CRIM's Audio Copy Detection System Using Nearest-Neighbor Fingerprints

Vishwa Gupta
COMPUTER RESEARCH INSTITUTE OF MONTREAL

TRECVID-2009
16 November 2009

CRIM's COPY DETECTION TEAM

- Maguelonne Héritier
- Vishwa Gupta
- Langis Gagnon
- Gilles Boulianne
- Samuel Foucher
- Patrick Cardinal

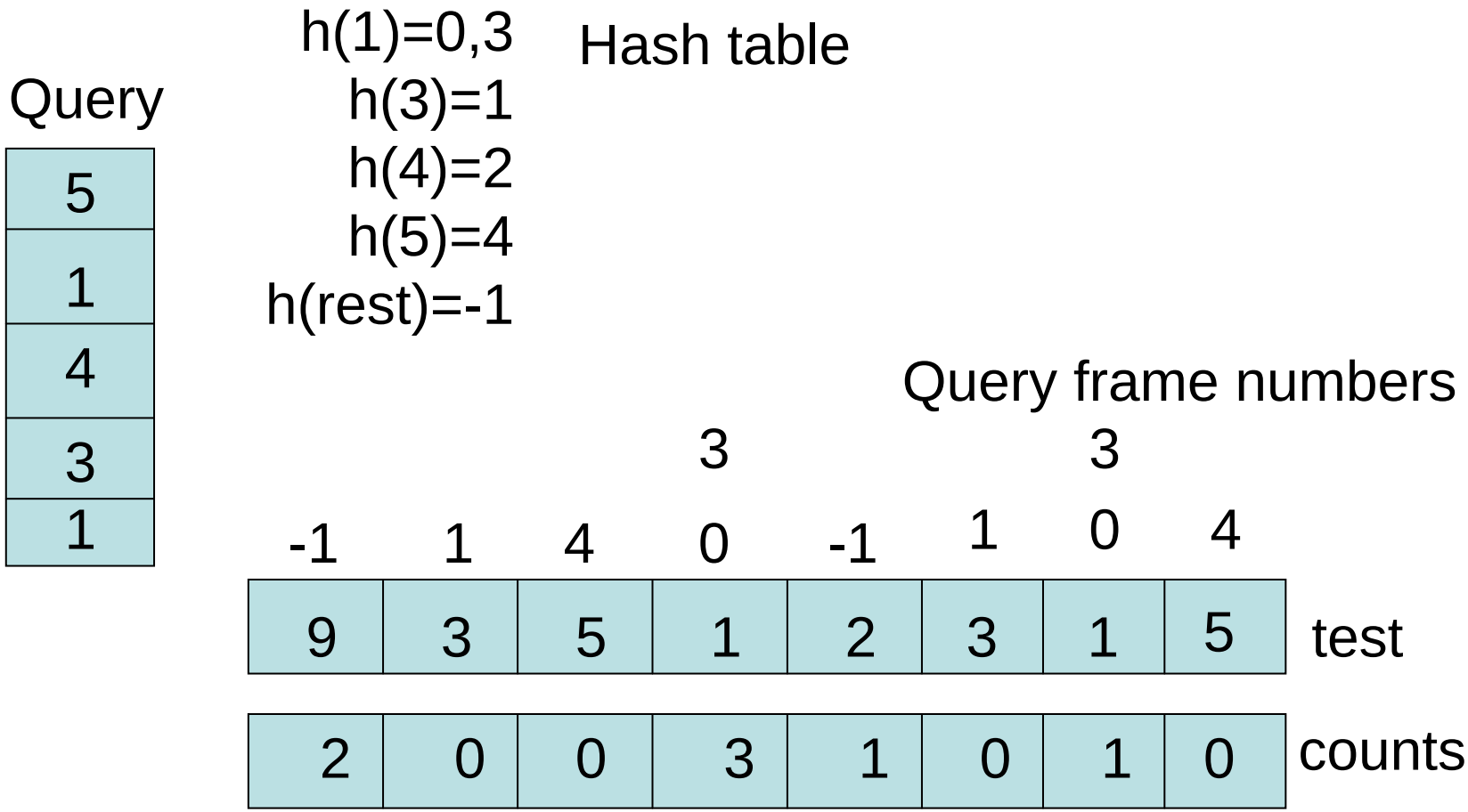
AUDIO FINGERPRINTS

- Energy-Difference Fingerprints
  - 15-bit fingerprints similar to those used by Ahmet Saracoğlu et al in TRECVID 2008
  - Very fast to compute
  - False alarms
- Nearest-Neighbor Fingerprints
  - Maps each test frame to query frame
  - Slow to compute (need GPU to speed-up computing)
  - Accurate
  - Low false-alarm rate

ENERGY-DIFFERENCE FINGERPRINTS

- Low-pass filter to 4 KHz
- Divide into 25 ms windows with 10 ms frame advance
- Each window: pre-emphasis->Hamming window->FFT
- Divide 300-3000 Hz band into 16 bands using Mel scale
- Compute energy in each band using triangular filters
- Use energy-difference in consecutive bands to assign values to 15 bits
- These 15-bits / frame are only used to test exact match between two different frames

Search with Energy-Difference Fingerprints



Query starting with frame 3 has highest count

NEAREST-NEIGHBOR (NN) FINGERPRINTS

- For each frame of test
  - Find the query frame closest to the test frame
  - Use absolute distance between the query and test cepstral features
    - $\Sigma|q(i)-t(i)|$
  - The fingerprint is simply the frame number of the query

Search with Nearest-Neighbor Fingerprints

Query

4
3
2
1
0

4	1	4	0	2	1	3	4	test
1	0	1	3	1	0	1	0	counts

Query starting with frame 3 has highest count

Energy-difference versus Nearest-Neighbor

- Both the fingerprints are consistent
  - Real copy has a higher score than the false segments
- Matching frame counts for false segments vary lot more for Energy-difference than for nearest-neighbor

Energy-difference versus Nearest-Neighbor

Count N	31	35	45	55	75	100
# of segments	738464	354898	133572	74480	16492	1796

For energy-difference fingerprints, segments with matching counts N for the 1400 Trecvid-2008 queries

Count N	11	20	25	30	35	40
# of segments	12147	71	61	22	36	28

For nearest-neighbor fingerprints, segments with matching counts N for the 1400 queries

Results for 2008 audio queries for no false-alarm case

Transform	1	2	3	4	5	6	7
min NDCR energy-diff 1 thresh/transform	.007	.007	.030	.022	.060	.053	.053
1 thresh all transforms -energy-diff fingerprints	.015	.037	.037	.022	.127	.135	.165
Rescore with nearest neighbor fingerprints	.007	0	.007	.007	.037	.03	.03
Search with nearest neighbor fingerprints	.007	0	.015	.015	.022	0	.03

Fusion of Energy-Difference and Nearest-Neighbor Fingerprints

- Combine counts for overlapping segments as follows:
  - Energy-diff counts/sec * 15 + Nearest-neighbor counts
- If Energy-diff and Nearest-Neighbor segments overlap
  - Take segment boundaries from Nearest-Neighbor
- Only combine the highest scoring segment per query

Comparative Results for 2008 queries using one threshold for all transforms (no FA case)

Method	Minimal NDCR	Avg CPU time
Energy-diff fingerprints	0.077	15 sec
Energy diff + NN-based 2nd pass	0.017	20 sec
Nearest neighbor fingerprints	0.016	360 sec
Fused results	0.008	375 sec

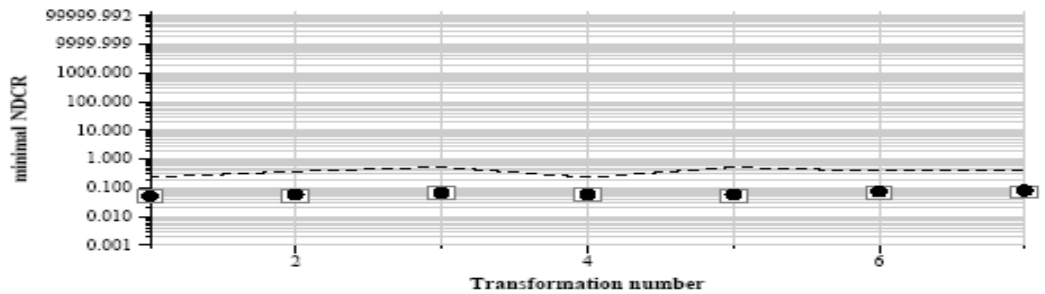
Comparative Results for 2009 queries

Method	Opt min NDCR	Actual min NDCR	Avg CPU time
Energy diff + NN-based 2nd pass	0.065	0.068	20.5 sec
Nearest neighbor fingerprints	0.061	0.066	376 sec
Fused results	0.057	0.070	390 sec

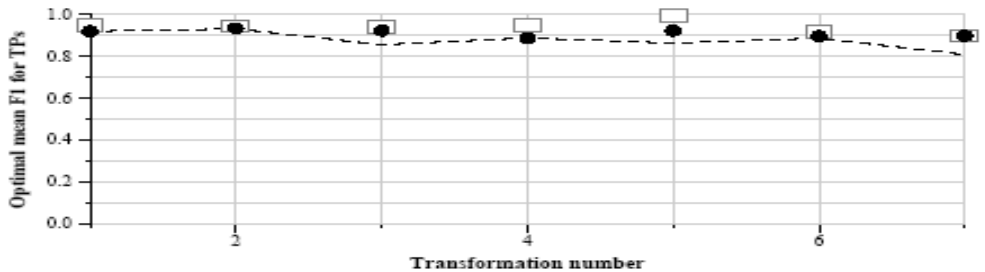
Comparative Results for 2009 queries

TRECVID 2009: copy detection results (no false alarms application profile)

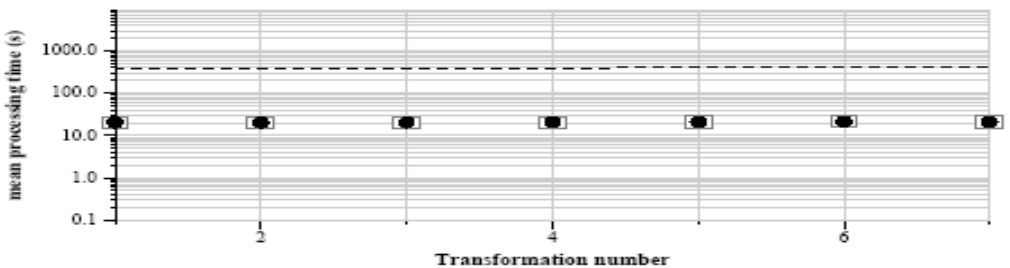
Run name: CRIM.a.nofa.BnNN2pass
Run type: audio-only



Run score (dot) versus median (---) versus best (box) by transformation

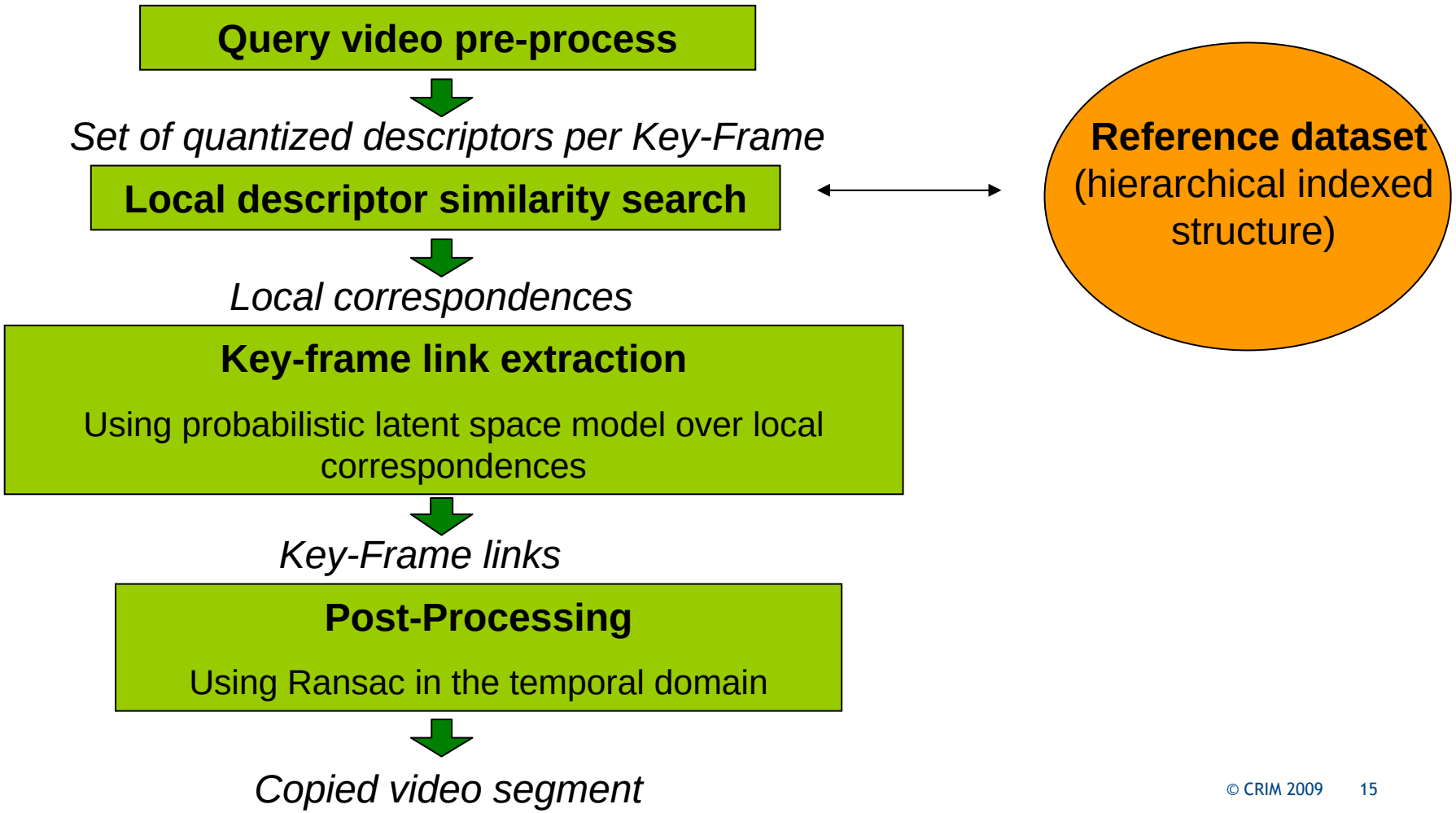


Run score (dot) versus median (---) versus best (box) by transformation



Run score (dot) versus median (---) versus best (box) by transformation

Video copy detection system



Pre-processes for video reference database

Key-frame selection and local descriptor extraction

Set of local descriptors (SIFT) per Key-frame.

Local descriptor quantizing

128 floating point vector to 17 int values

Set of quantized descriptors

Non relevant descriptor filtering

40% of non relevant descriptors are discarded

Hierarchical storage

Hierarchical indexed structure

Pre-processes for query video

Auto-Crop and Video Insert Detection

↓

Key-frames selection

↓ ↓

Key-Frames from background video

Key-Frames from inserted video

Local descriptor extraction (SIFT) and local descriptor quantizing

↓

Set of quantized descriptors for:

- *background Key-frames*
- *flipped background Key-frames*
- *inserted Key-frames*
- *flipped inserted Key-frames*

Video only copy detection results

- Problems when swapping data from the disk(very slow)
- Results are close to the median detection performance

Transform	1	2	3	4	5	6	7
N. queries	134	134	134	134	134	134	134
Miss rate	0.48	0.4	0.61	0.43	0.28	0.7	0.71
FA count	17	15	21	13	14	10	11
Mean F1	0.72	0.73	0.61	0.64	0.68	0.63	0.61
Mean time(s)	1374	796	989	765	780	1123	1136
Opt NDCR B	0.84	0.87	0.85	0.89	0.69	0.83	0.96
Opt NDCR NF	0.84	0.87	0.90	0.89	0.69	0.98	0.96

Video + Audio copy detection

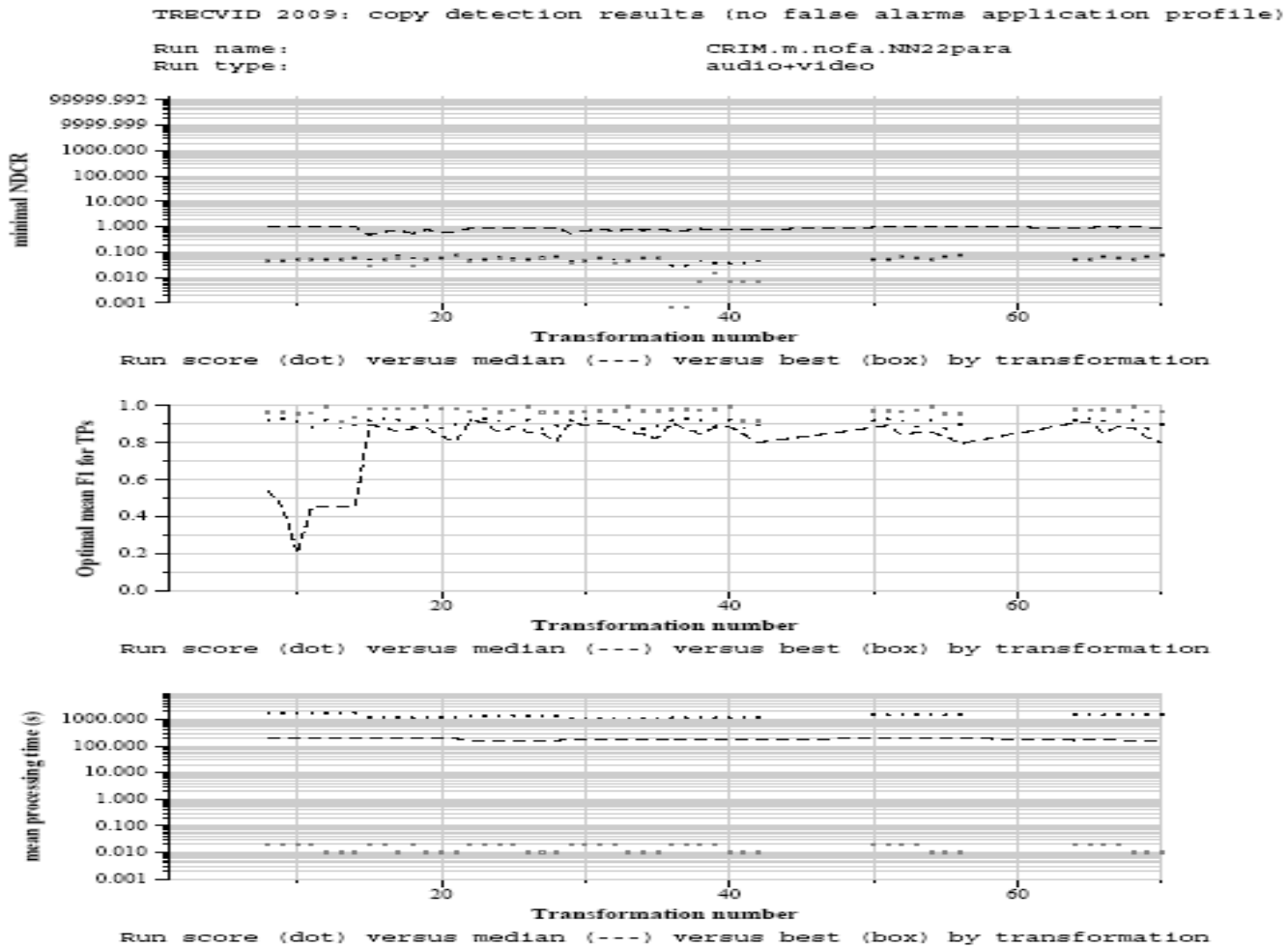
- Combine top segment from audio and video for each query
- Combine scores for overlapping segment as follows:
  - Audio Score + Video Score * N
  - Take segment boundaries from audio search
- If no overlap:
  - Output only highest scoring segment
- Results for 2008 A+V queries (averaged over all transforms):

Weight N	0	1	**2**	3	4
Minimal NDCR	0.017	0.016	**0.014**	0.016	0.017

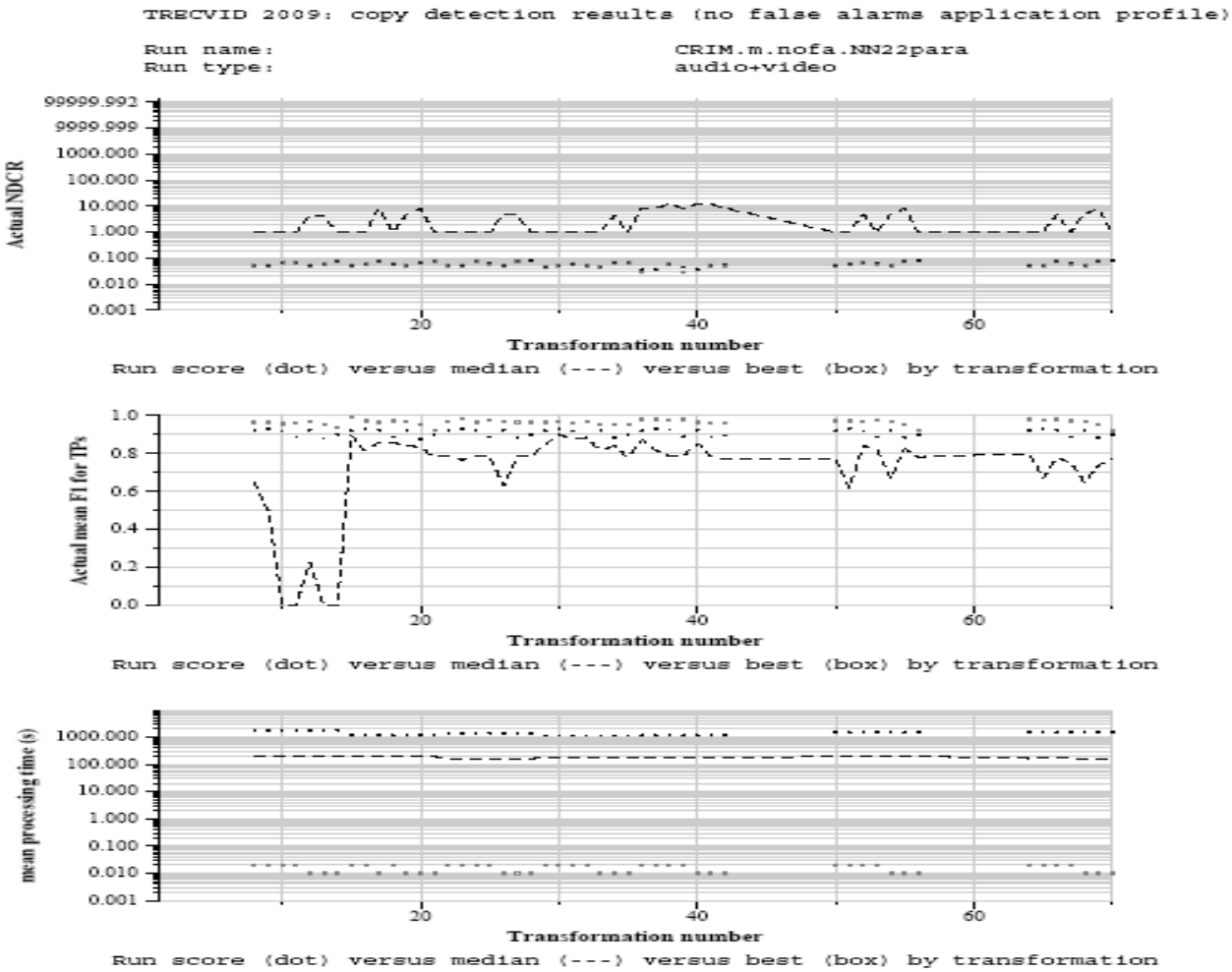
Video + Audio copy detection results for 2009

Method	Opt min NDCR	Actual min NDCR	Avg CPU time (sec)
NN-2pass (no FA)	0.056	1.34	1016
NN-2pass (balanced)	0.056	0.063	1016
NN-search (no FA)	0.055	0.06	1371
Fused energy+NN (balanced)	0.052	0.058	1385

Video + Audio copy detection opt results for 2009



Video + Audio copy detection actual results for 2009



CONCLUSIONS

- Nearest-Neighbor based audio fingerprints give the lowest min NDCR
- Nearest-Neighbor fingerprint computing is speeded-up by graphics processing unit.
- Combined with Energy-difference fingerprints, they give the fastest computing with the lowest min NDCR.
- When combined with video copy detection with median performance, they give the lowest NDCR for A+V copy detection for most of the transforms.