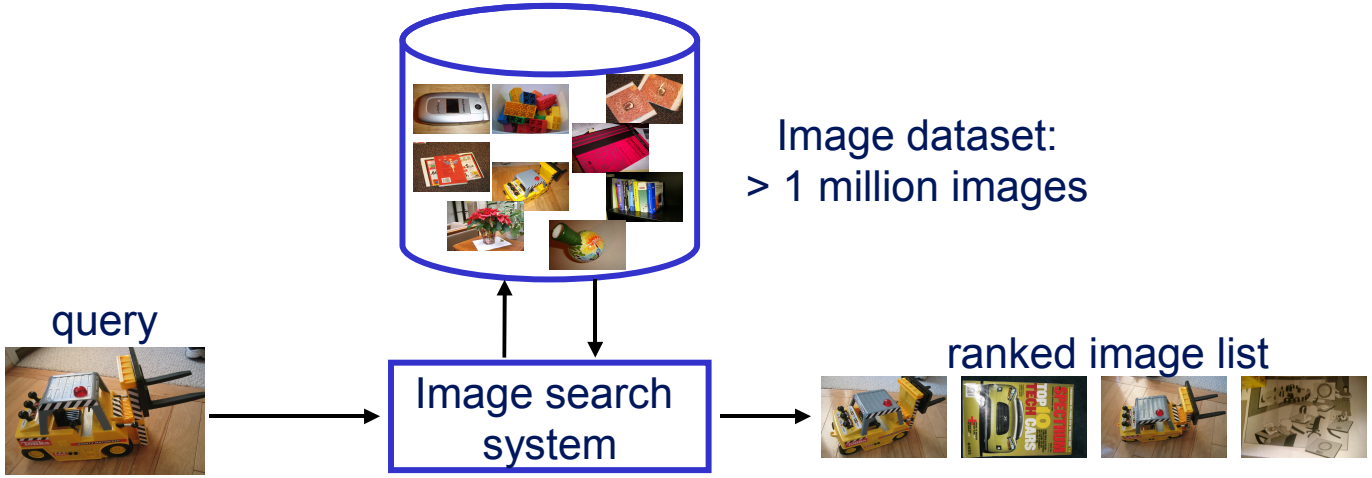
Hamming Embedding and Weak Geometry Consistency for large scale image search

Hervé Jégou, Matthijs Douze & Cordelia Schmid

Large scale object/scene recognition



- Each image described by approximately 2000 descriptors
  - 2 10^9 descriptors to index!

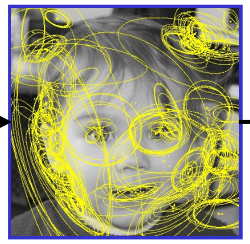
- Database representation in RAM:
  - Raw size of descriptors : 1 TB, search+memory intractable
  - Fast search with LSH: 800 GB, memory intractable

State-of-the-art: Bag-of-features (BOF) [Sivic & Zisserman'03]

Query image

Hessian-Affine regions + SIFT descriptors

[Mikolajezyk & Schmid 04]
[Lowe 04]



Set of SIFT descriptors

centroids (visual words)

Bag-of-features processing +*tf-idf* weighting

[Sivic & al 03, Nister & al 04, …]

sparse frequency vector

- "visual words":
  - 1 "word" (index) per local descriptor
  - only images ids in inverted file

  => 8 GB fits!

Inverted file

querying

ranked image short-list

Geometric verification

Re-ranked list

[Lowe 04, Chum & al 2007]

Outline

- Bag-of-features: approximate nearest neighbors (ANN) search interpretation
- Hamming Emdedding
- Weak geometry consistency

Bag-of-features as an ANN search algorithm

- Matching function of descriptors : *k*-nearest neighbors or ε-search

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

- Bag-of-features matching function $f_q(x,y) = \delta_{q(x),q(y)}$
  where q(x) is a quantizer, i.e., assignment to visual word and
  δa,b is the Kronecker operator (δa,b=1 iff a=b)

Approximate nearest neighbor search evaluation

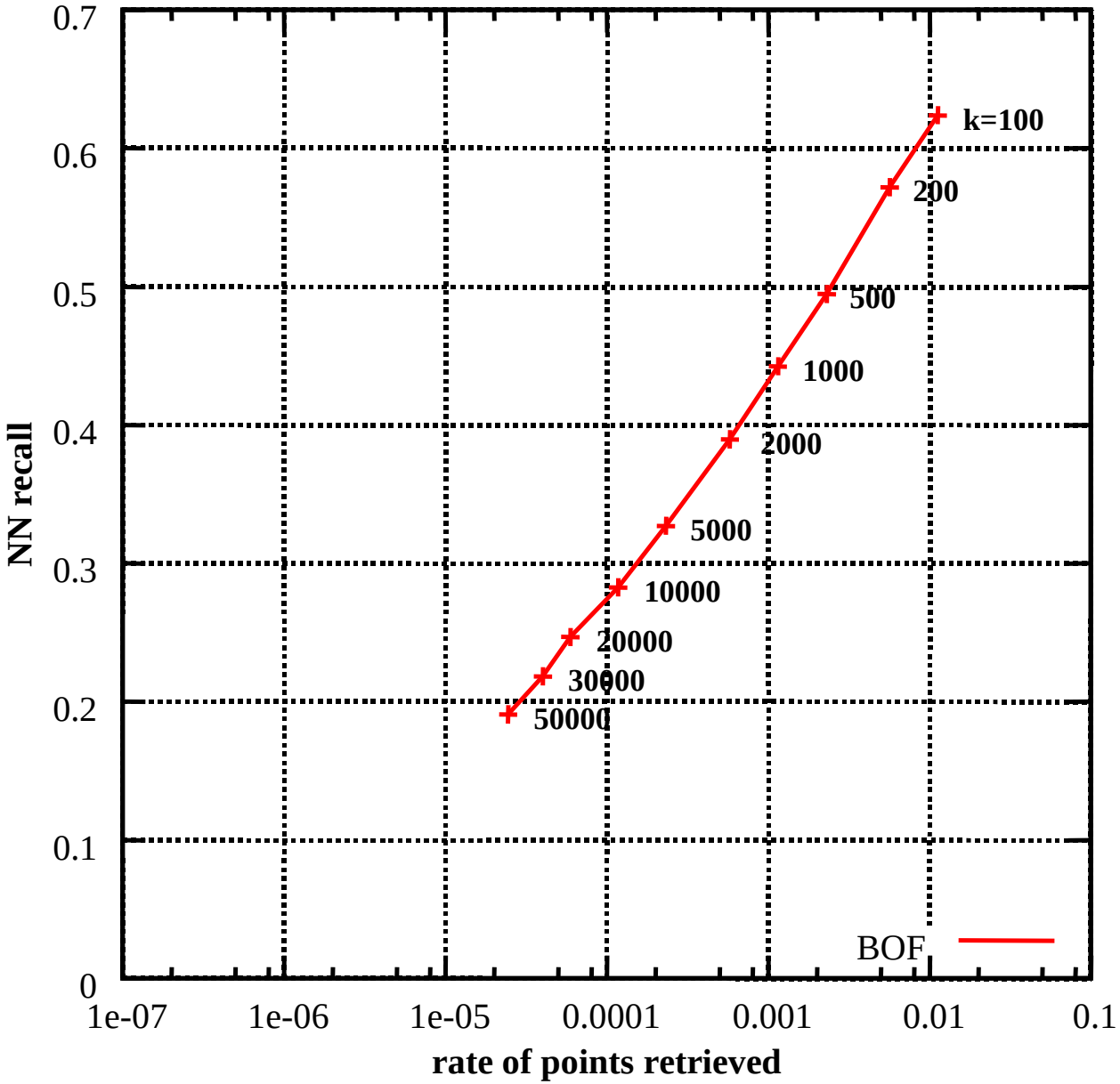
- ANN algorithms usually returns a short-list of nearest neighbors
  - this short-list is supposed to contain the NN with high probability
  - exact search may be performed to re-order this short-list

- Proposed quality evaluation of ANN search: trade-off between
  - **Accuracy**: **NN recall** = probability that *the* NN is in this list

    against

  - **Ambiguity removal** = proportion of vectors in the short-list
    - the lower this proportion, the more information we have about the vector
    - the lower this proportion, 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

Accuracy: **NN recall**
= probability that *the* NN is in this list

Ambiguity removal:
= proportion of vectors in the short-list

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

Outline

- Bag-of-features: voting and ANN interpretation
- Hamming Embedding
- Weak geometry consistency

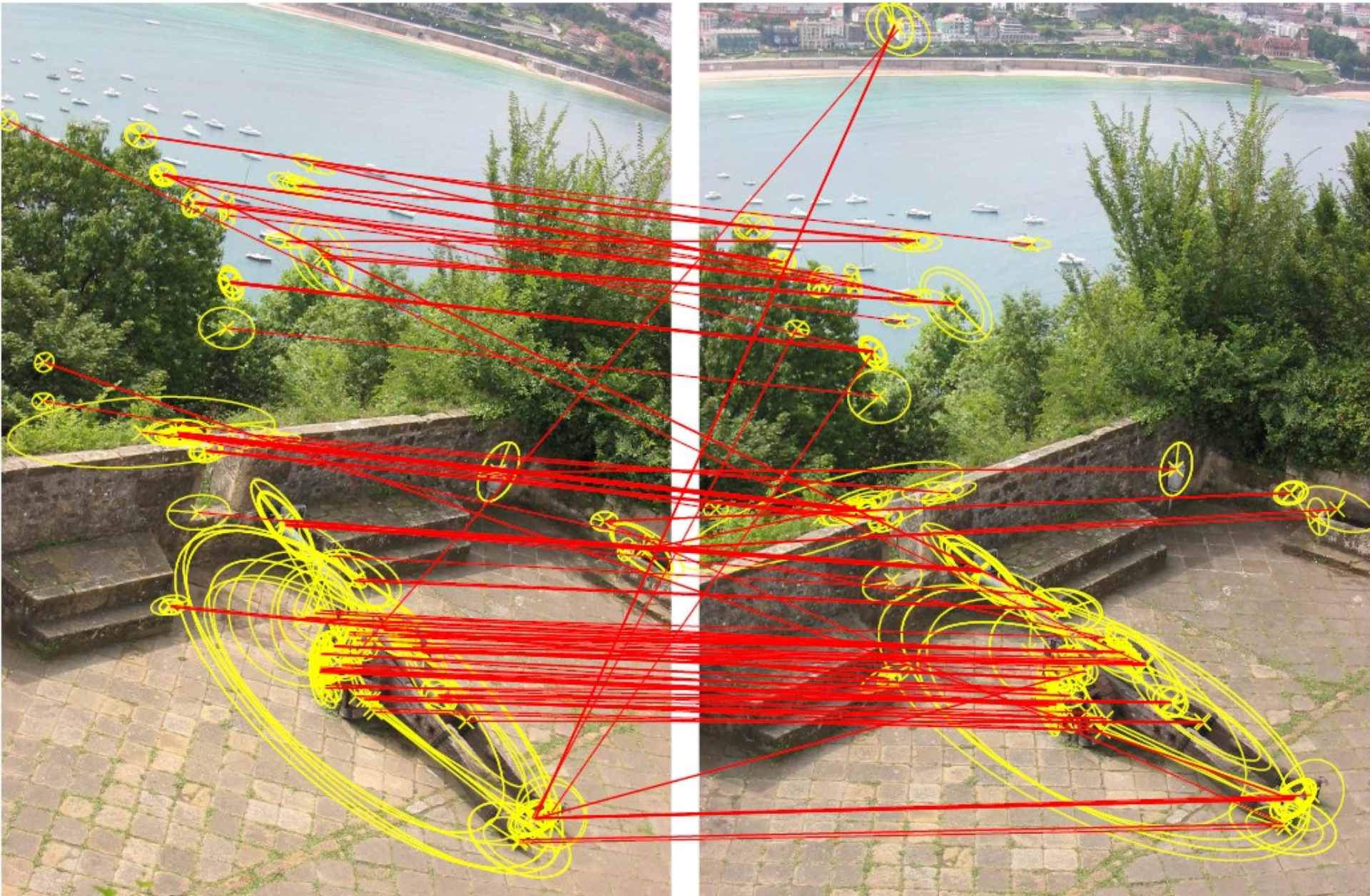
State-of-the art: First issue

- The intrinsic matching scheme performed by BOF is weak
  - for a “small” visual dictionary: too many false matches
  - for a “large” visual dictionary: many true matches are missed

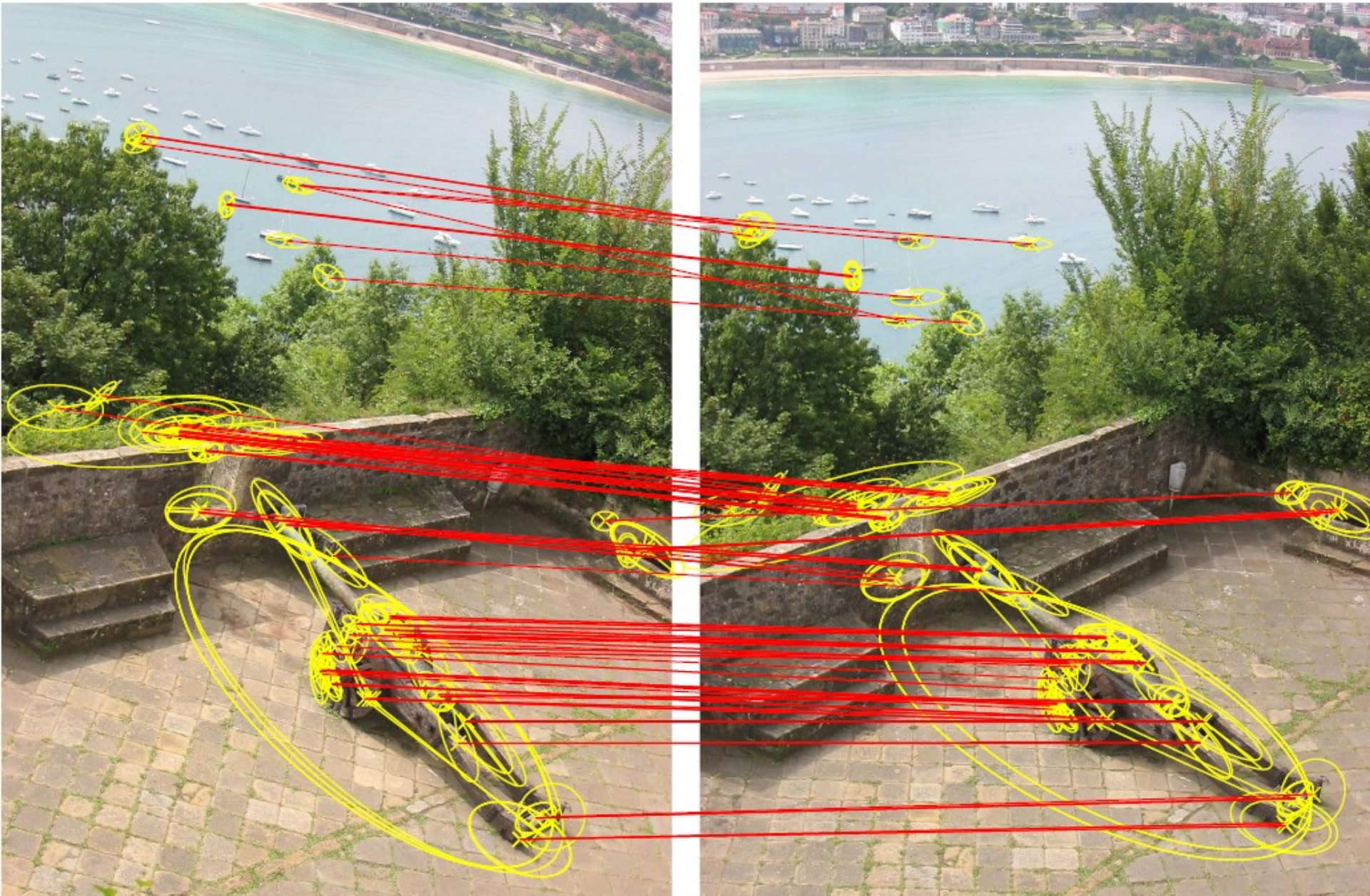
- No good trade-off between “small” and “large” !
  - either the Voronoi cells are too big
  - or these cells can’t absorb the descriptor noise

  $\rightarrow$ intrinsic approximate nearest neighbor search of BOF is not sufficient

20K visual word: false matchs



200K visual word: good matches missed

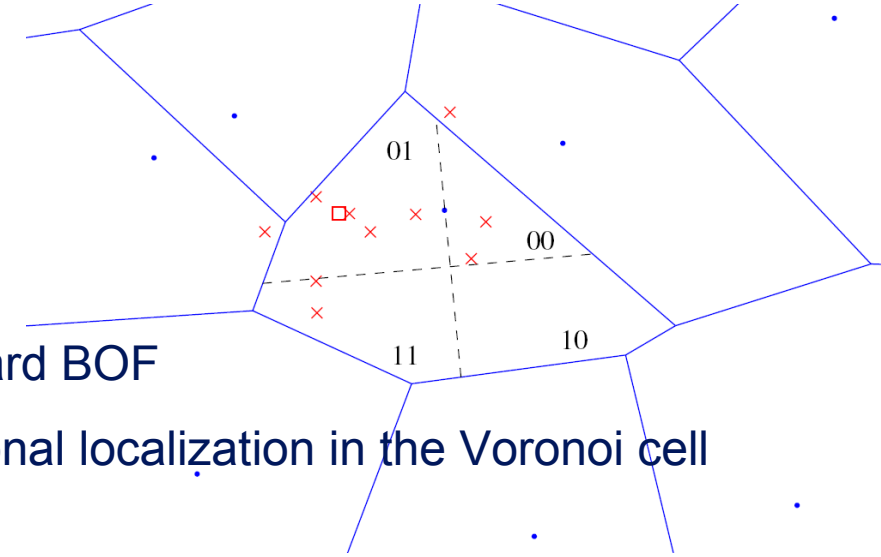


State-of-the art: First issue

- Need to fight against the quantization noise
  $\rightarrow$ several recent paper proposed methods to have a richer representation of (sets of) descriptors
- In image search: multiple or soft assignment of descriptors to visual words
  - Jegou *et* al, “A contextual dissimilarity measure for accurate and efficient image search”, CVPR’2007
  - Philbin *et* al., “Lost in quantization: improving particular object retrieval in large scale image databases”, CVPR’2008

- these methods reduce the sparsity of the BOF representation
  $\rightarrow$ negative impact on the search efficiency

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

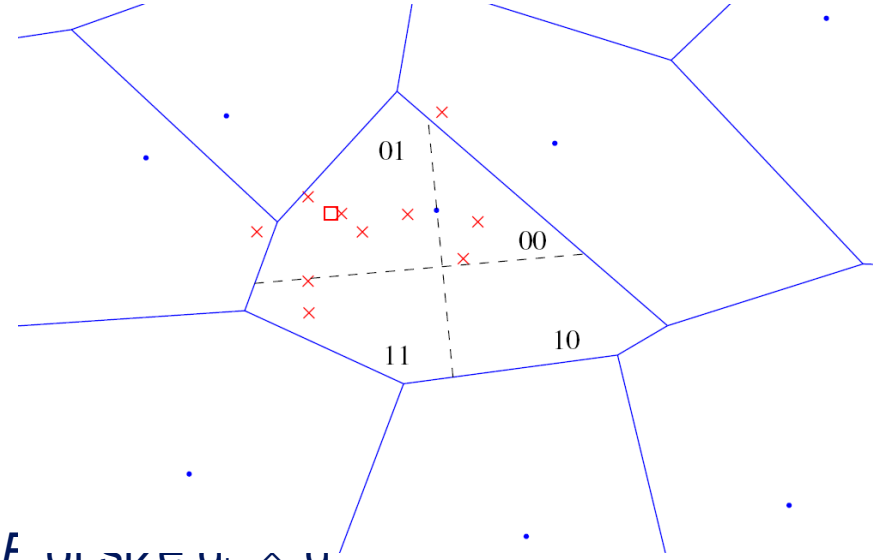
- Two descriptors x and y match iif $\begin{cases} q(x) = q(y) \\ \mathrm{h}(b(x),\ b(y)) < \tau \end{cases}$

  where h(*a*,*b*) is the Hamming distance

- Nearest neighbors for Hamming distance ≈ 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 that standard BOF with same dictionary size!

Hamming Embedding



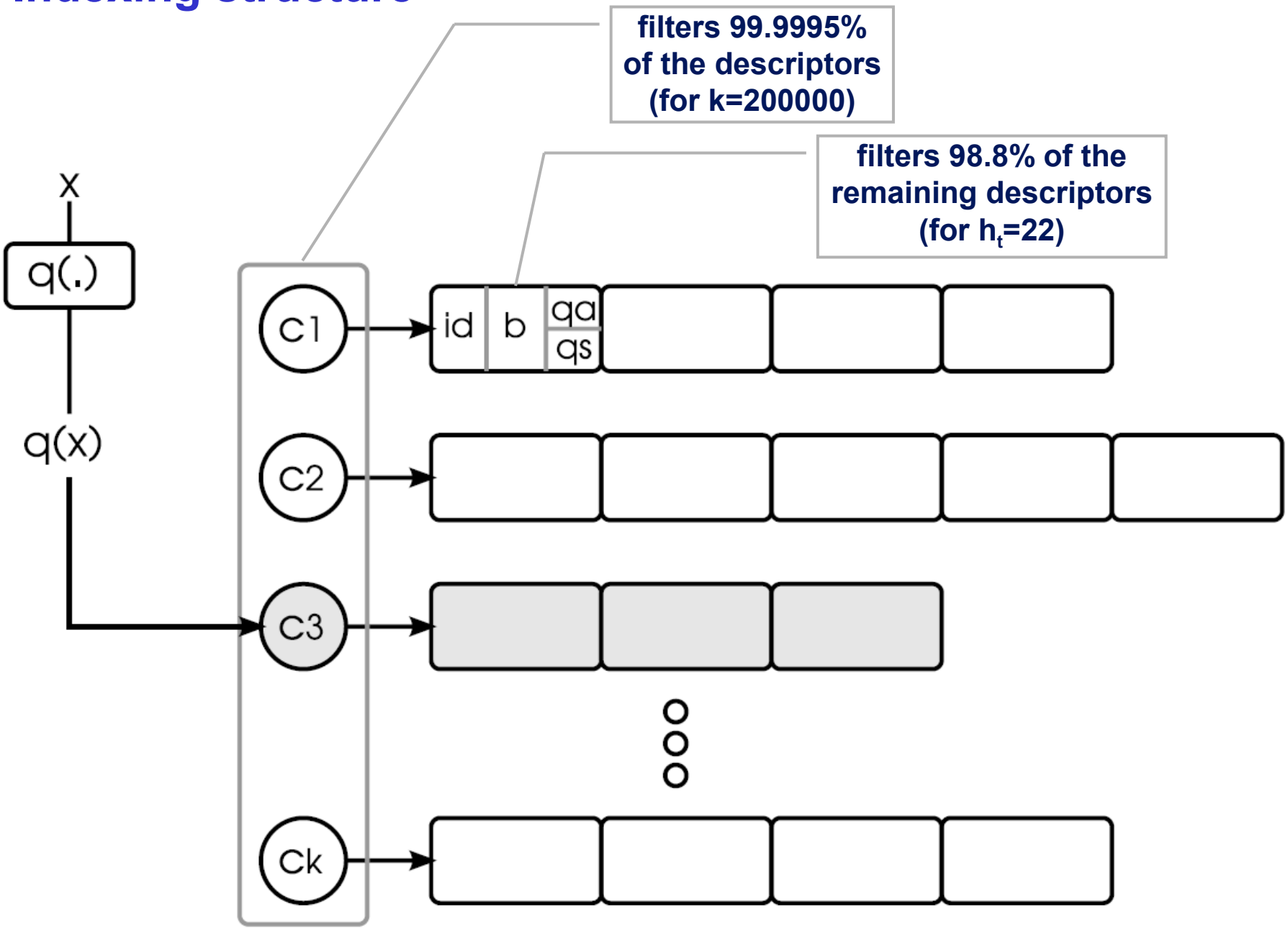
- **Off-line** (given a quantizer)
  - draw an orthogonal projection matrix P of size $d_b \times d$

  $\rightarrow$ this defines d_b random projection directions

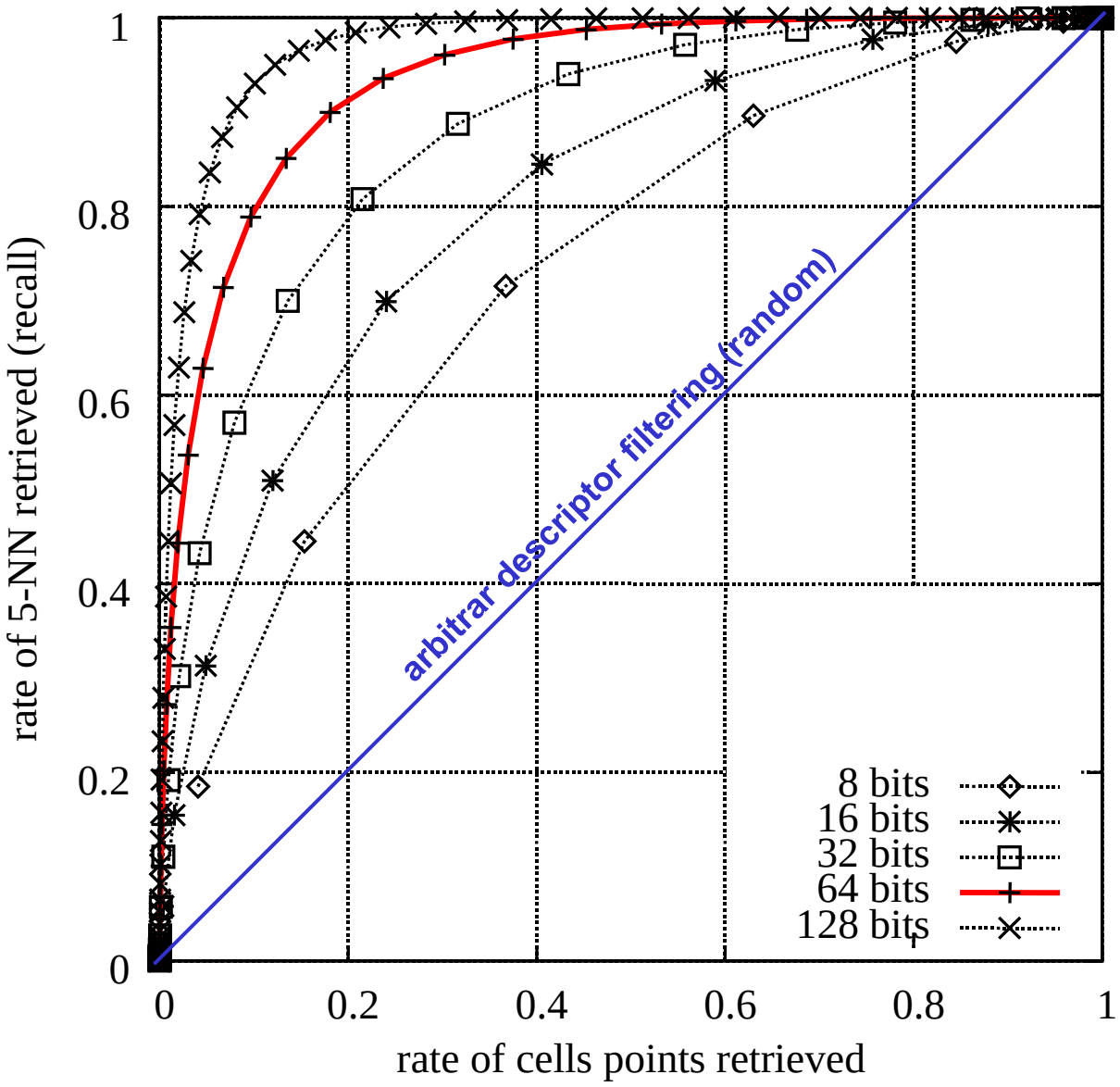
  - for each Voronoi cell and projection direction, compute the median value for a learning set

- **On-line**: compute the binary signature b(x) of a given descriptor
  - project x onto the projection directions as $z(x) = (z_1,\ldots z_{db})$
  - $b_i(x) = 1$ if $z_i(x)$ is above the learned median value, otherwise 0

Indexing structure



Hamming and Euclidean neighborhood

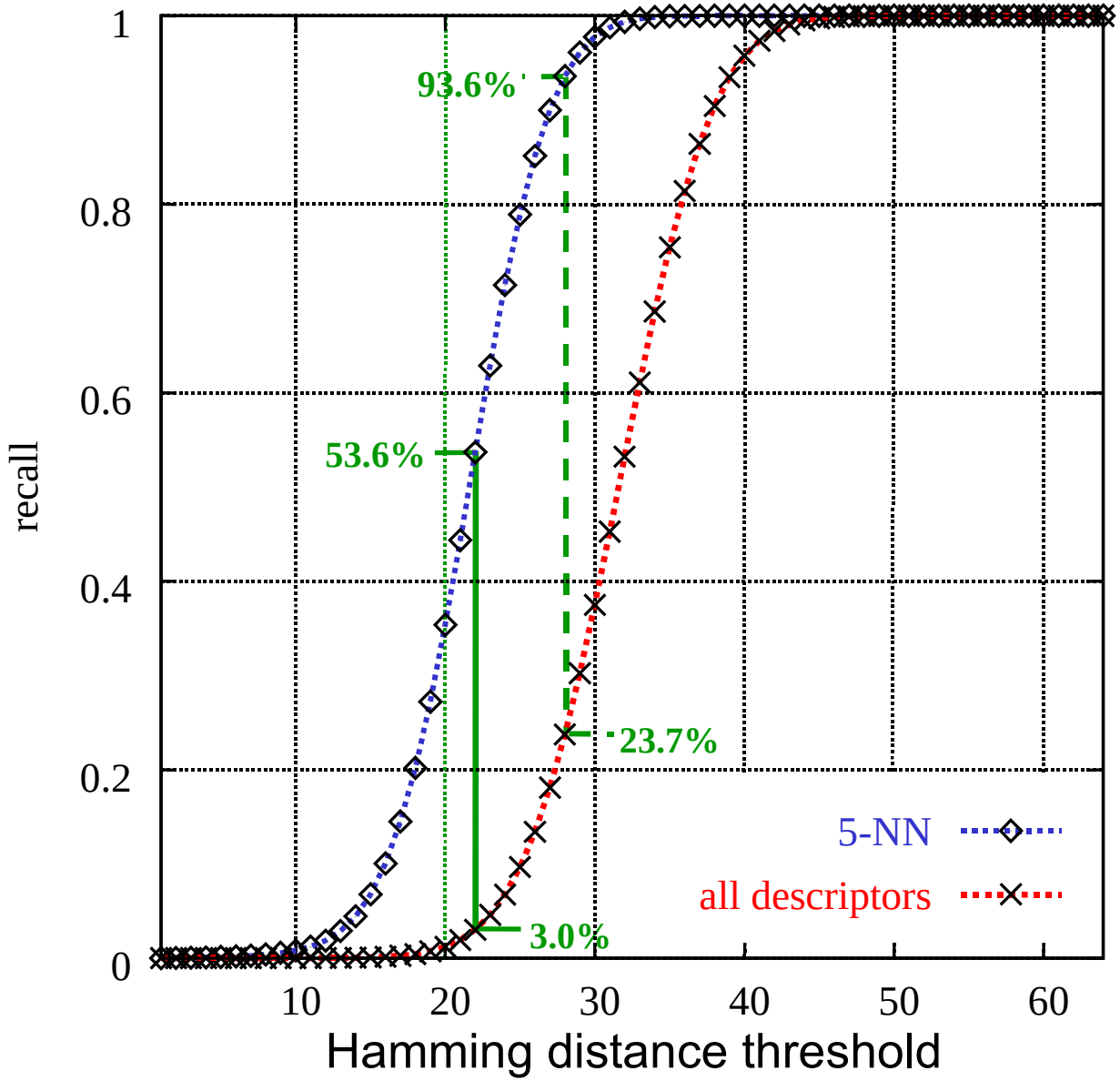


- trade-off between memory usage and accuracy

→ more bits yield higher accuracy

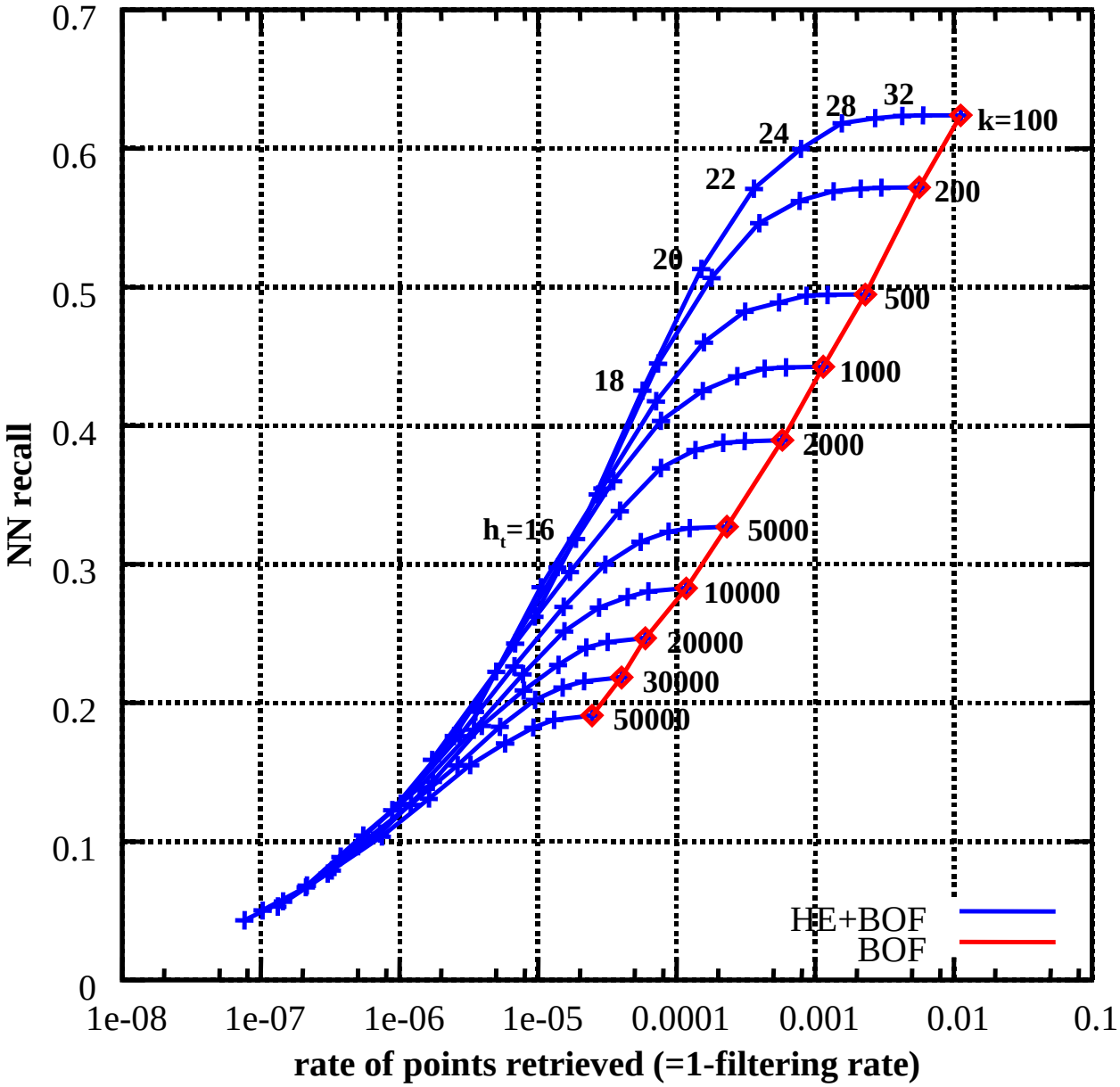
We used 64 bits (8 bytes)

Hamming Embedding: filtering matches



$b(x) \in \{0,1\}^{64}$

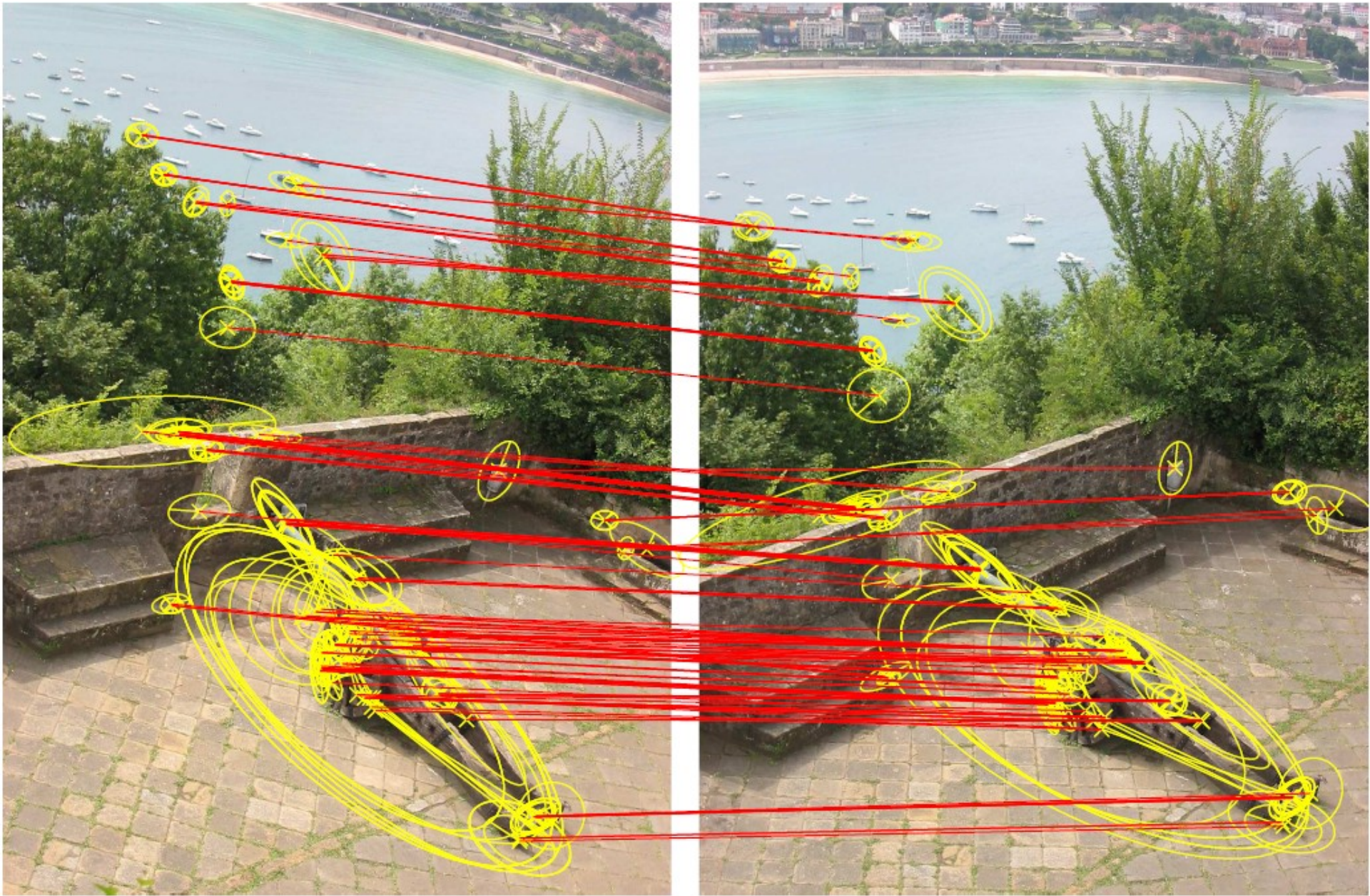
ANN evaluation of Hamming Embedding



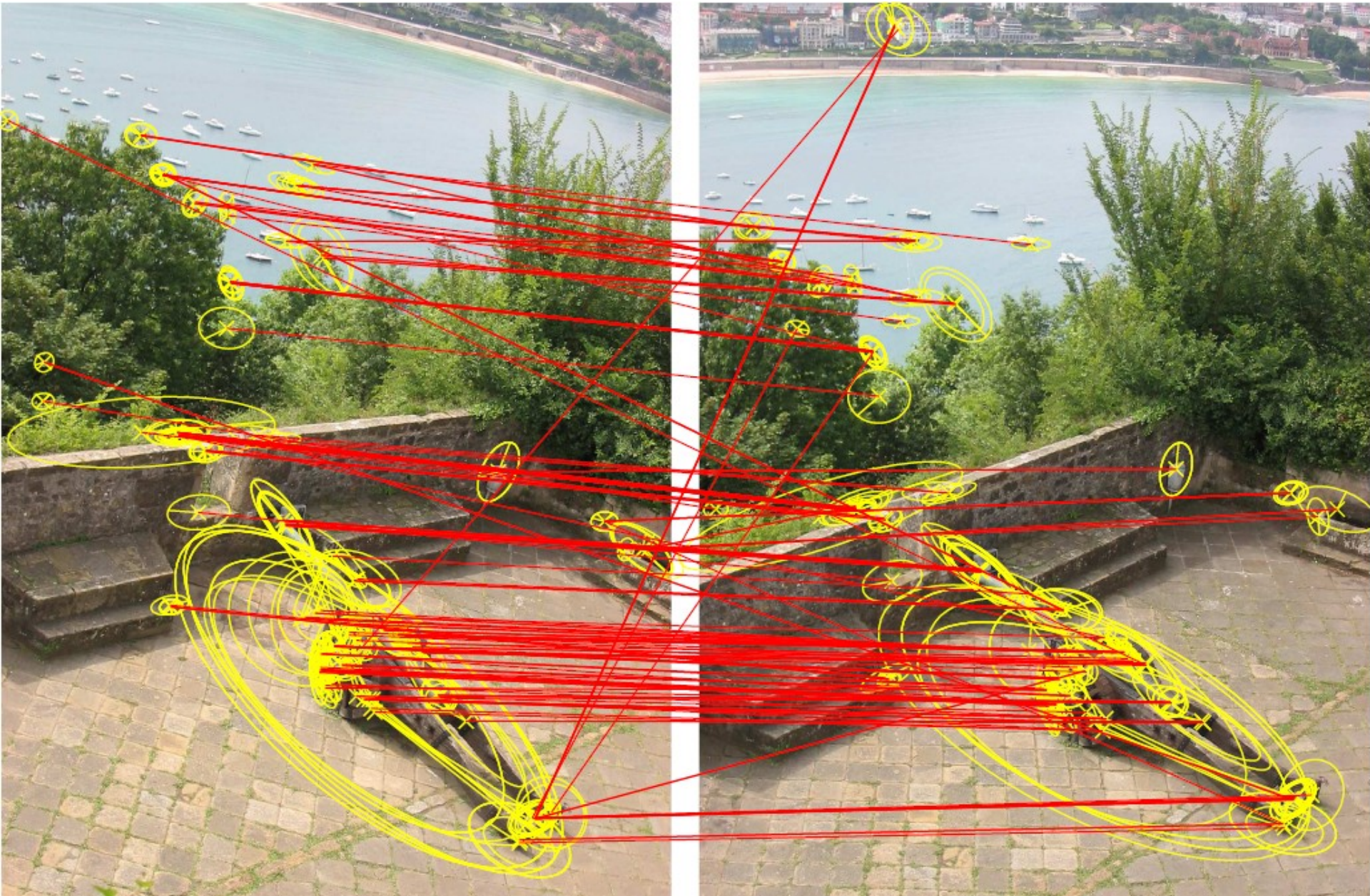
compared to BOF: at least 10 times less points in the short-list for the same level of accuracy

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

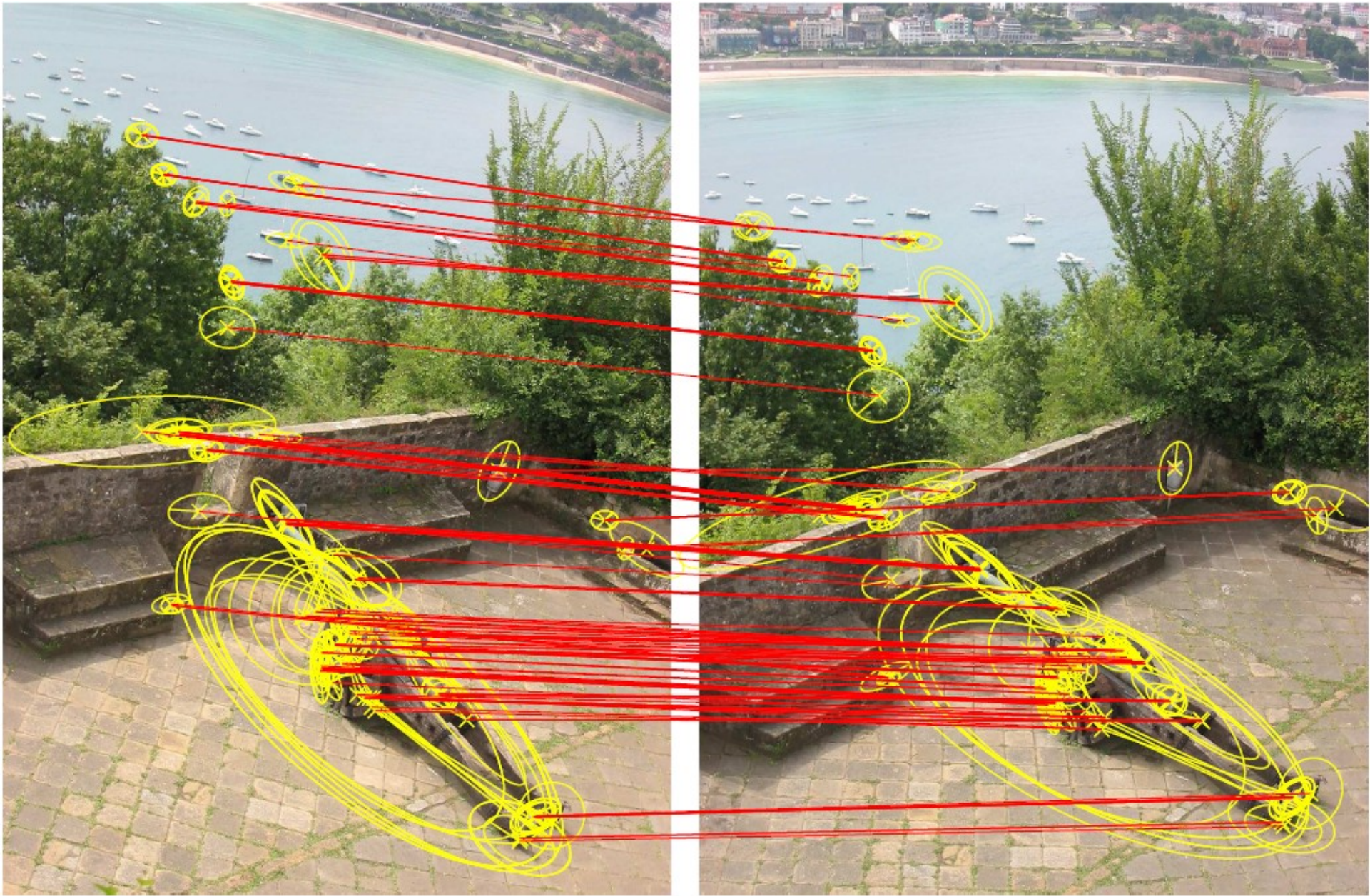
Hamming Embedding: Example



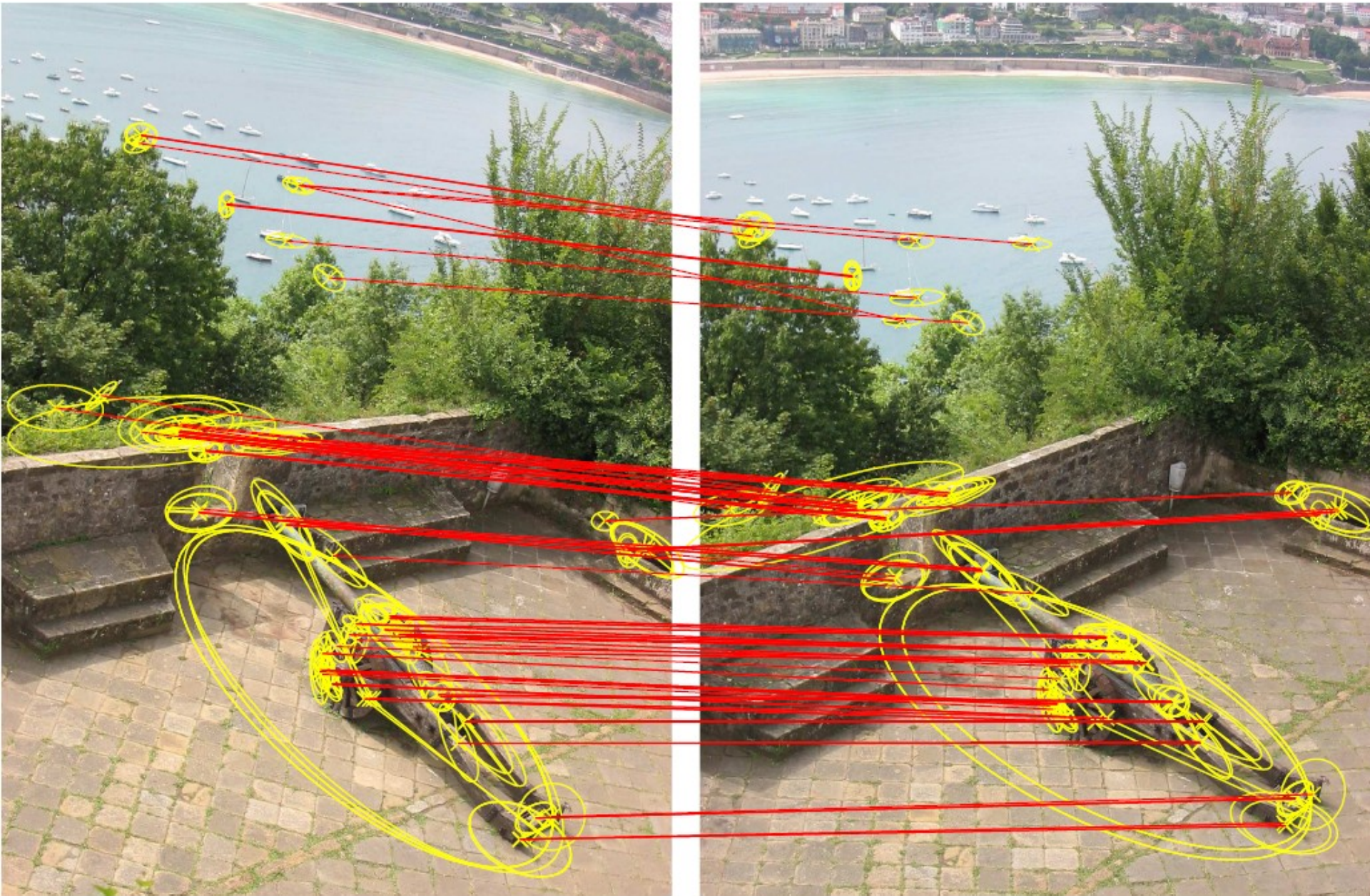
Compared with 20K dictionary : false matches



Hamming Embedding: Example



Compared with 200K visual word: good matches missed



Outline

- Bag-of-features: voting and ANN interpretation
- Hamming Emdedding
- Weak geometry consistency

State-of-the art: Second issue

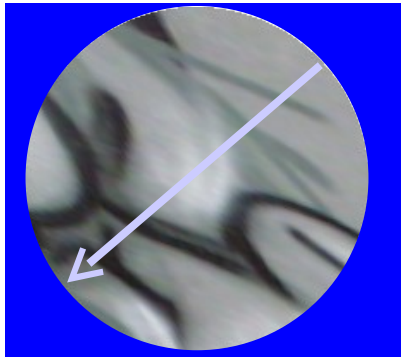
- Re-ranking based on full geometric verification [Lowe 04, Chum & al 2007]
  - works very well
  - but performed on a short-list only (typically, 100 images)

  → for very large datasets, the number of distracting images is so high that relevant images are not even short-listed!

Weak geometry consistency

- Weak geometric information used for **all** images (not only the short-list)

- Each invariant interest region detection has a scale and rotation angle associated, here characteristic scale and dominant gradient orientation



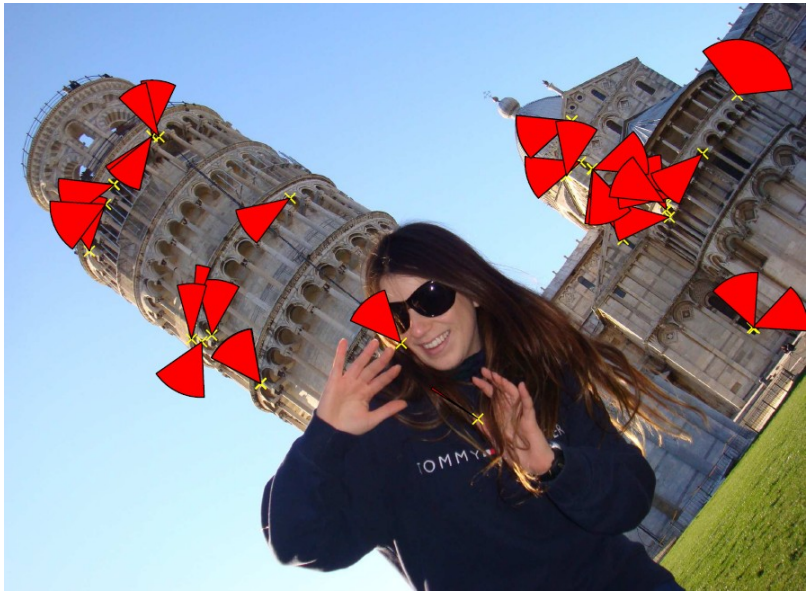
Scale change 2
Rotation angle ca. 20 degrees

- Each matching pair results in a scale and angle difference

- For the global image scale and rotation changes are roughly consistent

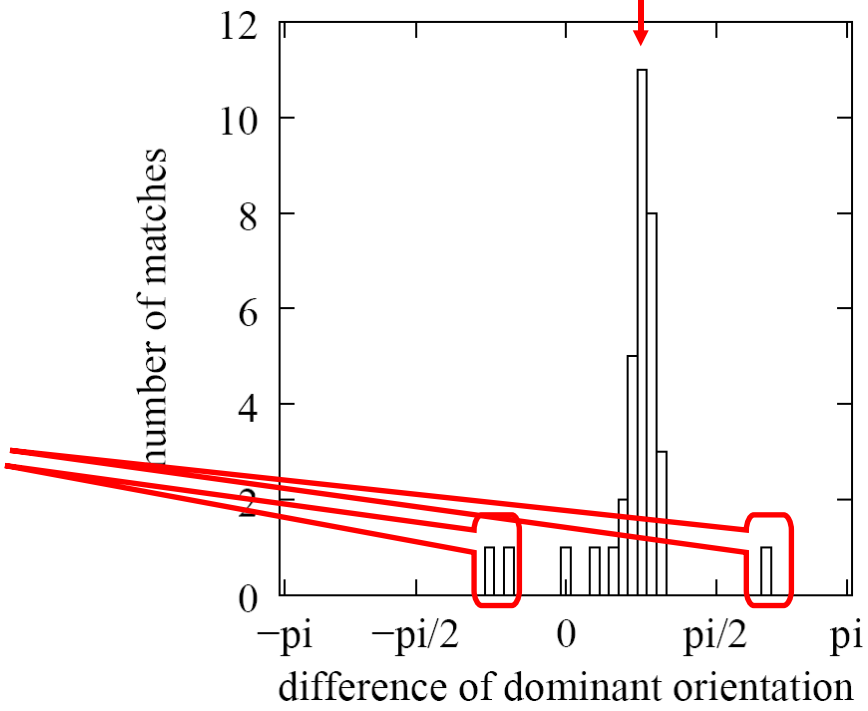
Pisa tower: Let analyze the dominent orientation difference of matching descriptors

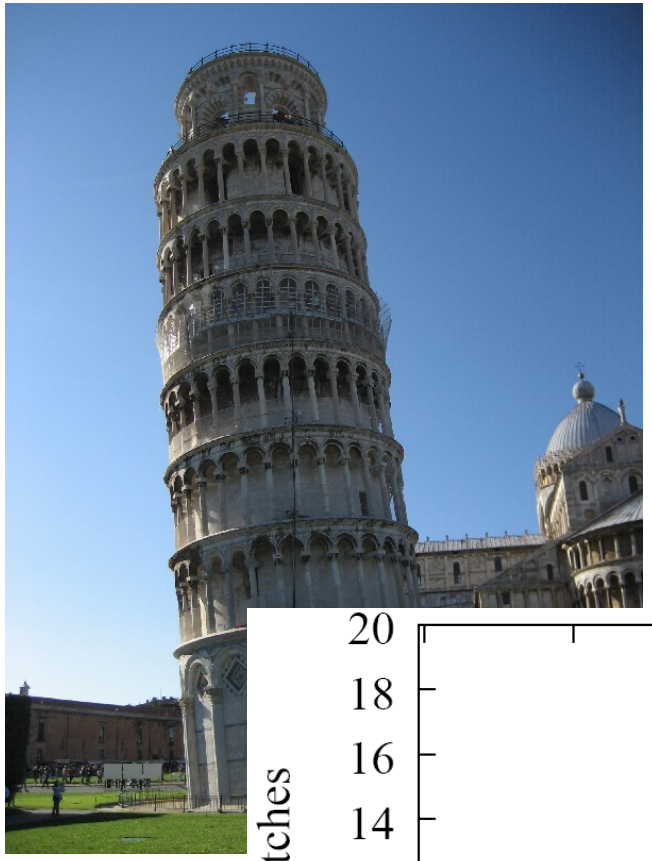
Orientation consistency



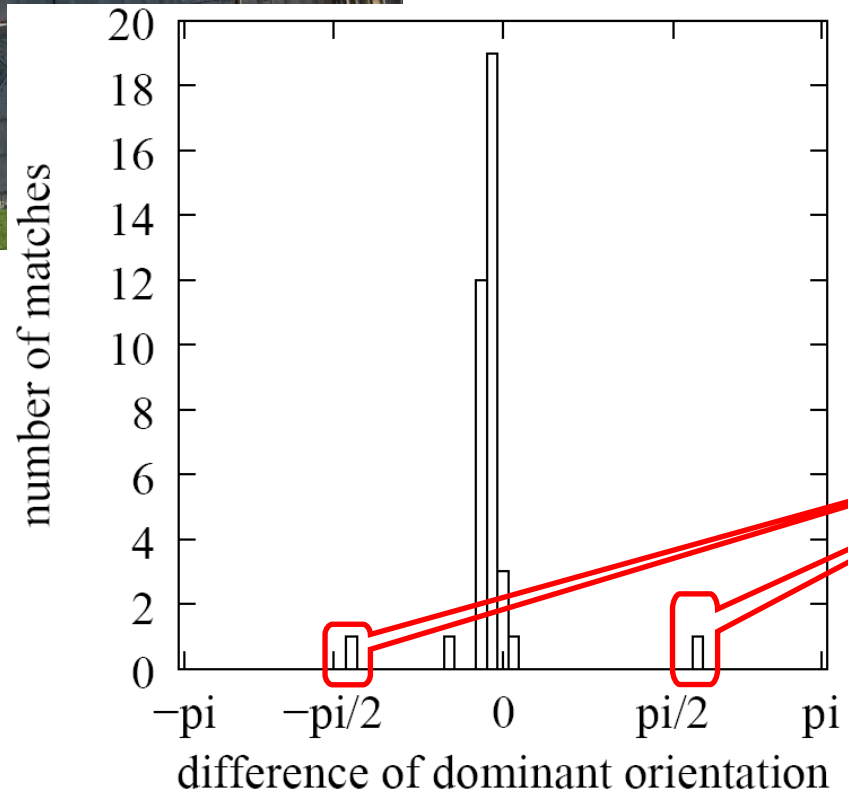
Max = rotation angle between images

FILTERED!

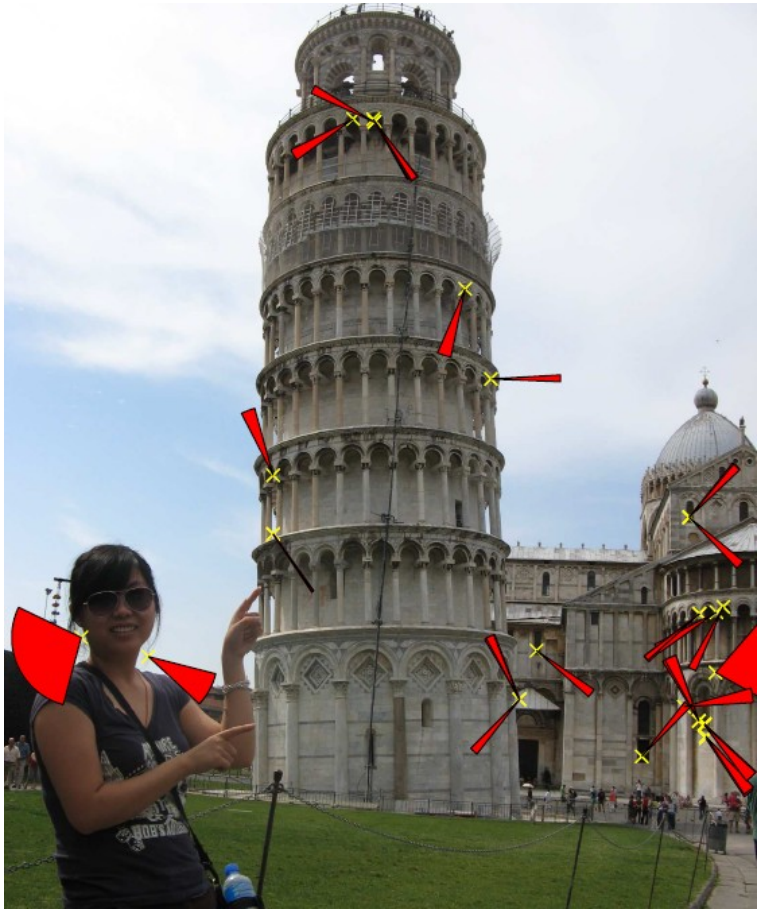




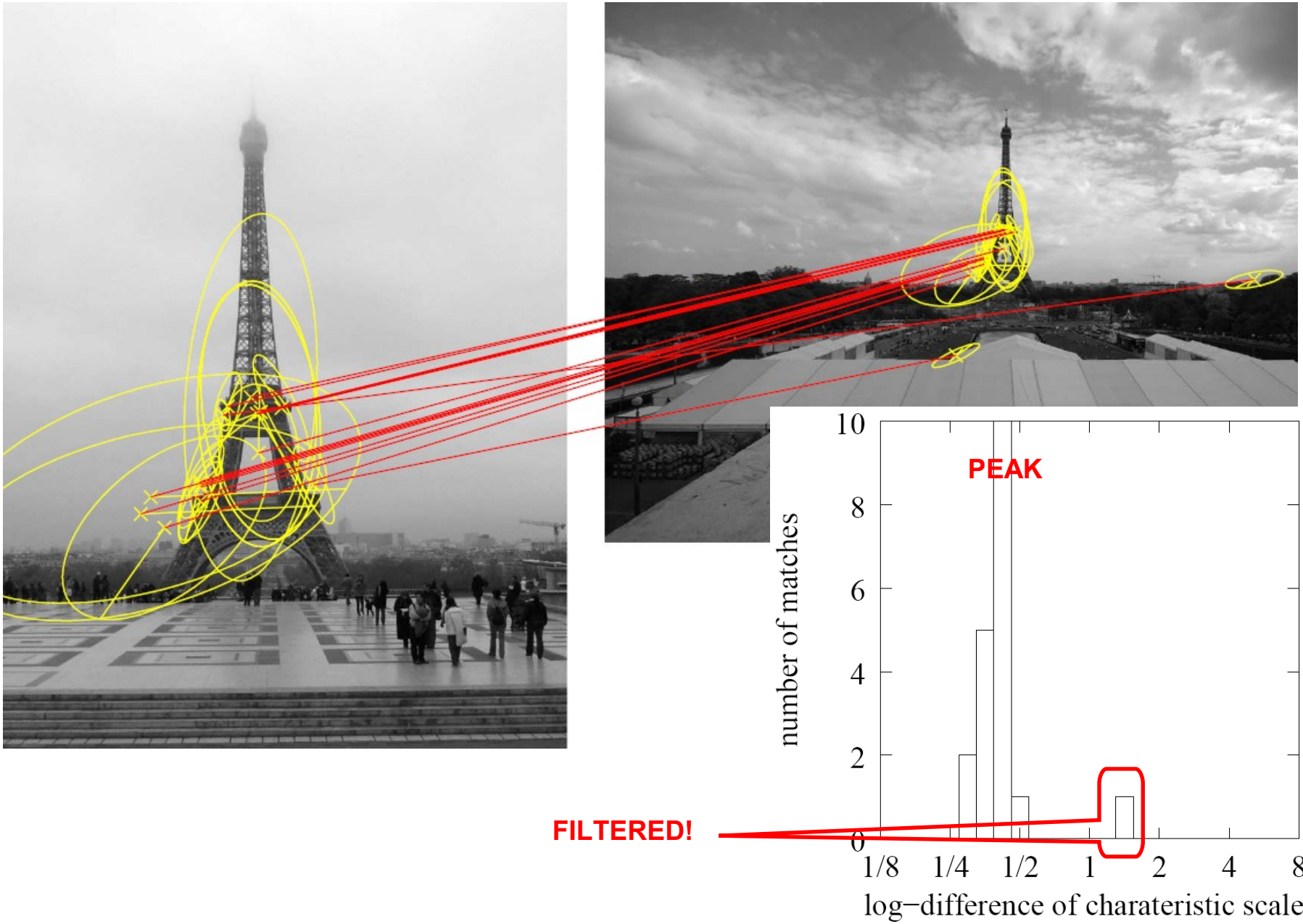
PEAK



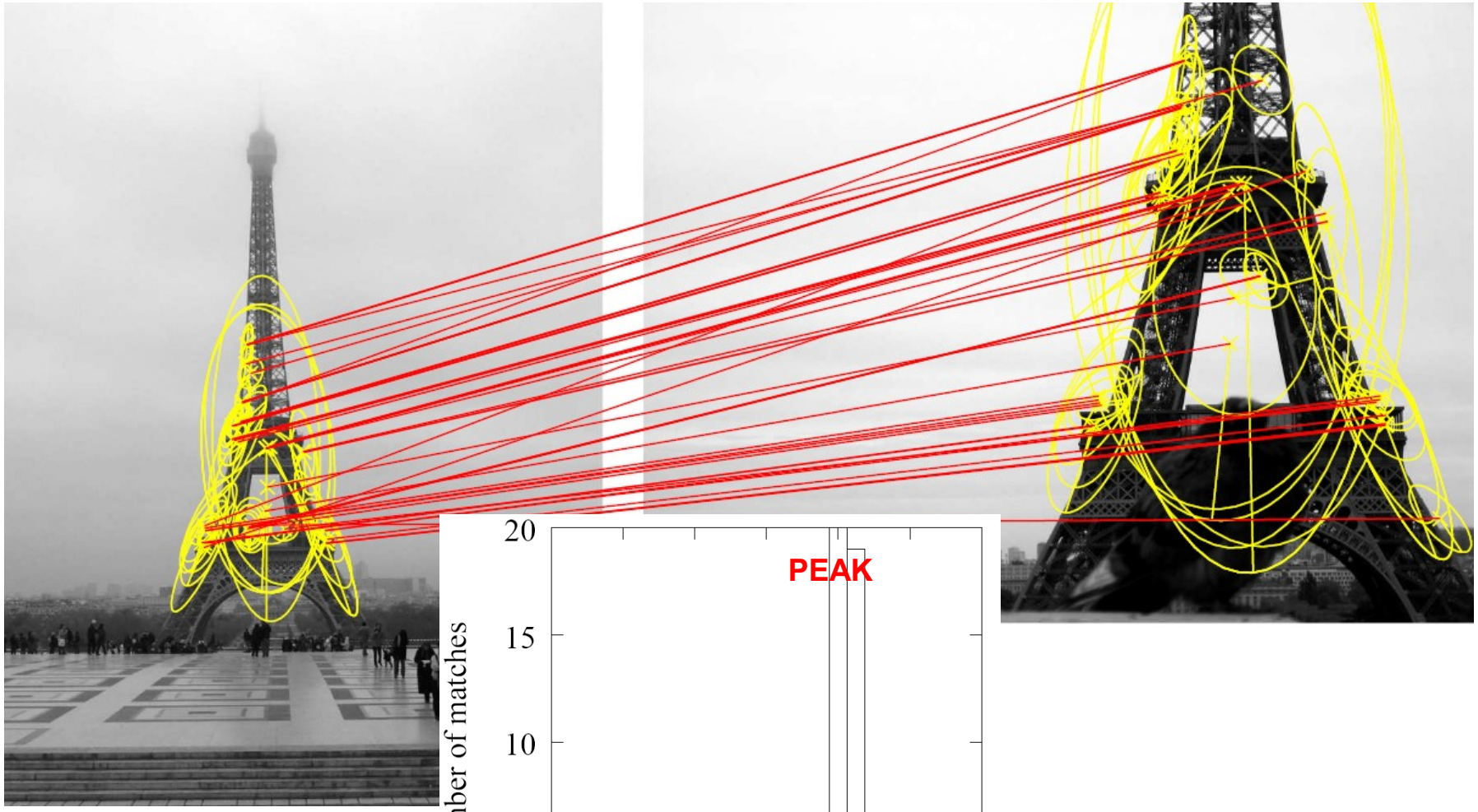
FILTERED!



WGC: scale consistency



WGC: scale consistency



Weak geometry consistency

- Integrate the geometric verification into the BOF representation
  - votes for an image projected onto two quantized subspaces, that is vote for an image at a given angle & scale

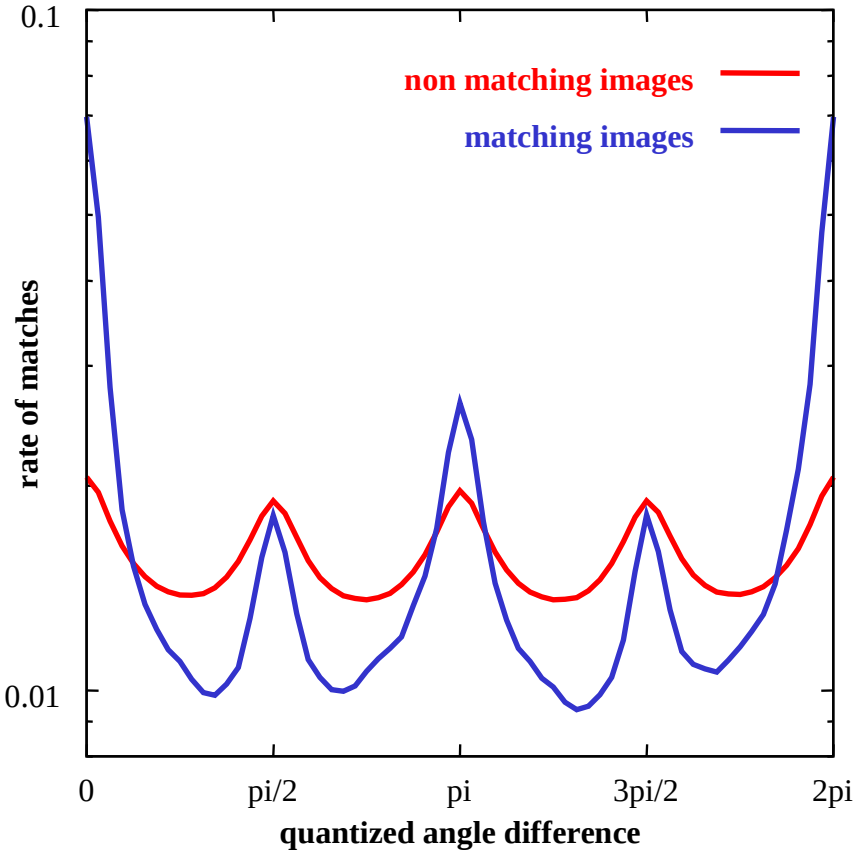
  $\rightarrow$ these subspace are shown to be independent
  - a score s_j for all quantized angle and scale differences for each image
  - final score: filtering for each parameter (angle and scale) and min selection

- Only matches that do agree with the main difference of orientation and scale will be taken into account in the final score

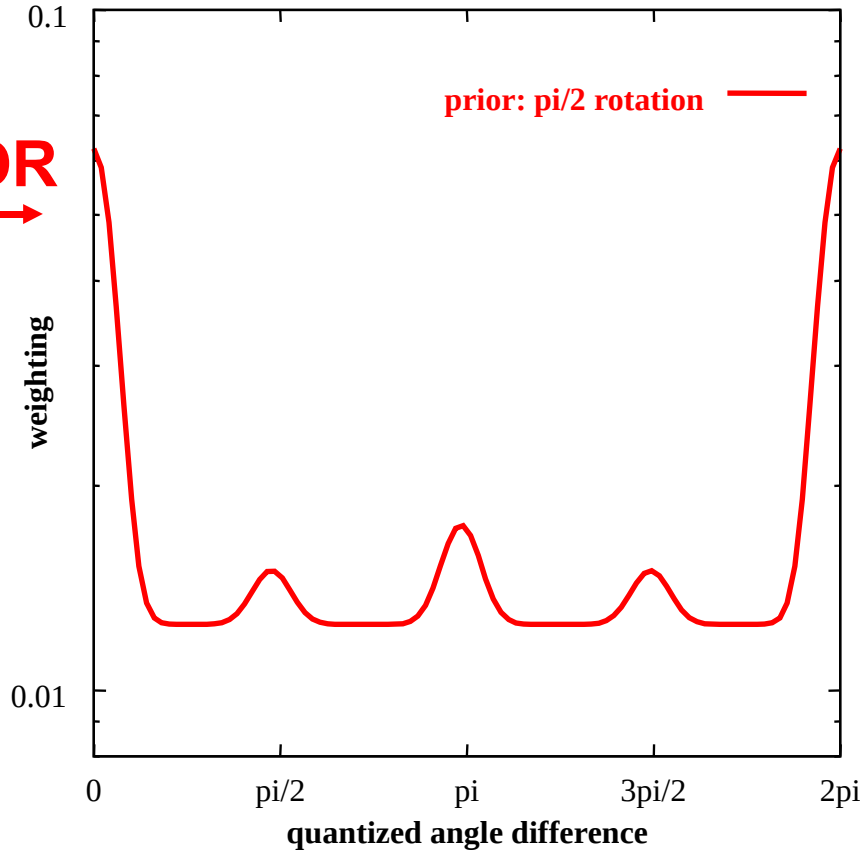
- Re-ranking using full geometric transformation still adds information in a final stage

Integrating geometric *a priori* information

- Images orientation difference is strongly non uniform
  - natural orientation for many images (but still π/2 rotation ambiguity)
  - human tendency to use the same orientation for the same place



PRIOR

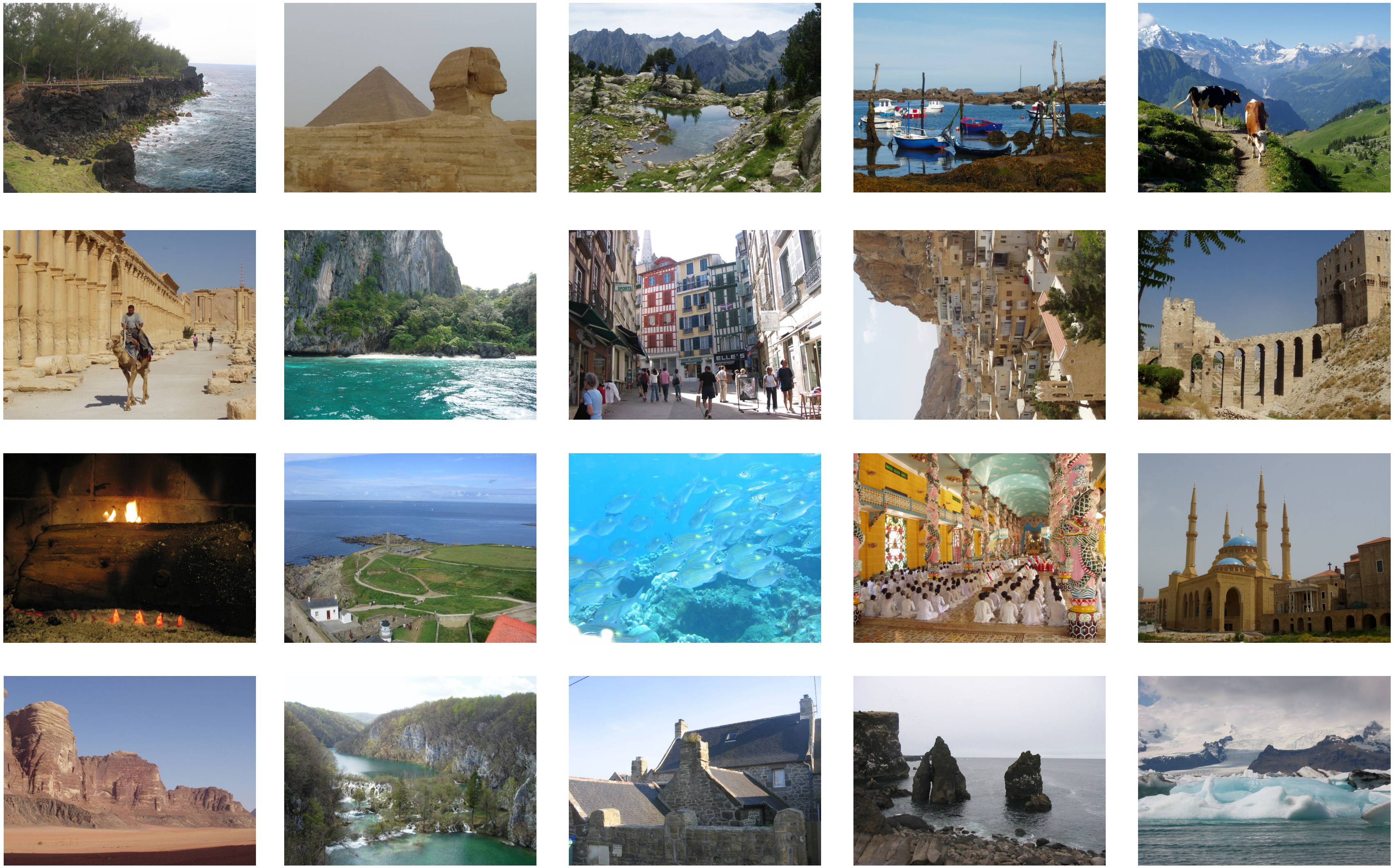


Experimental results

- 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 distractor images from Flickr
- Dataset size 1.001.491 images
- Vocabulary construction on a different Flickr set
- Almost real-time search speed, see retrieval demo

- Evaluation metric: mean average precision (in [0,1], bigger = better)
  - Average over precision/recall curve

Holidays dataset – example queries (out of 500)



Example query – response : Venice Channel

Example query – response : San Marco square

Results – Venice Channel

Query

BOF 2
Ours 1

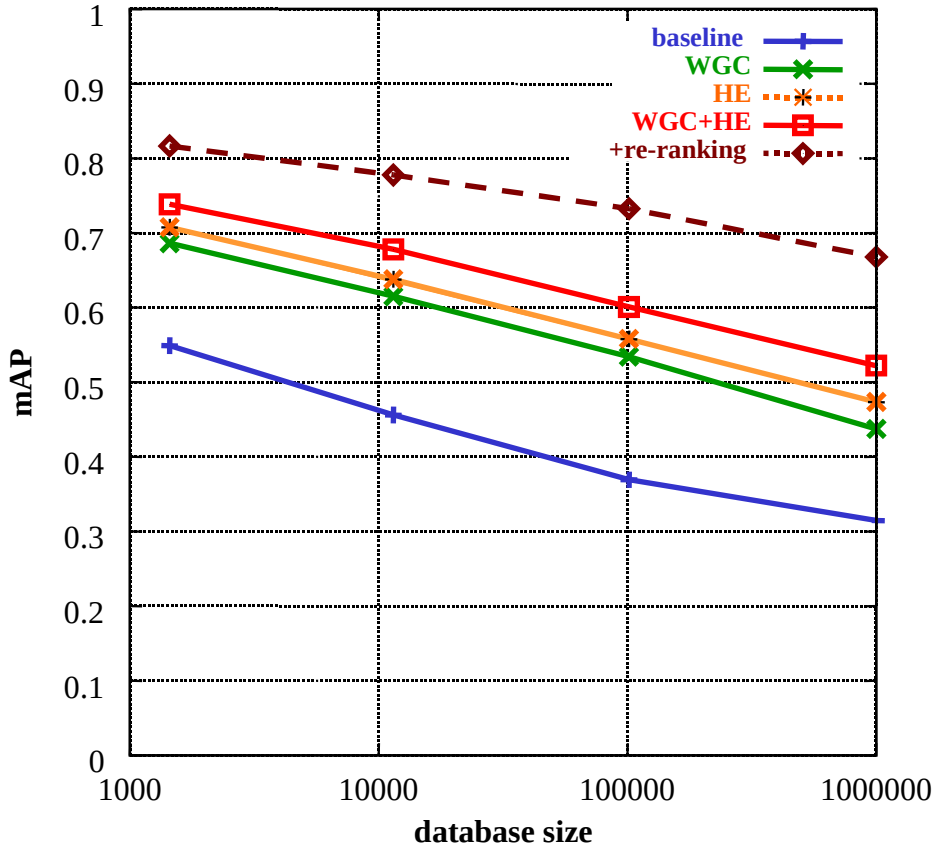
BOF 5890
Ours 4

BOF 43064
Ours 5

Results – San Marco

Comparison with state-of-the-art

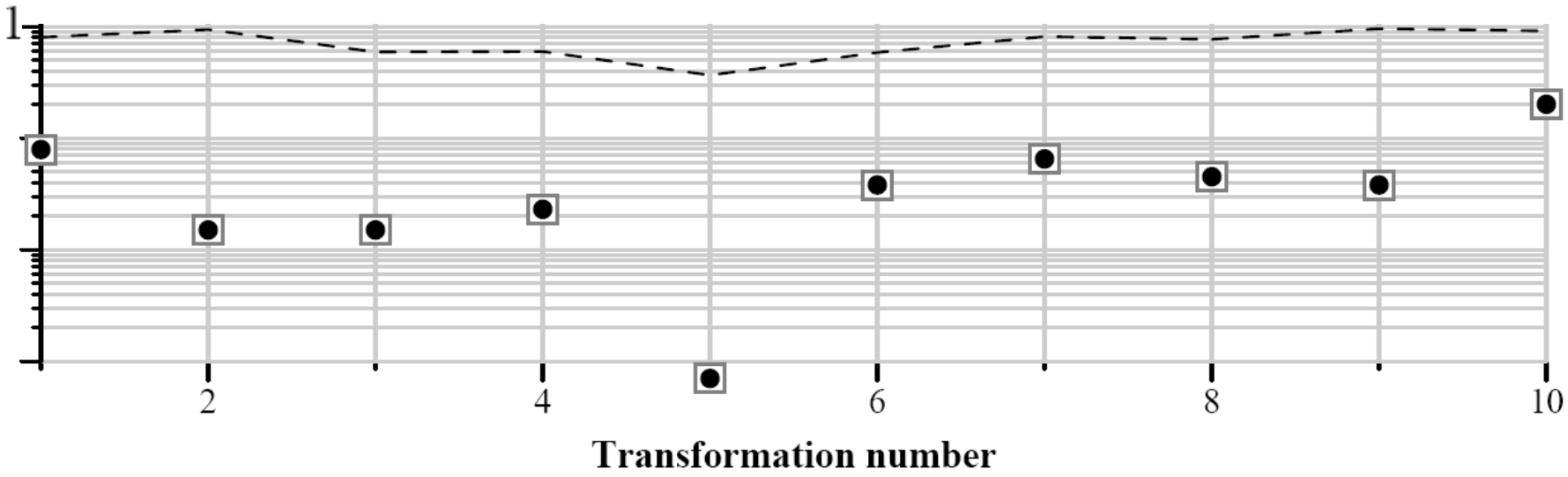
- Evaluation on our holidays dataset, 500 query images, **1 million** images in total
- Metric: mean average precision (in [0,1], bigger = better)



Average query time (4 CPU cores)	
Compute descriptors	880 ms
Quantization	600 ms
Search – baseline	**620** ms
Search – WGC	**2110** ms
Search – HE	**200** ms
Search – HE+WGC	**650** ms

Trecvid video copy detection task: evaluation results

- NDCR measure: the lower the best (0 = perfect)
- See our excellent results for all types of transformations below
  - circles: our result
  - squares: best results
  - dashed: medians of all runs (22 participants)



Conclusion

- HE: state-of-the-art representation of local descriptors
- WGC: use partial geometry information for billions descriptors

- See Internet demo (from http://lear.inrialpes.fr)
- Trecvid copyright detection task: we obtained excellent results

DEMO AND QUESTIONS