INRIA-IMEDIA TRECVID 2008: Video Copy Detection

Alexis Joly, Julien Law-to, and Nozha Boujemaa

INRIA Paris-Rocquencourt, IMEDIA Team, France

Abstract. This paper reports the participation of INRIA IMEDIA team in TRECVID 2008 Video Copy Detection task. Three runs were submitted using video-only content. Two of them correspond to two different techniques based on local visual features and the last one is a combination of them. In this paper we overview the underlying methodologies and technologies and we discuss the obtained results.

1 Overview of submitted runs

1. **INRIA-IMEDIA.v.joly**: This run is a combination of the following methods: Dissociated dipoles features extraction [1] in sampled keyframes, Features indexing and retrieval with distortion-based similarity search structure [2] and finally Spatio-temporal registration of retrieved features.
2. **INRIA-IMEDIA.v.ViCopT**: ViCopT [3] system is based on an assymetric processing between reference contents and queries. Offline reference set processing performs a tracking of visual local features and index them differently according to some labels of behaviour. Online Search is performed with distortion-based similarity search structure directly on the local features extracted in keyframes, without tracking. Finally, a robust voting algorithm based on labels of behavior [4] is applied to enhance spatio-temporal coherence.
3. **INRIA-IMEDIA.v.fusion**: Combination of two previous approaches. ViCopt is first applied to all queries. After thresholding, the remaining queries are processed by the dissociated dipoles based approach. After thresholding again, the remaining queries are resized (x2) and processed by the dissociated dipoles based approach

2 Discussion about submitted runs

The idea of the first run was to extend a previous technique dedicated to still images that did obtained very good performances in the ImagEval benchmark 2006 (Task1: Transformed images retrieval). Keyframes were thus simply sampled and processed as still images. Temporal aspects have been added in the search algorithm, to enforce spatio-temporal coherence and reduce time processing of this step. The second run was obtained with ViCopT, a system developped by IMEDIA in the last few years in collaboration with INA and dedicated to video copy detection. Due to temporal tracking during indexing step, it is much faster than the first technique while being as much or more robust for much encountered real attacks. The feature extraction used in ViCopT is however not invariant to the picture-in-picture transformations of TRECVID 2008 and we can thus expect some performance degradation. Submitting these two runs was a good way for us to compare both techniques and to decide wether we should integrate the dipole features in ViCopT. The idea of the last run was to benefit from both techniques by combining them. ViCopT being much faster, it is applied first. Queries without significant results are then processed by the first technique, in two steps: with normal queries and then with resized-and-split queries if no significant results were found within first step. Resizing and splitting the queries allows to improve the retrieval of very small picture-in-picture.

3 Underlying techniques

This section describes the techniques underlying the submitted runs. We precise for each in which run they were used.

3.1 Dissociated dipoles features extraction

Used in run INRIA-IMEDIA.v.joly and INRIA-IMEDIA.v.fusion
In [1], we proposed new local photometric descriptors based on dissociated dipoles for transformed images or rigid objects retrieval. Dissociated dipoles are non local differential operators which have been proved to be more stable than purely local standard differential operators. In this study, we define and compute specific oriented dissociated dipoles around multi-resolution color Harris points and we form 20-dimensional normalized features, invariant to rotation, affine luminance transformations, negative or flip. In a comparison with extensively used SIFT descriptors, we show that such descriptors are as much efficient while containing 6 times less information. This allows the complete retrieval to be both more efficient and faster. This strategy ranked first in **ImagEval**[1] benchmark 2006, which, as far as we know, was the first competition including a transformed image recognition task (or content-based copy retrieval task).

3.2 Distortion-based similarity search structure (used in all runs)

Used in all runs
In [2], we proposed a new approximate similarity search technique in which the selection of the feature space regions is not based on the distribution in the database but on the distribution of the features distortion. Since most robust content-based copy detection techniques are based on local features, the approximation can be strong and reduce drastically the amount of data to explore. This framework was already applied on very large datasets containing more than one billion local features corresponding to $30,000$ hours of video.

[1] http://www.imageval.org/

3.3 Matching improvements by spatio-temporal registration (used in run INRIA-IMEDIA.v.joly and INRIA-IMEDIA.v.fusion)

Used in run INRIA-IMEDIA.v.joly and INRIA-IMEDIA.v.fusion
Once the local features of a given temporal query window have been matched to their similar features in the database, we first rank all retrieved video sequences by voting and we use a stoplist to keep only the most relevant ones. We then filter the matches by a spatio-temporal registration computed on the spatio-temporal positions of the query features and the retrieved features. We first compute a temporal registration for each query segment to estimate the best temporal and filter the matches according to this parameter. We then apply geometric registration thanks to an affine model and a RANSAC algorithm and filter the matches that are not geometrically consistent. Then, we estimate the best final temporal offset of the entire query by the one ot the segment with the most remaining matches. We then prune all segments for which the temporal offset is far from the best one. Reference segment is then estimated by the minimum and maximum time codes of the remaining matches.

3.4 ViCopT

Used in run INRIA-IMEDIA.v.ViCopT and INRIA-IMEDIA.v.fusion
ViCopT [3] system is also based on local features extraction and distortion-based similarity search structure for matching candidate features in large datasets. However, the technique differs in several crucial aspects: First of all Vicopt involves estimating and characterizing trajectories of points of interest throughout the video sequence. It takes advantage of such trajectories to characterize the spatio-temporal content of videos: it allows the local description to be enriched by adding a spatial, dynamic and temporal behavior of this point. The aim is to provide a rich, compact and generic video content description where the behavior of a point along a trajectory can be seen as a temporal context of this point. The redundancy of the local description along the trajectory is efficiently summarized (the number of features is reduced by 50 in average in our experiments), with a reduced loss of information. By using the properties of the built trajectories, a label of behavior can be assigned to the corresponding local description: the categories of behaviors are simply obtained with heuristics and thresholds. This leads to different levels of description:

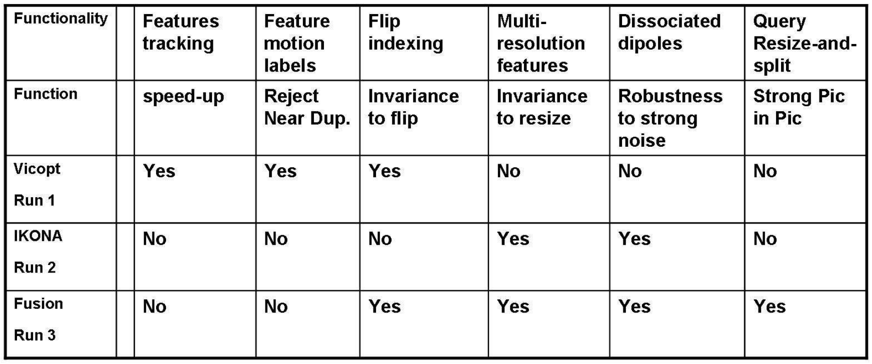
- Low-Level: spatio-temporal description of the signal.
- Mid-Level: trajectory parameters.
- High-Level: labels based on the behavior of points (temporal context).

The low and mid-level descriptors are obtained at the end of a purely bottom-up process, independent of the application: it is a generic description of the video, and is computed only once. The final high-level description follows from a top-down process, that is specific to the application.
As the off-line indexing part needs long time computational and as the system of retrieval needs to be in real-time, the whole indexing process with tracking

can not be done for the candidate video sequences. A more fundamental reason is that the system has to be robust to small video insertion, or to re-authored video. The retrieval approach is therefore asymmetric and queries are local descriptors selected as following: every p frames, n points of interest are extracted. The advantage of the asymmetric technique is an on-line choice of the number of queries and the temporal precision, which gives flexibility to the system.
Finally, a robust voting and registration algorithm based on labels of behavior, presented in [4], is used to efficiently discriminate the remaining candidate results.

4 Functional summary

Figure 1 gives a summary of the main functionalities used in the three runs.



Functionality		Features tracking	Feature motion labels	Flip indexing	Multi-resolution features	Dissociated dipoles	Query Resize-and-split
Function		speed-up	Reject Near Dup.	Invariance to flip	Invariance to resize	Robustness to strong noise	Strong Pic in Pic
Vicopt Run 1		Yes	Yes	Yes	No	No	No
IKONA Run 2		No	No	No	Yes	Yes	No
Fusion Run 3		No	No	Yes	Yes	Yes	Yes

Fig. 1. Functional summary of the three submitted runs

5 Results analysis

The three official trecvid 2008 notebook pages for each run are given in the annex of this paper. We give here some additional analysis we made on the results. We first report a comparative study between our three run, we then discuss the influence of two parameters of the experimental setup, i.e the length of the queries and the false alarams cost. We finally discuss the trade-off between speed and accuracy.

5.1 Comparative results between the three runs

Figure 2 gives the minimum NDCR for each run and for all transformations. NDCR (Normalized Detection Cost Rate) is a weighted linear combination of the

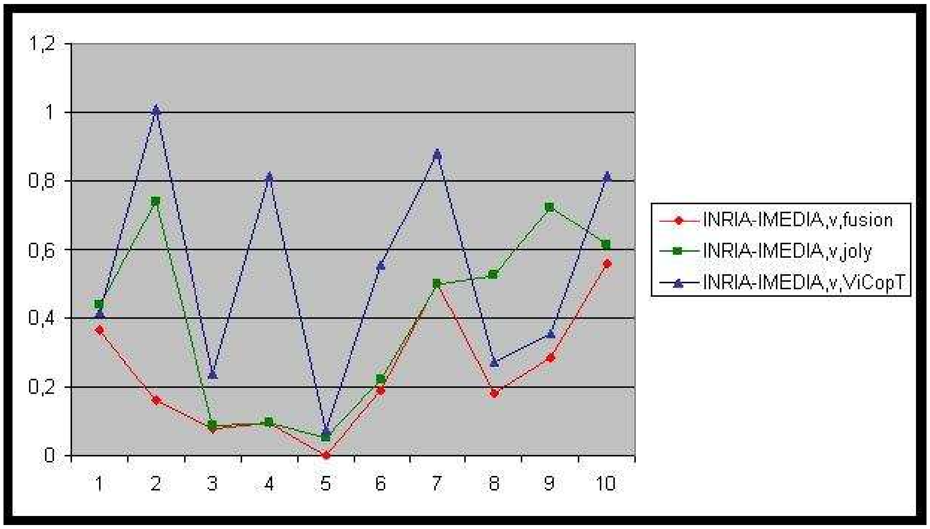


Fig. 2. min NDCR over all transformations

system's Missed Detection Probability and False Alarm Rate (measured per unit time). The constant parameters of NDCR represent both the richness of events in the source data and the relative detriment of particular error types to a hypothetical application. More information about this metric can be found here on trecvid website (http://www.nist.gov/speech/tests/trecvid/2008/doc/EventDet08-EvalPlan-v04.htm).

The figure shows that the fusion run performs always better than the two individual runs. The easiest transformation for all 3 runs is the changing of gamma parameter (transformation 5). The hardest transformation for all 3 runs is the combination of all possible attacks (transformation 10).

The highest gain of the fusion run against the two individual runs is obtained for the transformation 2, which corresponds to the Picture in Picture attack. This is mainly due to the fact that the fusion run is not only a fusion of two other runs but the fusion of three runs: the two submitted ones and a third not submitted one obtained by splitting and resizing the queries before retrieval. This run was dedicated to this Picture in Picture attack that appeared to be to strong for the scale invariant features we used.

The figure also shows that Vicopt system performs consistently worst than the still image strategy (INRIA-IMEDIA.v.joly) on transformation 4 and 7 involving strong photometric degradations (compression, noise, etc.). This is probably due to the instability of the tracking procedure to such attacks.

Table 1 gives the total recall of the three runs over all queries, using only on the best match of each query. It gives a more intuitive measure of the robustness of each method and confirms that the fusion run is much better.

run id	recall
INRIA-IMEDIA.v.joly	0.63
INRIA-IMEDIA.v.Vicopt	0.52
INRIA-IMEDIA.v.fusion	0.83

Table 1. Recall of the three runs over all queries

To analyse more precisely what is lost or won by each method, table 2 gives the same recall on the intersection and the union of results of each run. It shows that only 35% of the positive queries are retrieved by both Vicopt and the still image based technique whereas the union of their results cover 73% of the positive queries. 28% of the positive queries are retrieved only by the still image method and this correspond mainly to the strong photometric degradations discussed above. 17% of the positive queries are retrieved only by Vicopt with a wide part of them composed of the flip transformations that was not handled by the still image technique (Vicopt did index the flip version of each reference video). About 10% of the positive queries are retrieved only by the fusion run and correspons mainly to the Picture in Picture transformation as discussed above.

run ids	intersection recall	union recall
INRIA-IMEDIA.v.joly vs INRIA-IMEDIA.v.Vicopt	0.35	0.73
INRIA-IMEDIA.v.joly vs INRIA-IMEDIA.v.fusion	0.61	0.85
INRIA-IMEDIA.v.Vicopt vs INRIA-IMEDIA.v.fusion	0.49	0.84

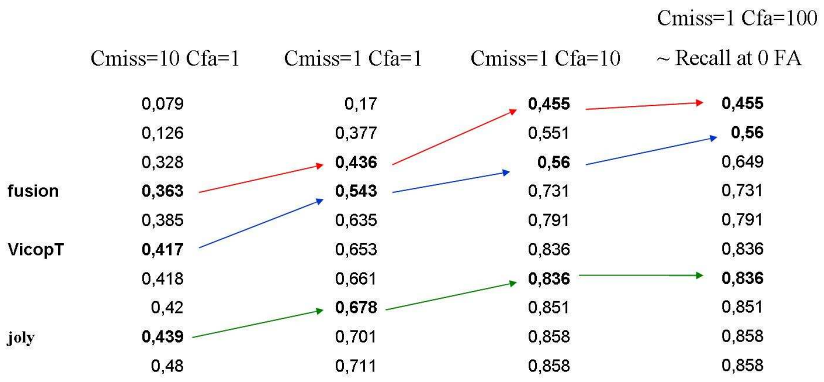
Table 2. Recall over union and intersection of run results

5.2 Influence of false alarms cost

Figure 3 illustrates the influence of the alarms cost of the min NDCR metric on the overall ranking among all runs. The TRECVID 2008 default relative costs between missed positive results and false alarms were set to $Cmiss = 10$ and $Cfa = 1$. Increasing Cfa emphasizes the cost of false alarms. The figure shows that all our three methods and specifically Vicopt performs better on high false alarms costs. Vicopt strategy labelling points trajectory was indeed specifically designed to reject ambiguous contents creating a lot of false alarms such as near duplicates. This experiment does not really proves its effiency in rejecting near duplicates but shows that it is efficient to reject false alarms in general.

5.3 Influence of query length

Figure 3 illustrates the influence of the query length on the accuracy of the detection. It plots the min NDCR measure at different query length intervals for



	Cmiss=10 Cfa=1	Cmiss=1 Cfa=1	Cmiss=1 Cfa=10	Cmiss=1 Cfa=100 ~ Recall at 0 FA
	0,079	0,17	**0,455**	**0,455**
	0,126	0,377	0,551	**0,56**
	0,328	**0,436**	**0,56**	0,649
fusion	**0,363**	**0,543**	0,731	0,731
	0,385	0,635	0,791	0,791
VicopT	**0,417**	0,653	0,836	0,836
	0,418	0,661	**0,836**	**0,836**
	0,42	**0,678**	0,851	0,851
joly	**0,439**	0,701	0,858	0,858
	0,48	0,711	0,858	0,858

Fig. 3. Influence of false alarms cost on runs ranking based on min NDCR metric

our three runs and for the median and the best run over all participants runs. It shows that the performances of most systems, including the best one and Vicopt, increase with the length of the queries, which is quite intuitive. On the other side our still image based technique (and consequently the fusion run) performs better on shorter queries. This is probably due to the fact that longer queries leads to higher temporal imprecision and/or confusion between the query frames and the reference ones.

5.4 Time vs Quality

Figure 5 plots the retrieval time provided by all participants versus the min NDCR measure. It shows that Vicopt system is much faster than our two other runs thanks to the tracking strategy used during indexing. Vicopt is about 10 times faster than our still image based method and obtains a very good time/quality tradeoff over all runs.

6 Conclusion

As a conclusion we give here a short list of the main concerns we did learn by this TRECVID participation:

- Local features based approaches with geometry consistency perform the best among all participant runs.
- Our dissociated dipoles local features give very good results while being 6 times smaller than SIFT features (used by most other techniques among the best ones).
- Tracking the local features in time allows to compress the dataset and to speed-up consequently the rerieval (by about a factor 10) but is less robust to strong photometric degradations.

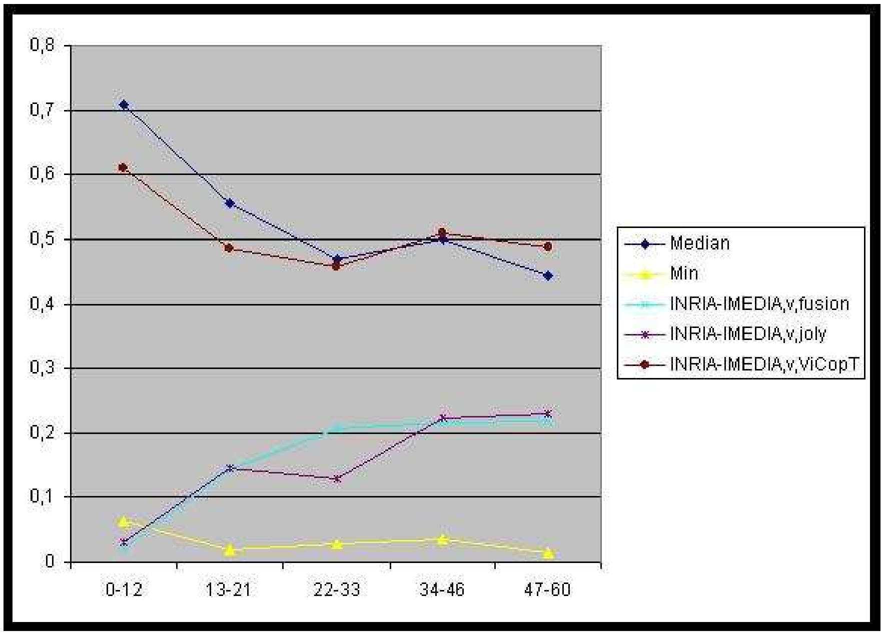


Fig. 4. min NDCR measure for different query lentgh intervals (sec)

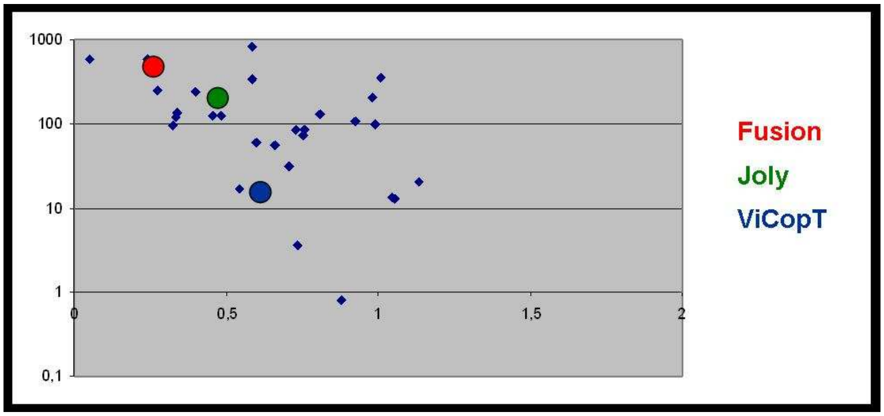


Fig. 5. Retrieval time vs min NDCR measure for all participant runs

- Strong Picture in Picture transformations are not retrieved by common scale invariant local features and must be processed by specific strategies such resizing the queries or the reference videos which are time and space consuming.
- Combining different complementary methods can improve drastically the results while offering nice time/quality tradeoffs in real life scenarios.

– Flipped videos are not retrieved by non-flip invariant local features despite the symetric nature of a lot of objects in natural scenes.

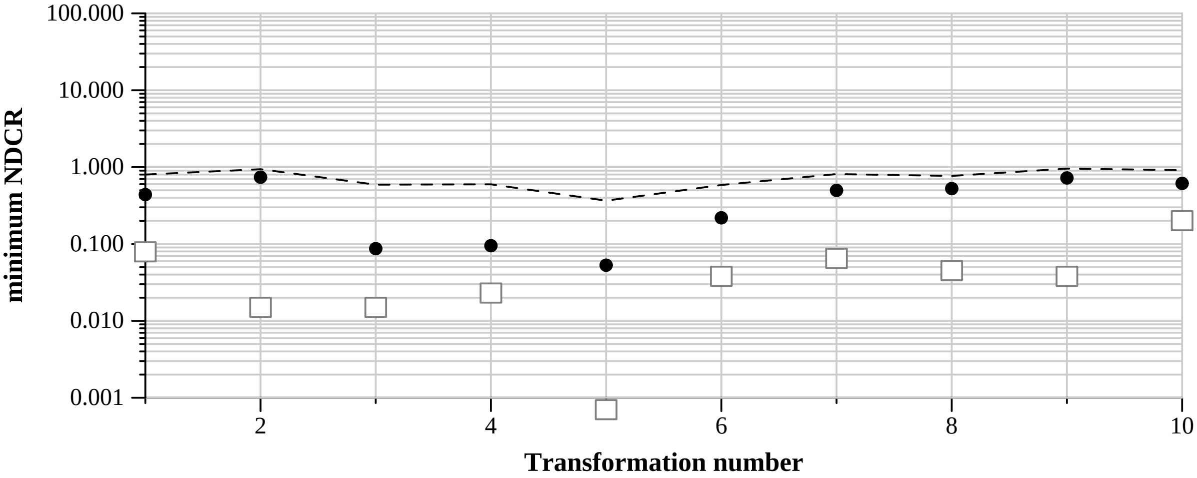
References

1. Alexis Joly, “New local descriptors based on dissociated dipoles,” in *CIVR ’07: Proceedings of the 6th ACM international conference on Image and video retrieval*, New York, NY, USA, 2007, pp. 573–580, ACM.
2. A. Joly, C. Frélicot, and O. Buisson, “Content-based copy retrieval using distortion-based probabilistic similarity search,” *IEEE Trans. on Multimedia*, vol. 9, no. 2, pp. 293–306, 2007.
3. Julien Law-To, Valrie Gouet-Brunet, Olivier Buisson, and Nozha Boujemaa, “Video copy detection on the internet: The challenges of copyright and multiplicity.,” in *ICME*, 2007, pp. 2082–2085.
4. Julien Law-To, Olivier Buisson, Valerie Gouet-Brunet, and Nozha Boujemaa, “Robust voting algorithm based on labels of behavior for video copy detection,” in *MULTIMEDIA ’06: Proceedings of the 14th annual ACM international conference on Multimedia*, 2006, pp. 835–844.

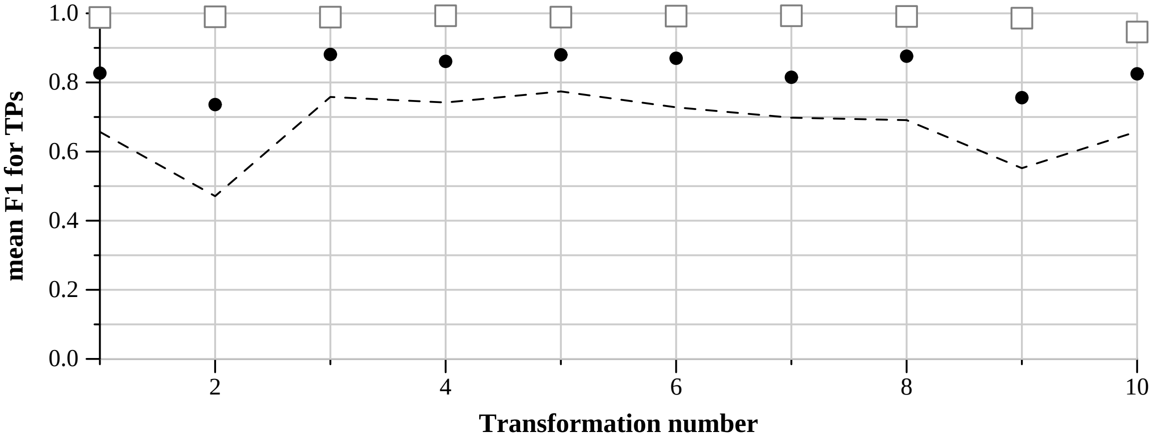
TRECVID 2008: copy detection results

Run name: INRIA-IMEDIA.v.joly
Run type: video-only

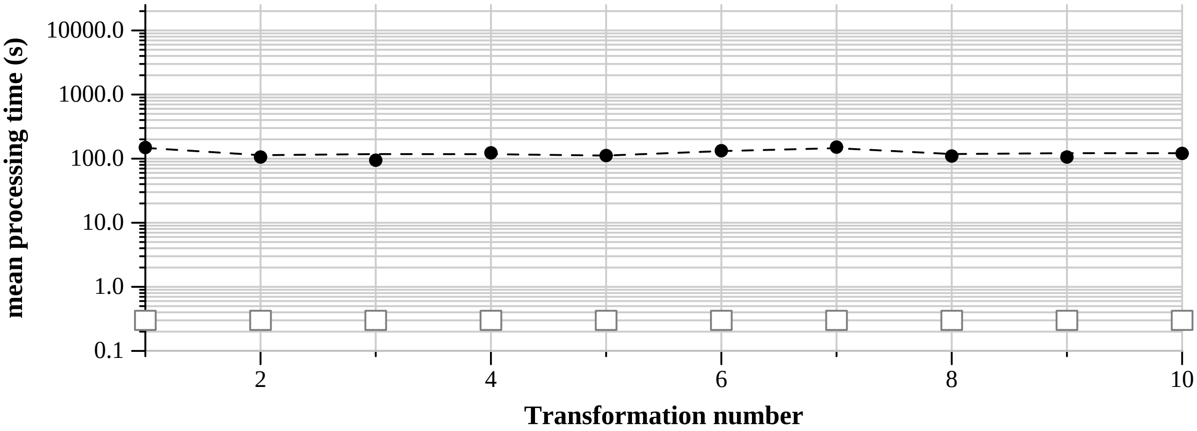
Run name:	INRIA-IMEDIA.v.joly
Run type:	video-only



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

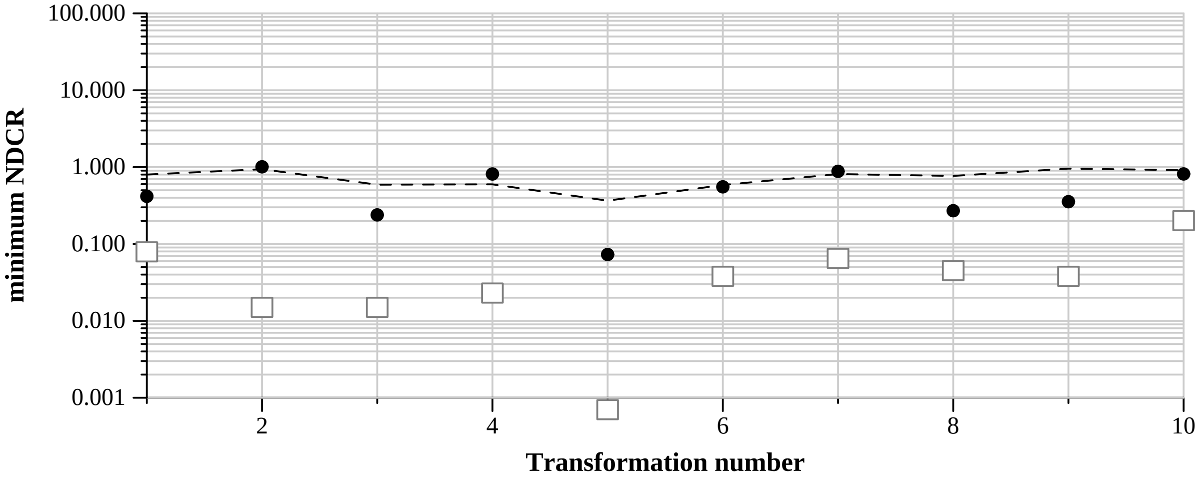


Run score (dot) versus median (---) versus best (box) by transformation

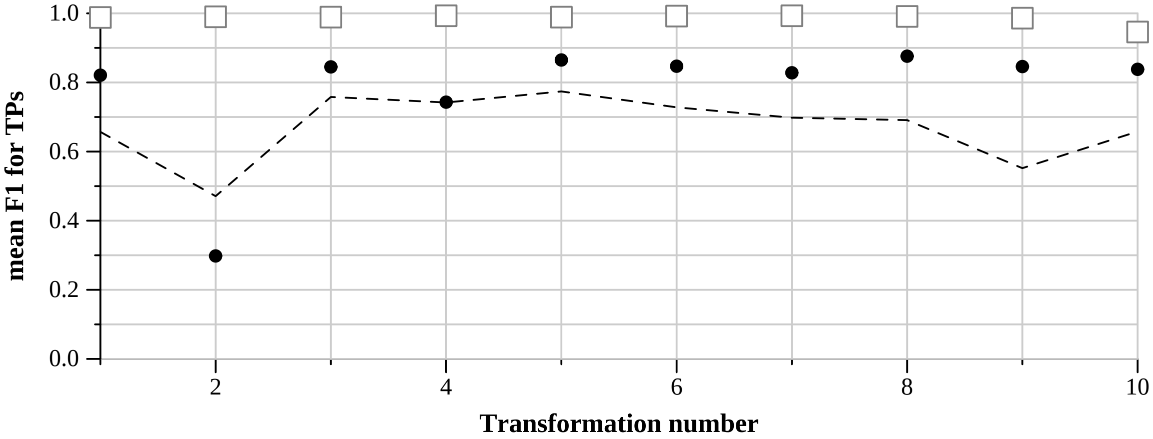


Run score (dot) versus median (---) versus best (box) by transformation

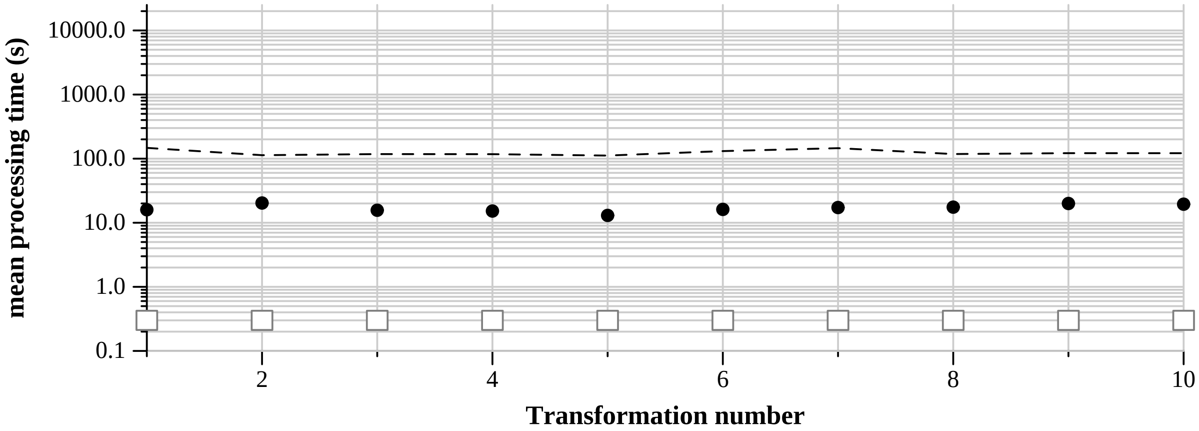
```
TRECVID 2008: copy detection results

Run name:                           INRIA-IMEDIA.v.ViCopT
Run type:                           video-only
```



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

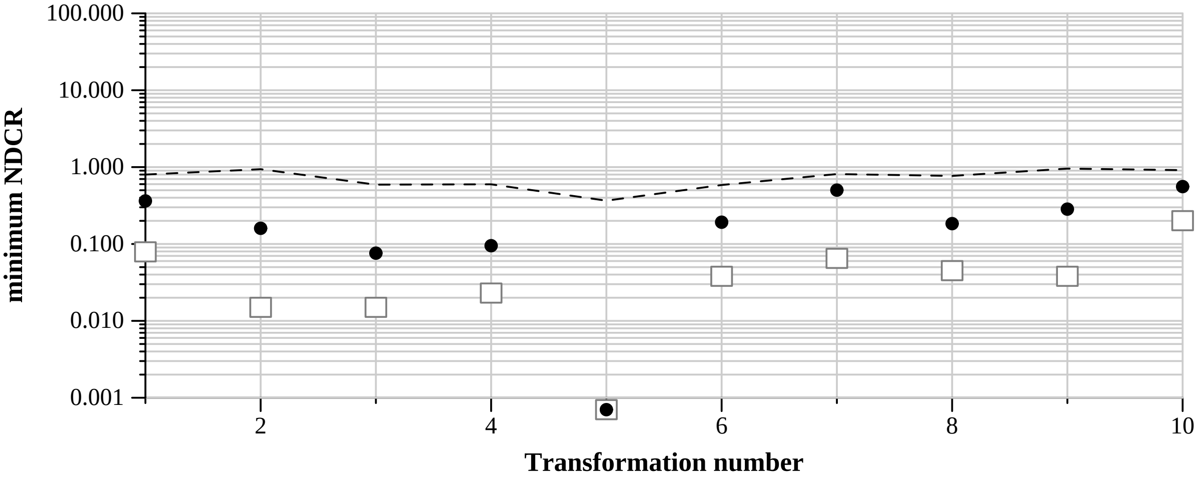


Run score (dot) versus median (---) versus best (box) by transformation

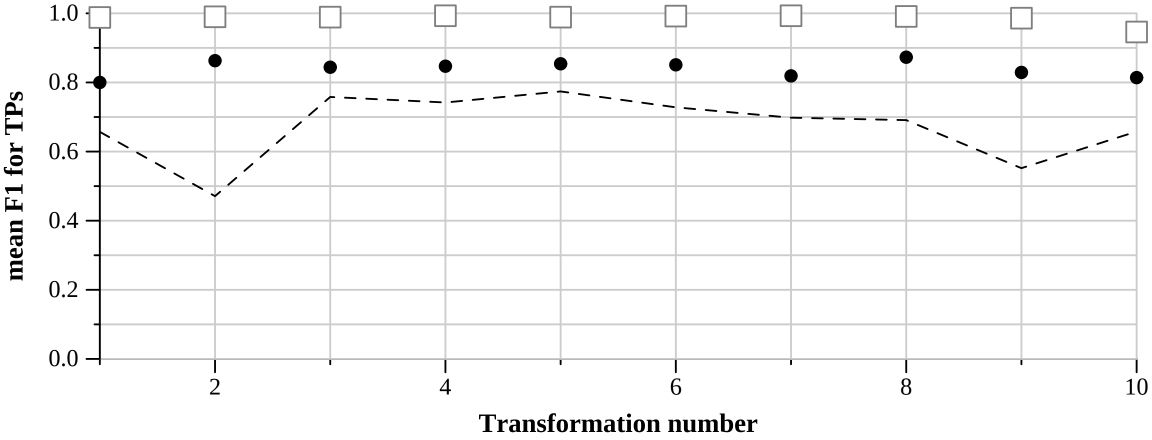


Run score (dot) versus median (---) versus best (box) by transformation

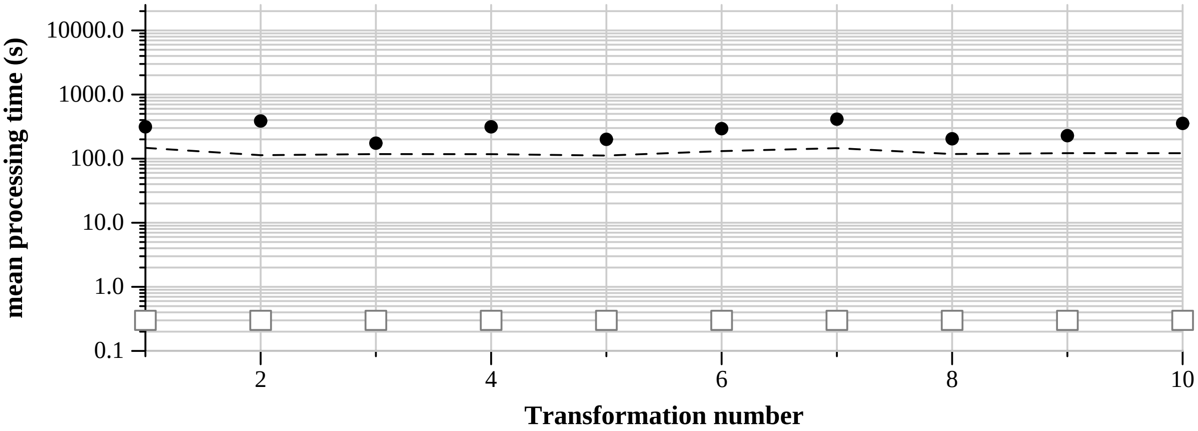
```
TRECVID 2008: copy detection results

Run name:                           INRIA-IMEDIA.v.fusion
Run type:                           video-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