

# Instance-level recognition

---

- 1) Local invariant features
- 2) Matching and recognition with local features
- 3) Efficient visual search**

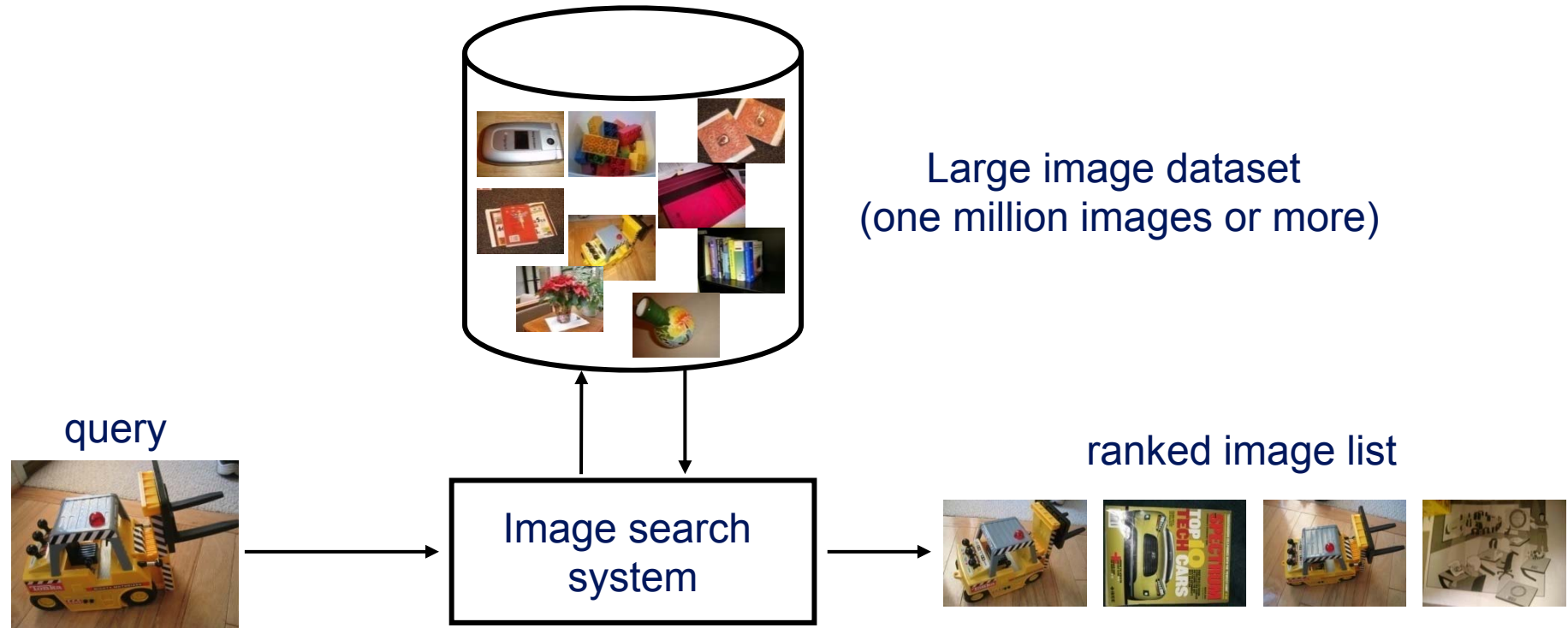
# Visual search

---



# Image search system for large datasets

---



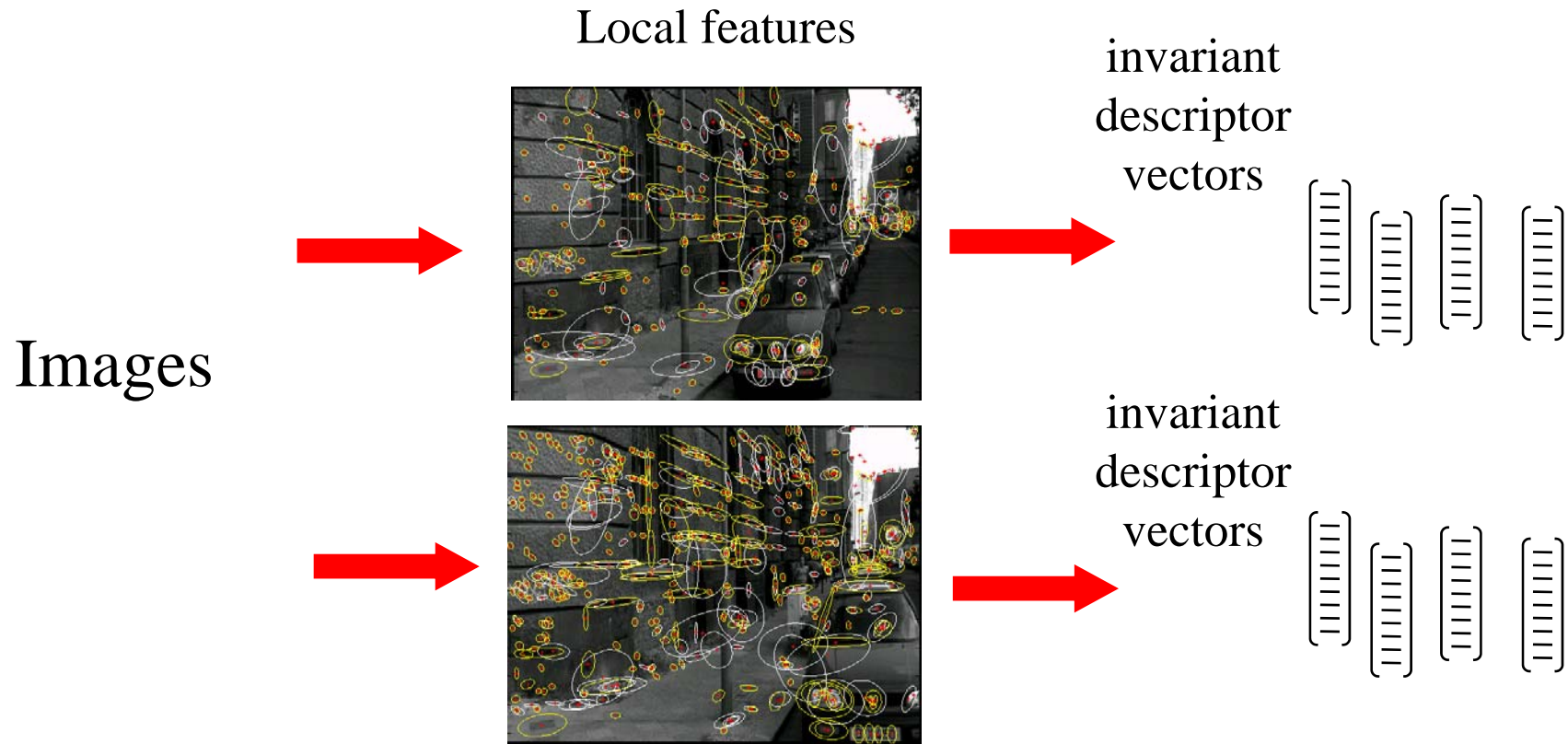
- **Issues** for very large databases
  - to reduce the query time
  - to reduce the storage requirements
  - with minimal loss in retrieval accuracy

## Two strategies

1. Efficient approximate nearest neighbor search on local feature descriptors
2. Quantize descriptors into a “visual vocabulary” and use efficient techniques from text retrieval  
(Bag-of-words representation)



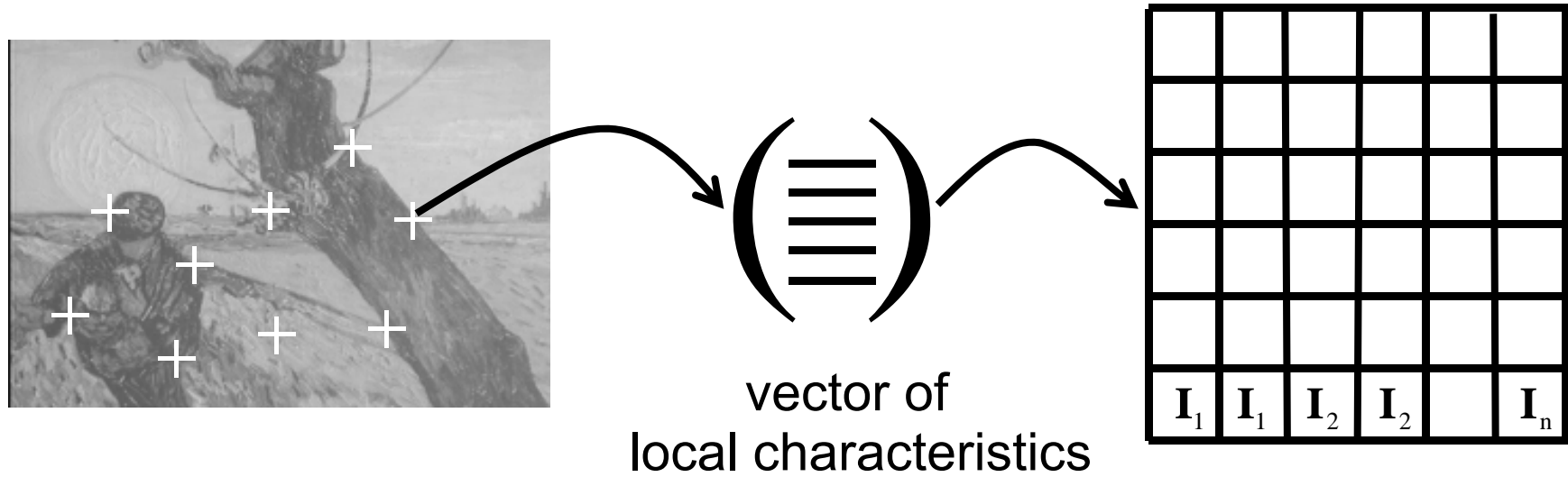
# Strategy 1: Efficient approximate NN search



1. Compute local features in each image independently
2. Describe each feature by a descriptor vector
3. Find nearest neighbour vectors between query and database
4. Rank matched images by number of (tentatively) corresponding regions
5. Verify top ranked images based on spatial consistency

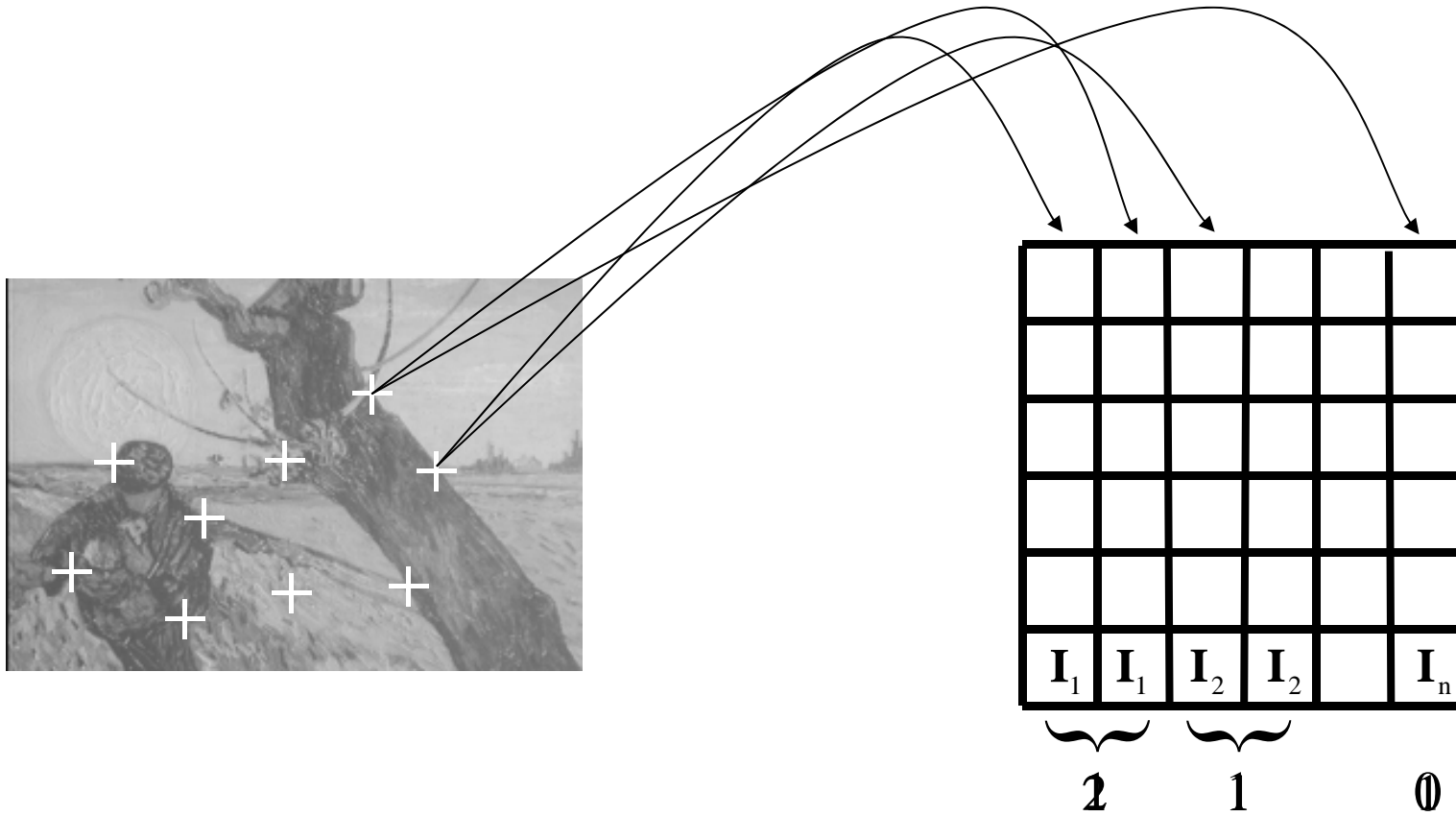
# Voting algorithm

---



# Voting algorithm

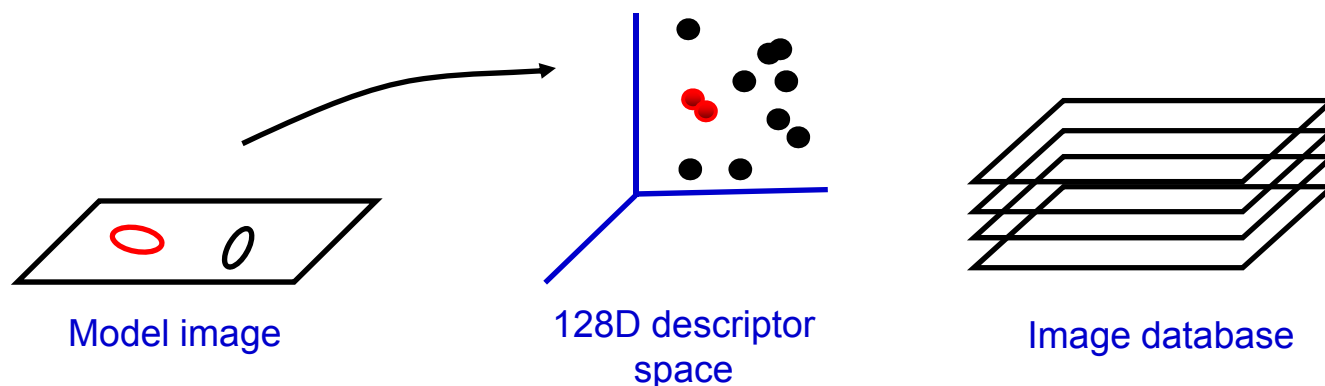
---



$I_1$  is the corresponding model image

# Finding nearest neighbour vectors

Establish correspondences between query image and images in the database by **nearest neighbour matching** on SIFT vectors



Solve following problem for all feature vectors,  $\mathbf{x}_j \in \mathcal{R}^{128}$ , in the query image:

$$\forall j \text{ NN}(j) = \arg \min_i \|\mathbf{x}_i - \mathbf{x}_j\|$$

where,  $\mathbf{x}_i \in \mathcal{R}^{128}$ , are features from all the database images.

# Quick look at the complexity of the NN-search

N ... images

M ... regions per image (~1000)

D ... dimension of the descriptor (~128)

Exhaustive linear search:  $O(M \cdot N \cdot D)$

Example:

- Matching two images (N=1), each having 1000 SIFT descriptors  
Nearest neighbors search: 0.4 s (2 GHz CPU, implementation in C)
- Memory footprint:  $1000 \cdot 128 = 128\text{kB}$  / image

# of images	CPU time	Memory req.
N = 1,000 ...	~7min	(~100MB)
N = 10,000 ...	~1h7min	(~ 1GB)
...		
N = $10^7$	~115 days	(~ 1TB)
...		
All images on Facebook:		
N = $10^{10}$ ...	~300 years	(~ 1PB)

# Nearest-neighbor matching

Solve following problem for all feature vectors,  $\mathbf{x}_j$ , in the query image:

$$\forall j \text{ } NN(j) = \arg \min_i ||\mathbf{x}_i - \mathbf{x}_j||$$

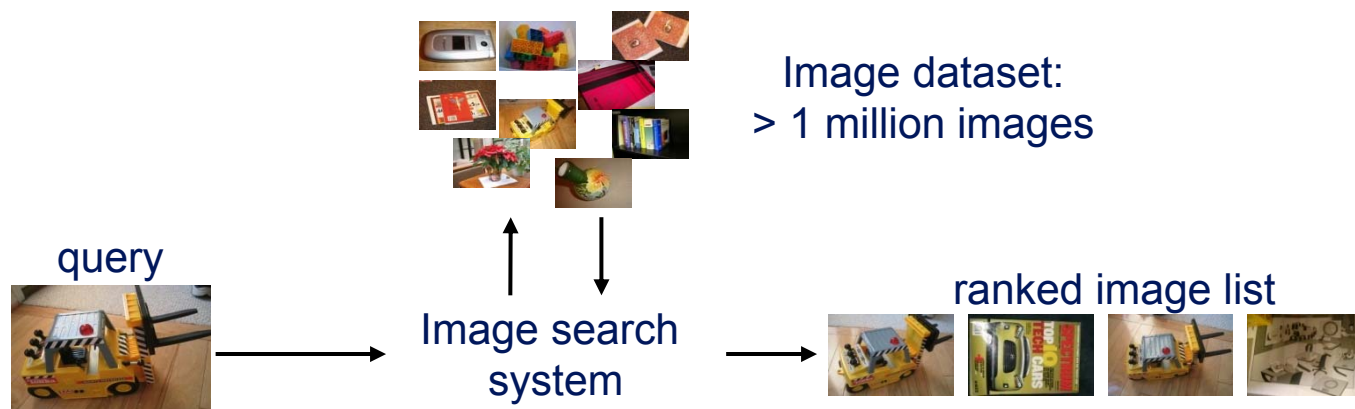
where  $\mathbf{x}_i$  are features in database images.

Nearest-neighbour matching is the major computational bottleneck

- Linear search performs  $dn$  operations for  $n$  features in the database and  $d$  dimensions
- No exact methods are faster than linear search for  $d > 10$
- Approximate methods can be much faster, but at the cost of missing some correct matches

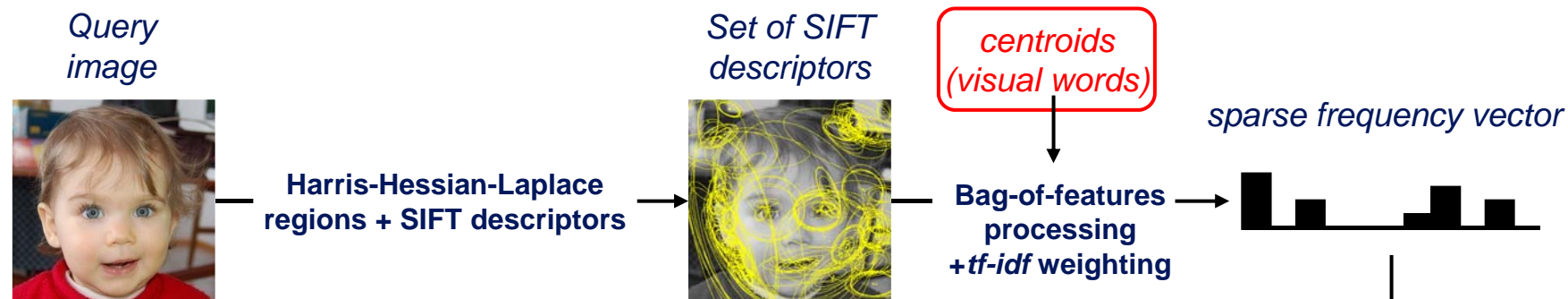
# Large scale object/scene recognition

---

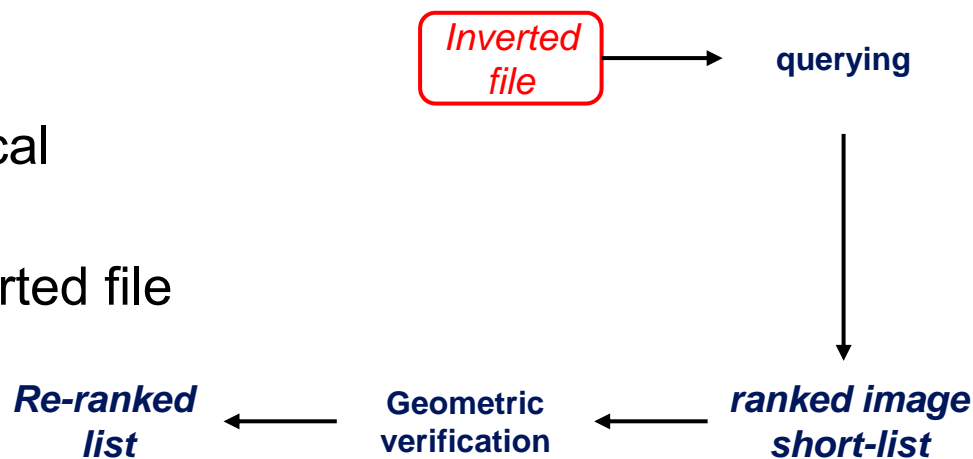


- Each image described by approximately 1000 descriptors
  - $10^9$  descriptors to index for one million images!
- Database representation in RAM:
  - Size of descriptors : 1 TB, search+memory intractable

# Bag-of-features [Sivic&Zisserman'03]



- “visual words”:
  - 1 “word” (index) per local descriptor
  - only images ids in inverted file  
→ 8 GB fits!

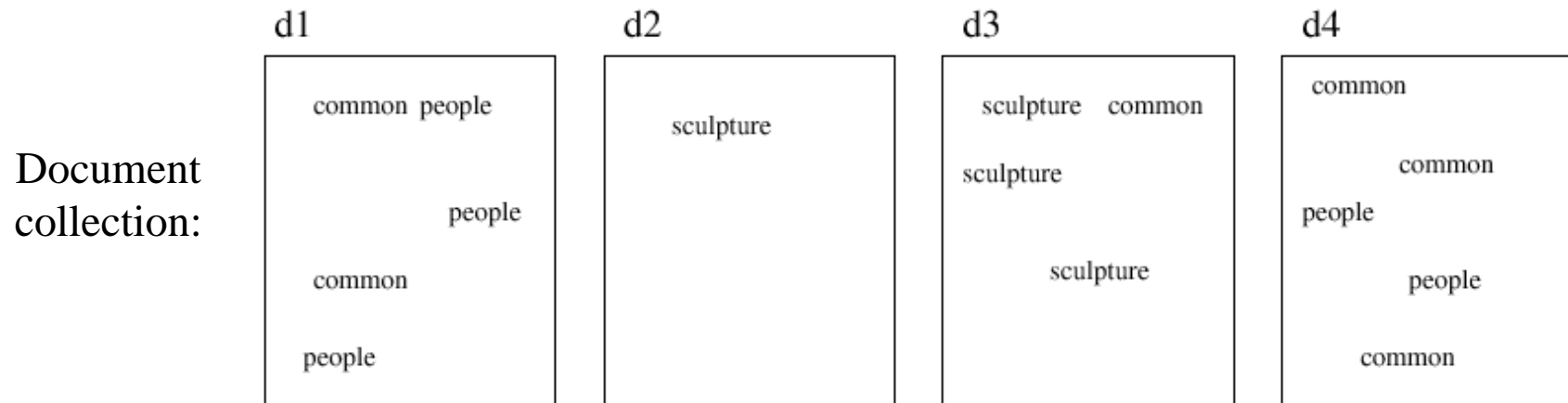


[Chum & al. 2007]



# Indexing text with inverted files

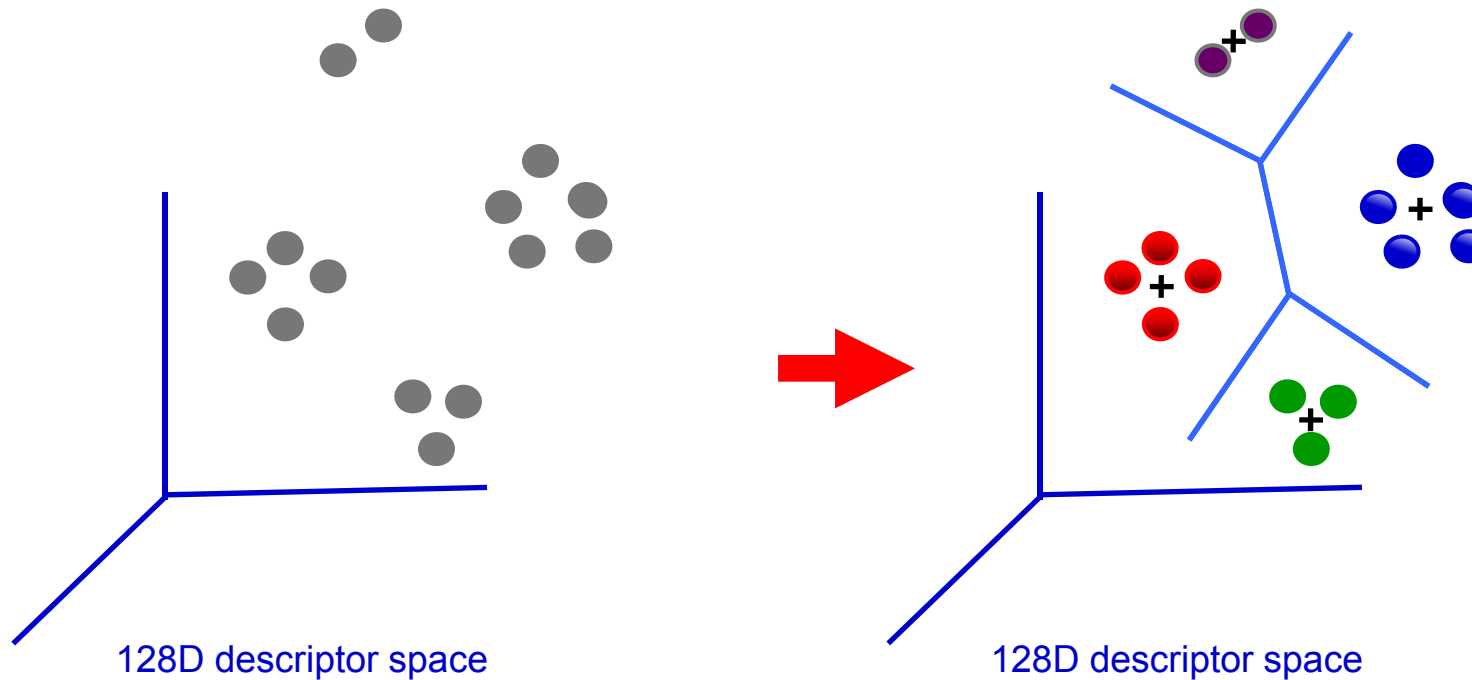
---



Inverted file:	<b>Term</b>	<b>List of hits</b> (occurrences in documents)
	People	[d1:hit hit hit], [d4:hit hit] ...
	Common	[d1:hit hit], [d3: hit], [d4: hit hit hit] ...
	Sculpture	[d2:hit], [d3: hit hit hit] ...

Need to map feature descriptors to “visual words”

## Build a visual vocabulary



Vector quantize descriptors

- Compute SIFT features from a subset of images
- K-means clustering (need to choose K)

[Sivic and Zisserman, ICCV 2003]

# K-means clustering

Minimizing sum of squared Euclidean distances  
between points  $x_i$  and their nearest cluster centers

## Algorithm:

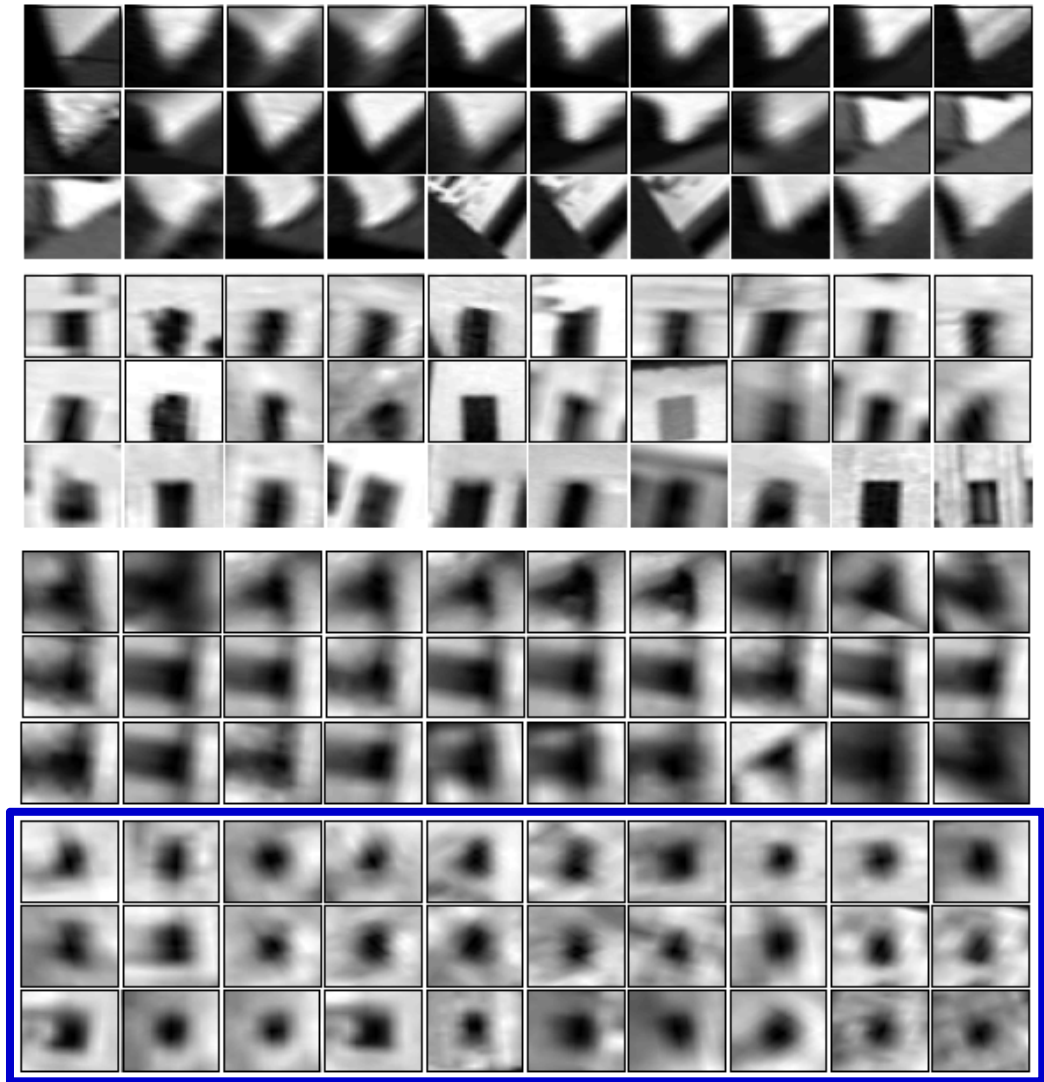
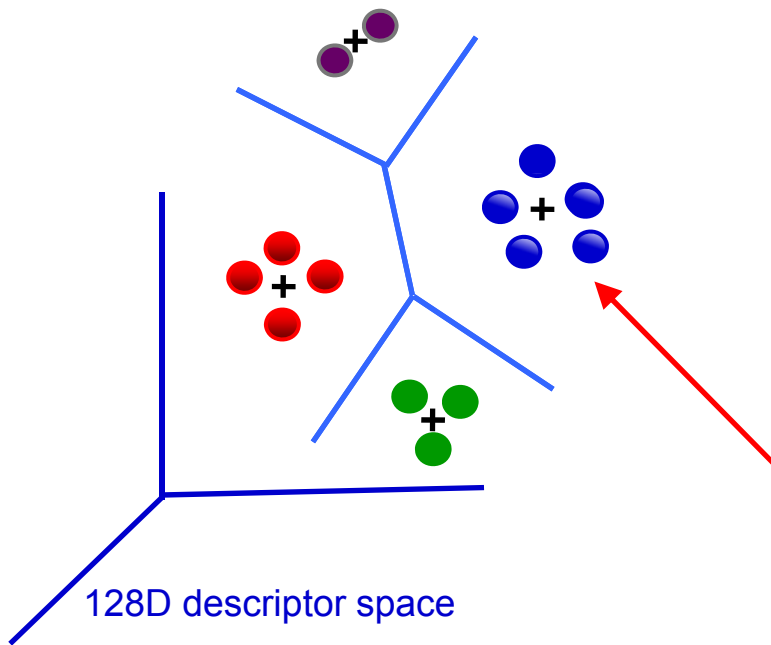
- Randomly initialize K cluster centers
- Iterate until convergence:
  - Assign each data point to the nearest center
  - Recompute each cluster center as the mean of all points assigned to it

Local minimum, solution dependent on initialization

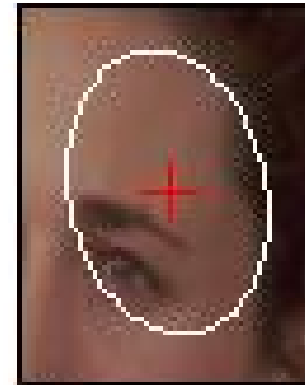
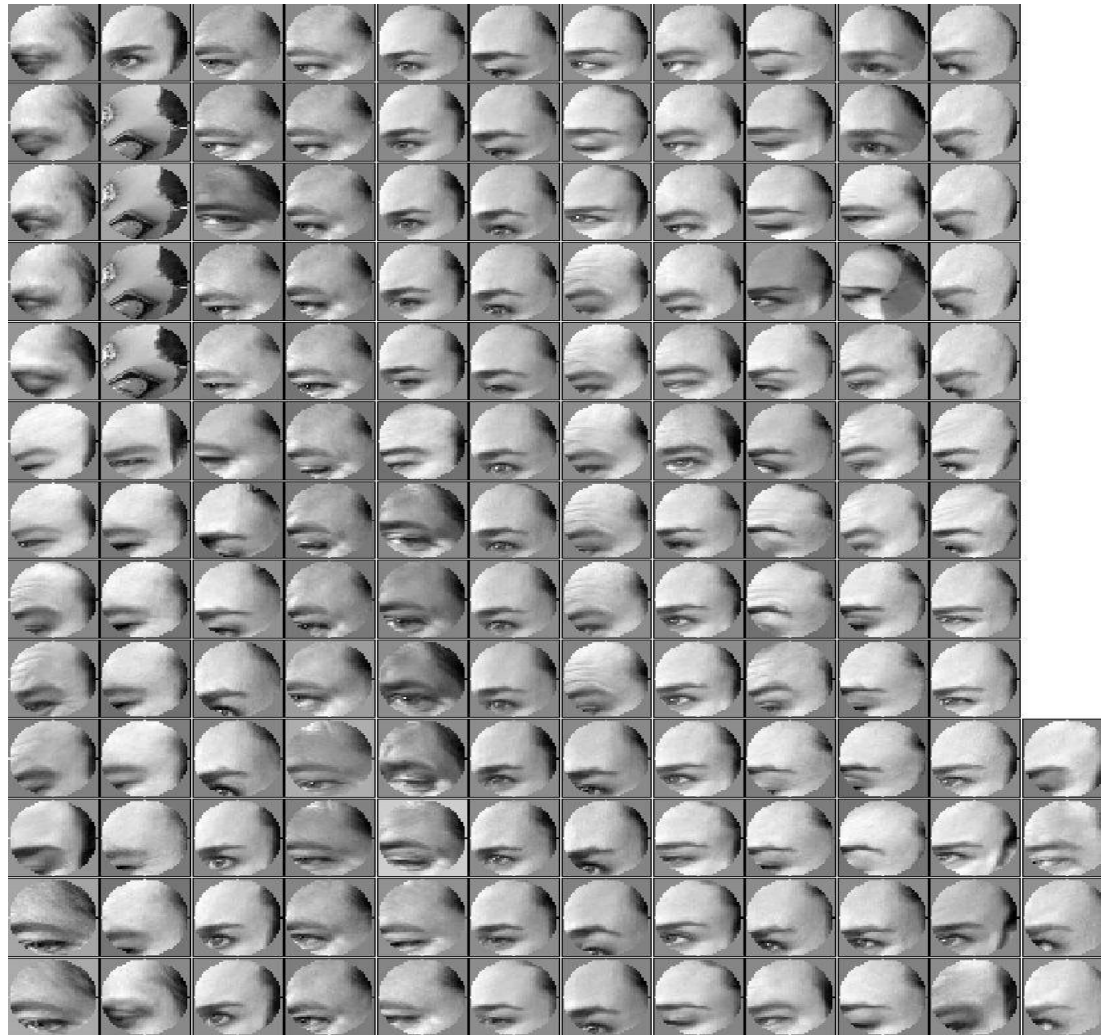
Initialization important, run several times, select best

# Visual words

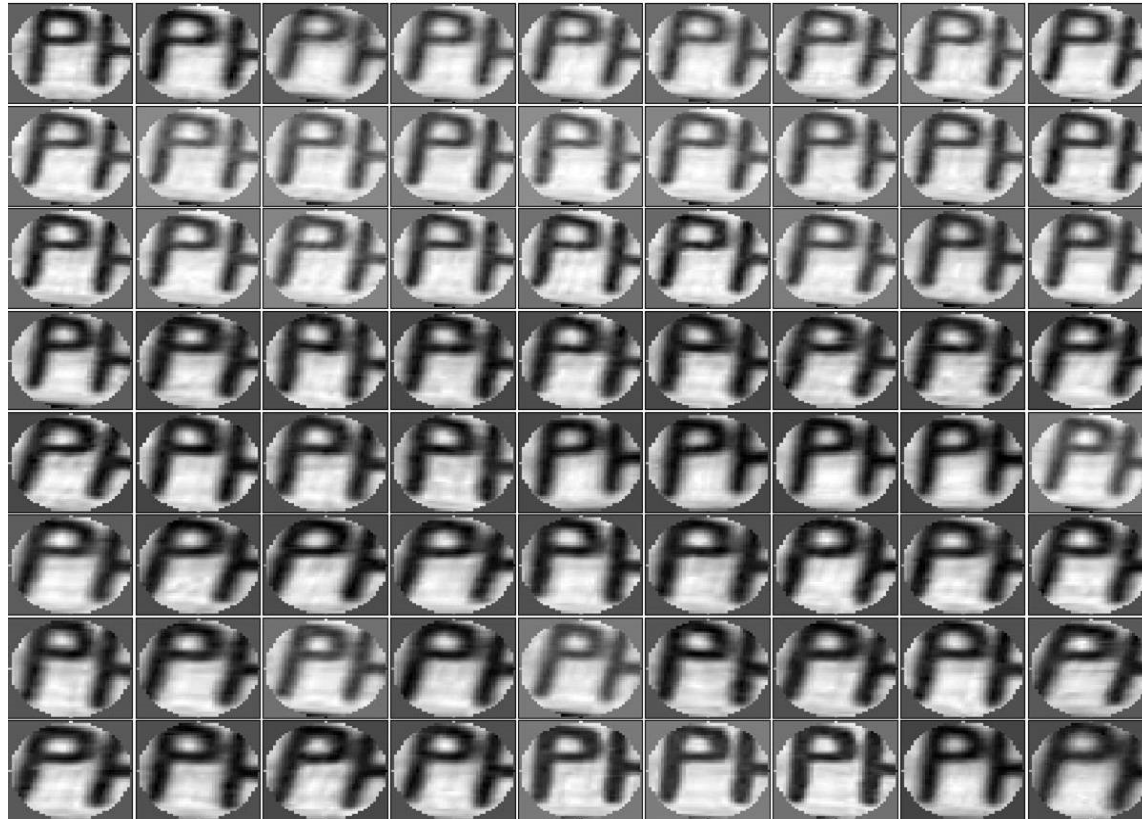
Example: each group of patches belongs to the same visual word



# Samples of visual words (clusters on SIFT descriptors):



Samples of visual words (clusters on SIFT descriptors):

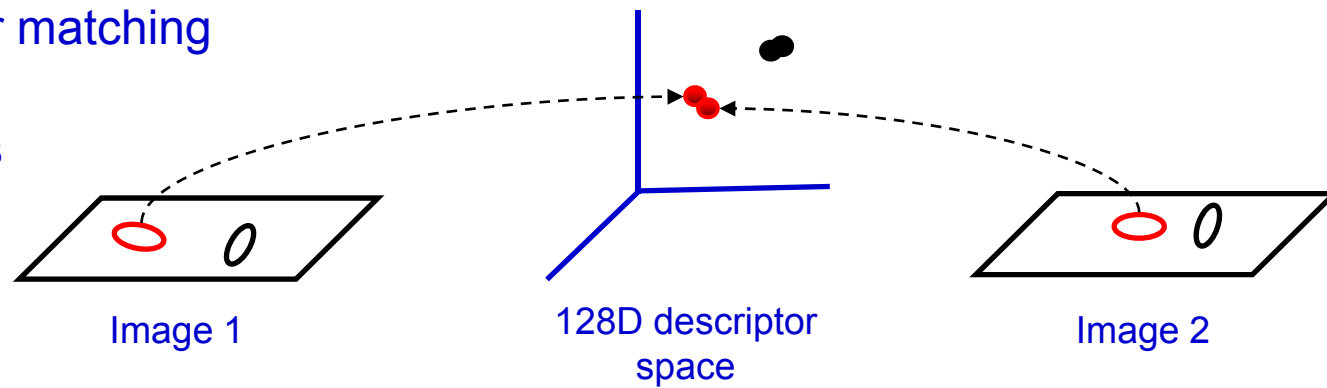


# Visual words: quantize descriptor space

Sivic and Zisserman, ICCV 2003

Nearest neighbour matching

- expensive to do for all frames

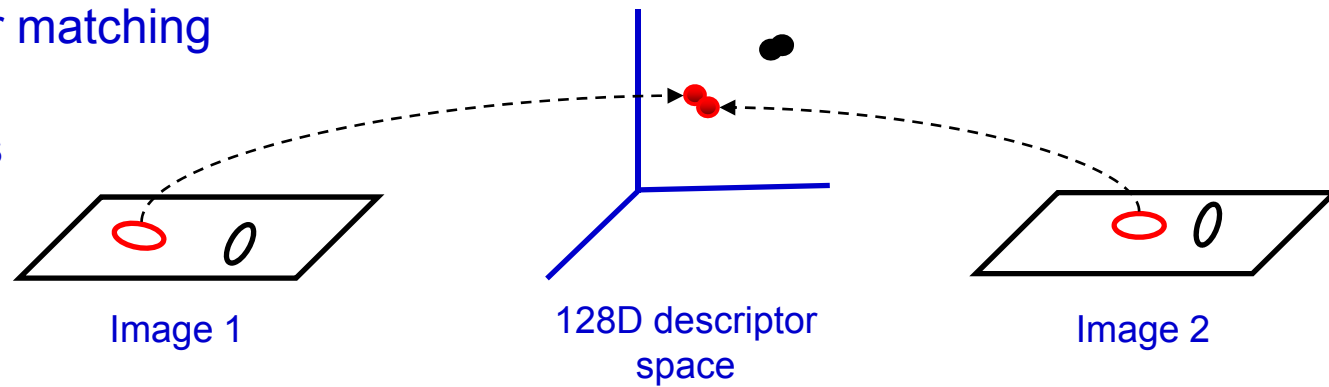


# Visual words: quantize descriptor space

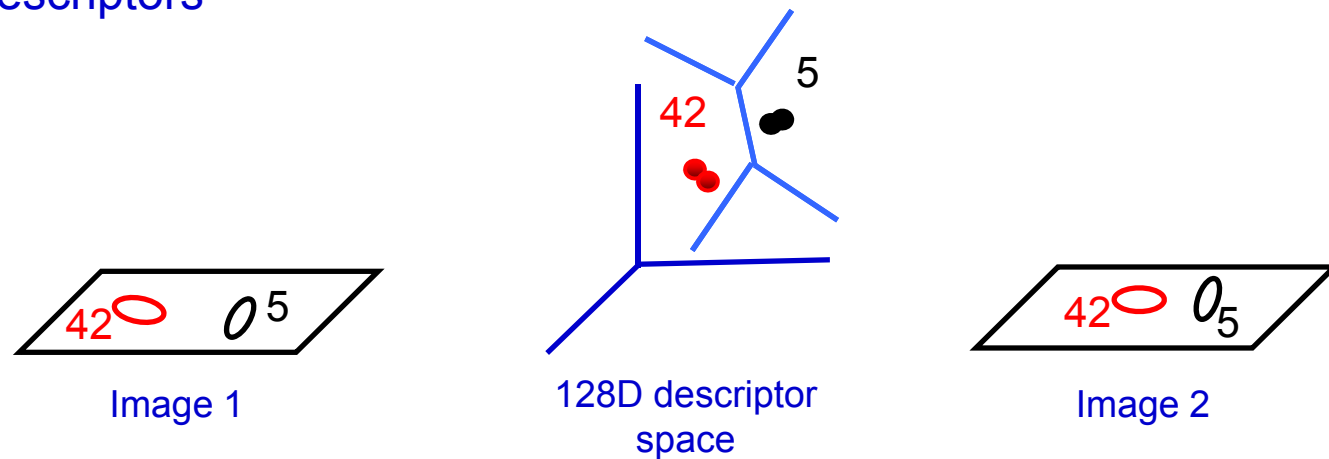
Sivic and Zisserman, ICCV 2003

## Nearest neighbour matching

- expensive to do for all frames



## Vector quantize descriptors



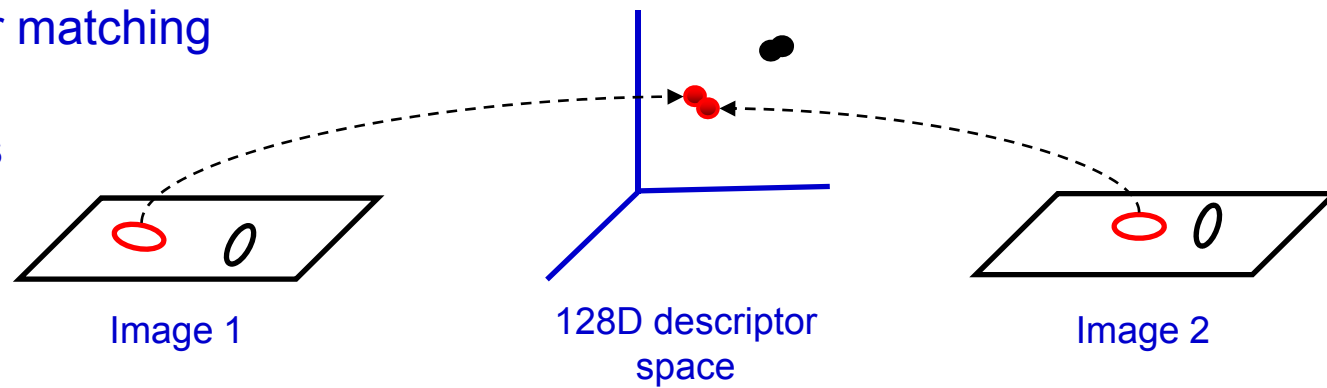


# Visual words: quantize descriptor space

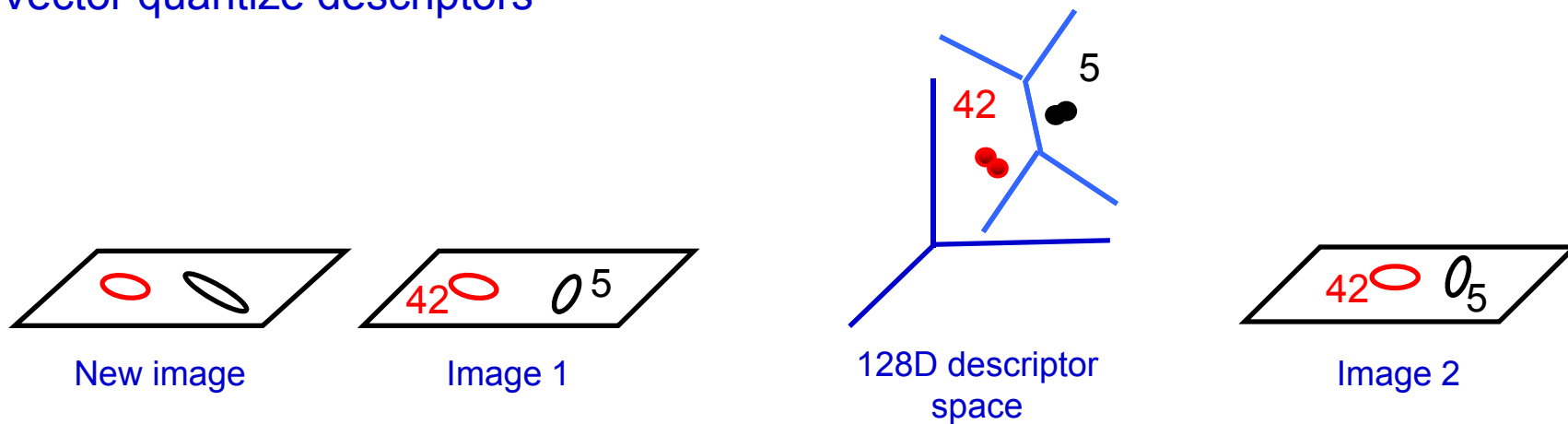
Sivic and Zisserman, ICCV 2003

## Nearest neighbour matching

- expensive to do for all frames



## Vector quantize descriptors

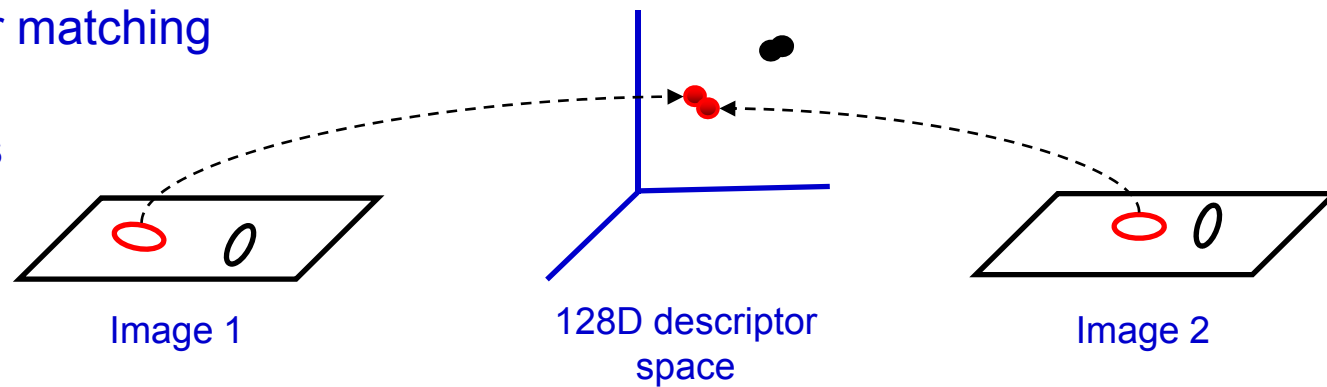


# Visual words: quantize descriptor space

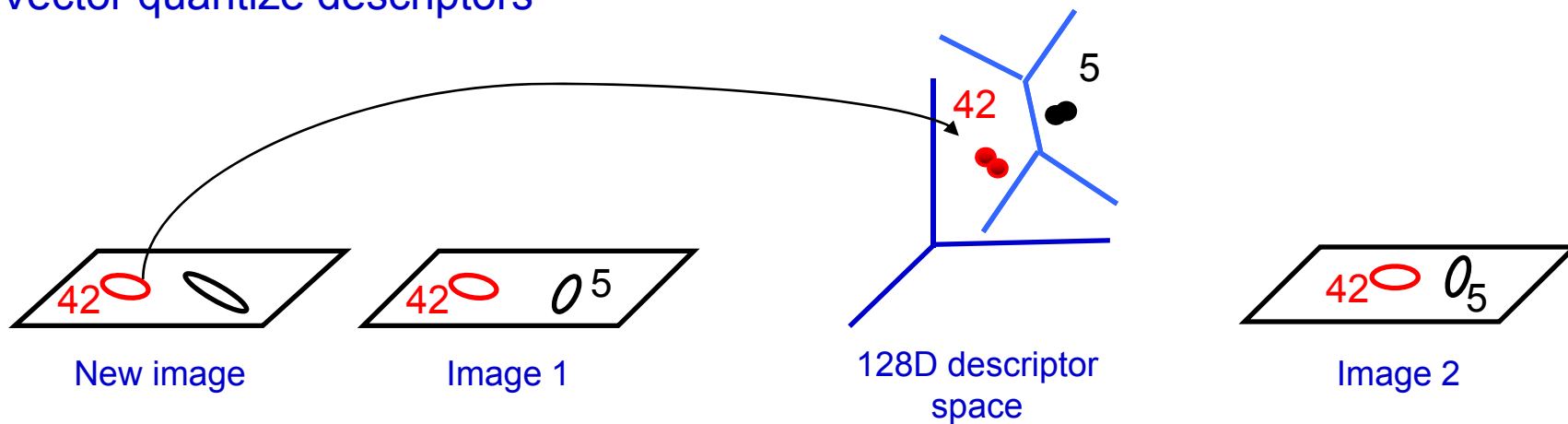
Sivic and Zisserman, ICCV 2003

## Nearest neighbour matching

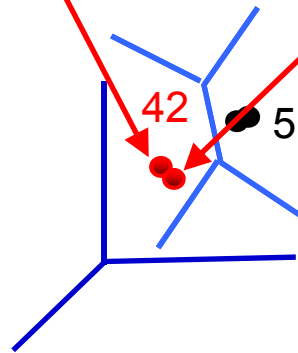
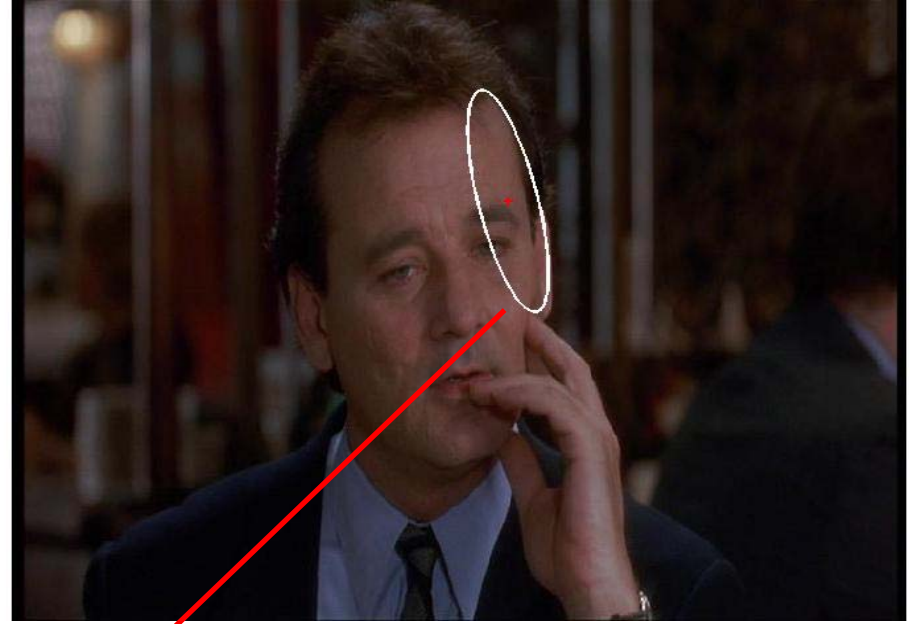
- expensive to do for all frames



## Vector quantize descriptors



# Vector quantize the descriptor space (SIFT)

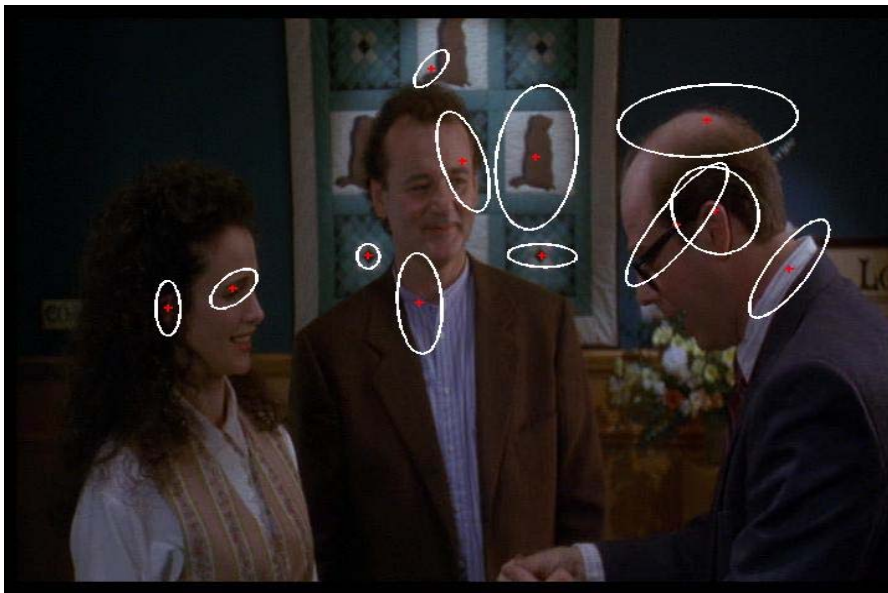


The same visual word

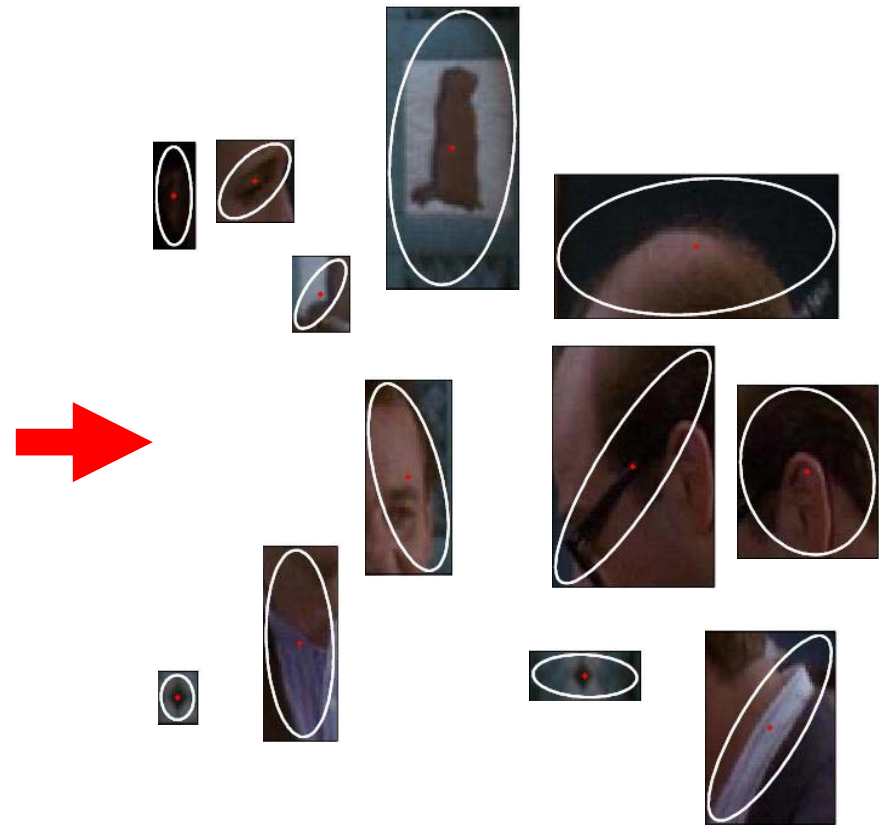
# Representation: bag of (visual) words

Visual words are 'iconic' image patches or fragments

- represent their frequency of occurrence
- but not their position

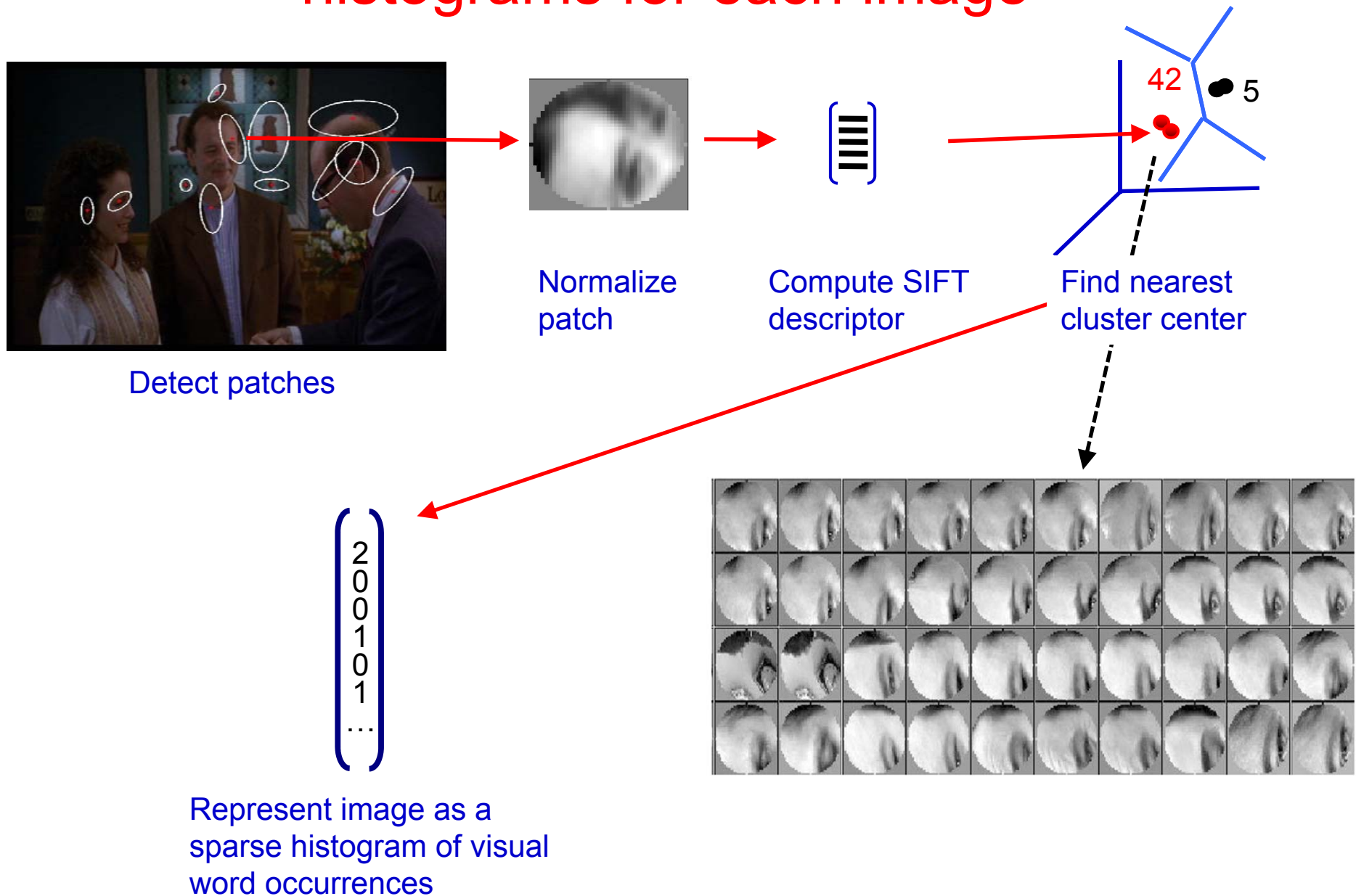


Image



Collection of visual words

# Offline: Assign visual words and compute histograms for each image



# Offline: create an index



frame #5



frame #10

Word number	Posting list
1	5, 10, ...
2	10, ...
...	...

- For fast search, store a “posting list” for the dataset
- This maps visual word occurrences to the images they occur in (i.e. like the “book index”)



# At run time



frame #5



frame #10

Word number	Posting list
1	5, 10, ...
2	10, ...
...	...

- User specifies a query region
- Generate a short-list of images using visual words in the region
  1. Accumulate all visual words within the query region
  2. Use “book index” to find other images with these words
  3. Compute similarity for images sharing at least one word

# At run time



frame #5



frame #10

Word number	Posting list
1	5, 10, ...
2	10, ...
...	...

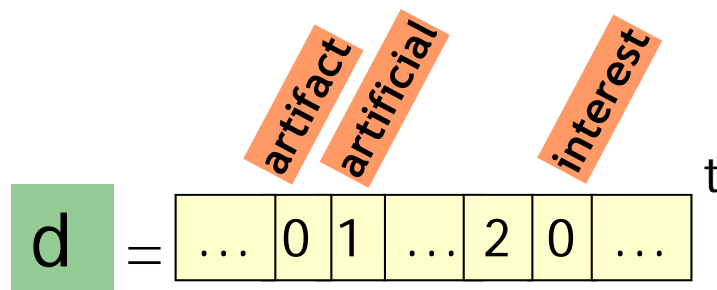
- Score each image by the (weighted) number of common visual words (tentative correspondences)
- Worst case complexity is linear in the number of images  $N$
- In practice, it is linear in the length of the lists ( $\ll N$ )



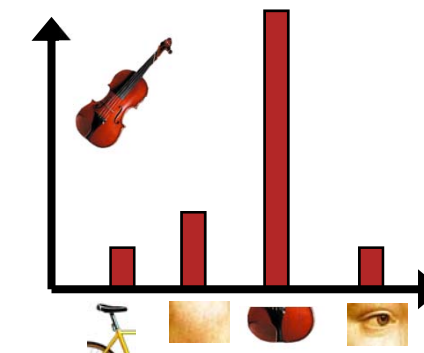
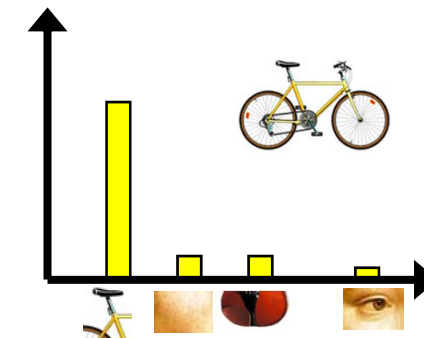
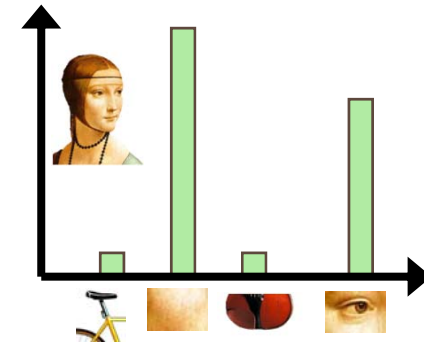
# Another interpretation: Bags of visual words

Summarize entire image based on its distribution (histogram) of visual word occurrences

Analogous to bag of words representation commonly used for text documents



Hofmann 2001



## Another interpretation: the bag-of-visual-words model

For a vocabulary of size  $K$ , each image is represented by a  $K$ -vector

$$\mathbf{v}_d = (t_1, \dots, t_i, \dots, t_K)^\top$$

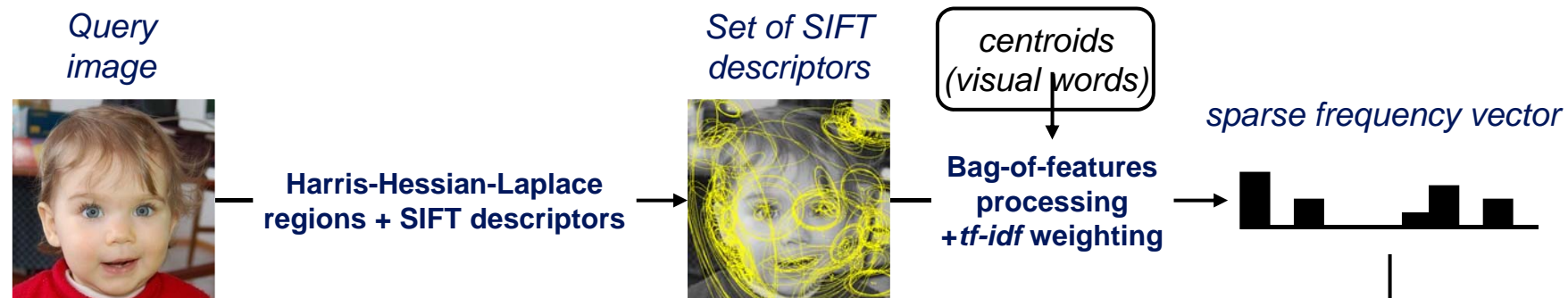
where  $t_i$  is the number of occurrences of visual word  $i$

Images are ranked by the normalized scalar product between the query vector  $\mathbf{v}_q$  and all vectors in the database  $\mathbf{v}_d$ :

$$f_d = \frac{\mathbf{v}_q^\top \mathbf{v}_d}{\|\mathbf{v}_q\|_2 \|\mathbf{v}_d\|_2}$$

Scalar product can be computed efficiently using inverted file

# Bag-of-features [Sivic&Zisserman'03]



## Results



Inverted file

querying

Re-ranked list

Geometric verification

ranked image short-list

[Chum & al. 2007]

# Geometric verification

---

Use the **position** and **shape** of the underlying features to improve retrieval quality



Both images have many matches – which is correct?

# Geometric verification

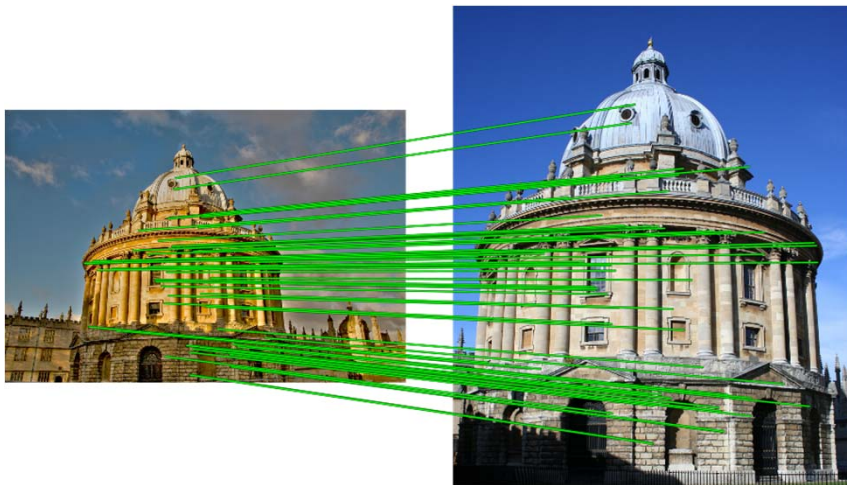
---

- Remove outliers, many matches are incorrect
- Estimate geometric transformation
- Robust strategies
  - RANSAC
  - Hough transform

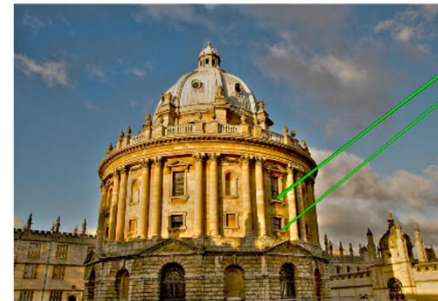
# Geometric verification

---

We can measure **spatial consistency** between the query and each result to improve retrieval quality, re-rank



Many spatially consistent matches – **correct result**

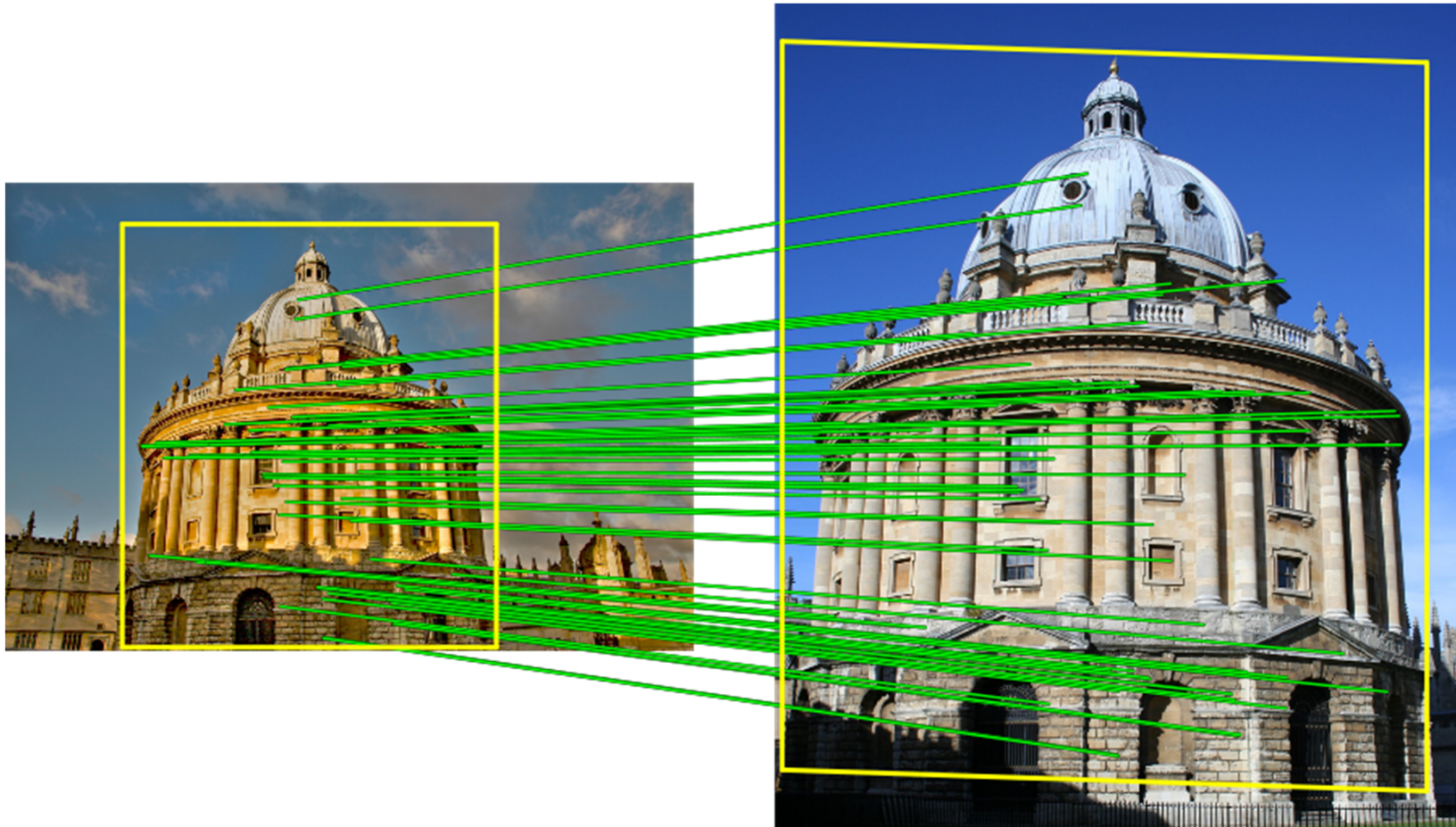


Few spatially consistent matches – **incorrect result**



# Geometric verification

Gives **localization** of the object



# Geometric verification – example

---

1. Query



2. Initial retrieval set (bag of words model)



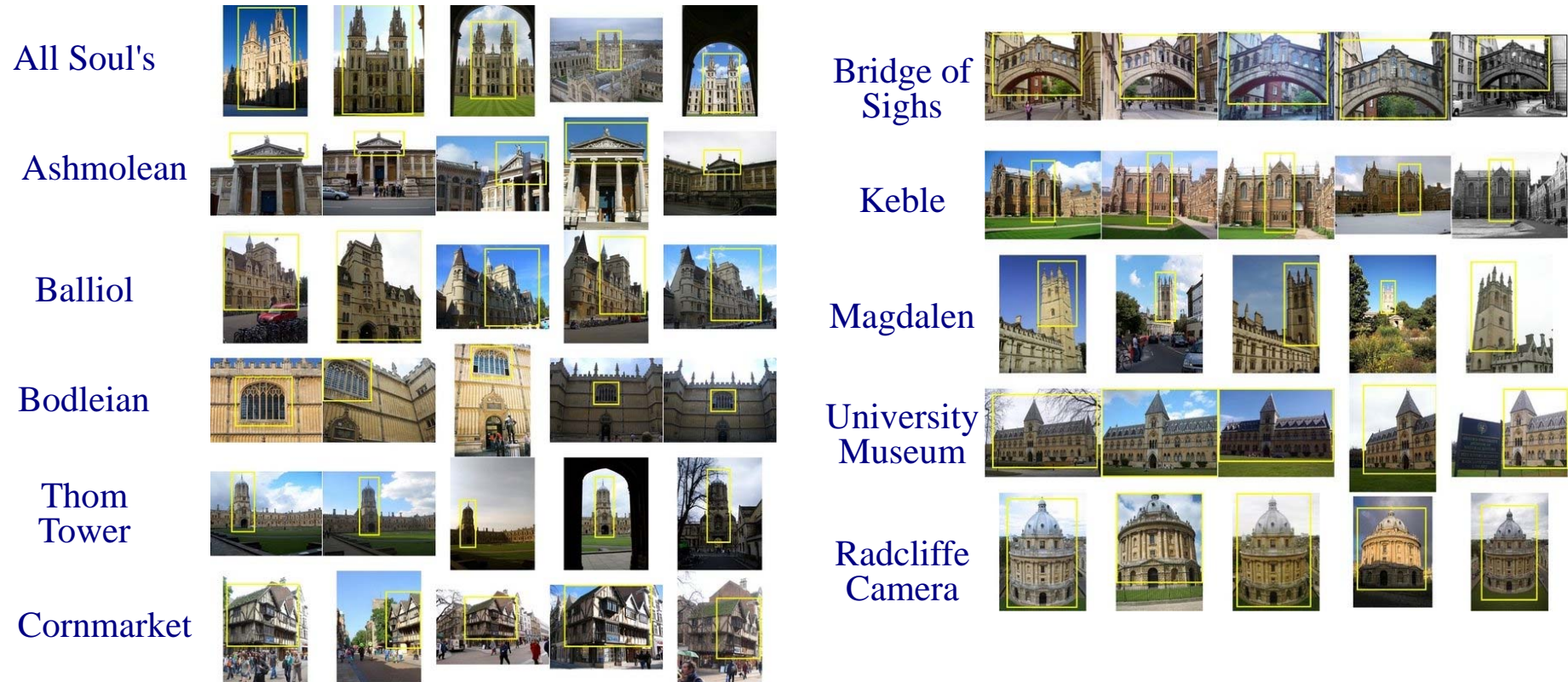
3. Spatial verification (re-rank on # of inliers)





# Evaluation dataset: Oxford buildings

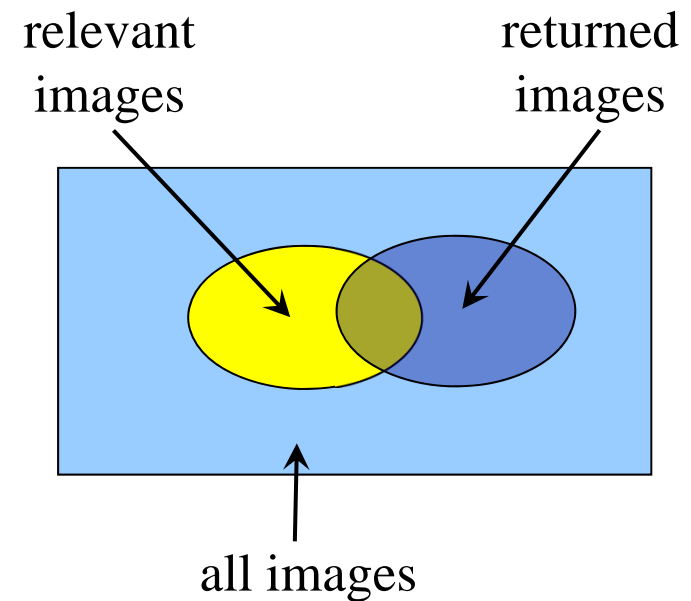
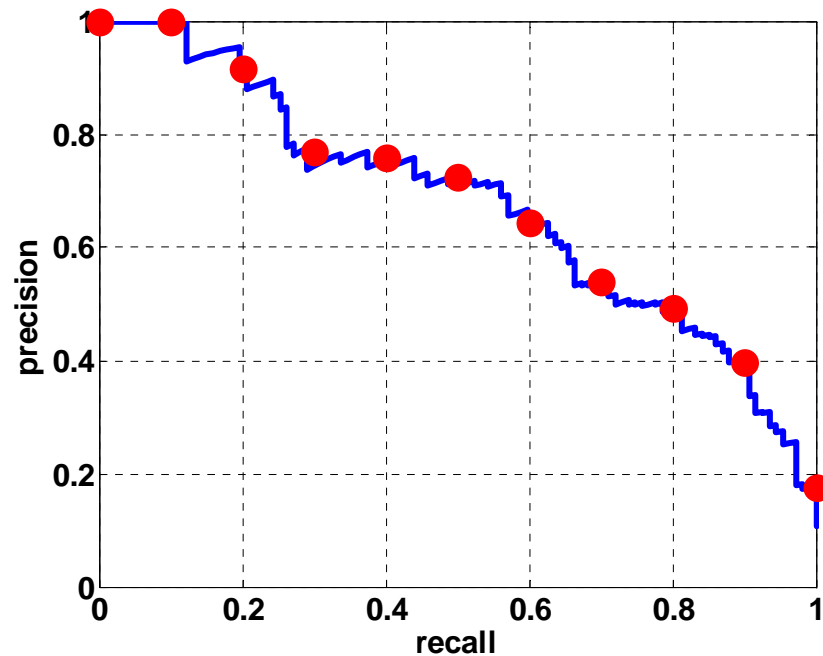
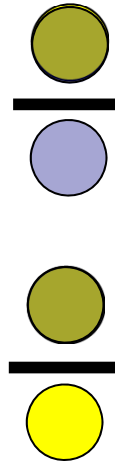
---



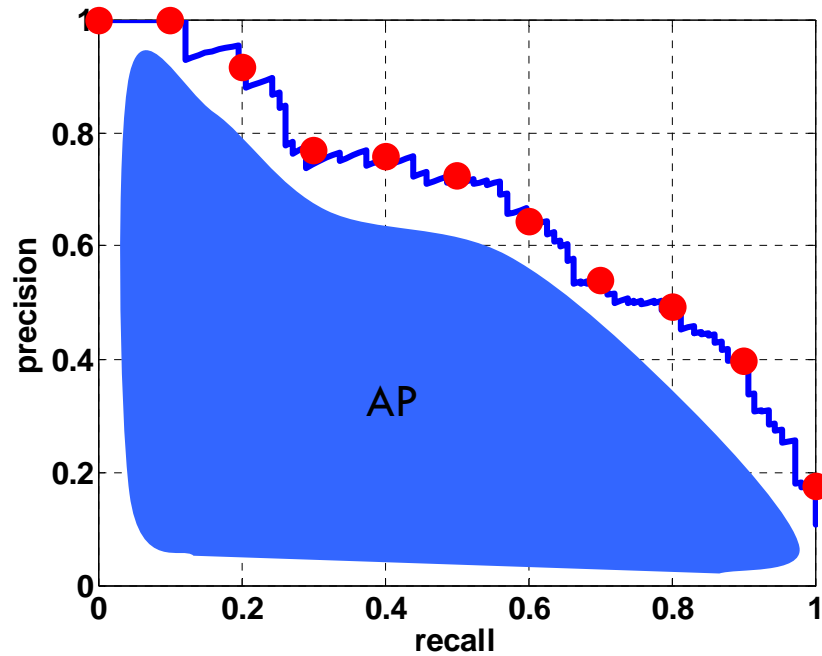
- Ground truth obtained for 11 landmarks
- Evaluate performance by mean Average Precision

# Measuring retrieval performance: Precision - Recall

- Precision: % of returned images that are relevant
- Recall: % of relevant images that are returned

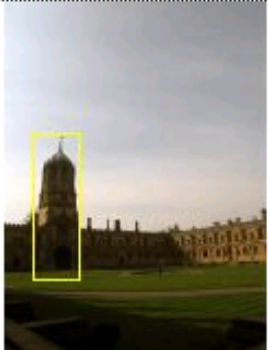


# Average Precision

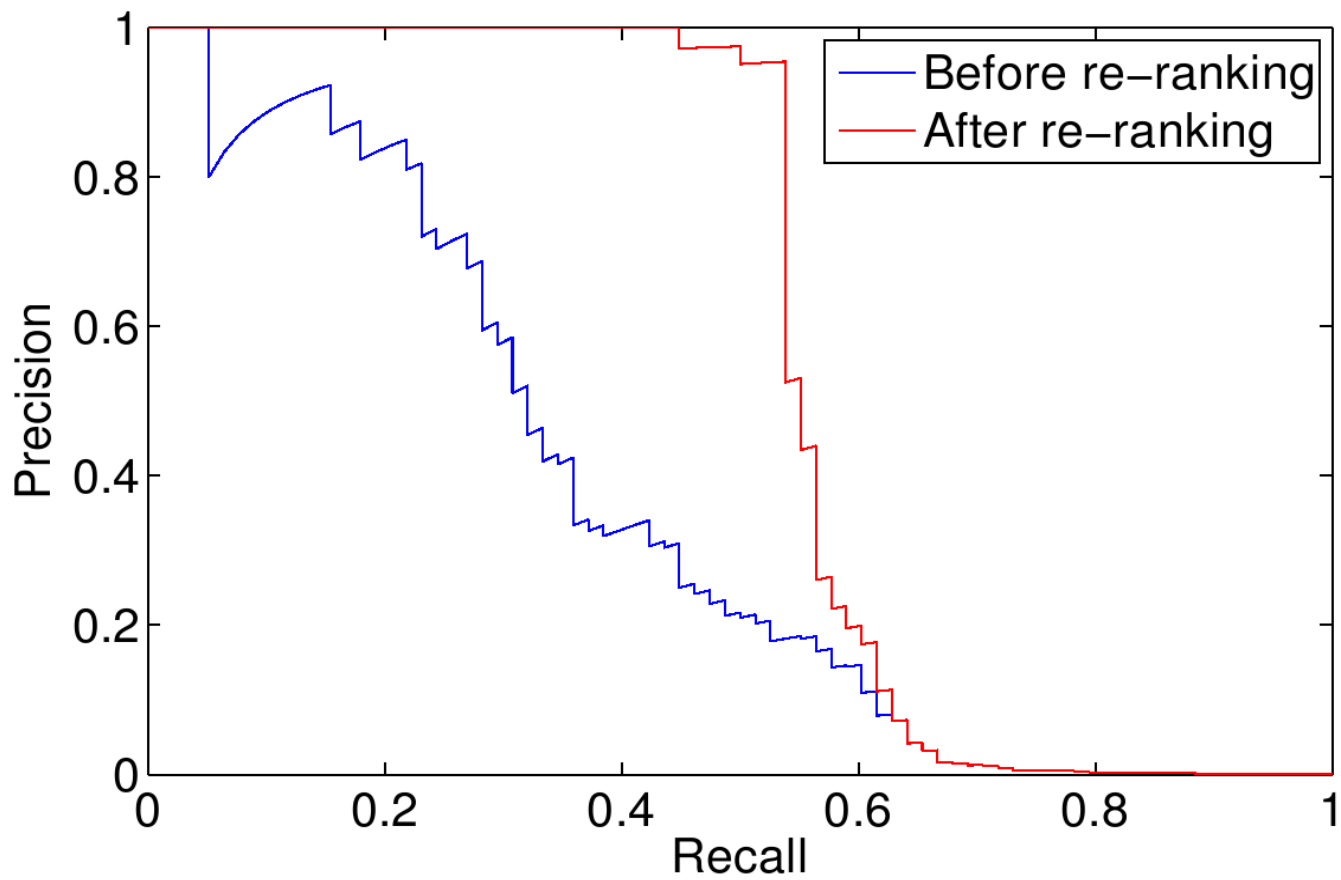


- A good AP score requires both high recall **and** high precision
- Application-independent

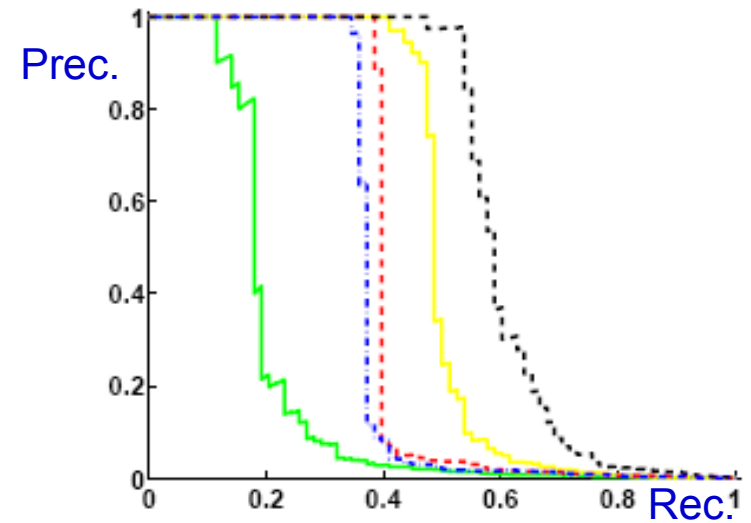
Performance measured by mean Average Precision (mAP)  
over 55 queries on 100K or 1.1M image datasets



Query: ChristChurch3

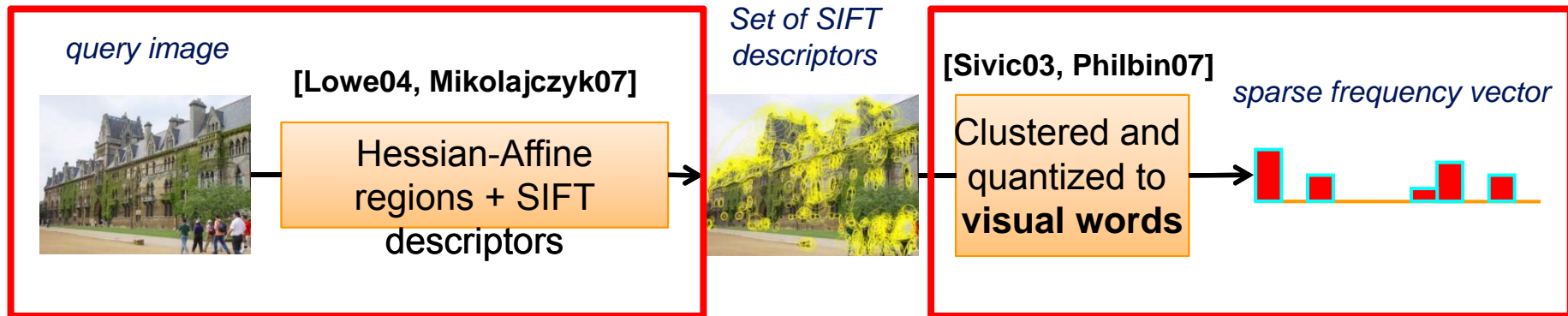


## Query images



- high precision at low recall (like google)
- variation in performance over queries
- does not retrieve all instances

# Why aren't all objects retrieved?



Obtaining visual words is like a sensor measuring the image

“noise” in the measurement process means that some visual words are missing or incorrect, e.g. due to

- Missed detections
- Changes beyond built in invariance
- Quantization effects

- 1. Query expansion
- 2. Better quantization

Consequence: Visual word in query is missing

# Query Expansion in text

## In text :

- Reissue top n responses as queries
- Blind relevance feedback
- Danger of topic drift

## In vision:

- Reissue **spatially verified** image regions as queries

# Automatic query expansion

Visual word representations of two images of the same object may differ (due to e.g. detection/quantization noise) resulting in missed returns

Initial returns may be used to add new relevant visual words to the query

Strong spatial model prevents 'drift' by discarding false positives

[Chum, Philbin, Sivic, Isard, Zisserman, ICCV'07;

Chum, Mikulik, Perdoch, Matas, CVPR'11]



# Visual query expansion - overview

1. Original query



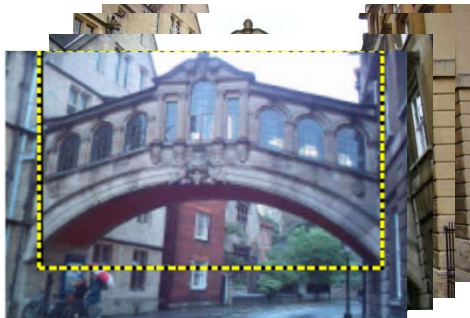
2. Initial retrieval set



3. Spatial verification



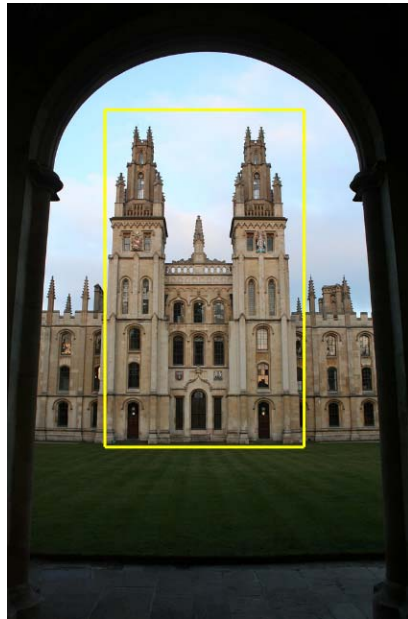
4. New enhanced query



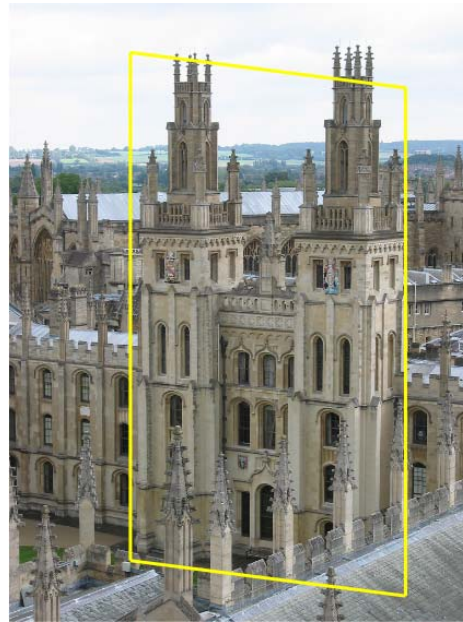
5. Additional retrieved images



# Query Expansion



Query Image

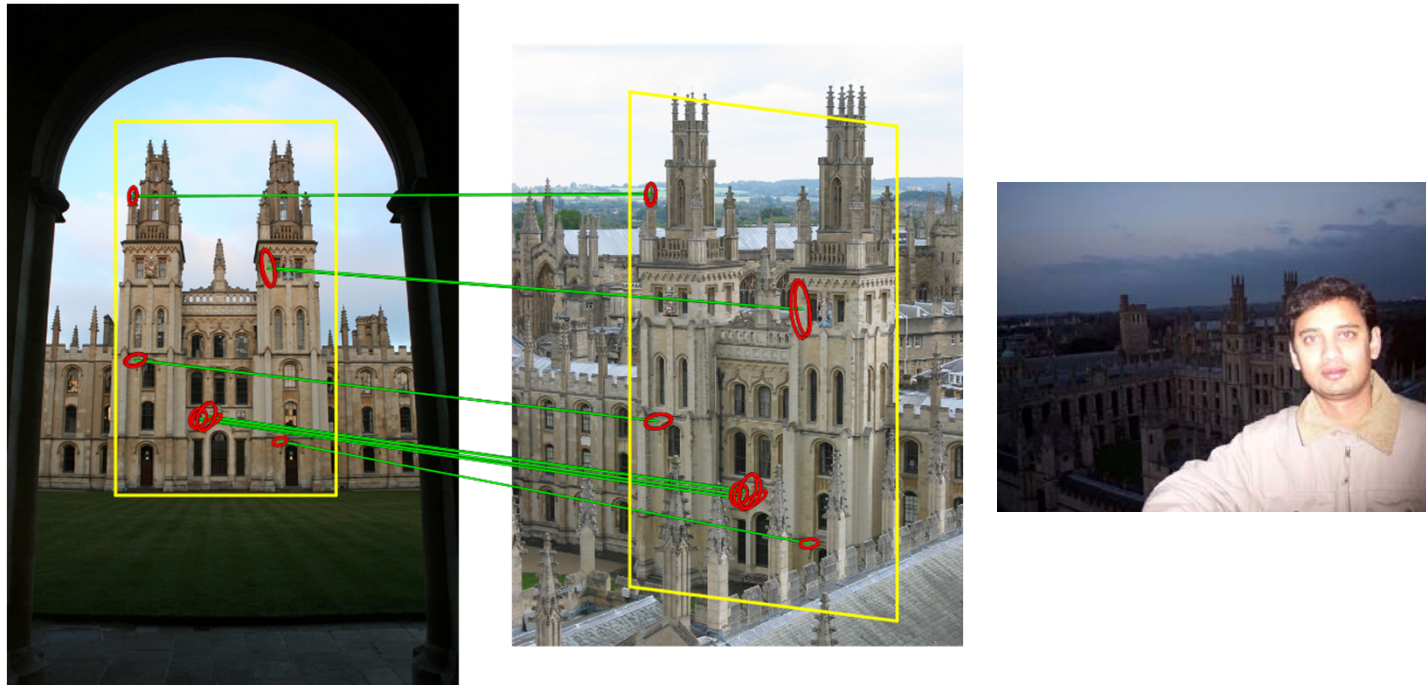


Originally retrieved image



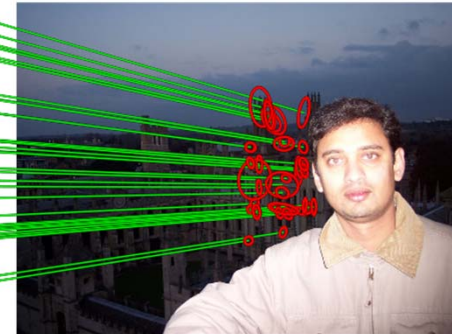
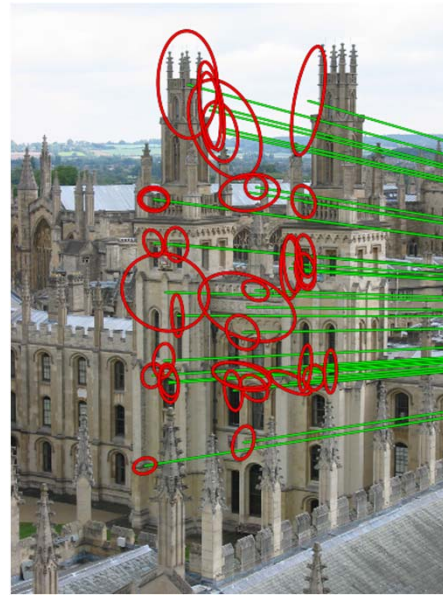
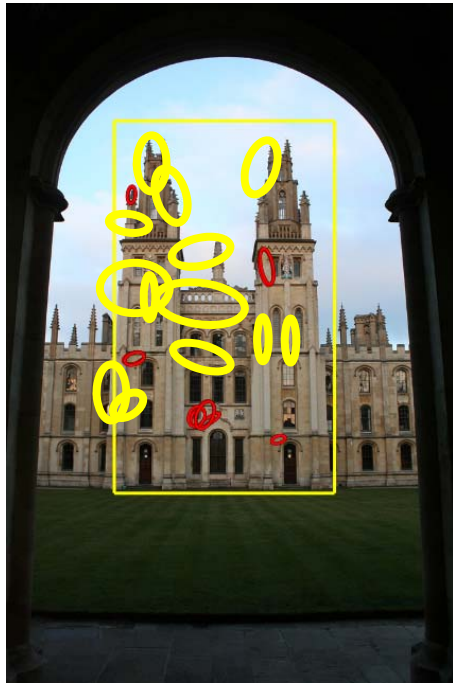
Originally not retrieved

# Query Expansion

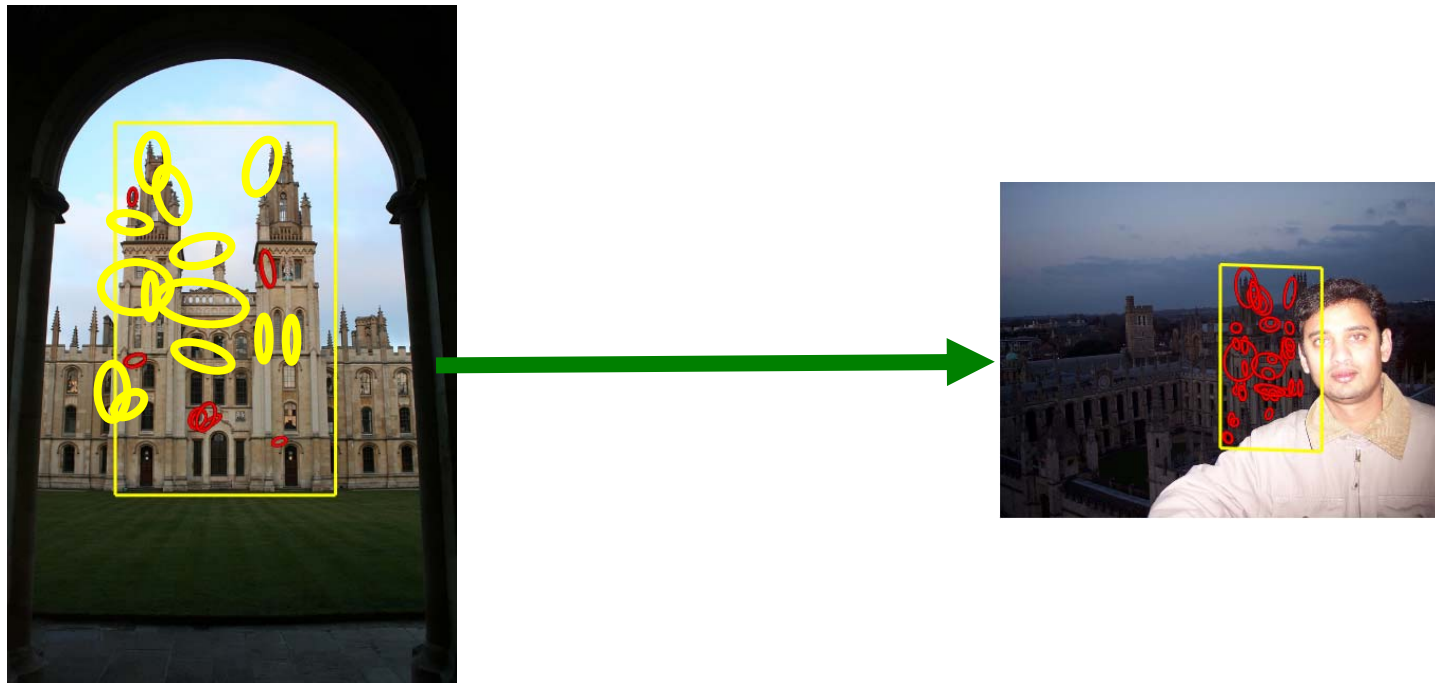




# Query Expansion

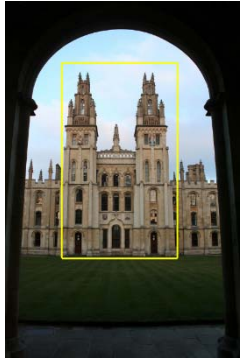


# Query Expansion

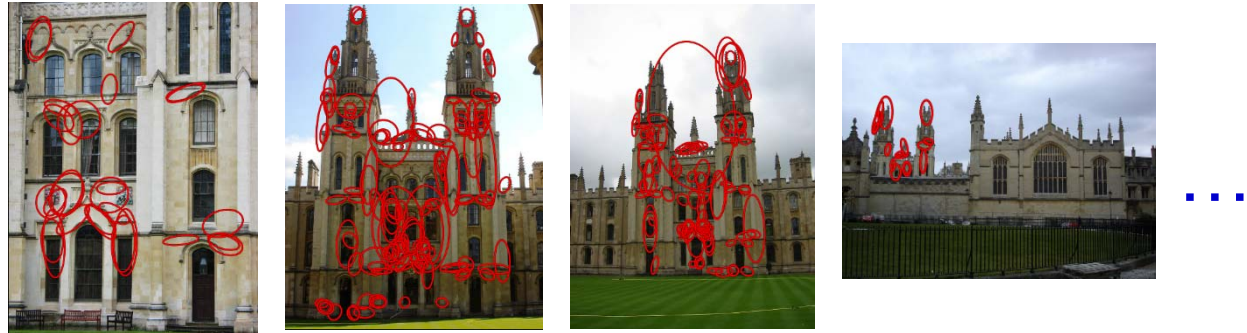


# Query Expansion

Query Image



Spatially verified retrievals with matching regions overlaid



New expanded query

New expanded query is formed as

- the average of visual word vectors of spatially verified returns
- only inliers are considered
- regions are back-projected to the original query image



# Query Expansion

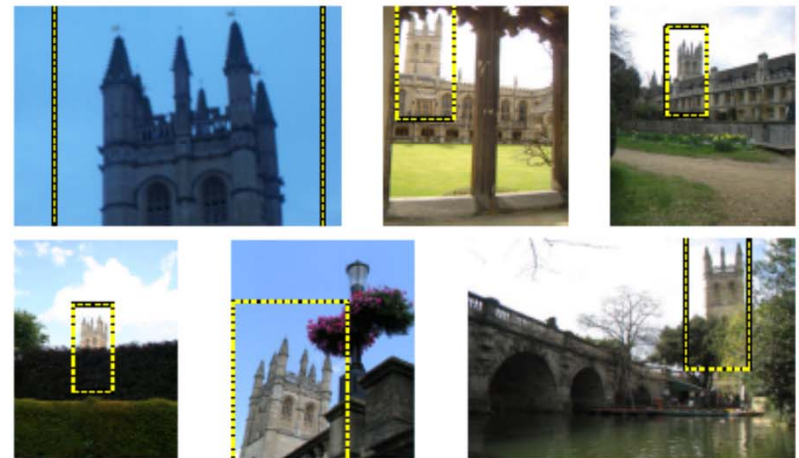
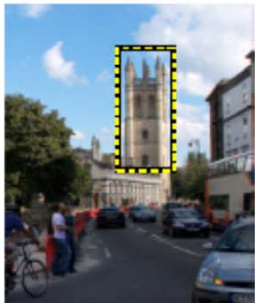
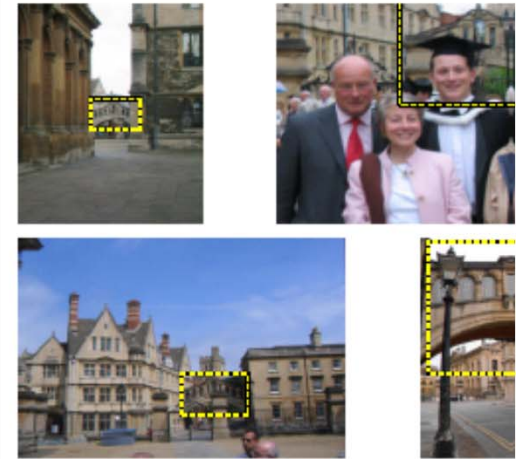
Query image



Originally retrieved



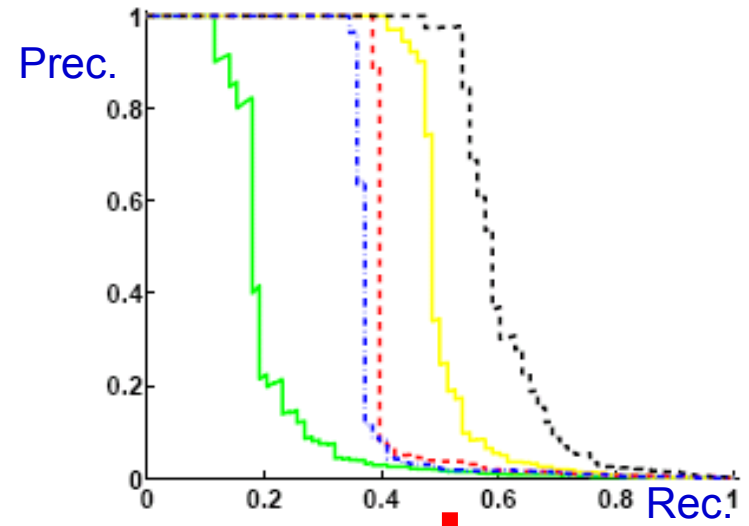
Retrieved only after expansion



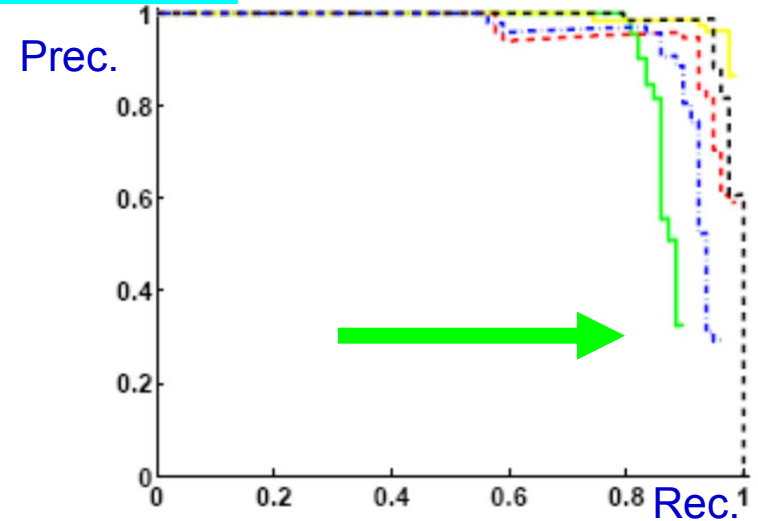


Query image

### Original results



### Expanded results (improved)

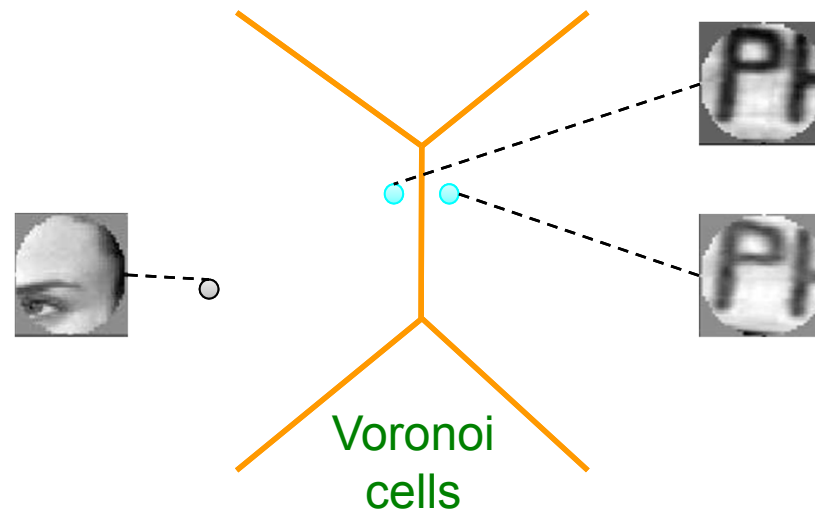




# Quantization errors

Typically, quantization has a significant impact on the final performance of the system [Sivic03,Nister06,Philbin07]

Quantization errors split features that should be grouped together and confuse features that should be separated



# Visual words – approximate NN search

---

- Map descriptors to words by quantizing the feature space
  - Quantize via k-means clustering to obtain visual words
  - Assign descriptors to closest visual words
- Bag-of-features as approximate nearest neighbor search

Descriptor matching with  $k$ -nearest neighbors

$$f_{k\text{-NN}}(x, y) = \begin{cases} 1 & \text{if } x \text{ is a } k\text{-NN of } y \\ 0 & \text{otherwise} \end{cases}$$

Bag-of-features matching function  $f_q(x, y) = \delta_{q(x), q(y)}$

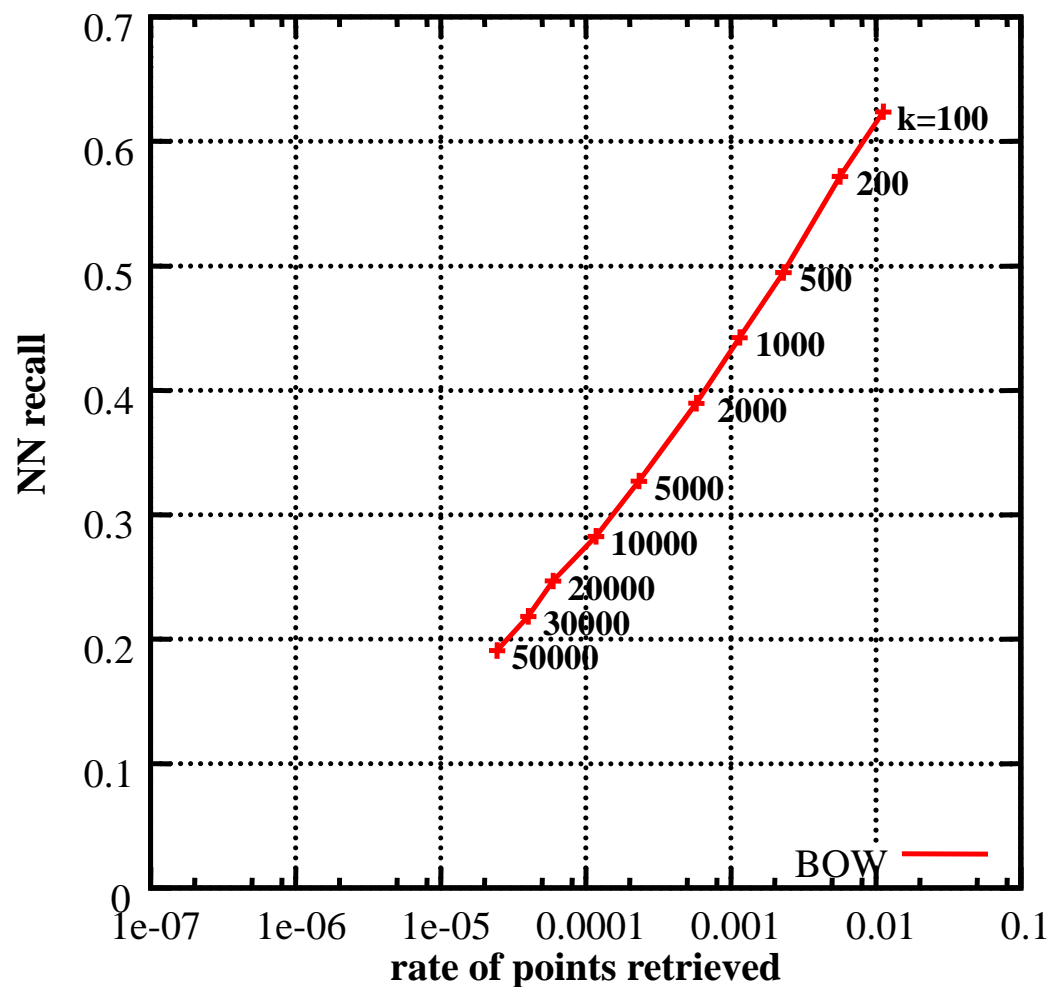
where  $q(x)$  is a quantizer, i.e., assignment to a visual word and  $\delta_{a,b}$  is the Kronecker operator ( $\delta_{a,b}=1$  iff  $a=b$ )

# Approximate nearest neighbor search evaluation

---

- ANN algorithms usually returns a short-list of nearest neighbors
  - this short-list is supposed to contain the NN with high probability
  - exact search may be performed to re-order this short-list
- Proposed quality evaluation of ANN search: trade-off between
  - **NN recall** = probability that *the* NN is in this list
  - against*
  - **NN precision** = proportion of vectors in the short-list
    - the lower this proportion
      - the more information we have about the vector
      - the lower the complexity if we perform exact search on the short-list
- ANN search algorithms usually have some parameters to handle this trade-off

# ANN evaluation of bag-of-features



- ANN algorithms returns a list of potential neighbors

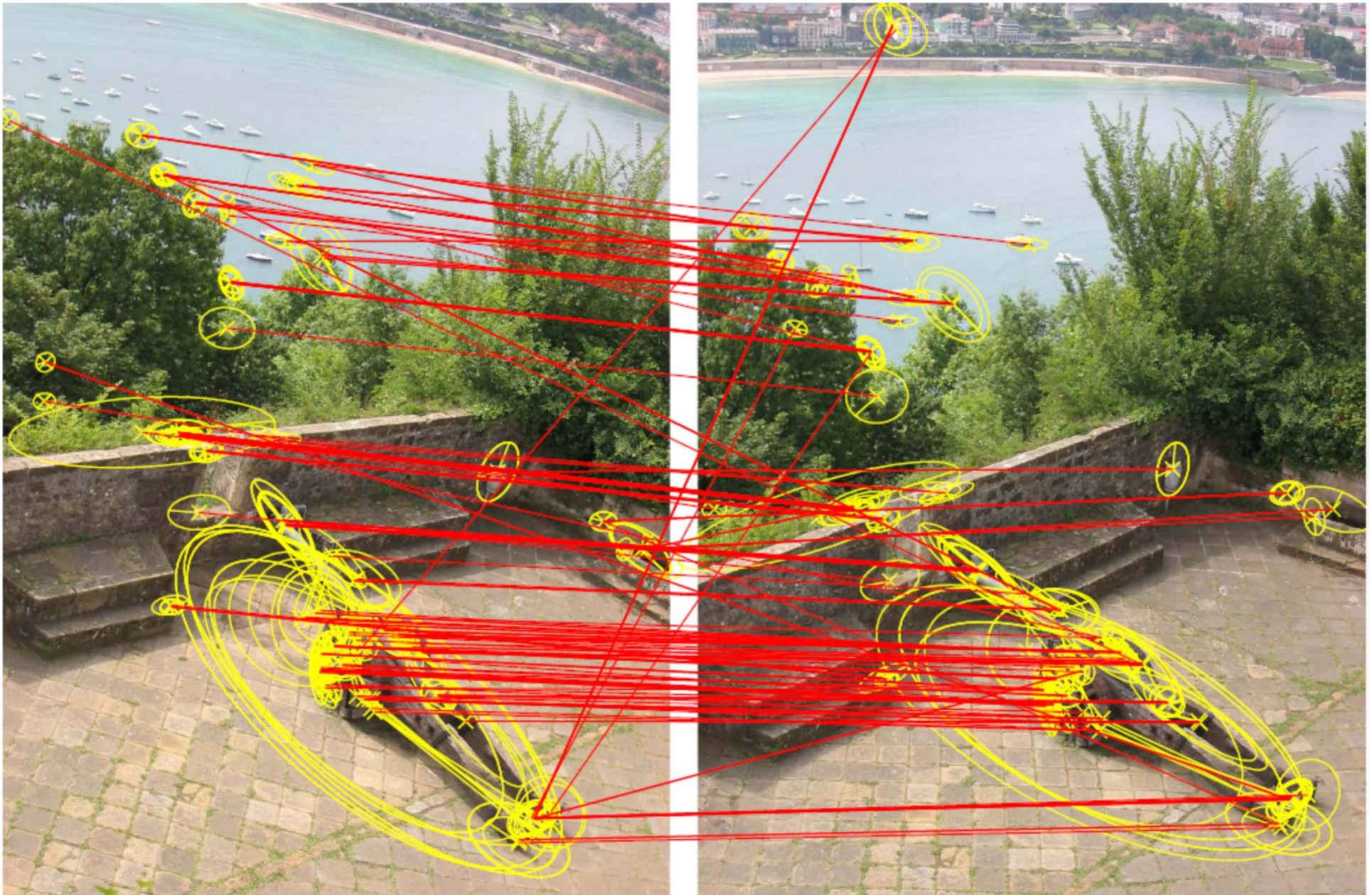
- NN recall**  
= probability that *the* NN is in this list

- NN precision:**  
= proportion of vectors in the short-list

- In BOF, this trade-off is managed by the number of clusters  $k$

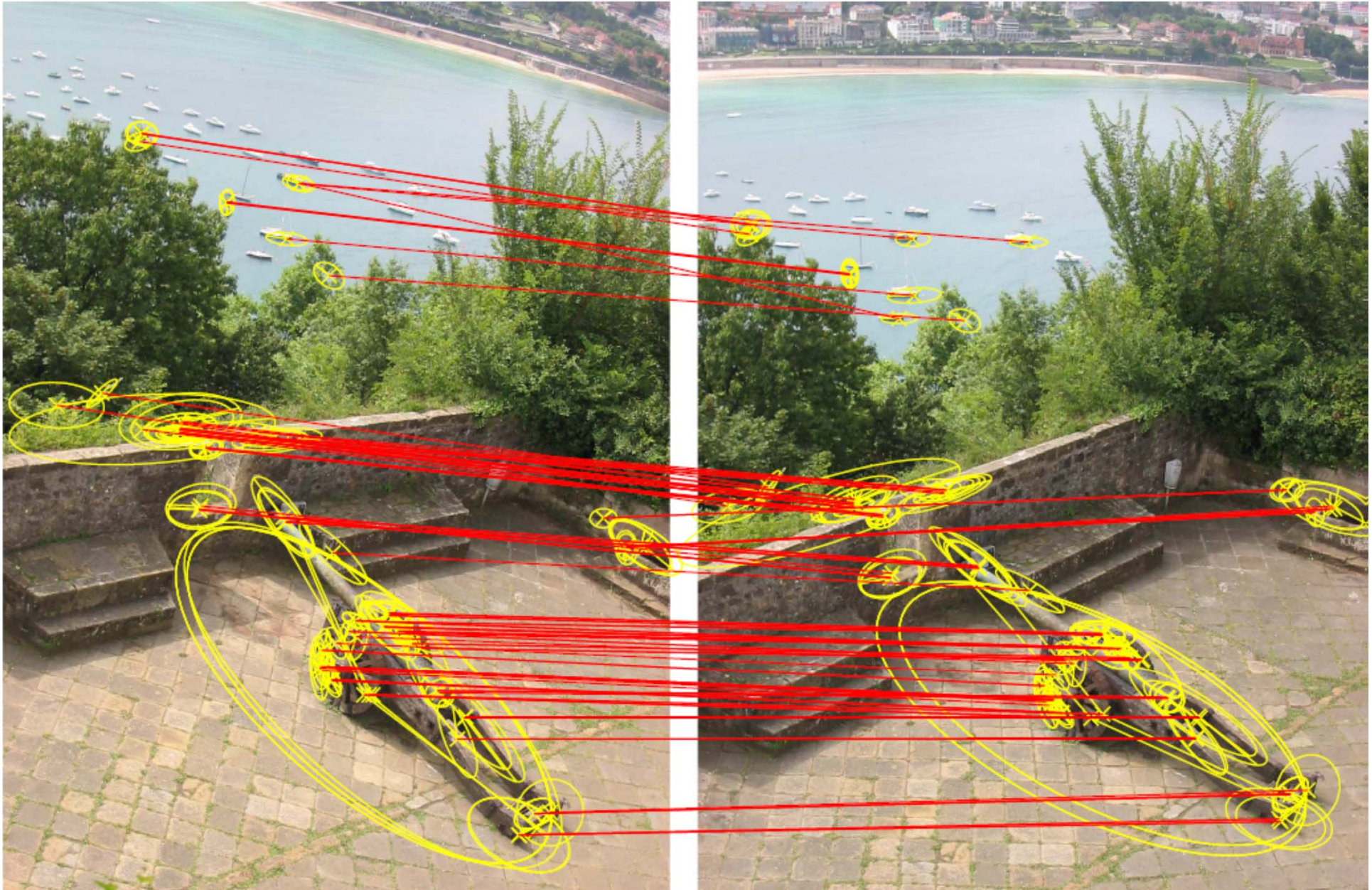


# 20K visual word: false matches





# 200K visual word: good matches missed



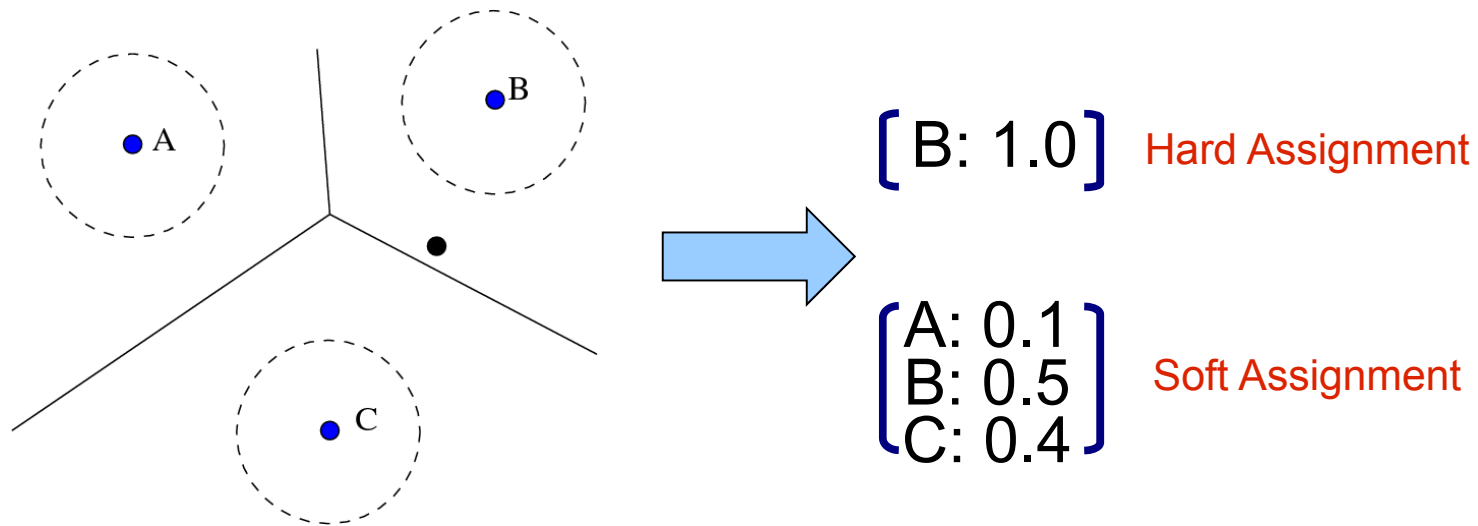
# Problem with bag-of-features

---

- The matching performed by BOF is weak
  - for a “small” visual dictionary: too many false matches
  - for a “large” visual dictionary: many true matches are missed
- No good trade-off between “small” and “large” !
  - either the Voronoi cells are too big
  - or these cells can’t absorb the descriptor noise
  - intrinsic approximate nearest neighbor search of BOF is not sufficient
  - possible solutions
    - soft assignment [Philbin et al. CVPR’08]
    - additional short codes [Jegou et al. ECCV’08]

# Beyond bags-of-visual-words

- Soft-assign each descriptor to multiple cluster centers [Philbin et al. 2008, Van Gemert et al. 2008]

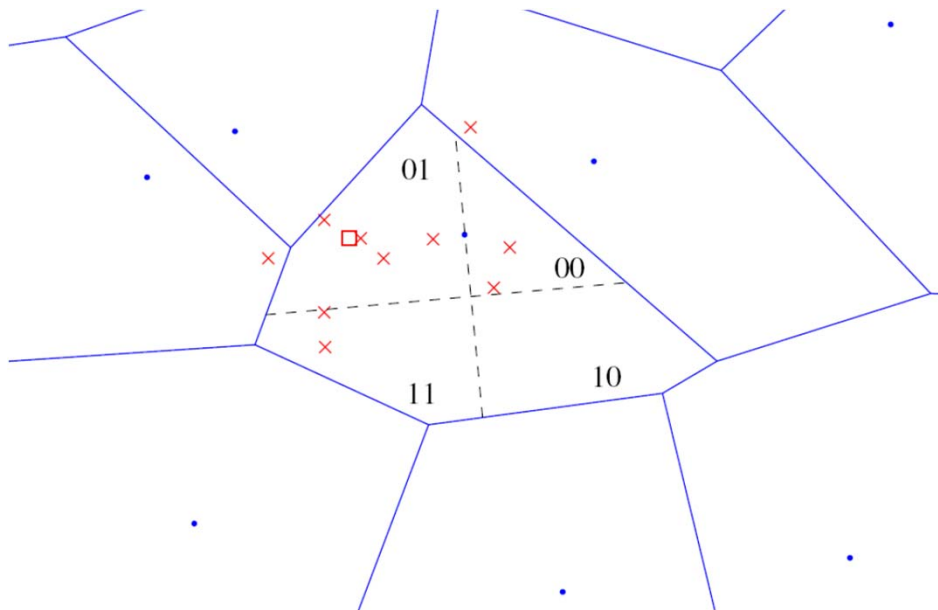




# Beyond bag-of-visual-words

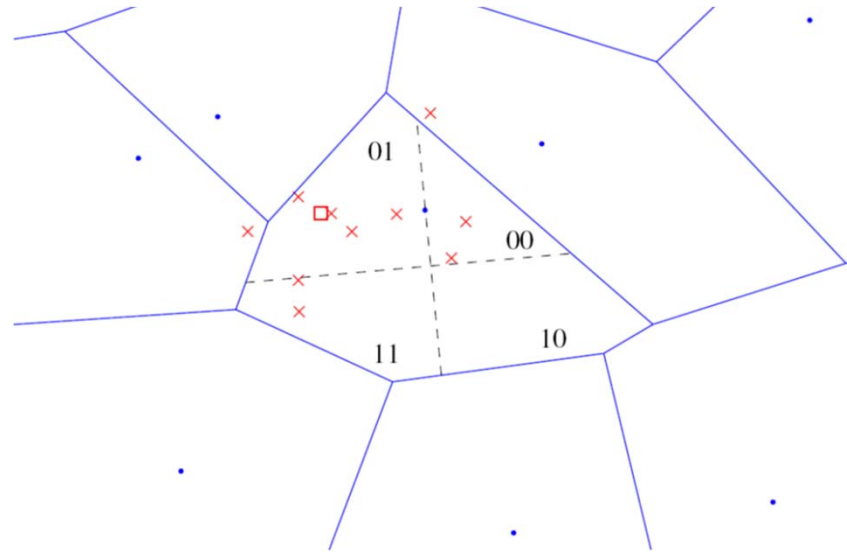
## Hamming embedding [Jegou et al. 2008]

- Standard quantization using bag-of-visual-words
- Additional localization in the Voronoi cell by a binary signature



# Hamming Embedding

---



Representation of a descriptor  $x$

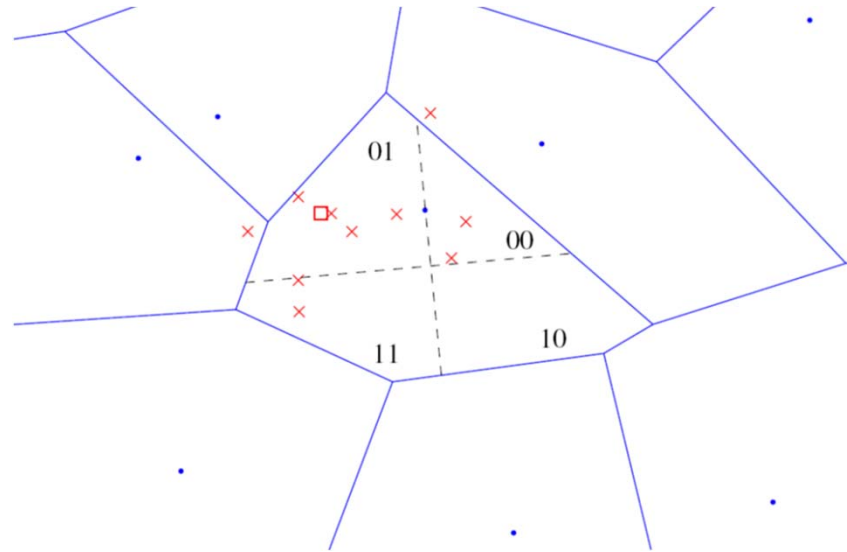
- Vector-quantized to  $q(x)$  as in standard BOF
- + **short binary vector  $b(x)$**  for an additional localization in the Voronoi cell

Two descriptors  $x$  and  $y$  match iff

$$f_{\text{HE}}(x, y) = \begin{cases} (\text{tf-idf}(q(x)))^2 & \text{if } q(x) = q(y) \\ & \text{and } h(b(x), b(y)) \leq h_t \quad \text{where } h(a, b) \text{ Hamming distance} \\ 0 & \text{otherwise} \end{cases}$$

# Hamming Embedding

---



- Nearest neighbors for Hamming distance  $\approx$  those for Euclidean distance  
→ a metric in the embedded space reduces dimensionality curse effects
- Efficiency
  - Hamming distance = very few operations
  - Fewer random memory accesses: 3 x faster than BOF with same dictionary size!

# Hamming Embedding

---

- **Off-line** (given a quantizer)

- draw an orthogonal projection matrix  $P$  of size  $d_b \times d$

- this defines  $d_b$  random projection directions

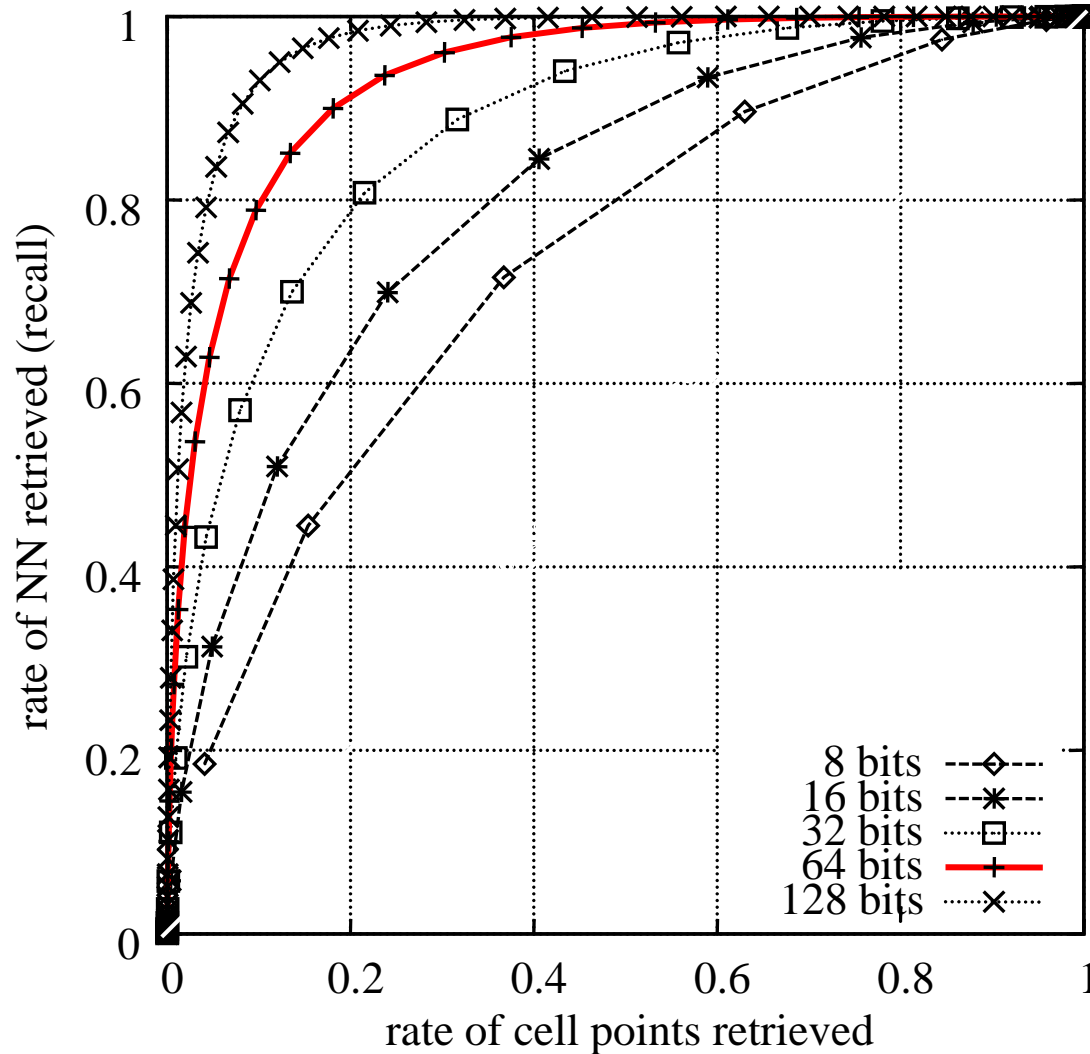
- for each Voronoi cell and projection direction, compute the median value for a training set

- **On-line:** compute the binary signature  $b(x)$  of a given descriptor

- project  $x$  onto the projection directions as  $z(x) = (z_1, \dots, z_{d_b})$

- $b_i(x) = 1$  if  $z_i(x)$  is above the learned median value, otherwise 0

# Hamming neighborhood

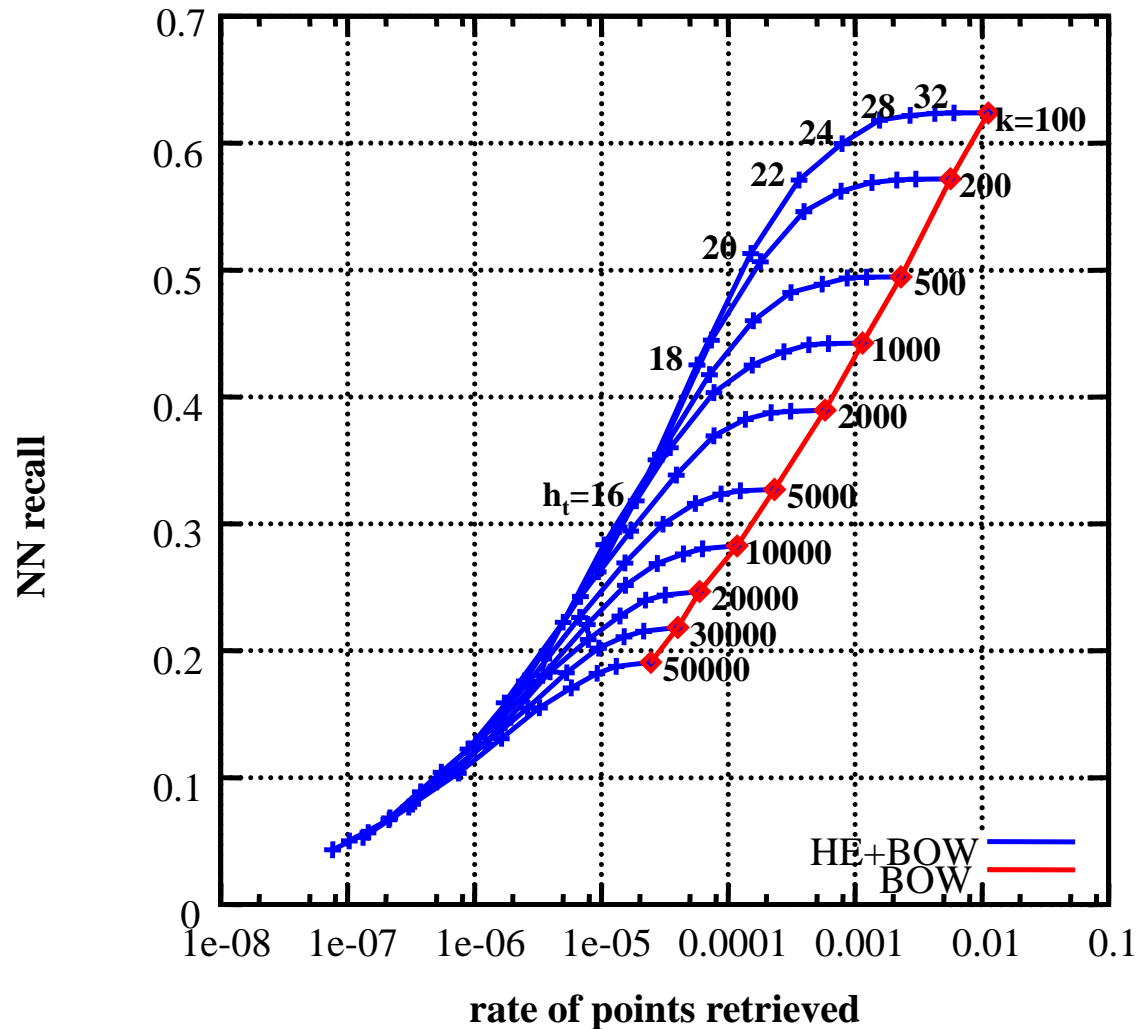


Trade-off between memory usage and accuracy

→ More bits yield higher accuracy

In practice, 64 bits (8 byte)

# ANN evaluation of Hamming Embedding



compared to BOW: at least 10 times less points in the short-list for the same level of NN recall

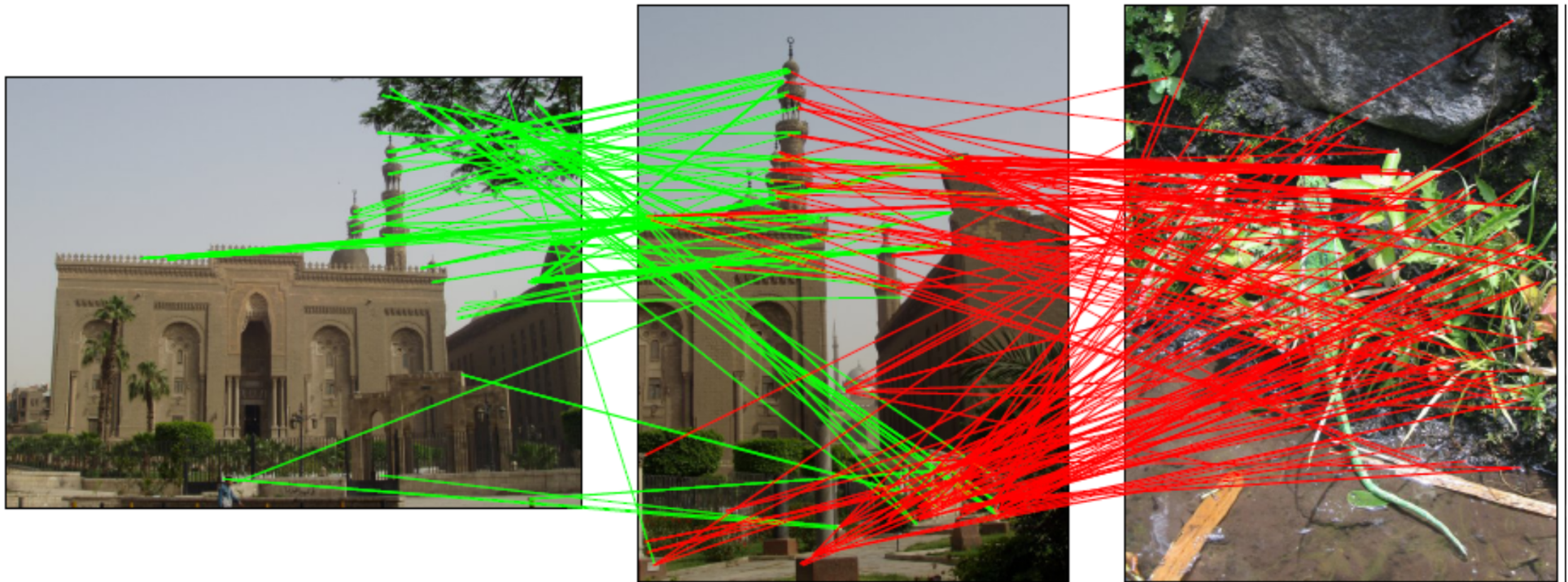
Hamming Embedding provides a much better trade-off between recall and ambiguity removal

# Matching points - 20k word vocabulary

---

201 matches

240 matches



Many matches with the non-corresponding image!



# Matching points - 200k word vocabulary

---

69 matches

35 matches



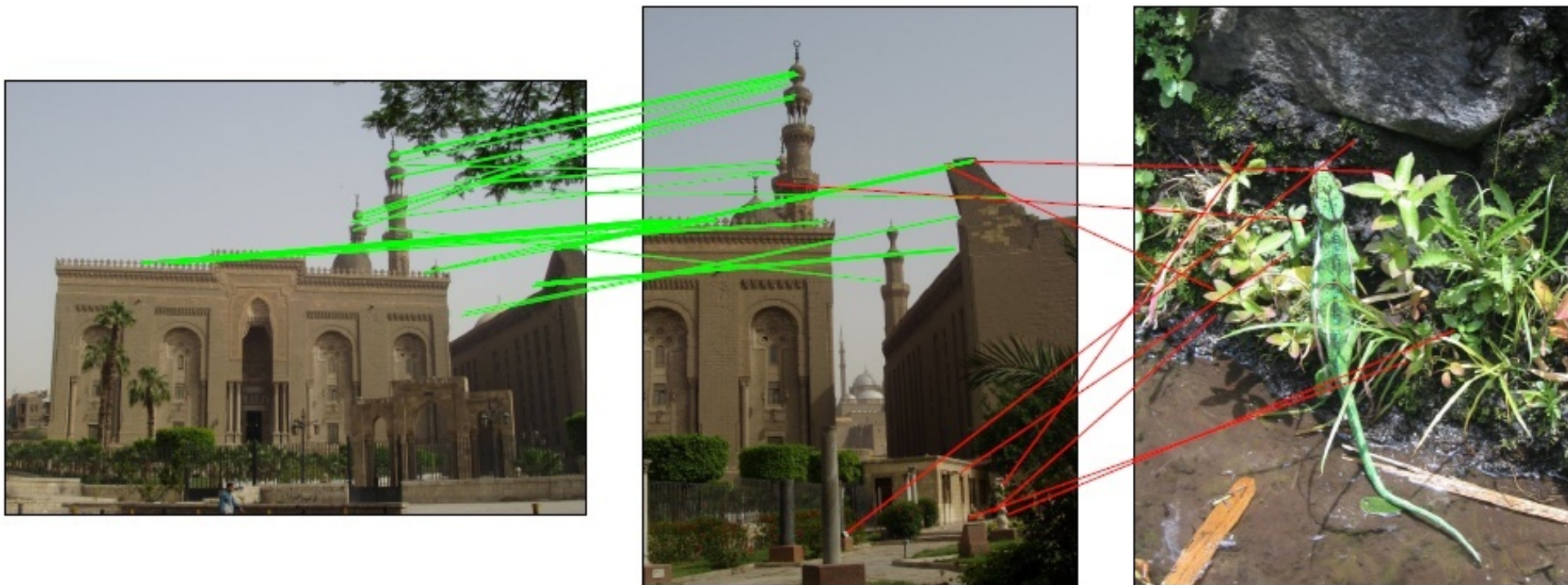
Still many matches with the non-corresponding one

# Matching points - 20k word vocabulary + HE

---

83 matches

8 matches



10x more matches with the corresponding image!

# INRIA holidays dataset

---

- Evaluation for the INRIA holidays dataset, 1491 images
  - 500 query images + 991 annotated true positives
  - Most images are holiday photos of friends and family
- 1 million & 10 million distractor images from Flickr
- Vocabulary construction on a different Flickr set
  
- Evaluation metric: mean average precision (in  $[0,1]$ , bigger = better)
  - Average over precision/recall curve



# Holiday dataset – example queries

---



# Dataset : Venice Channel

---





# Dataset : San Marco square

---



# Example distractors - Flickr

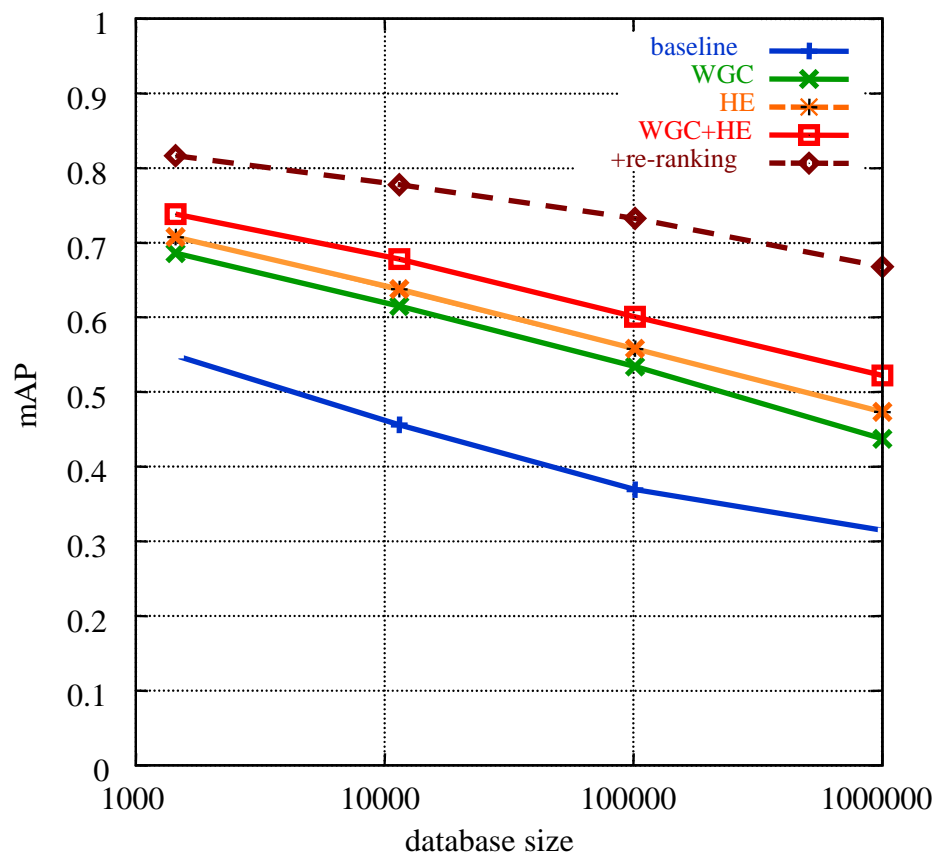
---





# Experimental evaluation

- Evaluation on our holidays dataset, 500 query images, 1 million distracter images
- Metric: mean average precision (in [0,1], bigger = better)

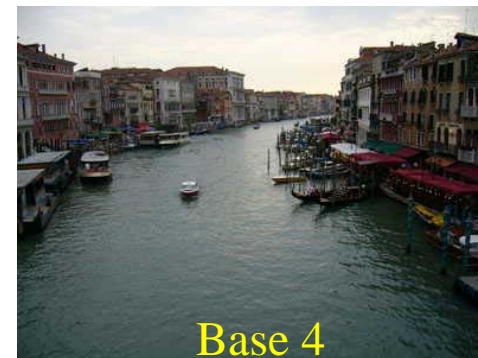


## Average query time (4 CPU cores)

Compute descriptors	880 ms
Quantization	600 ms
Search – baseline	<b>620 ms</b>
Search – WGC	<b>2110 ms</b>
Search – HE	<b>200 ms</b>
Search – HE+WGC	<b>650 ms</b>

# Results – Venice Channel

---



# Image retrieval - products


---

- Search for places and particular objects
  - For example on a smart phone




Courtesy Google

# Google image search

Google  cours\_exemple.png X mannequin tache de rousseur

Tous **Images** Maps Shopping Plus ▾ Outils de recherche

Environ 25 270 000 000 résultats (1,63 secondes)

 Taille de l'image :  
183 X 275

Trouver d'autres tailles de l'image :  
Toutes les tailles - Petite - Moyennes - Grandes

Hypothèse la plus probable pour cette image : [mannequin tache de rousseur](#)

[Les taches de rousseur font leur come-back ! - Cosmopolitan.fr](#)  
[www.cosmopolitan.fr](#) > Beauté > Maquillage > Tendances maquillage ▾  
En effet, les taches de rousseur sont en passe de devenir le it-truc beauté du moment. Entraînées sur le visage de quelques mannequins lors de la Fashion ...

[Les taches de rousseur ne se sont jamais aussi bien portées](#)  
[www.20minutes.fr](#) > Style ▾  
16 janv. 2016 - La tache de rousseur, ça peut être un complexe. ... en font un atout », explique Sylvie Fabre, directrice de l'agence de mannequins Wanted.

[Images similaires](#) [Signaler des images inappropriées](#)

