Indexing Local Configurations of Features for Scalable Content-Based Video Copy Detection

Sebastien Poullot, Xiaomeng Wu,
and Shin'ichi Satoh
National Institute of Informatics (NII)
Michel Crucianu,
Conservatoire National des Arts et Metiers (CNAM)

Goals and choices

- Priority: speed → scalability
- Quality, MinDCR = 0.5
- Choices
  - Frame selection → keyframes (3000 per hour)
    - Depending on global activity changes
  - Flipped keyframes in ref database
    - Descriptors not invariant

Goals and choices

- Priority: speed → scalability
- Quality, MinDCR = 0.5
- Choices
  - PoI → Harris corner
    - Fast computation, but noise and blur sensitive
  - Local descriptors → spatio-temporal local jets
    - Fast computation, but not scale invariant, and frame drop sensitive
  - Global description → scalability
    - Smaller database → search faster
    - No vote process at frame level
  - Indexing → scalability

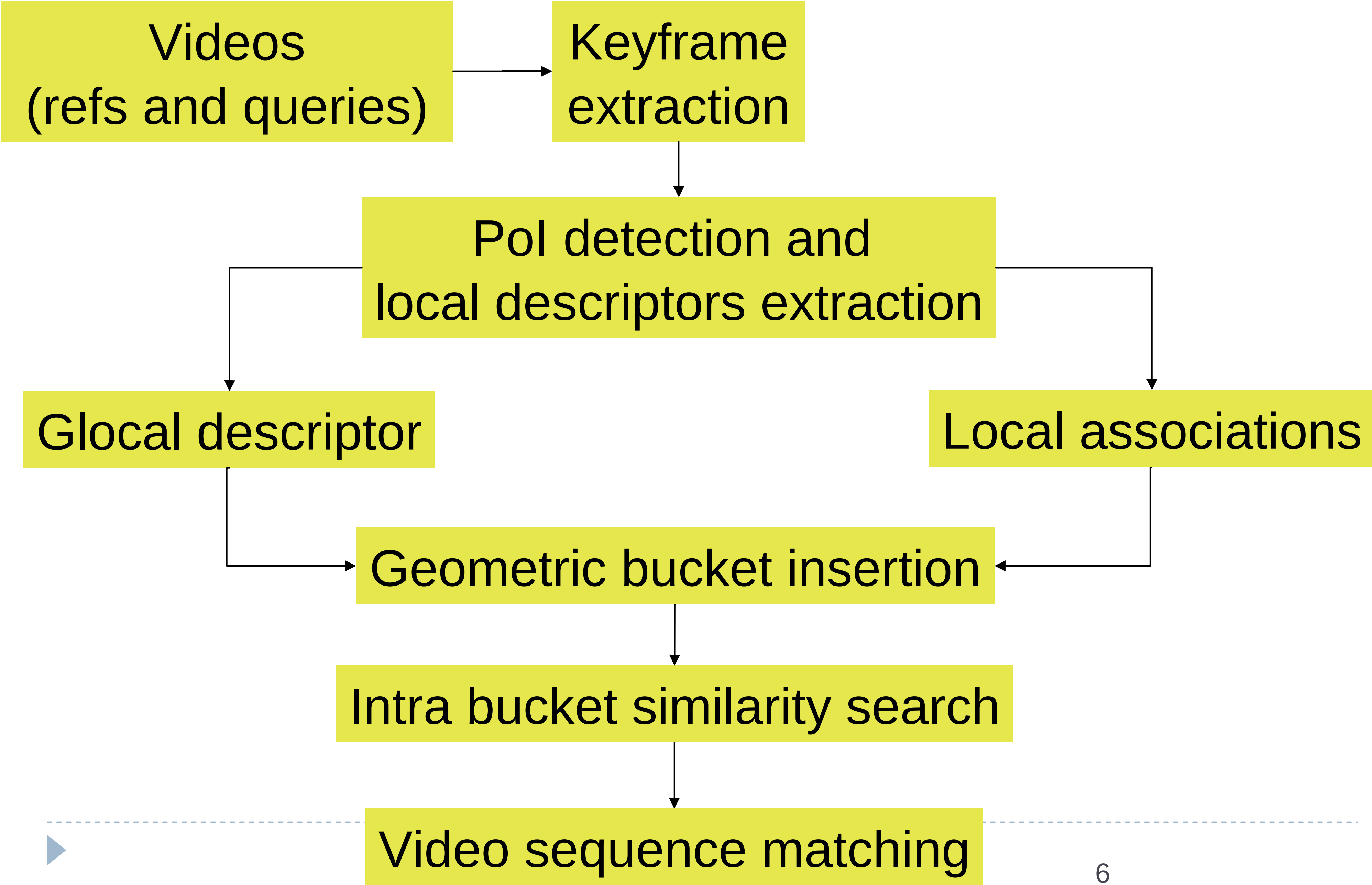
Goals

- A video description at frame level using local features: Glocal (alternative to BoF)
  - An interesting trade off scalability / accuracy
- An indexing scheme based on associations of local features
  - Reduce bad collisions
- A simple shape descriptor
  - Filter out remaining bad collisions

→ scalability and accuracy

Method

Processings



Local features

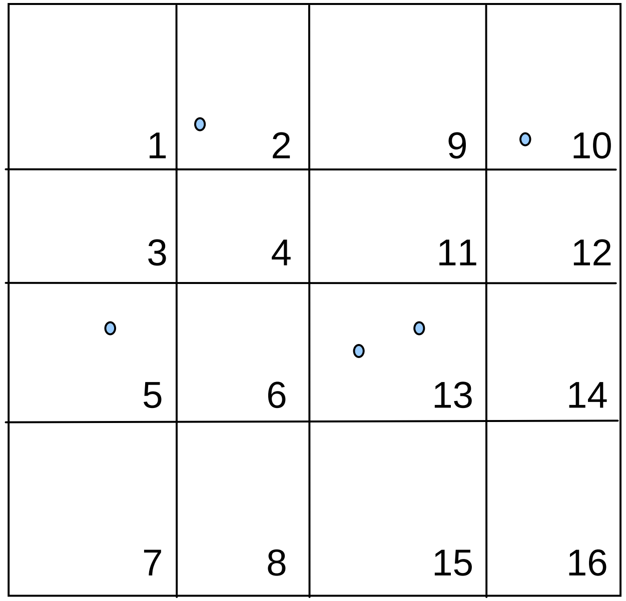
- Points of Interest: Harris corner (could be DoG, Hessian, etc)
  - Local Descriptors at these positions: SpatioTemporal Local Jets (could be dipoles, SIFT, GLOH, etc)

→ a set of descriptors associated to
a set of positions
(d1,p1), (d2,p2),..., (dn,pn)

Quantization of local features

- Quantization of the descriptors (d_i,p_i,q_i)

→ use a parameterized Zgrid (based on distributions)



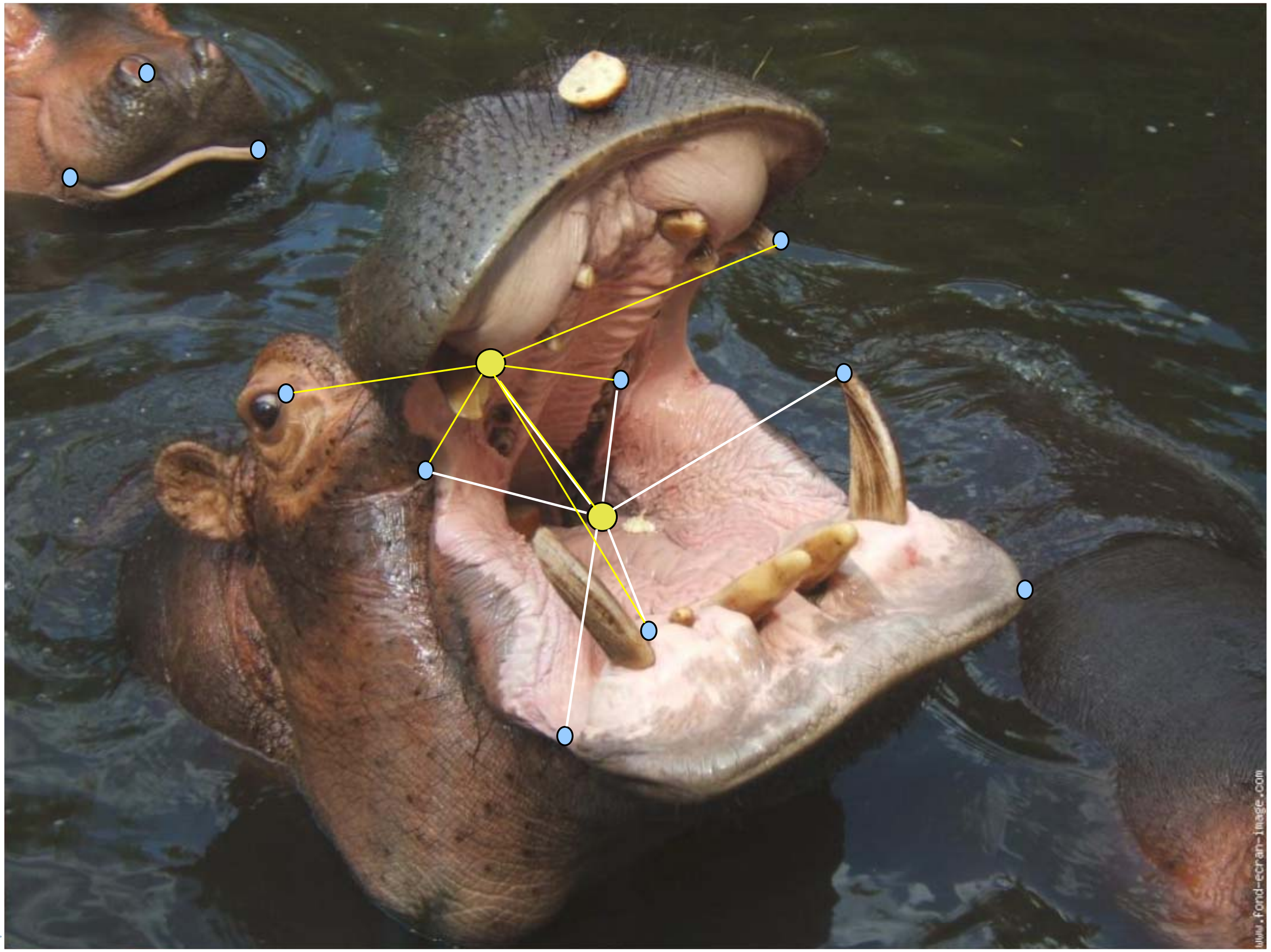
0100000000000000 D=4

0100100001001000 D=4

- Keyframe Glocal description = sum of quantizations of features
- Small descriptor and vocabulary (D=10, 1024 bits / 1024 words)
- No clustering needed

Combining local features

Construction of N-tuples using K-NN in image plane



$P_1 – P_{1NN1} – P_{2NN1}$

$P_1 – P_{3NN1} – P_{4NN1}$

$P_1 – P_{5NN1} – P_{6NN1}$

$P_2 – P_{1NN2} – P_{2NN2}$

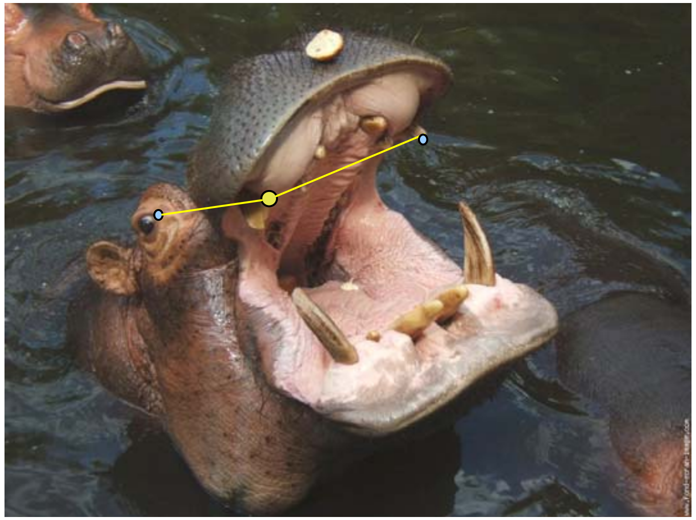
$P_2 – P_{3NN2} – P_{4NN2}$

$P_2 – P_{5NN2} – P_{6NN2}$

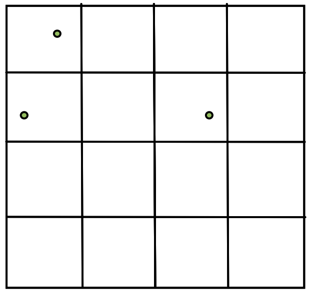
Combining local features

- PoI: up to 150 / keyframe
- Up to 5 triplets / PoI (1NN&2NN,..., 9NN&10NN)
- Up to 750 associations per keyframe
- Some redundancy appears → average = 650 associations
  - Glocal descriptors inserted in 650 buckets
    - Bucket choice depends on PoI
  - Buckets defined by quantization of descriptors
    - Bucket definition depends on local descriptors

Bucket definition

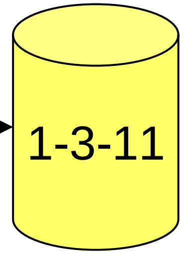


Local descriptors quantified
in description space



1010000000100000
Positions 1, 3 & 11

Bucket

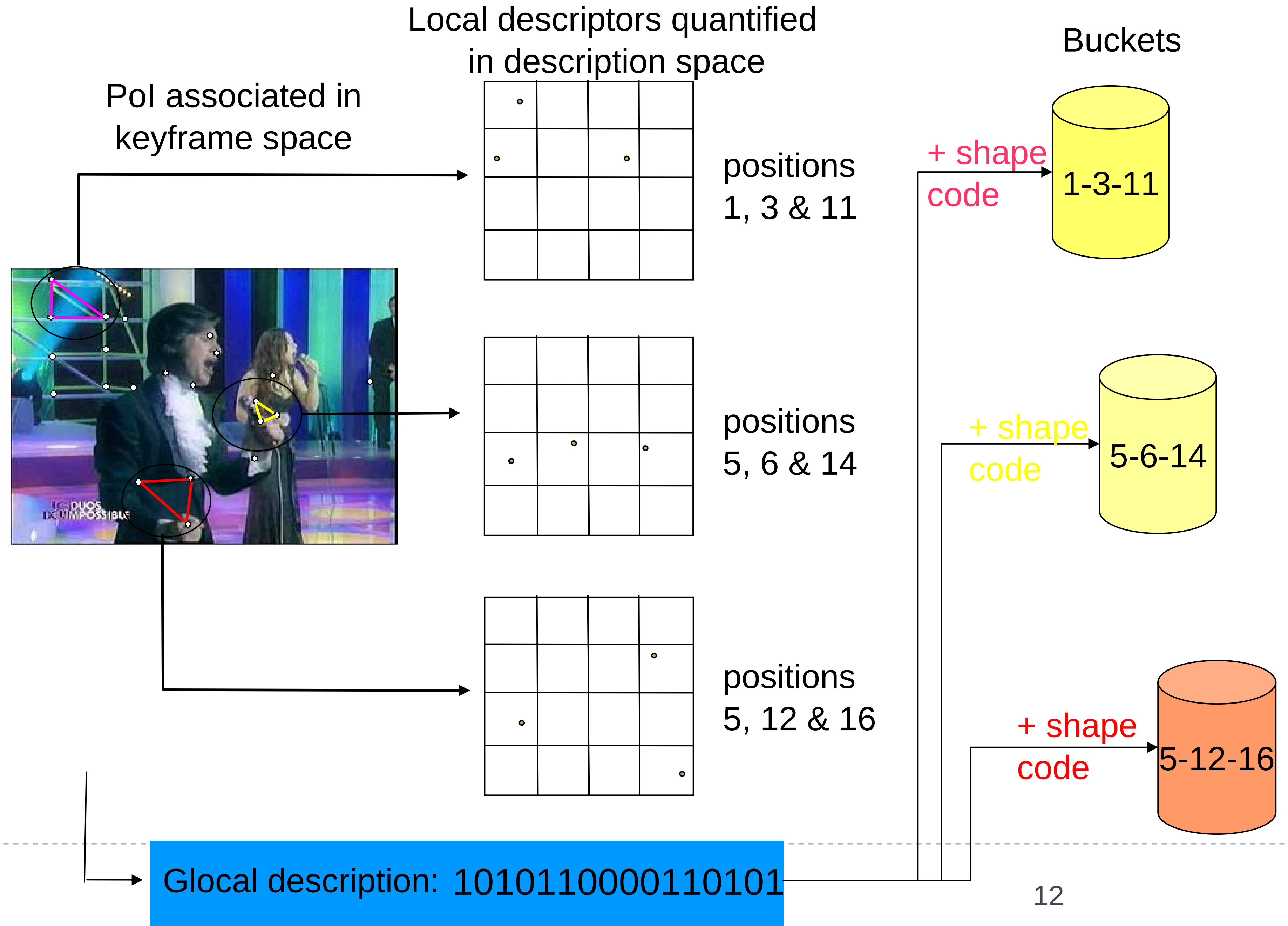


Glocal descriptor

Number of possible buckets $N_B = \frac{\left(2^d\right)^3}{L!}$ where L = sentence length

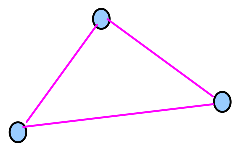
Trecvid: d=10, L=3 → N_B = 178.10e6

Indexing method

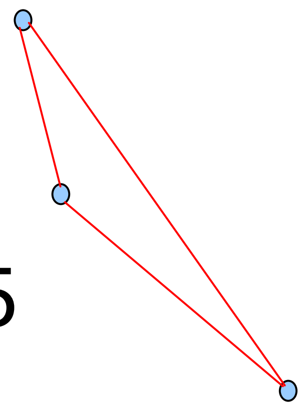


Weak shape code

- Ratio between longer and smaller side (>=1)



~ 1



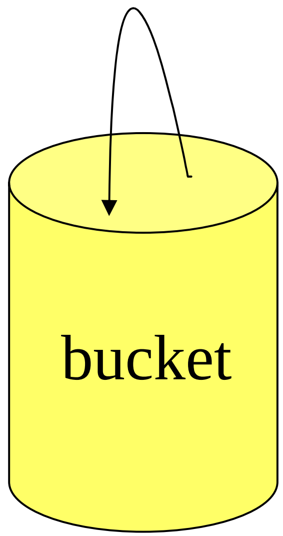
~ 2.5

- Allow to distinguish different local configurations: more or less flat

Intra bucket similarity search

- Bucket = list of Glocal Descriptor Gi.(q, sc, tc)

- In each bucket, only between refs and queries, compute:
- - correspondence between shape codes
- (filtering)
- - similarity



For each couple of Glocal descriptor (G_x, G_y)

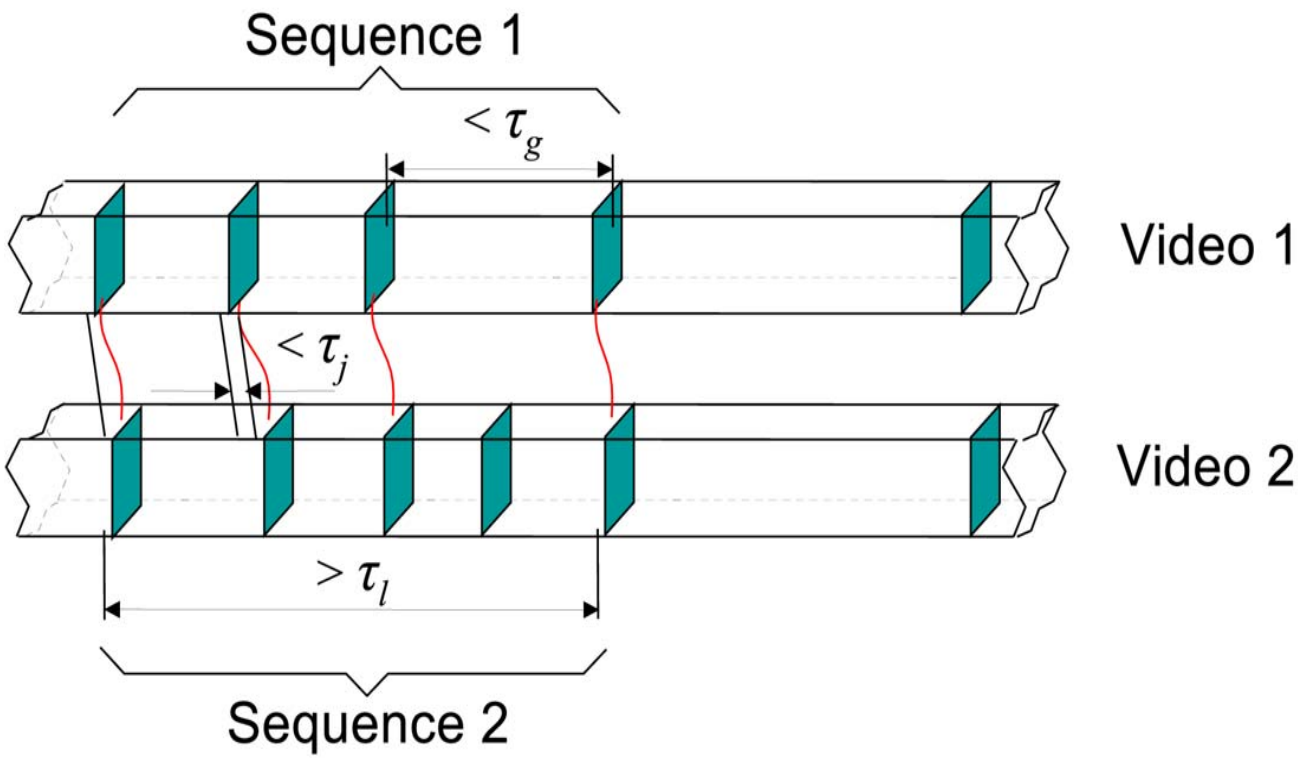
if (G_x.sc ~ G_y.sc)

then if (Sim(G_x.q, G_y.q) > Th)

Keep (G_x.(id,tc), G_y.(id,tc))

Matching Video Sequence

Between two videos find temporal consistency of keyframes



- Number of couples of matching keyframe >= τ_l
- Blank between two successive pairs of matching keyframes <= τ_g
- Offset between two successive pairs of keyframes <= τ_j

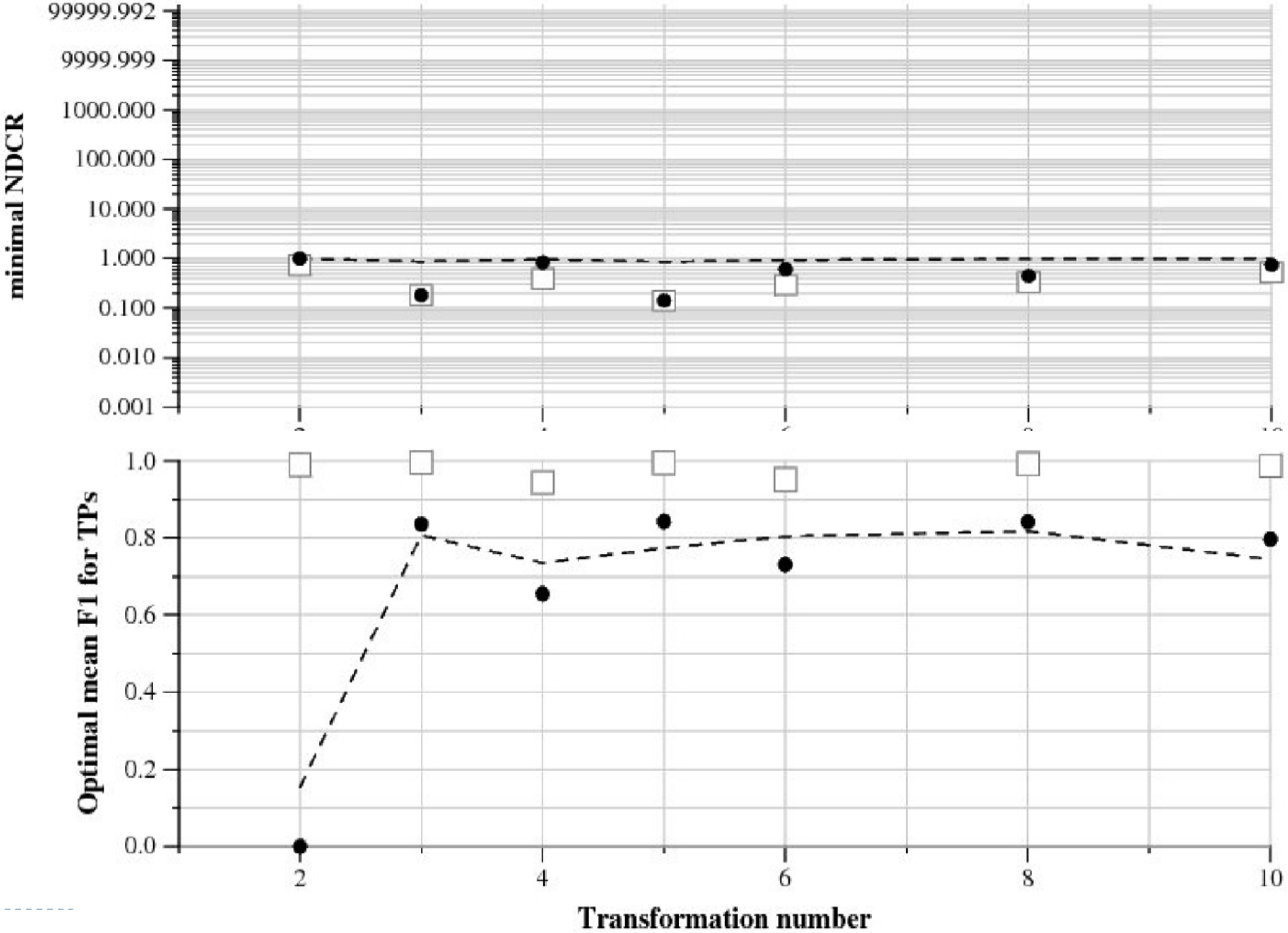
Computation costs

- Extraction of keyframes: 1/25 of real time (rl)
- Computation of descriptors: 1/50 rl
- Construction of reference database: 1/200 rl (offline)
- Query: 1/150 rl

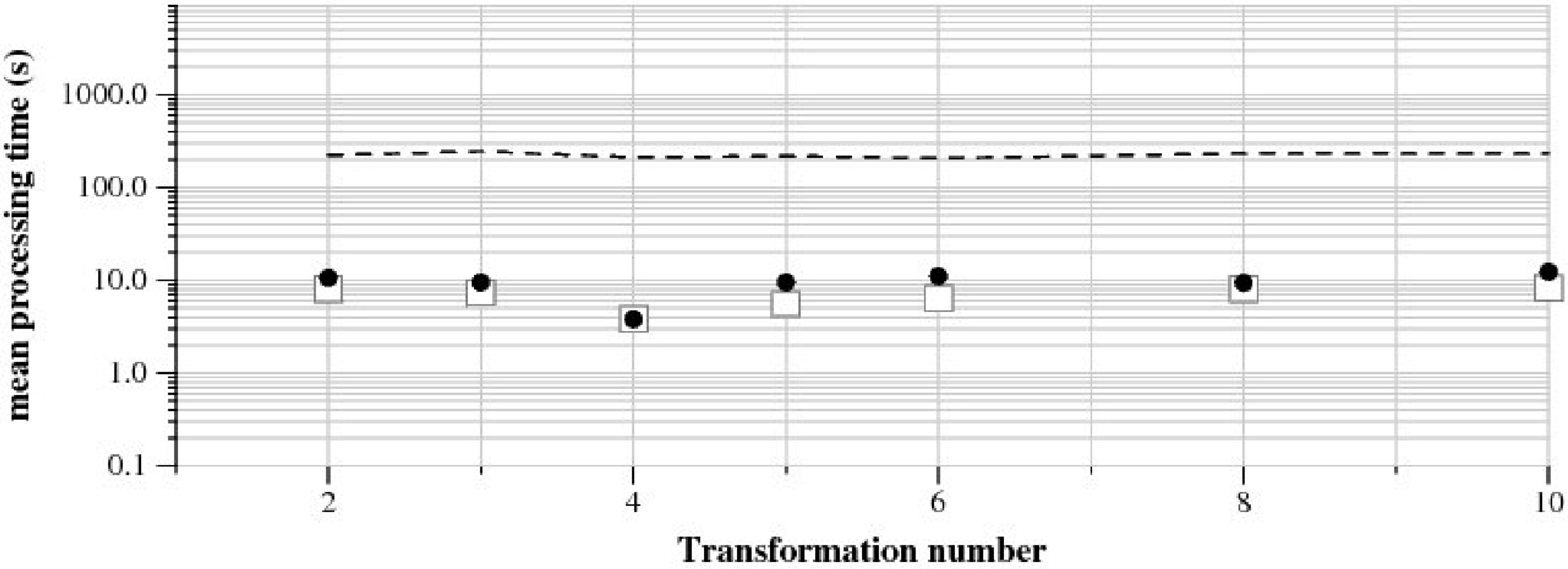
→ limits: keyframes extraction process and descriptor computation

Results

Results - Balanced

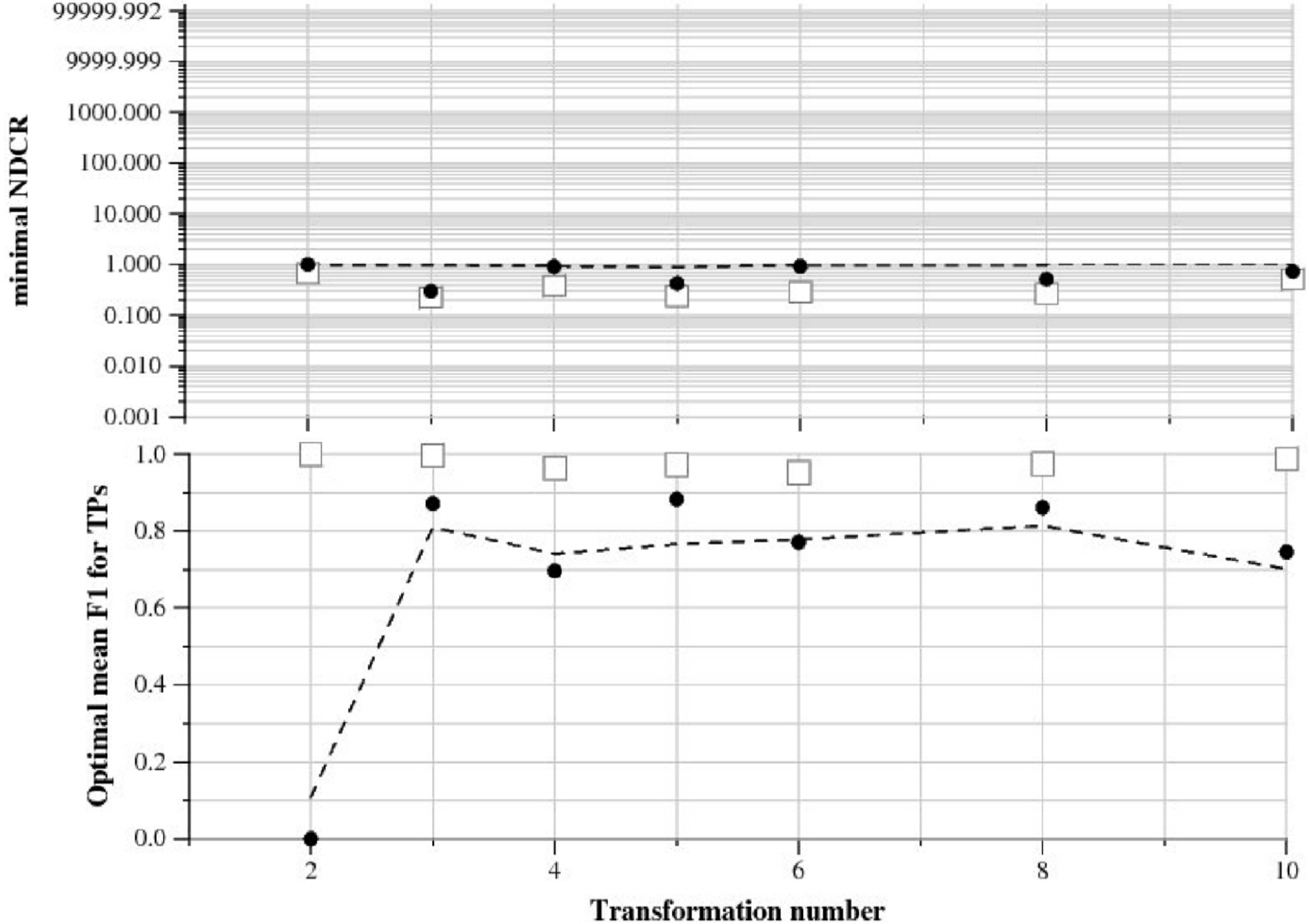


Results - Balanced

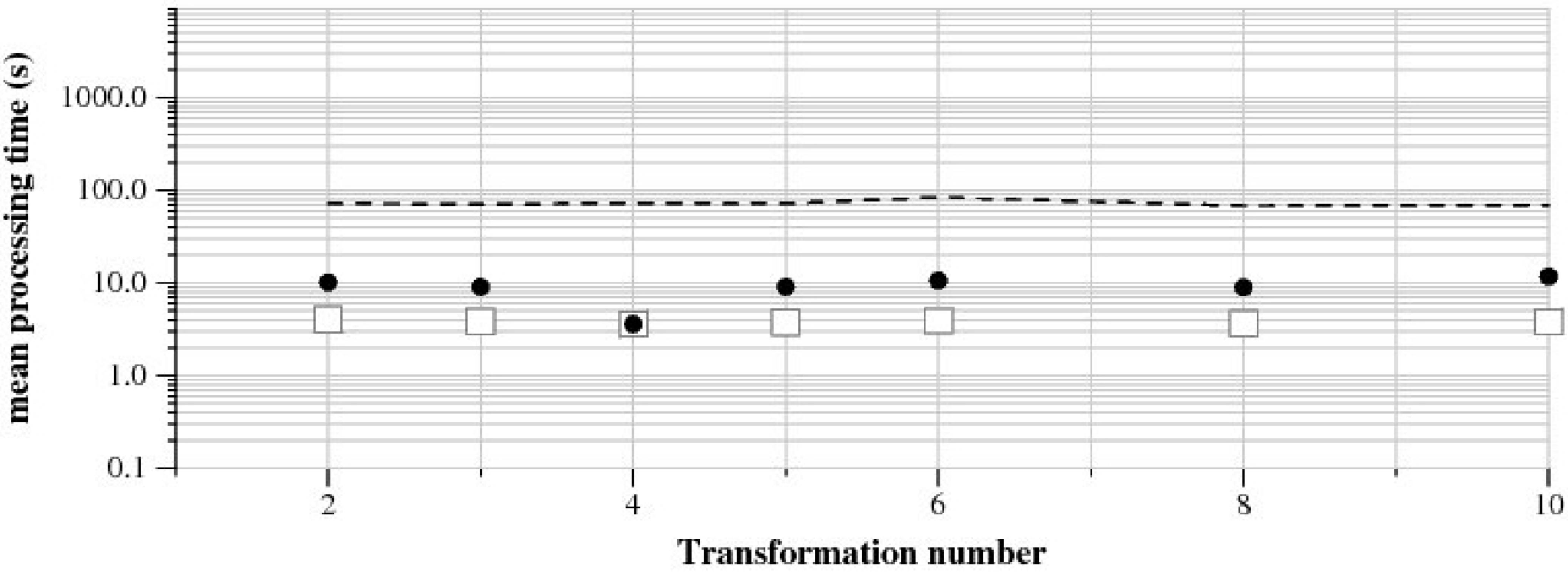


Computer: laptop - core2Duo@2.6Ghz - 4Gb RAM – HD 5400RPM

Results – No False Alarm



Results – No False Alarm



Computer: laptop - core2Duo@2.6Ghz - 4Gb RAM – HD 5400RPM

Conclusion

- Glocal description is relevant
- Local associations of features for indexing gives nice accuracy and good scalability to CDVCB
- Weak shape embedding dramatically scales up CDVCB with small loss of recall and high gain of precision (2/3 of similarities avoided, FA/10)
- Method has proven its possibility
  - TRECVID09 CBVCD task
  - 3000h database similarity self join (global 6 hours)

Future works

- Further association of PoI and Descriptors to test (Hessian, SURF, Dipoles, etc)
- Other weak geometric concept
- Try the method to other fields
  - Objects (BoF) – near duplicates
  - Pictures
- Extraction of knowledge on large databases

Thank you for attention