INRIA LEAR-TEXMEX: Copy detection task

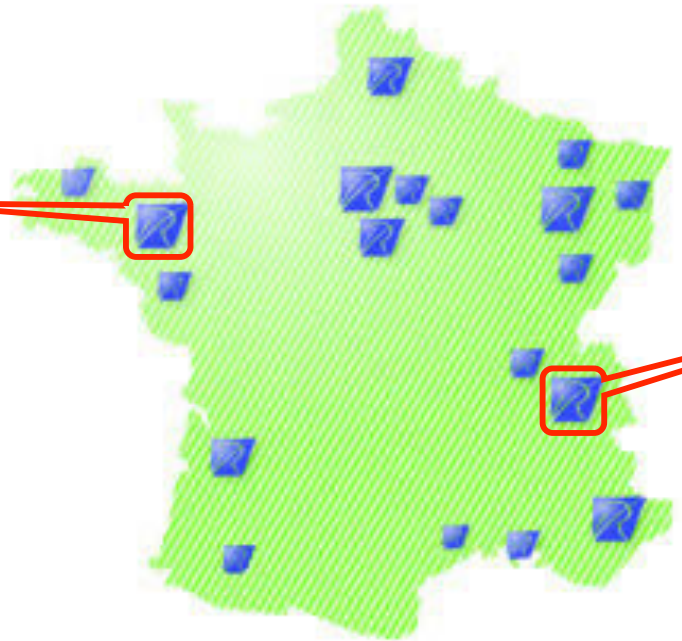
INRIA TEXMEX

(RENNES)

Hervé Jégou

Guillaume Gravier

Patrick Gros



INRIA Research centers

INRIA LEAR

(GRENOBLE)

Matthijs Douze

Cordelia Schmid

INRIA

Introduction

- INRIA participation in 2008: top results on all transformations
  - focus on accuracy + localization

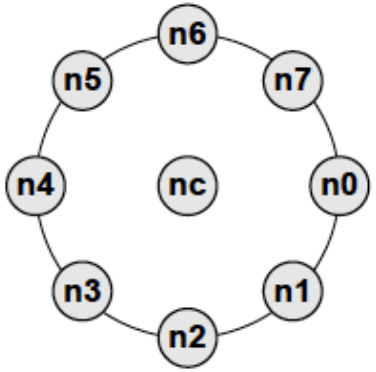
- Video:
  - same system as in 2008:
    An image-based approach to video copy detection with spatio-temporal filtering
    Douze, Jégou & Schmid, IEEE Trans. Multimedia 2010
  - + parameter's optimization

- Audio: new system (no audio in 2008's evaluation)
  - audio descriptors computed with standard package (spro)
  - novel approximate nearest neighbor search method

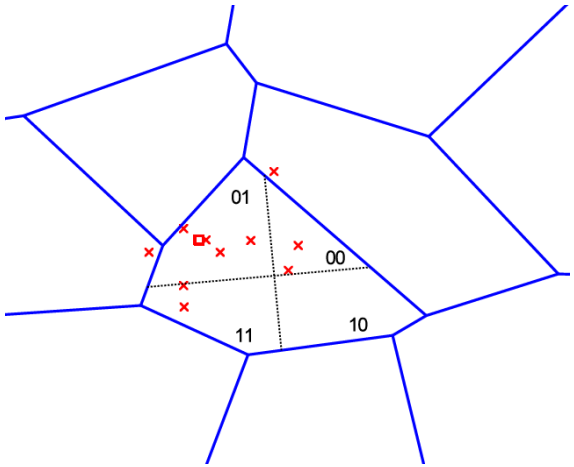
- In this talk:
  - brief overview of our video and audio systems
  - focus on our ANN method
  - comments on our results

Short overview of our video system: key components

- Local descriptors: CS-LBP
  - Heikkila et al., PR'2010



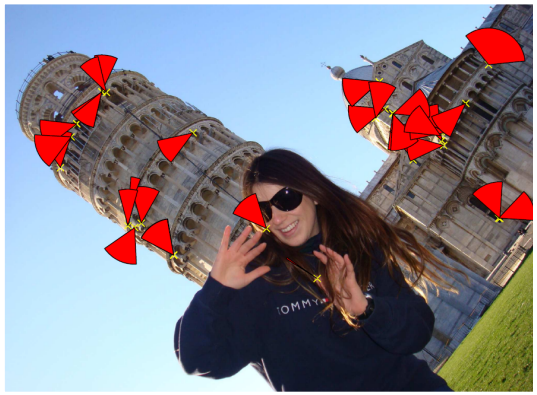
- ANN search: Hamming Embedding
  - Jégou et al., ECCV'08



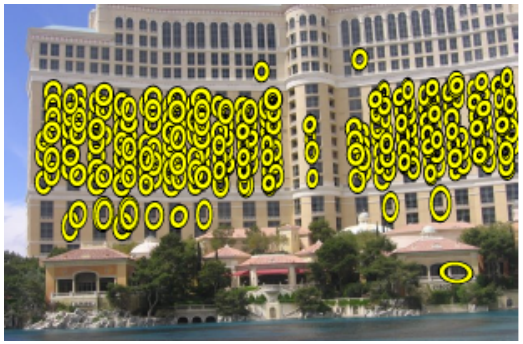
- Score regularization:

$$s_i = s_i \times \left(\frac{s_i}{\max_j s_j} \right)^{\alpha}$$

- Weak geometric consistency
  - Jégou et al., ECCV'08



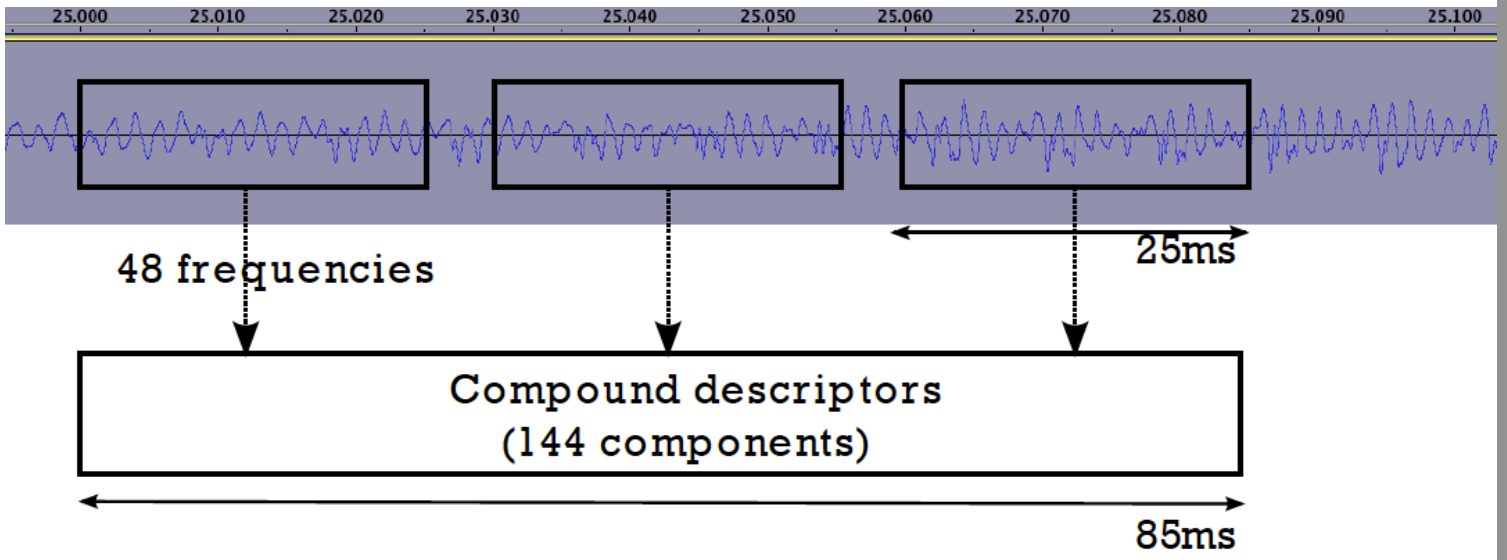
- Burstiness strategy + Multi-probe
  - Jégou et al., ICCV'09



- Spatio-temporal fine post-verification
  - Douze et al., IEEE TMM'10

Short overview of our audio system: key components

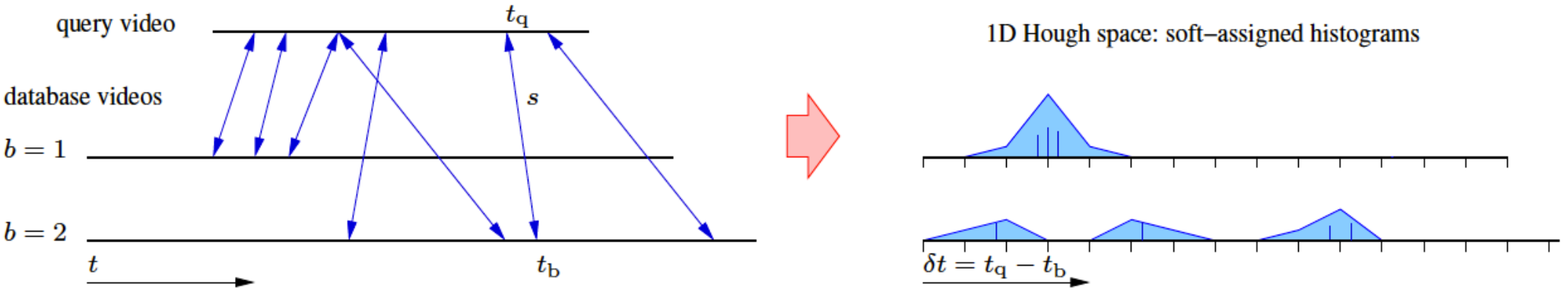
- Descriptors
  - filter banks
  - Compounding
  - energy invariance
  - 1 vector /10 ms



  - online package: https://gforge.inria.fr/projects/spro, filter banks, MFCC, etc

- Novel ANN search based on compression paradigm: see next slides

- Temporal integration: Hough voting scheme (votes in histogram Δt=tb-tq)



Video parameter optimization

OBJECTIVE: improve precision with "reasonable" cost w.r.t. efficiency

- Decreasing detector threshold
  - number of descriptors ↗
  - complexity ↗
  - precision ↗ (with HE)
  - threshold: T200 or T100
- Describe flip/half-sized frames
  - on database side only
  - threshold: H200 or H100
- Multiple assignment (=multi-probe)
  - on query side only

mAP on a validation dataset

query	database		
	T200	T200 +H200	T200 +H100
T200	0.483		
T100	0.514	0.568	0.583
T100+flip	0.627	**0.719**	0.738
T100+flip, MA10	0.683	0.749	0.737
T100+flip, MA3	0.650	0.755	**0.761**

Observation:
- half sized and flipped frame help a lot
- small multi-probe (x3) is sufficient

Note: generic system
- only flipped is specifically to

INRIA

Huge volumes to index: approximate nearest neighbor search

index size (database)		
Video, T200	d=128	2.48 billion descriptors
Video (half, H100)	d=128	0.97 billion descriptors
Audio	d=144	140 million descriptors

→ Need for powerful approximate search

- Locality Sensitive Hashing: memory consuming, need for post-verification on disk, not very good trade-off between precision/efficiency
- FLANN: excellent results, memory consuming, need for post-verification (on disk given the dataset size)
- We used:
  - Video: Hamming Embedding with 48 bits signature (10B/descriptors+geometry)
  - Audio: Compression based approach → Product quantization method

Indexing algorithm: searching with quantization [Jegou et al., TPAMI'11]

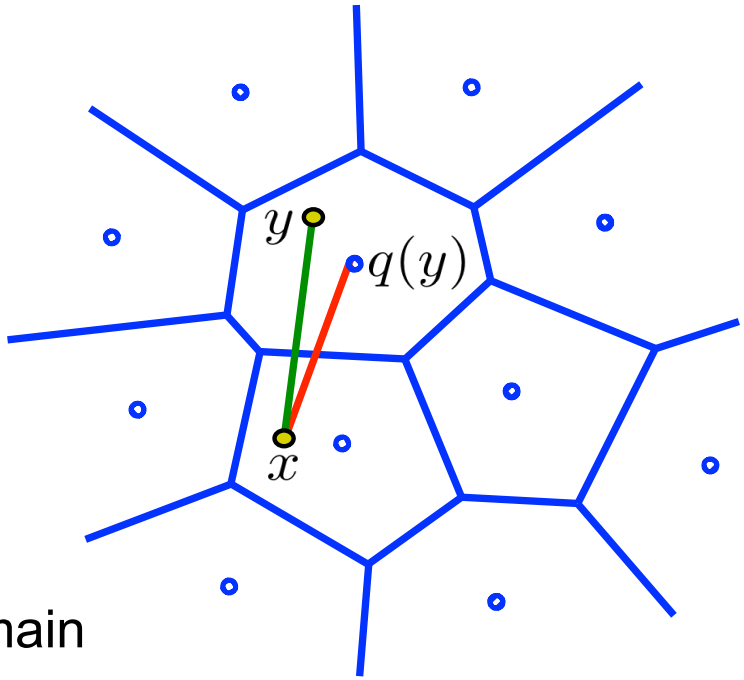
Purpose: approximate NN search **with limited memory** (and no disk access)

- Search/Indexing = distance approximation problem
- The distance between a query vector *x* and a database vector y is estimated by

$$d(x, y) \approx d(x, q(y))$$

where q(.) is a fine quantizer

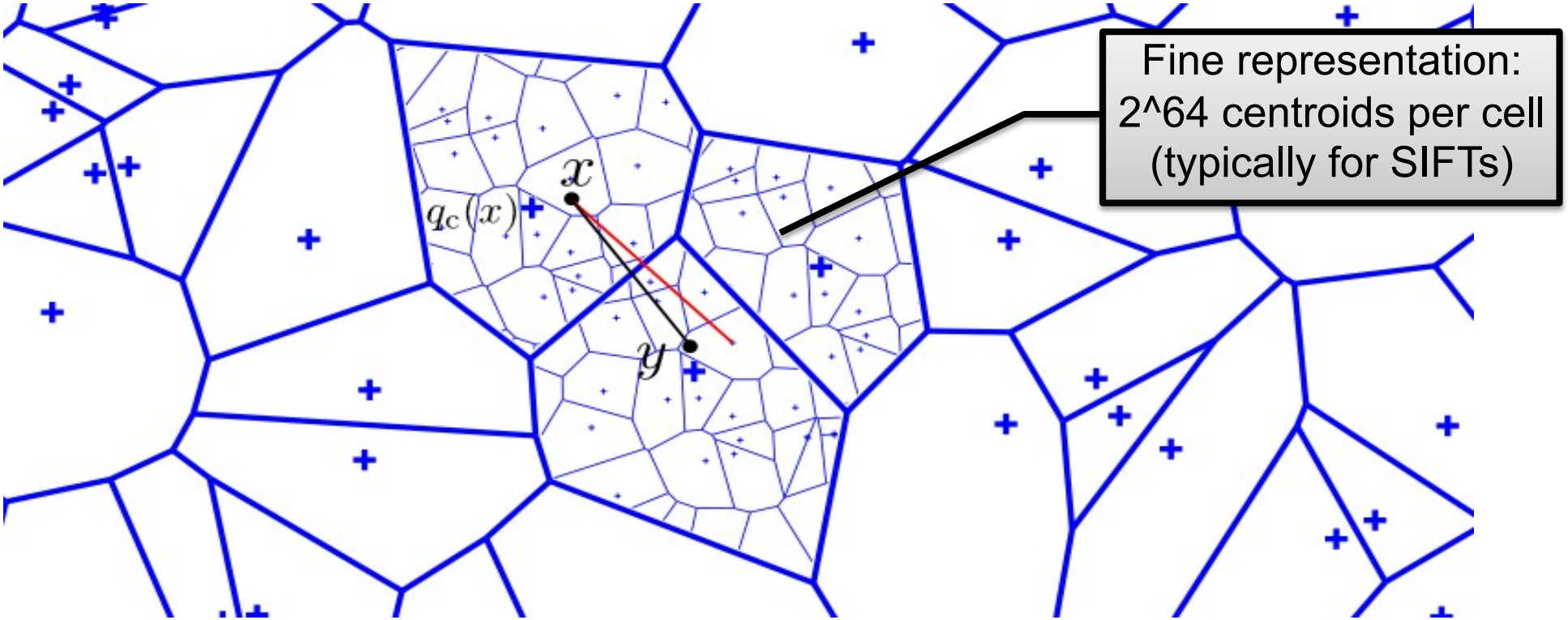
→ vector-to-code distance



- Distance is approximated in compressed domain
  - typically 8 table look-ups and additions per distance estimation (for SIFTs)
  - **proved statistical upper bound** on distance approximation error

Indexing algorithm: searching with quantization [Jegou et al., TPAMI'11]

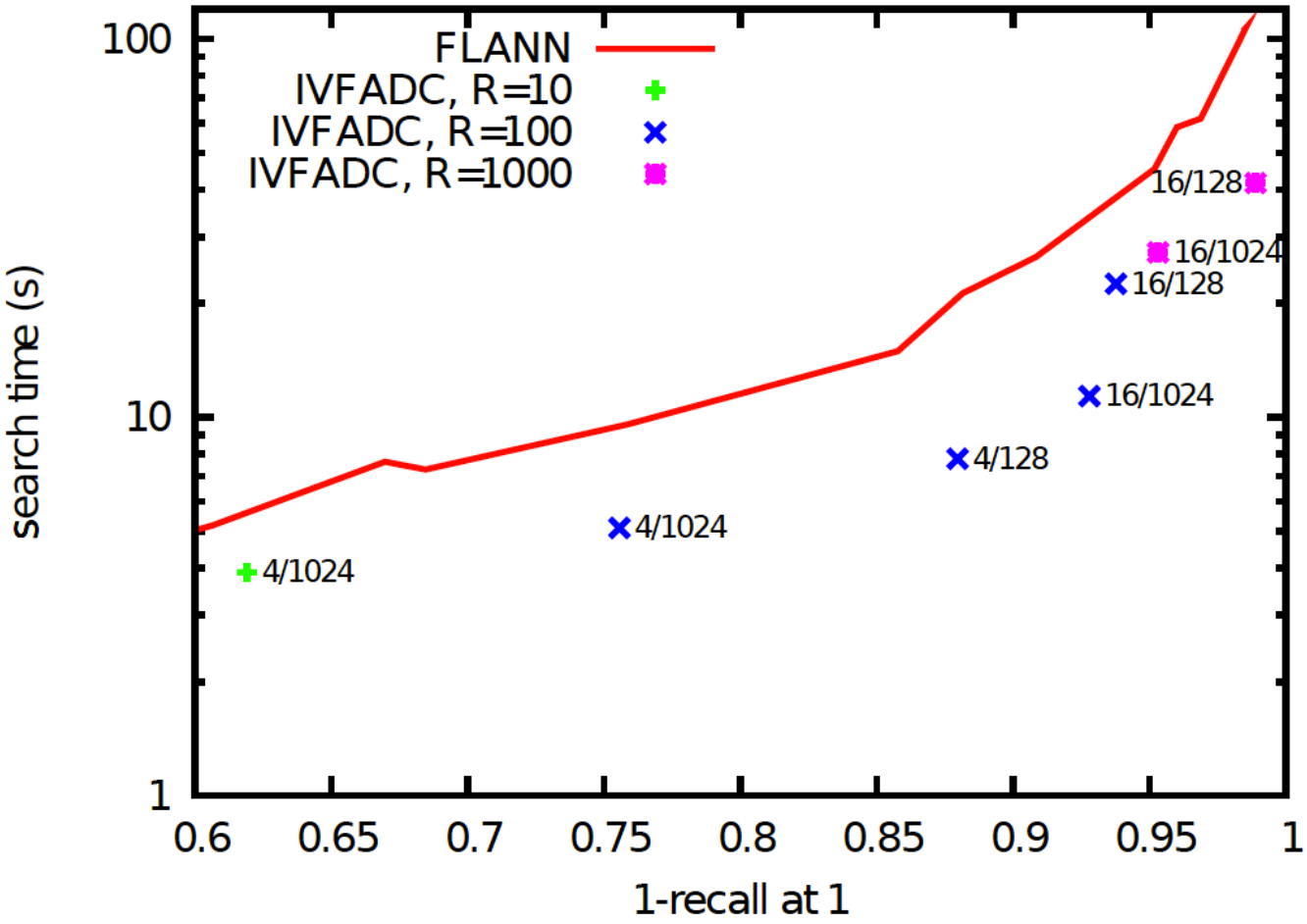
- Combination with inverted file: coarse quantizer to avoid scanning all elements
- Here: MA=3



- Efficient search: searching in 2 billion SIFT vectors (with MA=1)
  - This method: 3.4 ms / query vector
  - HE: 2.8 ms / query vector

Comparison with FLANN [Muja & Lowe'09]

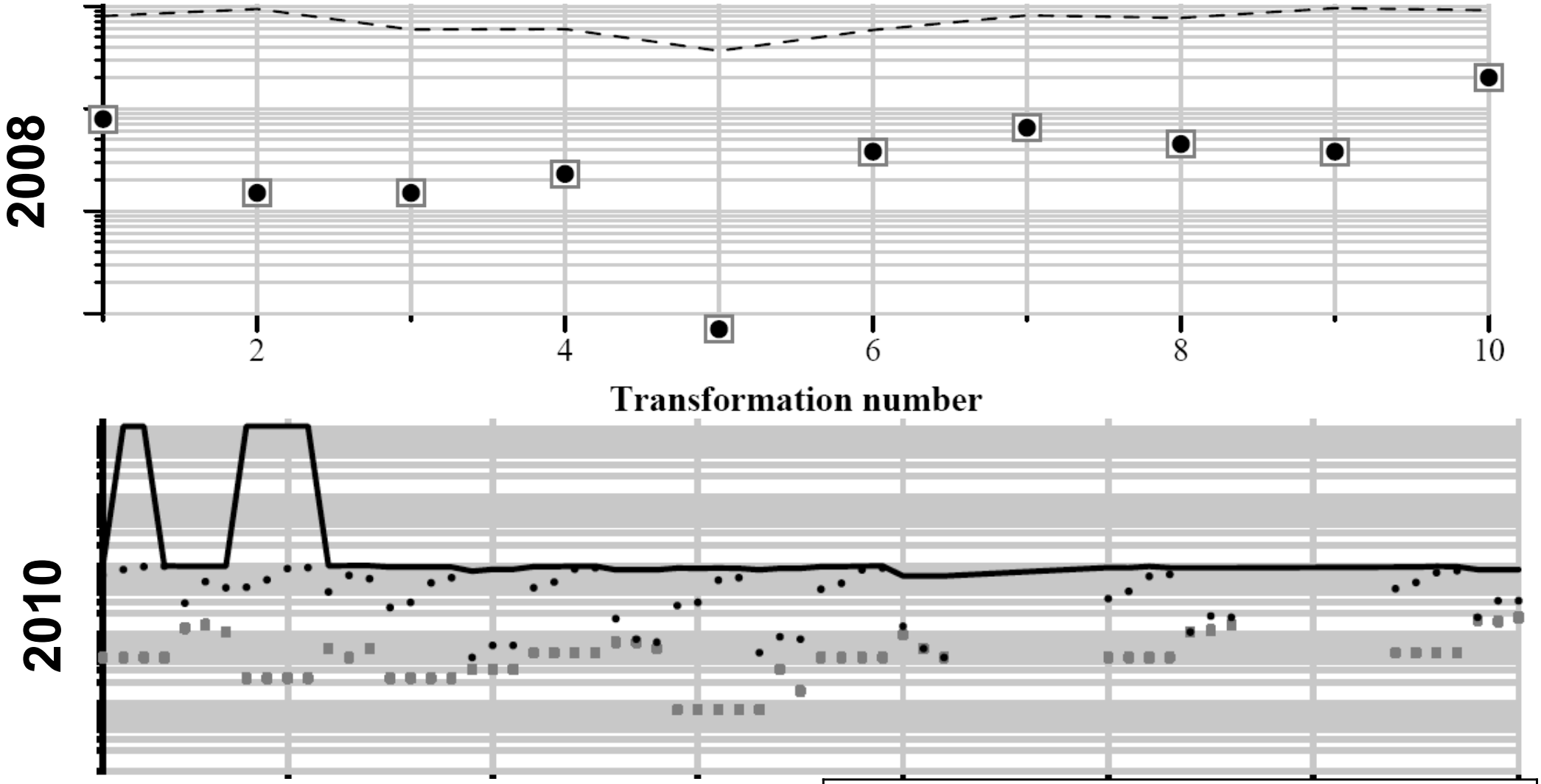
- Tested on 1 million SIFTs



- 1.5 to 2 faster than FLANN for same accuracy

- Memory usage for 1M vectors (according to "`top`" command):
  - FLANN: > 250MB
  - Ours: < 25MB

NDCR: Comparison between 2008 and 2010



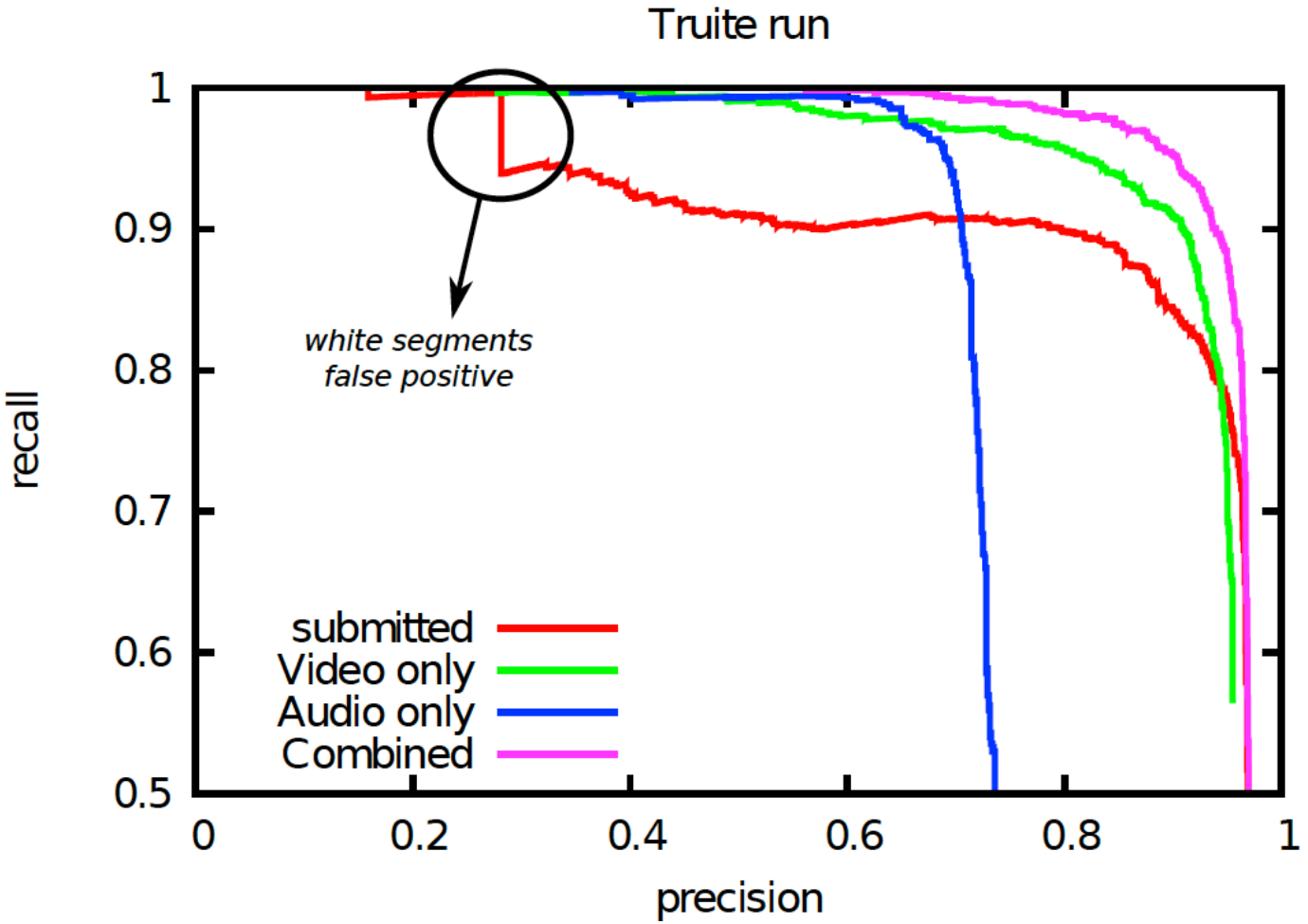
Ranks / 22 participants (BAL, Opt_NDCR)					
Rank	1st	2nd	3rd	4th	5th
#	6	10	19	18	2

- Huh?! What's the problem?

  "Bug": a few false positive videos are returned frequently with very high scores

INRIA

Results on Trecvid: sub-optimality of our approach



- Problem with audio: pseudo-white segments → corrupts similarity measure
- Fusion based on invalid assumptions:
  - two first runs: audio and video assumed to have similar performance
  - two last runs: audio assumed to be better than video

Conclusion

- We have learned many things this year:
  - actual decision threshold: need for « cross-databases » setting method
  - audio helps a lot (when working)
  - fusion module is very important
    - audio ≠ video, room for improvement by score normalization
    - strong bonus when both agree

- What's might interest the other participants in what we have done
  - approximate nearest neighbor method for billion vectors

- Online resources:
  - spro: library for audio descriptors
  - Matlab toy implementation of our compression based search method
  - BIGANN: a billion sized vector set to evaluate ANN methods
  - GIST descriptor in C: OK for several copy transformations
    [Douze et al., CIVR'09, IBM Trecvid'10]