Telefonica Research
Multimodal Video copy detection

Xavier Anguera, Tomasz Adamek and
Ehsan Younessian*

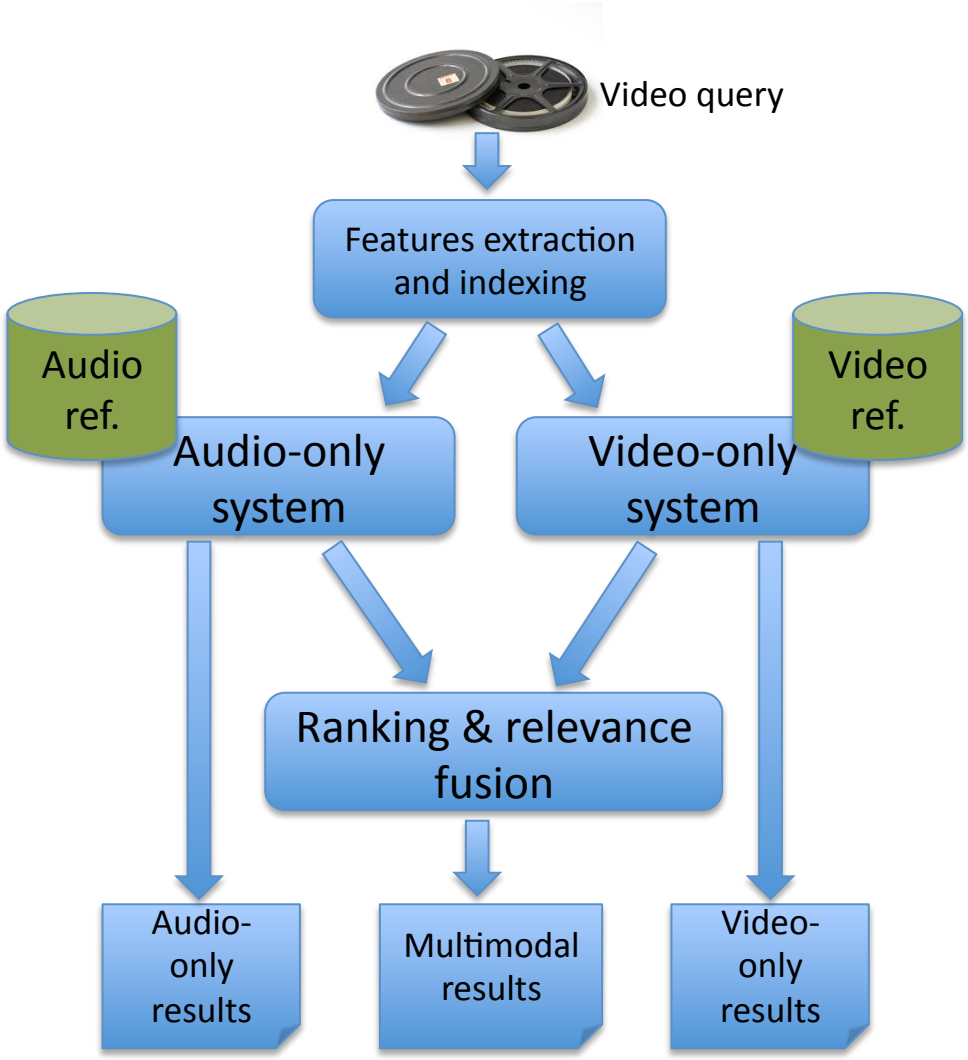
*School of Computer Engineering, Nanyang Technology Univ., Singapore, Singapore

Who we are?

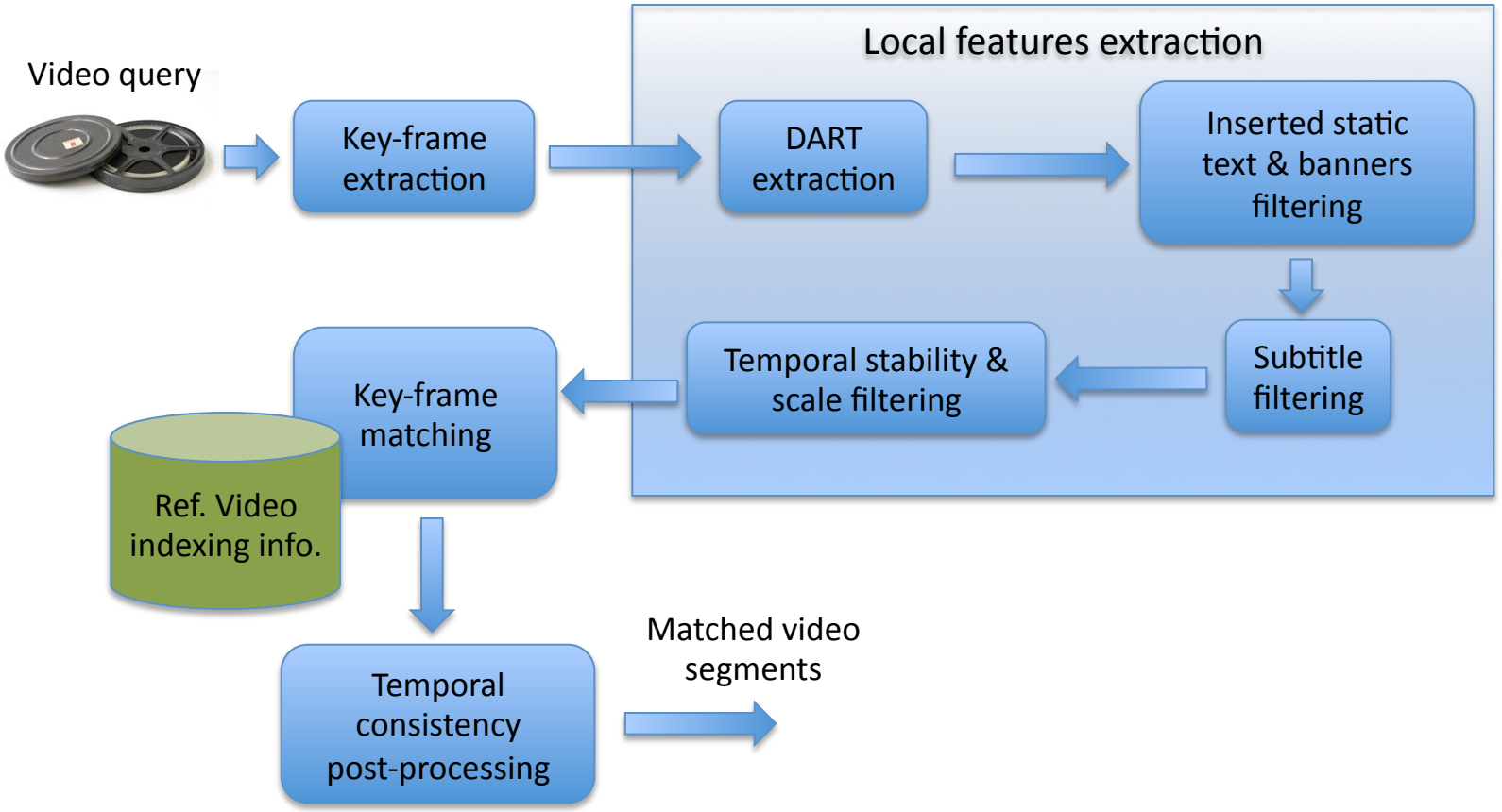
- Telefónica Research is the innovation company of the Telefónica Group
- Telefónica Research is the largest private R&D centre in Spain

- Telefónica is one of the world's largest telecommunications companies by market cap
  - operates in 25 countries
  - customer base 277.8 million

Multimodal Video Copy detection



Video-based block diagram



DART* local features (advantages)

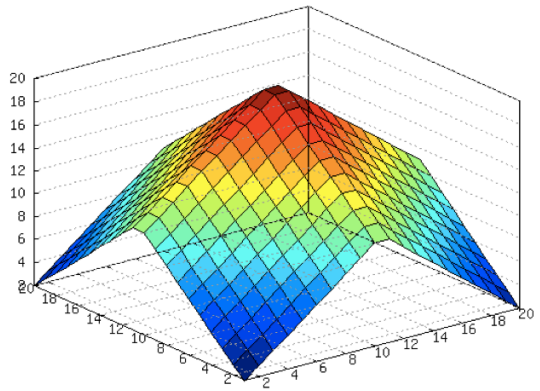
- Superior to SIFT or SURF
  - good repeatability of key-points
  - precision vs. recall
- Attractive for the video copy detection task:
  - very low computational cost
    - 6x faster that SIFT and 3x faster than SURF
  - compact descriptor
    - only 68 components

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

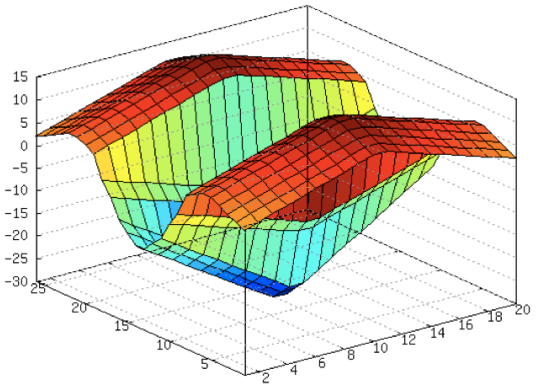
DART: key-point selection

- Efficient computation of the scale-space using piece-wise triangle filters*

2D triangle-shaped kernel



Approximation of the 2nd derivative of Gaussian

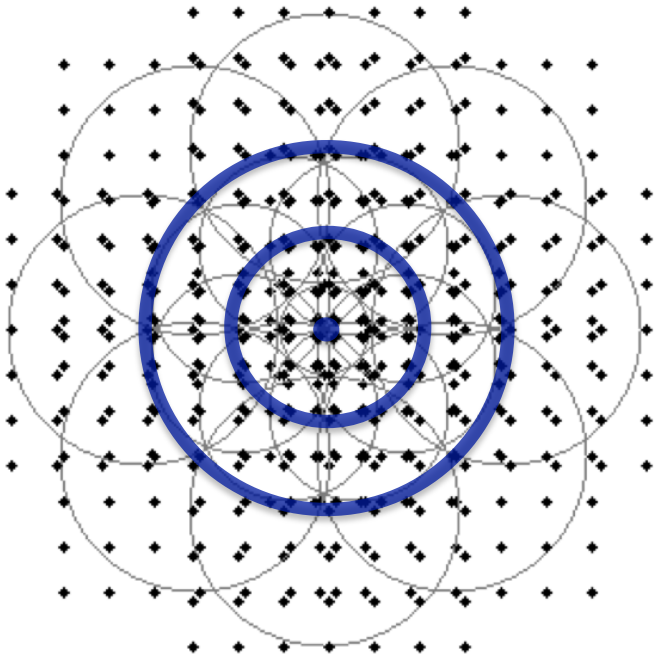


- Information reused for key-points orientation assignment and description computation

* P. Heckbert, “Filtering by repeated integration” SIGGRAPH 1986

DART: key-point description

- DAISY*-like descriptor
- Layout:
  - 2 rings, each with 8 segments
- Each segment represented by four values:
  - {|∂x |−∂x; |∂x |+∂x; |∂y|−∂y; |∂y|+∂y}
  - (1 + 2x8) x 4 = 68 components
- Segments overlap
- Re-grouping near samples into a single sample



* S. Winder, G. Hua, and M. Brown, “Picking the best daisy”, CVPR 2009.

Inserted static text and banner detection

- Sliding a temporal window of 15 key-frames
- Detection of pixels with zero standard deviation intensity
- Morphological filtering used to fill out holes
- Designed for longer videos with multiple shots
  - Problematic with short videos with static scenes

Subtitles detection

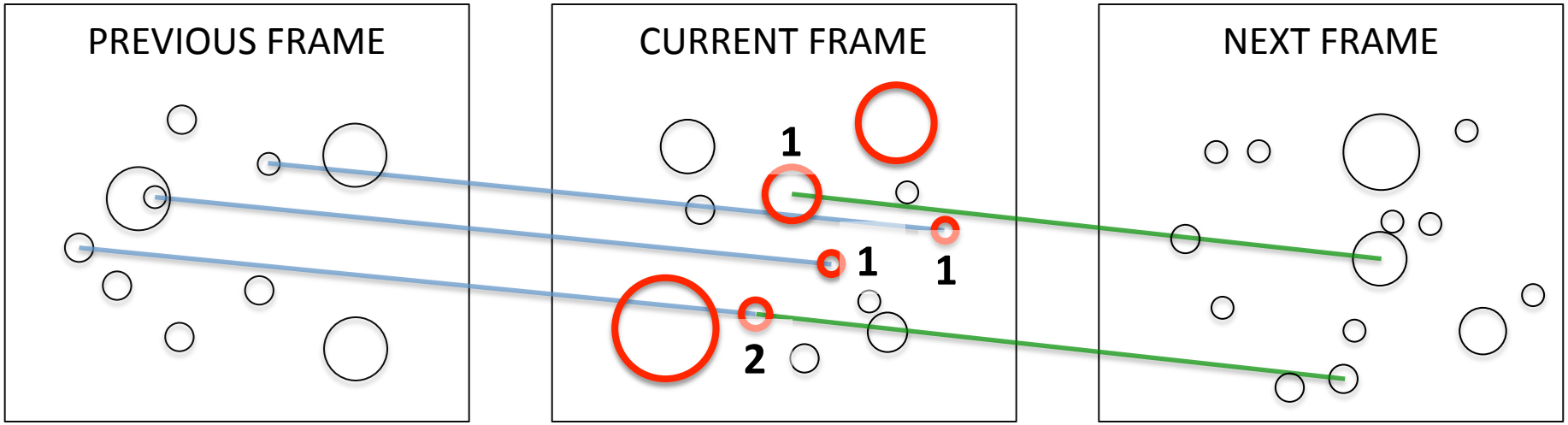
- Detecting spatial regions with high density of vertical edges
- Vertical edges computed using Sobel operator
- Edge density computed within a sliding window
- Morphological filtering filling out holes between letters

Key-point scale & temporal filtering (1/2)

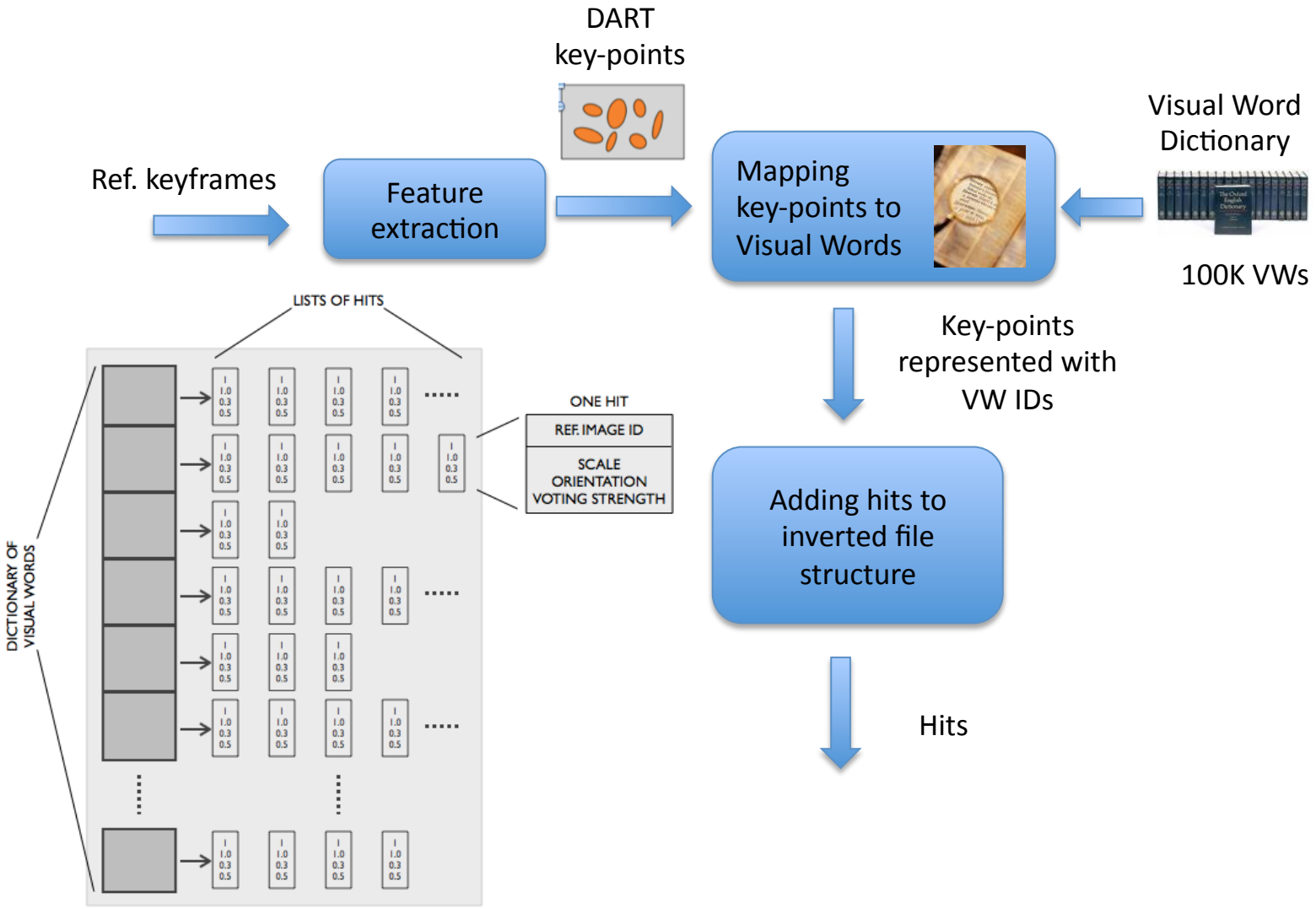
- Key-point number limits:
  - Queries: 1200 KPs
  - Reference: 400 KPs

- Not all key-points are equally useful:
  - Key-points extracted at higher scales are given more importance
  - Favoring temporarily stable key-points
    - Key-point trail length

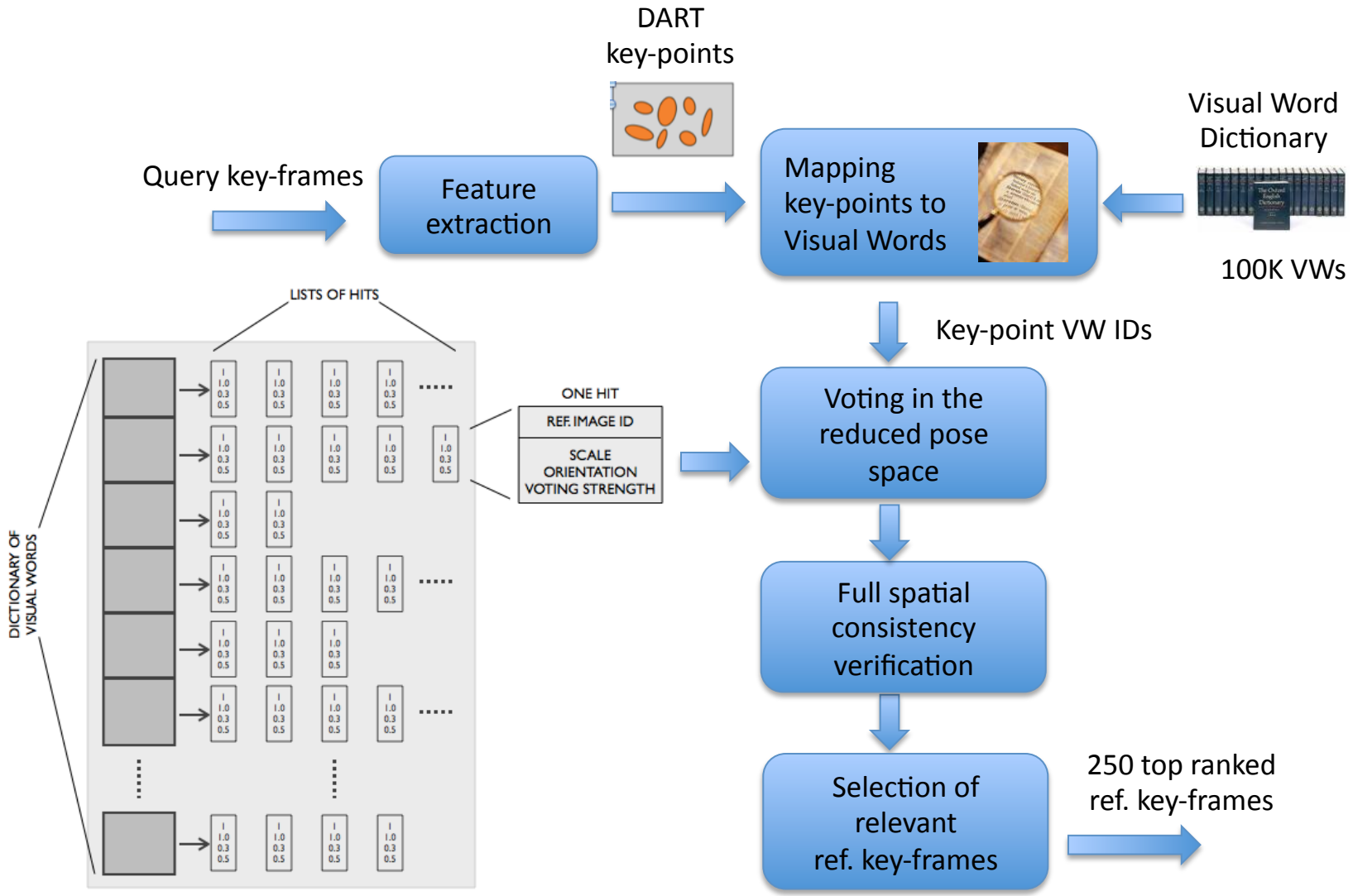
Key-point scale & temporal filtering (2/2)



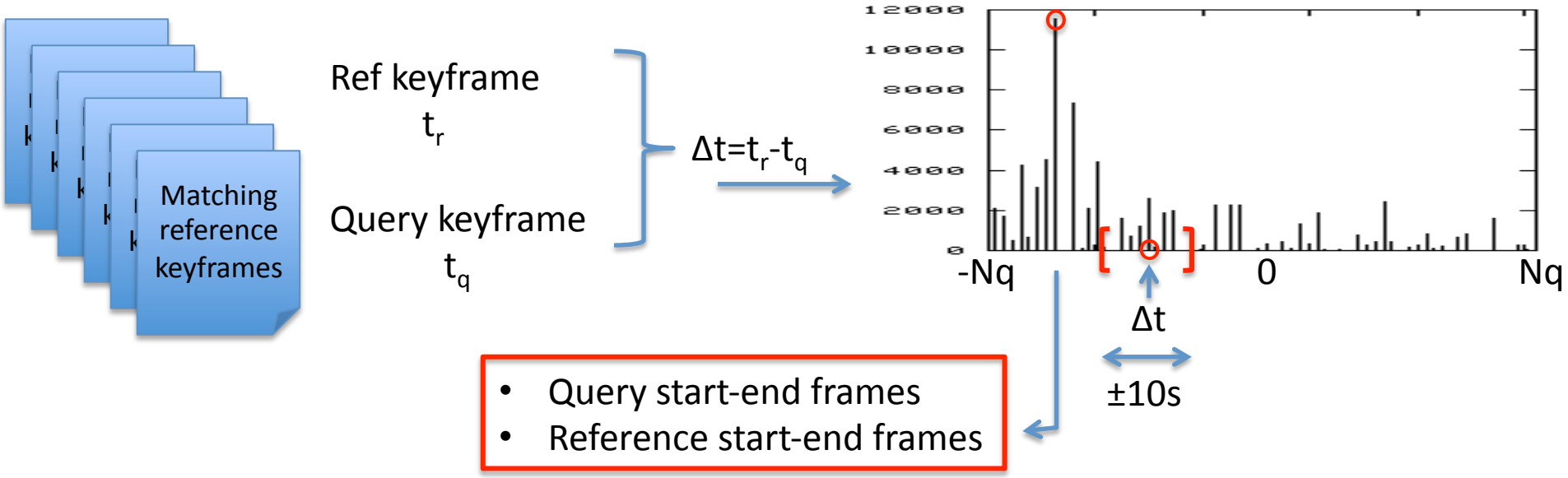
Ref. key-frame indexing



Query key-frame matching



Matching keyframes temporal consistency

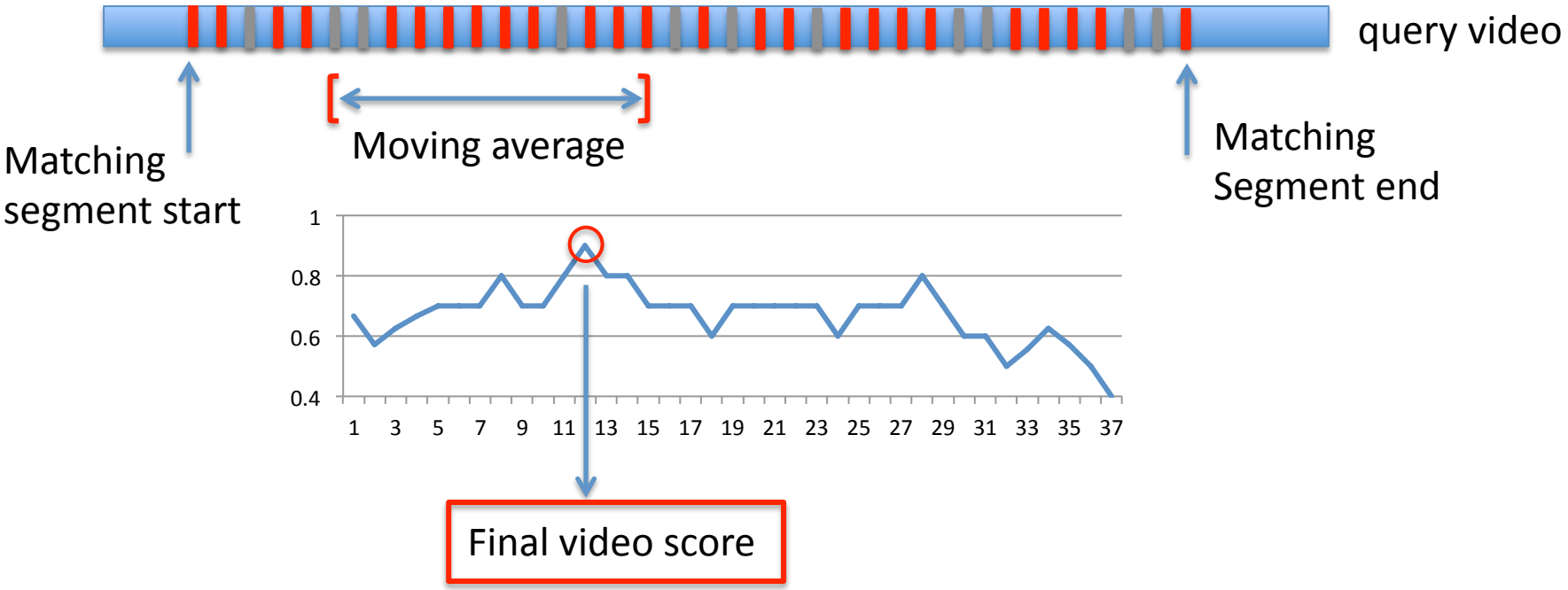


Step 1: insert all matches into a histogram based on relative times and select the 20 biggest matches

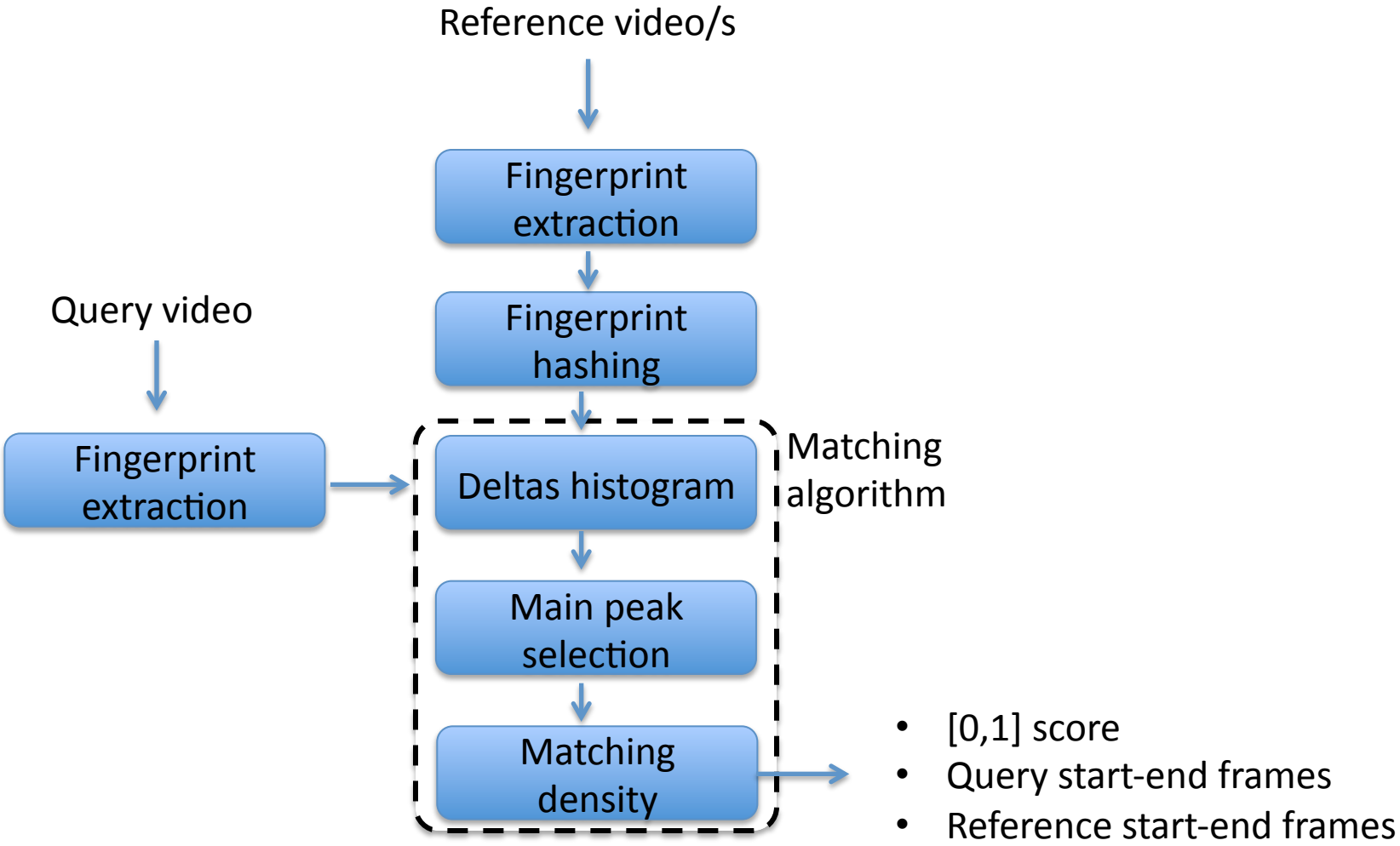
Matching keyframes temporal consistency

Step 2: compute an output score as the density of matches along a 10s window

Foreach matching video (out of 20):

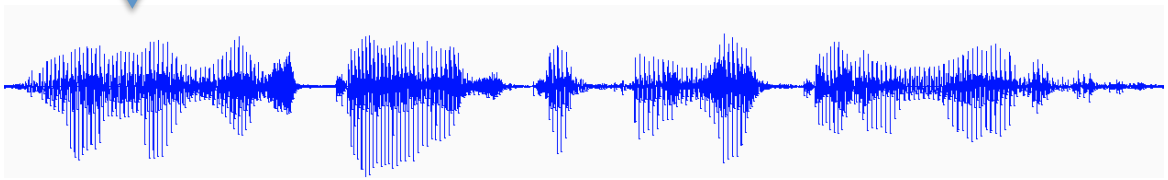


Audio-based system blocks diagram



Acoustic fingerprint extraction*

1) Audio track extraction using FFMPEG

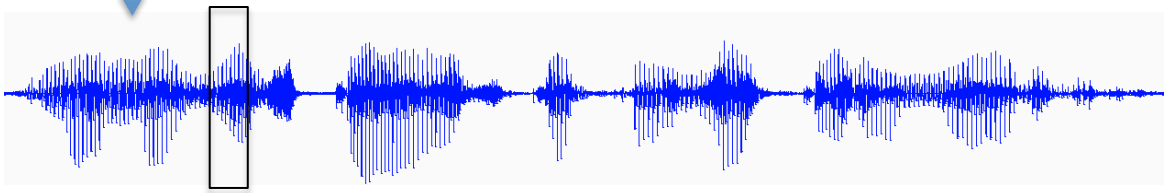


*T. Kalker and J. Haitsma. A highly robust audio finger- printing system. In *Proceedings of ISMIR'2002*, pages 144–148, 2002.

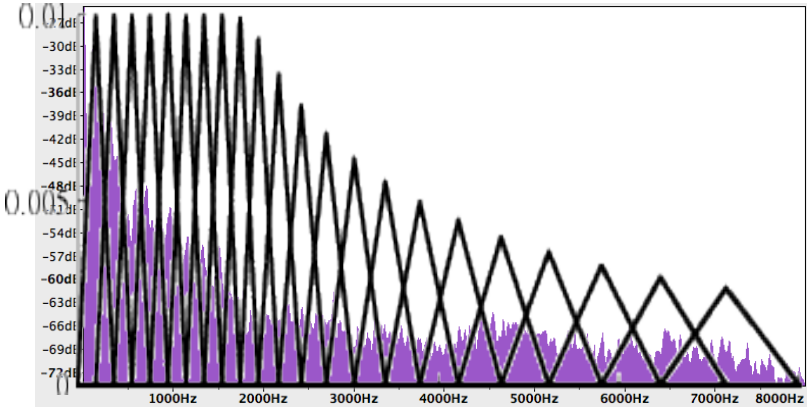
Acoustic fingerprint extraction

1) Audio track extraction using FFMPEG

10ms



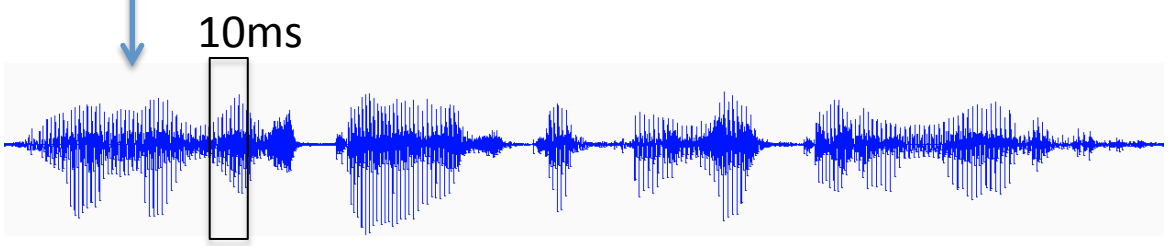
2) FFT, bandwidth limited to 300-3KHz



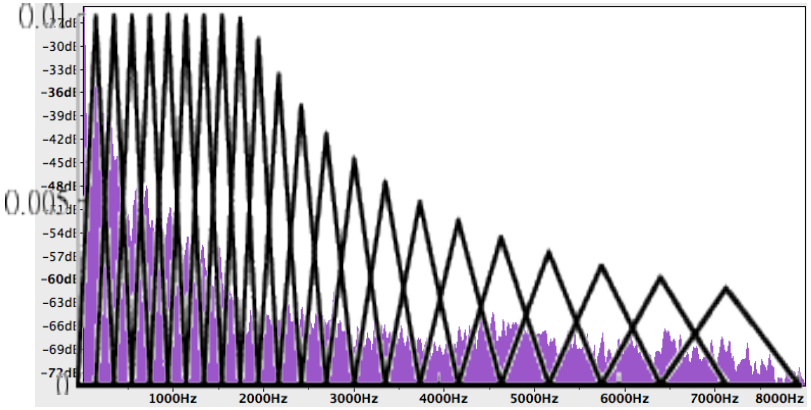
17 MEL-spectrum bands

Acoustic fingerprint extraction

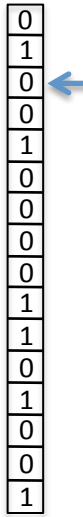
1) Audio track extraction using FFMPEG



2) FFT, bandwidth limited to 300-3KHz



17 MEL-spectrum bands

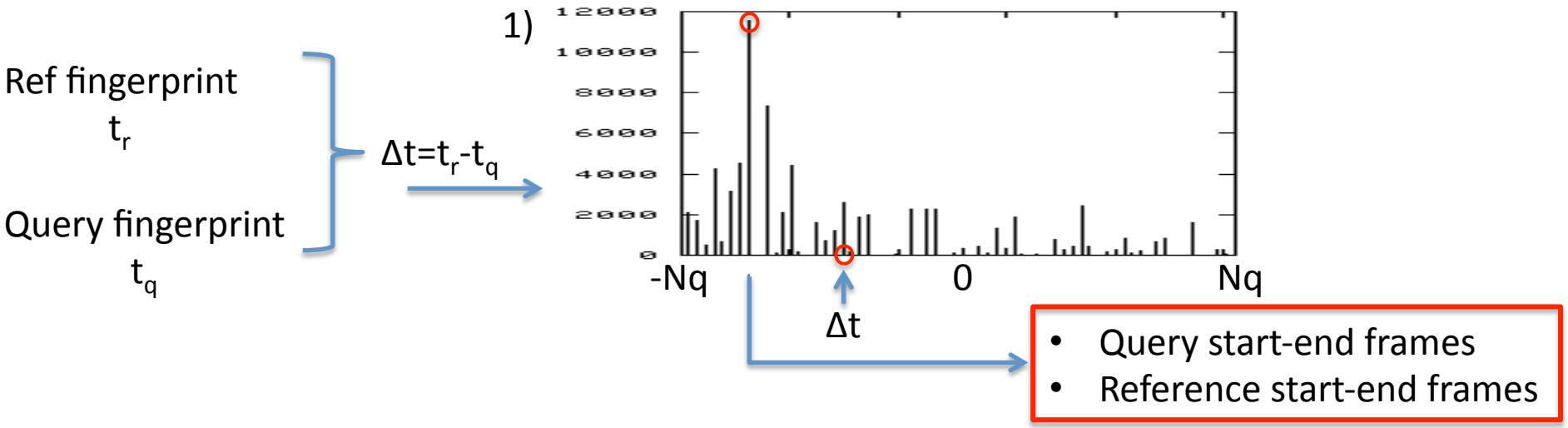


16bits

$$X[i] = \begin{cases} 0 & iif\ E_i \geq E_{i-1} \\ 1 & otherwise \end{cases}$$

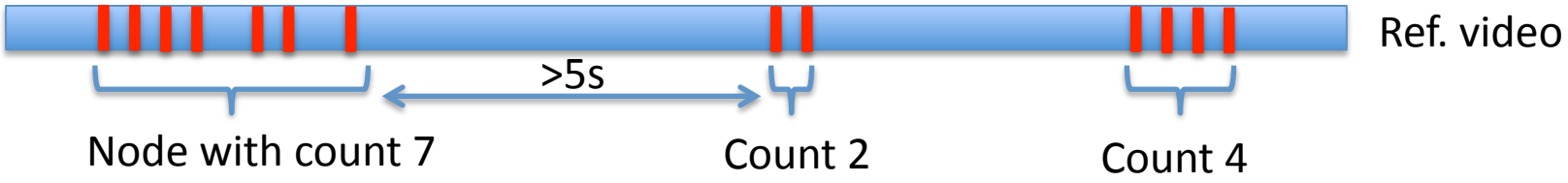
3) Contiguous bands energy comparison

Acoustic matching algorithm



Step 1: insert all matches into a histogram based on relative times and select the biggest
For every relative time a different node is created if:

- No previous reference video was found at that relative time OR
- Time difference between two matches is small (less than 5s)



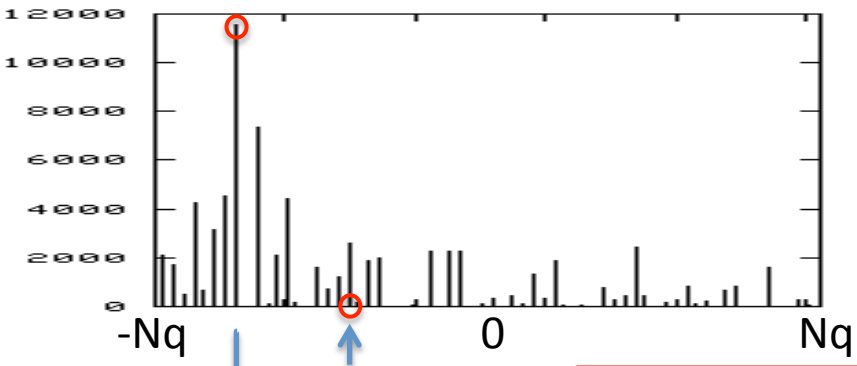
Acoustic matching algorithm

Ref fingerprint
t_r

Query fingerprint
t_q

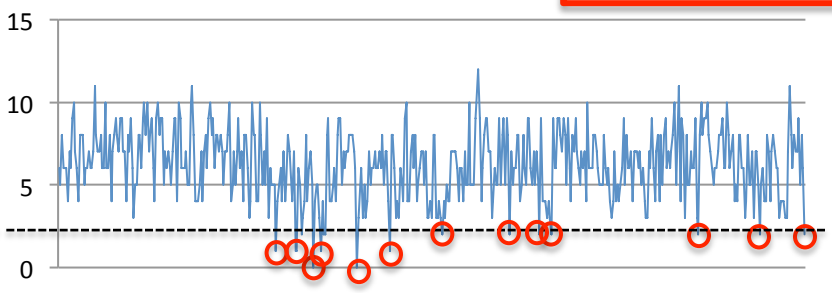
$\Delta t = t_r - t_q$

1)



- Query start-end frames
- Reference start-end frames

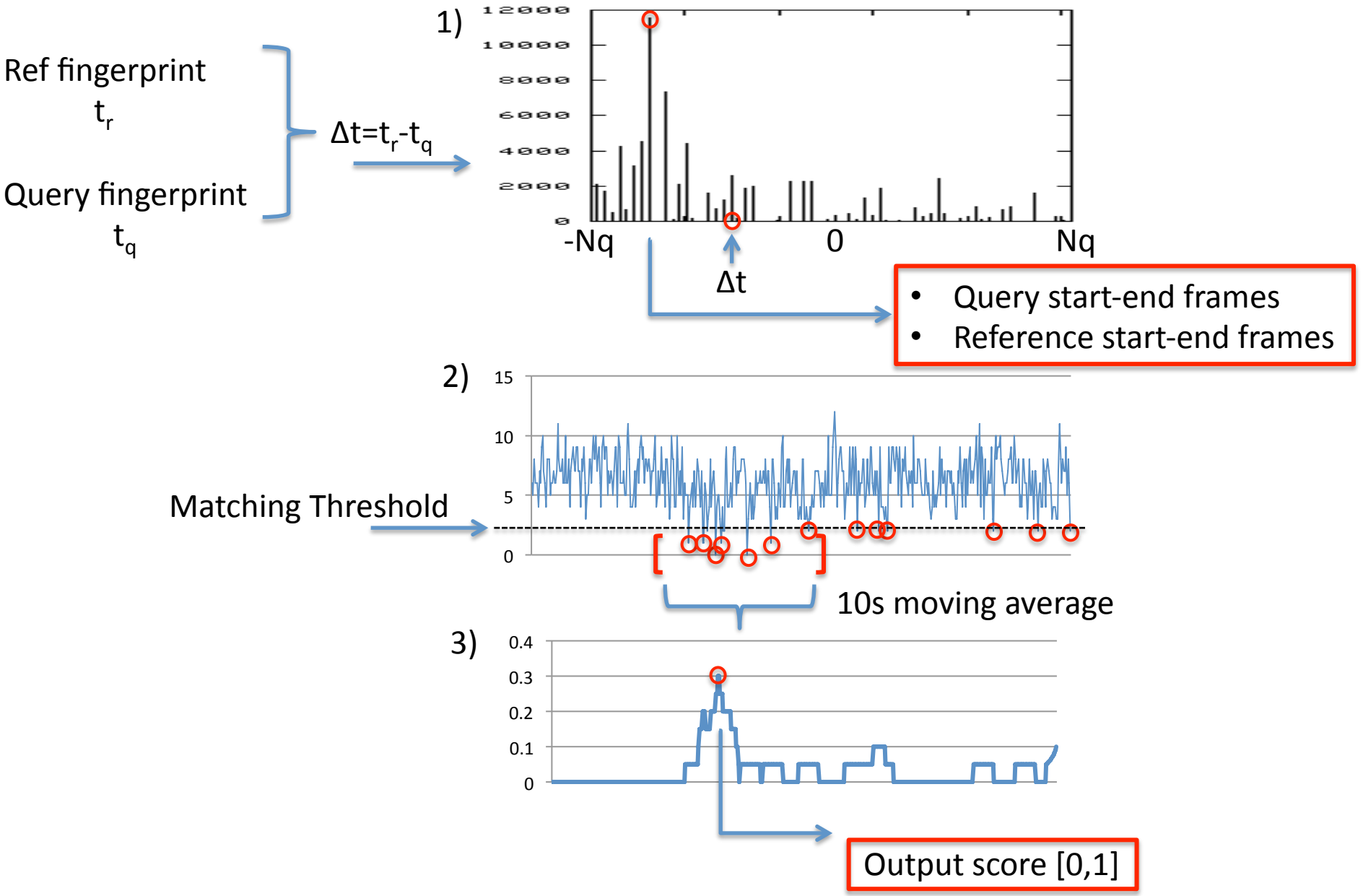
2)



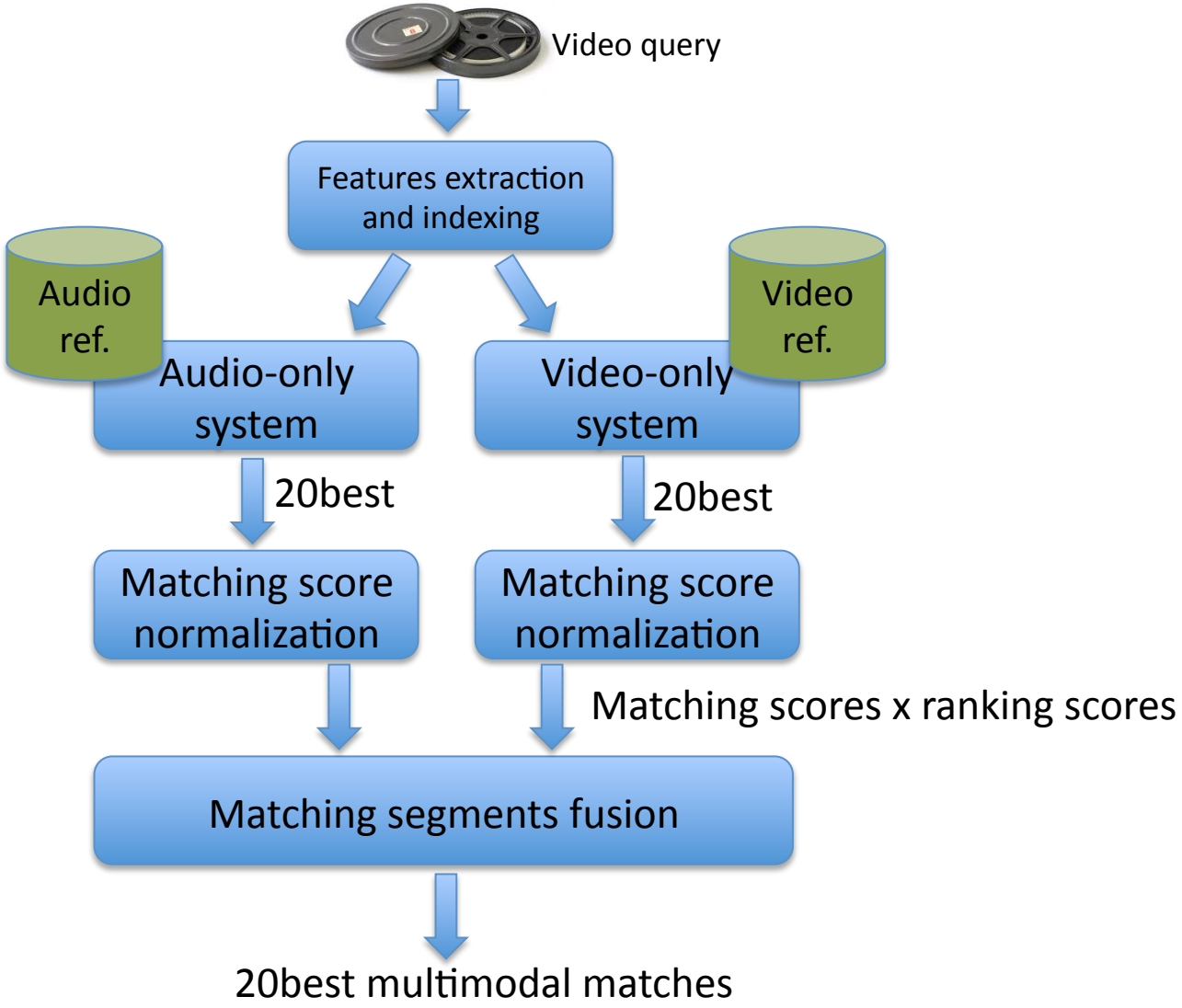
Matching Threshold

Hamming distance

Acoustic matching algorithm

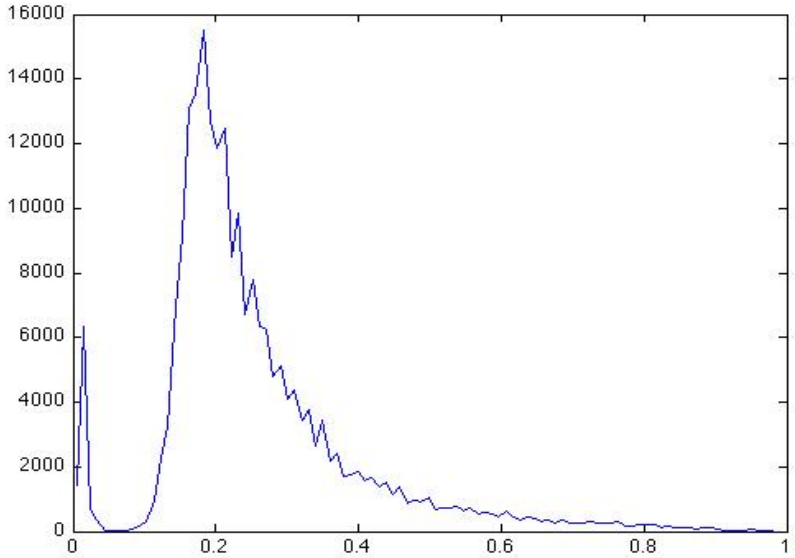


Fusion system general blocks*

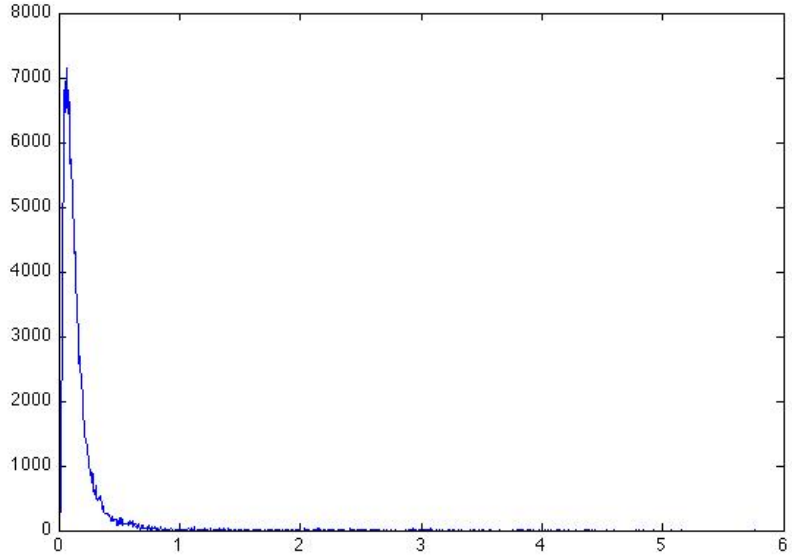


X. Olivares, M. Ciaramita, and R. van Zwol, "Boosting image retrieval through aggregating search results based on visual annotations," in Proc. ACM MM, 2008.

Fusion steps



Audio Matches score histogram



Video Matches score histogram

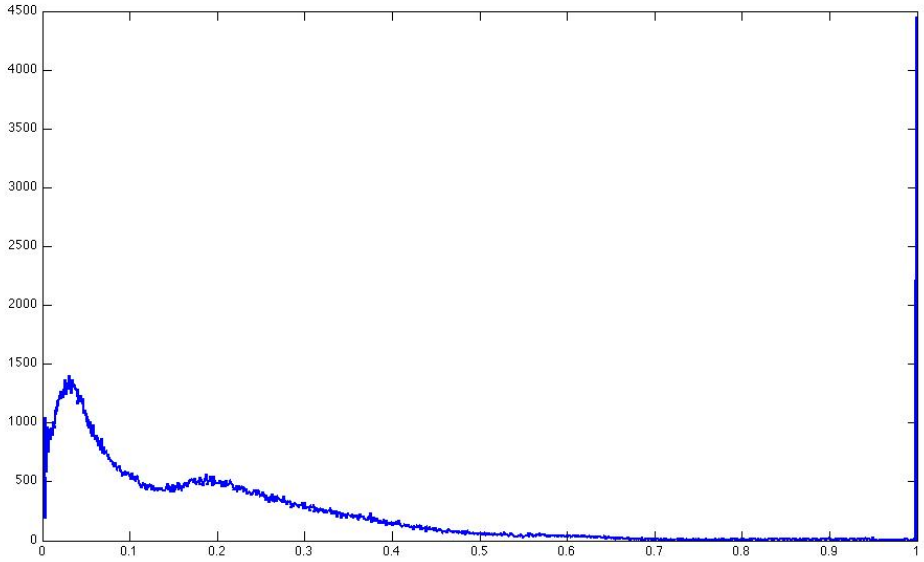
- Matching score L1 normalization

$$\overline{MScore_i} = \frac{MScore_i}{\sum_{j=1}^{20} MScore_j}$$

Fusion steps

- We consider segments with overlap > 50% between both modalities
- Combination of ranking and matching scores

$$FScore_i = \frac{\sum_k \frac{21 - rank_i^k}{20} \cdot \overline{MScore_i^k}}{\sum_k \overline{MScore_1^k}}$$

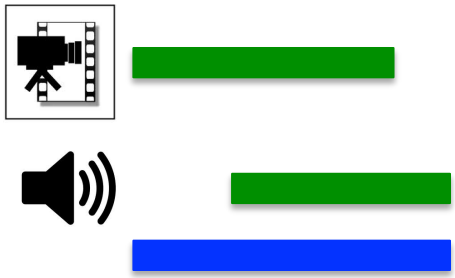
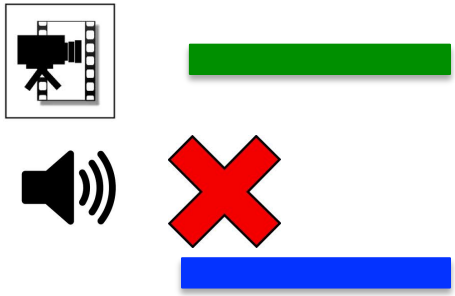
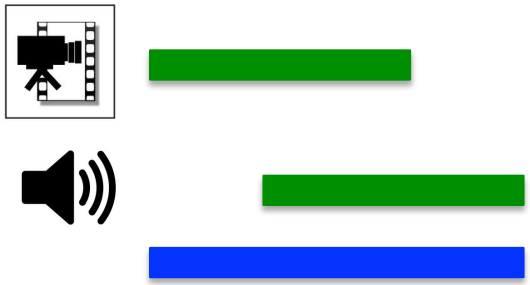


Fusion scores histogram

4422 queries with same audio & video best match
With only 2,3% FA

Fusion examples

Query Video

	Matching segments	$\overline{MScore}$	Rank score	Final score
Ref. 1		0.8 0.4	1 -> 1 2 -> 0.95	$\frac{(0.8\cdot 1+0.4\cdot 0.95)}{(0.8+0.95)}=0.67$
Ref. 2		0.5 n/a	10 -> 0.55 n/a	$\frac{(0.5\cdot 0.55+0)}{(0.8+0.95)}=0.16$
Ref. 3		0.5 0.95	10 -> 0.55 1 -> 1	$\frac{(0.5\cdot 0.55+0.95\cdot 1)}{(0.8+0.95)}=0.7$

Official evaluation results

Actual scores (averaged over all transformations), balanced profile

	NDCR	FA count	Miss count	True positives	F1 score
Audio only	43.95	407.57	30.86	**90.14**	0.93
Video only	4.83	41.63	19	81.63	0.93
Fusion	1.2	8.84	7.77	97.20	**0.91**
Position	8	10	4	8	3

Out of 134 copies per transformation

Only case where the fusion did not work better

Take home messages from the results

- Fusion is always helping to detect copies
- We got many false alarms in both video and audio, mostly due to lack of tuning
  - In general, audio fingerprints need some extra work.
- F1 is very good for videos we do detect
- Processing time… we better not report on that

Analysis of errors in audio: misses

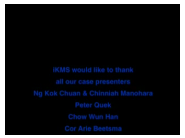
- Music getting very distorted within the 300-3KHz bands.
  - Original signal 
  - Band-limited to 300-3KHz 
- Very short audio segments (sometimes with silences) 
- Strong audio overlap + reencodings

Analysis of errors in video

REFERENCE QUERY OUR RESULT

- False alarms:
  - Wrong shot boundaries
    - static shots
    - semi-static shots
  - Wrongly matched dark blue text



- Misses:
  - Horizontal flip
  - Very small Picture in Picture
  - Heavy compression
  - Very dark and/or empty scenes



Conclusions and future work

- Fusion of multiple modalities greatly improves copy detection
  - Need to be smarter when fusing segment boundaries
- DART features are suitable for the task
- Audio fingerprints need some extra work to make them robust to IACC data
- In general, we need to reduce false alarms