TRECVID 2010 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
- Extended at TV08, consolidated at TV09 (actual vs optimal NDCR)
- TV10 changes:
  - 2010: first year using internet videos (IACC). Dataset composed of much shorter videos with variable frame rates.
  - Camcorder feature back
  - just AV runs
  - adjusted 'balanced' profile

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
- For 2010 only one task type:
  - Copy detection of <u>video + audio</u> (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:

Copies {
    - 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)

    - Type 3: composed of a non-reference video only. (1/3)
  - 201 total original queries. 67 queries for each type.
  - Type 1 & 2 durations (~ 3.6 – 59 sec)
  - Type 3 durations (~ 30.4 – 162.3 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
- 5 original queries (280 transformed queries) were dropped for evaluation due to:
  - Query corruptions
  - Identifying duplicate answers within the reference set or within the same original reference video (e.g loops)

Video transformations

- Camcording transformation was restored this year (thanks to Matthijs Douze, INRIA-LEAR-TEXMEX).
- 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.

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.

22 Participants (finishers)

Asahikasei Co.	CCD	---	---	---	---	---
AT&T Labs - Research	CCD	INS	***	***	---	***
Beijing University of Posts and Telecom.-MCPRL	CCD	INS	KIS	---	SED	SIN
Brno University of Technology	CCD	***	---	***	---	SIN
City University of Hong Kong	CCD	---	KIS	---	***	SIN
IBM Watson Research Center	CCD	***	***	MED	---	***
Istanbul Technical University	CCD	---	---	---	---	---
INRIA-TEXMEX	CCD	***	***	***	***	***
KDDI R&D Labs and SRI International	CCD	---	---	***	***	***
National Institute of Informatics	CCD	INS	***	***	***	SIN
National Chung Cheng University	CCD	---	---	---	---	---
Nanjing University	CCD	INS	---	---	***	---
NTT Communication Science Laboratories-CSL	CCD	---	---	---	---	---
NTNU and Academia Sinica	CCD	---	---	---	---	---
Peking University-IDM	CCD	---	---	---	SED	---
Shandong University	CCD	---	***	---	---	***
Sun Yat-sen University - GITL	CCD	---	---	---	***	***
Telephonica Research	CCD	---	---	---	---	---
Tsinghua University-IMG	CCD	***	***	***	***	***
TUBITAK - Space Technologies Research Inst.	CCD	***	---	---	***	SIN
University of Brescia	CCD	---	---	---	---	---
University of Chile	CCD	---	---	---	---	---

--- : group didn’t participate

** : group applied but didn’t submit

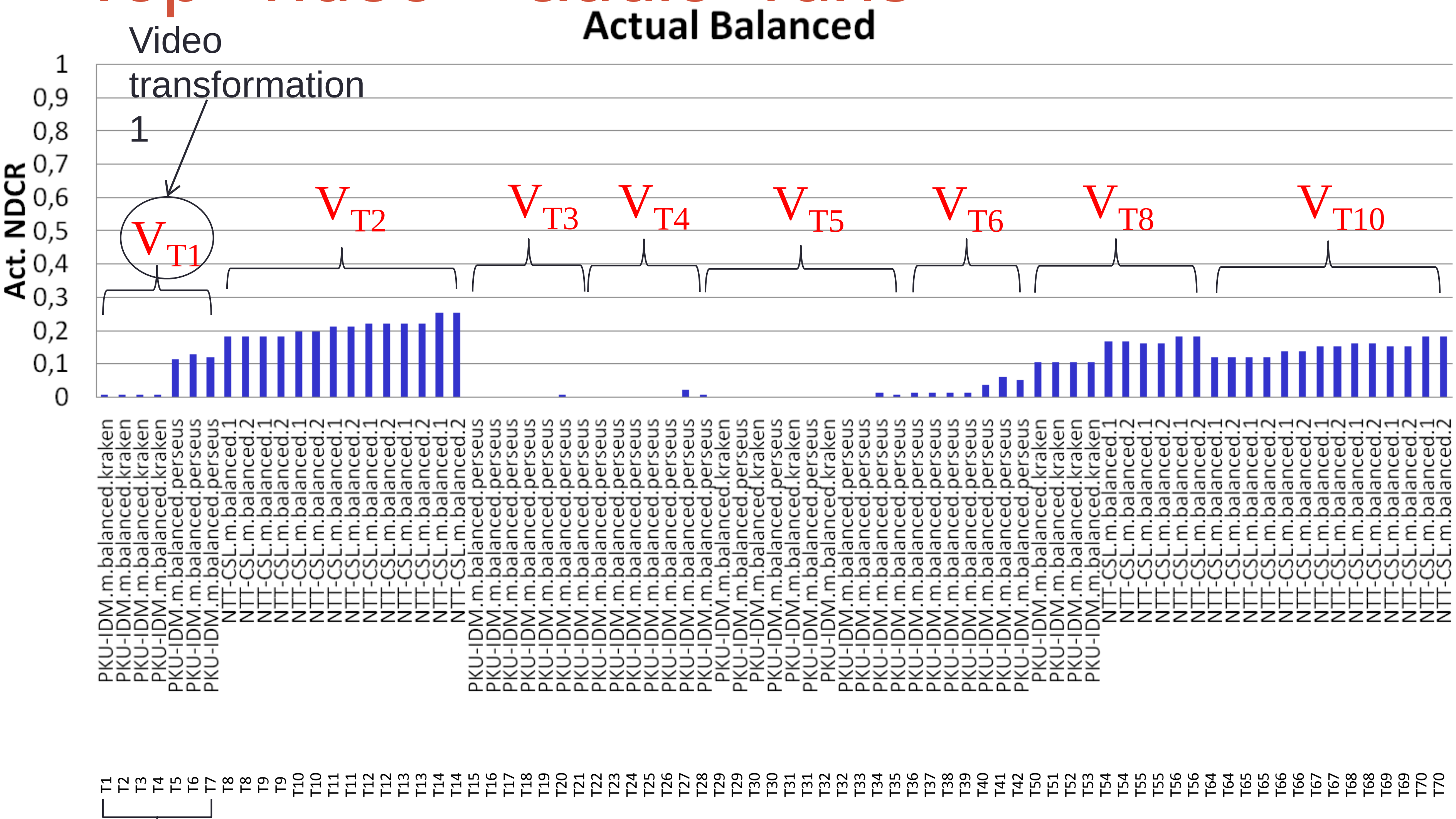
Submission types and counts

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

2009		2010	
Type M (Balanced)	Type M (NoFa)	Type M (Balanced)	Type M (NoFa)
22	20	41	37

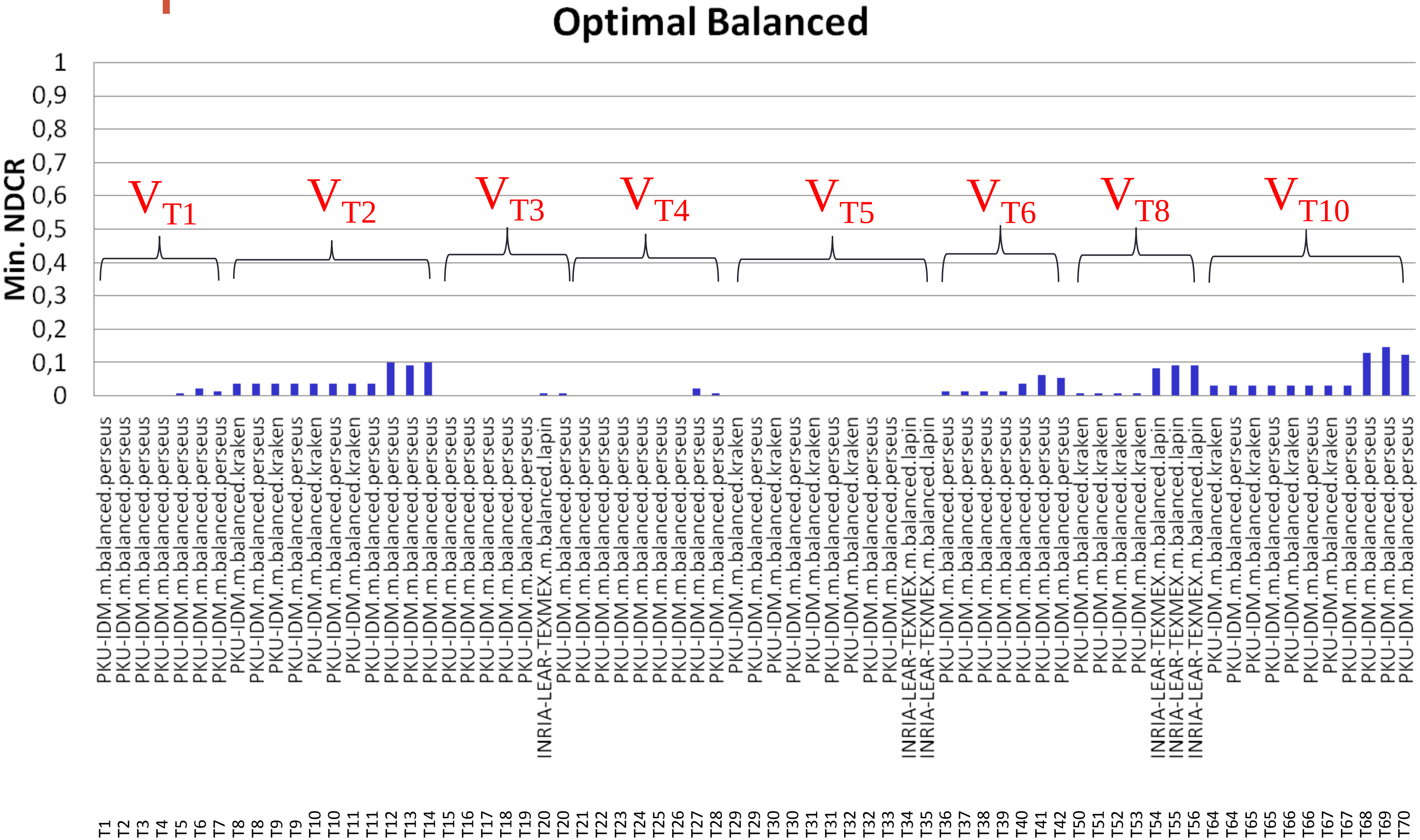
Balanced submissions between the two application profiles

Top “video + audio” runs

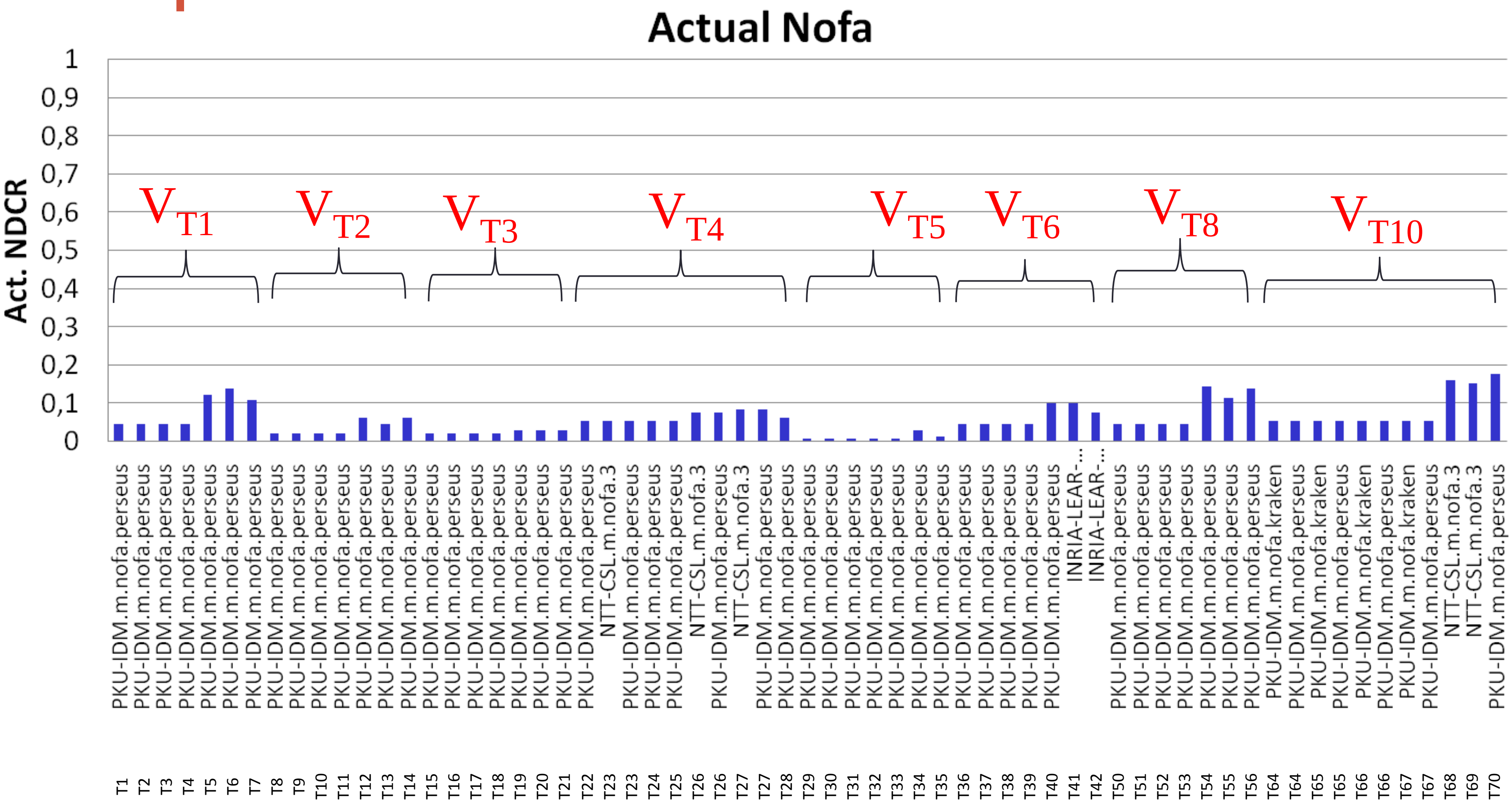


7 Audio transformations

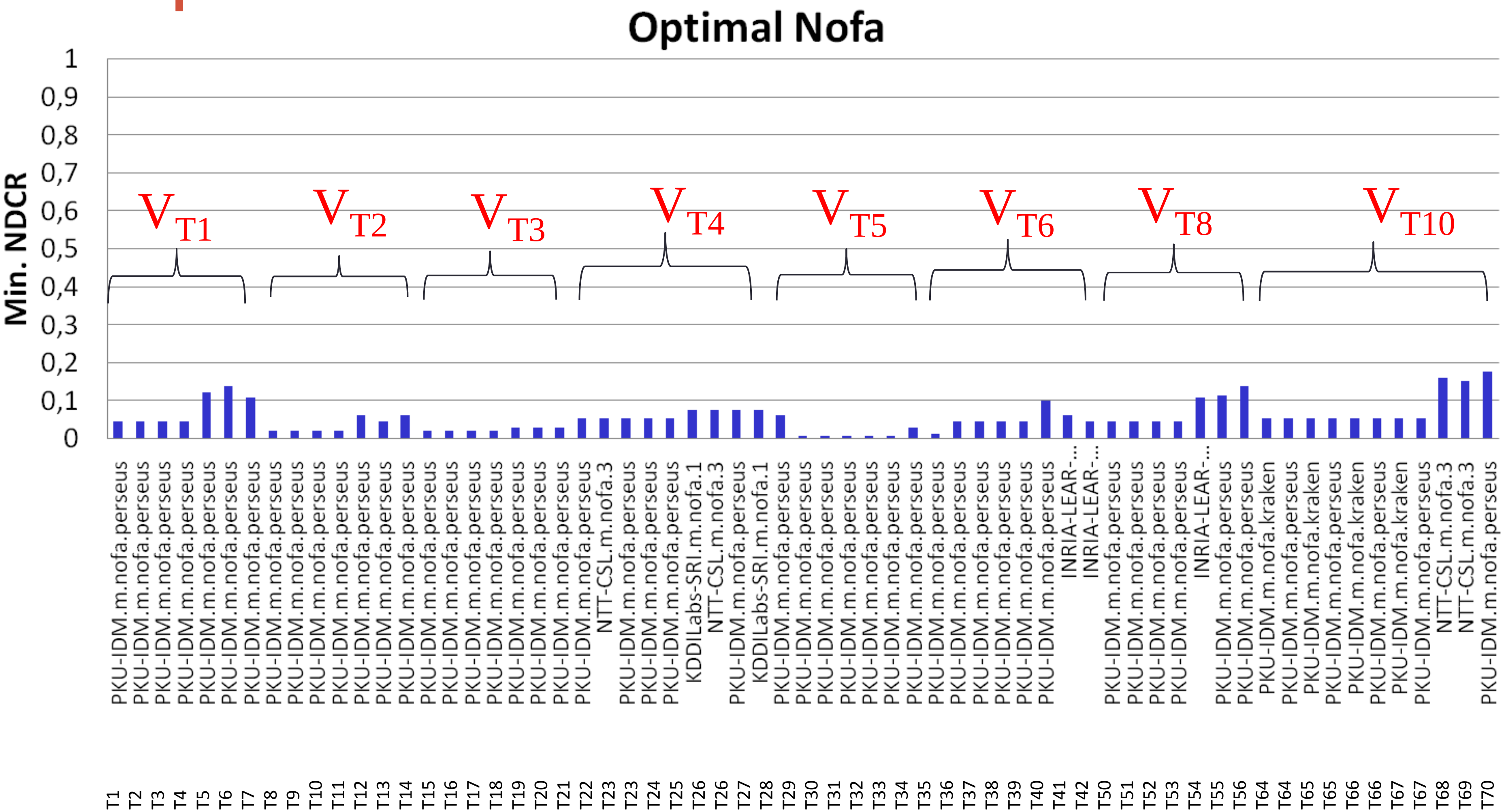
Top “video + audio” runs



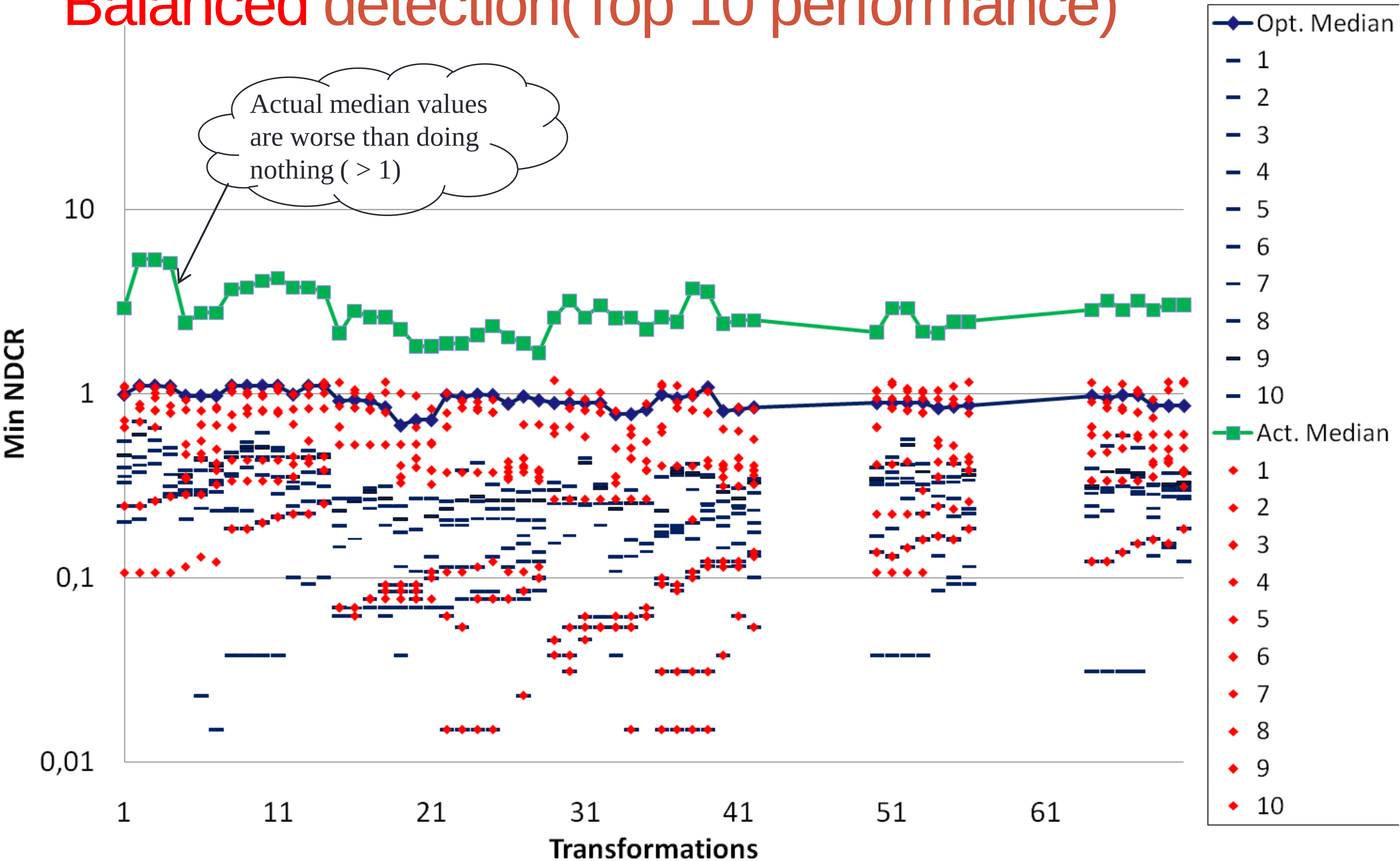
Top "video+audio" runs



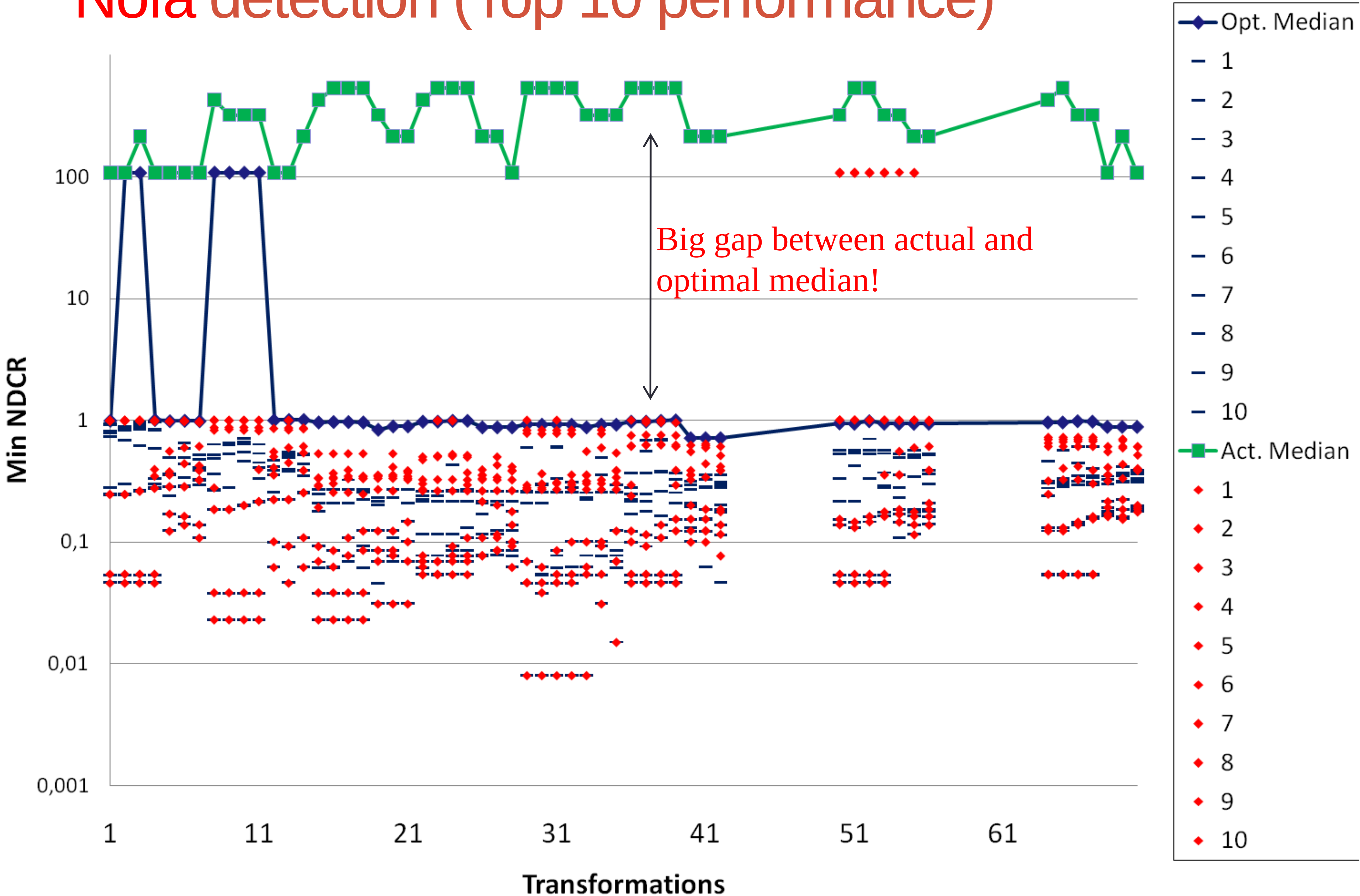
Top “video+audio” runs



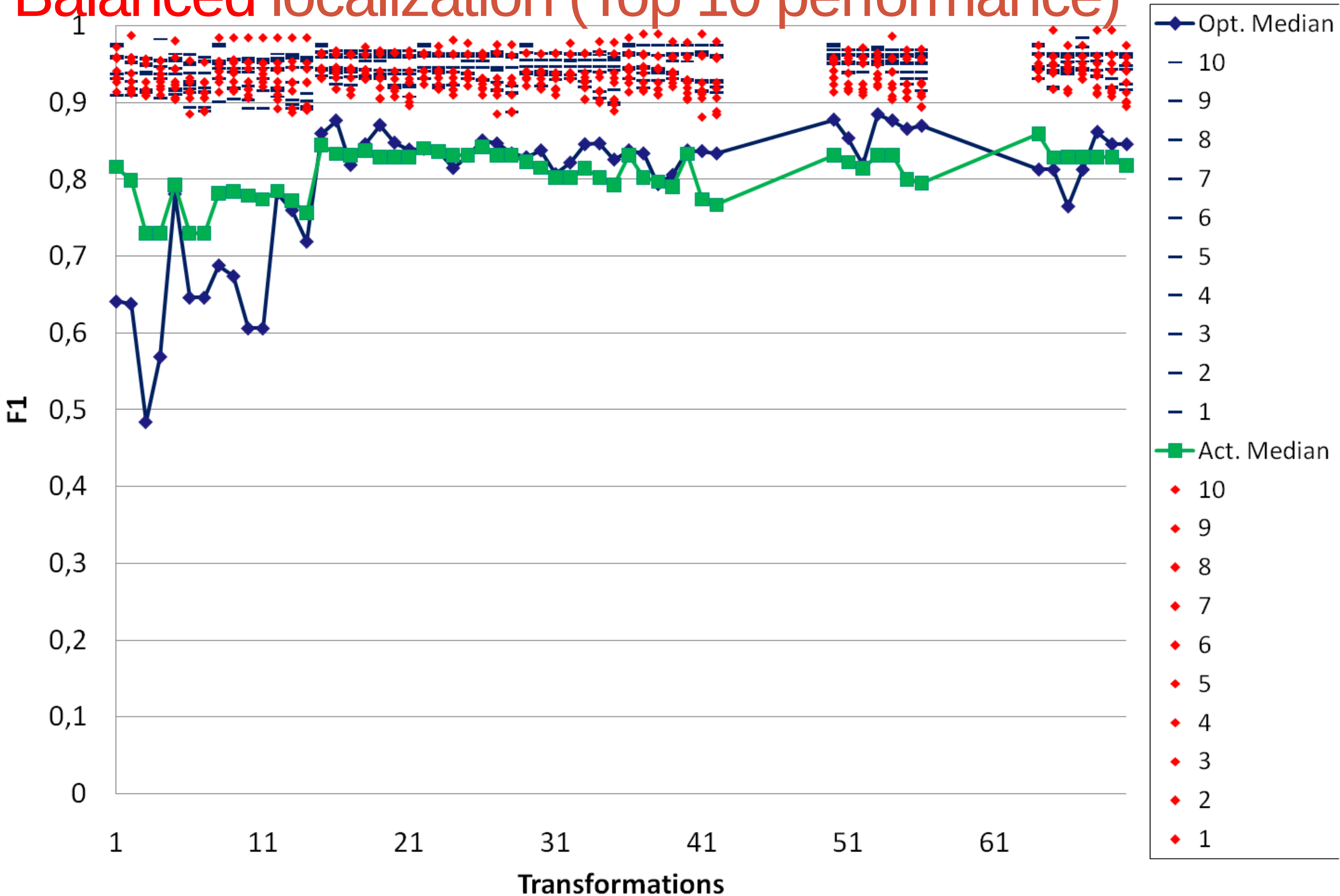
Balanced detection(Top 10 performance)



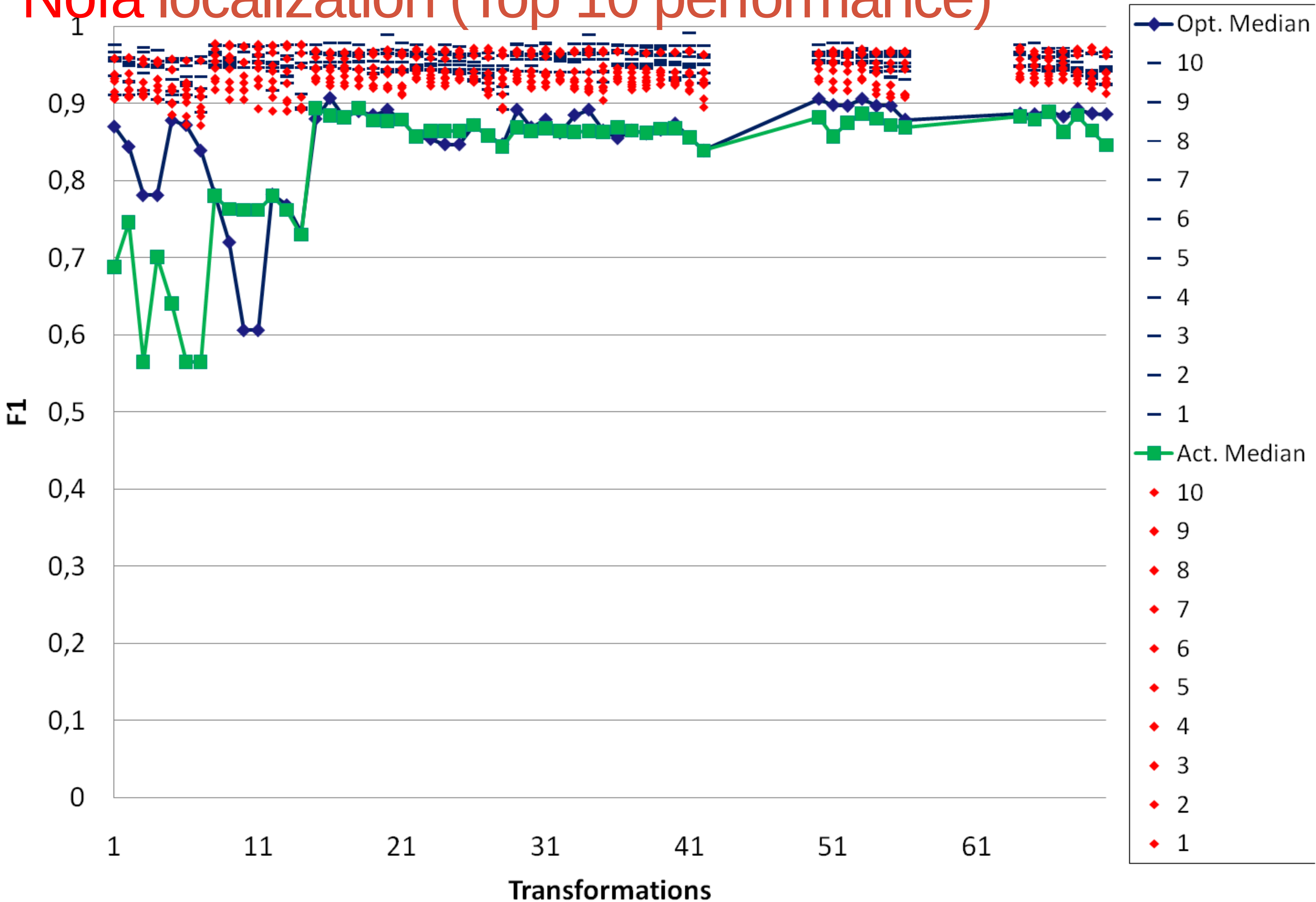
Nofa detection (Top 10 performance)



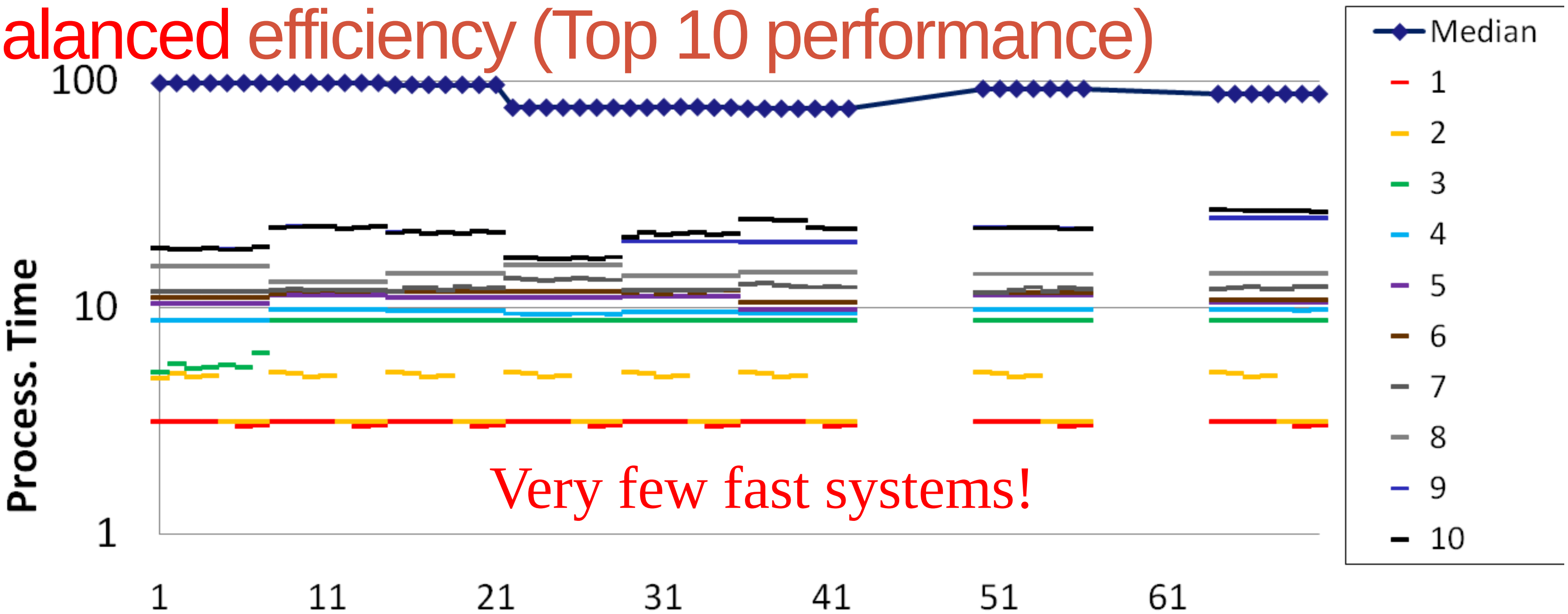
Balanced localization (Top 10 performance)



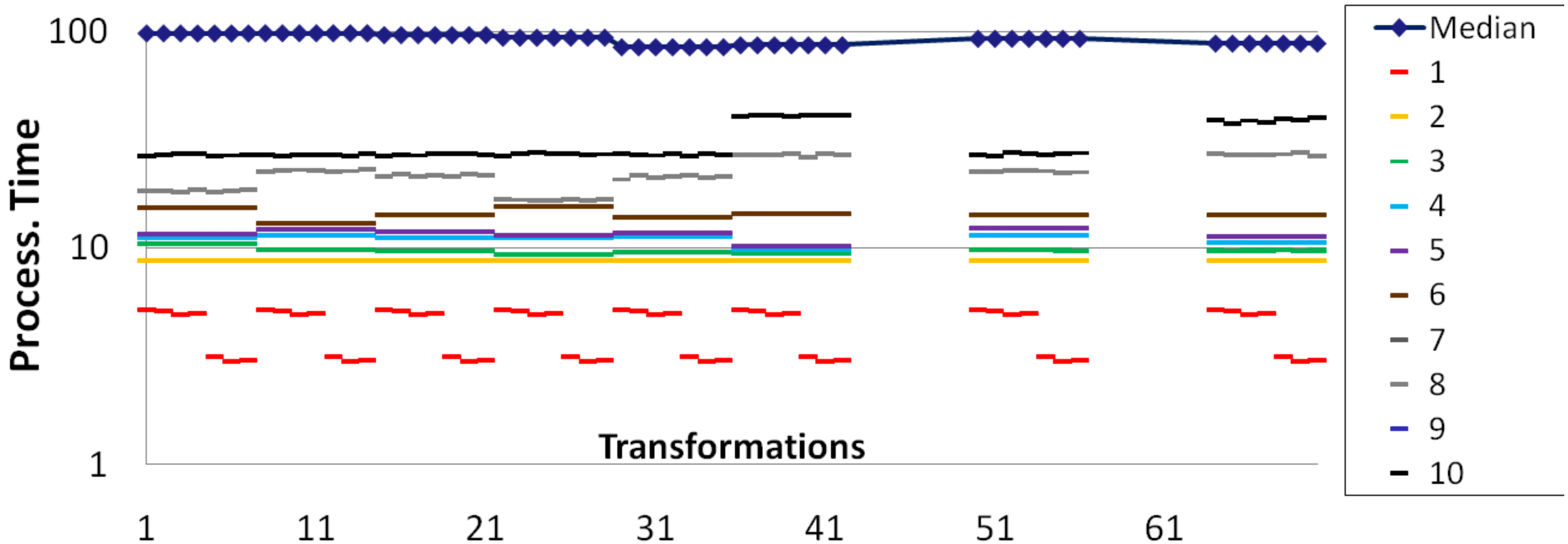
Nofa localization (Top 10 performance)



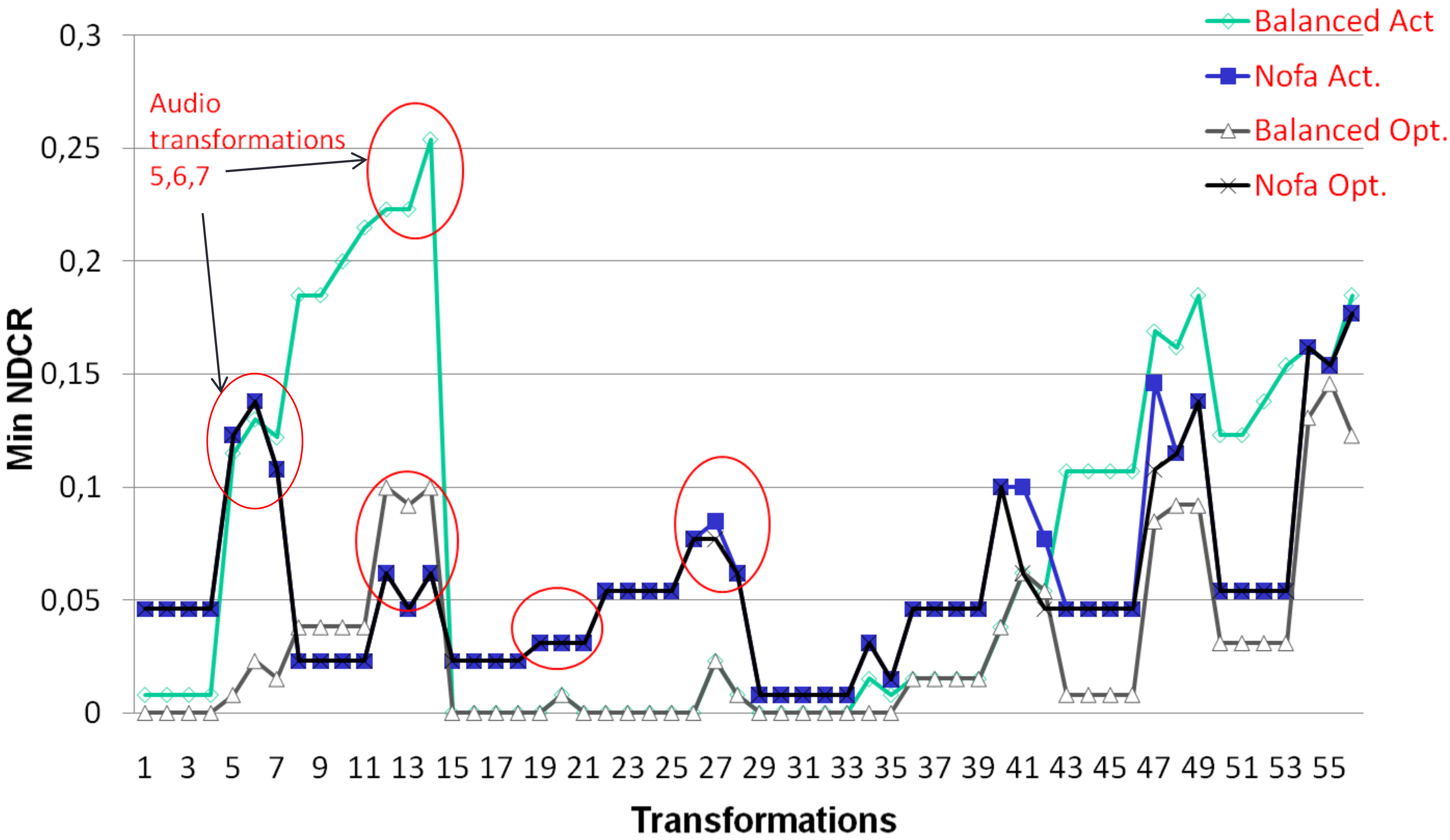
Balanced efficiency (Top 10 performance)



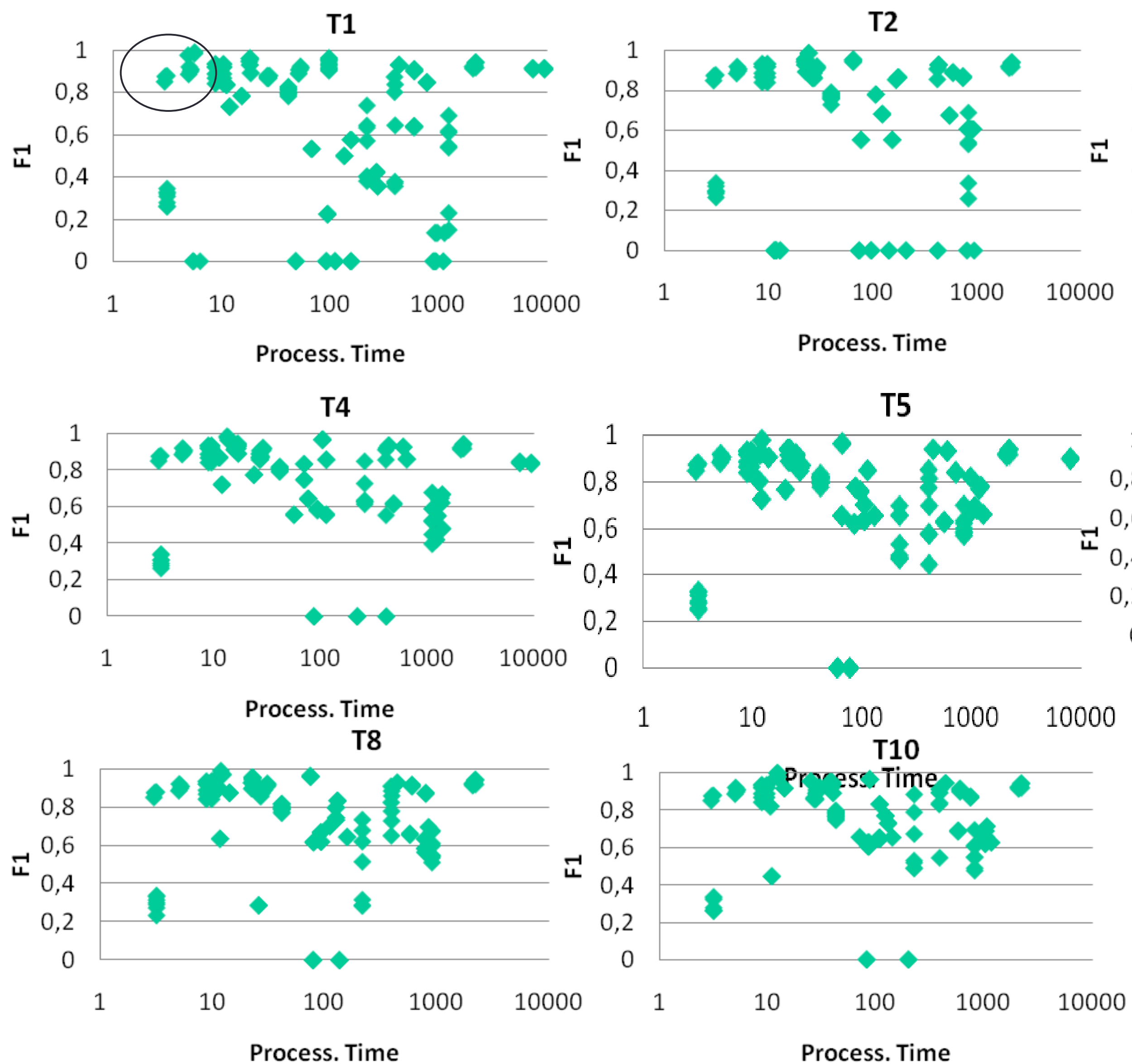
Nofa efficiency (Top 10 performance)



Comparing best runs (detection)

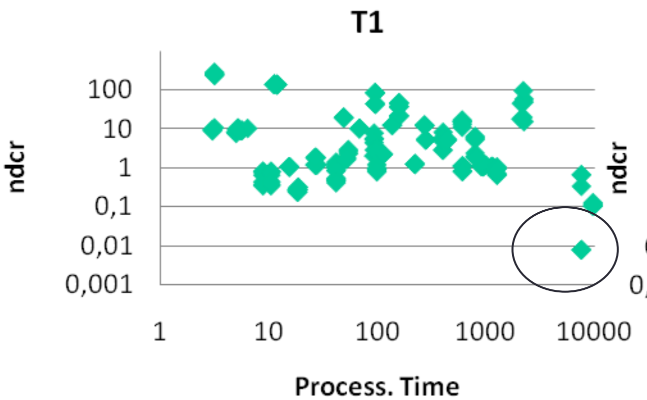


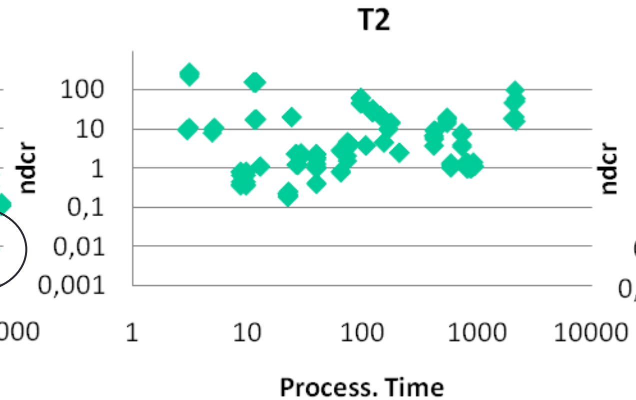
Actual Balanced runs by video transformations (across all audio transformations)

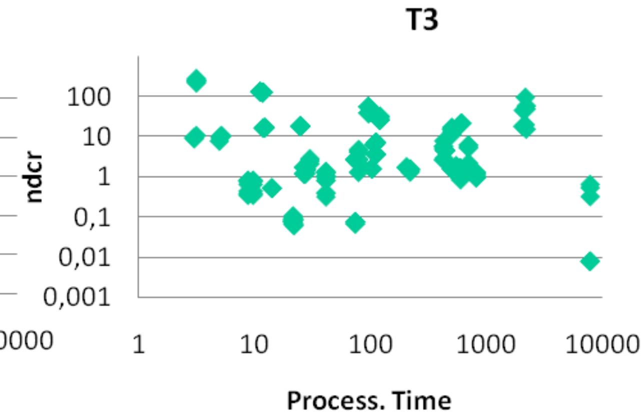


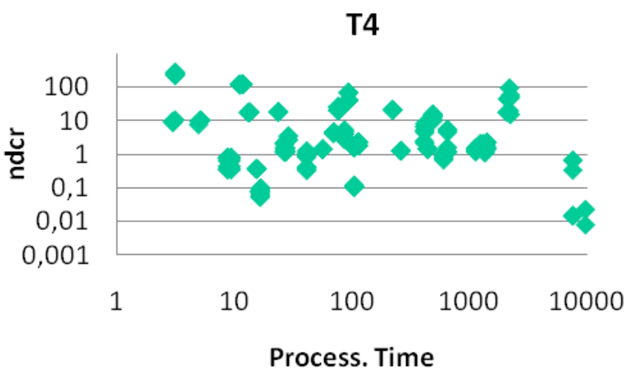
Increasing proc. time did not enhance localization. Few systems achieved high localization in small proc. time. Strong systems are efficient and precise.

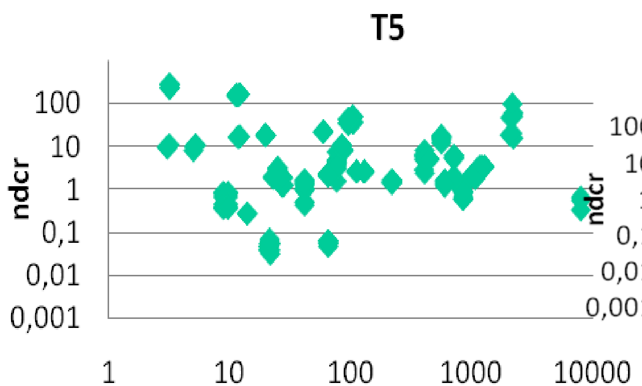
Actual Balanced runs by video transformations (across all audio transformations)

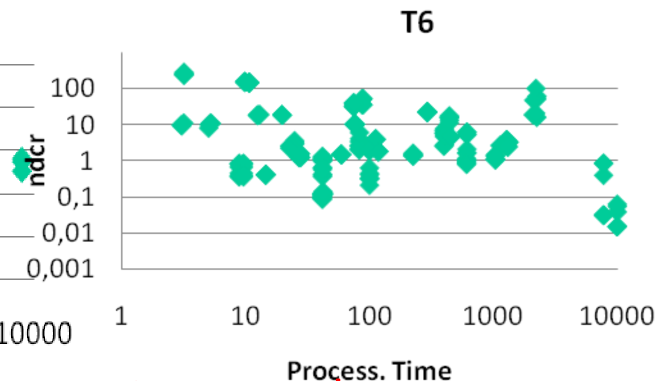


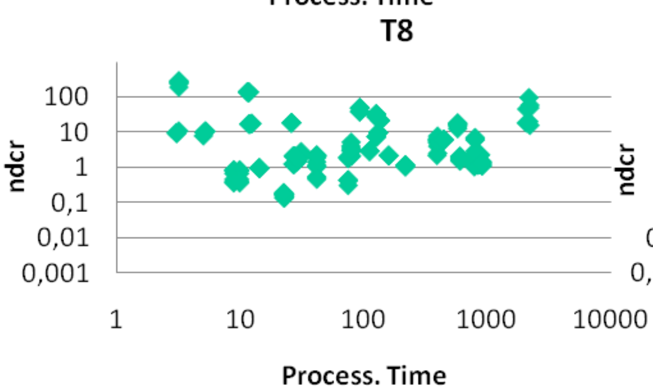


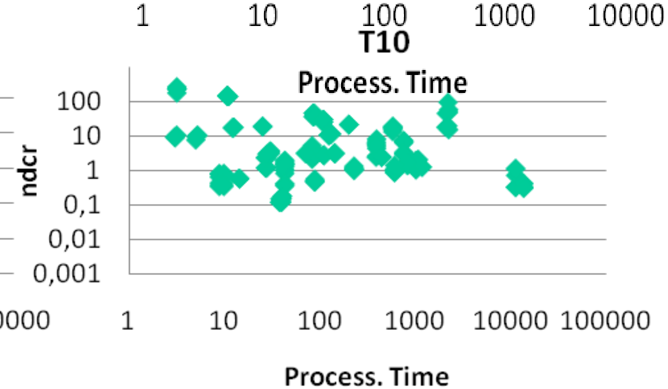






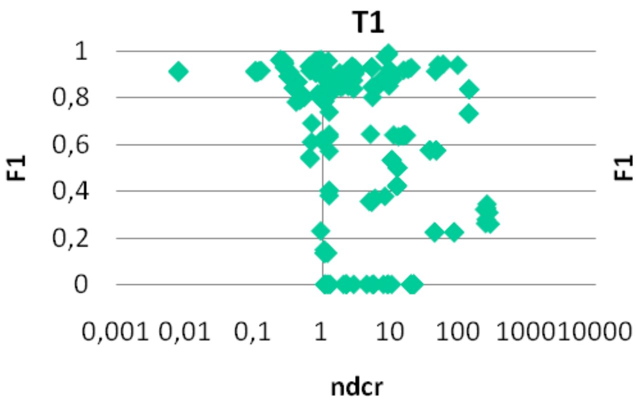


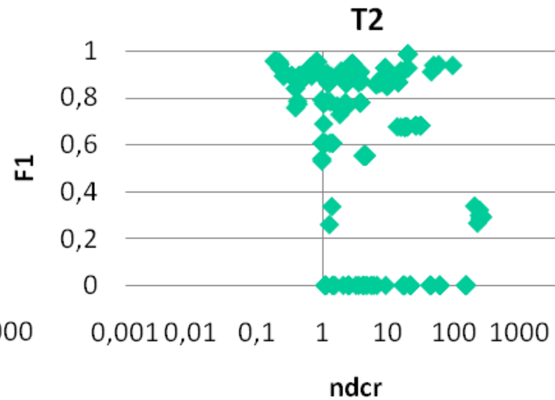


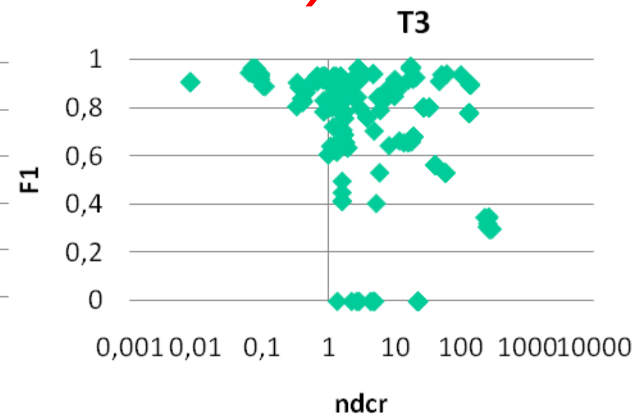


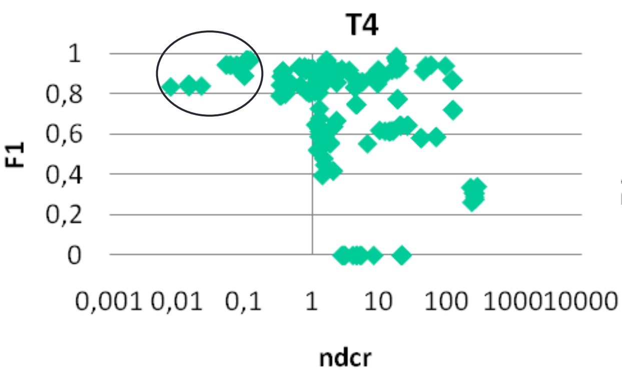
In general, more processing time does not improve detection

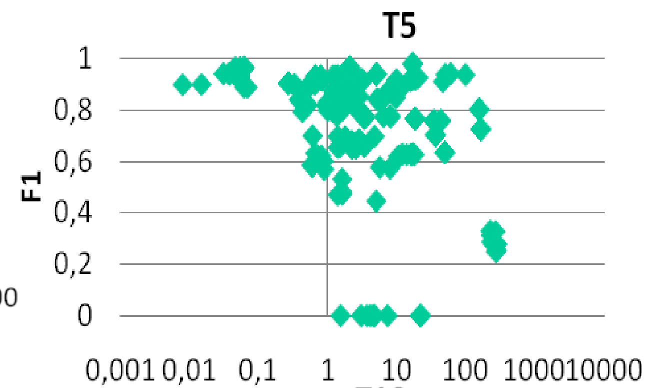
Actual Balanced runs by video transformations (across all audio transformations)

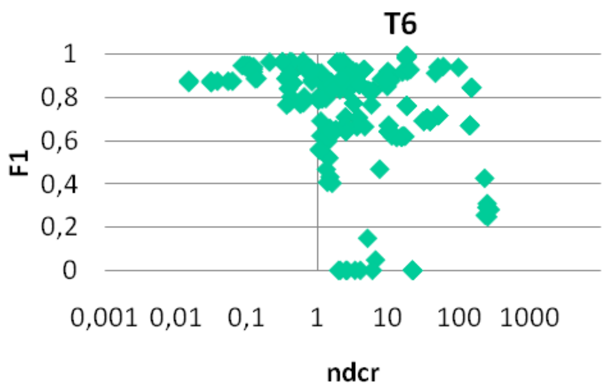


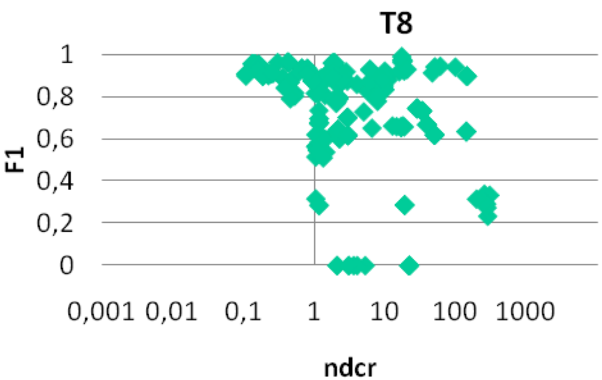


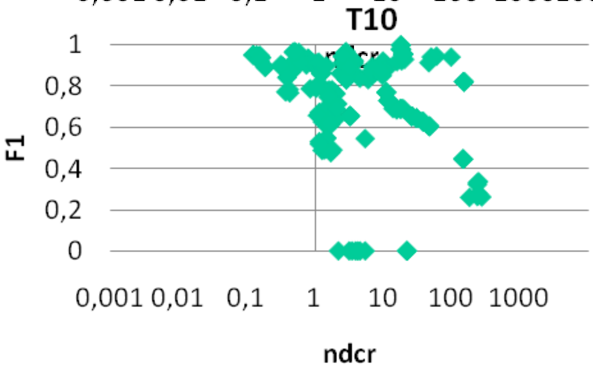












Most of the systems that are good in separating copies from non-copies (low NDCR) are also good in localization.

Observations (1)

- Some systems (incuding first-timers) have achieved very good results, the task has been difficult for many others.
- Substantial room for improvement is available for the ‘balanced’ condition indicated by difference between actual vs optimal results and difference across top runs.
- Determining the optimal treshold is still a major hurdle.
- Some systems achieved better NDCR scores compared to 2009. However the median values are higher as the dataset is very different.
- Most of the systems are still far from real-time detection.

Observations (2)

- Good detecting systems are also good in localization.
- Complex transformations (audio or video) are indeed more difficult.
- Camcording is a difficult transformation for some systems
- Some submissions were using only the video modality (eg IBM, NJU, NTNU, Univ of Chile, CUHK)
- Audio modality helped to reduce the FAR for PiP video transformations
- Most (all?) teams fuse audio and video at the decision level
- Queries with short copied segments tend to be missed

Questions

- Regarding this year:
  - How difficult/easy was the IA dataset compared to S&V ? Why ?
  - Did any one run comparison between a+v vs video-only or audio-only? (Telefonica and ..)
  - Did anybody cross check TV09 and TV10 systems on TV09 and TV10 datasets?
  - Any attempts/idea to fuse audio and video at a lower level?