Telefónica
Telefónica Investigación y Desarrollo

# Telefonica Research @ Trecvid 2011

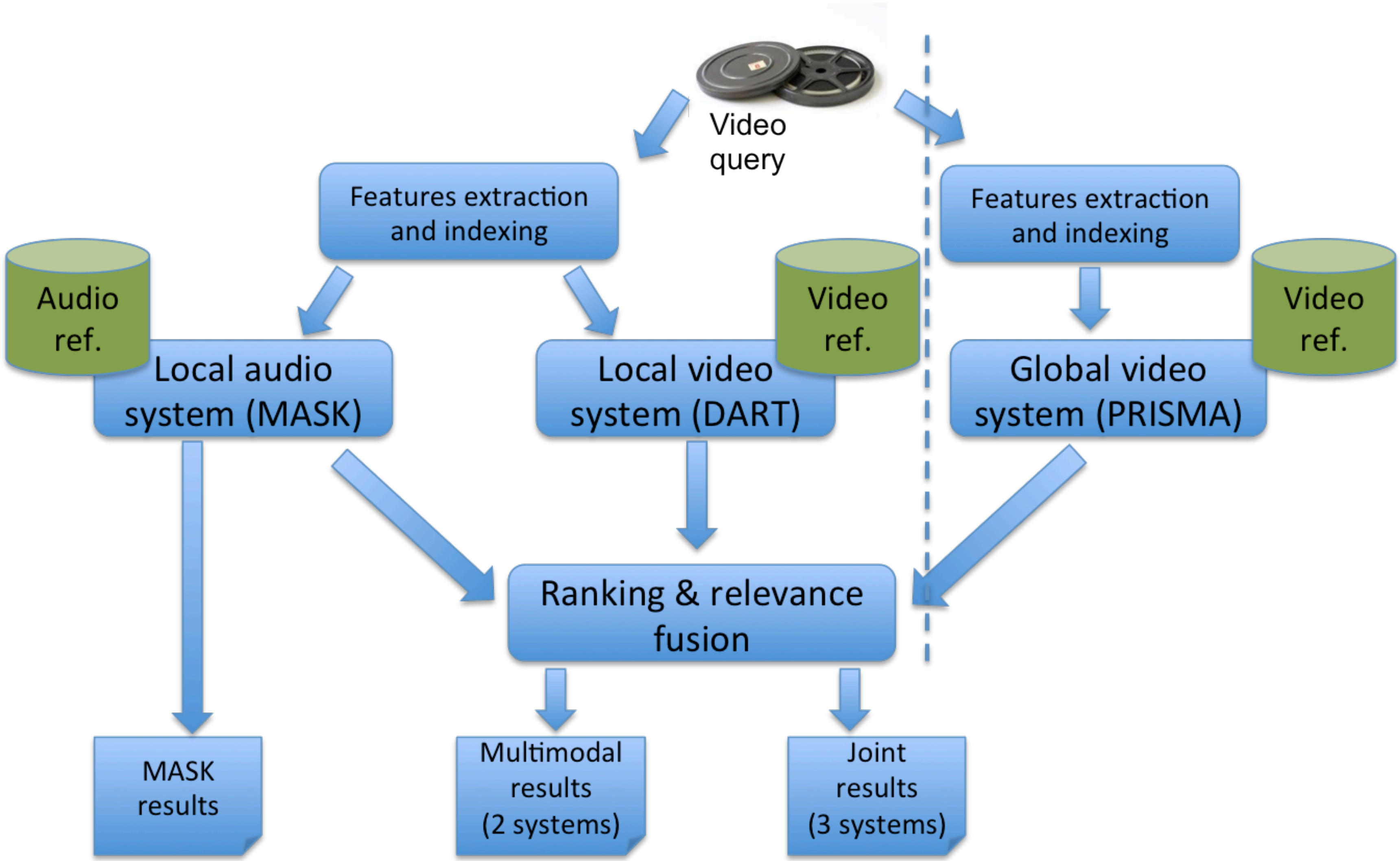
Xavier Anguera, Daru Xu[1] and Tomasz Adamek
(With the collaboration of Juan Manuel Barrios, Prisma Group)

[1]Daru Xu is a graduate student at the Ming-Hsieh Department of Electrical Engineering, University of Southern California, USA

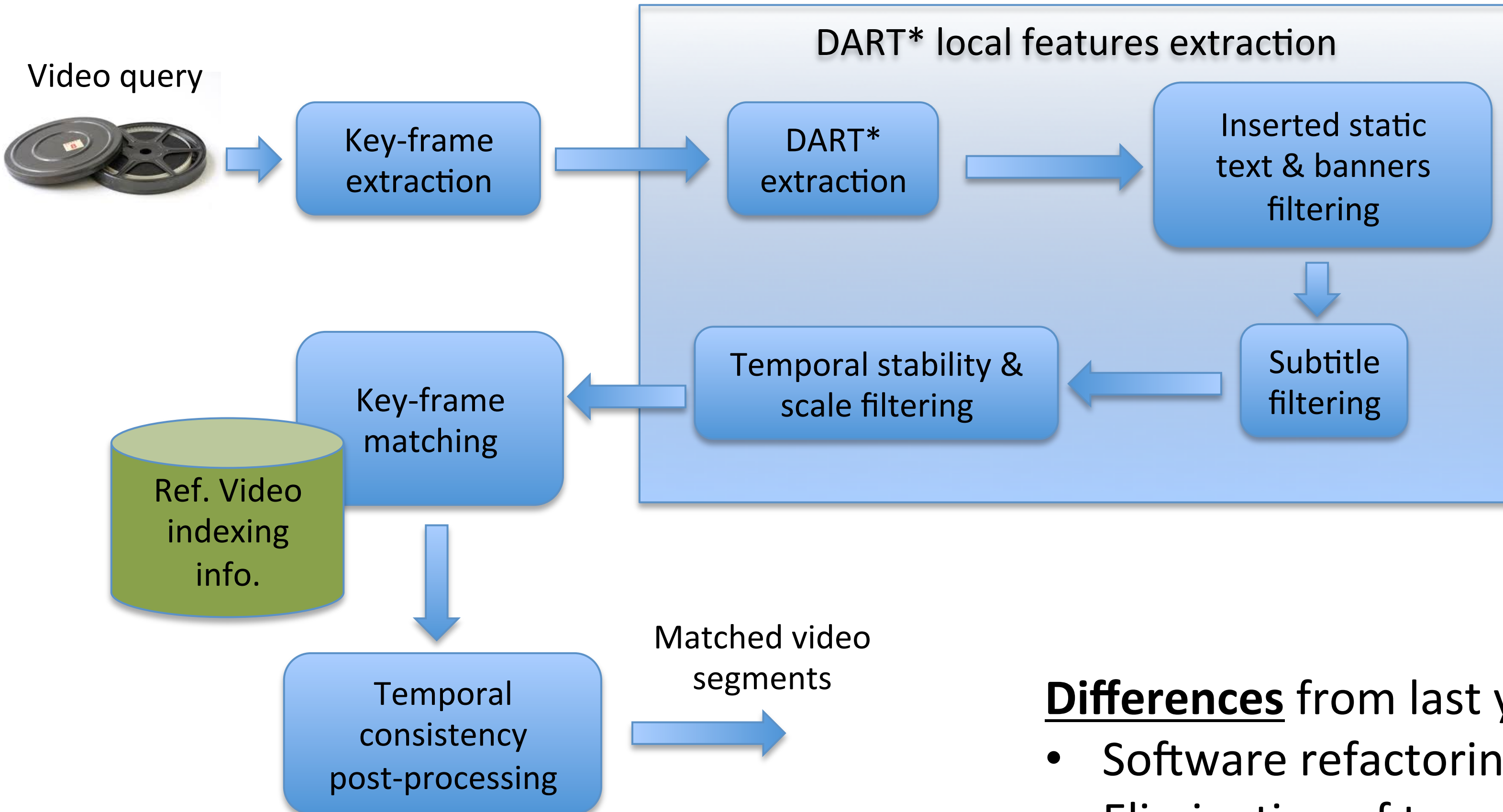
# Outline of the talk

- Telefonica 2011 Video-copy detection system
  - Overall system
  - Video-copy detection
  - Audio-copy detection
  - Fusion algorithm
  - Results
- Multi-systems fusion experiment

# Multimodal Video-copy detection



# Video-based System

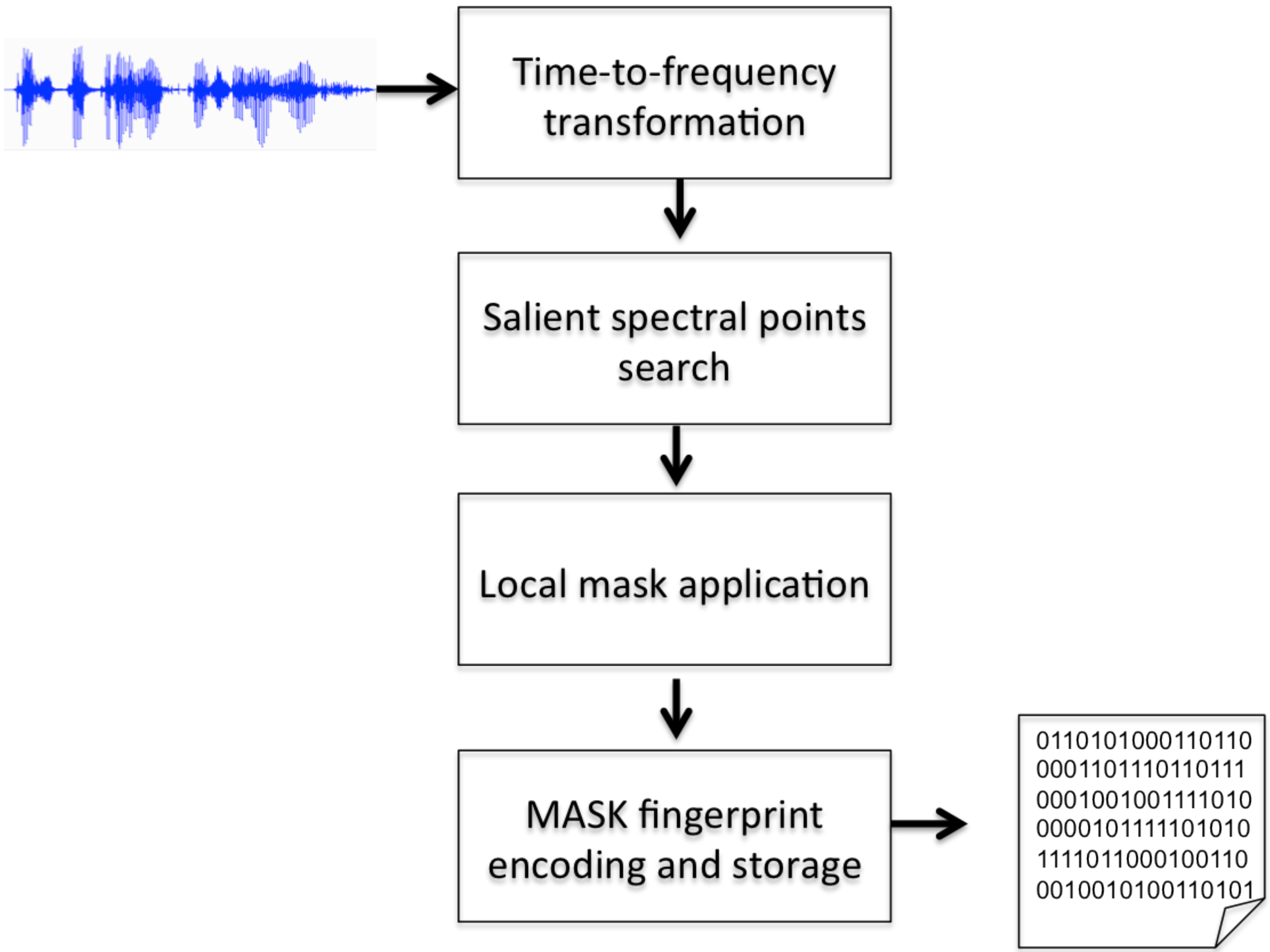



**<u>Differences</u>** from last year:

- Software refactoring
- Elimination of temporary files

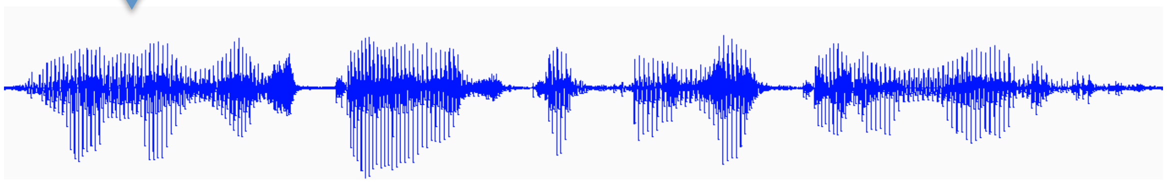
* D. Marimon, A. Bonnin, T. Adamek, and R .Gimeno, “DARTs:Efficient scale-space extraction of daisy key-points”, CVPR 2009.

# Audio-based System



# MASK fingerprint extraction (I)

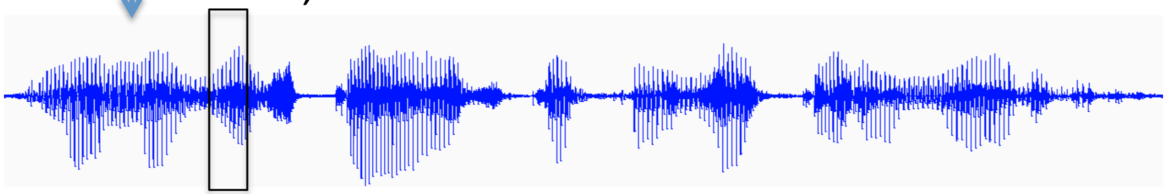
1) Audio track extraction using FFMPEG



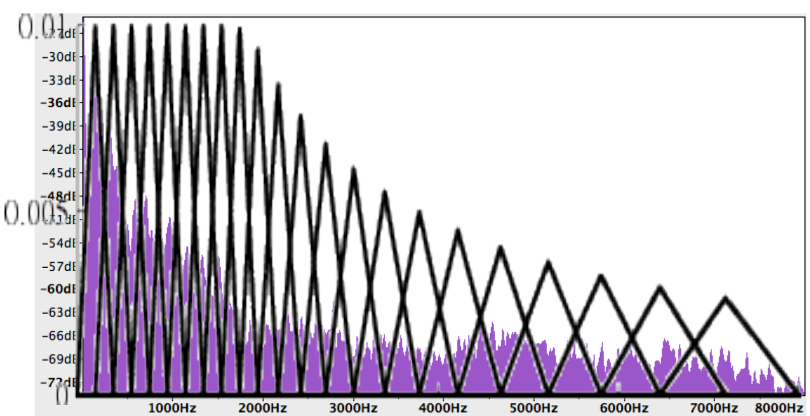
# Acoustic fingerprint extraction (I)

1) Audio track extraction using FFMPEG

10ms, 100ms window



2) FFT, bandwidth limited to 300-3KHz

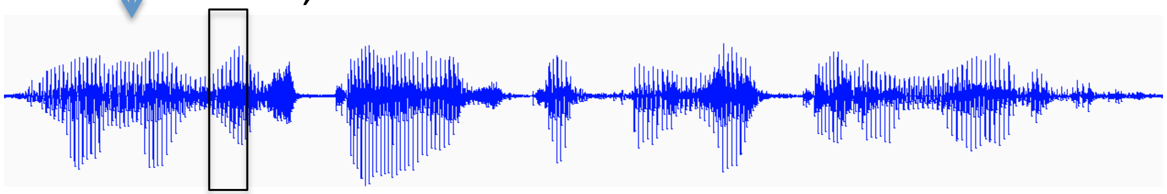



32 MEL-spectrum bands

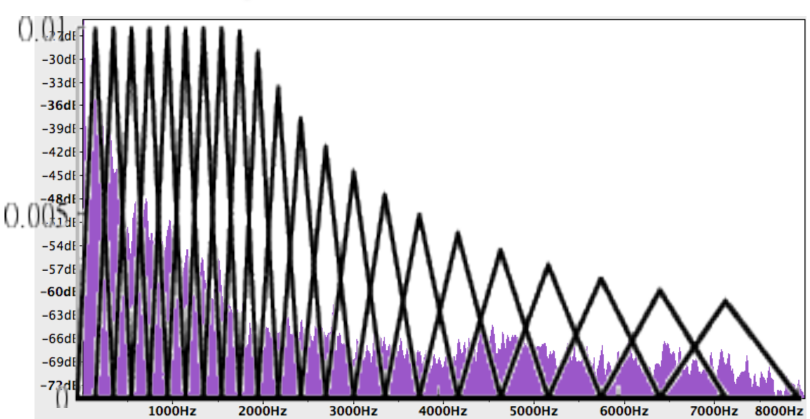
# Acoustic fingerprint extraction (I)

1) Audio track extraction using FFMPEG

10ms, 100ms window

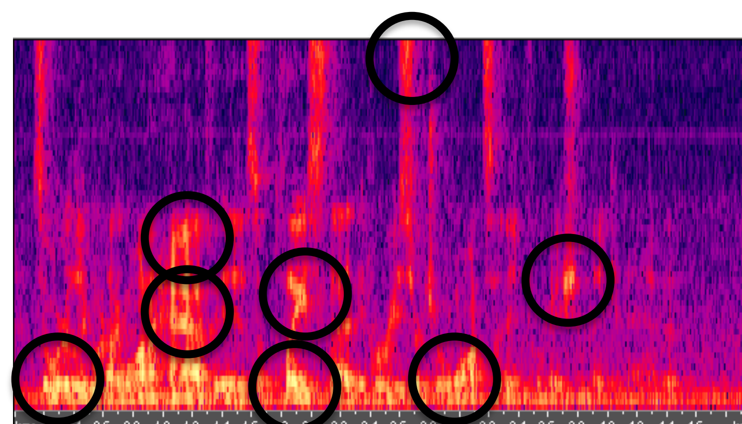


2) FFT, bandwidth limited to 300-3KHz



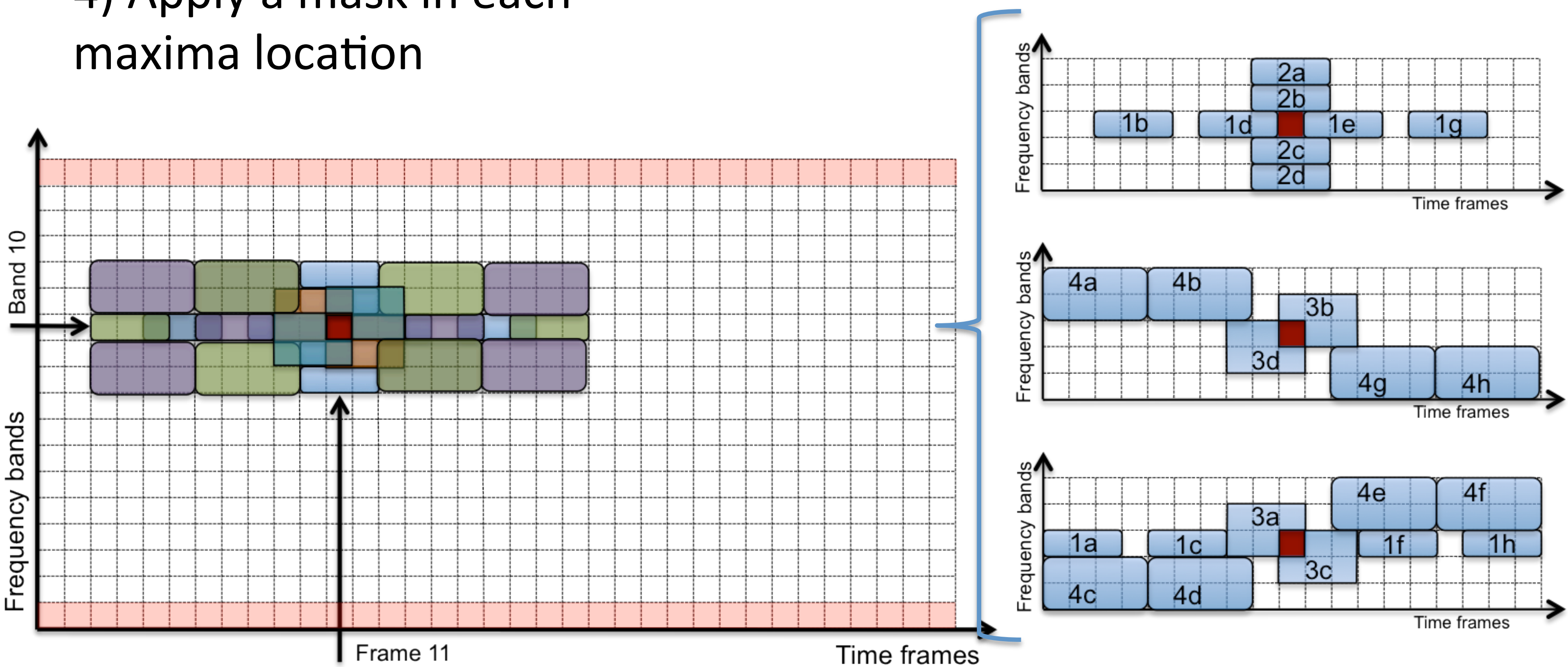

32 MEL-spectrum bands

3) Find spectrogram peaks.



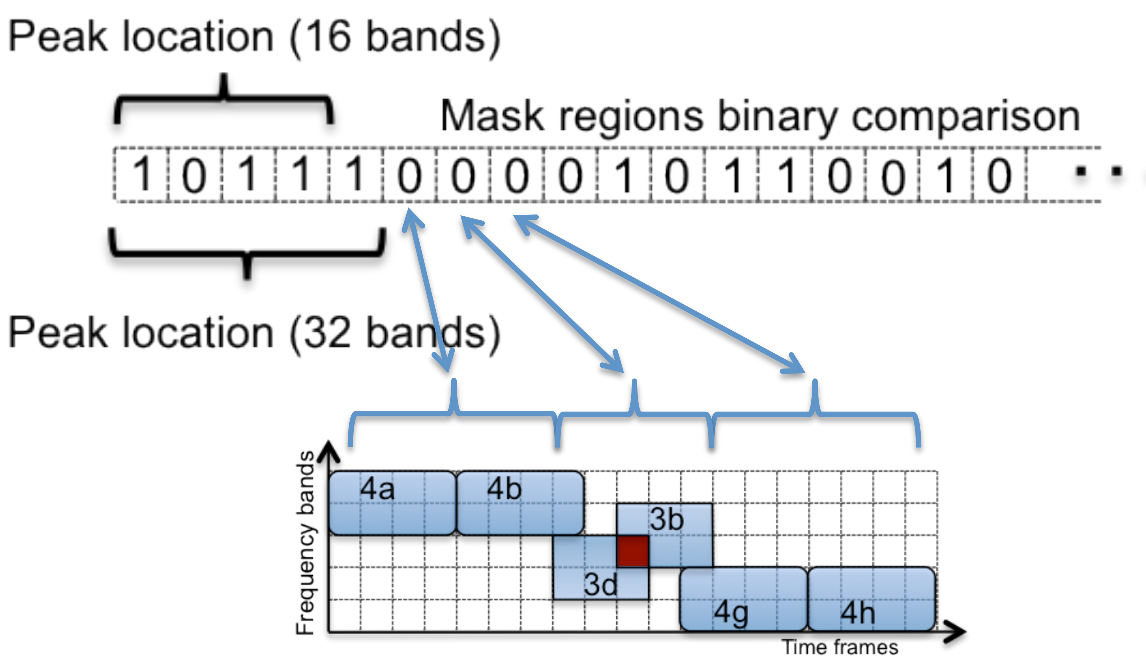
# Acoustic fingerprint extraction (II)

4) Apply a mask in each maxima location



# Acoustic fingerprint extraction (II)

5) Construct the fingerprint



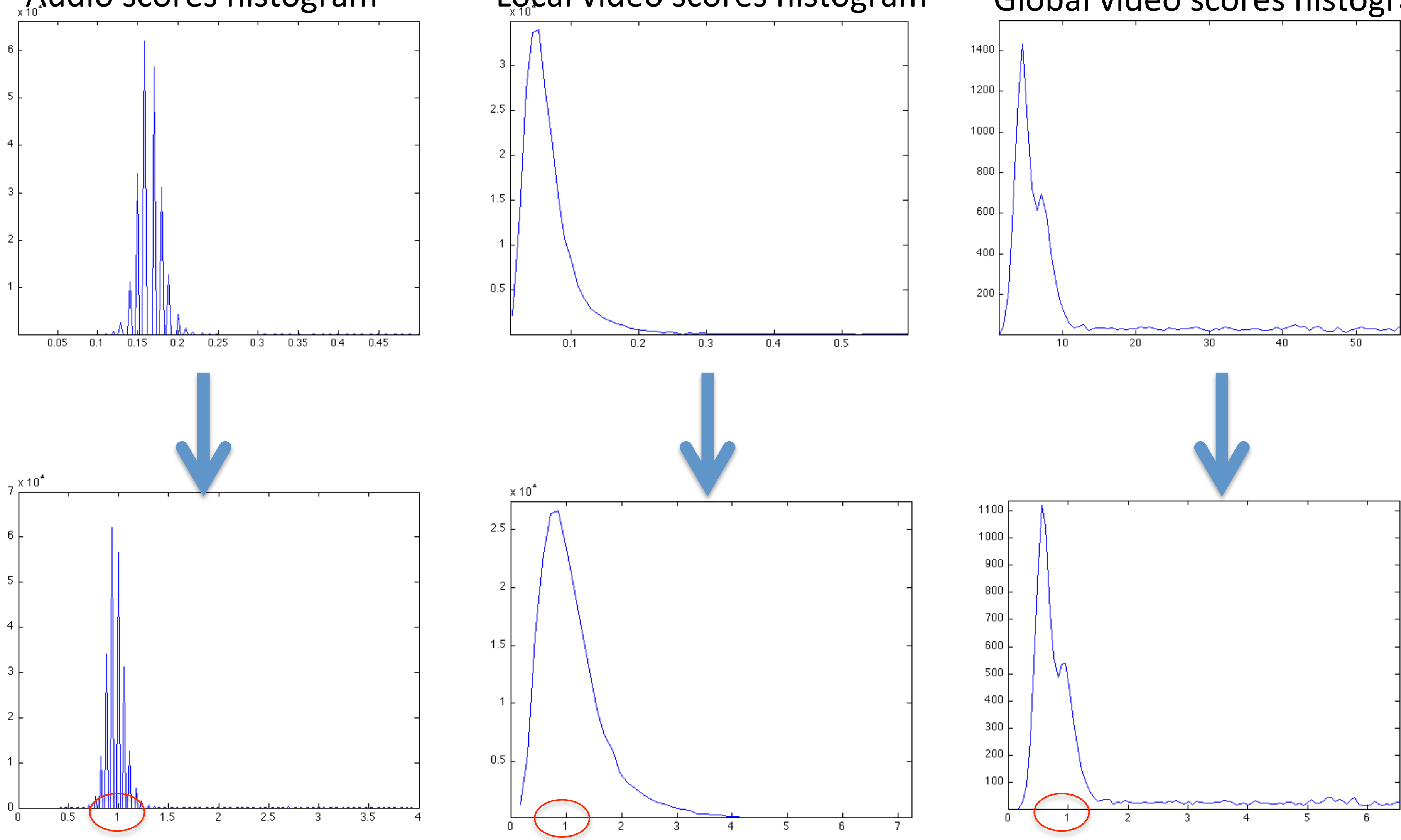
# Multimodal Fusion Algorithm

- Fusion of different modalities at decision level
  - Agnostic of internal system's behaviors
- No limit on the number of systems to be combined
  - provided each system is better than random
- To work optimally it needs N-best matches from each system. It returns the best fused matches (N=20)
  - Makes use of the individual scores and the rank within each modality.

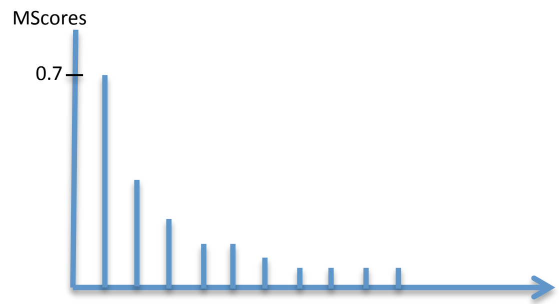
Paper on ACM MM 2011: "Multimodal Fusion for Video Copy Detection", Xavier Anguera, Juan Manuel Barrios, Tomasz Adamek and Nuria Oliver

# Data preprocessing

Audio scores histogram | Local video scores histogram | Global video scores histogram



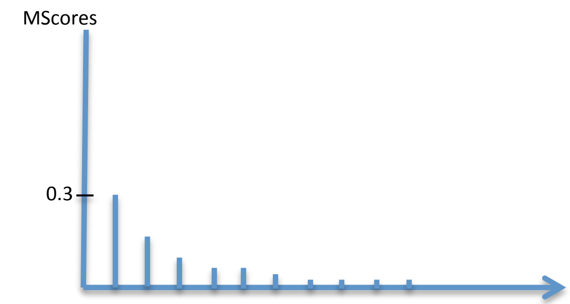
# N-best flooring and L1 Normalization (I)





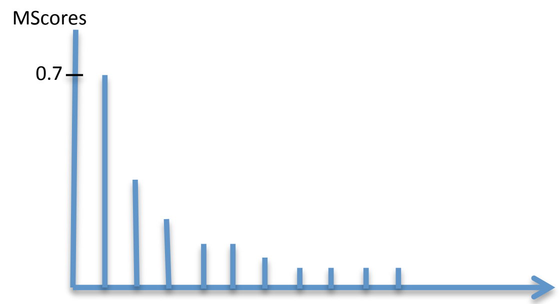
L1 normalization

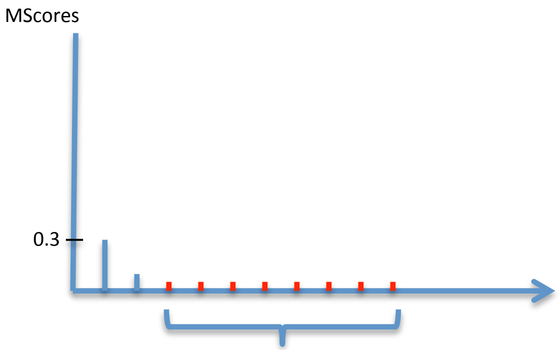
$$\overline{MScore_i} = \frac{MScore_i}{\sum_{j=1}^{Nbest} MScore_j}$$





# N-best flooring and L1 Normalization (II)

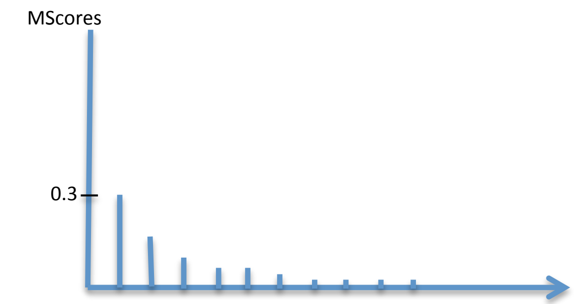


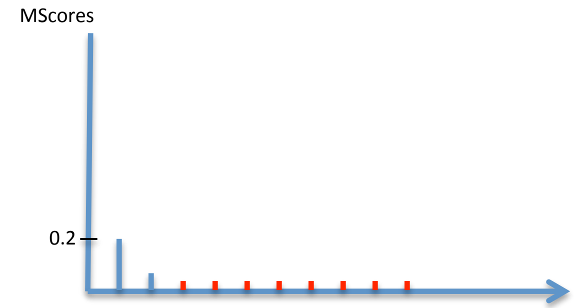




N-best Flooring

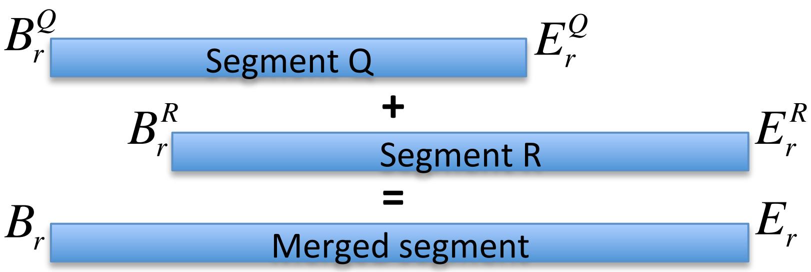
N-best flooring and
L1 normalization

$$\overline{MScore_i} = \frac{MScore_i}{\sum_{j=1}^{Nbest} MScore_j}$$



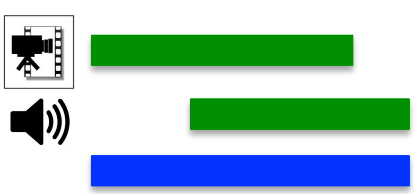




# Overlapping Segments Merge

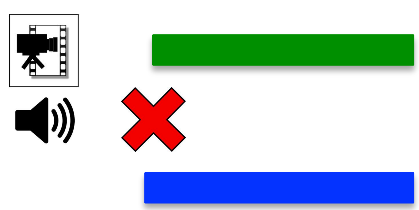



$$\frac{\min\{E_k^Q(r), E_k^R(r)\} - \max\{B_k^Q(r), B_k^R(r)\}}{\max\{E_k^Q(r), E_k^R(r)\} - \min\{B_k^Q(r), B_k^R(r)\}} > 0.5$$

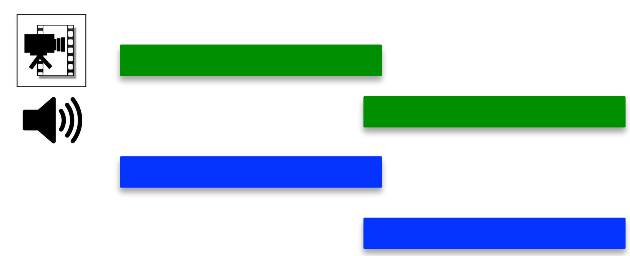
Examples:



Mutimodal overlap



Missing modality



Non-overlapping modalities

# Output score computation

Number of matches

Rank [1 to $N_k$]

Resulting score for fused match

$$S(c_l) = \frac{\sum_{c_k(r) \epsilon c_l} W_k \cdot \frac{N_k - r + 1}{N_k} \cdot \hat{S}_k(r)}{\sum_{k=1}^{K} (W_k \cdot \hat{S}_k(1))}$$

Normalized matching score at rank r

A-priori weight for each modality

Best normalized matching score for each modality

# Official evaluation results

Optimum scores, balanced profile:

| | Profile | Min NDCR | FA count | Miss count | True positives | Opt F1 score |
|---|---|---|---|---|---|---|
| Audio system | BALANCED | 0.662 | 0.66 | 54.75 | 54.78 | 0.729 |
| Multimodal | BALANCED | 0.610 | 0.80 | 11.73 | 63.69 | 0.947 |
| Joint | BALANCED | 0.268 | 0.23 | 4.71 | 101.4 | 0.957 |

Choosing only 1$^{st}$-best results:

| | Profile | Min NDCR | FA count | Miss count | True positives | Opt F1 score |
|---|---|---|---|---|---|---|
| Audio system | BALANCED | 0.477 | 0.14 | 55.89 | 72.05 | 0.712 |

# Multi-systems fusion experiment

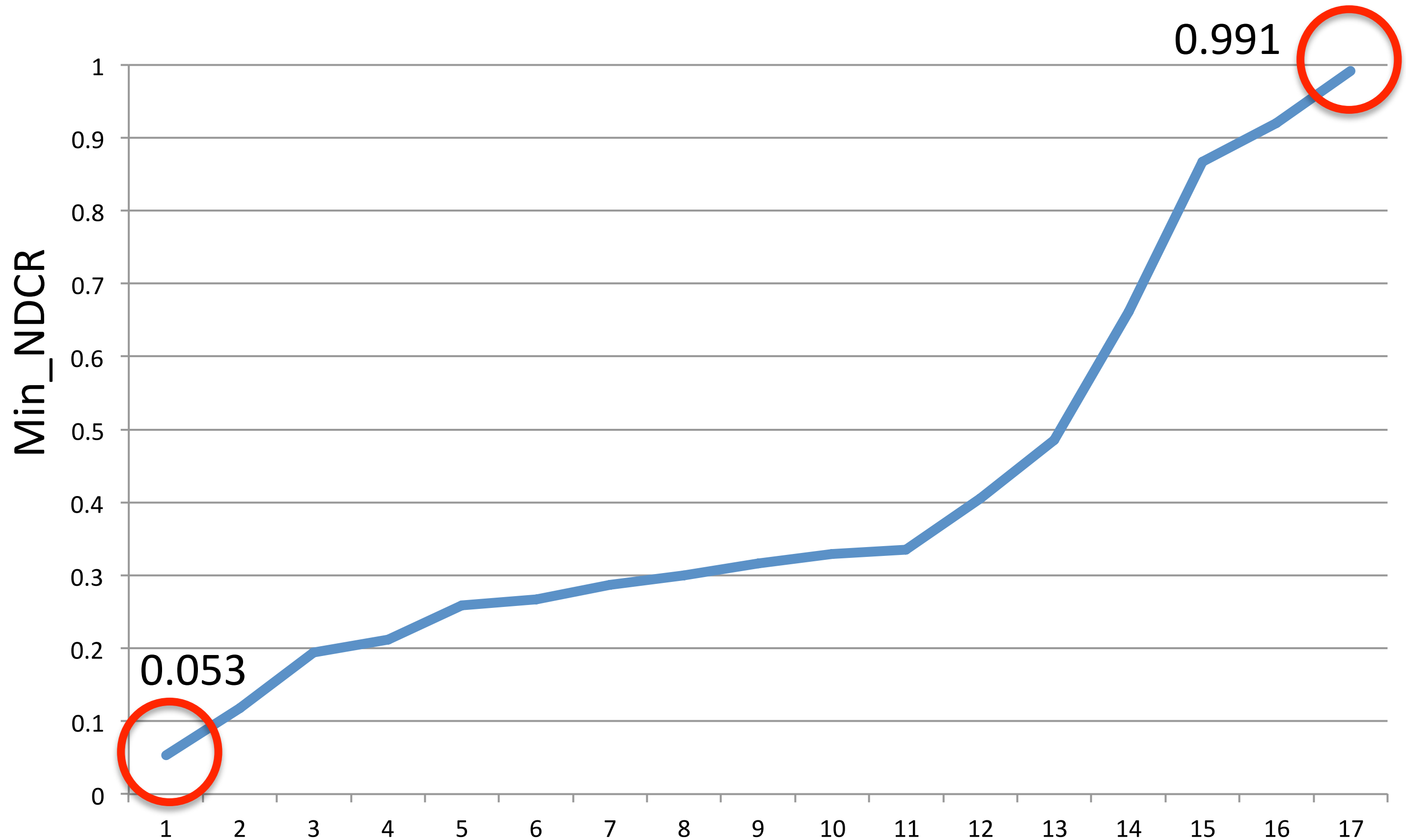
- We tested the fusion algorithm with many system outputs

- We asked participants in TRECVID 2011 for their submitted runs
  - 10 teams contributed their results: PKU-IDM, CRIM, INRIA-TEXMEX/LEAR, FT, prisma, ATTLabs, kddi, iupr-dfki, brno, Telefonica Research
  - I used the “Balanced” runs: 17 runs

# Status of the runs

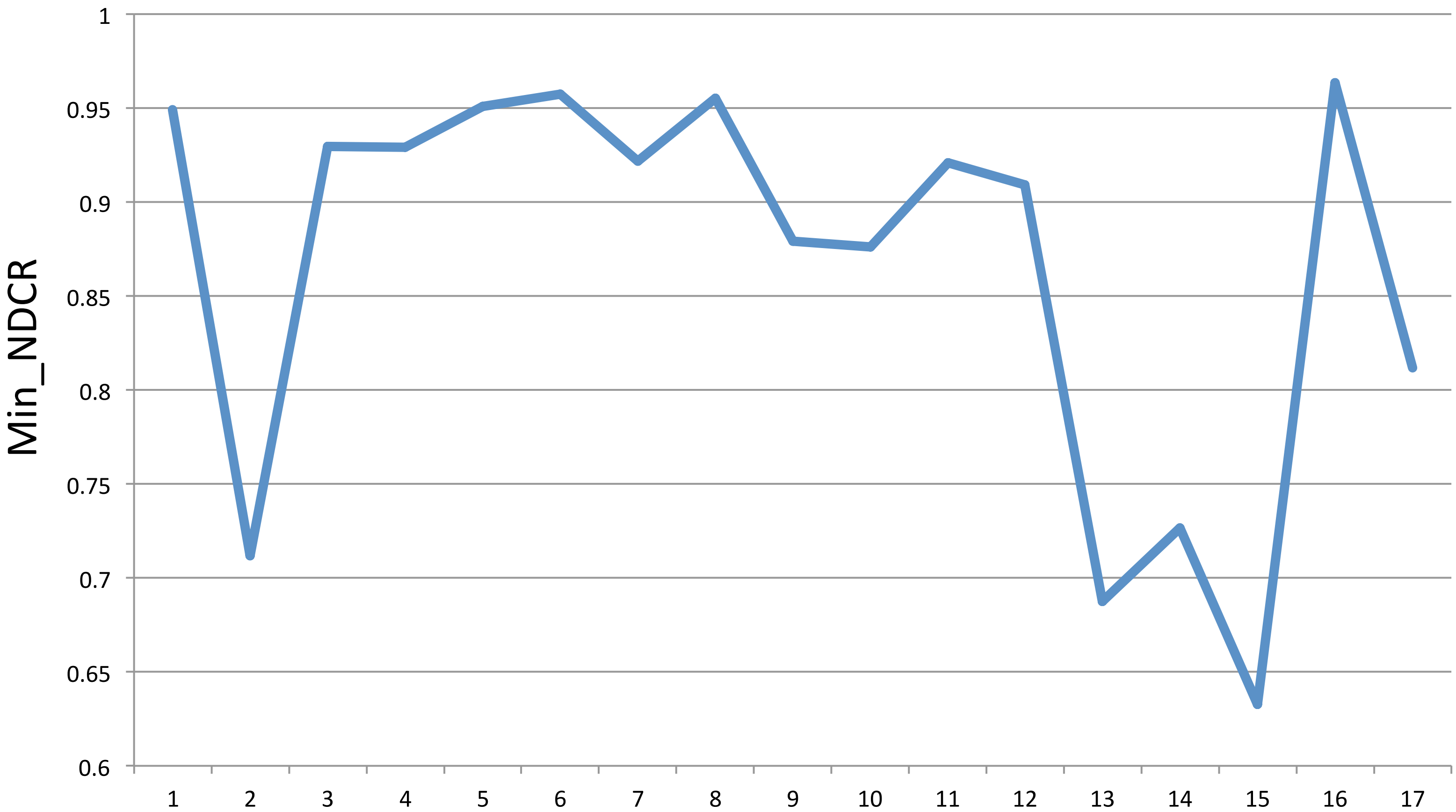
- The fusion algorithm works optimally when Nbest results are available for each fused output.
  - Results for the used systems had (many times) only 1best results, resulting suboptimal for the fusion.

# Individual results (Min NDCR)

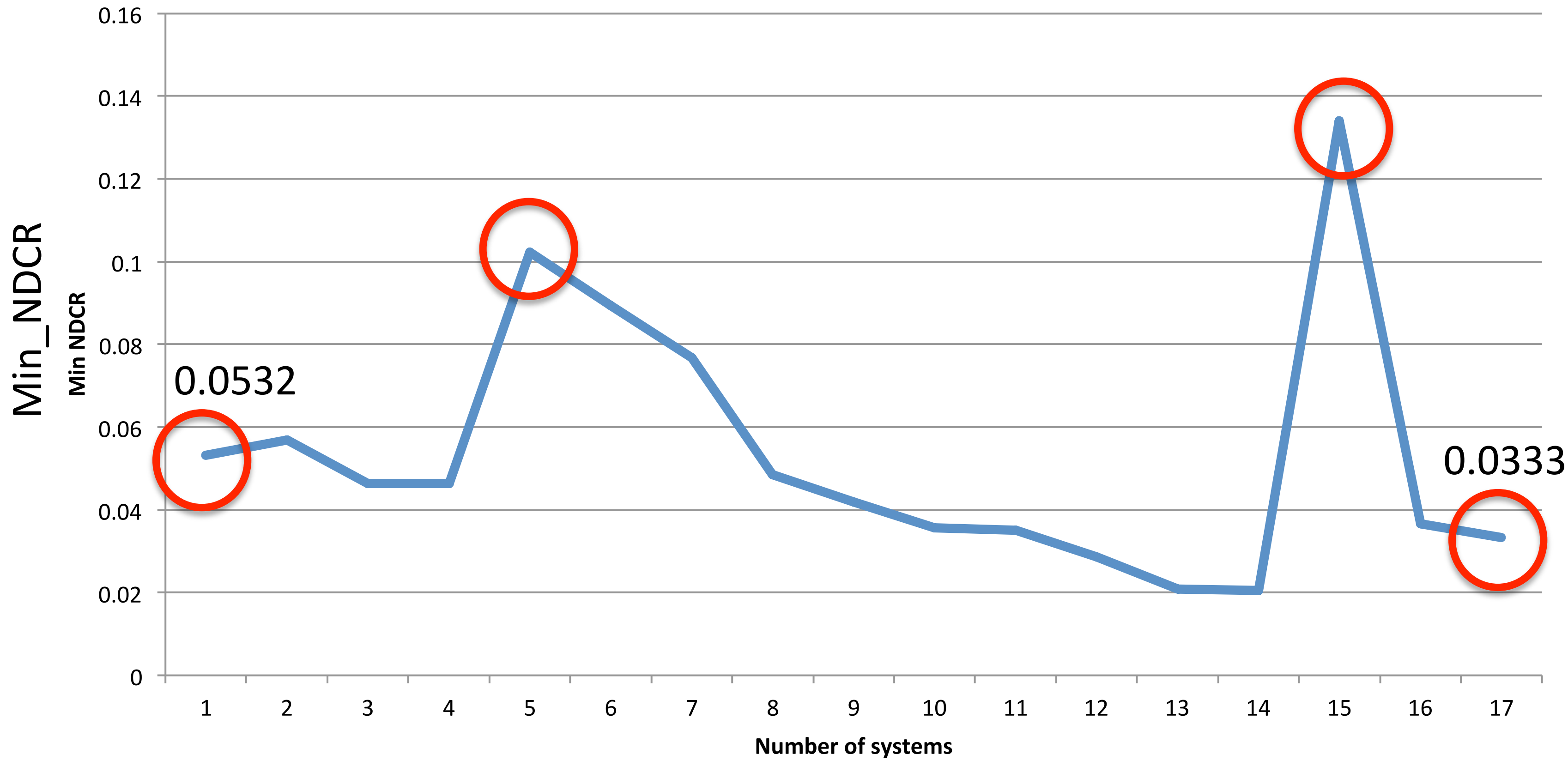



- Labeled from 1 to 17, to anonymize them.

# Individual results (optimum F1)

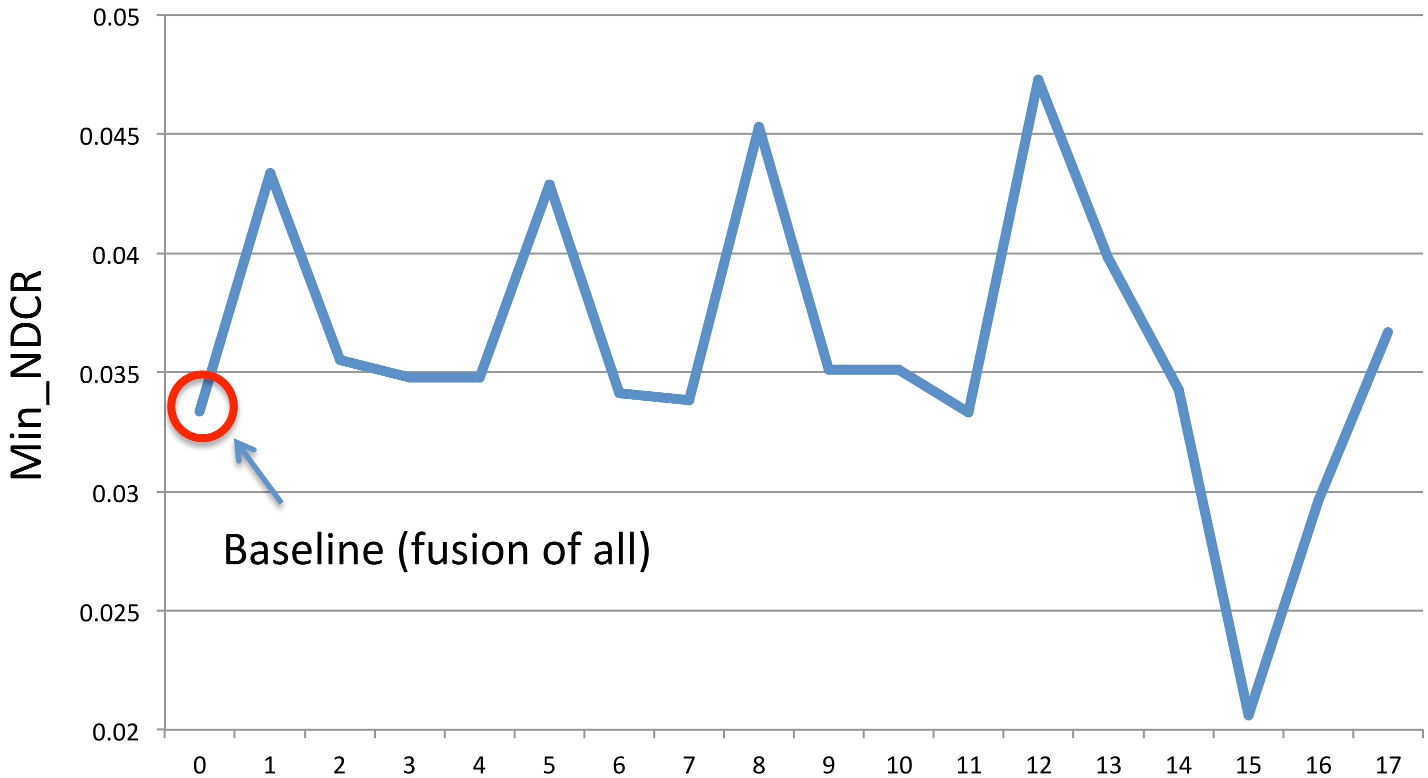


# Incremental fusion



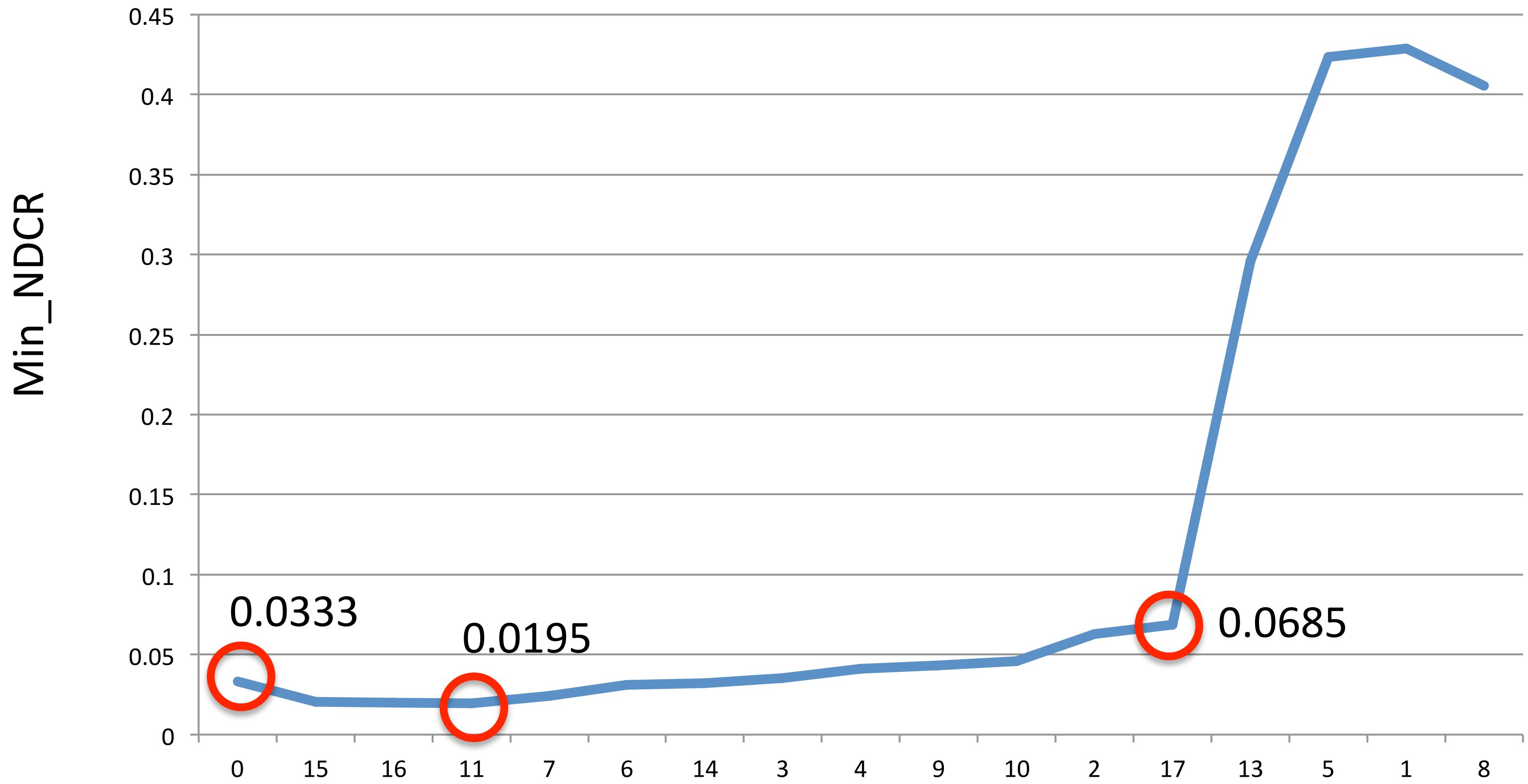

- We incrementally added systems and computed the fusion
- Systems 5 and 15 are the only ones making the fusion worse
- Final Min_NDCR=0.0333

# Fusion of all minus 1




We obtain an order from worse to best in the fusion (worse in here is system 15)

# Incremental elimination




- With only 5 systems we achieve pretty decent results
- The best result is 0.0195, although this is “cheating”

# Conclusions

- The fusion algorithm can extract knowledge and make results better
  - Even if fusing systems which have weaker NDCR results, the fusion results in good scores.
- FUTURE WORK: automatically identify which modalities bring novelty.