TRECVID 2011
Content Based Copy Detection
task overview

Wessel Kraaij
TNO, Radboud University Nijmegen

George Awad
NIST

Background

- Copy detection is applied in several real-word tasks:
  - television advertisement monitoring
  - detection of copyright infringement
  - detection of known (illegal) content
- Initial framework developed by NoE MUSCLE/INRIA
- Evaluation procedure re-defined at TV08, consolidated at TV09 (FA/Miss rate; actual vs optimal NDCR), renewed in TV10 (used internet videos (IACC) of much shorter videos with variable frame rates, Camcorder feature back, just AV runs, and adjusted 'balanced' profile).
- TV11:
  - Same dataset as in TV10 to be able to compare performance.

CBCD task overview

- Goal:
  - Build a benchmark collection for video copy detection methods
- Task:
  - Given a set of reference (test) video collection and a set of 11256 queries,
  - determine for each query if it contains a copy, with possible transformations, of video from the reference collection,
  - and if so, from where in the reference collection the copy comes
- As in 2010, only one task type:
  - Copy detection of video + audio (11256) queries
- At least 2 runs are required representing two application profiles (“no false alarms”, “balanced”).

Datasets and queries

- Dataset:
  - Reference video collection:
    - Testing data: IACC.1.A (~8000 videos, 200 hr, < 3.5min)
    - Development data : IACC.1.tv10.training(~3200 videos, 200 hr, 3.6 - 4.1min)
  - Non-reference video collection :
    - Internet Archives videos (~12480 videos, ~4000 hr, 10 – 30min)
- Queries: (Developed by INRIA-IMEDIA software run at NIST)
  - Types:
    - Type 1: composed of a reference video only. (1/3)
    - Type 2: composed of a reference video embedded in a non-reference video. (1/3)

Copies {

    - Type 3: composed of a non-reference video only. (1/3)
  - 201 total original queries. 67 queries for each type.
  - Type 1 & 2 durations (average of 27 sec)
  - Type 3 durations (average of 89 sec)

Datasets and queries

- After creating the queries, each was transformed by NIST
  - 8 video transformations using tools developed by Laurent Joyeaux (independent agent at INRIA)
  - 7 audio transformations using tools developed by Dan Ellis (Columbia University)
- Yielding…
  - 8 * 201 = 1608 video queries
  - 7 * 201 = 1407 audio queries
  - 8 * 7 * 201 = 11256 audio+video queries
- 8 groups of duplicate videos were found:
  - Systems were asked to submit all found duplicate result items.
  - Before evaluation, runs were filtered by removing duplicate result items that doesn't match the G.T reference video.

Some important task details/ assumptions

- Detection systems submit a run threshold.
- Systems are asked to output a list of possible copies (each associated with a desicion score).
- The run threshold is used to determine the asserted copies.
- A query can yield just one true positive
- A query can give rise to many false alarms
- Consequence:
  - Type I error modeled as *false alarm rate*
  - Type II error modeled as *Pmiss*

Video transformations

- 8 Transformations were selected:
  - Simulated camcording (T1) – by perspective transform, automatic gain control, and blurring effects.
  - Picture in picture (T2)
  - Insertions of pattern (T3)
  - Strong re-encoding (T4)
  - Change of gamma (T5)
  - Decrease in quality (T6) - by introducing 3 randomly selected combination of *Blur, Gamma, Frame dropping, Contrast, Compression, Ratio, White noise*
  - Post production (T8) – by introducing 3 randomly selected combination of *Crop, Shift, Contrast, Text insertion, Vertical mirroring, Insertion of pattern, Picture in picture,*
  - Combination of 3 randomly selected transformations (T10) chosen from T2-T5, T6 *and* T8.

Audio transformations

- T1: nothing
- T2: mp3 compression
- T3: mp3 compression and multiband companding
- T4: bandwidth limit and single-band companding
- T5: mix with speech
- T6: mix with speech, then multiband compress
- T7: bandpass filter, mix with speech, compress

Evaluation metrics

Three main metrics were adopted:

1. Normalized Detection Cost Rate (NDCR)
   - measures error rates/probabilities on the test set:
     - Pmiss (probability of a missed copy)
     - Rfa (false alarm rate)
   - combines them using assumptions about two possible realistic scenarios:

     1 - No False Alarm profile:
     - *Copy target rate (Rtarget)* = 0.005/hr
     - *Cost of a miss (CMiss)* = 1
     - *Cost of a false alarm (CFA)* = 1000

     2 – Balanced profile:
     - *Copy target rate (Rtarget) = 0.005/hr*
     - *Cost of a miss (CMiss) = 1*
     - *Cost of a false alarm (CFA) = 1*
2. F_1 (how accurately the copy is located, harmonic mean of P and R)
3. Mean processing time per query

[Kraaij, Over, Fiscus, Joly,2009] Final CBCD Evaluation plan TRECVID 2009 v1.3

Evaluation metrics (2)

General rules:

- No two query result items for a given video can overlap.
- For multiple result items per query, one mapping of submitted extents to ref extents is determined based on a combination of F1-score and the decision score (using the Hungarian solution to the Bipartite Graph matching problem).
- The reference data has been found if and only if: the asserted test video ID is correct AND asserted copy and ref. video overlap.

Actual vs. Optimal

- Actual NDCR:
  - Each transformation is evaluated on the single submiitted threshold value. This models the real situation where systems do not know the transformation in advance
- Optimal NDCR:
  - The minimal NDCR is computed by sweeping through the result list, for each combination of transformations
  - This way, teams can see where there is still room for improvement

22 Participants (finishers)

AT&T Labs - Research	CCD	INS	---	---	---	---
Beijing University of Posts and Telecom.-MCPRL	CCD	INS	KIS	---	SED	SIN
Brno University of Technology	CCD	---	---	***	SED	SIN
CRIM - Vision & Imaging team	CCD	---	---	---	SED	---
France Telecom Orange Labs (Beijing)	CCD	---	---	---	---	SIN
Osaka Prefecture University	CCD	***	---	---	---	---
INRIA-LEAR	CCD	---	***	MED	SED	***
INRIA-TEXMEX	CCD	***	---	---	---	---
Istanbul Technical University	CCD	---	---	---	---	---
University of Kaiserslautern	CCD	---	---	---	---	SIN
KDDILabs	CCD	***	---	---	---	---
NTT Communication Science Laboratories-CSL	CCD	---	---	---	---	---
Peking University-IDM	CCD	---	---	---	SED	---
PRISMA-University of Chile	CCD	---	---	---	---	---
RMIT University School of CS&IT	CCD	---	---	---	---	---
Sun Yat-sen University – GITL	CCD	---	---	---	SED	---
Telephonica Research	CCD	---	---	---	---	---
Tokushima University	CCD	INS	---	---	---	---
University of Queensland	CCD	---	---	---	---	SIN
USC Viterbi School of Engineering	CCD	***	---	---	---	***
Xi'an Jiaotong University	CCD	***	***	***	SED	***
Zhejiang University	CCD	***	---	---	SED	---

--- : group didn't participate

** : group applied but didn't submit

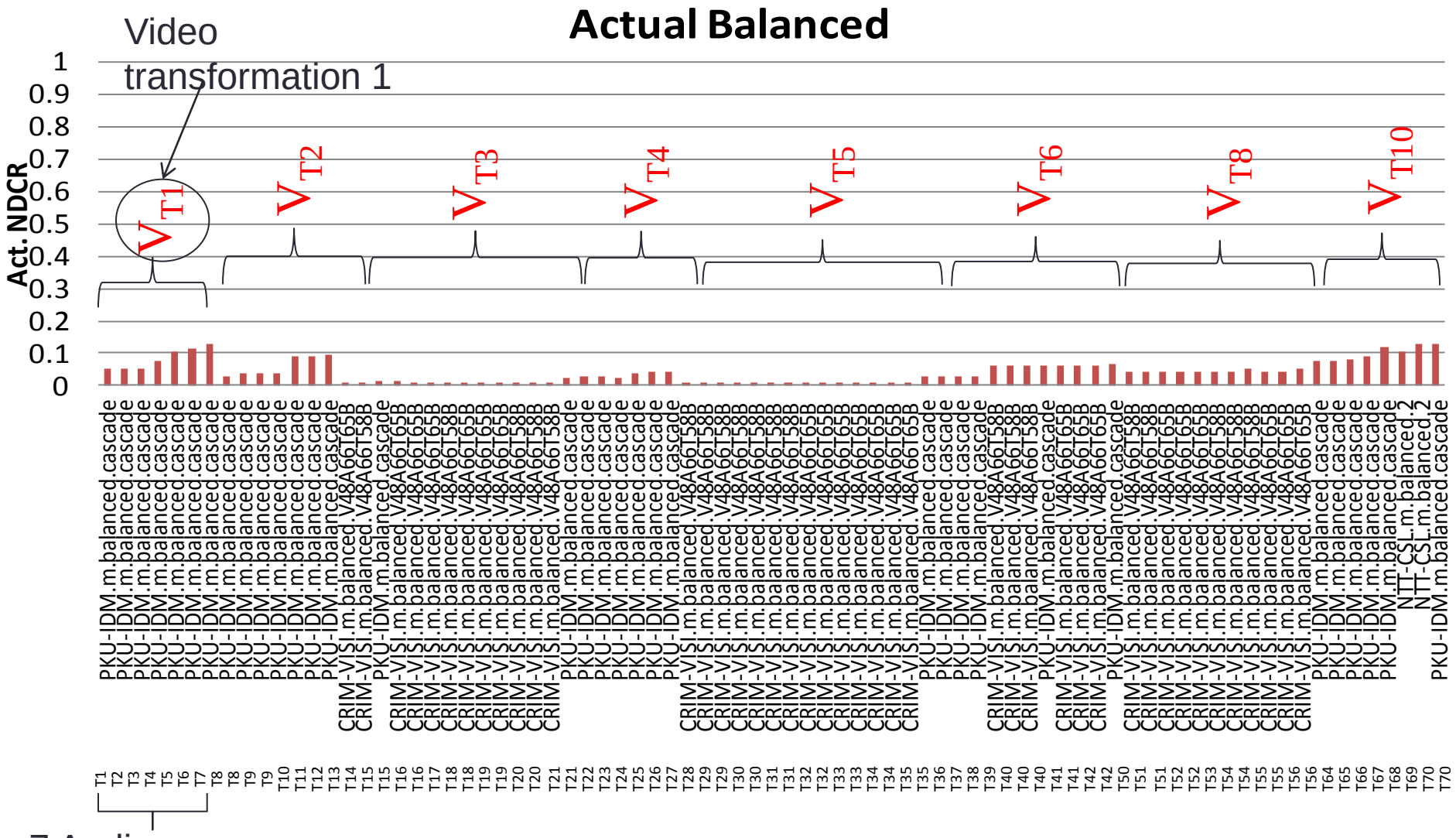
Submission types and counts

Run type	2008	2009	2010	2011
V (video only)	48	53	-	-
A (audio only)	1	12	-	-
M (video + audio)	6	42	78	73
Total runs	55	107	78	73

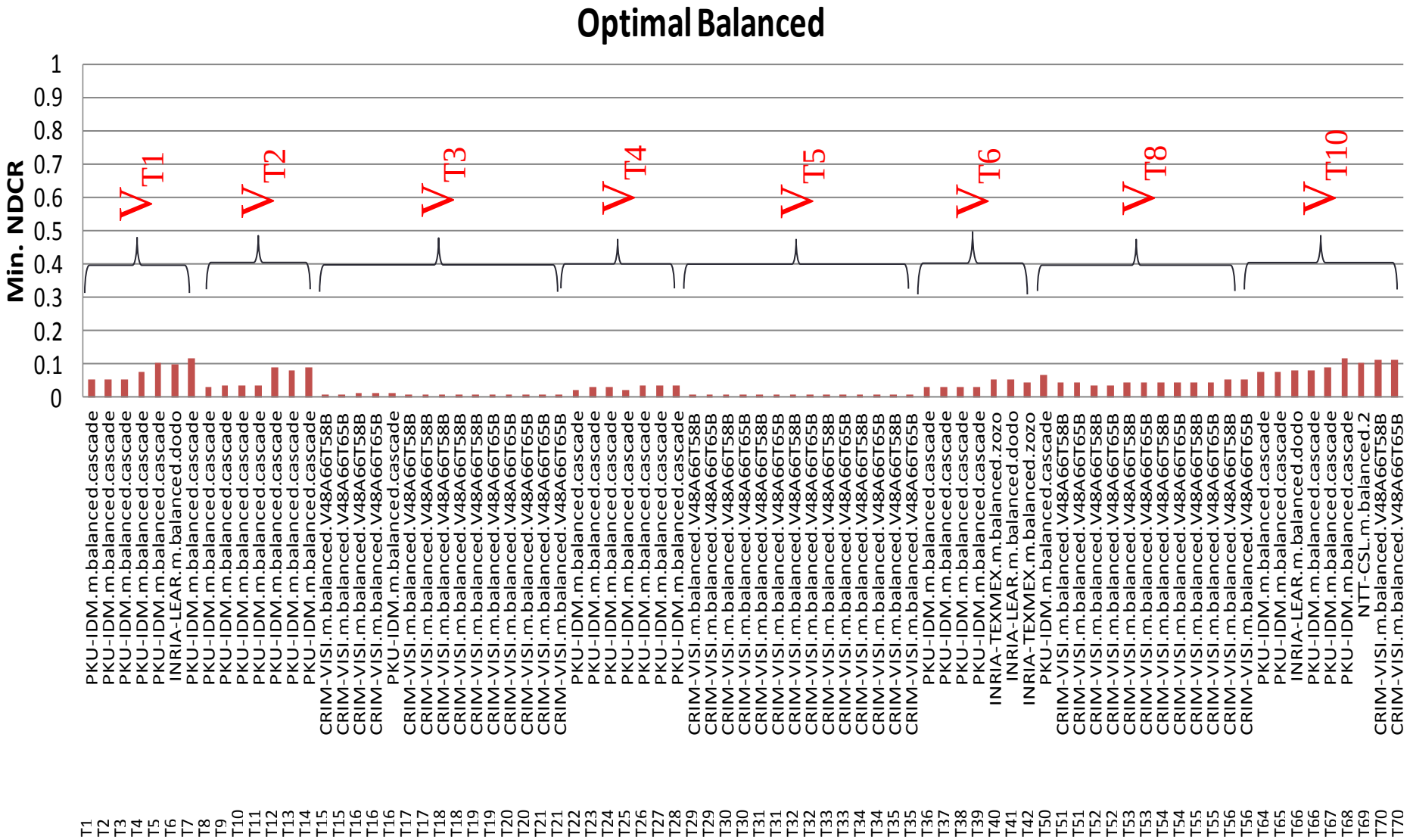
2010		2011	
Type M (Balanced)	Type M (NoFa)	Type M (Balanced)	Type M (NoFa)
41	37	41	32

Balanced number of submission types

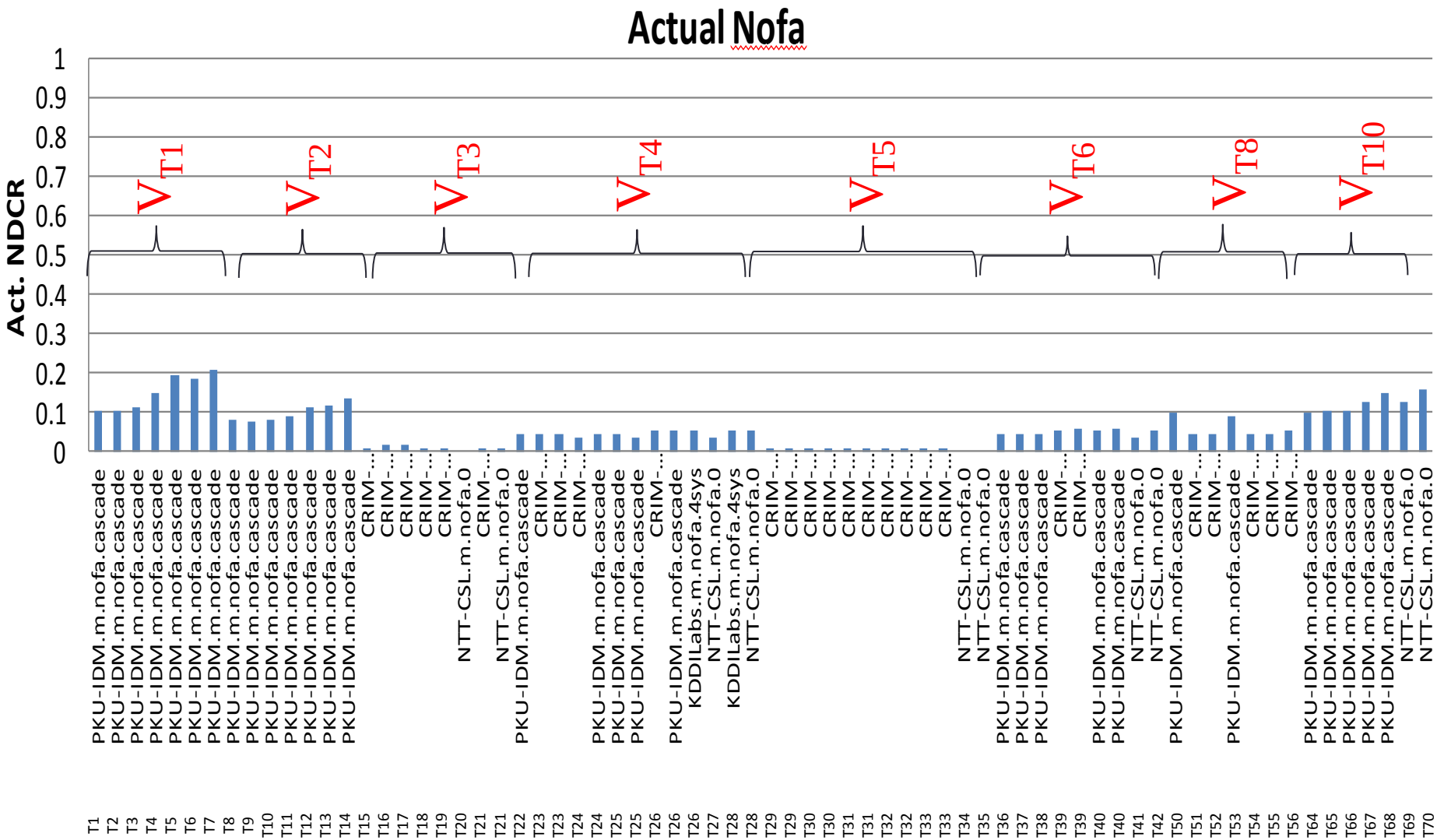
Top “video + audio” runs



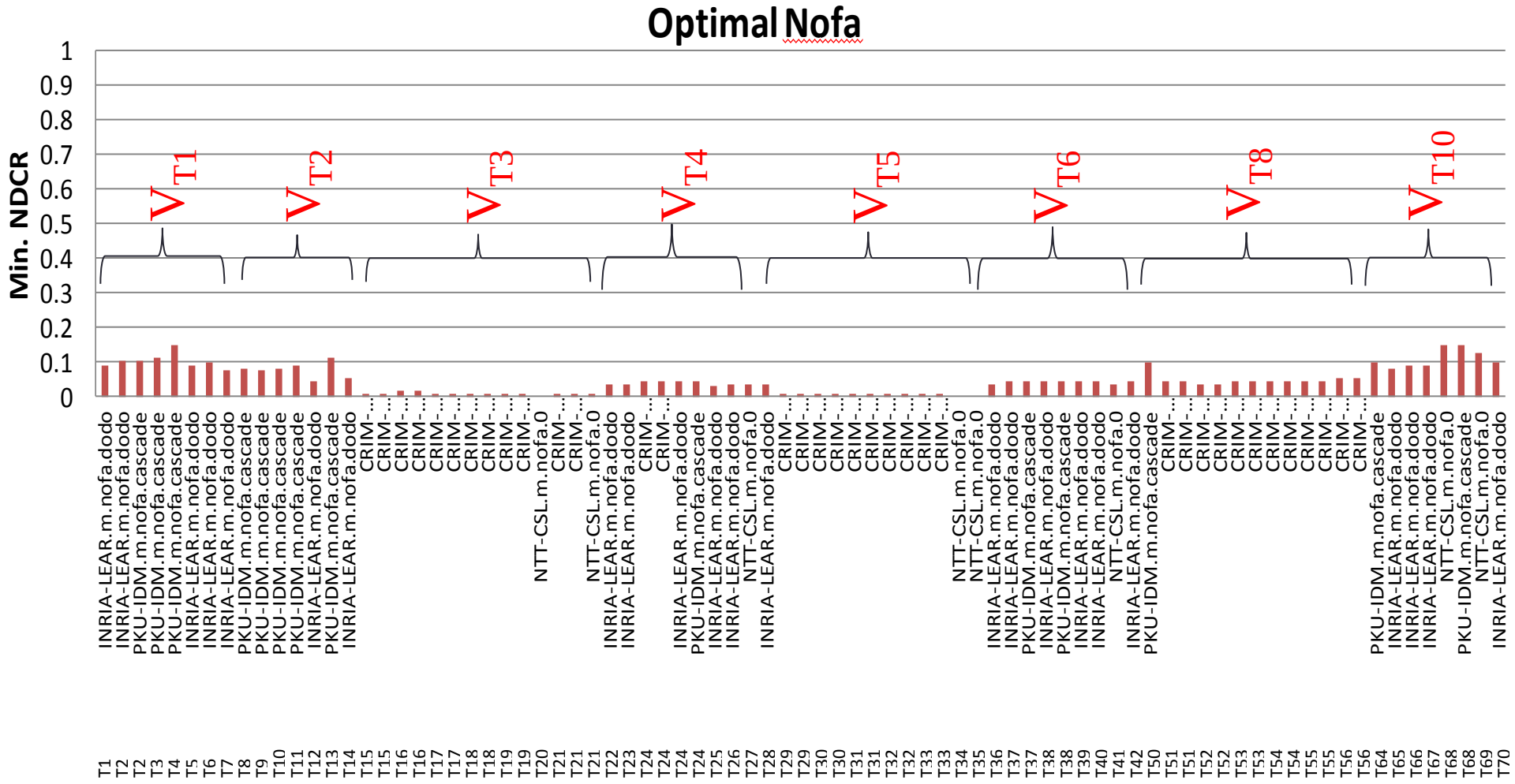
Top “video + audio” runs



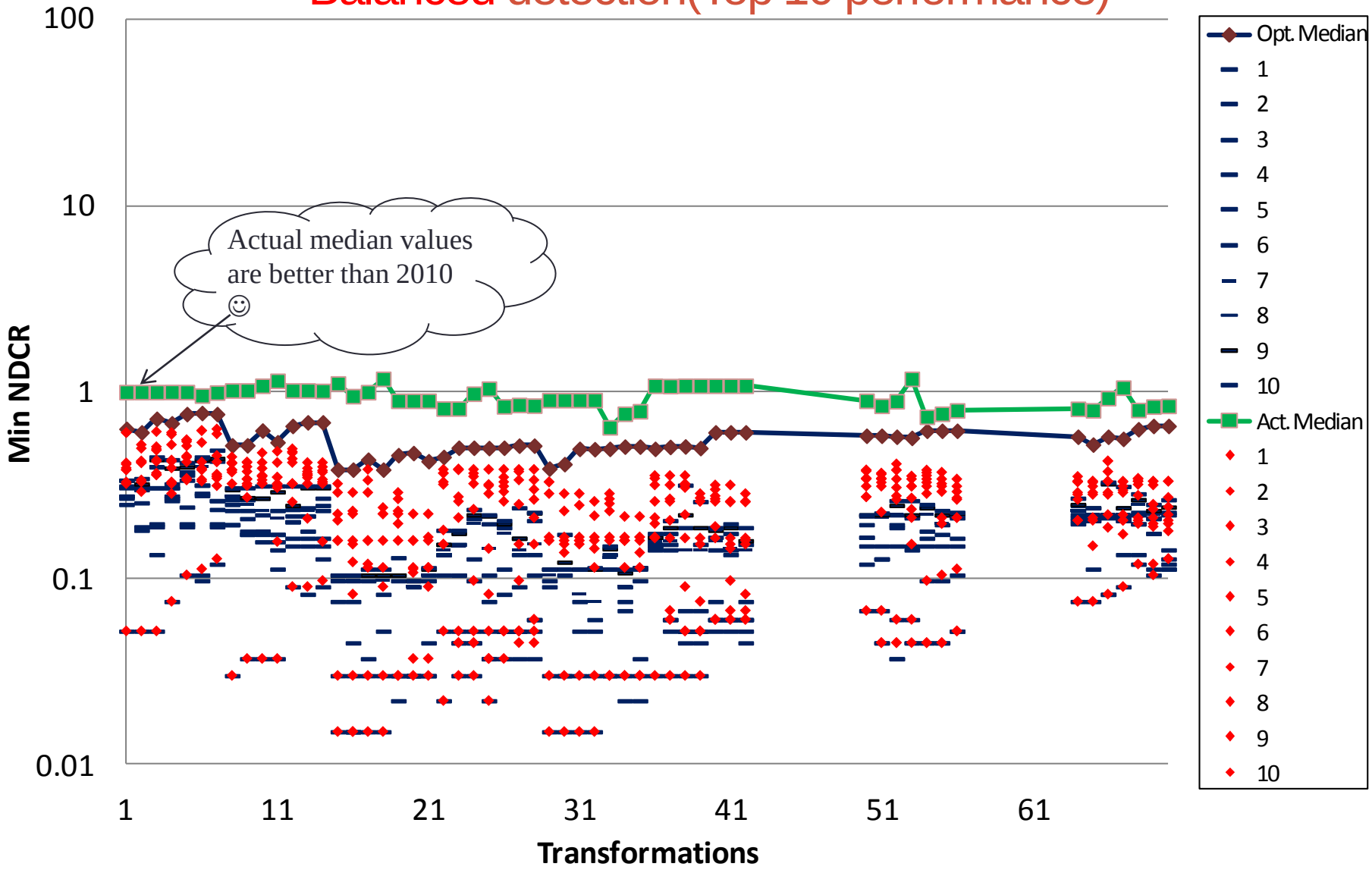
Top “video+audio” runs



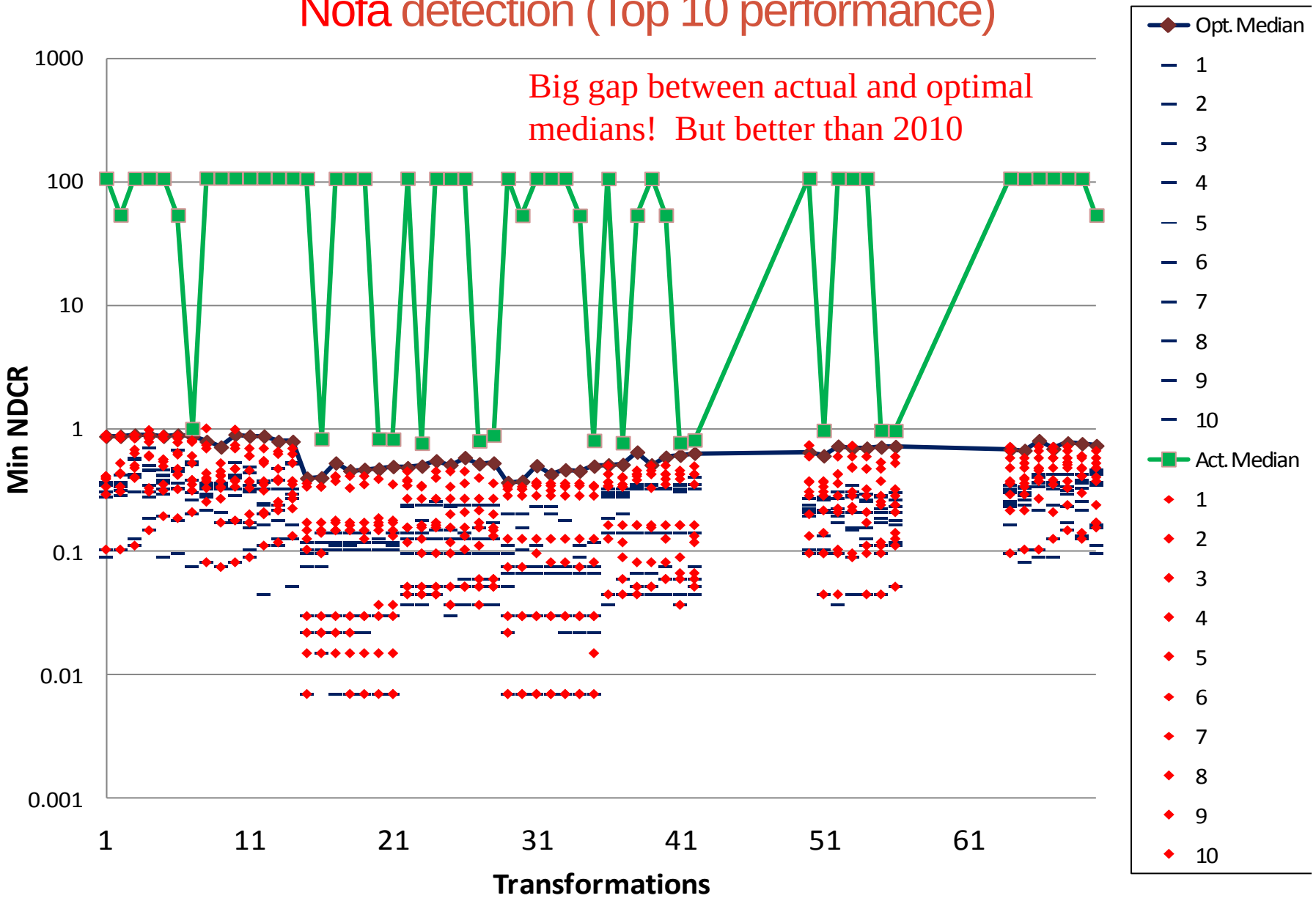
Top “video+audio” runs



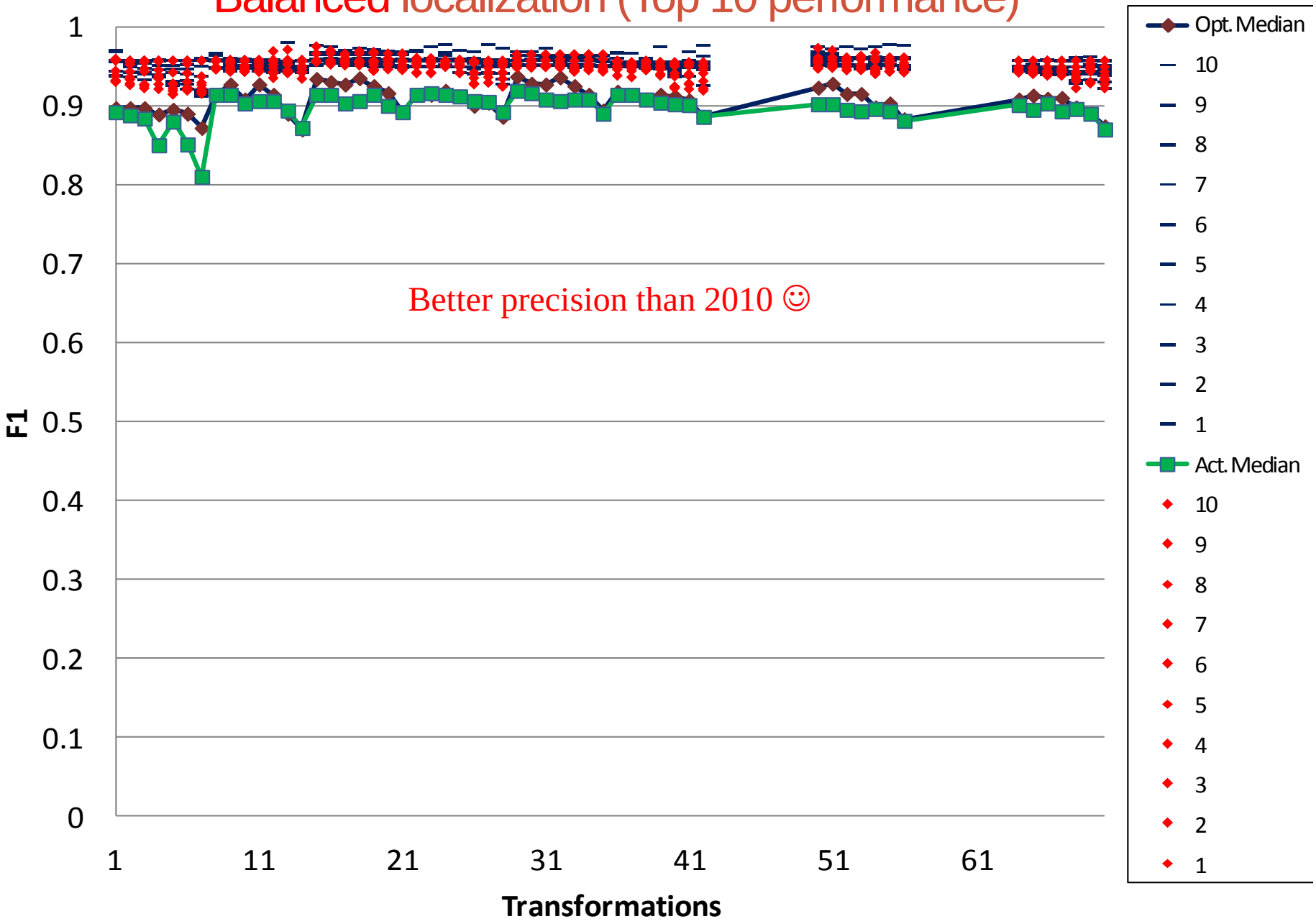
Balanced detection(Top 10 performance)

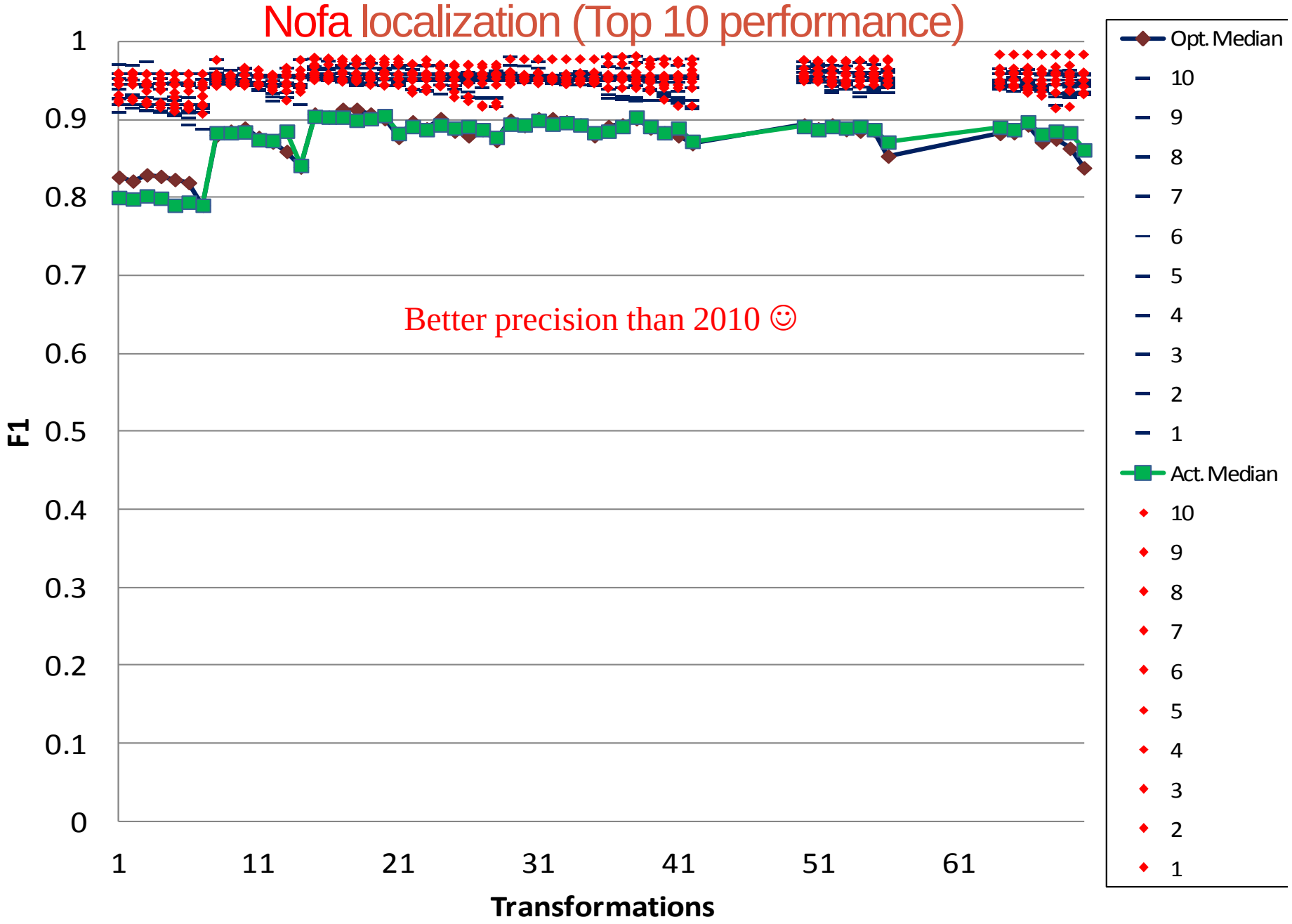


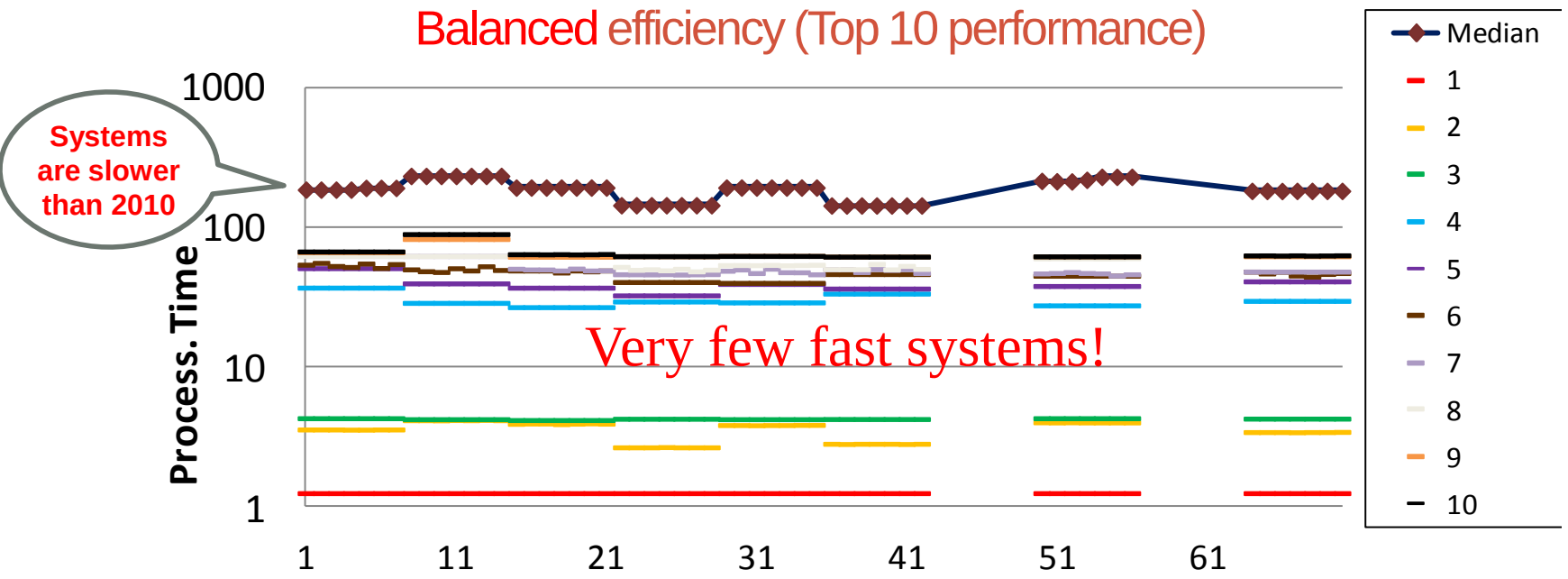
Nofa detection (Top 10 performance)

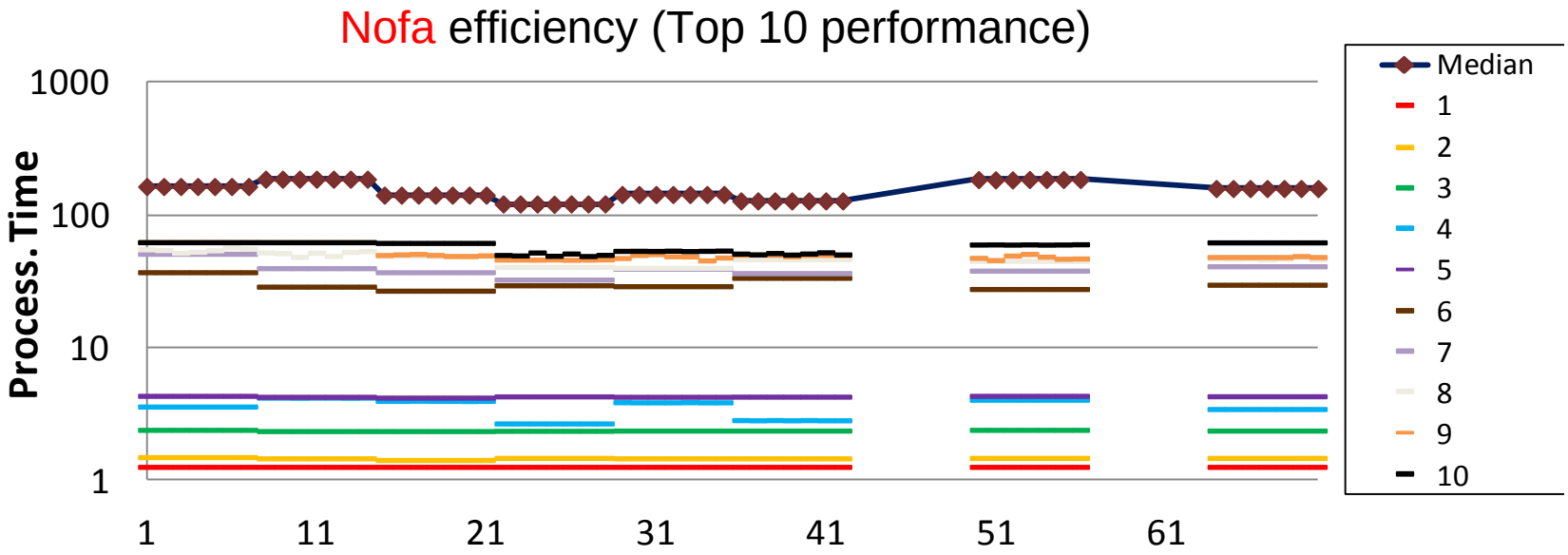


Balanced localization (Top 10 performance)

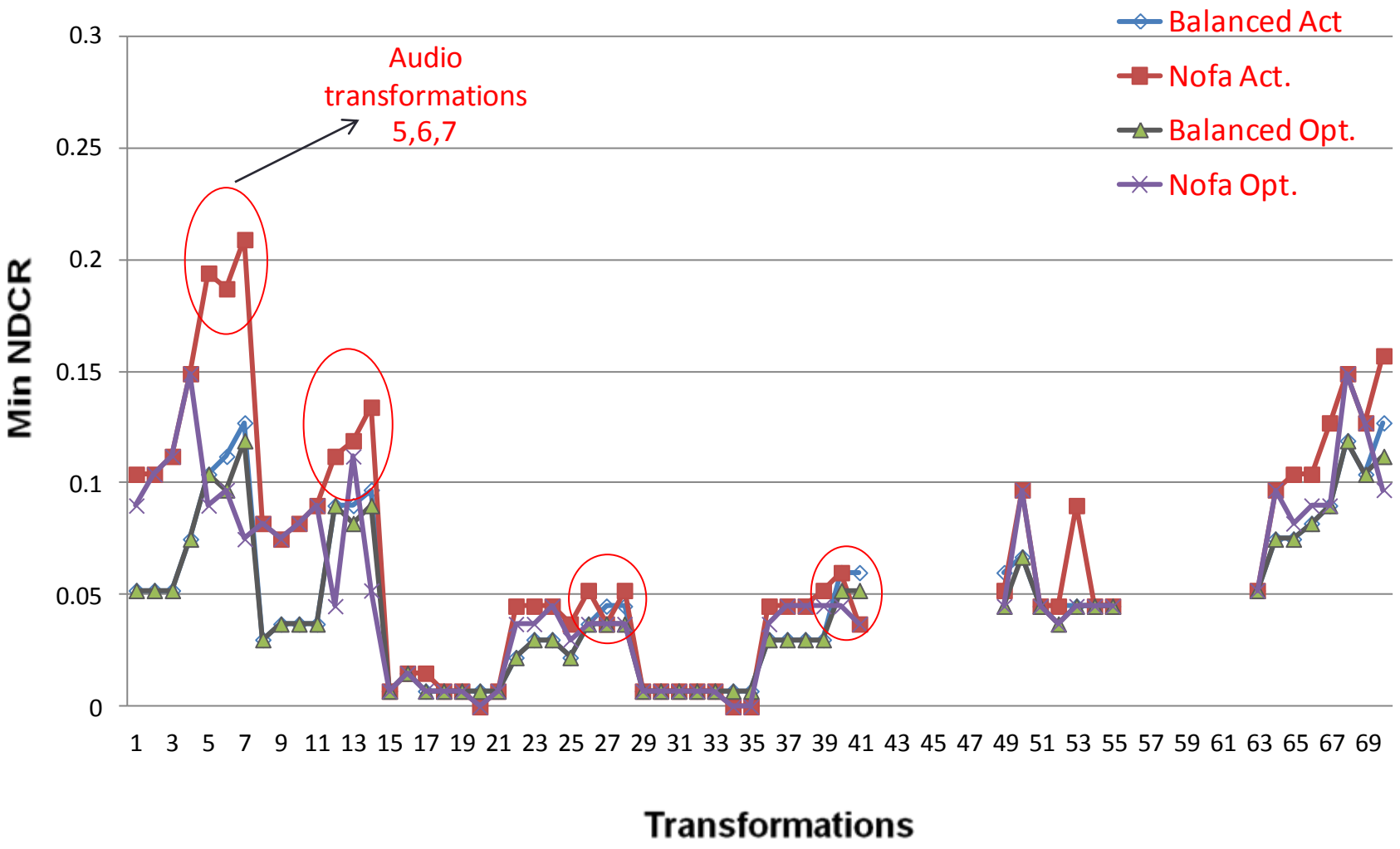



Nofa localization (Top 10 performance)
F1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
11
21
31
41
51
61
Transformations
Better precision than 2010 ☺
Opt. Median
10
9
8
7
6
5
4
3
2
1
Act. Median
10
9
8
7
6
5
4
3
2
1


Balanced efficiency (Top 10 performance)
Systems are slower than 2010
Very few fast systems!
Process. Time
1000
100
10
1
1
11
21
31
41
51
61
Median
1
2
3
4
5
6
7
8
9
10

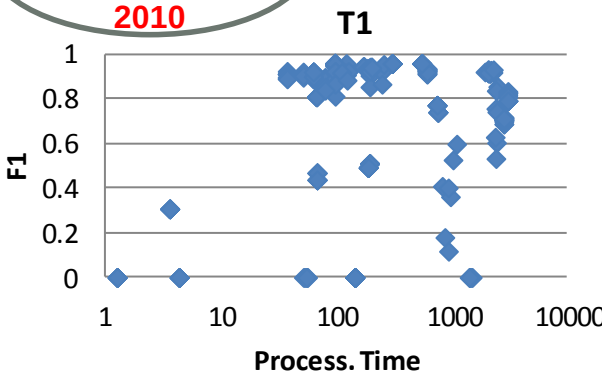

Nofa efficiency (Top 10 performance)
Process. Time
1000
100
10
1
1
11
21
31
41
51
61
Median
1
2
3
4
5
6
7
8
9
10

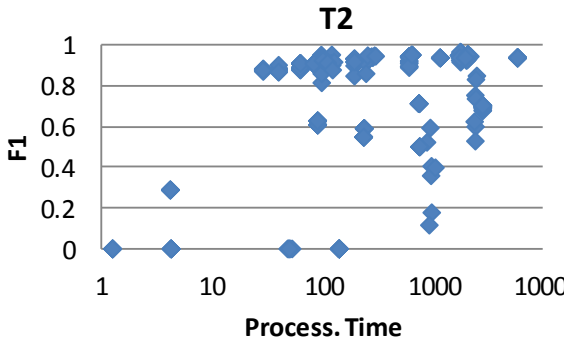
Comparing best runs (detection)

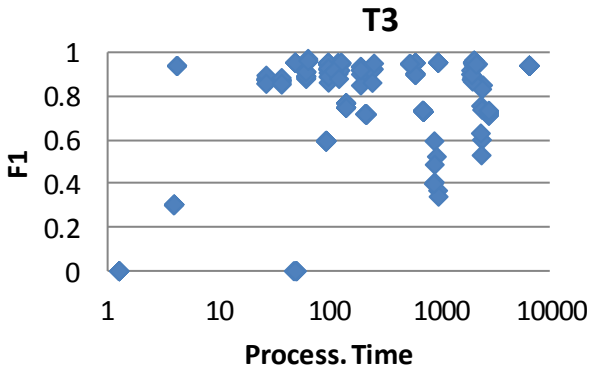


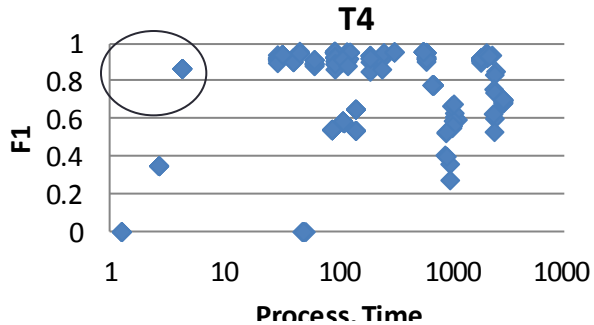
Actual Balanced runs by video transformations (across all audio transformations)

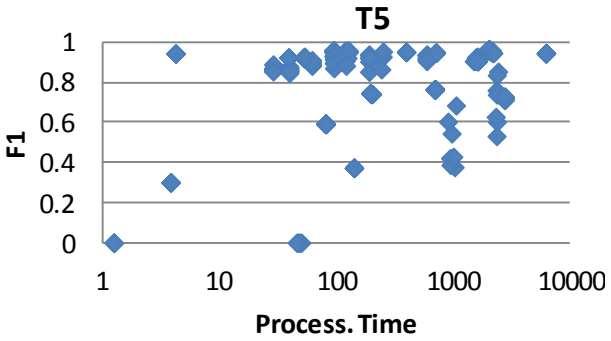
On average, systems are slower than in 2010

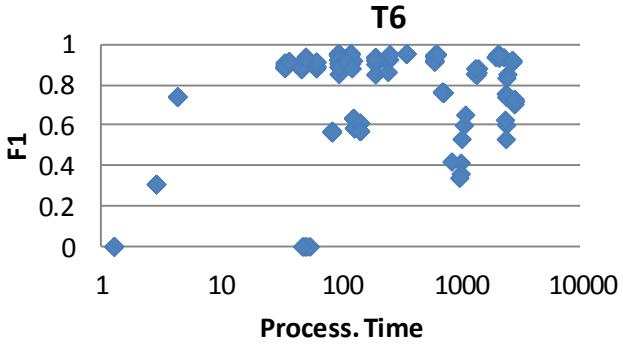


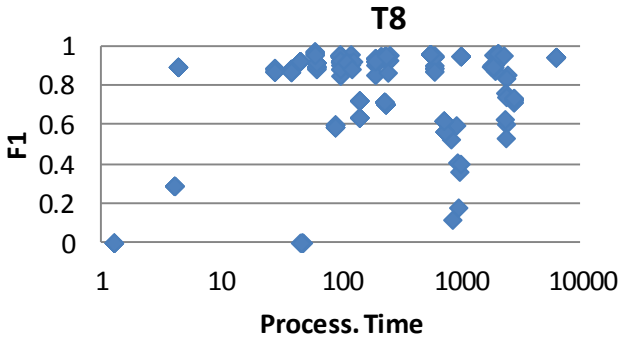


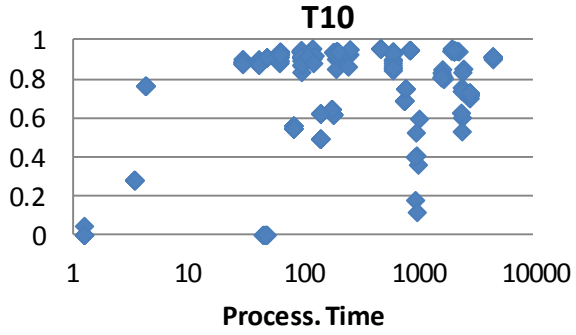






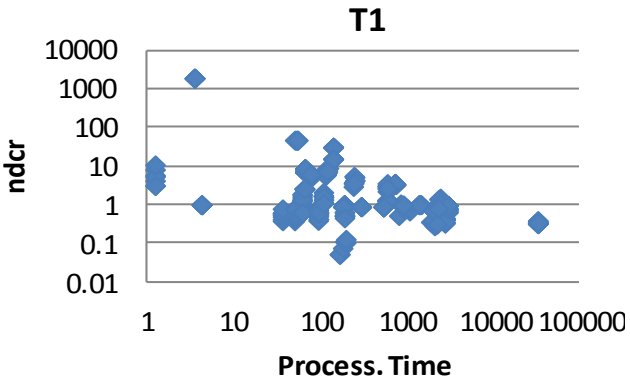


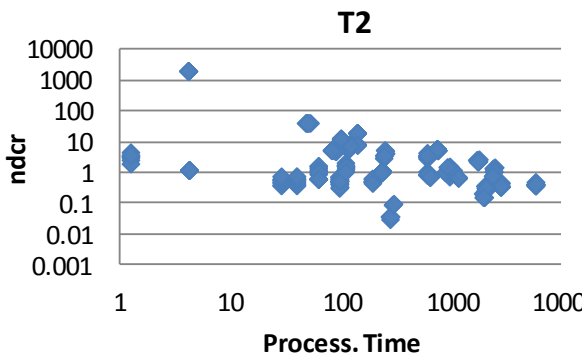


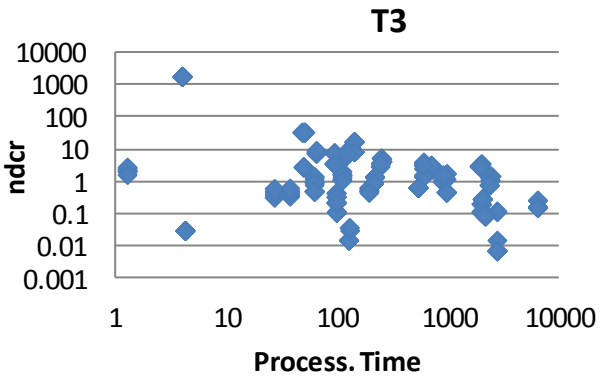


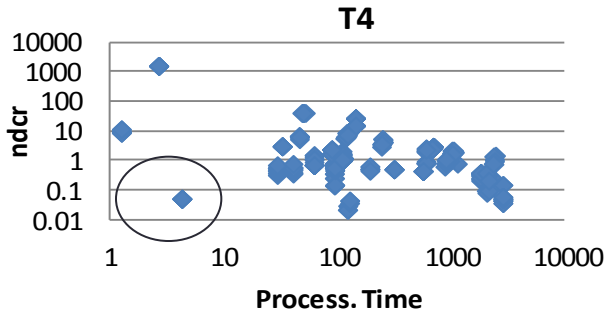
Increasing proc. time did not significantly enhance localization. Very few systems achieved high localization in small proc. time.

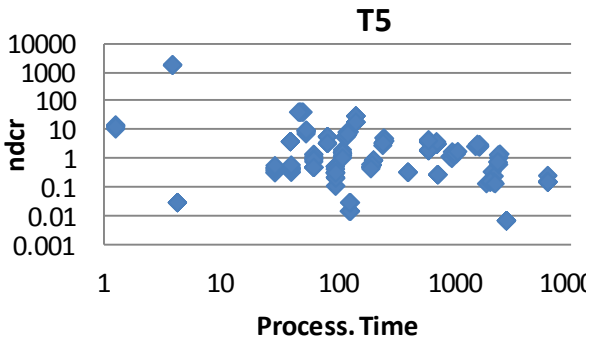
Actual Balanced runs by video transformations (across all audio transformations)

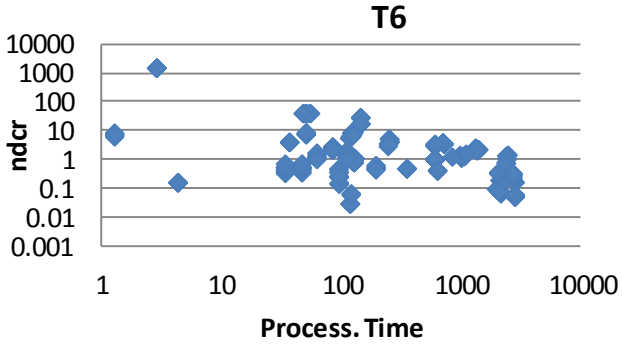


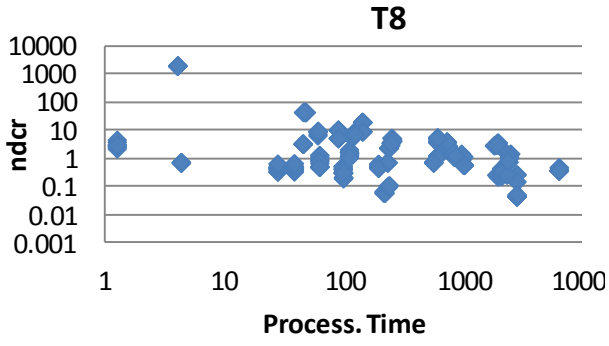


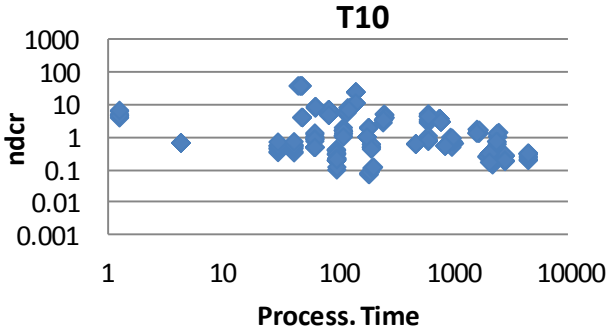






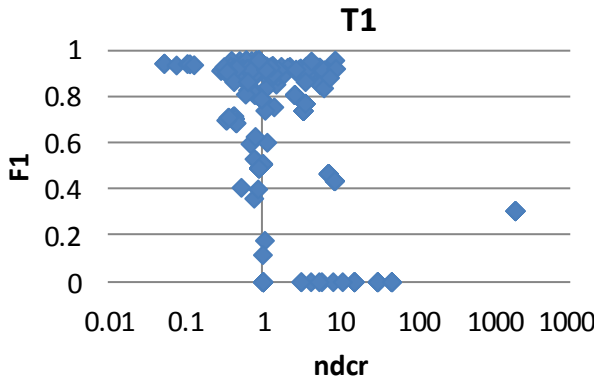


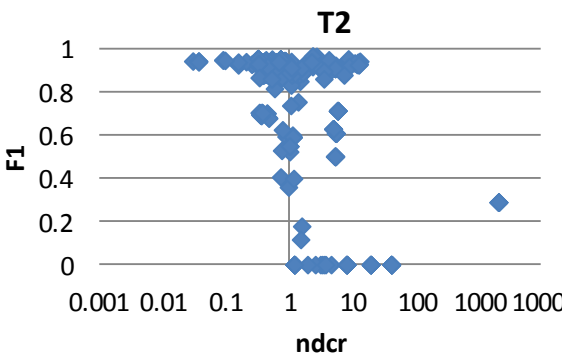


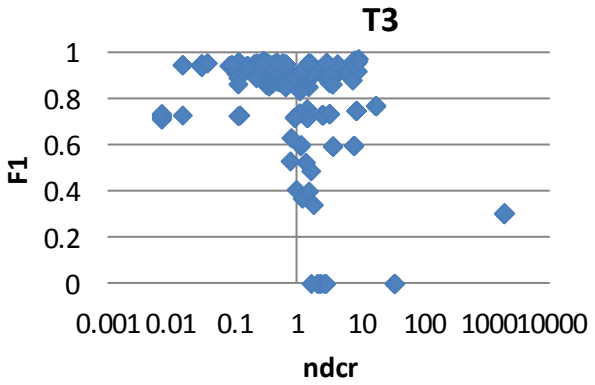


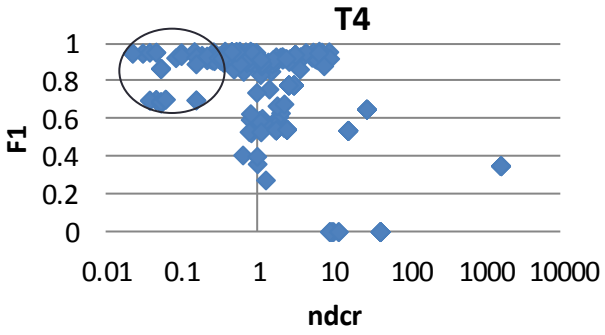
In general, more processing time does not significantly improve detection. Very few strong & fast systems!

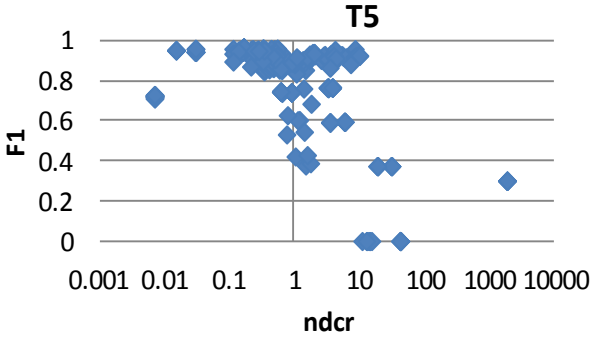
Actual Balanced runs by video transformations (across all audio transformations)

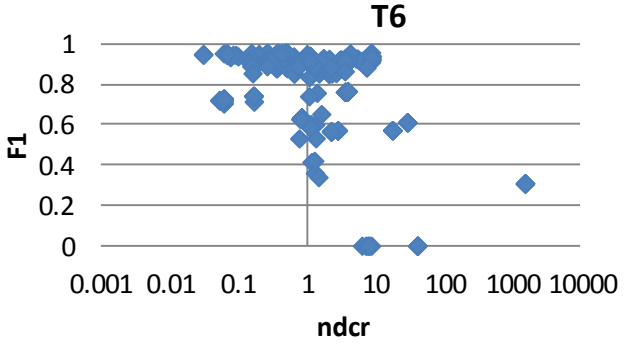


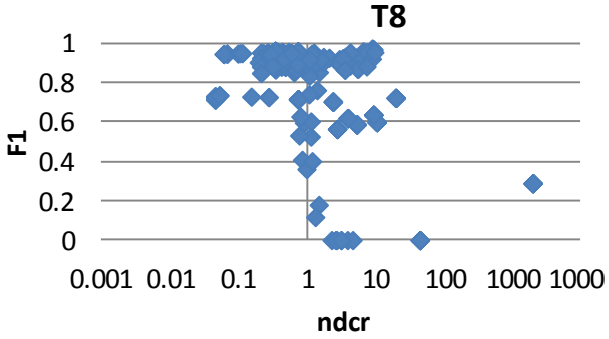


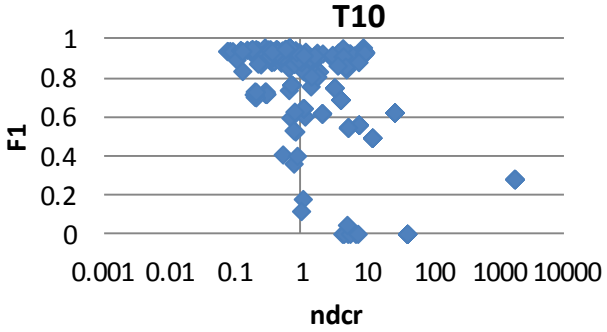






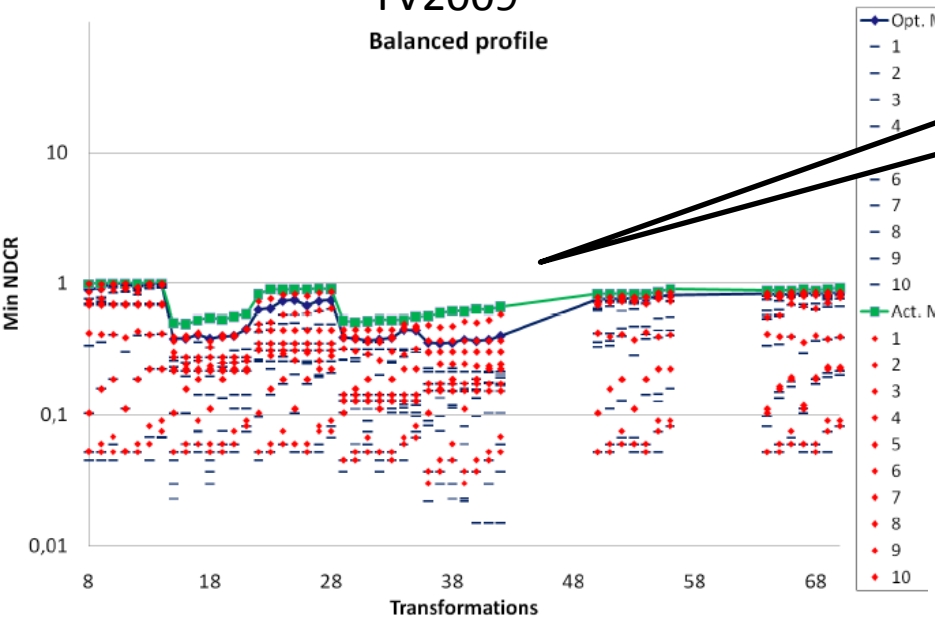






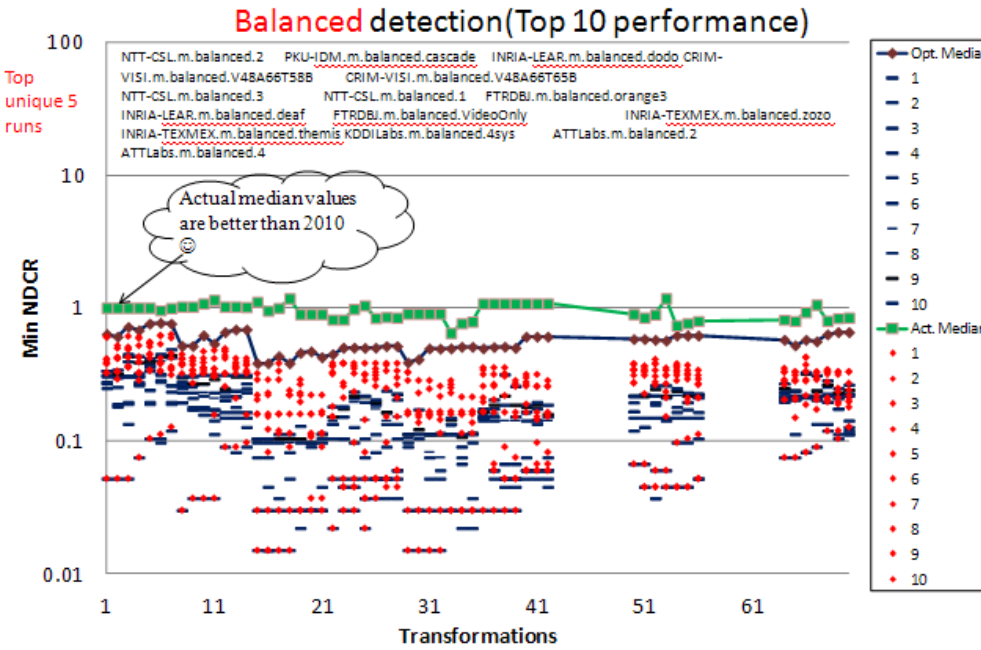
Most of the systems that are good in separating copies from non-copies (low NDCR) are also good in localization.

TV2009

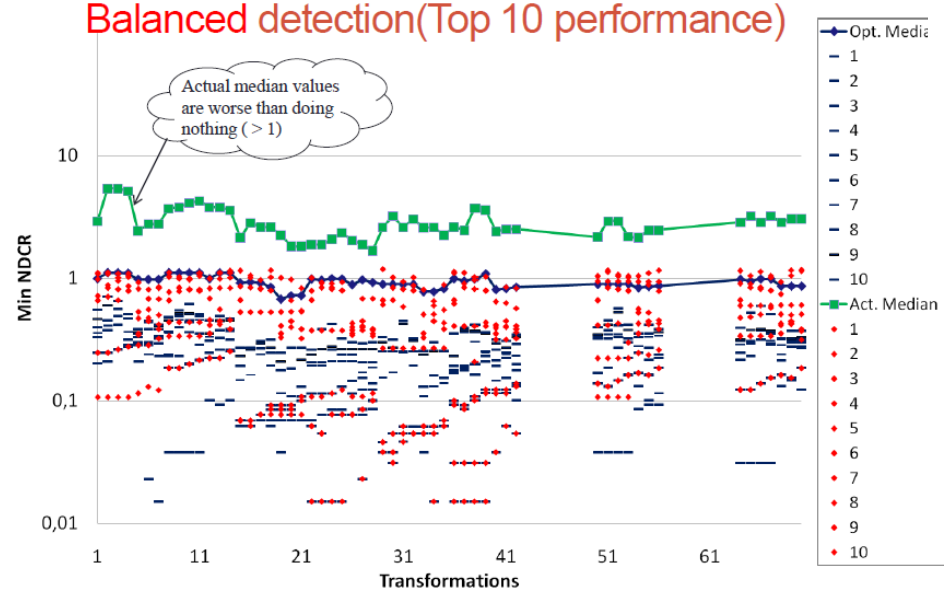


Caution!... different dataset

TV2011

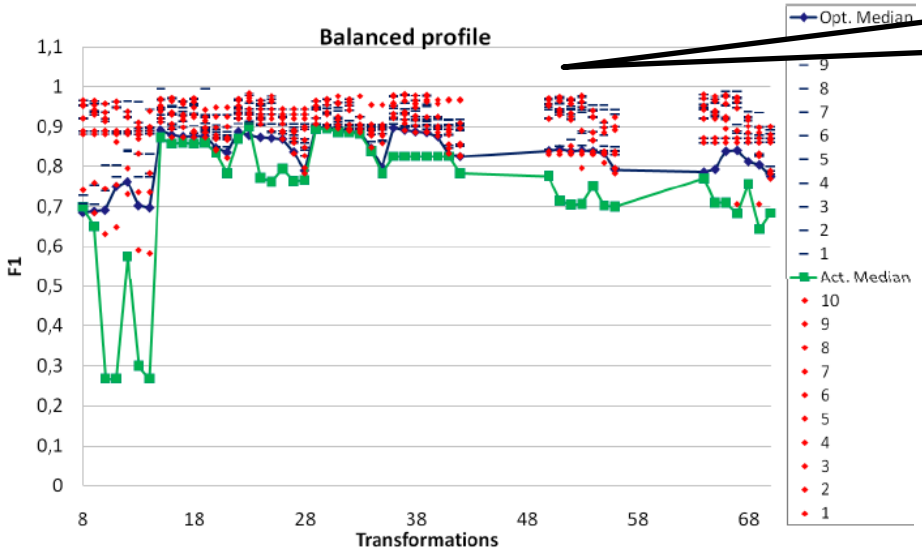


TV2010



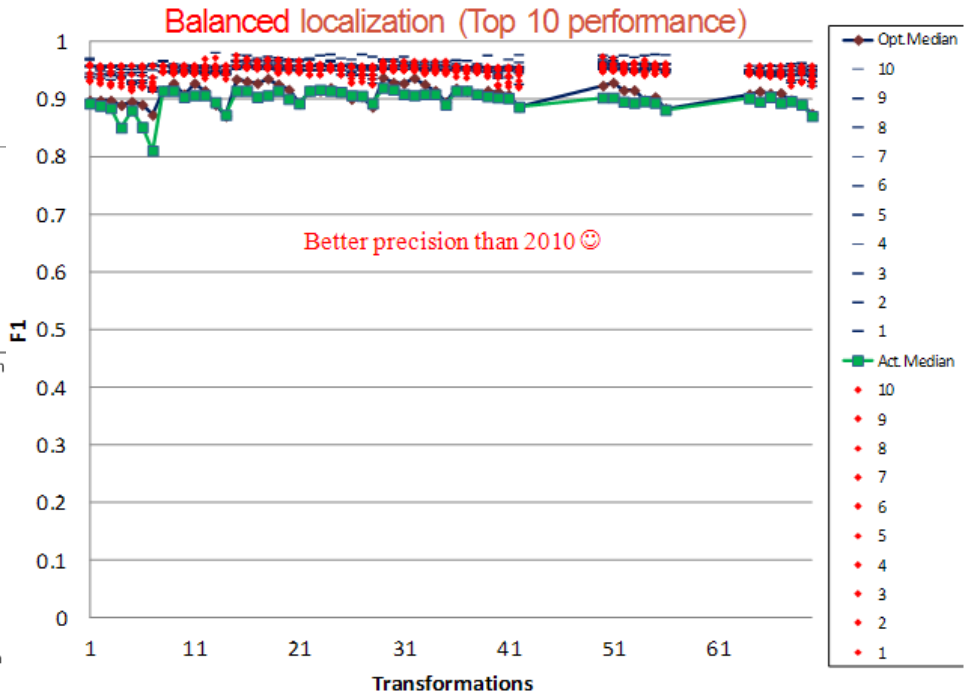
Detection progress across 3 years

TV2009

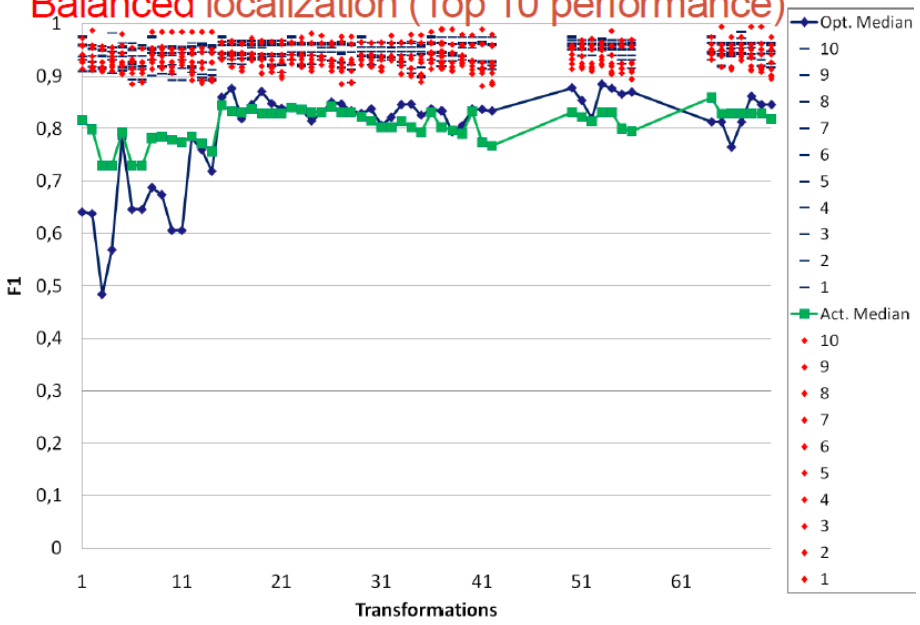


Caution!... different dataset

TV2011



TV2010



Localization progress across 3 years

Focus of participating teams

- **AT&T**: speed enhancements, new audio features, PiP detector

- **BUPT**: fusion of SIFT, color corellogram, LBP, audo, PiP detector

- **Brno**: BoW based on SIFT and SURF descriptors

- **CRIM**: added video processing, based on audio KNN approach, using visual features proposed by NTT@tv10

- **FT Orange**: 50000 visual words, spatial consistency filtering, EDF/CEPS audio features, audio has preference in results merging

Focus of participating teams

-**Osaka prefecture u.:** study performance of existing image recognition. Hashes of PCA reduced SIFT

-**INRIA tex-mex:** - presentation follows-

-**Univ. Kaiserslautern:** BoW approach plus Reciprocal geometry verification

-**KDDI:** Focus on non-geometric transformations. DCT-sign based feature representations. IDF weighting, temporal burstiness aware scoring

-**NTT:** TV10 system: Coarsely-quantized area matching (CAM) video, Divide and Locate (DAL) audio,improved to detect very short copied segments. Simple merging (audio OR video) match

Focus of participating teams

-**PRISMA (Univ Chile+Telefonica):** early fusion of audio and video features – talk follows-

-**RMIT:** apply techniques from image recognition based on global features: divide frame in border (intensity histogram) and central region (auto-correlogram)

-**Telefonica:** video system from TV10 (DART), new audio system, exploiting peak structure preservation under transformations (MASK). Median bases score normalization and fusion experiments – talk follows –

-**University of Queensland:** Combination of local binary patterns and global HSV Color histogram. No audio analysis

Observations (NDCR)

-Top actual balanced detection scores are very near to top optimal balanced detection scores

-In general top balanced detection results are better than no false alarm.

- TV11 top balanced detection results are better than 2009 and 2010. Median scores for 2011 (<= 1) are lower than 2010.

-Bigger gap between actual and optimal medians for no false alarm detection results. However, TV11 looks better than TV10

Observations

-TV11 top localization scores for both profiles are better than TV10

- On average, TV11 systems are slower than TV10!

- Good detecting systems are also good in localization.

- Audio transformations 5,6 & 7 still seems to be the hardest. while video transformations 3,4,5 & 6 seems to be the easiest.

Observations

- CCD task community is stable in size
- CCD attracts new TrecVID participants
- CCD platform is a good starting point for the INS task
- NDCR and F1 results do approach a ceiling

Questions

-Did any one cross check their TV10 & TV11 system on 2010 & 2011 queries?

- It appears that efficiency is not an important goal for participating groups

-Did any one run comparison between a+v vs v-only or a-only?

-Do teams feel that their systems are getting better? Why?

-What did systems learn over the past 4 years in this task?

-Are there any remaining challenges?

CfP Special Issue IEEE Multimedia

- **Web-Scale Near-Duplicate Search: Techniques and Applications**

- **Important Dates:**
  - Submission Deadline: 29 June 2012
  - Publication Issue: July–September 2013

- **More information at:**
  - http://www.computer.org/portal/web/computingnow/2013/mmcfp3