ multiple models
\+ can be very fast
\- memory demanding
\- distances measured in the parameter space

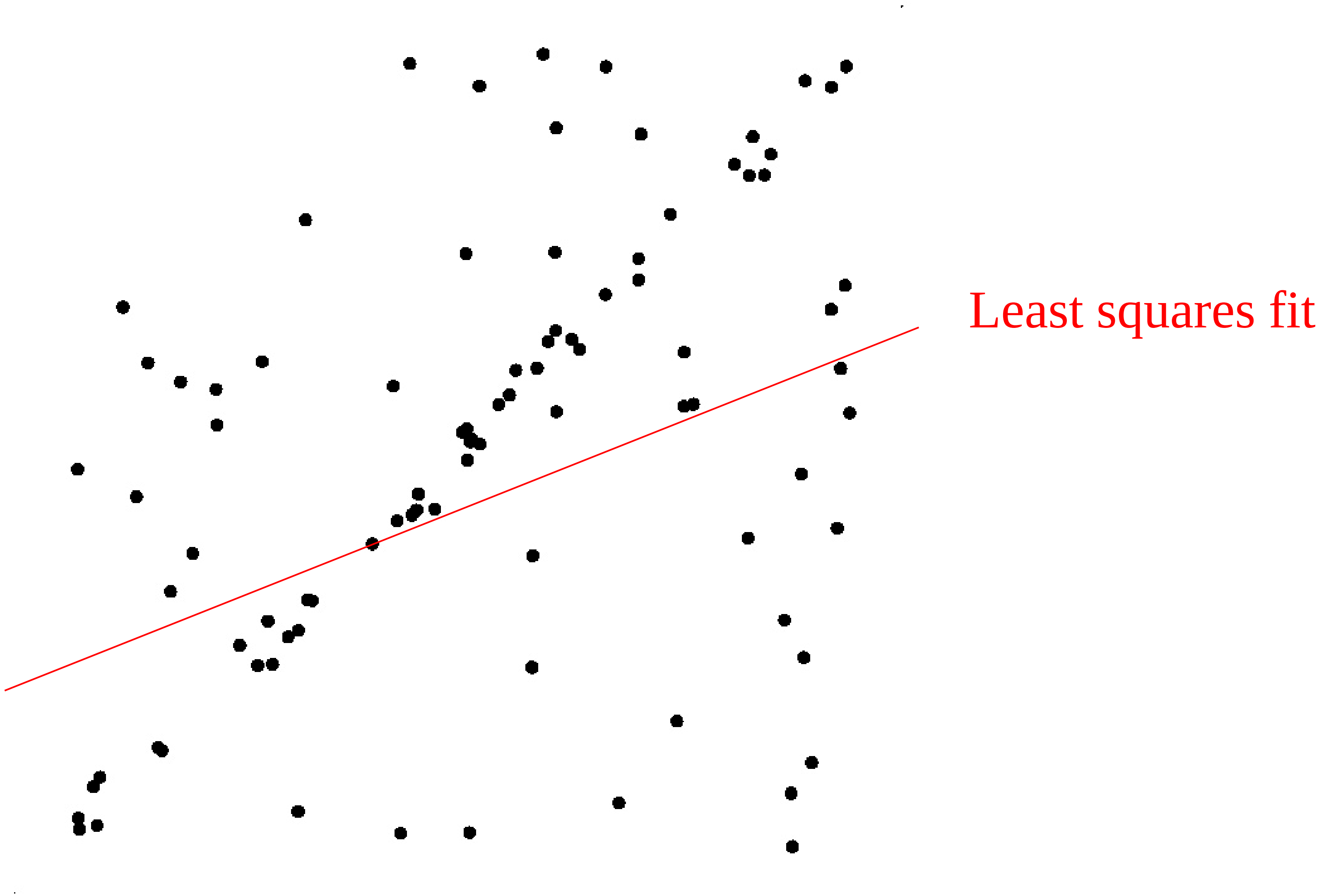
RANSAC:

- hypothesize and verify loop

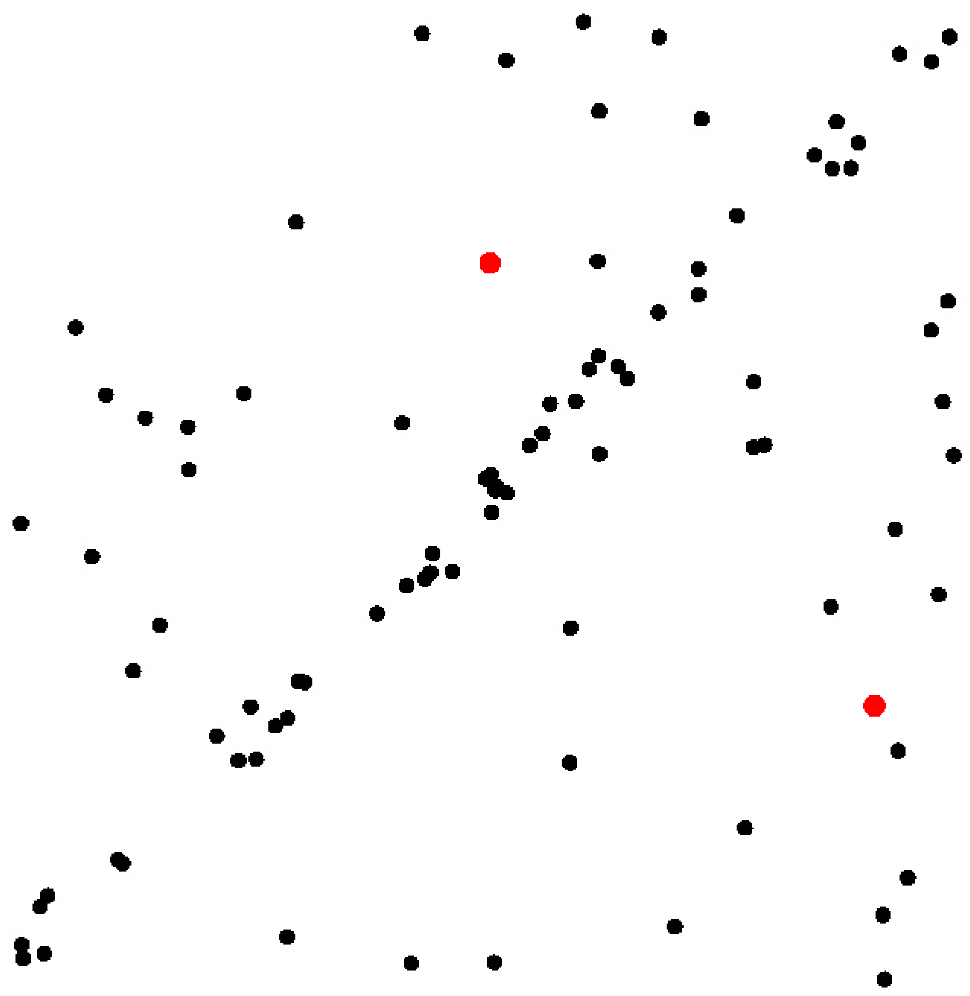
\- randomized (unless you try it all)
\- typically slower than voting
\+ no extra memory required
\+ measures distances in pixels!

RANSAC

Fitting a Line

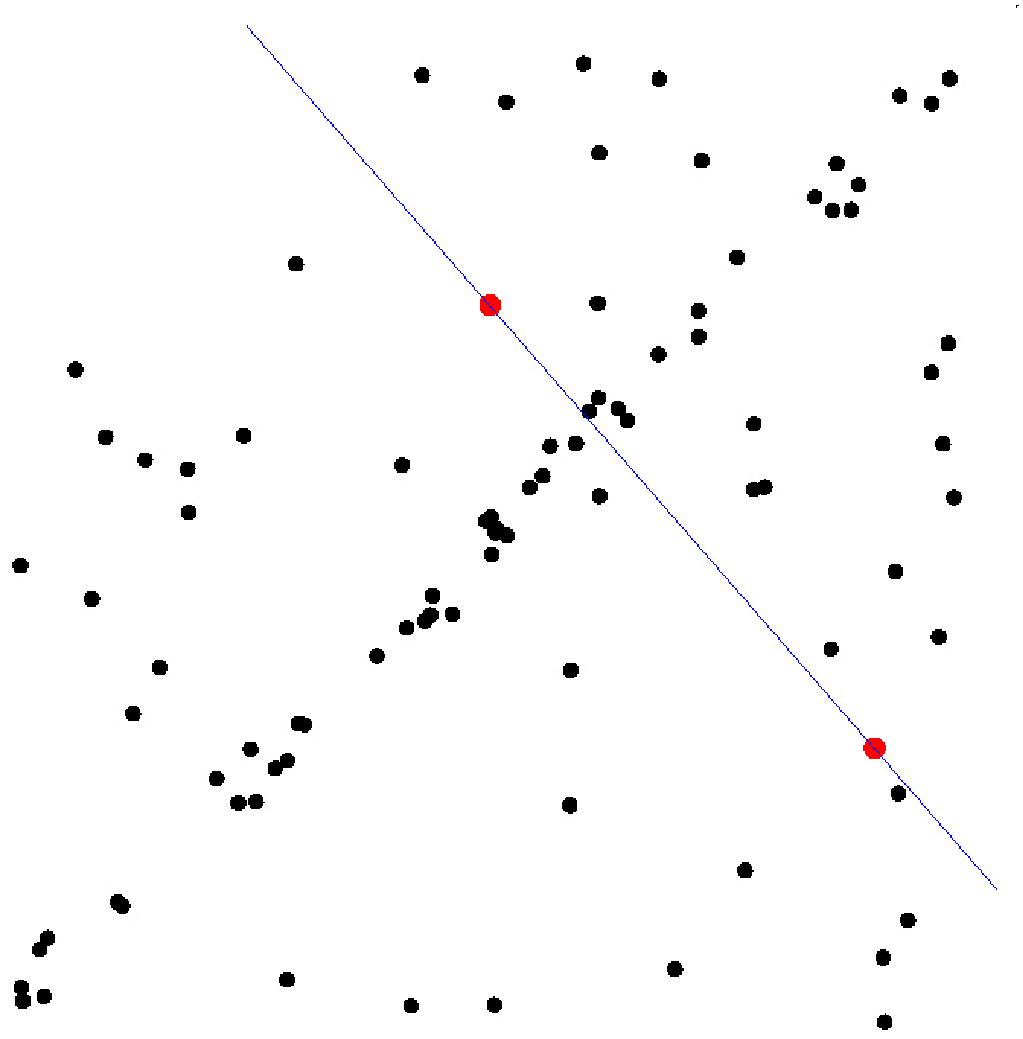


RANSAC



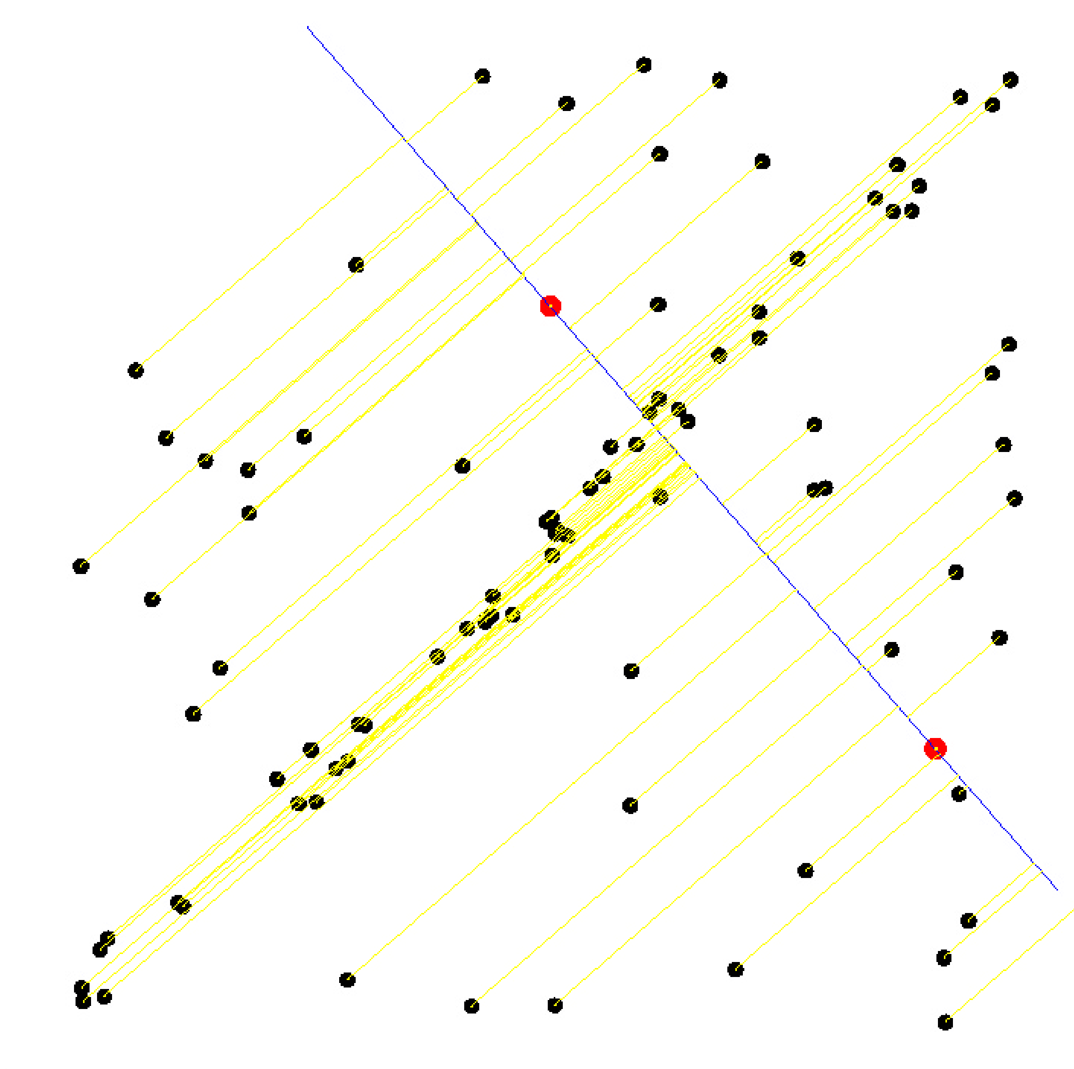
- **Select sample of m points at random**

RANSAC



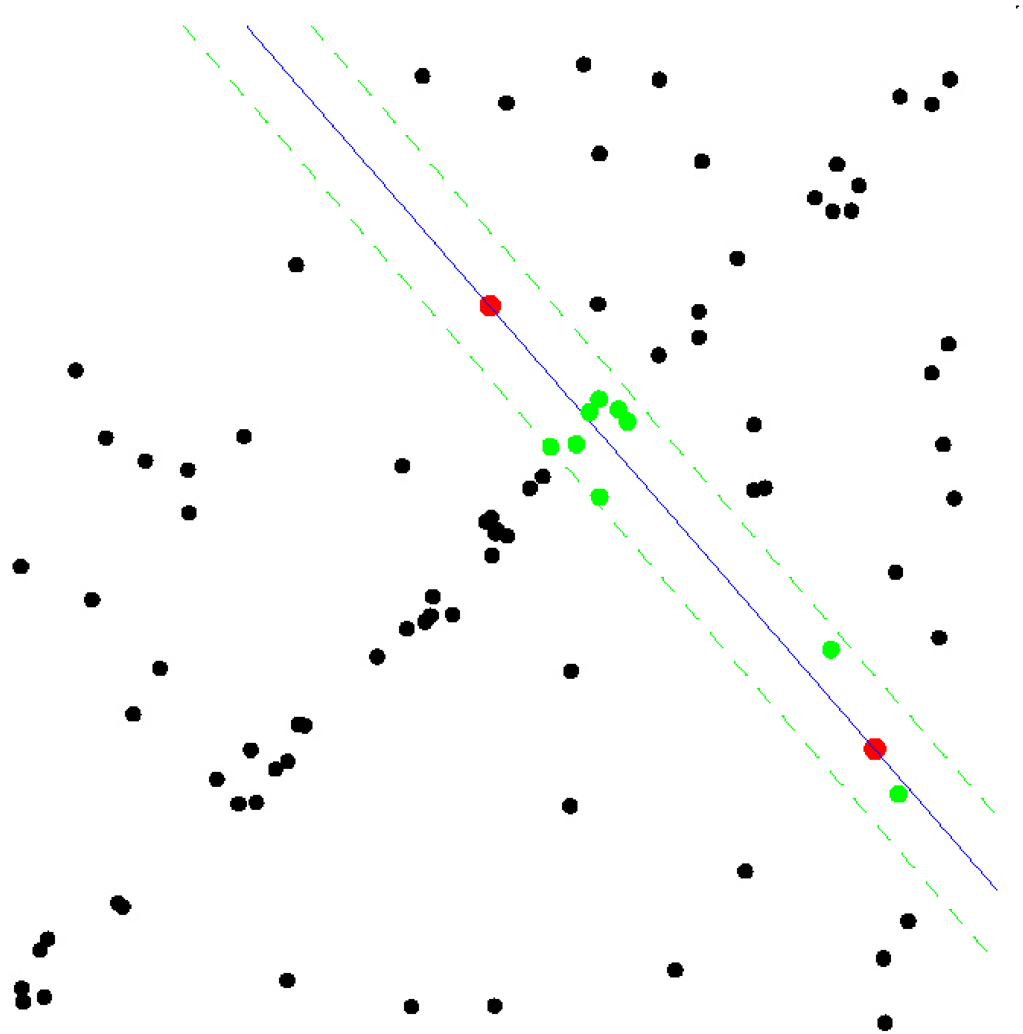
- Select sample of m points at random
- **Calculate model parameters that fit the data in the sample**

RANSAC



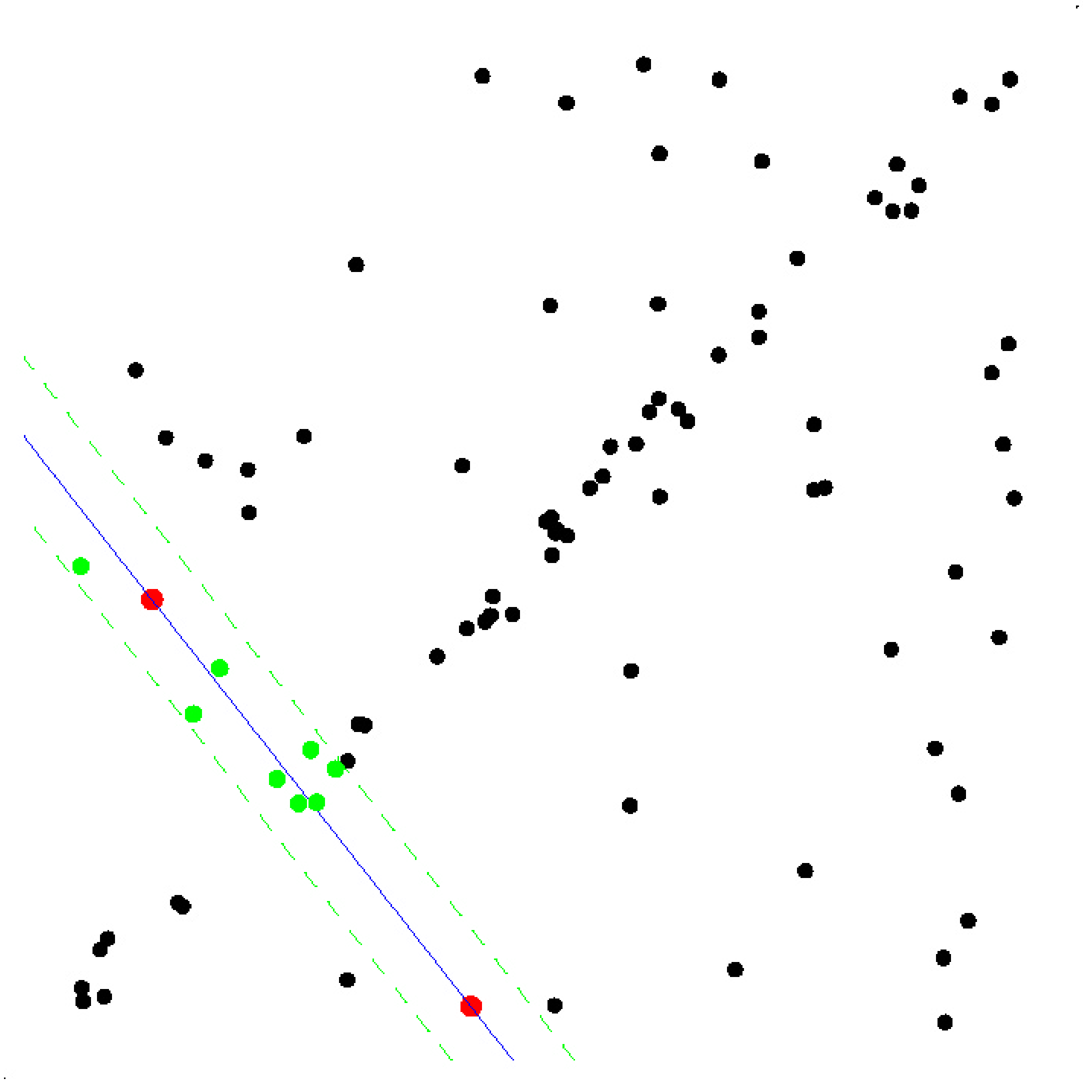
- Select sample of m points at random
- Calculate model parameters that fit the data in the sample
- **Calculate error function for each data point**

RANSAC



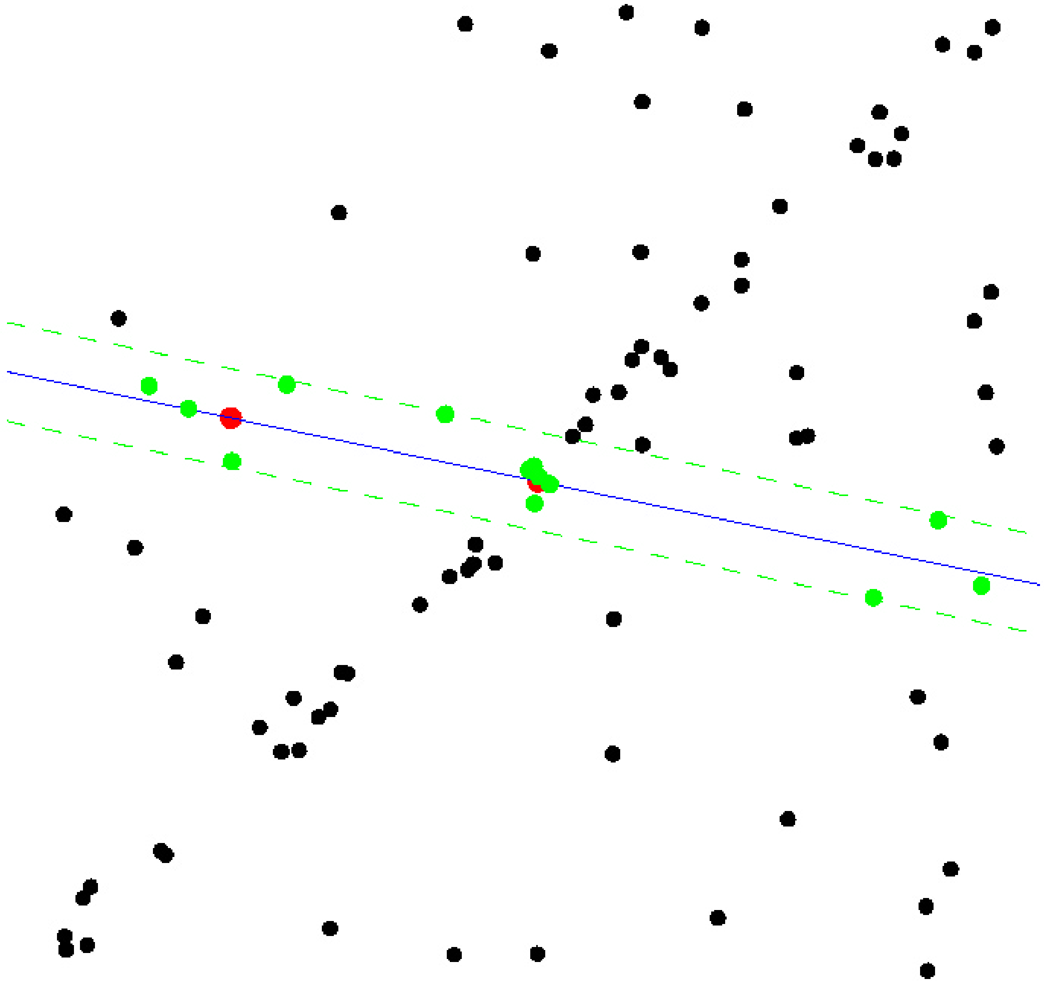
- Select sample of m points at random
- Calculate model parameters that fit the data in the sample
- Calculate error function for each data point
- **Select data that support current hypothesis**

RANSAC



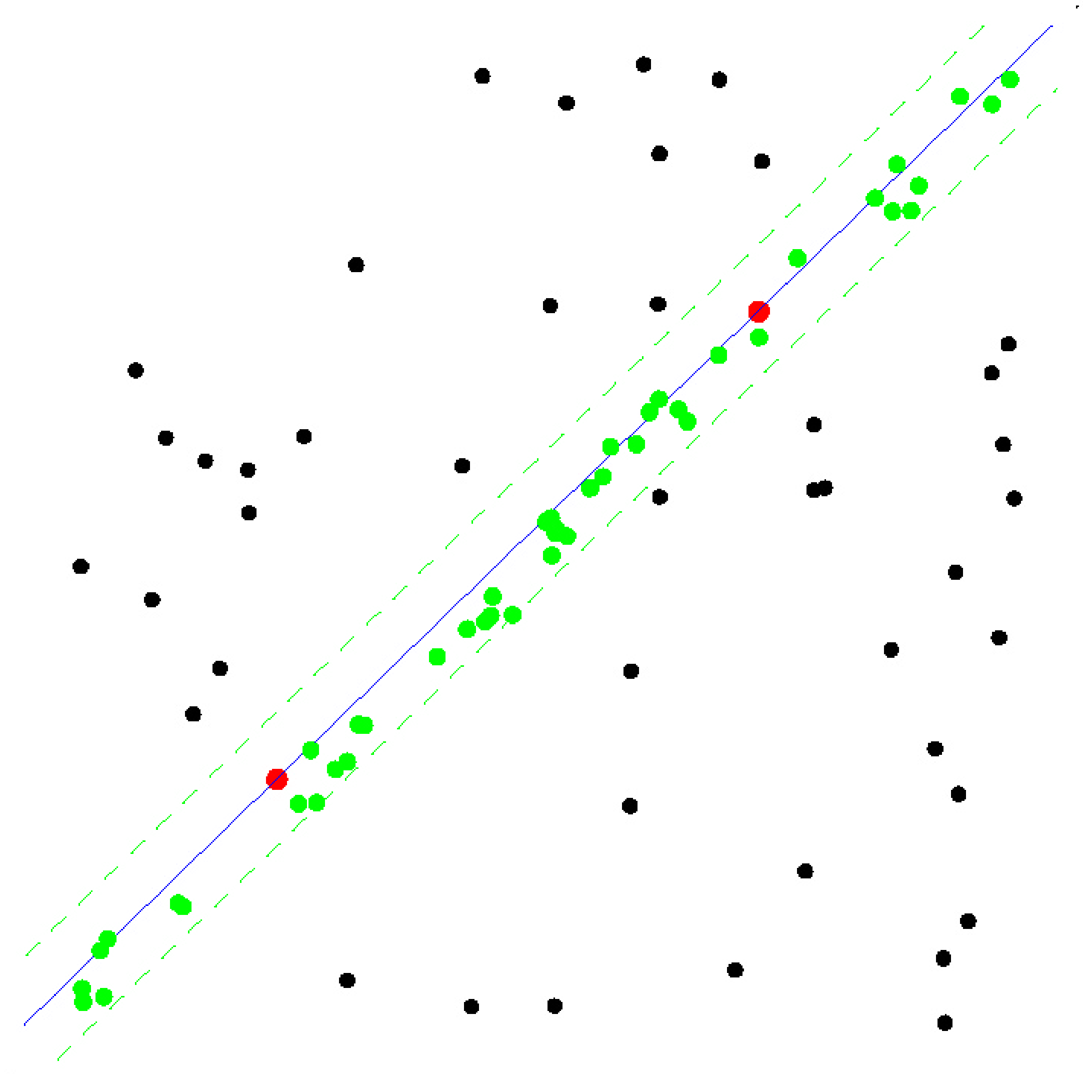
- Select sample of m points at random
- Calculate model parameters that fit the data in the sample
- Calculate error function for each data point
- Select data that support current hypothesis
- **Repeat sampling**

RANSAC



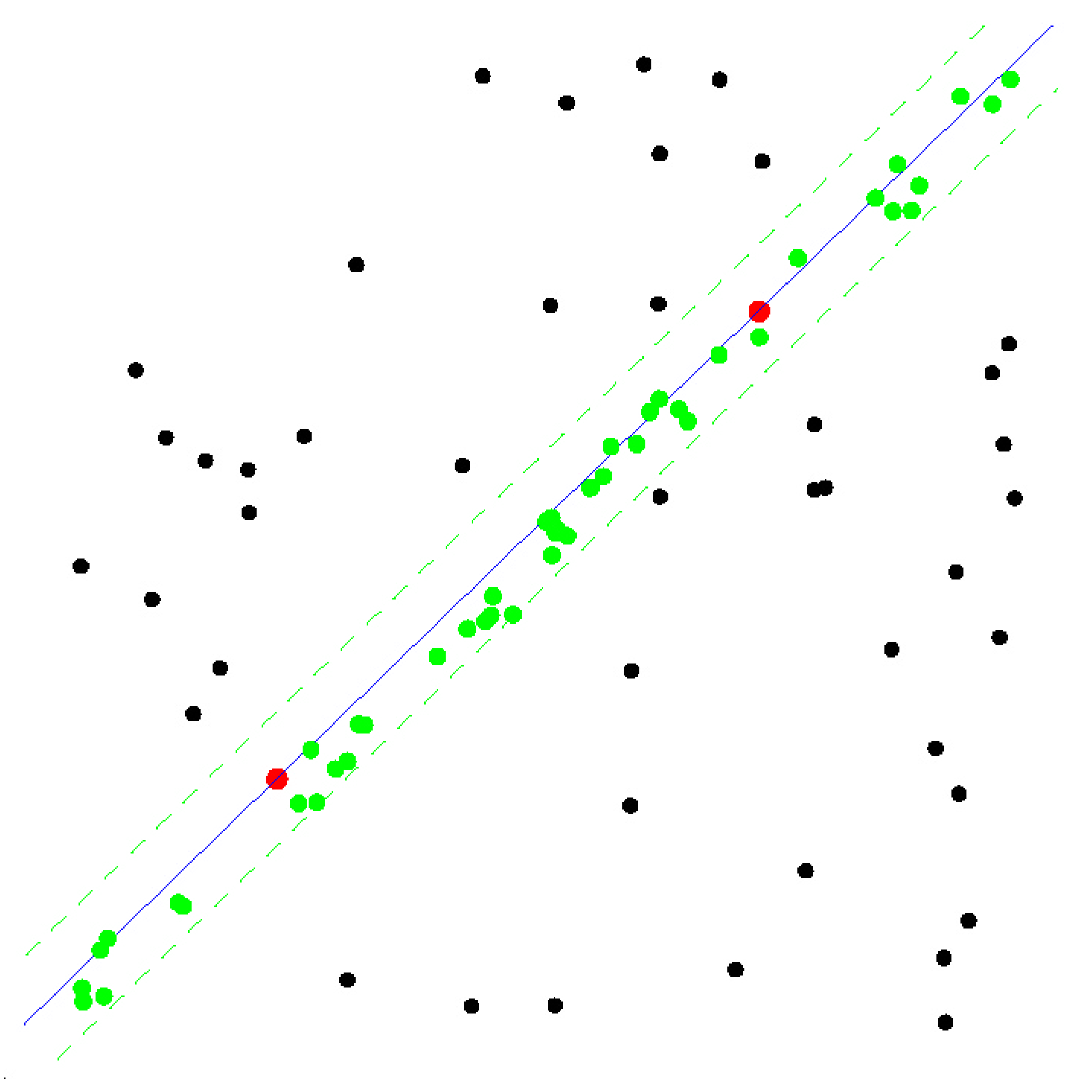
- Select sample of m points at random
- Calculate model parameters that fit the data in the sample
- Calculate error function for each data point
- Select data that support current hypothesis
- **Repeat sampling**

RANSAC



- Select sample of m points at random
- Calculate model parameters that fit the data in the sample
- Calculate error function for each data point
- Select data that support current hypothesis
- **Repeat sampling**

RANSAC



$$k = \frac{\log(1-p)}{\log\left(1-\frac{I^m}{N^m}\right)}$$

k … number of samples drawn

m … minimal sample size

N … number of data points

I … time to compute a single model

p … confidence in the solution (.95)

How Many Samples

I / N [%]

Size of the sample m

	15%	20%	30%	40%	50%	70%
2	132	73	32	17	10	4
4	5916	1871	368	116	46	11
7	$1.75 \cdot 10^6$	$2.34 \cdot 10^5$	$1.37 \cdot 10^4$	1827	382	35
8	$1.17 \cdot 10^7$	$1.17 \cdot 10^6$	$4.57 \cdot 10^4$	4570	765	50
12	$2.31 \cdot 10^{10}$	$7.31 \cdot 10^8$	$5.64 \cdot 10^6$	$1.79 \cdot 10^5$	$1.23 \cdot 10^4$	215
18	$2.08 \cdot 10^{15}$	$1.14 \cdot 10^{13}$	$7.73 \cdot 10^9$	$4.36 \cdot 10^7$	$7.85 \cdot 10^5$	1838
30	∞	∞	$1.35 \cdot 10^{16}$	$2.60 \cdot 10^{12}$	$3.22 \cdot 10^9$	$1.33 \cdot 10^5$
40	∞	∞	∞	$2.70 \cdot 10^{16}$	$3.29 \cdot 10^{12}$	$4.71 \cdot 10^6$

RANSAC [Fischler, Bolles '81]

In: $U = \{x_i\}$ set of data points, $|U| = N$

$f(S) : S \rightarrow p$ function f computes model parameters p given a sample S from U

$\rho(p, x)$ the cost function for a single data point x

Out: p^* p^*, parameters of the model maximizing the cost function

k := 0

Repeat until P{better solution exists} < η (a function of C^* and no. of steps k)

k := k + 1

I. Hypothesis

(1) select randomly set $S_k \subset U$, sample size $|S_k| = m$

(2) compute parameters $p_k = f(S_k)$

II. Verification

(3) compute cost $C_k = \sum_{x \in U} \rho(p_k, x)$

(4) if $C^* < C_k$ then $C^* := C_k$, $p^* := p_k$

end

Advanced RANSAC

In: $U = \{x_i\}$ set of data points, $|U| = N$

$f(S): S \rightarrow p$ function f computes model parameters p given a sample S from U

$\rho(p, x)$ the cost function for a single data point x

Out: p^* p^*, parameters of the model maximizing the cost function

Preemptive scoring

$k := 0$

Repeat until P{better solution exists} < η (a function of C^* and no. of steps k)

$k := k + 1$

Non-uniform sampling

I. Hypothesis

(1) select randomly set $S_k \subset U$, sample size $|S_k| = m$

(2) compute parameters $p_k = f(S_k)$

Error scale estimation

II. Verification

(3) compute cost $C_k = \sum_{x \in U} \rho(p_k, x)$

Randomized verification

(4) if $C^* < C_k$ then $C^* := C_k$, $p^* := p_k$

Improving precision

end

Potential degeneracy tests

*SAC

RANSAC [Fischler'81], **MLESAC** [Torr'00], **R-RANSAC** [Chum'02], **NAPSAC** [Myatt'02], **Guided MLESAC** [Tordoff'02], **LO-RANSAC** [Chum'03], **Preemtive RANSAC** [Nister'03], **PROSAC** [Chum'05], **RANSAC with bail-out** [Capel'05], **DegenSAC** [Chum'05], **WaldSAC** [Matas'05], **QDEGSAC** [Frahm'06], **GASAC** [Rodehorst'06], **ARRSAC** [Raguram'08] **GroupSAC** [Ni'09], **Cov-RANSAC** [Raguram'09], **...**

Lebeda, Matas, and Chum: **Fixing the Locally Optimized RANSAC**, BMVC 2012

images, data, executables:
http://cmp.felk.cvut.cz/software/LO-RANSAC/index.xhtml

Raguram, Chum, Pollefeys, Matas, Frahm:
"**USAC: A Universal Framework for Random Sample Consensus**", PAMI 2013

code, data:
http://cs.unc.edu/~rraguram/usac/

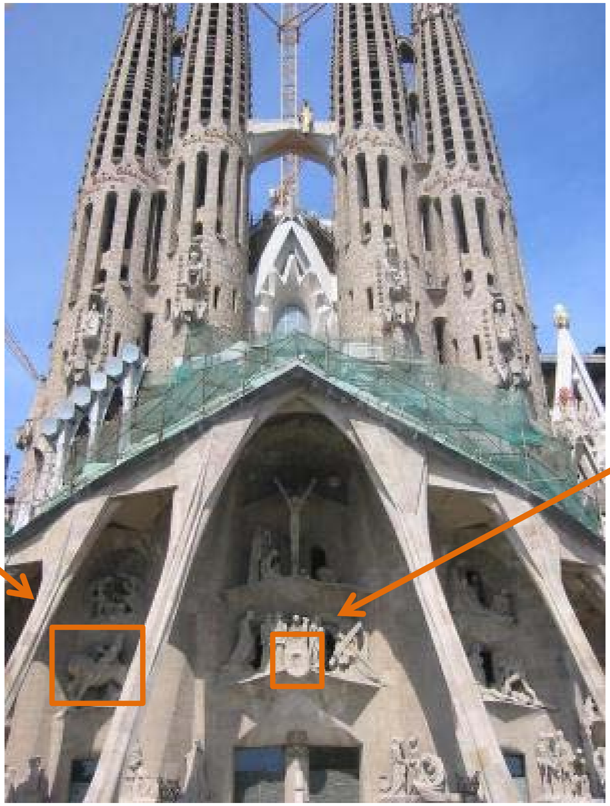
BEYOND VISUAL NEAREST NEIGHBOR SEARCH

RETRIEVAL WITH (GEOMETRIC) CONSTRAINTS

Retrieval for Browsing

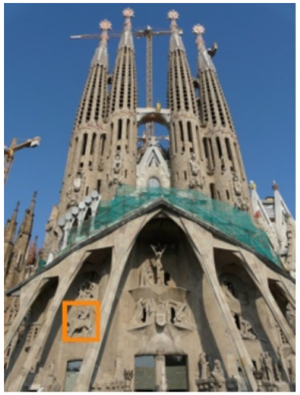
What is this?

... and what is that?

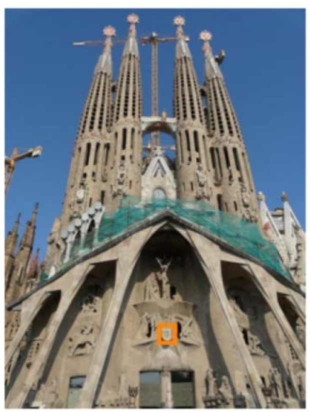


Let's query!

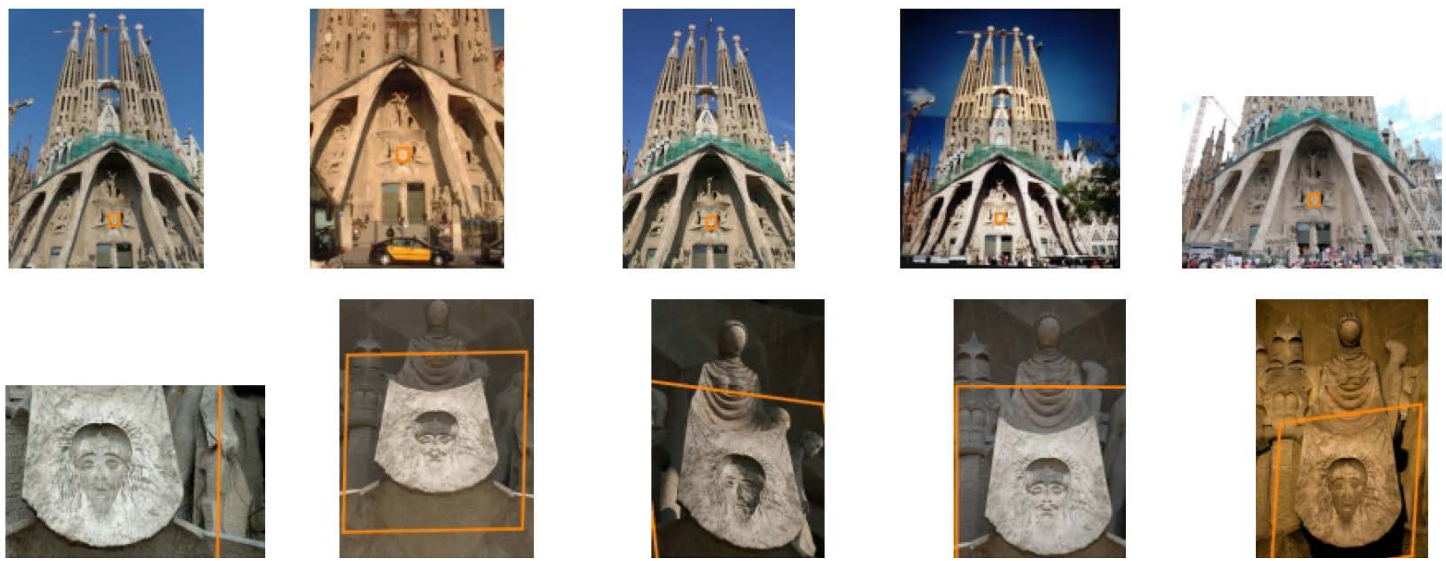
Retrieval for Browsing



Query 1



Query 2



Mikulik, Chum, Matas: Image Retrieval for Online Browsing in Large Image Collections, SISAP 2013.

New Problem Formulation

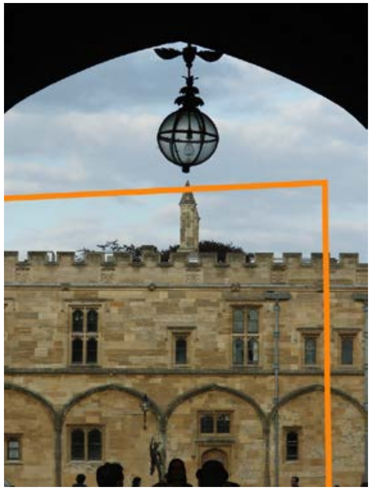
Retrieve relevant images subject to a constraint

- **Geometric**
  - Maximize number of relevant pixels
  - Maximize scale change
  - Change of viewpoint
- Other
  - High photometric change (day / night)

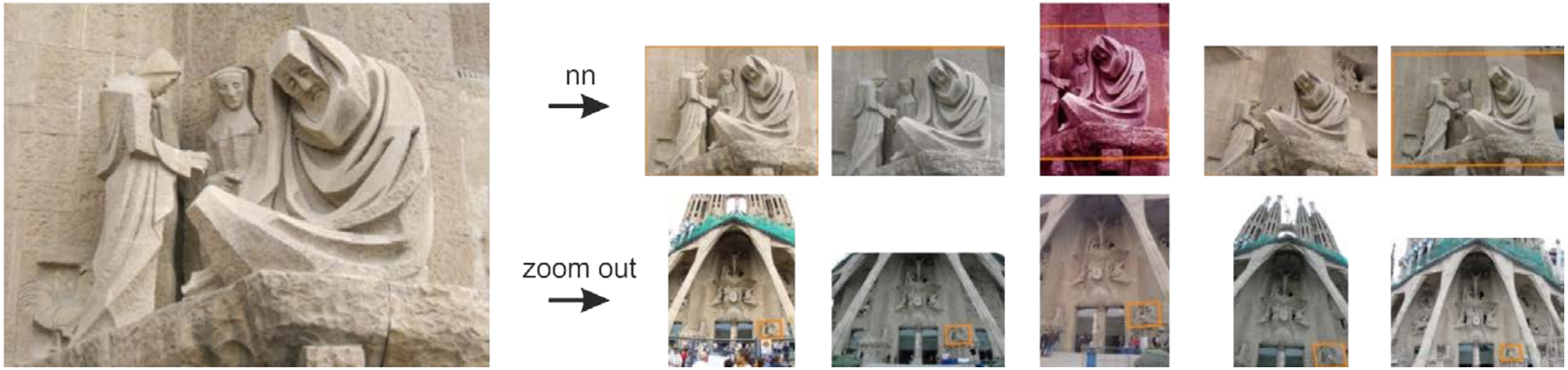
New Problem Formulation

Results

- Low rank in standard similarity measure
  - Geometry for verification and constraint enforcement
  - Geometry in the inverted file (DAAT)
- Standard similarity measure can be 0
  - Matching through a path of images (query expansion)



“Where is this” example



Query Image

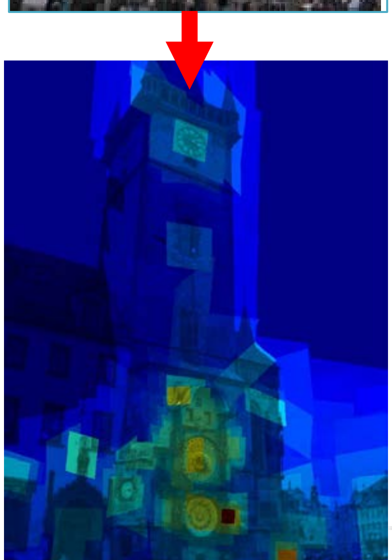
What is interesting here?

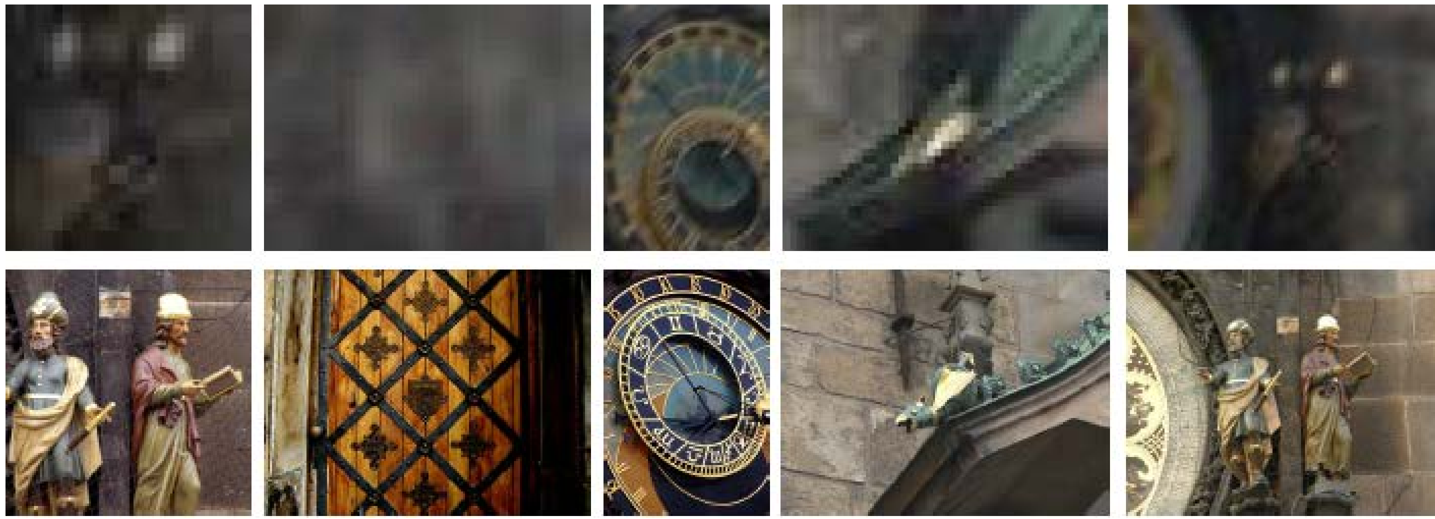
All Details on the Landmark

Mikulik, Radenovic, Chum, and Matas : Efficient Image Detail Mining, ACCV 2014

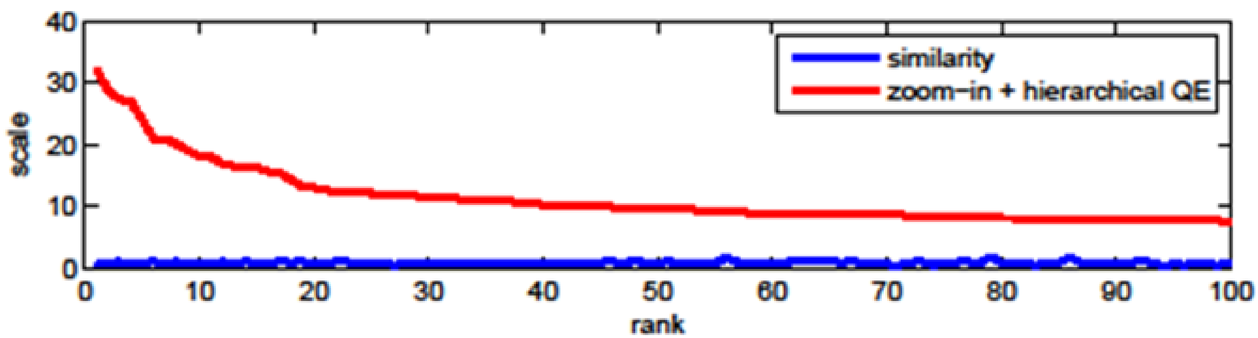
Highest Resolution Transform

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





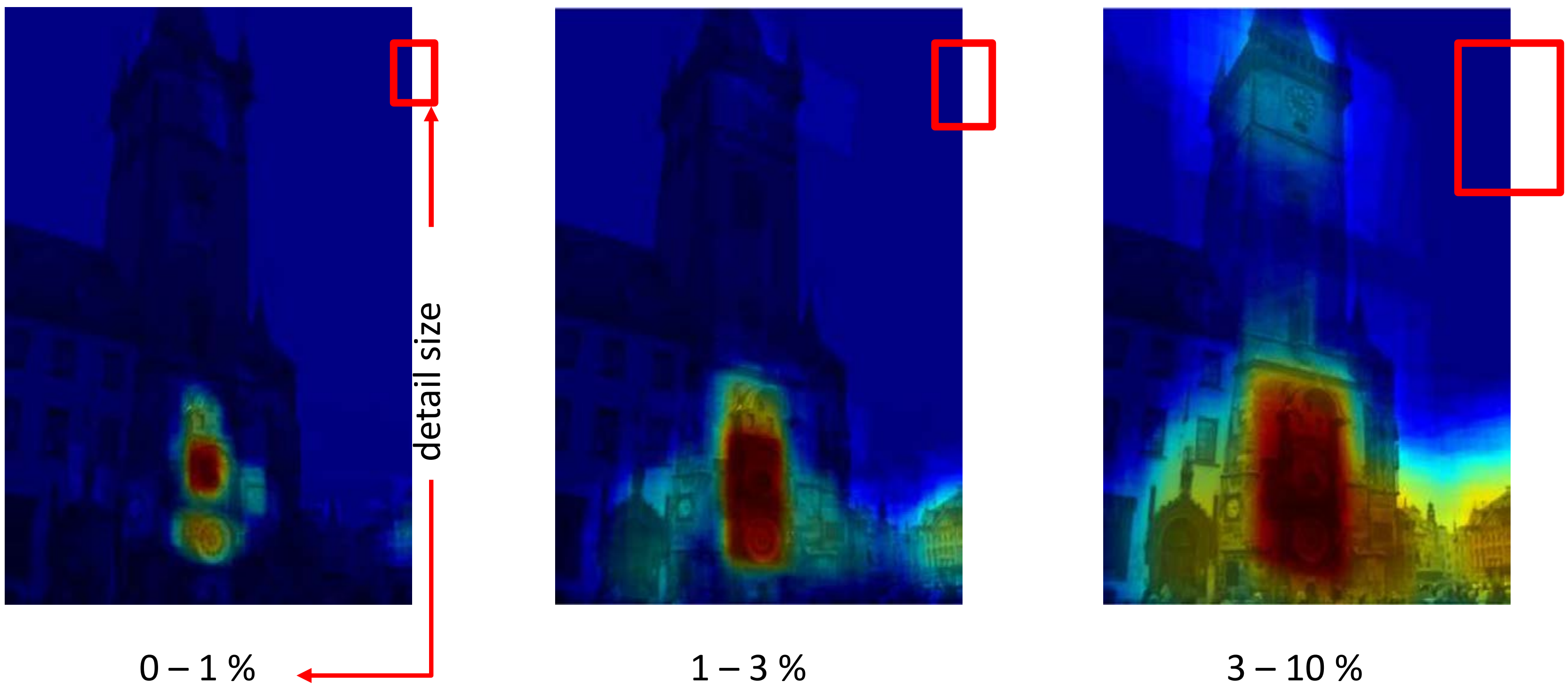
37.3x 27.0x 22.8x 21.9x 21.6x

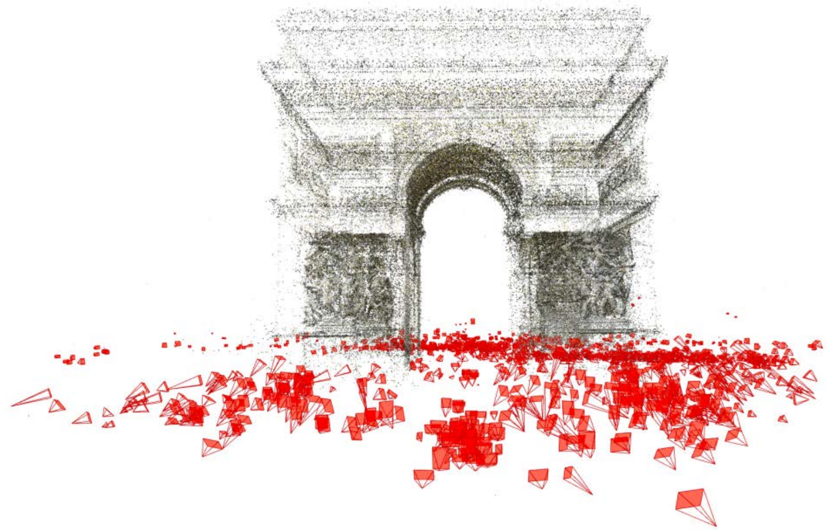


Highest Details

Level of Interest Transform

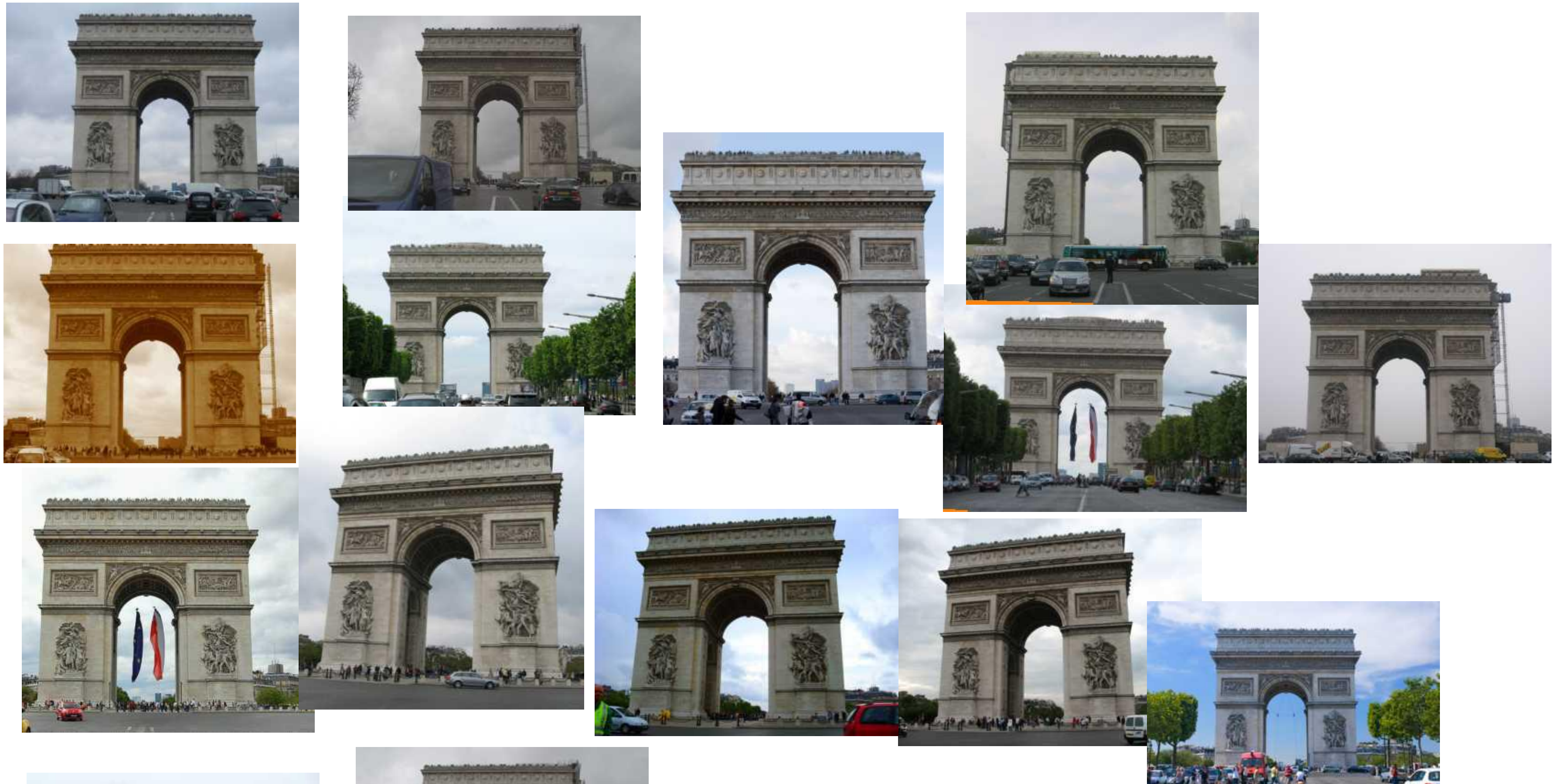
Given a query and a dataset, for every pixel in the query image:
Find the frequency with which it is photographed in detail

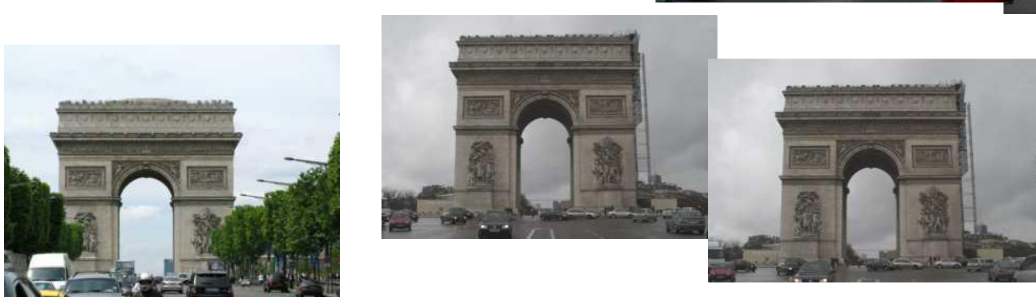




FROM SINGLE IMAGE QUERY TO DETAILED 3D RECONSTRUCTION

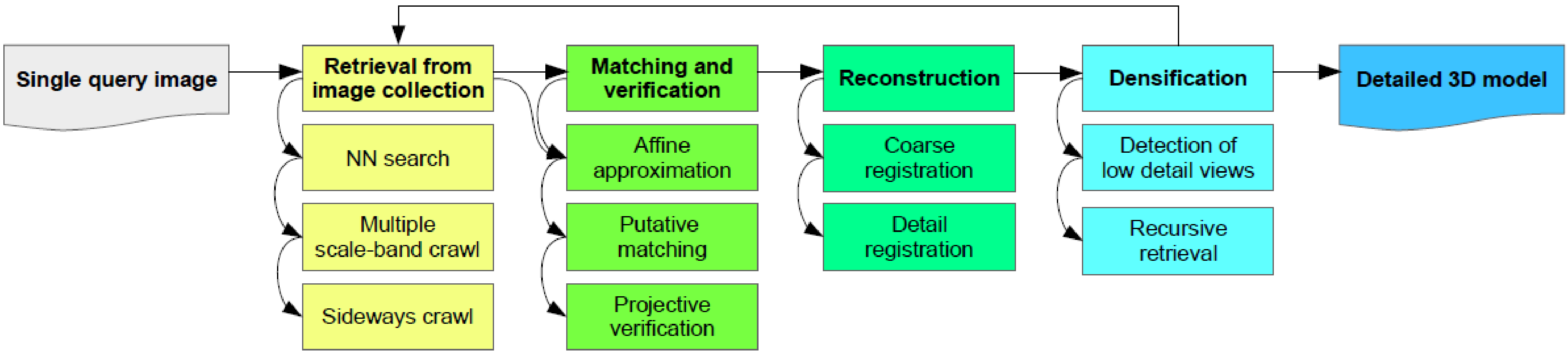
Retrieval and SfM



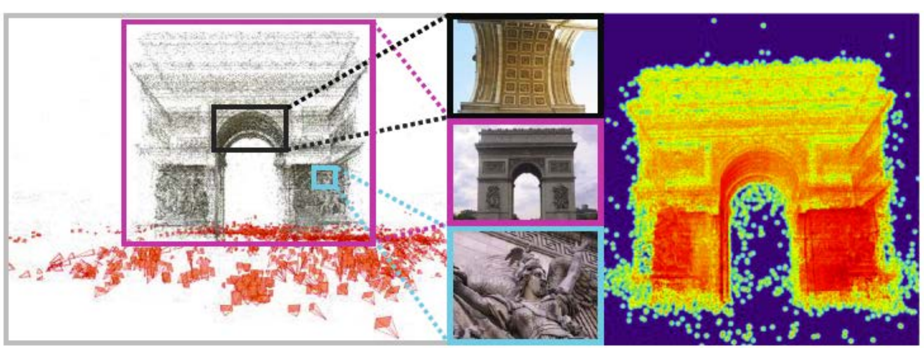


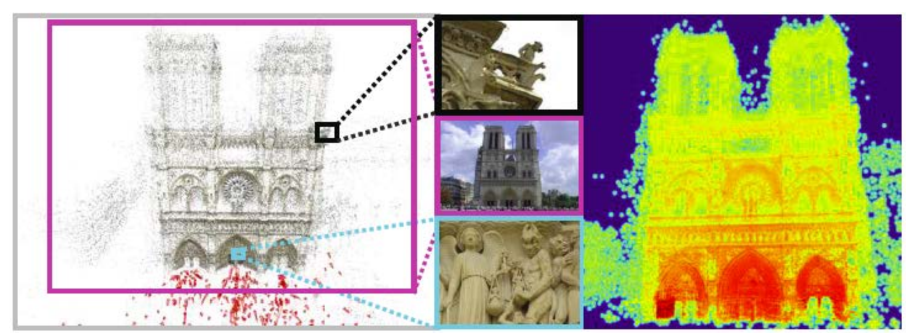
k-NN search often find small connected components

Tight Coupling of Retrieval and SfM



Schoenberger, Radenovic, Chum, and Frahm:
From Single Image Query to Detailed 3D Reconstruction , CVPR'15





Beyond Nearest Neighbour

- Zoom out – getting a context of the image
- All details – getting transition to the object details
- Sidewise crawl

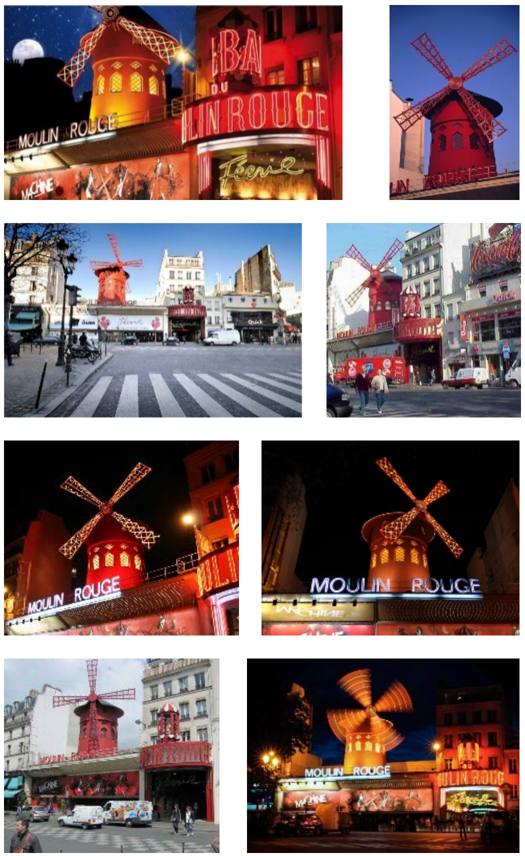
Looking around the corner

Some Results …

FROM DUSK TILL DOWN MODELLING IN THE DARK

Separate Day & Night Dense Reconstructions

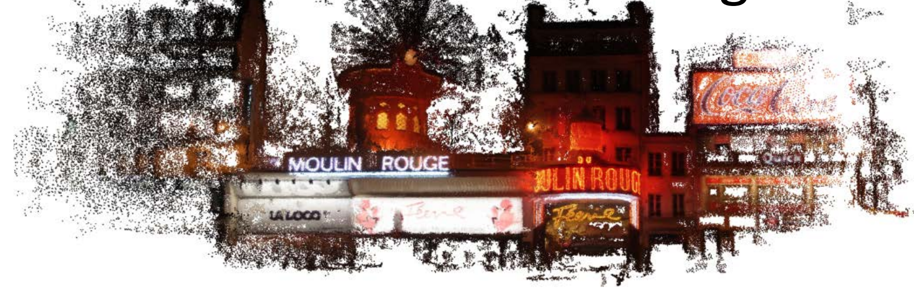
Day & Night Images



Standard Dense

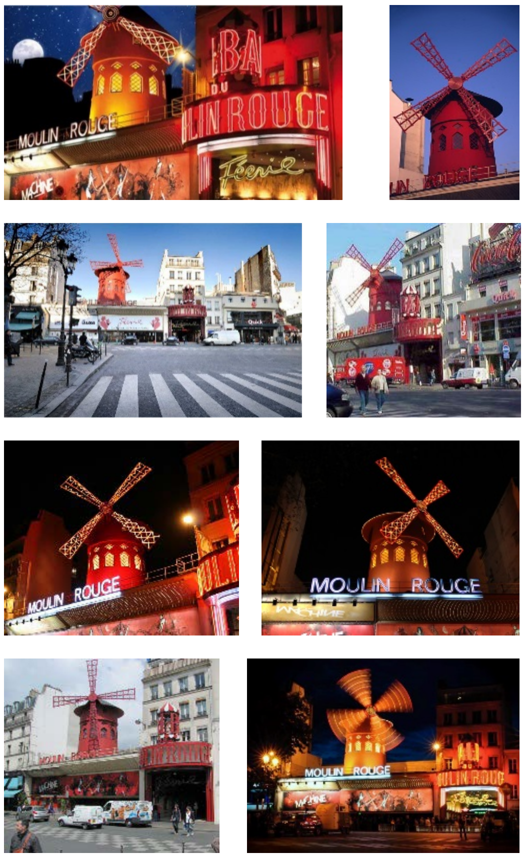
Day Dense

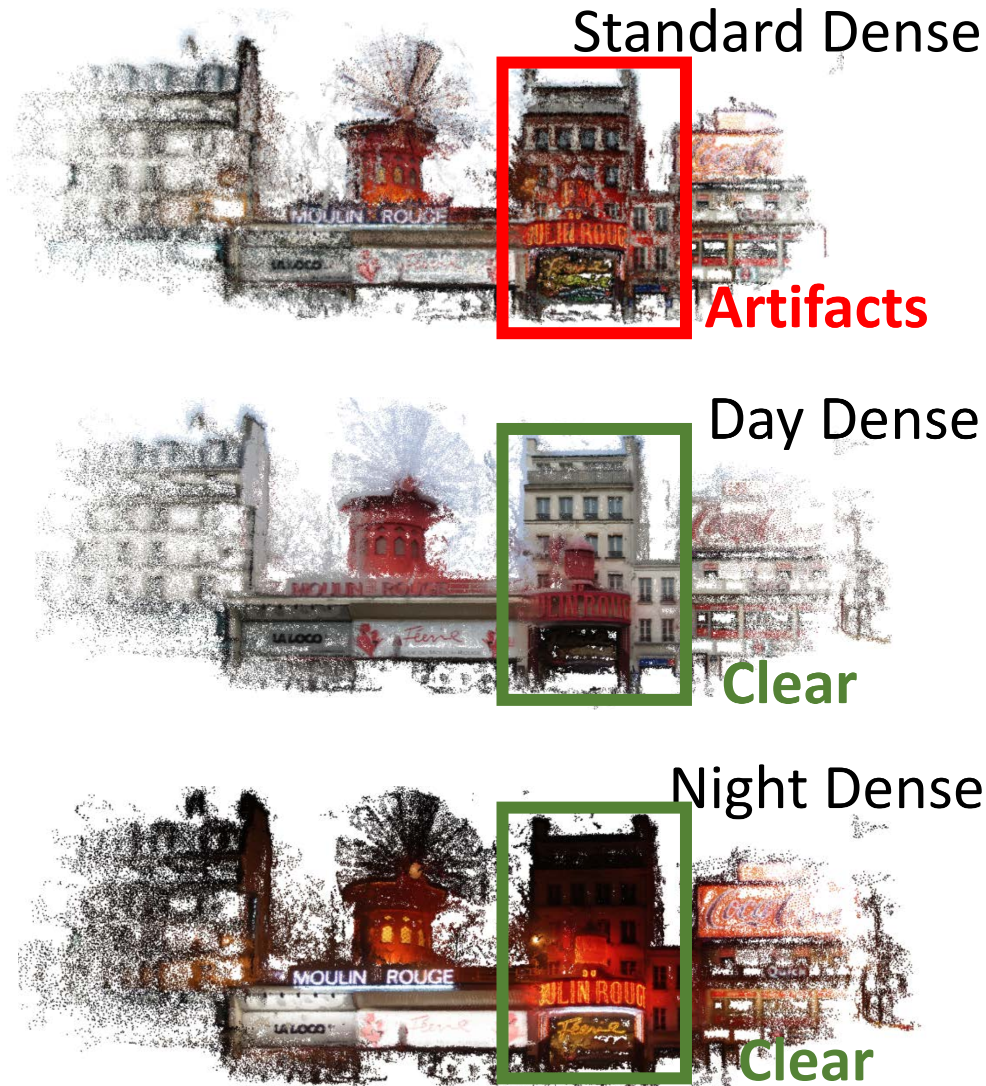
Night Dense



Separate Day & Night Dense Reconstructions

Day & Night Images



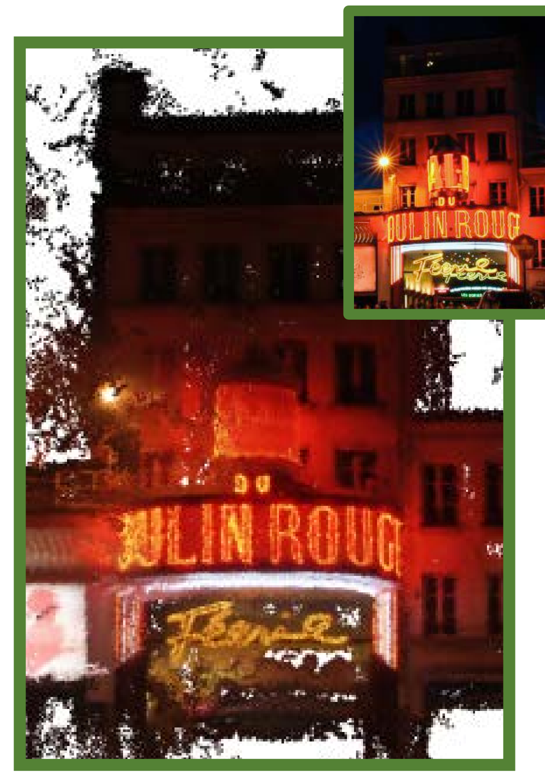


Separate Day & Night Dense Reconstructions

Standard Dense

Day Dense

Night Dense



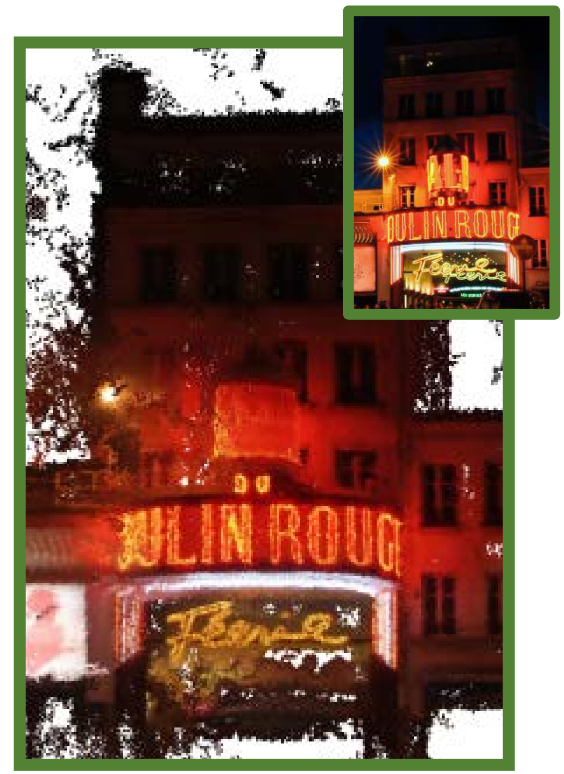
Separate Day & Night Dense Reconstructions

Standard Dense

Artifacts

Day Dense

Night Dense

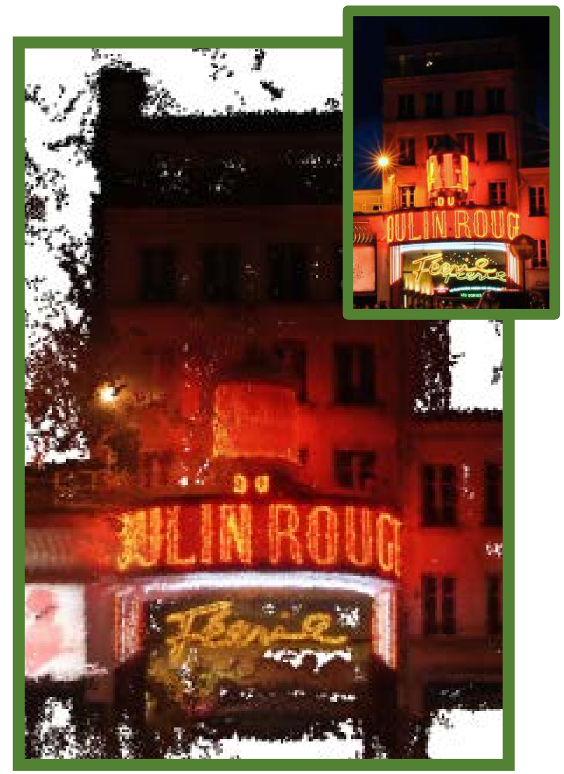


Separate Day & Night Dense Reconstructions

Standard Dense

Day Dense

Night Dense



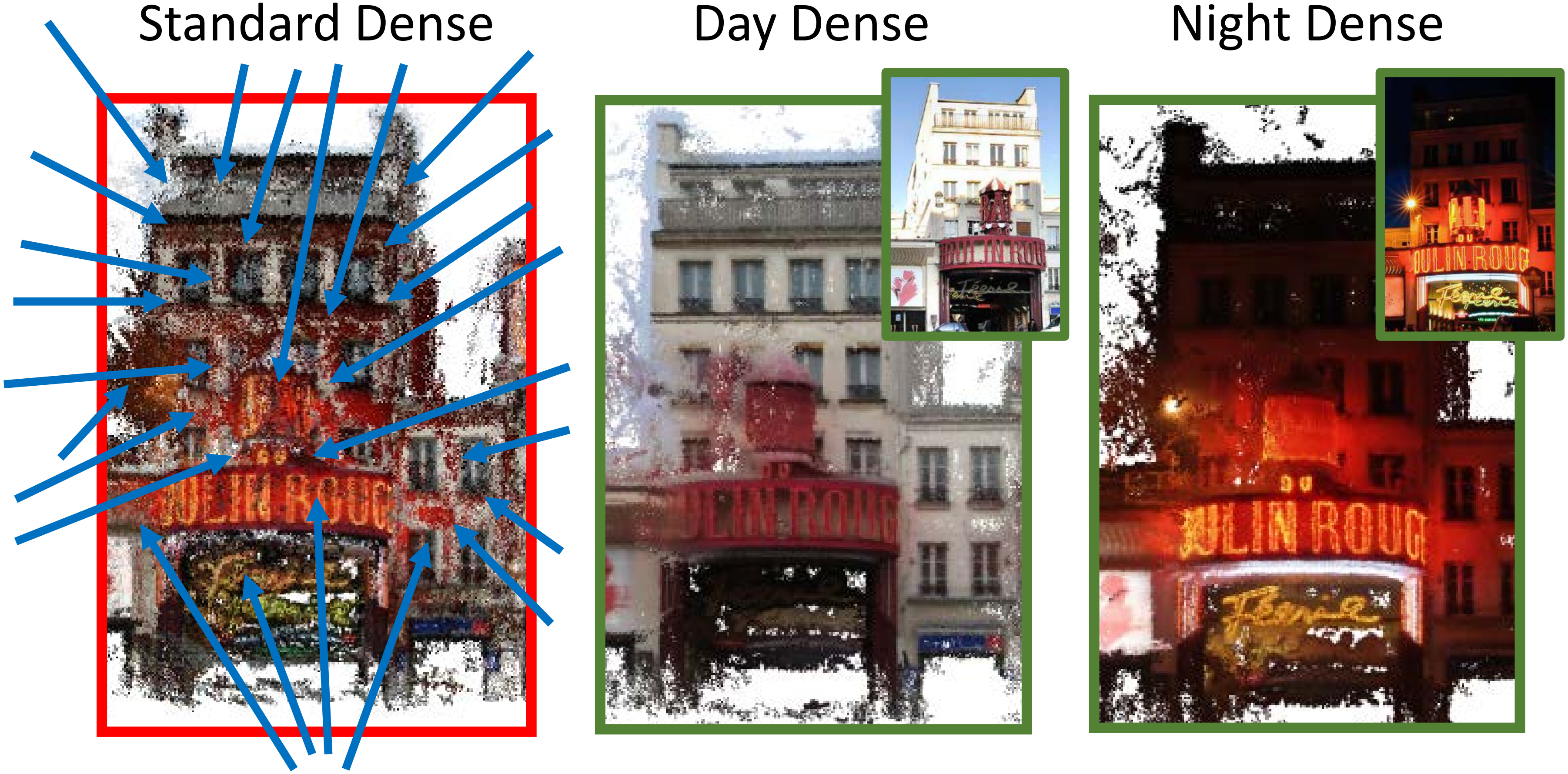
Artifacts

Separate Day & Night Dense Reconstructions

Standard Dense

Day Dense

Night Dense



Artifacts

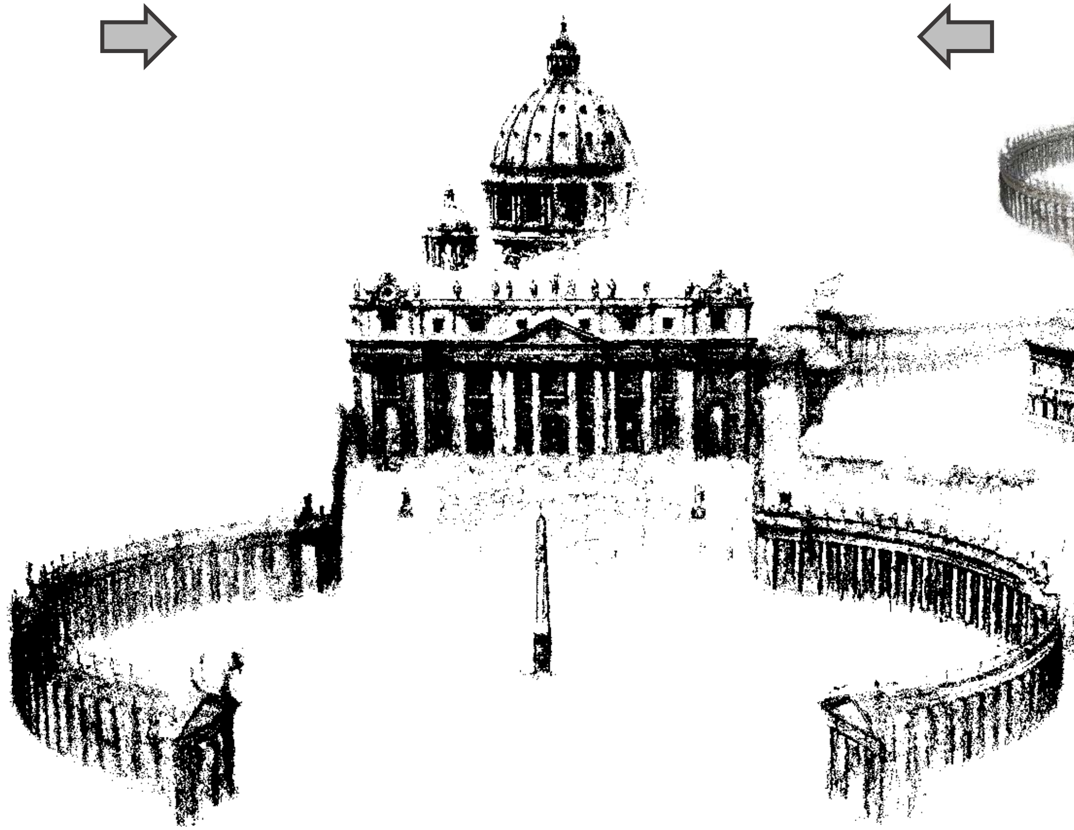
Day & Night Dense Models

Night Model

Day Model

Geometric Fusion of Day & Night Models

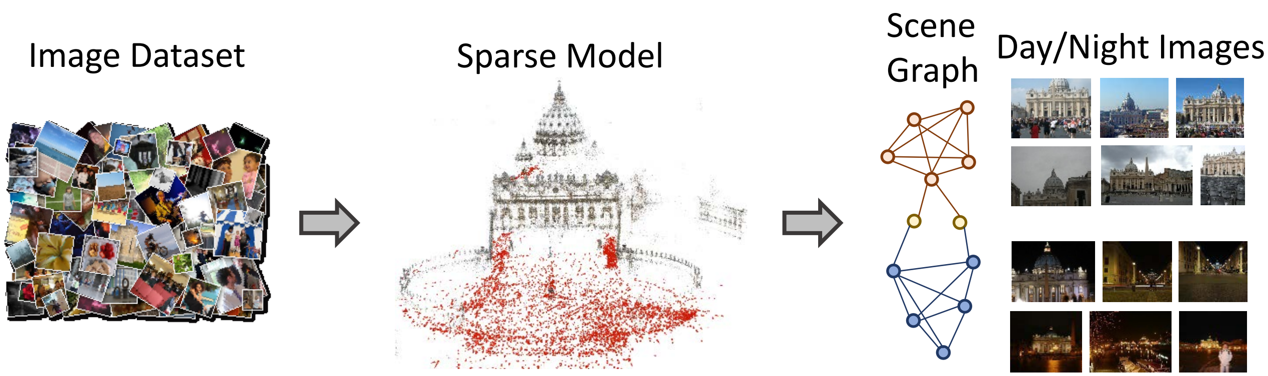
Fused Geometry



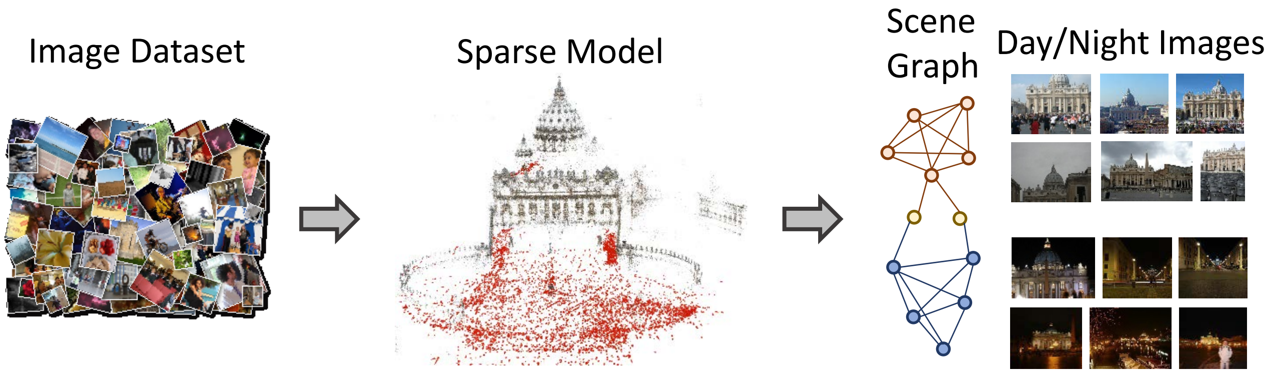
Recoloring of Day & Night Models

Fused Geometry Night Illumination

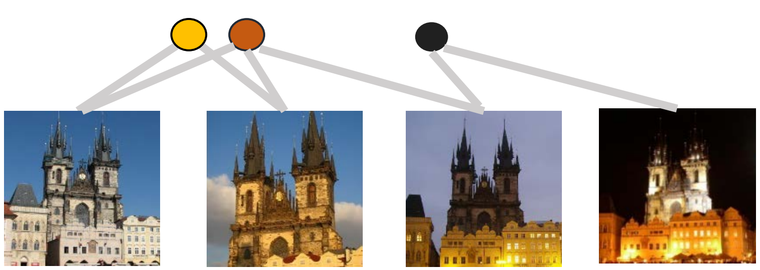
Clustering into Day & Night



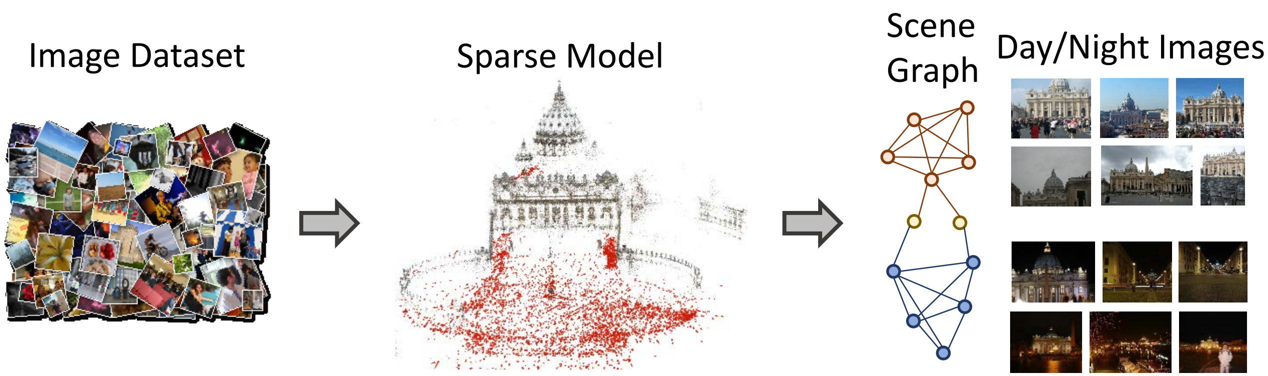
Clustering into Day & Night



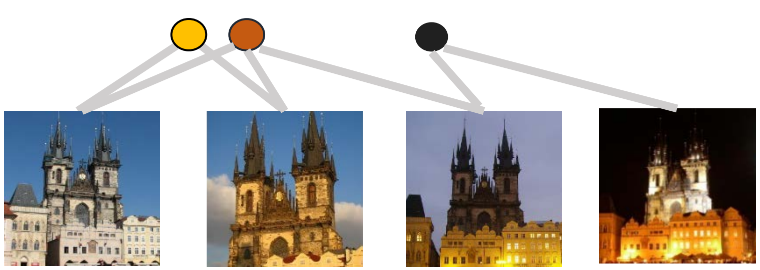
- Images and 3D points in sparse scene graph

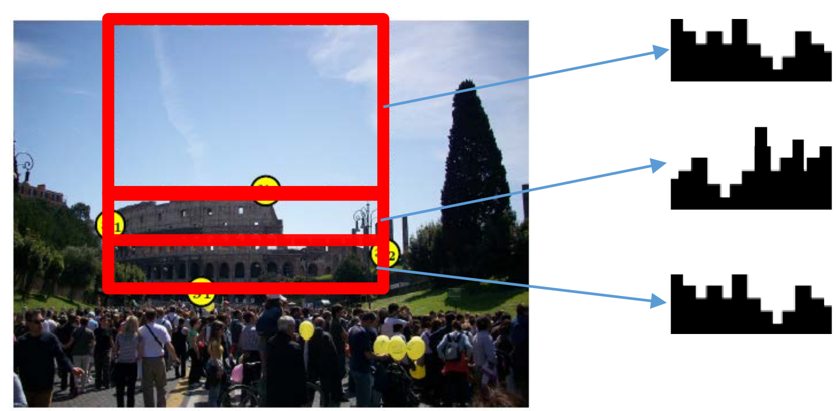


Clustering into Day & Night

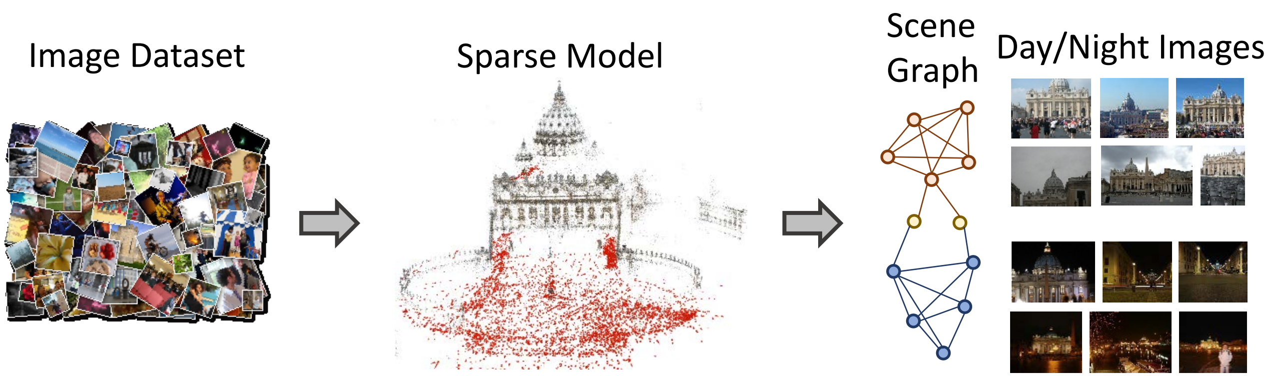


- Images and 3D points in sparse scene graph
- Train SVM on a single model (Colosseum)

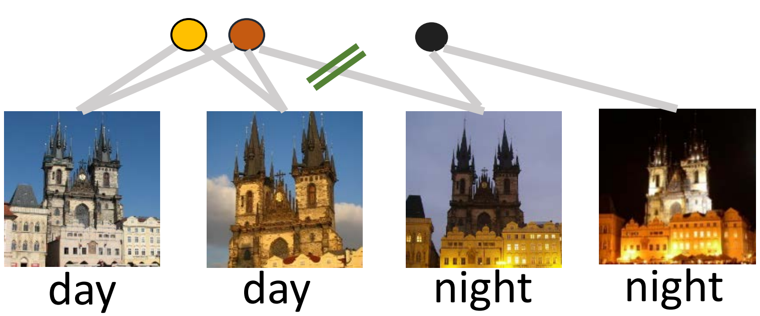


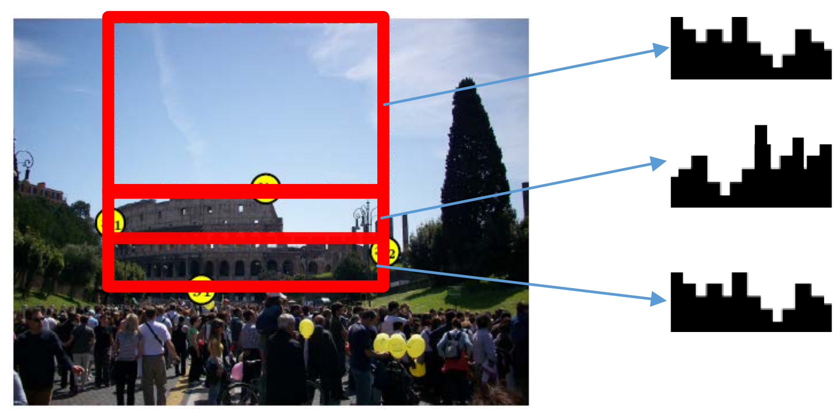


Clustering into Day & Night

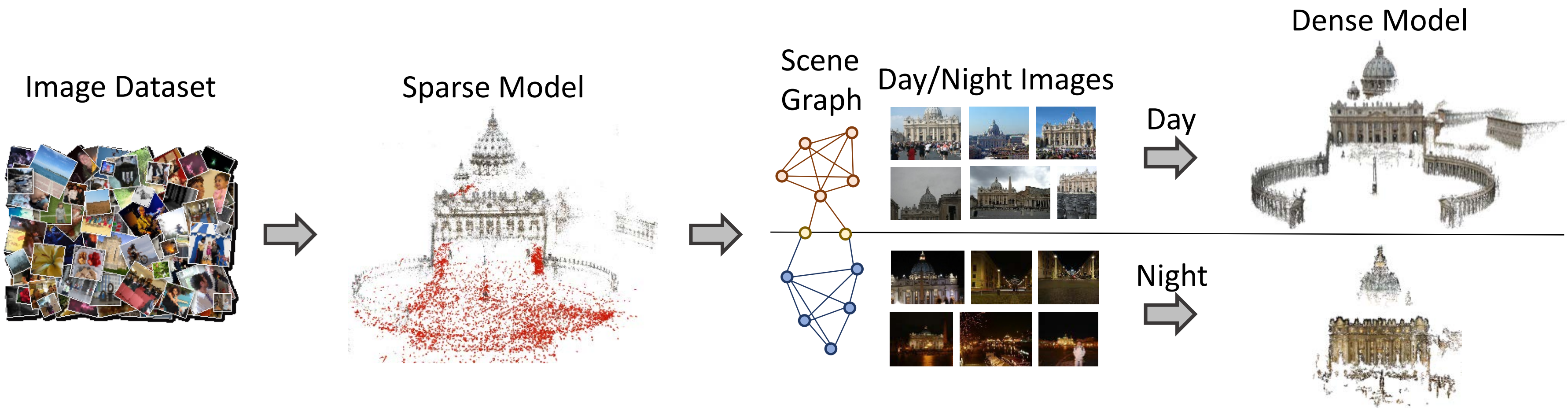


- Images and 3D points in sparse scene graph
- Train SVM on a single model (Colosseum)
- Graph-cut

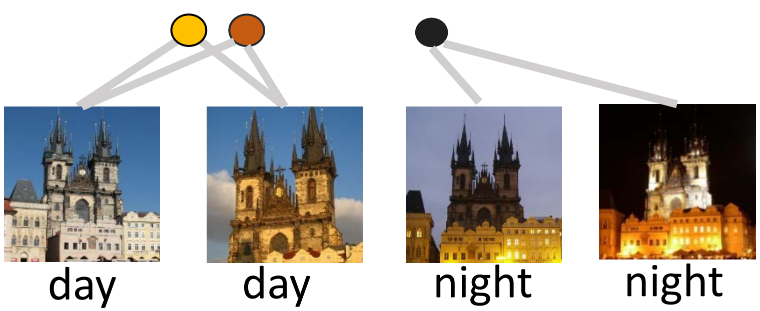


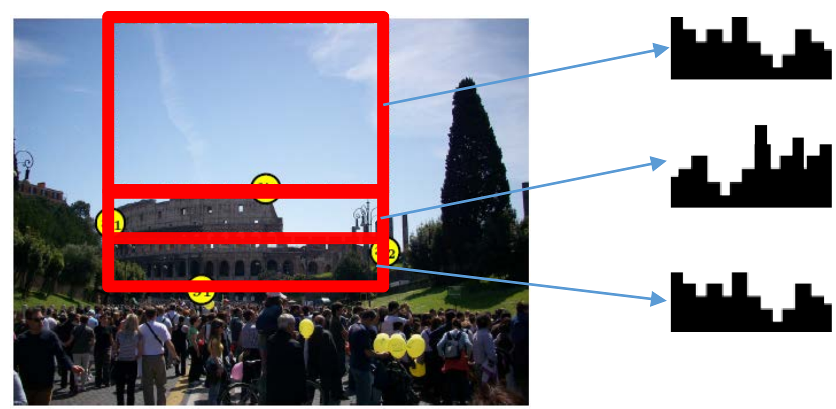


Separate Dense Reconstruction of Day & Night

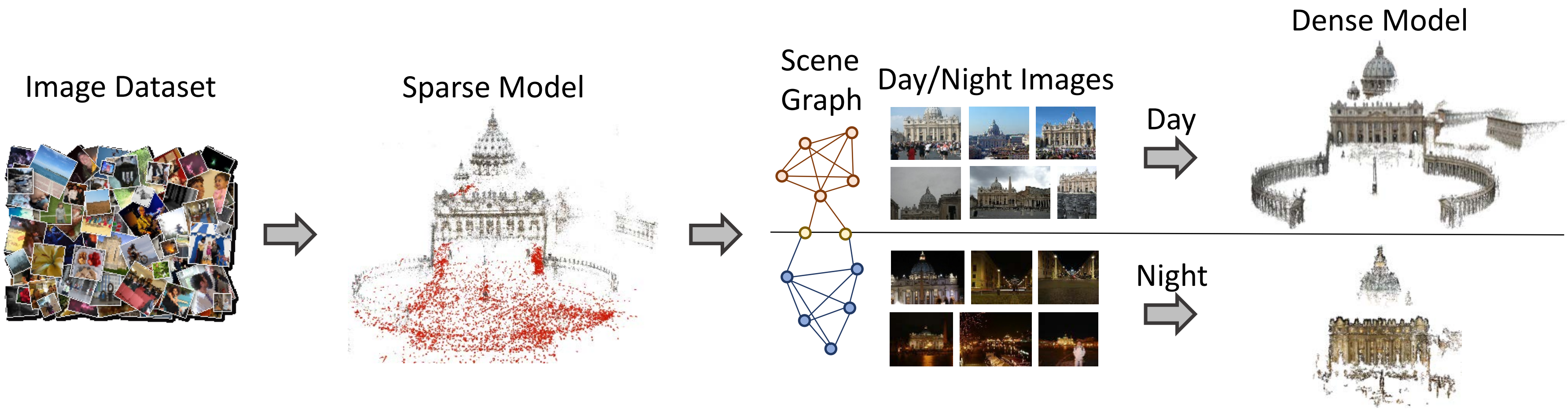


- Images and 3D points in sparse scene graph
- Train SVM on a single model (Colosseum)
- Graph-cut

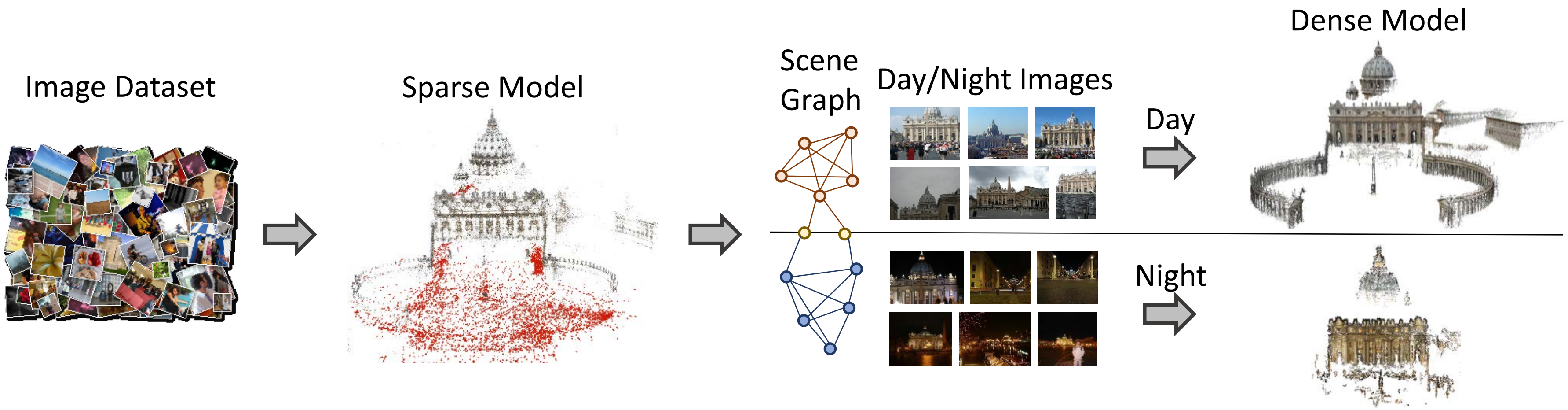




Geometric Fusion of Structure & Recoloring

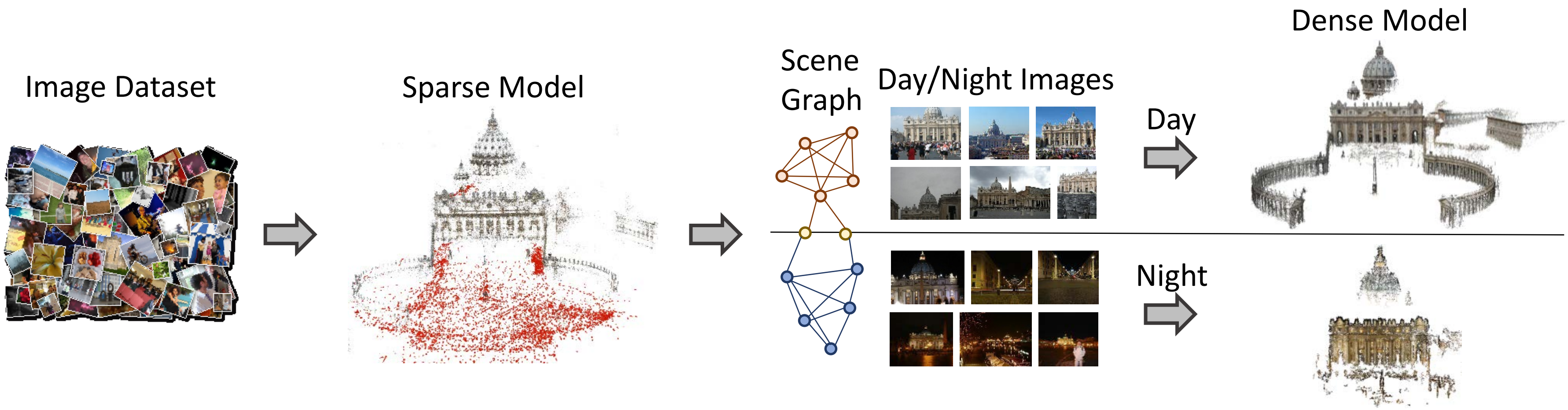


Geometric Fusion of Structure & Recoloring

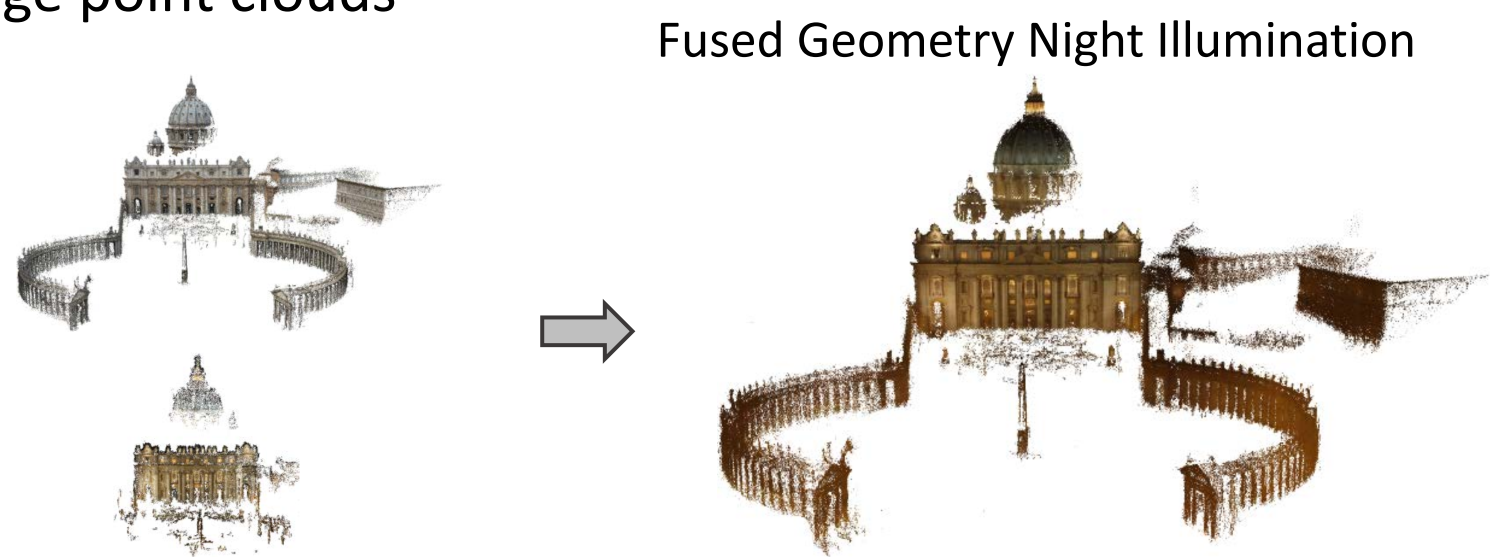


- Merge point clouds

Geometric Fusion of Structure & Recoloring



- Merge point clouds

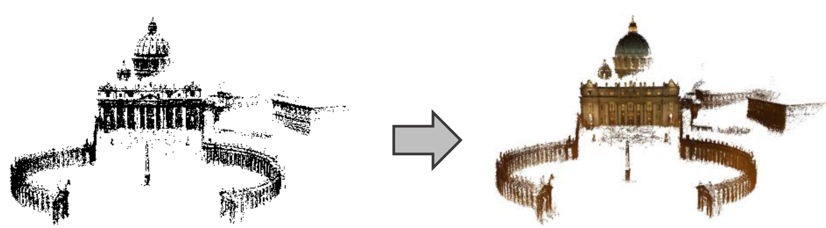


Summary

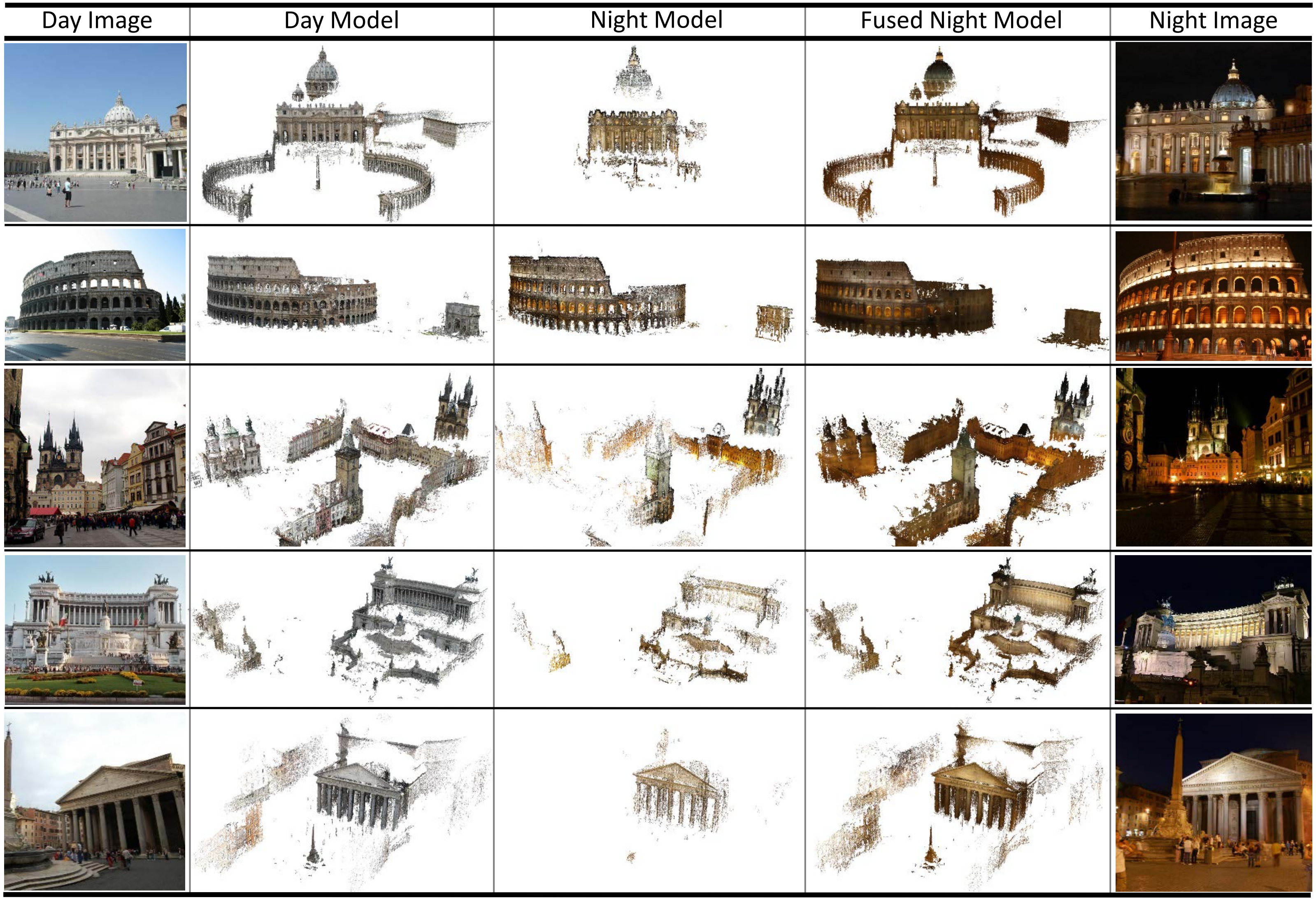
- Automatic separation of day and night images

- Geometric fusion of day & night dense models

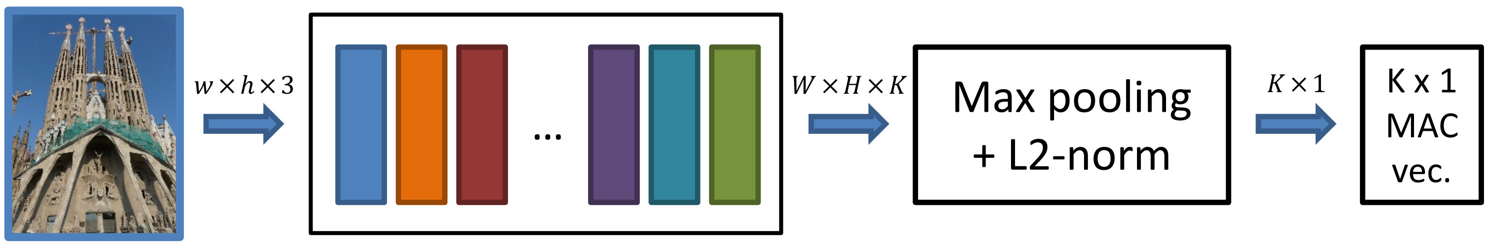
- Color transfer to recolor unreconstructed areas



Some Results



Radenovic, Schoenberger, Ji, Frahm, Chum, and Matas:
From Dusk till Dawn: Modeling in the Dark , CVPR 2016



CNN IMAGE RETRIEVAL LEARNS FROM BOW

Retrieval Challenges

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





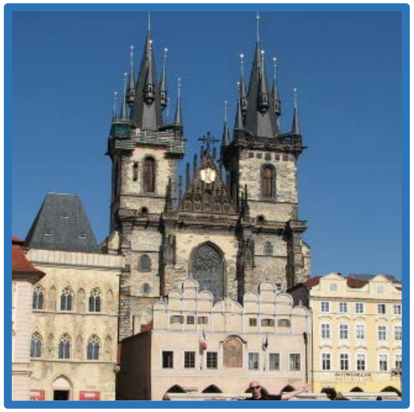
Retrieval Challenges

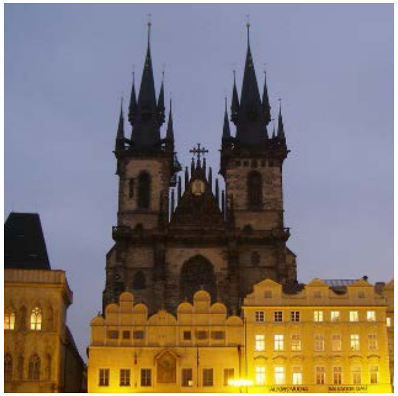
Significant viewpoint and/or scale change

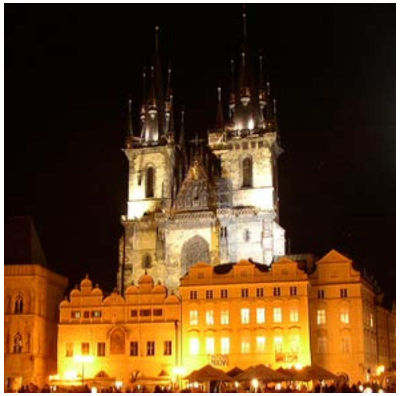
Significant illumination change

Severe occlusions

Visually similar but different objects







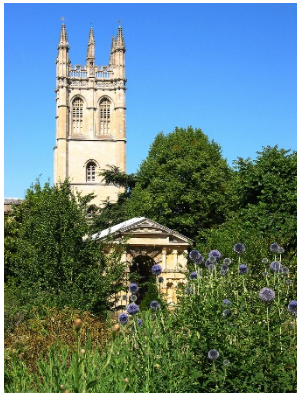
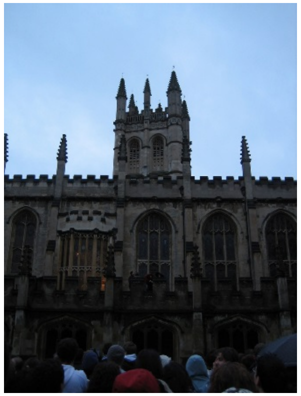
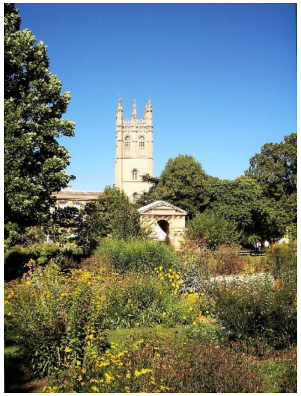
Retrieval Challenges

Significant viewpoint and/or scale change

Significant illumination change

Severe occlusions

Visually similar but different objects


Retrieval Challenges

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




CNN Image Retrieval

- Image representation created from CNN activations of a network pre-trained for classification task

  [Gong et al. ECCV'14, Razavian et al. arXiv'14, Babenko et al. ICCV'15, Kalantidis et al. arXiv'15, Tolias et al. ICLR'16]

\+ Retrieval accuracy suggests generalization of CNNs

\- Trained for image classification, **NOT** retrieval task

Image from ImageNet.org

CNN Image Retrieval

- Ima ... tions
  of a

  [Go ... l.
  ICCV

+ Ret ... NNs

- Tra ... ask

Image from ImageNet.org

CNN Image Retrieval

- CNN network re-trained using a dataset that contains landmarks and buildings as object classes.

  [Babenko et al. ECCV’14]

+ Training dataset closer to the target task

- Final metric different to the one actually optimized

- Constructing training datasets requires manual effort

Image from [Babenko et al. ECCV’14]

CNN Image Retrieval

Image from [Babenko et al. ECCV'14]

CNN Image Retrieval

- NetVLAD: end-to-end fine-tuning for image retrieval. Geo-tagged dataset for weakly supervised fine-tuning.

  [Arandjelovic et al. CVPR'16]

+ Training dataset corresponds to the target task

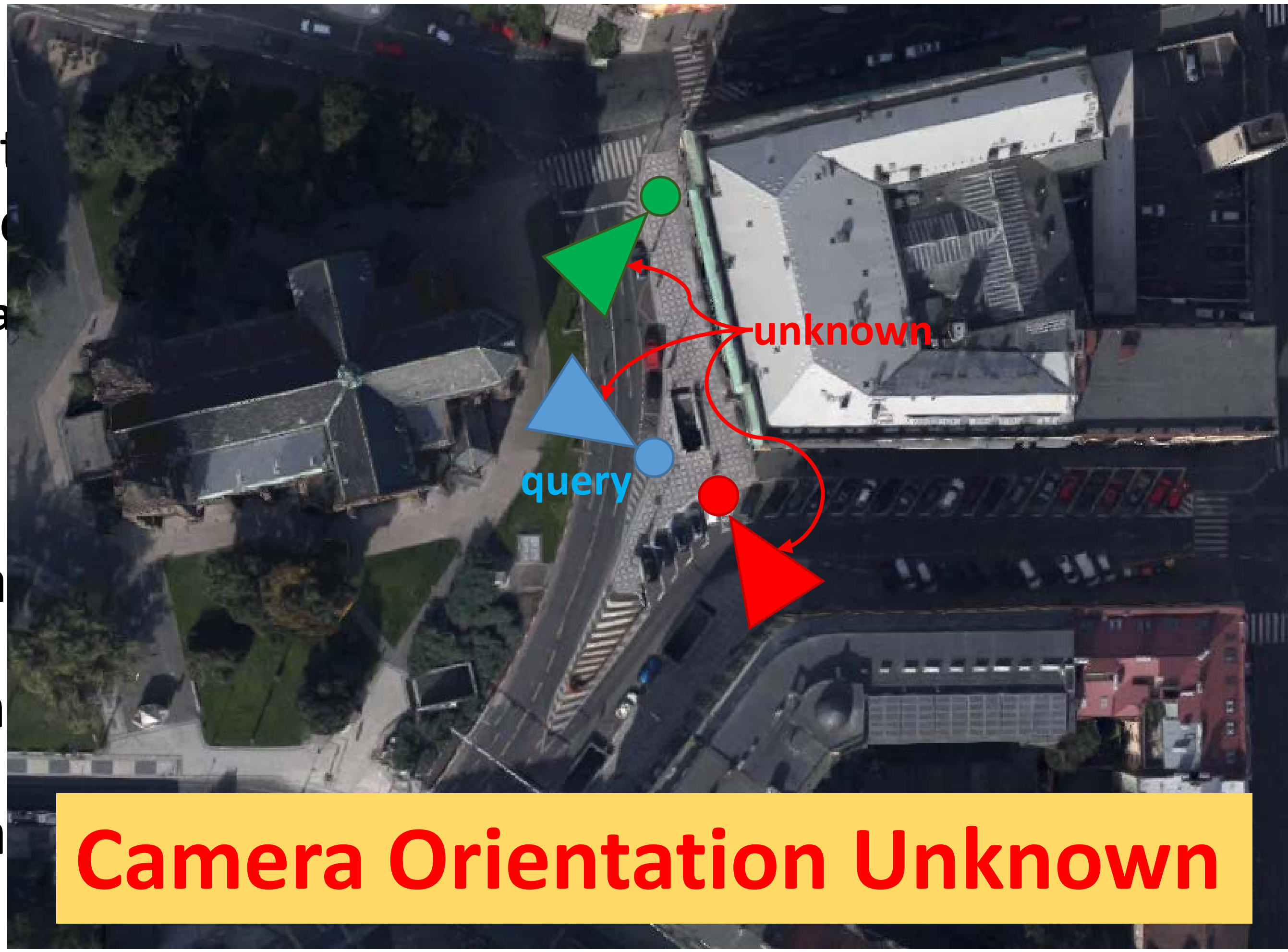
+ Final metric corresponds to the one actually optimized

− Training dataset requires geo-tags

CNN Image Retrieval

- Net al.
  Ge ng.

  [Ara



+ Tra
+ Fin zed
- Tra

CNN learns from BoW – Training Data

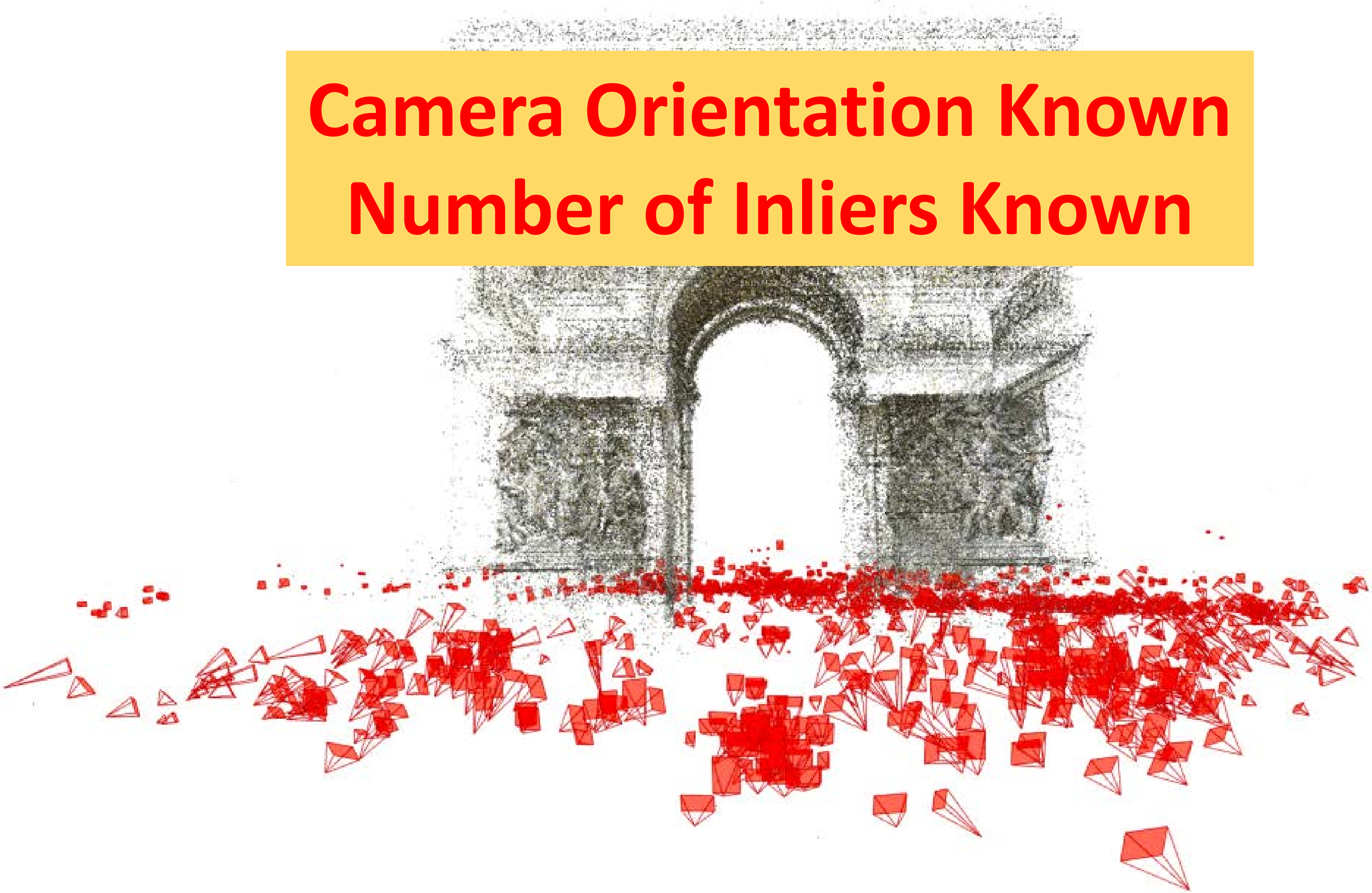
Input: Large unannotated dataset

1. Initial clusters created by grouping of spatially related images **[Chum & Matas PAMI'10]**
2. Clustered images used as queries for a retrieval-SfM pipeline **[Schonberger et al. CVPR'15]**

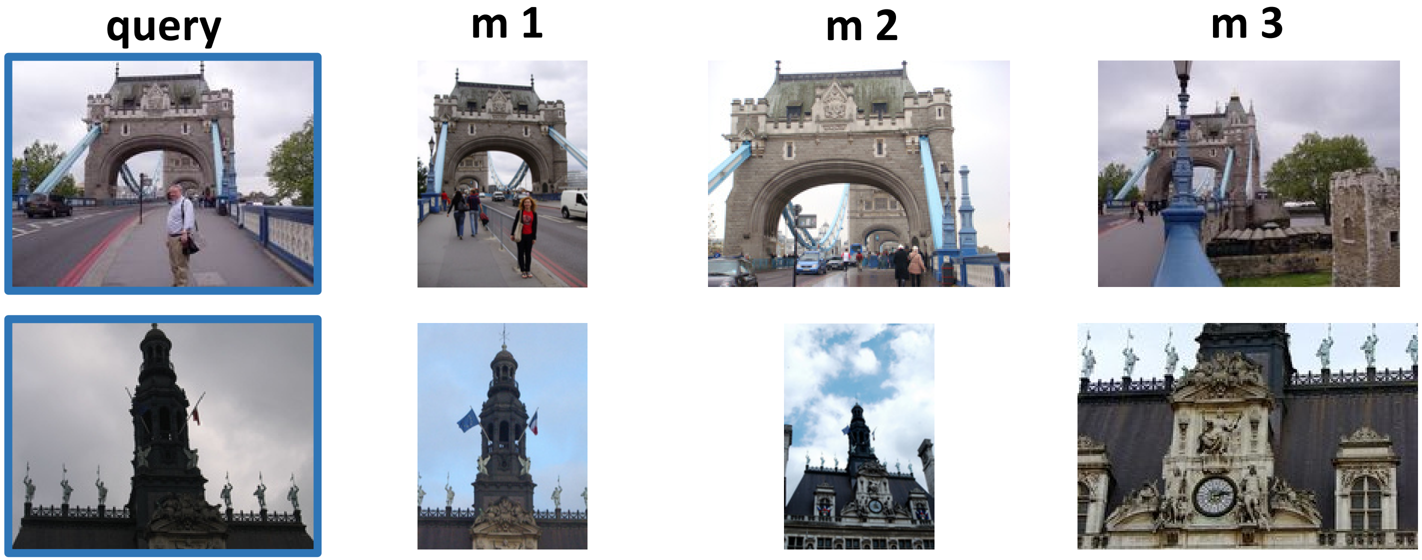
Output: Non-overlapping 3D models

551 (134k) training / 162 (30k) validation

CNN learns from BoW – Training Data



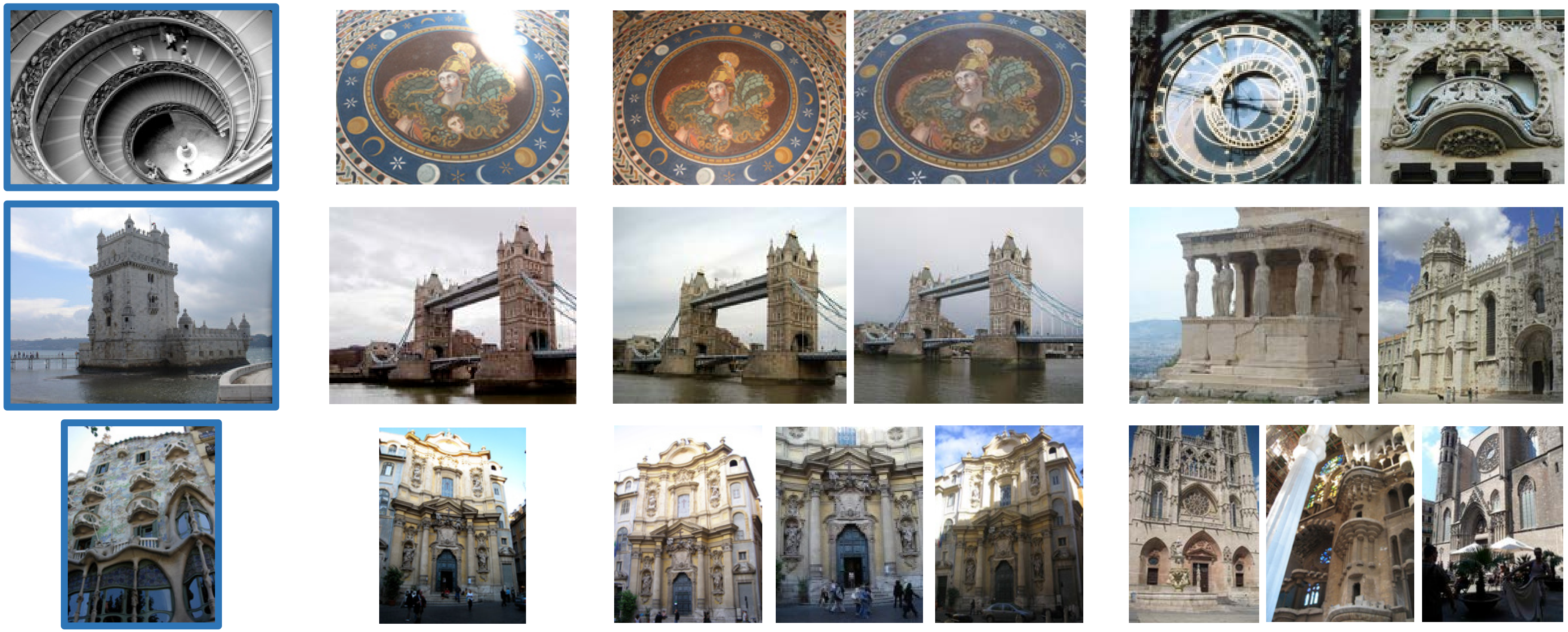
CNN learns from BoW – Positives



1. Descriptor distance: Image with the lowest global descriptor distance is chosen (NetVLAD use this)
2. Maximum inliers: Image with the highest number of co-observed 3D points with the query image is chosen
3. Relaxed inliers: Random image close to the query, with enough inliers and not an extreme scale change is chosen

CNN learns from BoW – Negatives

query **hardest negative** **N 1** **N 2**

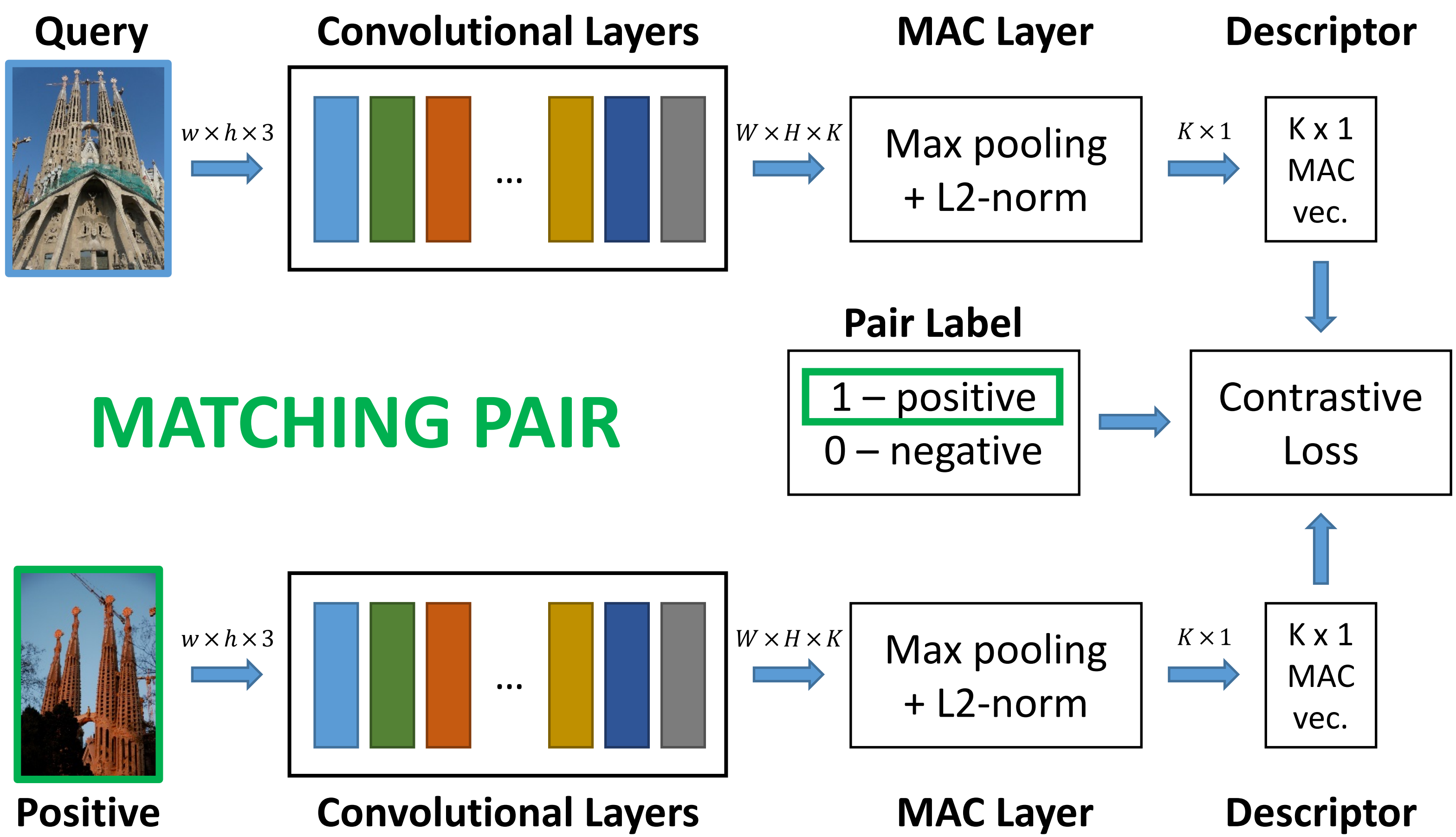


K-nearest neighbors of the query image are selected from all non-matching clusters, using different methods:

1. No constraint: chosen images often near identical.
2. At most one image per cluster: higher variability.

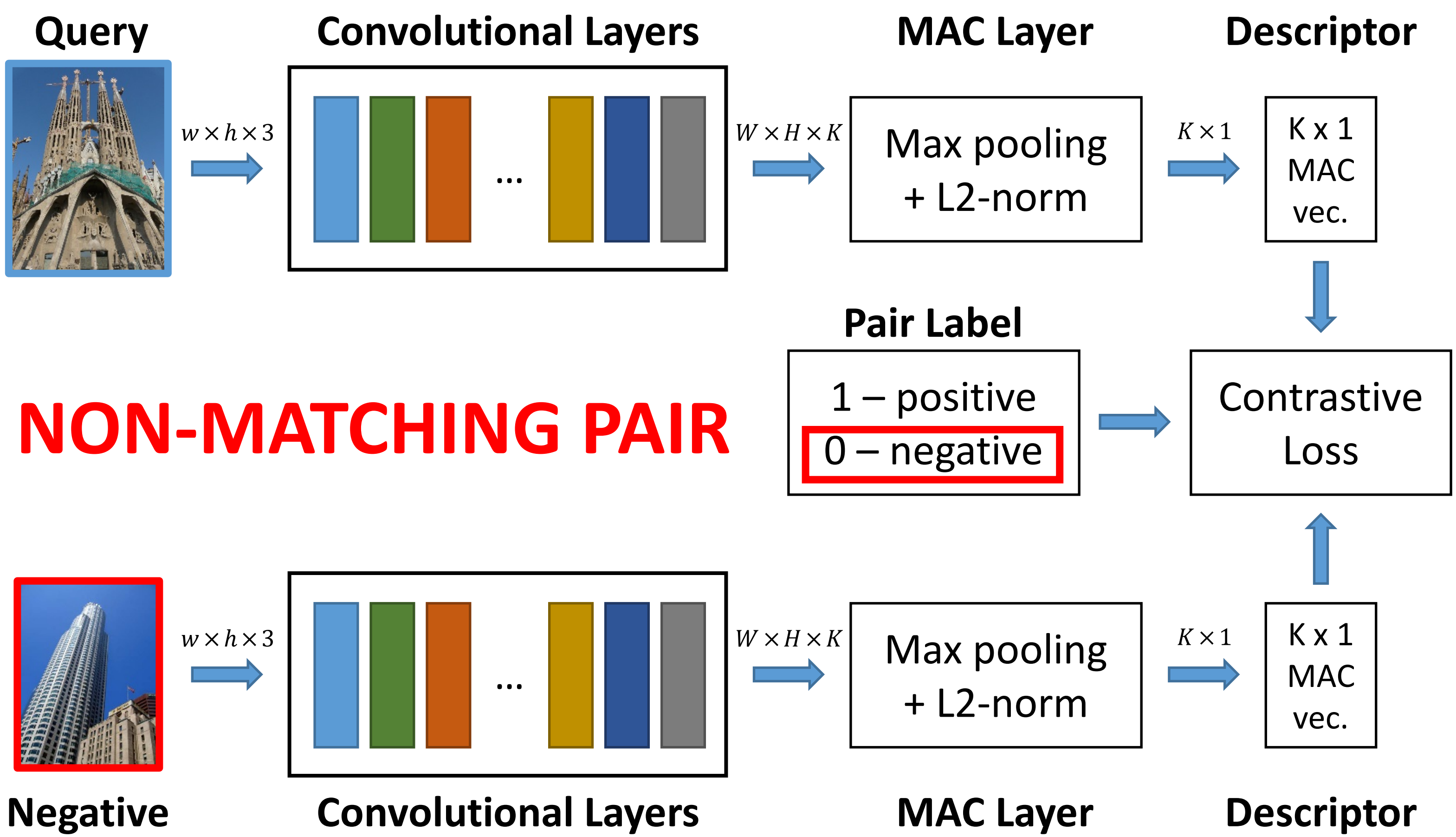
CNN Siamese Learning

MAC – Maximum Activations of Convolutions
$w \times h$ – image width and height
$W \times H$ – number of activations for feature map $k \in \{1 \dots K\}$
K – number of feature maps in the last convolutional layer



CNN Siamese Learning

MAC – Maximum Activations of Convolutions
$w \times h$ – image width and height
$W \times H$ – number of activations for feature map $k \in \{1 \dots K\}$
K – number of feature maps in the last convolutional layer



Contrastive Loss

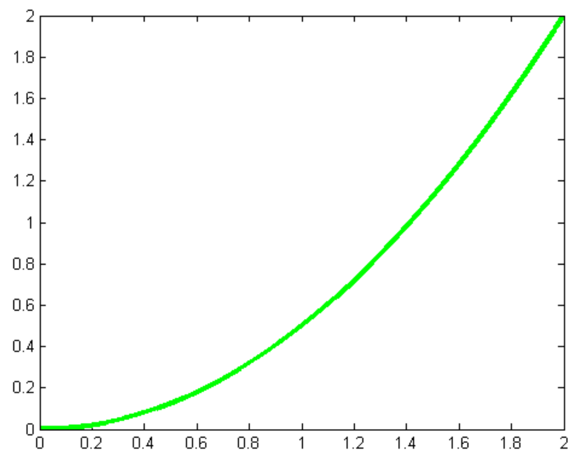
$\bar{\boldsymbol{f}}(i)$ – MAC vector for image i
$Y(i,j)$ – Label for image pair (i,j), 1 – positive, 0 – negative
τ – defining when a negative pair is far enough not to influence the loss

$$L(i,j) = \frac{1}{2}\left(Y(i,j)\left\|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right\|^2 + \cancel{(1-Y(i,j))\left(\max\{0, \tau - \|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\|\}\right)^2}\right)$$

POSITIVE PAIR

$$L(i,j) = \frac{1}{2}\left\|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right\|^2$$

$L(i,j)$



$\left\|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right\|$

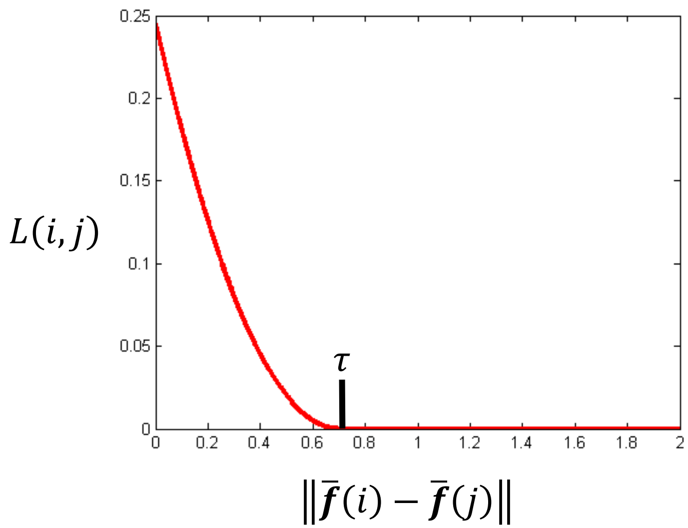
Contrastive Loss

$\bar{\boldsymbol{f}}(i)$ – MAC vector for image i
$Y(i,j)$ – Label for image pair (i,j), 1 – positive, 0 – negative
τ – defining when a negative pair is far enough not to influence the loss

$$L(i,j) = \frac{1}{2}\left(\cancel{Y(i,j)\|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\|^2} + (1 - Y(i,j)\left(\max\{0, \tau - \|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\|\}\right)^2\right)$$

NEGATIVE PAIR

$$L(i,j) = \frac{1}{2}\max\{0, \tau - \|\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\|\}^2$$



Whitening and dimensionality reduction

1. PCA_w – PCA of an independent set of descriptors used for whitening and dimensionality reduction
   [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 part is the inverse of the square-root of the intraclass (matching pairs) covariance matrix ${C_S}^{-1/2}$

   $$C_S = \sum_{Y(i,j)=1} \left(\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right)\left(\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right)^\top$$

   - Rotation part is the PCA of the interclass (non-matching pairs) covariance matrix in the whitened space $eig\left({C_S}^{-1/2} C_D {C_S}^{-1/2}\right)$

   $$C_D = \sum_{Y(i,j)=0} \left(\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right)\left(\bar{\boldsymbol{f}}(i) - \bar{\boldsymbol{f}}(j)\right)^\top$$

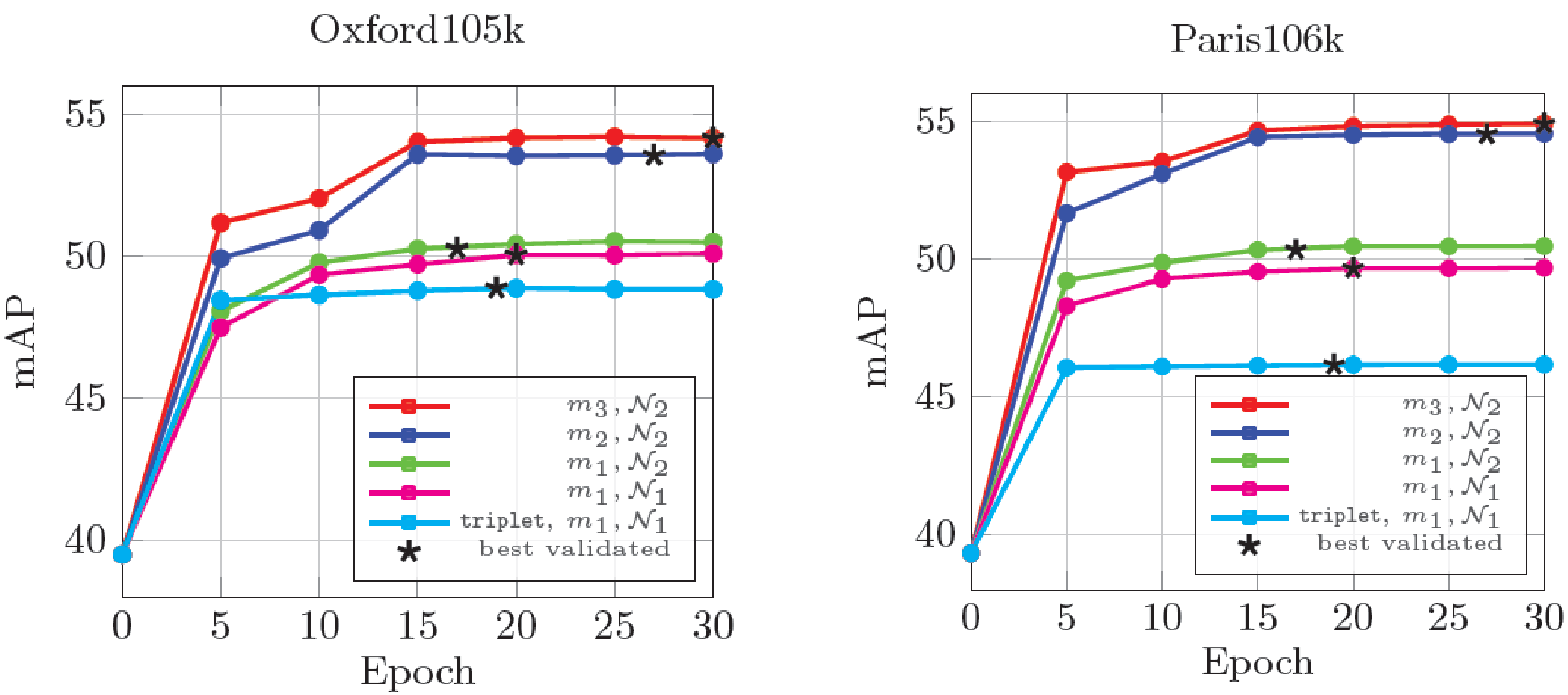
   - Dimensionality reduction is done by using only D largest eigenvalues

Experiments – datasets

- **Oxford 5k dataset** (1024 x 768) **[Philbin et al. CVPR'07]**
  - 55 queries, 5.062 database images
- **Paris 6k dataset** (1024 x 768) **[Philbin et al. CVPR'08]**
  - 55 queries, 6.300 database images
- **Holidays dataset** (1024 x 768) **[Jegou et al. ECCV'10]**
  - 500 queries, 1.491 database images
- **Oxford 100k dataset** (1024 x 768) **[Philbin et al. CVPR'07]** Combined with previous datasets to create:
  - **Oxford 105k:** 55 queries, 104.844 database images
  - **Paris 106k:** 55 queries, 106.082 database images
  - **Holidays 101k:** 500 queries, 101.273 database images
- **Protocol:** mean Average Precision (mAP)

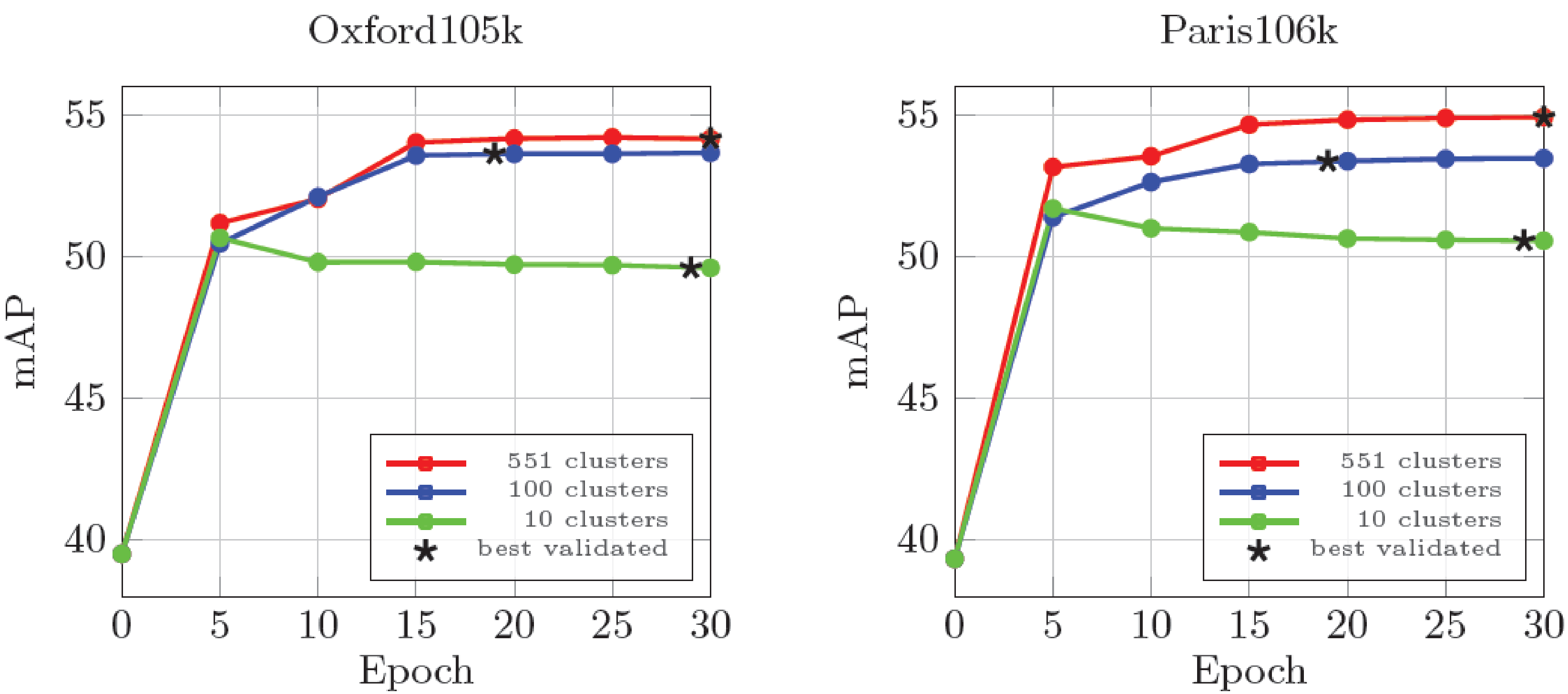
Experiments – Learning (AlexNet)

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



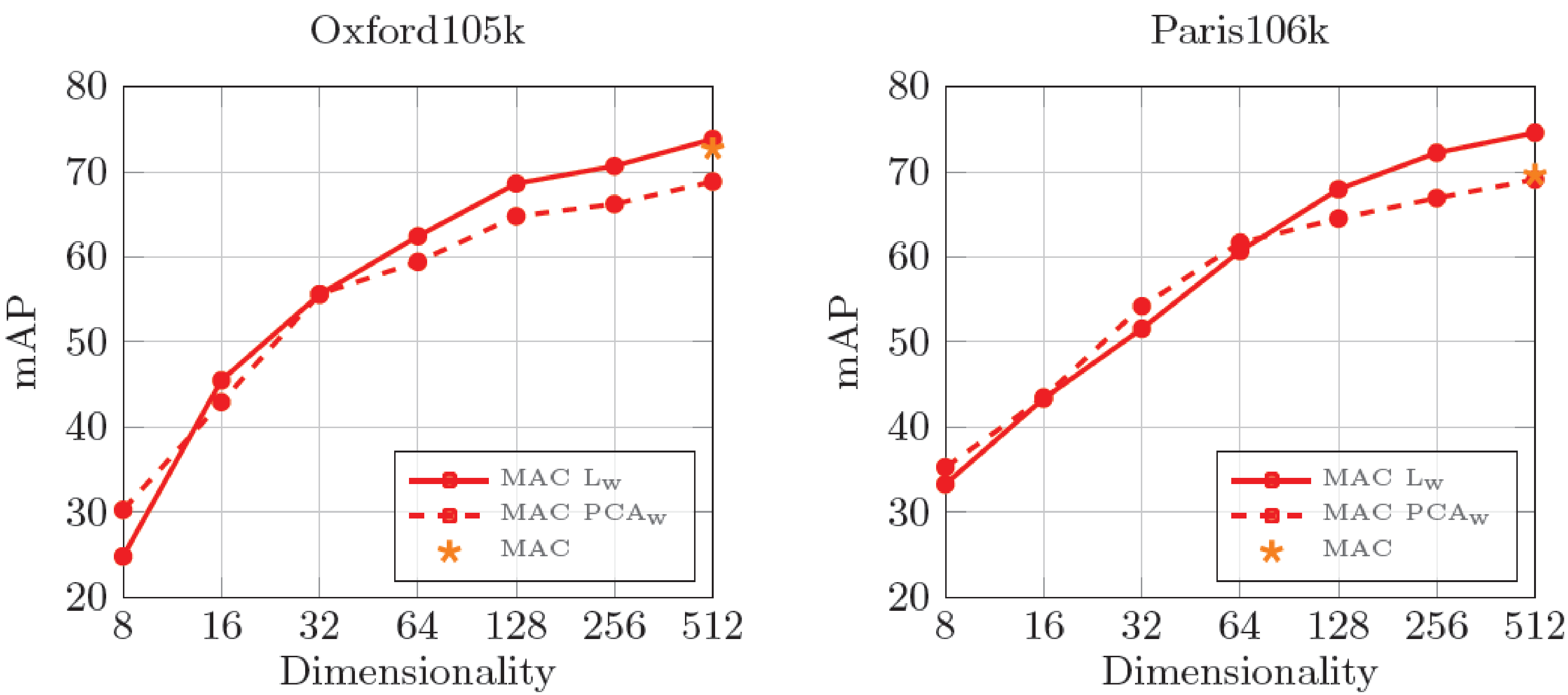
Experiments – Dataset variability (AlexNet)

- More 3D models leads to higher performance
- Remarkable improvements even with 10 models



Experiments – Dimensionality reduction (VGG)

- Our 32D comparable with previous state-of-the-art on 256D
- Oxford5k: Our 32D MAC **69.2** vs. 256D NetVLAD **63.5** mAP
- Paris6k: Our 32D MAC **69.5** vs. 256D NetVLAD **73.5** mAP



Experiments – Overfitting / Generalization

- We added Oxford and Paris landmarks as 3D models and repeated fine-tuning
- Negligible difference in the performance of the network on Oxford and Paris evaluation results

Only +0.3 mAP on average over all testing datasets

State-of-the-art

Method		D	Oxf5k		Oxf105k		Par6k		Par106k		Hol	Hol
			crop$_{\mathcal{I}}$	crop$_{\mathcal{X}}$	crop$_{\mathcal{I}}$	crop$_{\mathcal{X}}$	crop$_{\mathcal{I}}$	crop$_{\mathcal{X}}$	crop$_{\mathcal{I}}$	crop$_{\mathcal{X}}$		101k
Compact representations												
mVoc/BoW [11]		128	48.8	–	41.4	–	–	–	–	–	65.6	–
Neural codes† [14]	(f A)	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	(f V)	128	**75.8**	**76.8**	**68.6**	**70.8**	77.6	78.8	68.0	69.0	73.2	58.8
⋆ R-MAC	(f V)	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	–	55.5	–	–	–	67.7	–	–	**86.0**	–
NetVlad [35]	(f V)	256	–	**63.5**	–	–	–	**73.5**	–	–	84.3	–
⋆ MAC	(f A)	256	62.2	65.4	52.8	58.0	68.9	72.2	54.7	58.5	76.2	63.8
⋆ R-MAC	(f A)	256	62.5	68.9	53.2	61.2	74.4	76.6	61.8	64.8	81.5	**70.8**
⋆ MAC	(f V)	256	**77.4**	**78.2**	**70.7**	**72.6**	80.8	81.9	72.2	73.4	77.3	62.9
⋆ R-MAC	(f V)	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	(f V)	512	**79.7**	80.0	**73.9**	**75.1**	82.4	82.9	74.6	75.3	79.5	67.0
⋆ R-MAC	(f V)	512	77.0	**80.1**	69.2	74.1	**83.8**	**85.0**	**76.4**	**77.9**	82.5	**71.5**

State-of-the-art

Method	D	Oxf5k		Oxf105k		Par6k		Par106k		Hol	Hol 101k
		Crop$_\mathcal{I}$	Crop$_\mathcal{X}$	Crop$_\mathcal{I}$	Crop$_\mathcal{X}$	Crop$_\mathcal{I}$	Crop$_\mathcal{X}$	Crop$_\mathcal{I}$	Crop$_\mathcal{X}$		
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	–	**51.5**	–	**46.7**	–	–	–	–	**72.9**	–
⋆ MAC (**fV**)	32	**65.3**	**69.2**	**55.6**	**59.5**	**63.9**	**69.5**	51.6	**56.3**	62.4	41.8
⋆ R-MAC (**fV**)	32	58.4	64.2	50.1	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 [51]	–	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**	–	–

Summary

- Introduction to image retrieval and BoW
- Discovering image clusters and co-occurring features with min-Hash
- Retrieval with geometric constraints helps to get better 3D reconstruction
  - more details
  - more stable – less mismatched structures
- Automated 3D models provide great training data for CNN retrieval

Vision and Sports Summer School

Prague August 2017

Vittorio Ferrari Jiri Matas Ondra Chum Giorgos Tolias

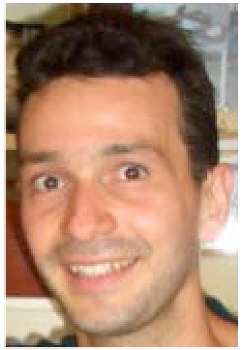
cmp.felk.cvut.cz/summerschool2016

Lecturers 2016


Jiri Matas


Victor Lempitsky


Christoph Lampert


Vittorio Ferrari

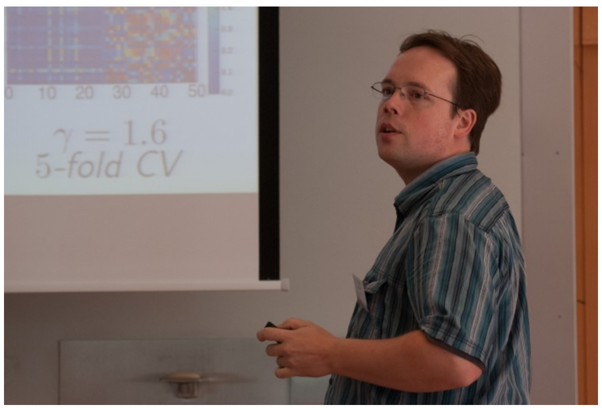

Andrew Fitzgibbon

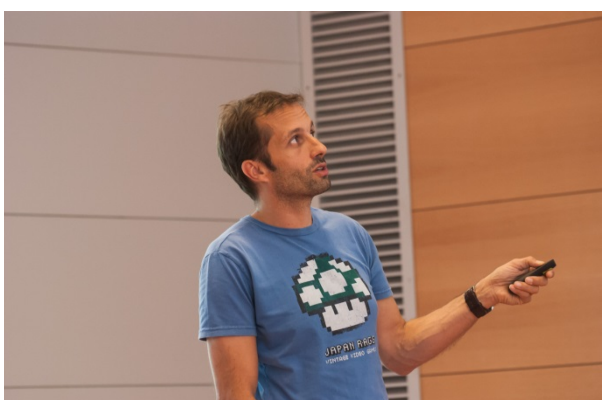

Daniel Cremers


Raquel Urtasun


Ondrej Chum





0 10 20 30 40 50
γ = 1.6
5-fold CV



JAPAN RAGS