# Combining Features at Search Time: PRISMA at TRECVID 2011

**Juan Manuel Barrios**[1], **Benjamin Bustos**[1], **and Xavier Anguera**[2]

[1] PRISMA Research Group, Department of Computer Science, University of Chile.

[2] Telefónica Research, Barcelona, Spain.

Content-Based Video Copy Detection Task, TRECVID.
December 7, 2011

# P-VCD Overview

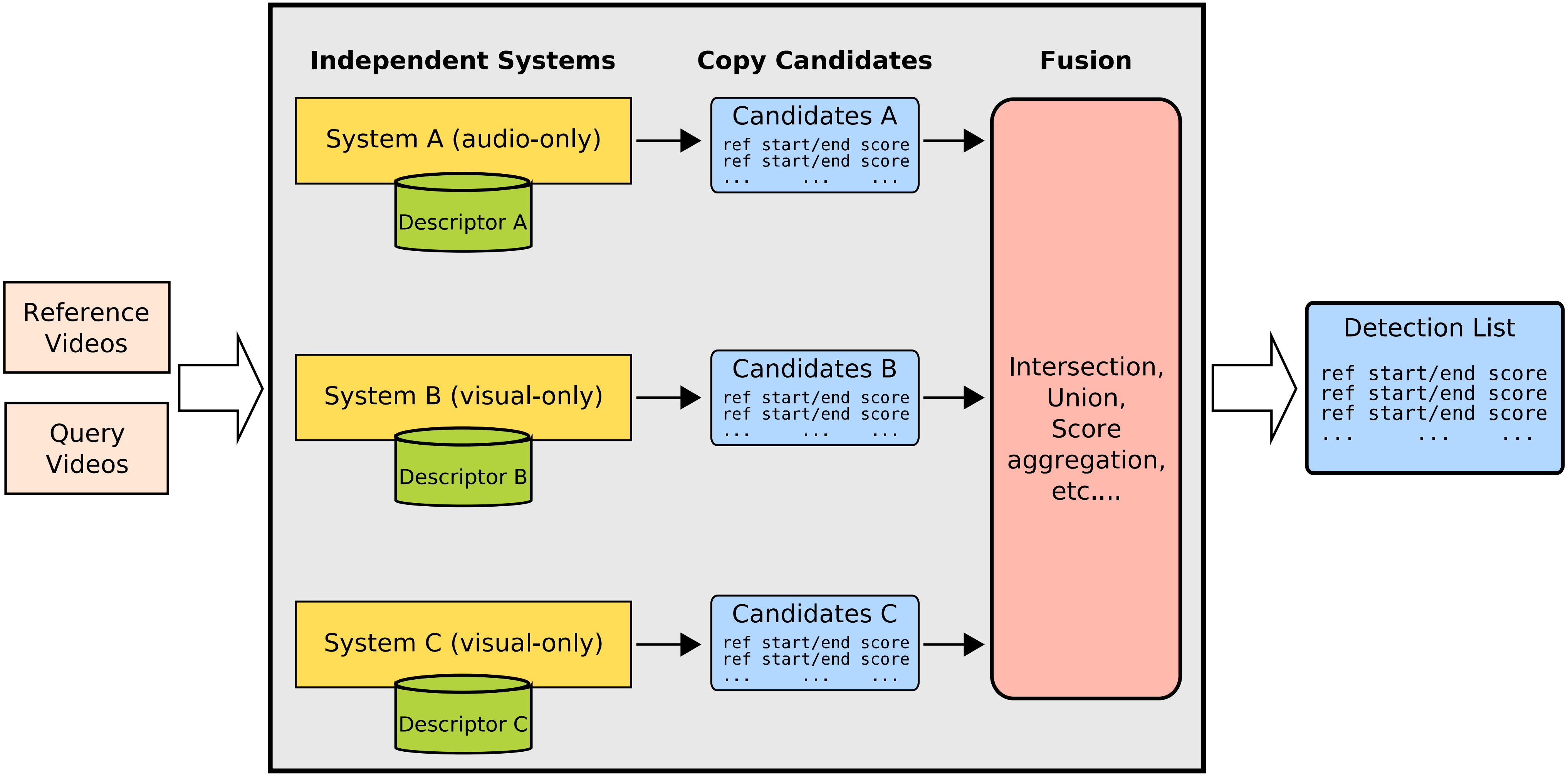
- P-VCD System developed for TRECVID 2010. [1]

- **2010**: Visual-only detection.
  - Global descriptors.
  - Approximate k-NN search using pivots.

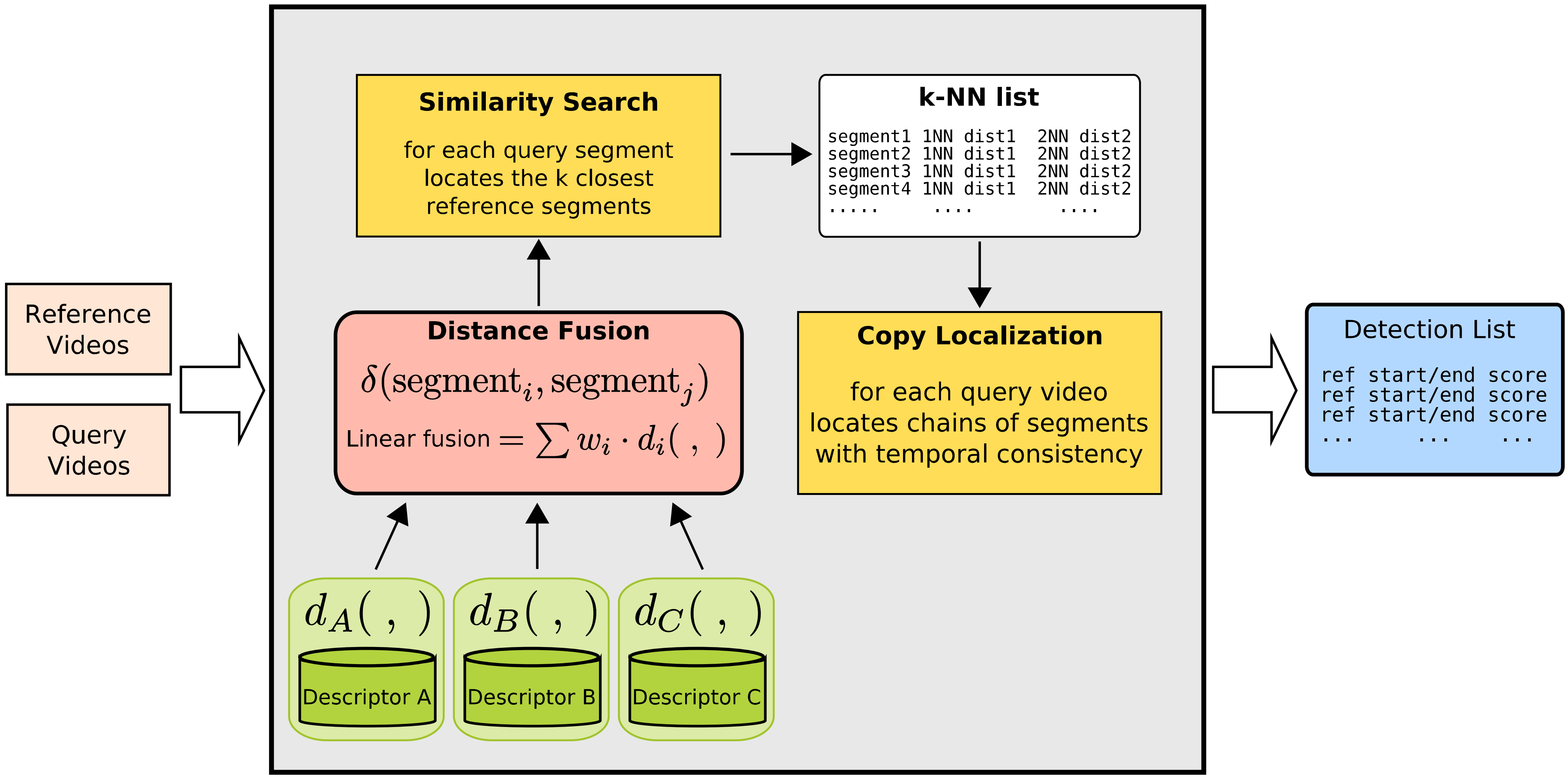
- **2011**: Audio+Visual detection.
  - Fusion of audio and global descriptors at the similarity search: “distance fusion”.
  - Approximate search as a filtering step.
  - Sequential (exact) A+V search.

[1] J.M.Barrios and B.Bustos. *Competitive content-based video copy detection using global descriptors*. Multimedia Tools and Applications. Springer, 2011.

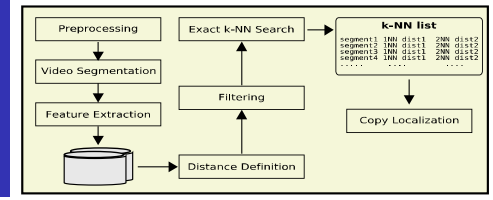
# Fusion at Decision Level

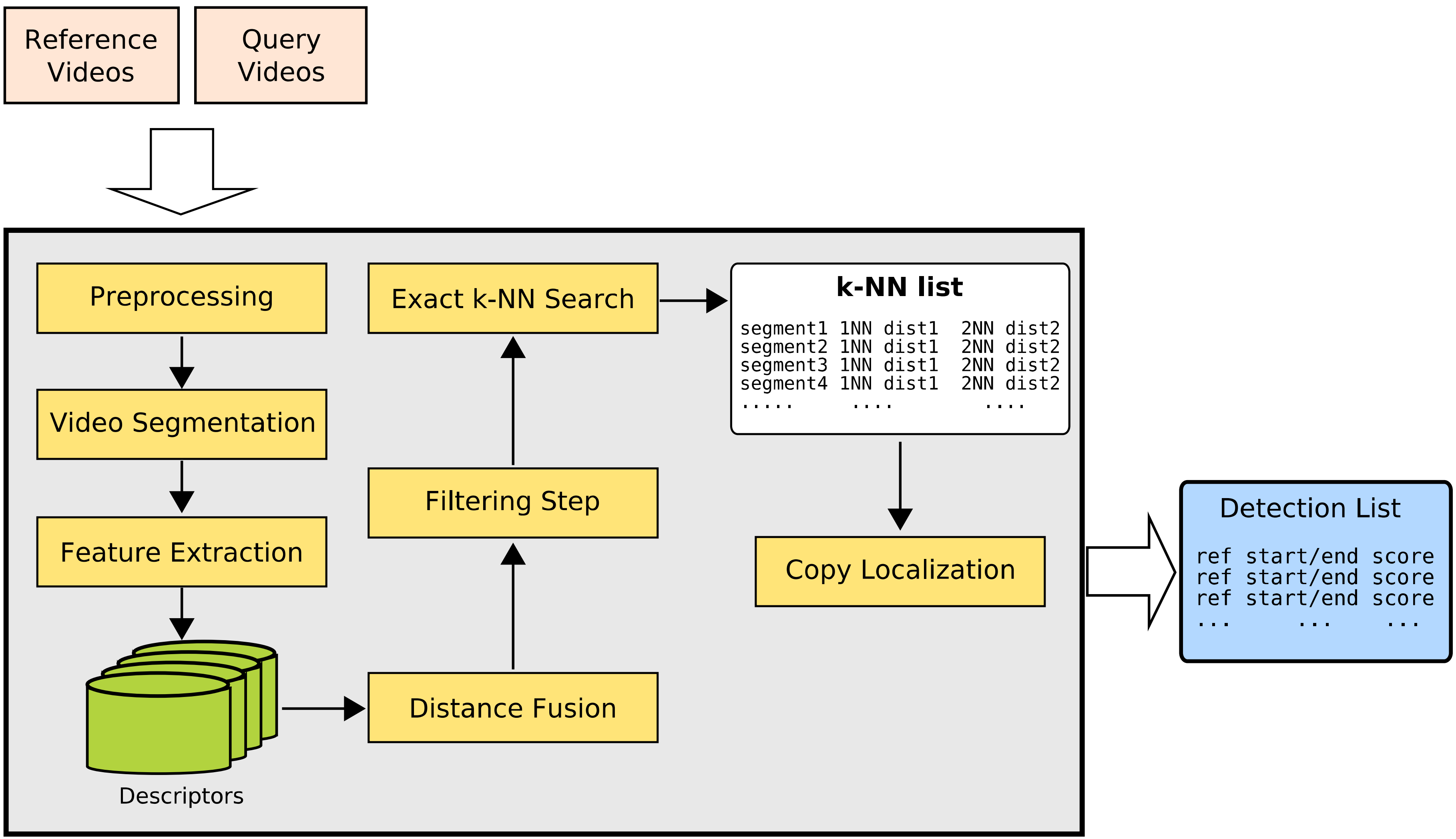


# Fusion at Similarity Search Level

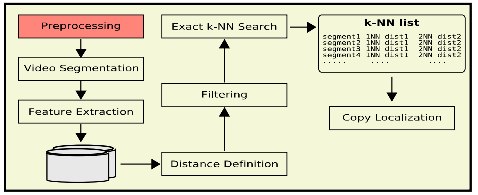


# P-VCD 2011 Overview

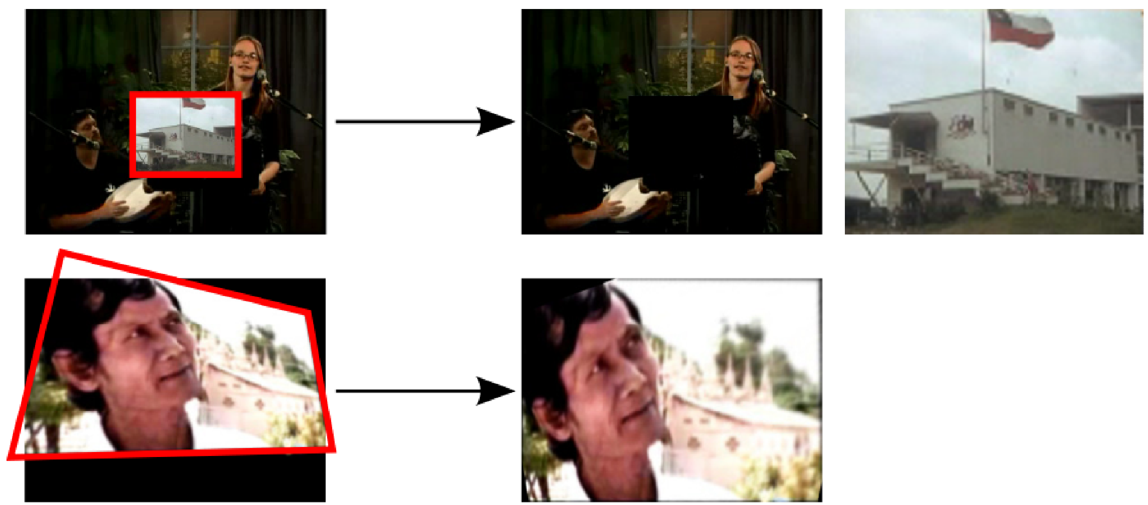


# 1. Preprocessing

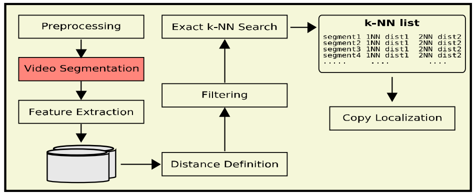



- Removes black borders and noisy frames from each query and reference video.
- For each query video, it creates a flipped version and detects and reverts PIP and camcording.

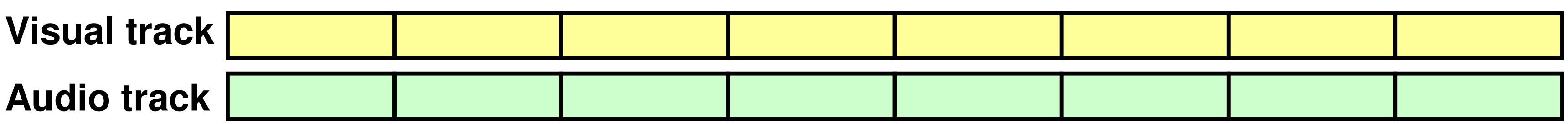


| | Audio | Visual | Audio+Visual |
|---|---|---|---|
| **Original Queries** | 1,407 | 1,608 | 11,256 |
| **New Queries** | - | 3,539 | - |
| **Total Queries** | **1,407** | **5,147** | **36,029** |

# 2. Video Segmentation

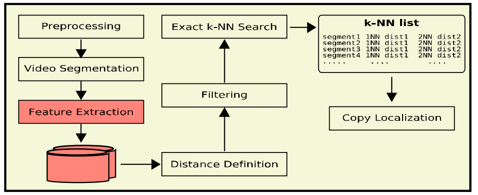



- Partitions every query and reference video into segments of 0.333 ms length (visual and audio track).

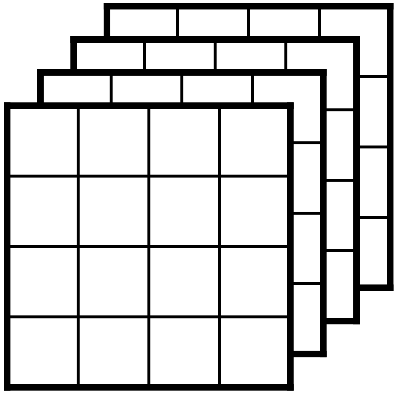



| | Audio segments | Visual segments | Audio+Visual segments |
|---|---|---|---|
| **Query collection** | 306,304 | 1,120,455 | 7,840,587 |
| **Reference collection** | 4,441,717 | 4,522,262 | 4,387,633 |

# 3. Feature Extraction

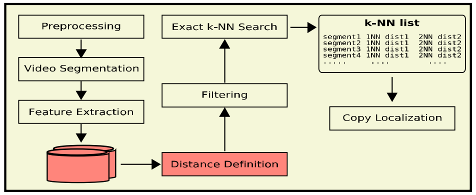



- Three Visual-Global descriptors per segment:
  - Edge Histogram (Ehd): 4x4x10 =160d.
  - Gray Histogram (Gry): 4x4x12 = 192d.
  - Color Histogram (Rgb): 4x4x12 = 192d.
- The descriptor for a visual segment is the average descriptor for every frame.



- One Audio Descriptor (Aud), 160d.

# 4. Distance Fusion




- Distance between two descriptors: Manhattan distance (city-block)

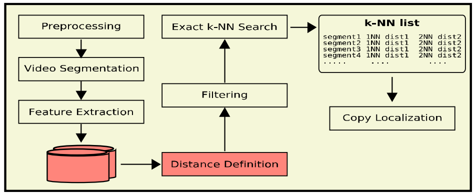
$$L_1(\vec{x},\vec{y}) = \sum_{i=0}^{dim} |x_i - y_i|$$

- Distance between any two Audio+Visual segments:

$$\begin{aligned} d_{av}(q,r) &= \frac{w_1}{\tau_1} * L_1(\mathrm{Ehd}(q),\mathrm{Ehd}(r)) + \frac{w_2}{\tau_2} * L_1(\mathrm{Rgb}(q),\mathrm{Rgb}(r)) \\ &+ \frac{w_3}{\tau_3} * L_1(\mathrm{Aud}(q),\mathrm{Aud}(r)) \end{aligned}$$

- Normalization factors $\tau_i$ and weighting factors $w_i$ are calculated by the "$\alpha$-Normalization" and "weighting by max-$\tau$" algorithms. [1]

# 4. Distance Fusion (cont.)




- For efficiency, we define two more distances:
  - Between two audio segments:

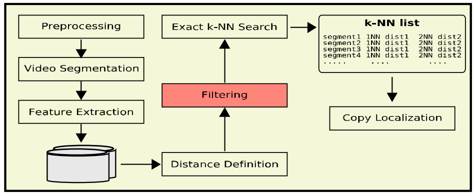
$$d_a(q,r) = L_1(\text{Aud}(q), \text{Aud}(r))$$

  - Between two visual segments:

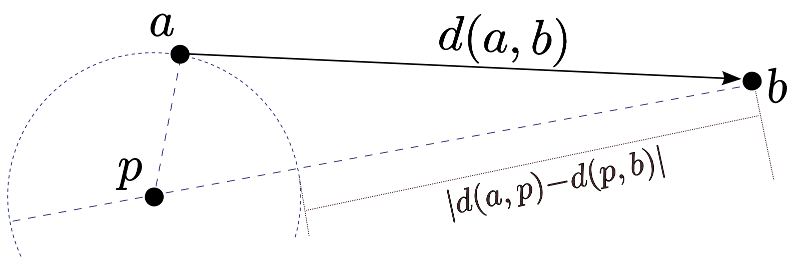
$$d_v(q,r) = \frac{w_1}{\tau_1} * L_1(\text{Ehd}(q), \text{Ehd}(r)) + \frac{w_2}{\tau_2} * L_1(\text{Rgb}(q), \text{Rgb}(r))$$

$$d_v(q,r) = \frac{w_1}{\tau_1} * L_1(\text{Ehd}(q), \text{Ehd}(r)) + \frac{w_2}{\tau_2} * L_1(\text{Gry}(q), \text{Gry}(r))$$

# 5. Search Domain Filtering




- It performs approximate k-NN searches [1] using visual-only distance and audio-only distance.
  - Requirement: $d$ complies the triangle inequality.

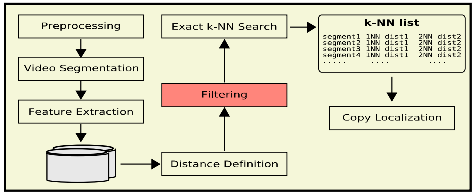



- Distance approximation: $d(a,b) \approx |d(a,p) - d(p,b)|$

- For many pivots: $d(a,b) \approx \max_{p \in \mathcal{P}} |d(a,p) - d(p,b)|$

- It evaluates the actual distance only for the pairs with lowest approximated distance.

# 5. Search Domain Filtering




- Perform approximate k-NN searches for each query segment using visual-only distance and audio-only distance (k=30).

| Query Video |
|---|
| Query1-Segm1 |
| Query1-Segm2 |
| Query1-Segm3 |
| Query1-Segm4 |
| Query1-Segm5 |
| Query1-Segm6 |
| Query1-Segm7 |
| Query1-Segm8 |
| ... |

**k-NN list** $d_a$

| 1st NN | | 2nd NN | |
|---|---|---|---|
| Vid07-Segm54 | dist | Vid08-Segm73 | dist |
| Vid09-Segm13 | dist | Vid02-Segm34 | dist |
| Vid07-Segm34 | dist | Vid03-Segm54 | dist |
| Vid09-Segm15 | dist | Vid02-Segm13 | dist |
| Vid01-Segm88 | dist | Vid01-Segm12 | dist |
| Vid09-Segm54 | dist | Vid09-Segm17 | dist |
| Vid01-Segm45 | dist | Vid03-Segm43 | dist |
| Vid09-Segm19 | dist | Vid01-Segm12 | dist |
| ... | | ... | |

**k-NN list** $d_v$

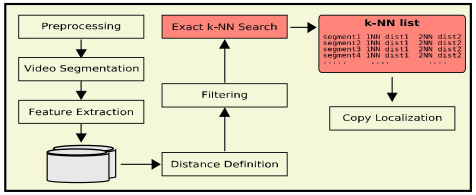
| 1st NN | | 2nd NN | |
|---|---|---|---|
| Vid07-Segm54 | dist | Vid08-Segm73 | dist |
| Vid09-Segm13 | dist | Vid02-Segm34 | dist |
| Vid07-Segm34 | dist | Vid03-Segm54 | dist |
| Vid09-Segm15 | dist | Vid02-Segm13 | dist |
| Vid01-Segm88 | dist | Vid01-Segm12 | dist |
| Vid09-Segm54 | dist | Vid09-Segm17 | dist |
| Vid01-Segm45 | dist | Vid03-Segm43 | dist |
| Vid09-Segm19 | dist | Vid01-Segm12 | dist |
| ... | | ... | |

- For each query video, it selects the *D* reference videos that have more segments in the k-NN lists (*D*=40).

Query1 $\longrightarrow$ {Vid01, Vid02, Vid03, Vid07, Vid08, Vid09}

Query2 $\longrightarrow$ {Vid02, Vid04, Vid06, Vid07}

# 6. Exact k-NN Search




- For each query segment performs an exact k-NN search using the audio+visual distance (k=10).
- The search space domain depends on each query video.

Query1 $\longrightarrow$ {Vid01, Vid02, Vid03, Vid07, Vid08, Vid09}

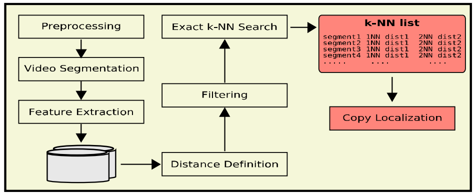
**k-NN list** $d_{av}$

| Query Video |
|---|
| Query1-Segm1 |
| Query1-Segm2 |
| Query1-Segm3 |
| Query1-Segm4 |
| Query1-Segm5 |
| Query1-Segm6 |
| Query1-Segm7 |
| Query1-Segm8 |
| ... |

$\Rightarrow$

| 1st NN | | 2nd NN | | 3rd NN | |
|---|---|---|---|---|---|
| Vid07-Segm54 | dist | Vid08-Segm73 | dist | Vid01-Segm68 | dist |
| Vid09-Segm13 | dist | Vid02-Segm34 | dist | Vid02-Segm33 | dist |
| Vid07-Segm34 | dist | Vid03-Segm54 | dist | Vid09-Segm14 | dist |
| Vid09-Segm15 | dist | Vid02-Segm13 | dist | Vid03-Segm65 | dist |
| Vid01-Segm88 | dist | Vid01-Segm12 | dist | Vid07-Segm58 | dist |
| Vid09-Segm54 | dist | Vid09-Segm17 | dist | Vid07-Segm59 | dist |
| Vid01-Segm45 | dist | Vid03-Segm43 | dist | Vid03-Segm20 | dist |
| Vid09-Segm19 | dist | Vid01-Segm12 | dist | Vid07-Segm61 | dist |
| ... | | ... | | ... | |

# 7. Copy Localization




- Locates chains of NN with temporal consistency. [1]

**k-NN list** $d_{av}$

| Query Video | 1st NN | 2nd NN | 3rd NN |
|---|---|---|---|
| Query1-Segm1 | Vid07-Segm54 dist | Vid08-Segm73 dist | Vid01-Segm68 dist |
| **Query1-Segm2** | **Vid09-Segm13 dist** | Vid02-Segm34 dist | Vid02-Segm33 dist |
| **Query1-Segm3** | Vid07-Segm34 dist | Vid03-Segm54 dist | **Vid09-Segm14 dist** |
| **Query1-Segm4** | **Vid09-Segm15 dist** | Vid02-Segm13 dist | Vid03-Segm65 dist |
| Query1-Segm5 | Vid01-Segm88 dist | Vid01-Segm12 dist | Vid07-Segm58 dist |
| **Query1-Segm6** | Vid09-Segm54 dist | **Vid09-Segm17 dist** | Vid07-Segm59 dist |
| Query1-Segm7 | Vid01-Segm45 dist | Vid03-Segm43 dist | Vid03-Segm20 dist |
| **Query1-Segm8** | **Vid09-Segm19 dist** | Vid01-Segm12 dist | Vid07-Segm61 dist |
| ... | ... | ... | ... |

- No False Alarms profile:
  - It reports the candidate with the highest score.
- Balanced profile:
  - It reports the two candidates with highest scores.

Detection List

| | | | | |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |
| **Query1** | **Segm2-Segm8** | **Vid09** | **Segm13-Segm19** | **score** |
| ... | ... | ... | ... | ... |

# TRECVID 2011 Results

# No False Alarms profile

- Analysis focused on optimal threshold and average result for all transformations.
- No False Alarms profile:
  - One candidate per query.
  - `EhdGry`: Combination of two global descriptors
    - Average Optimal NDCR=**0.374**
    - Average Optimal F1=**0.938**
    - Average Processing Time=**50** s
  - `EhdRgbAud`: Combination of two global descriptors and audio
    - Average Optimal NDCR=**0.286**
    - Average Optimal F1=**0.946**
    - Average Processing Time=**64** s

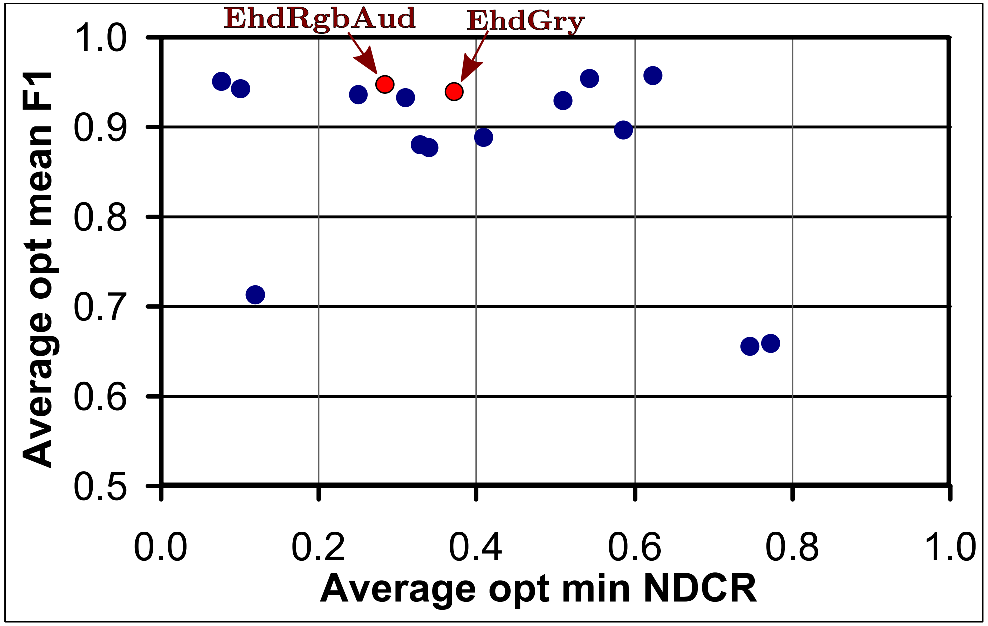
**TRECVID 2010**

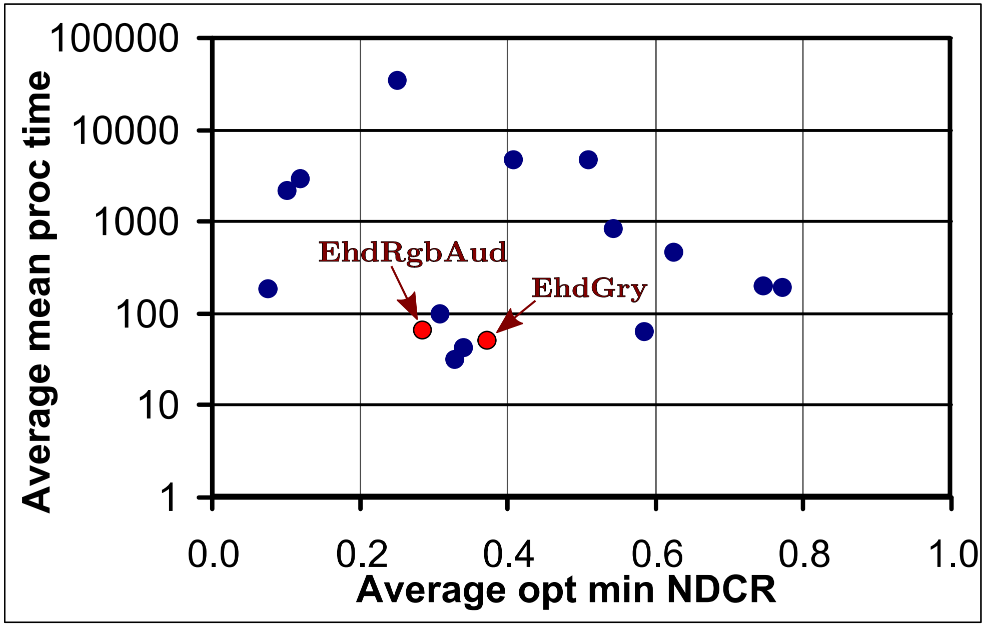
Avg.Opt.NDCR=**0.611**

Avg.Opt.F1=**0.828**

Avg.Proc.Time=**128** s

# No False Alarms profile



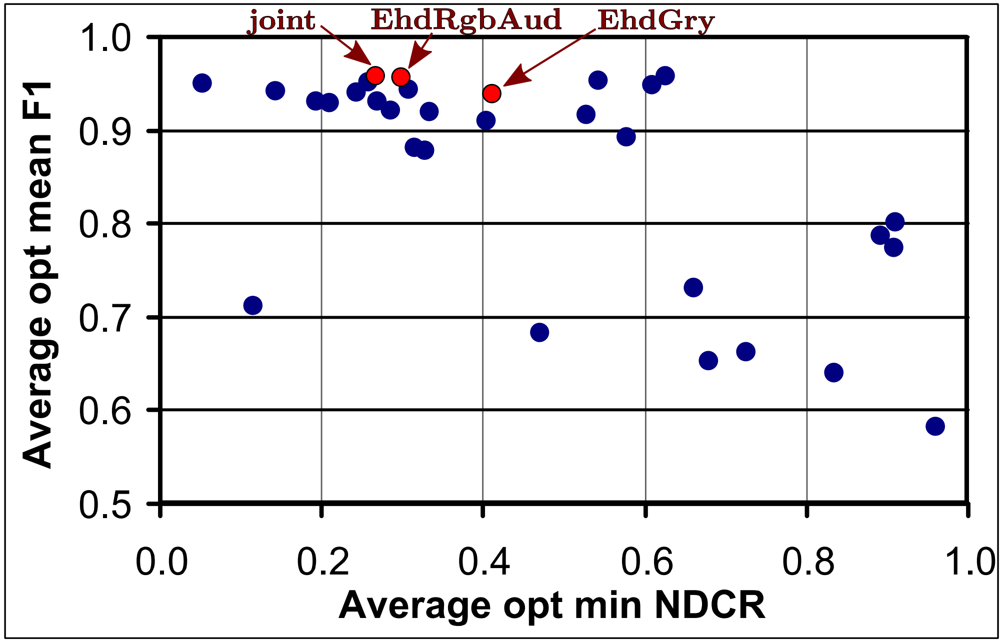

- Multimodal detection outperforms visual-only detection.
- The exact search step increases the accuracy for copy localization.
- Good tradeoff between effectiveness and efficiency.
- Global descriptors can achieve good performance in NoFA profile.

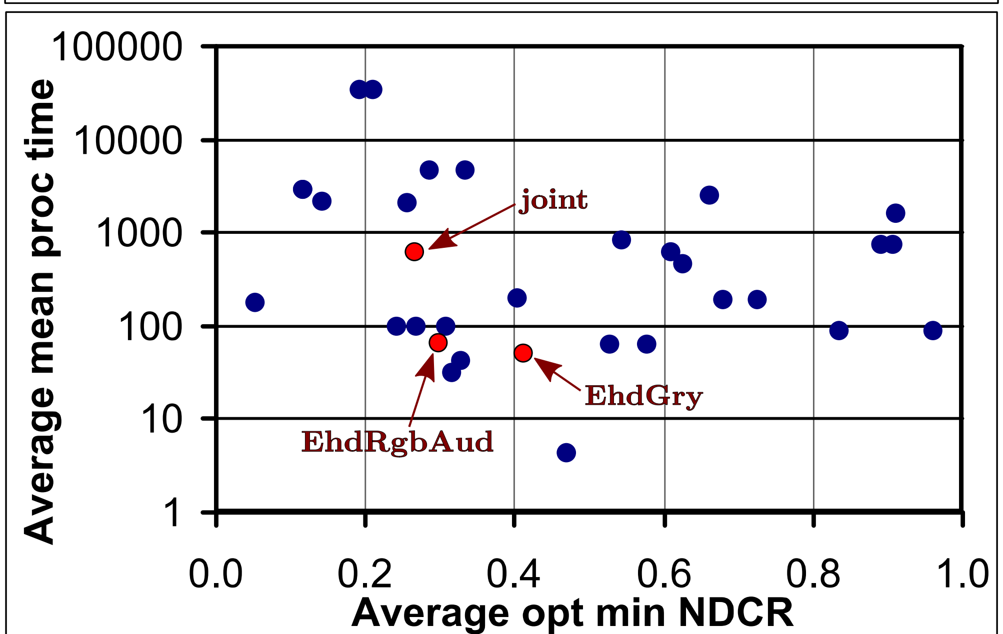
# Balanced profile

- Balanced profile:
  - Two candidates per query.
  - `EhdGry`: Combination of two global descriptors
    - Average Optimal NDCR=**0.412**
    - Average Optimal F1=**0.938**
    - Average Processing Time=**50** s
  - `EhdRgbAud`: Combination of two global descriptors and audio
    - Average NDCR=**0.300**
    - Average F1=**0.955**
    - Average Processing Time=**64** s
  - `Joint` submission with Telefonica team.
    - `EhdRgb` with twenty candidates per query.
    - Late fusion with Telefonica's audio and local descriptors.

**TRECVID 2010**
Avg.Opt.NDCR=**0.597**
Avg.Opt.F1=**0.820**
Avg.Proc.Time=**128** s

# Balanced profile



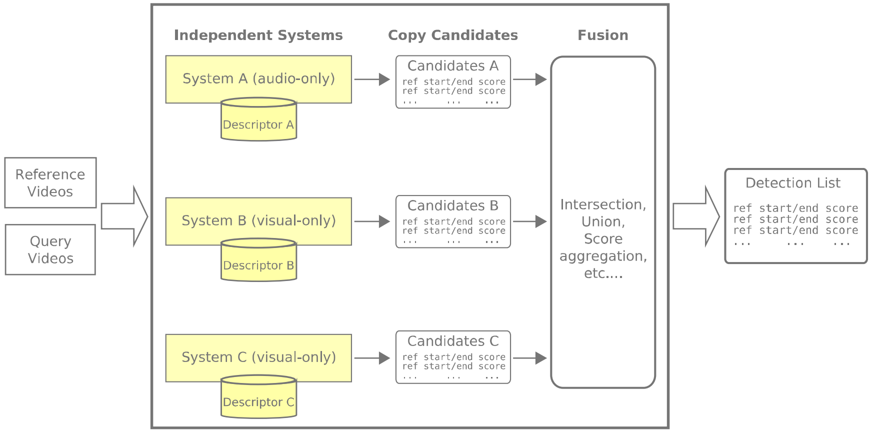

- Good localization accuracy.
- Good tradeoff between effectiveness and efficiency.
- Global descriptors achieve better performance in NoFA profile than in Balanced profile.
- All these tests were run on a desktop computer:
  - Intel Core i7-2600k
  - 8 GB RAM

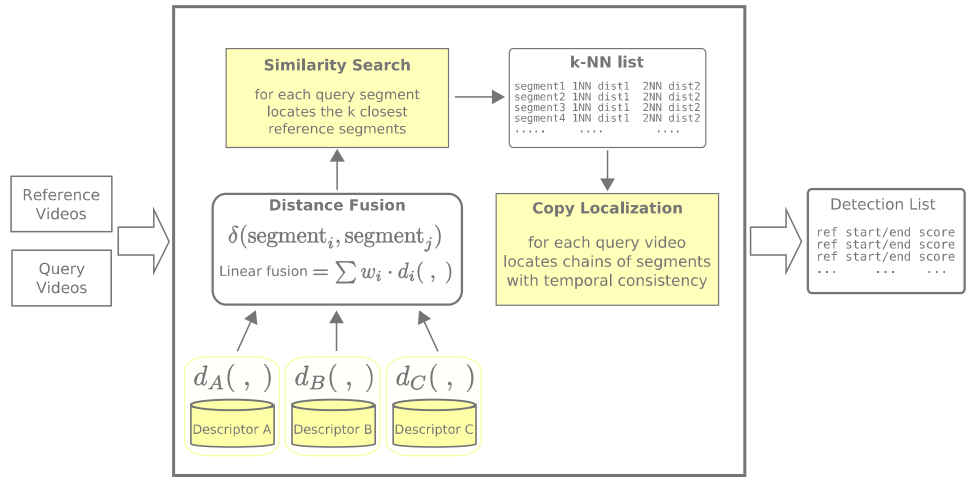
# Conclusions

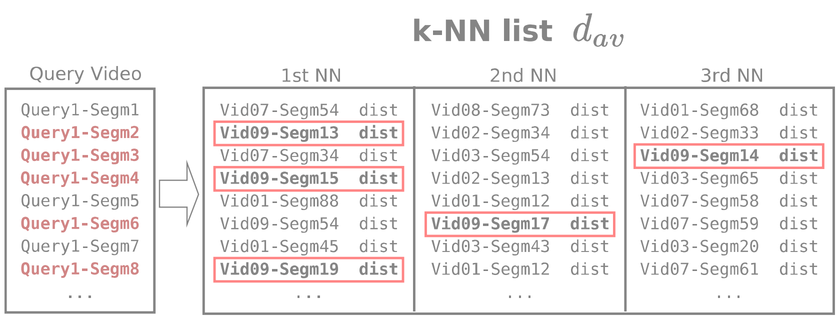
- We have presented the "distance fusion" approach for combining global and audio descriptors.
  - It automatically fixes a good set of weigths.
- The approximate search can avoid most of the distance evaluations while achieving a good detection performance.
  - The analysis of the approximate search is in [1].
- The exact search step increases the accuracy for the copy localization.

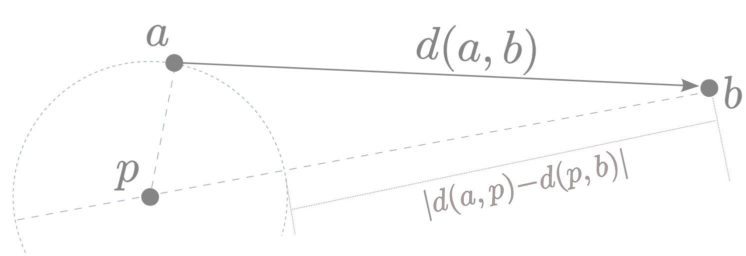
- Future work:
  - Fuse audio, global and local descriptors following this approach.
  - Test non-metric distances at the exact search step.
  - Test a segmentation with overlaps.

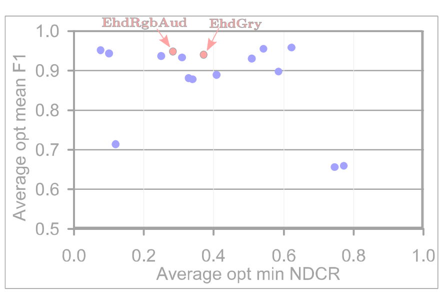
# Thank you!

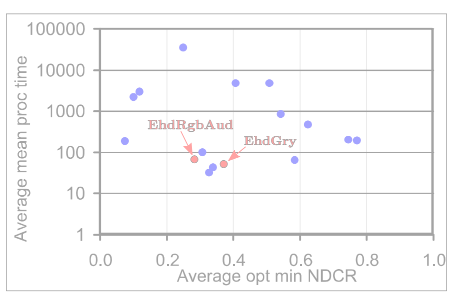

Independent Systems
Copy Candidates
Fusion
Reference Videos
Query Videos
System A (audio-only)
Descriptor A
System B (visual-only)
Descriptor B
System C (visual-only)
Descriptor C
Candidates A
Candidates B
Candidates C
ref start/end score
Intersection, Union, Score aggregation, etc....
Detection List
ref start/end score


Reference Videos
Query Videos
Similarity Search
for each query segment locates the k closest reference segments
k-NN list
segment1 1NN dist1 2NN dist2
segment2 1NN dist1 2NN dist2
segment3 1NN dist1 2NN dist2
segment4 1NN dist1 2NN dist2
Distance Fusion
$\delta(\text{segment}_i, \text{segment}_j)$
Linear fusion $= \sum w_i \cdot d_i(\ ,\ )$
Copy Localization
for each query video locates chains of segments with temporal consistency
$d_A(\ ,\ )$
$d_B(\ ,\ )$
$d_C(\ ,\ )$
Descriptor A
Descriptor B
Descriptor C
Detection List
ref start/end score


k-NN list $d_{av}$
Query Video
1st NN
2nd NN
3rd NN
Query1-Segm1
Query1-Segm2
Query1-Segm3
Query1-Segm4
Query1-Segm5
Query1-Segm6
Query1-Segm7
Query1-Segm8
Vid07-Segm54 dist
Vid09-Segm13 dist
Vid07-Segm34 dist
Vid09-Segm15 dist
Vid01-Segm88 dist
Vid09-Segm54 dist
Vid01-Segm45 dist
Vid09-Segm19 dist
Vid08-Segm73 dist
Vid02-Segm34 dist
Vid03-Segm54 dist
Vid02-Segm13 dist
Vid01-Segm12 dist
Vid09-Segm17 dist
Vid03-Segm43 dist
Vid01-Segm12 dist
Vid01-Segm68 dist
Vid02-Segm33 dist
Vid09-Segm14 dist
Vid03-Segm65 dist
Vid07-Segm58 dist
Vid07-Segm59 dist
Vid03-Segm20 dist
Vid07-Segm61 dist


$a$
$d(a,b)$
$b$
$p$
$|d(a,p)-d(p,b)|$


EhdRgbAud
EhdGry
Average opt mean F1
Average opt min NDCR


EhdRgbAud
EhdGry
Average mean proc time
Average opt min NDCR