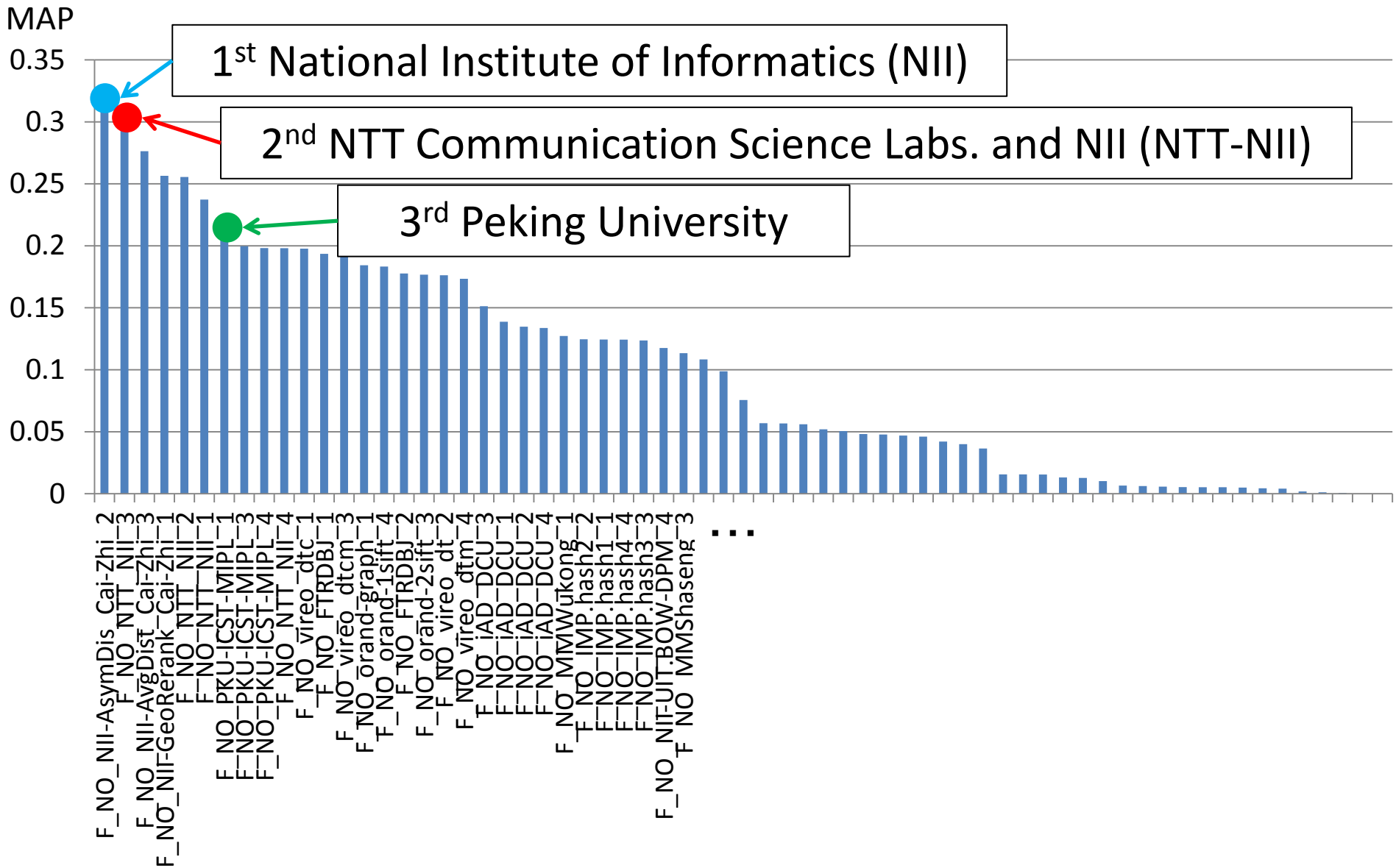
NTT Communication Science Laboratories and National Institute of Informatics at TRECVID2013 Instance Search Task

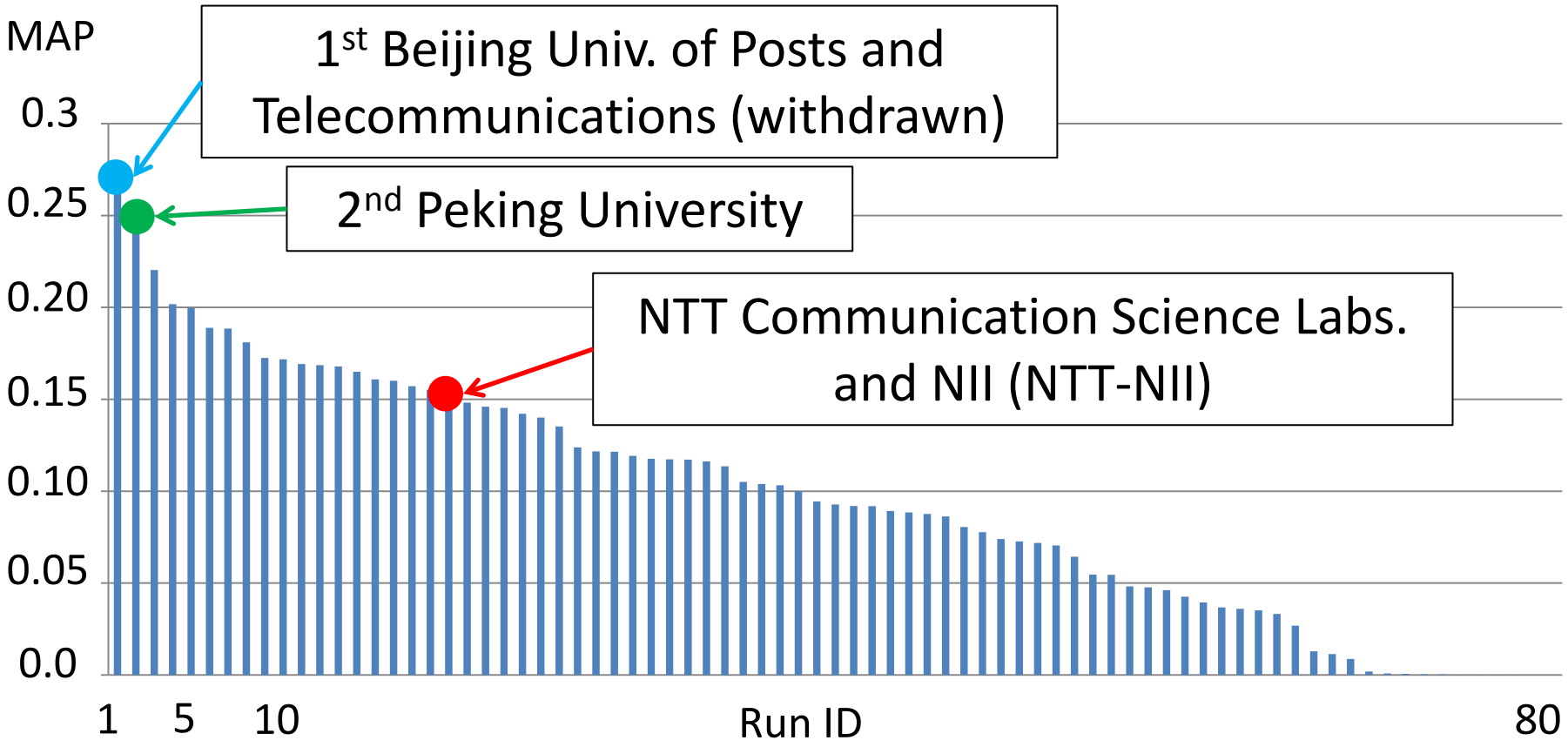
Masaya Murata, Hidehisa Nagano, Kashino Kunio and Shin'ichi Satoh

Speaker: Masaya Murata (NTT CSL, Japan)

This year, our best run is ranked 2nd



Previous year, our run was ranked at upper-middle



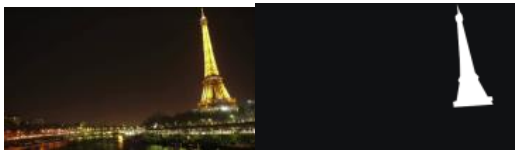
Primary differences from the previous year's instance search method

- Previous method
  - Used only one descriptor to feature image key-points: Color SIFT vector
  - Used cosine similarity value of 0.95 as a threshold of the key-point matching
  - Used standard BM25 video ranking function
    - key-point weight is based on standard Inversed Document Frequency (IDF)
- New method
  - Used two descriptors to feature image key-points: SIFT and Color SIFT vectors
  - Used cosine similarity value of 0.9(softer) as a threshold of the key-point matching
  - Used our new BM25 video ranking function called exponential BM25
    - key-point weight is based on exponential IDF which is designed to suppress the effect of the noisy (unnecessary) key-points
- Trial run
  - Multimodal instance search using audio and textual information of videos

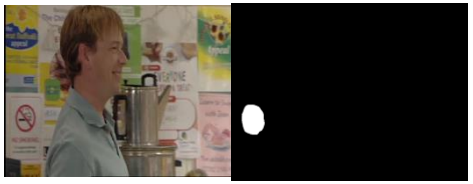
Dataset has also changed

- Previous year
  - Queries: 21 instances
    - object 15, people 1, place 5
    - five images per instance on average
  - Videos: about 77,000 clip movies from Flickr.com
    - consumer generated media
    - average clip duration is 10 sec
- This year
  - Queries: 30 instances
    - object 26, people 4, place 0
    - four images per instance on average
  - Videos: about 470,000 shots from BBC's drama
    - EastEnders (professional media)
    - average shot duration is 1~3 sec

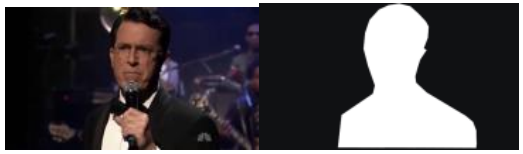
instance "Eiffel tower"



instance "a circular 'no smoking' logo"



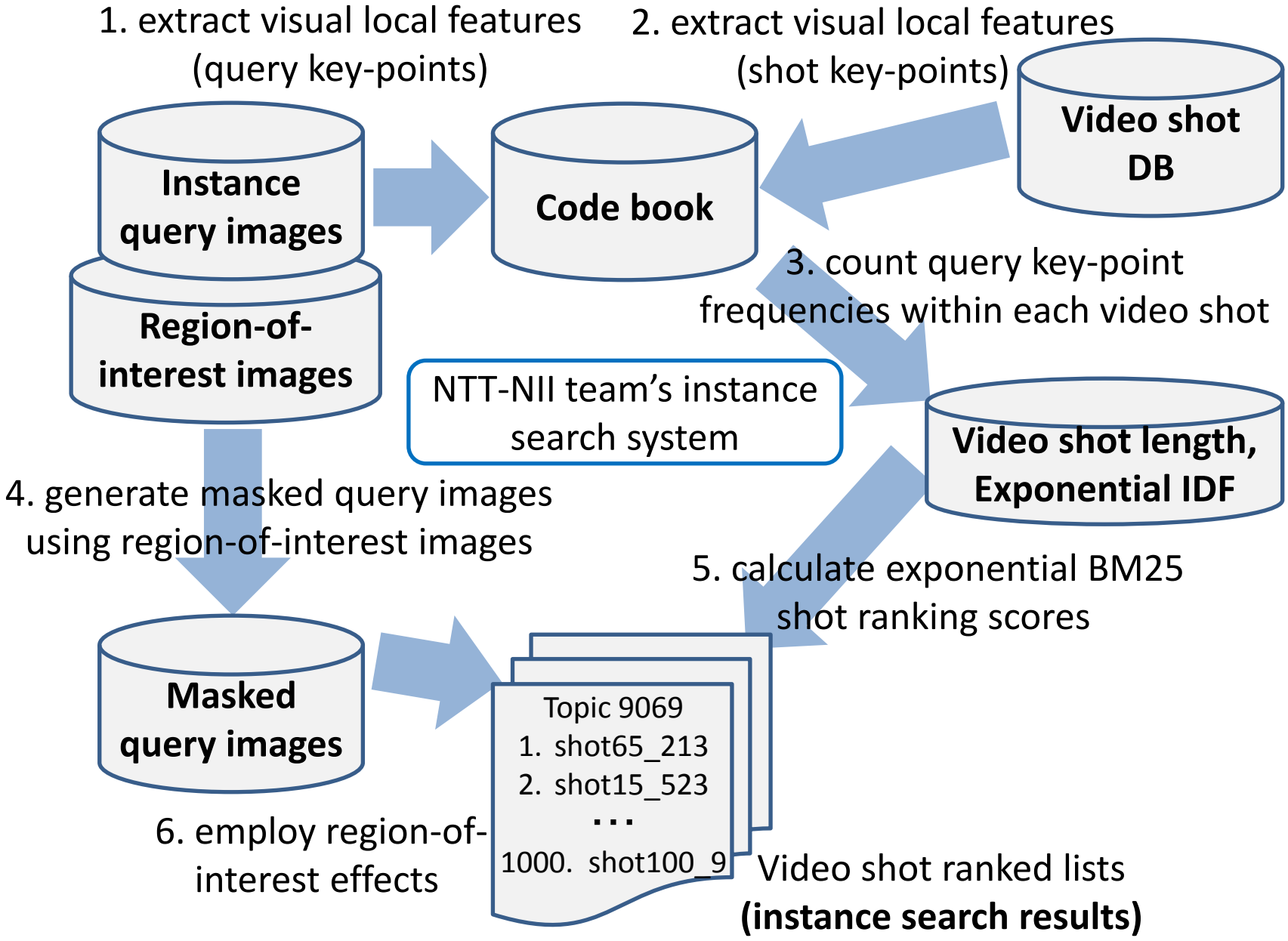
instance "Stephen Colbert"



instance "an Audi logo"

instance "this man"

Our instance search system



1. extract visual local features (query key-points)

- Key-point detector
  - Harris-Laplace detector
- Key-point descriptor
  - 128-dimensional SIFT vector
  - 192-dimensional color SIFT(compact color SIFT) vector
- Code book
  - key-points extracted from all of the query images (duplicated key-points are removed)
  - merit: no quantization errors are included on the query key-points → we prioritize search precision
  - demerit: query images are necessary a priori

2. extract visual local features (shot key-points)

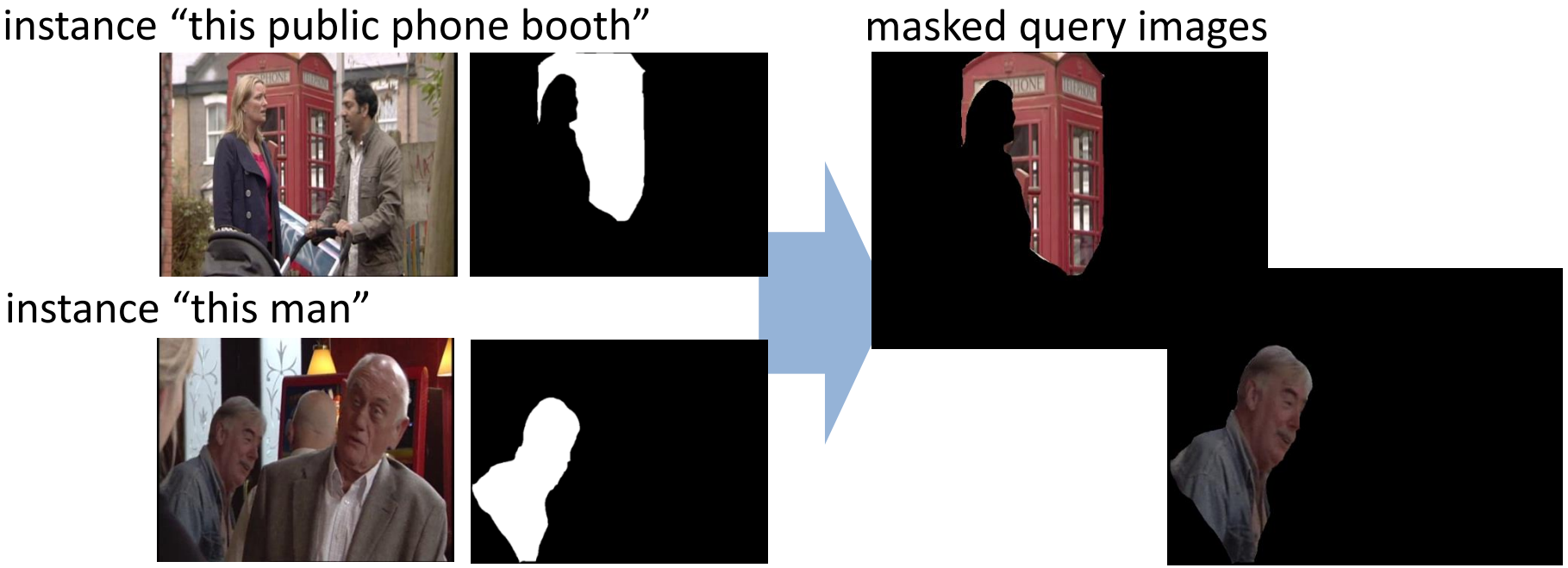
- Key-point detector
  - Harris-Laplace detector
- Key-point descriptor
  - 128-dimensional SIFT vector
  - 192-dimensional color SIFT(compact color SIFT) vector
- Matching against Code book
  - key-point pairs whose cosine similarity values are 0.95 or more are considered as matched
  - 360,000 (code book) vs 4,100,000,000 (video shot DB) pair-wise high dimensional vector matching for SIFT and color SIFT!

3. count query key-point frequencies within each video shot

- Count the number of times the query key-points (visual words) appear in each video shot
- Then,
  - obtain the within-shot query key-point frequencies
  - calculate the query key-point importance weights such as the inversed document frequency (IDF)
    - Our method used an extended version of the IDF, called exponential IDF, which is designed to suppress the contributions from noisy (unwanted) key-points to the instance search results

4. generate masked query images using region-of-interest images

- Region-of-interest (ROI) images distinguish the instance search task from the other tasks such as near-duplicate or similar video searches.
- Superimpose the ROI images on the instance query images to generate the masked images (right examples)



5. calculate exponential BM25 shot ranking scores

- Probabilistic information retrieval model (PIR)

$$P(REL = rel \mid v, q) \propto_q \sum_{(q, kf_i > 0)} \frac{kf_i}{kf_i + \kappa} \log\left(\frac{P(e_i \mid rel)}{P(e_i \mid irrel)} \frac{P(\bar{e}_i \mid irrel)}{P(\bar{e}_i \mid rel)} \right)$$

$$REL = \begin{cases} rel & \text{(relevance)} \\ irrel & \text{(irrelevance)} \end{cases}$$

$v = (KF_1, KF_2, \cdots, KF_L)$:query and video shot

$q = (kf_1, kf_2, \cdots, kf_L)$ key-point frequency vectors

$\sum_{(q, kf_i > 0)}$:summation over within-shot query key-points

e_i :eliteness (aboutness) of i th key-point

$\bar{e}_i$:non-eliteness (non-aboutness) of i th key-point

- Interpretation of the key-point eliteness
  - If the key-point is elite, the key-point is related to the main object in the video
  - If the key-point is non-elite, the key-point is not related to the main object in the video

5. calculate exponential BM25 shot ranking scores

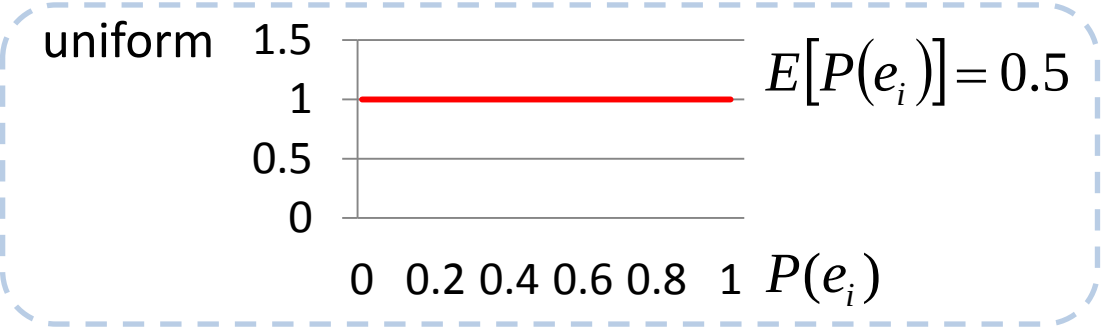
- Probabilistic information retrieval model

$$P(REL = rel \mid v, q) \propto_q \sum_{(q, kf_i > 0)} \frac{kf_i}{kf_i + \kappa} \log\left(\frac{P(e_i \mid rel)}{P(e_i \mid irrel)} \frac{P(\bar{e}_i \mid irrel)}{P(\bar{e}_i \mid rel)} \right)$$

・BM25 with standard IDF

Suppose $P(e_i) \sim uniform$

$\left(= P(e_i) \sim beta(1,1)\right)$



prior distributions that i th key-point becomes elite

then the log value and the ranking function becomes

$$P(REL = rel \mid v, q) \propto_q \sum_{(q, kf_i > 0)} \frac{kf_i'}{kf_i' + \kappa} \log\left(\frac{N - n_i + 1}{n_i + 1} \right)$$

$kf_i' = kf_i / (1 - b + b(vl/avdl))$, $vl = \sum_j KF_j$, $avdl = average\ of\ vl$, $\kappa = 2,\ b = 0.75$

N :total number of shots in DB n_i :total number of shots containing i th key-point

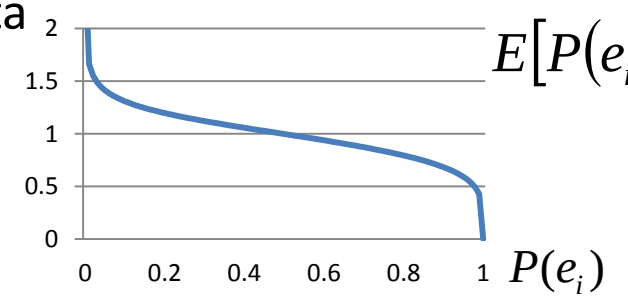
5. calculate exponential BM25 shot ranking scores

- BM25 with <u>exponential IDF</u>

n_i :total number of shots containing i th key-point

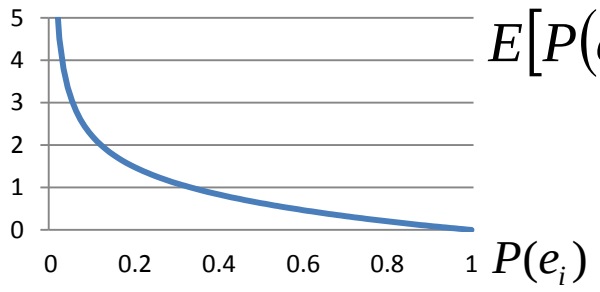
Suppose $P(e_i) \sim beta\left(e^{-n_i/\gamma}, e^{n_i/\gamma} - e^{-n_i/\gamma} + 1\right)$

beta



$E[P(e_i)] = 0.43$

$n_i = 100,\ \gamma = 1000$



$E[P(e_i)] = 0.23$

$n_i = 500,\ \gamma = 1000$

<u>prior distributions that i th key-point becomes elite</u>

then the log value and the ranking function becomes

$$P(REL = rel \mid v, q) \propto_q \sum_{(q, kf_i > 0)} \frac{kf_i'}{kf_i' + \kappa} \log\left(\frac{e^{-n_i/\gamma}}{(e^{n_i/\gamma} - e^{-n_i/\gamma} + 1)} \frac{\left(N - n_i + e^{n_i/\gamma} - e^{-n_i/\gamma} + 1\right)}{(n_i + e^{-n_i/\gamma})} \right)$$

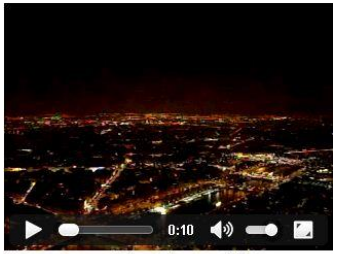
※contributions from key-points that frequently appear in DB are suppressed by this design (IDF retains the same feature, but it is not sufficient for image/video retrieval)

Instance images and the region-of-interest images of “Eiffel tower”

- Search results using BM25

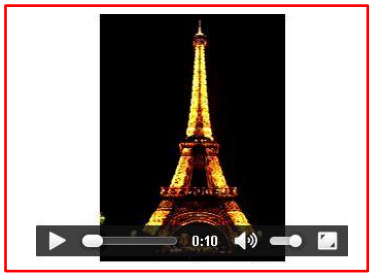
rank 1.

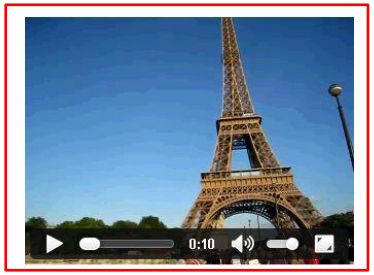
rank 2.

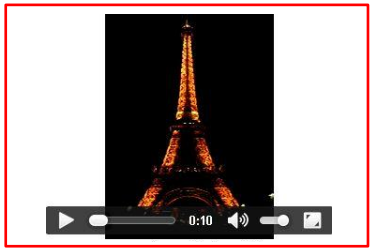


rank 3.

rank 4.

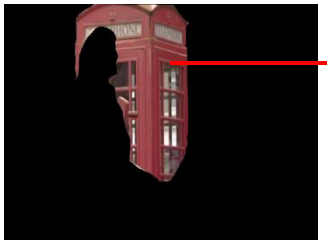






- Search results using exponential BM25

6. employ region-of-interest effects



masked query images

$$\sum_{(q,kf_i>0)} ROI_i \frac{kf_i'}{kf_i'+\kappa} \log\left(\frac{e^{-n_i/\gamma}}{(e^{n_i/\gamma}-e^{-n_i/\gamma}+1)} \frac{\left(N-n_i+e^{n_i/\gamma}-e^{-n_i/\gamma}+1\right)}{(n_i+e^{-n_i/\gamma})}\right),$$

region of Interest

within-shot key-point frequency

Key-point exponential IDF weight

$$where\ ROI_i = \begin{cases} \lambda\ (if\ \ keypoint\ i\ is\ within\ \ ROI) \\ 1\ (otherwise) \end{cases}$$

- Combine SIFT and CSIFT (late fusion)

$$eBM\,25(v_{SIFT}, q_{SIFT}) =$$

$$\sum_{(q_{SIFT},kf_{i,SIFT}>0)} ROI_{i,SIFT} \frac{kf_{i,SIFT}'}{kf_{i,SIFT}'+\kappa} \log\left(\frac{e^{-n_{i,SIFT}/\gamma}}{(e^{n_{i,SIFT}/\gamma}-e^{-n_{i,SIFT}/\gamma}+1)} \frac{\left(N-n_{i,SIFT}+e^{n_{i,SIFT}/\gamma}-e^{-n_{i,SIFT}/\gamma}+1\right)}{(n_{i,SIFT}+e^{-n_{i,SIFT}/\gamma})}\right)$$

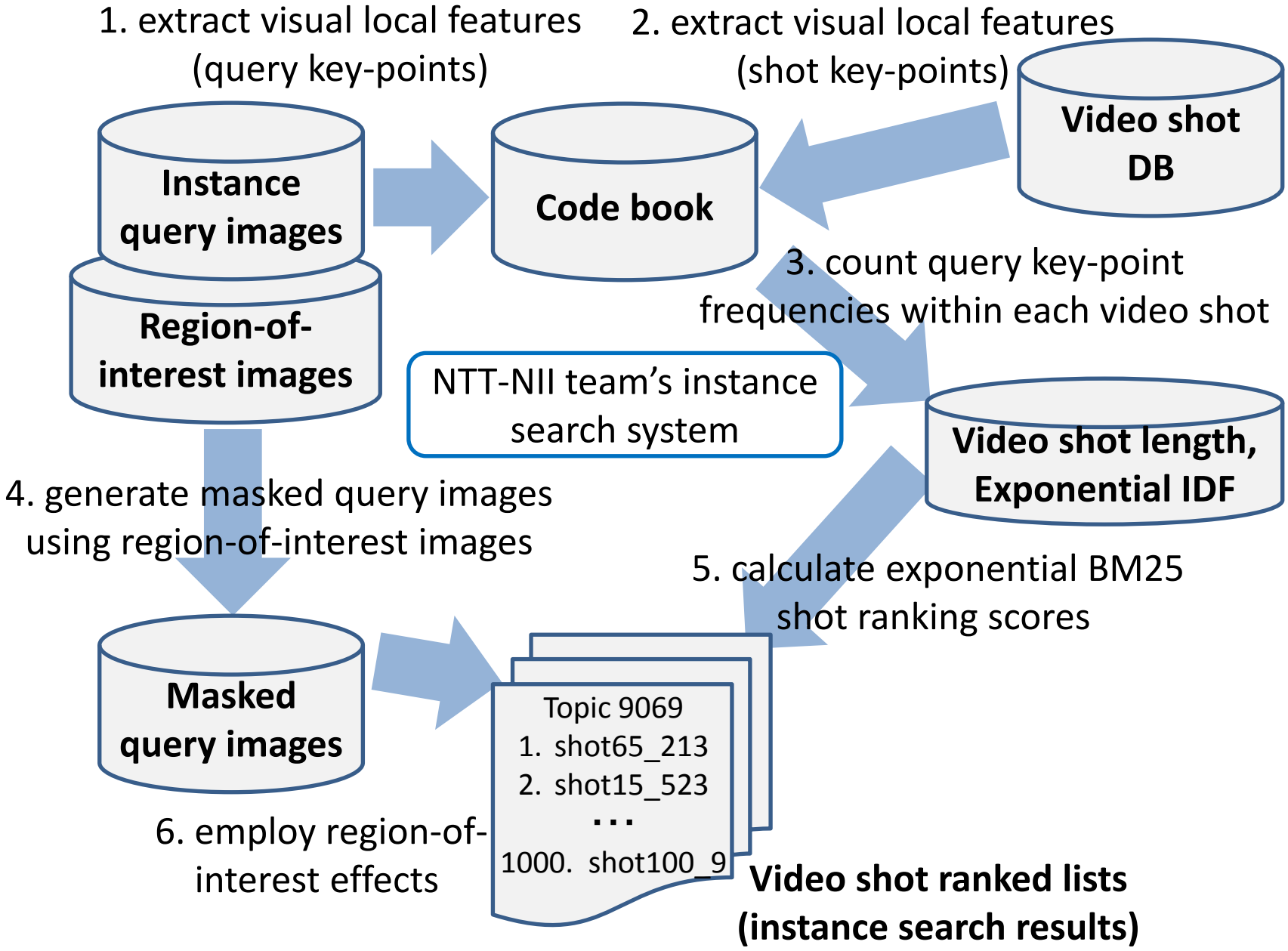
$$eBM\,25(v_{CSIFT}, q_{CSIFT}) =$$

$$\sum_{(q_{CSIFT},kf_{i,CSIFT}>0)} ROI_{i,CSIFT} \frac{kf_{i,CSIFT}'}{kf_{i,CSIFT}'+\kappa} \log\left(\frac{e^{-n_{i,CSIFT}/\gamma}}{(e^{n_{i,CSIFT}/\gamma}-e^{-n_{i,CSIFT}/\gamma}+1)} \frac{\left(N-n_{i,CSIFT}+e^{n_{i,CSIFT}/\gamma}-e^{-n_{i,CSIFT}/\gamma}+1\right)}{(n_{i,CSIFT}+e^{-n_{i,CSIFT}/\gamma})}\right)$$

$$score(v,q) = eBM\,25(v_{SIFT}, q_{SIFT}) + eBM\,25(v_{CSIFT}, q_{CSIFT})$$

(video shots in DB are ranked by the decreasing order of $score(v,q)$)

Our instance search system



Evaluation results (submitted runs)

$$P(REL = rel \mid v, q) \propto_q \sum_{(q, kf_i > 0)} ROI_i \frac{kf_i'}{kf_i' + 2} \log\left(\frac{e^{-n_i/\gamma}}{(e^{n_i/\gamma} - e^{-n_i/\gamma} + 1)} \frac{\left(N - n_i + e^{n_i/\gamma} - e^{-n_i/\gamma} + 1\right)}{(n_i + e^{-n_i/\gamma})} \right)$$

Region of Interest

Within-shot key-point frequency

Key-point exponential IDF weight

NTT_NII_1: EBM25(SIFT+CSIFT) $\gamma = 100,\ \lambda = 2$

NTT_NII_2: EBM25(SIFT+CSIFT) $\gamma = 100,\ \lambda = 10$

NTT_NII_3: EBM25(SIFT+CSIFT) $\gamma = 1000,\ \lambda = 10$

Multimodal (linear combination of multimodal BM25 scores)

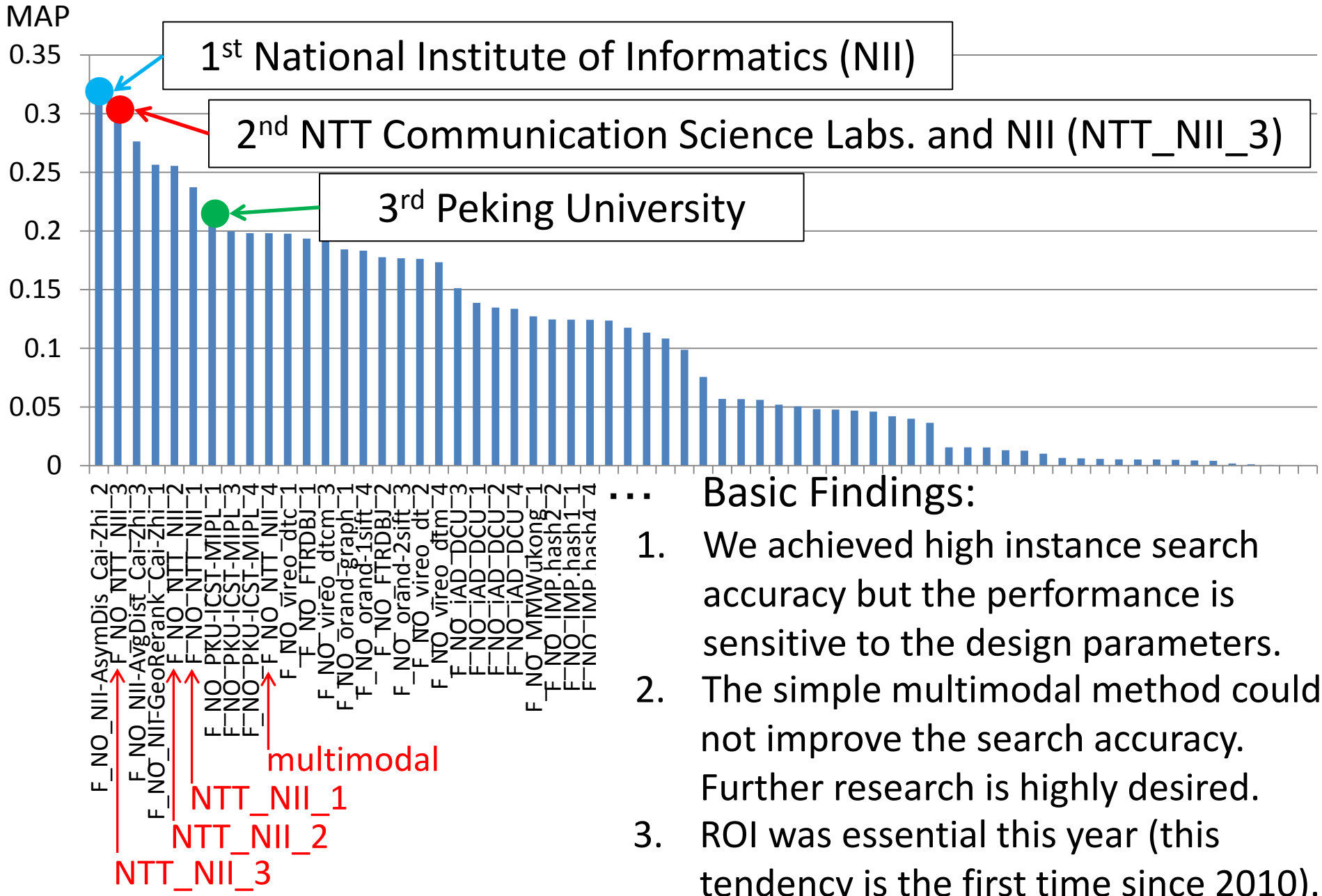
NTT_NII_4: BM25(SIFT+CSIFT)+BM25(audio)+BM25(text)

Multimodal run

1. Execute instance search using SIFT and CSIFT (using visual information only)
   – based on BM25 with standard IDF
2. Assume the top 10 ranked shots as relevant to the instance and use their audio and textual information to re-rank the original search results
   – shots include audio and textual (actors'/actresses' line) information
   – multimodal BM25 based on bag-of-audio-words (used MFCC) and on bag-of-textual-words
   – linear combination of the multimodal BM25 scores

NTT_NII_4: BM25(SIFT+CSIFT)+BM25(audio)+BM25(text)

Overall results



…

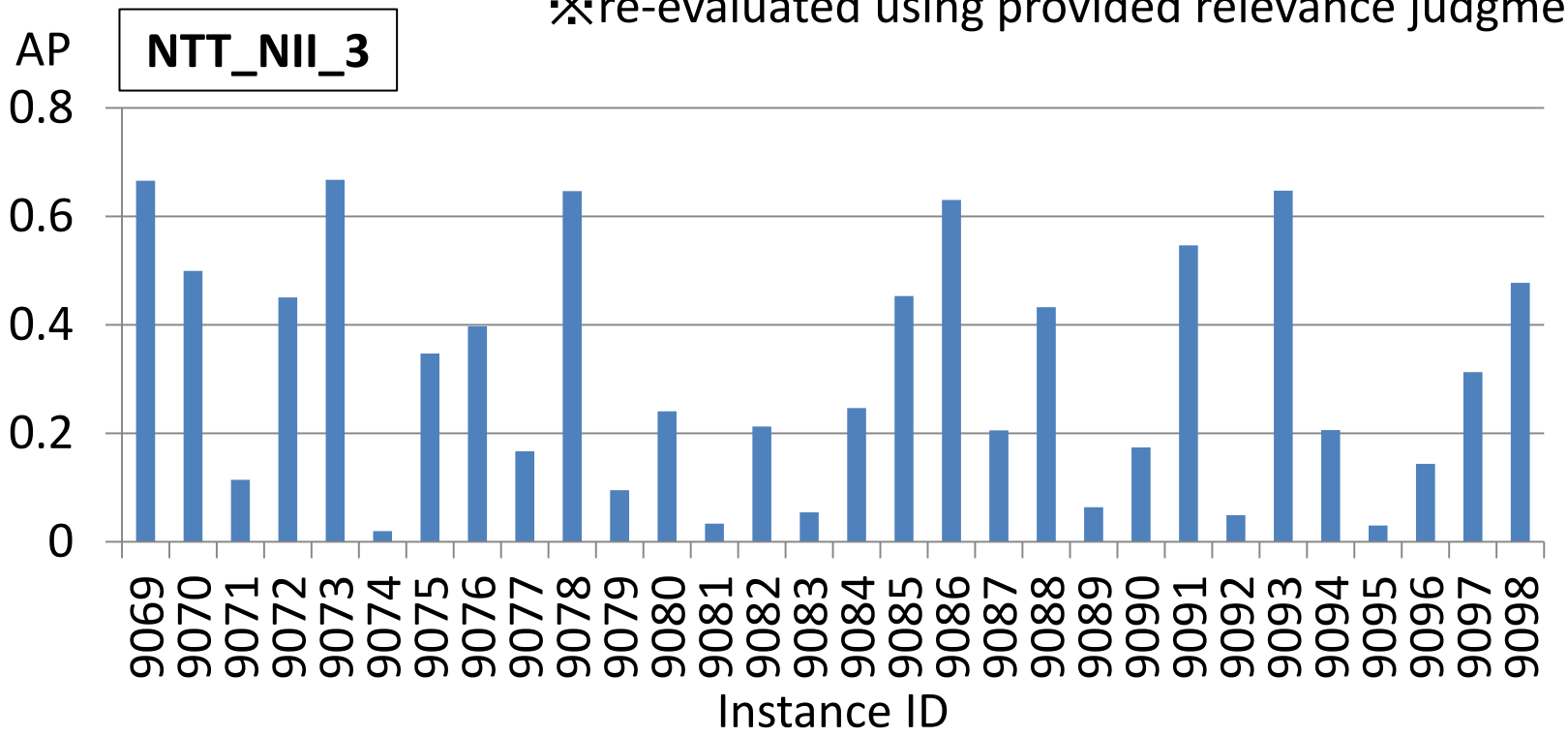
Basic Findings:

1. We achieved high instance search accuracy but the performance is sensitive to the design parameters.
2. The simple multimodal method could not improve the search accuracy. Further research is highly desired.
3. ROI was essential this year (this tendency is the first time since 2010).

Detailed comparison

	MAP	P@10	P@20
NTT_NII_1 $(\gamma = 100,\ \lambda = 2)$	0.25($p < 0.01$)	0.65	0.57($p < 0.05$)
NTT_NII_2 $(\gamma = 100,\ \lambda = 10)$	0.27	0.71	0.65
NTT_NII_3 $(\gamma = 1000,\ \lambda = 10)$	**0.31**	**0.7**	**0.65**
NTT_NII_4 (multimodal)	0.21($p < 0.01$)	0.56	0.51($p < 0.01$)

※re-evaluated using provided relevance judgment data



Instance search result examples

Concluding remarks and our suggestions for the next year's INS

- eBM25 (exponential BM25) was effective.
  - But needs the careful parameter setting.
- Region-of-interest images worked well.
  - To the best of our knowledge, since 2010, this is the first time that ROI was essential for improving search accuracy.
- Suggestions for the evaluation methodology
  - We have to say that this year's evaluation was too coarse.
    - Too many video shots that remain unjudged
    - Prevent us from further research using this year's dataset because the relevance judgment data is considerably incomplete.
  - **Can the member of participating teams volunteer to make better relevance judgment data?**
    - **Re-evaluate the top N (e.g. N=100) search results of all submission runs.**

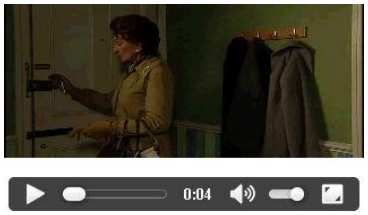
Instance images and the region-of-interest images of "this dog"

4.

5.

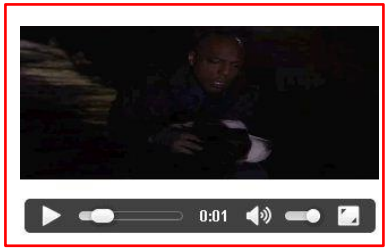
6.

7.



8.

・Search results using exponential BM25 (1-3 ranked videos are correct)



・Search results using exponential BM25 **<u>+ region-of-interest</u>** (1-3 ranked videos are correct)

This year's relevance judgment data overview