Content-Based Video Copy Detection Benchmarking at TRECVID

George Awad, Dakota Consulting, Inc.
Paul Over, National Institute of Standards and Technology
Wessel Kraaij, TNO, Radboud University

This paper presents an overview of the video copy detection benchmark which was run over a period of 4 years (2008-2011) as part of the TREC Video Retrieval (TRECVID) workshop series. The main contributions of the paper include i) an examination of the evolving design of the evaluation framework and its components (system tasks, data, measures); ii) a high-level overview of results and best-performing approaches; and iii) a discussion of lessons learned over the four years. The content-based copy detection (CCD) benchmark worked with a large collection of synthetic queries, which is atypical for TRECVID, as was the use of a normalized detection cost framework. These particular evaluation design choices are motivated and appraised.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Information filtering, search process; I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision - Applications

General Terms: algorithms, experimentation, measurement, performance

Additional Key Words and Phrases: video copy detection, multimedia, evaluation, TRECVID

ACM Reference Format:
George Awad, Paul Over, and Wessel Kraaij, 2013. Content-Based Video Copy Detection Benchmarking at TRECVID *ACM Trans. Inf. Syst.* 99, 9, Article 99 (Month 2099), 36 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The move from analogue to digital video processing has greatly facilitated copying, reuse and redistribution of video content. One could say that this has caused disruptive changes in many domains of our society. Video has become more important and more pervasive. Nowadays, recording high definition video on a smartphone is doable, whereas such a quality and miniaturization would not have been thought possible at the end of the previous century. Large aggregations and sharing of online video, both commercial and non-commercial, are now commonplace and advances in bandwidth (wired and wireless) make ubiquitous viewing practical and, among many populations, the norm. Assessing whether new video material contains (near) copies of material from a reference video database (perhaps embedded in other non-reference video) is a very important problem in various application settings, such as:

— Tracking the exact number of broadcasts of a certain commercial (airing a commercial is costly, so monitoring is a business goal) [Sadlier et al. 2002] , [Albiol et al. 2004].
— Following an advertisement campaign of competitors as it appears in a number of versions tailored to different timeslots, with different local information overlays, etc. [Huang et al. 2010]. In this case detecting subtle changes in the campaign is of crucial importance.
— Detecting copyright infringement: most video sharing portals have technology in place to detect uploaded videos that are in fact illegal copies. This technology is used as a means to protect against lawsuits e.g. [Breen 2007].
— Detecting video material with abuse-related footage [Eendebak et al. 2008]. Owning copies of and redistribution of this material is a criminal offense in most countries. Note such illicit video will likely *not be watermarked*.

Authors' addresses: G. Awad, Dakota Consulting, Inc., 1110 Bonifant Street, Suite 310, Silver Spring, MD 20910, {gawad@nist.gov}; P. Over, Information Access Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8940, USA, {over@nist.gov}; Wessel Kraaij, TNO, Delft, the Netherlands, Radboud University Nijmegen, Nijmegen, the Netherlands, {kraaijw@acm.org}.

DOI: http://dx.doi.org/10.1145/0000000.0000000

— Improving video search by suppressing or clustering near duplicates on the search engine result page to promote diversity in search results [Liu et al. 2013; Wu et al. 2007].

Liu et al. 2013 discuss at length the varying definitions of "near-duplicate video" (NDV) found in the literature and the relationship of NDV retrieval to copy detection. This paper follows [Law-To et al. 2007b] in defining a video "copy" narrowly as a segment of video derived from another video, usually by means of various transformations such as addition, deletion, modification (of aspect, color, contrast, encoding, ...), camcording, etc. *Content-based copy detection* (CCD) is then the challenge of answering the question whether a video clip contains any (transformed) segment derived from any original reference video, not just segments *of* the same or similar subject. If the answer is yes, than this is relevant information in most of the scenarios sketched before. Note that our scoping of the problem excludes other cases such as the detection that e.g., two news videos have been shot of the same event, but from a different camera angle and with a different camera [Huang et al. 2010]. It also makes no attempt to support research in the use of watermarks to detect copies but rather reflects a scenario in which no mark was added to the original before copies could be made, "the case for a large part of the existing content" according to [Wan et al. 2008].

There are a number of dimensions that are often used when discussing the quality of CCD systems. There is system effectiveness: How good is the system in finding copies? How many (or better, how few) false alarms are raised by the system? There is also system efficiency. For large scale deployment it is important that the system is fast, processing can be distributed somehow and that the data structures used are not too large (since they might need to be transported for evaluating a sample video against a reference database).

TRECVID, the international conference on benchmarking technology for content-based video indexing and retrieval, organized a special Content-Based Copy Detection track in order to facilitate open benchmarking for research groups from academia and industry. The CCD task ran for four years, from 2008 through 2011. The track was started because only a small public benchmark collection was available for evaluating CCD systems at the time [[Law-To et al. 2007b]]. A proprietary benchmark for CCD was under development by Movielabs[1], but this dataset is not public, preventing public scientific research. Movielabs' main interest was to benchmark commercial grade CCD systems for the large movie studios. In 2007, when the track was proposed at the TRECVID workshop, sufficient challenges were identified for advanced CCD systems, that were not met by commercial systems. The main reason was that copy detection of full feature films was not a large challenge, since e.g., a database of shot length sequences is already quite unique [Indyk et al. 1999]. However detecting copies in very short fragments (under one minute, or even seconds) was much more challenging, since less data is available. In addition, the introduction of copy detection technology for user-generated video sharing portals (e.g., Youtube, Vimeo etc) had initiated a challenge to circumvent these systems with new transformations that only marginally affect the perceptual quality of the material. All application settings mentioned before had aspects such as short query length and variation in transformations that limited the effectiveness of commercial CCD systems developed for longer video clips.

The following paper is an introductory overview of the TRECVID copy detection track. It traces the development of the TRECVID evaluation framework from its origins in the Multimedia Understanding through Semantics, Computation and Learning (MUSCLE) evaluation, describes the data used and the test queries created. A general description of the approaches used by the participating research teams is followed by exploratory (graphical) summaries of the main results and reflections on the entire effort. For detailed discussions linking particular algorithms to their performance the reader is directed to papers by the participating research teams, which explain the results in the full context of their experiments.

2. DESIGN OF THE EVALUATION

As mentioned in the previous section, the main goals of the TRECVID CCD were to create an open benchmark for CCD systems that was suitable for the different application scenarios that were identified and which were especially challenging since the copy segments are short.

The MUSCLE evaluation showcase (which took place as a side-event of Computing Machinery's Conference on Image and Video Retrieval (CIVR) 2007 served as a starting point.

[1] www.movielabs.com

2.1. The MUSCLE evaluation

The European Union's Network of Excellence MUSCLE, coordinated by the Institut National de Recherche en Informatique et en Automatique (INRIA) IMEDIA group organized a "video copy detection showcase", with a workshop at CIVR2007 [Law-To et al. 2007b]. Three application scenarios were identified:

— Transformed full-length videos with no post-production and a possible decrease of quality (camcording);
— Short segments on TV streams with possibly extensive large post-production transformations;
— Short videos on the Internet with various transformations (may be extracted from a TV stream);

Looking at the vast variety of possible transformations used in the various application scenarios, it is hard to create a test collection for video copy detection that is representative for real world tasks, for two reasons: i) real applications involve copyrighted material, which makes test set distribution for research purposes very difficult, especially for a government organisation such as the National Institute of Standards and Technology (NIST) and ii) it is hard to gather a representative sample of transformed videos. We would like to mention the availability of the CC_WEB_VIDEO test collection composed by City University Hong Kong and Carnegie Mellon University, which consists of the crawled result pages of 24 queries to popular video search engines. This test collection is available from the organizers and a declaration of non-liability has to be signed to avoid risks due to inadvertent use or redistribution of the copyrighted material in the collection [Wu et al. 2007]

The MUSCLE organizers took a different approach: working with open access video material and a simulation of the copy creation process. Examples of transformations applied are: camcording, overlay of subtitles, color change, blurring, flip, crop and strong re-encoding in a low bitrate format.

These transformations used in MUSCLE – similarly in [Meng et al. 2003],[Ke and Sukthankar 2004],[Chiu et al. 2006], and [Joly et al. 2007] – seem to represent a variety of plausible effects that might realistically follow from re-encoding for alternate platforms/devices, editing for inclusion in other video, or, more freely, for artistic changes to the basic perceptual characteristics of images and video. The authors are unable find any systematic empirical studies describing the nature and number of actual transformations in video copies as defined in this paper; [Kennedy and Chang 2008] analyze Internet image edit differences in real data to create ground truth but do not provide statistics.

The evaluation consisted of two tasks:

(1) **ST1 Video Query** The task was to determine whether a test video was a copy of a video in a reference database of 101 video clips, with a total length of 80 h. The videos were downloaded from the Internet Archive and Open Video project. The test set consisted of 15 videos. Ten videos were copies of reference videos that had been transformed in some way with re-encoding or post-production artifacts. Five videos were not in the reference collection. Systems were evaluated on classification accuracy $C \in \{copy, \textit{non-copy}\}$:

$$Quality = \frac{\#TP}{\#queries} \tag{1}$$

The total length of the test videos was 2 h and 30 min.

(2) **ST2 Video Stream Query** This task modeled a more difficult situation where some reference material has been inserted in non-reference material, where the reference material can be transformed. The objective of this task was to model a video stream (e.g., a TV channel) being monitored for reference content, such as certain TV commercials or re-use of archival copyrighted content. Three video streams were part of this task: i) a non-reference video with short segments of reference videos inserted, untransformed; ii) as in the first condition, but with transformations; and iii) as in the second condition, but with shorter extracted segments. Systems had to detect re-use of reference data and supply the offsets of start and end of the segment at frame level. ST2 used the following metrics[2]:

$$QualitySegment = \frac{\#TP - \#FP}{\#refsegments} \tag{2}$$

[2]interpretation by the authors

$$QualityFrame = 1 - \frac{\#incorrectframes}{\#refframes} \quad (3)$$

The total length of the three simulated query video streams was 15 min.

The evaluation took place 'live' at CIVR2007 (test videos were distributed at the workshop).

The MUSCLE evaluation has the following strong elements:

— Participants received the test queries at the workshop
— Participants did not know the transformations upfront, systems could not be tailored.
— It was easy to evaluate the effectiveness of copy detection systems for different classes of transformations (ST1).
— Using the INRIA parameterized transformation tool it was easy to generate queries.
— ST2 made it possible to compare systems on different levels of difficulty.
— ST2 measured effectiveness, location accuracy (1-proportion of incorrectly classified frames) and processing speed.

The MUSCLE evaluation, positioned as a small-scale show-case, had the following limitations:

— The evaluation was based on simulated, synthetic copies. It was hard to generalize evaluation results to real world settings.
— Using classification accuracy as an effectiveness measure did not allow for individual measurement of precision and recall. In addition, accuracy was not an informative measure in the case of skewed class distributions.
— The reference database was small.
— The number of queries was small.
— The number of non-reference query videos was probably too small for a reliable assessment of the false positive rate in real circumstances.
— The video query length (full clips) was long (ST1)
— The QualitySegment metric (part of ST2) could have negative values.

Five teams participated in the evaluation. The best result for ST1 was a system with a quality score of 0.86, which meant that 2 of the 15 query videos were not correctly classified (no information was available, whether this was due to false positives or false negatives). The processing time was 40 min, meaning 0.26xRT. Especially camcording turned out to be a difficult transformation. The ST2 processing speed was slower, probably due to the location task. The intuition behind the ST2 quality metric seems not completely well defined. It is clear that ST2 can become negative (if the number of false positives is larger than the number of true positives), which probably models the idea that a system that returns more false alarms than true positives is not a useful system. A negative metric is rather unusual when discussing fractions in a contingency table, but it could model a certain usage scenario, however at https://www.rocq.inria.fr/imedia/civr-bench/Results.html, the metric is referred to as "recall". It is clear that ST2 quality is not equal to standard recall, since this would be defined as $QualitySegment = \#TP/\#refsegments$. Perhaps a better name for the ST2 quality metric would have been "discounted recall", or even better, replacing the measure by a combination of precision and recall or the balanced F measure.

Overall, the MUSCLE evaluation showcase achieved its goal. i.e., demonstrating the feasibility of an evaluation framework for video copy detection using synthetic transformations. However the small scale of the evaluation did limit the impact of the conclusions drawn from the experimental results.

For TRECVID CCD, NIST, working with INRIA, decided to create an improved copy detection framework. This meant that some of the design elements of MUSCLE were left unchanged in TRECVID and some elements were adjusted in order to overcome some of the limitations of the evaluation showcase.

The TRECVID CCD task for the systems was as follows: given a test collection of reference videos and a set of short queries (video segments), determine for each query whether it contains any video, with possible audio and/or video transformations, from the reference collection. And if the query contains reference video, locate the start and end of that reference video sequence in the query and the reference collection. Two thirds of the queries contained copies.

As indicated earlier, various considerations made the use of real copies impractical. Beyond the legal problems of using real copies, which likely comprise copyright-protected material, finding such copies in sufficient varieties and quantities presuposes solutions to the very research problems TRECVID

CCD was trying to address! The CCD evalution design did incorporate existing TRECVID test collections of real video created independent of TRECVID and the CCD task. Externally motivated audio and video transformations were applied to randomly selected segments from the test videos to create the video copy queries. These choices enabled a very tight level of control on the video copy test collection (the queries) and by-passed the problem of obtaining a "real" test collection and negotiating the rights for distribution. In addition, TRECVID chose to work with the model of ST2, where a segment of a copy clip was inserted into some non-copied material. Detection whether a full clip was a pirated video could already be sufficiently be addressed by existing methods, some even not requiring sophisticated visual analysis (e.g. [Indyk et al. 1999], [Law-To et al. 2007a]). The challenge of detecting copied material inserted into non-copied material was a much harder one, but had several realistic application scenarios as elaborated in the introduction. The changes that were made to the original MUSCLE design of the task are motivated in the following paragraph. The actual instantiation of the TRECVID CCD over the years is documented in sections 3 and 4.

2.2. Key changes in TRECVID CCD evaluation compared to MUSCLE

From its inception, TRECVID has worked mainly with tasks modeled and evaluated as an information retrieval (IR) problem. That is, each video in the test collection was usually accompanied by a master shot segmentation, effectively transforming the test collection into a collection of shots. The search task and the semantic indexing task (formerly high-level feature detection) were cast as shot ranking tasks, evaluated with standard mean average precision (MAP). Such a framework has many attractive properties, e.g., the unit of retrieval / detection is standardized, allowing pooling of runs, making it feasible to do a reasonable estimate of recall, at least reliable enough to rank systems on their MAP values. For the TRECVID CCD evaluation, we wanted to work with a framework that would support multiple application scenarios with varying density of copies among non-copies and different costs associated with misses versus false alarms. Indeed [Liu et al. 2013] confirm that different application scenarios have a profound impact on system design, notably on the specificity or compactness of video signatures and processing times.

We denoted the prior probability of some test clip being a copy as R_{target}. This prior probability varied significantly across application scenarios. For example R_{target} would be quite high for a news channel (when comparing with the news archive of the channel) and low for a television channel programming classical movies and sitcoms. Ideally we wanted to model different application scenarios given a single test collection. This effectively required R_{target} to be a parameter in the evaluation measure, that can be set independently from the distributed test collection. This requirement ruled out the use of classical IR measures for evaluation, since precision depends on the class distribution. Precision is defined as

$$P = \frac{tp}{tp + fp} \tag{4}$$

where the denominator holds elements from both the target and non target class (true positives and false positives). An alternative was to create a measure using the probability of a missed detection of a copy

$$P_{Miss} = \frac{fn}{tp + fn} \tag{5}$$

and the probability of a false alarm:

$$P_{FA} = \frac{fp}{fp + tn} \tag{6}$$

Both metrics are entirely defined within a single class, i.e., within the class of target items and non-target items respectively. In fact these are class-level error metrics. The advantage of the class level error metrics was that they allowed us to simulate different application scenarios by setting R_{target}. Given this advantage, we chose to reuse and adapt the normalized detection cost framework that had been developed in the topic detection (TDT) benchmark [Fiscus and Doddington 2002].

2.3. Metrics

2.3.1. Normalized detection cost rate (NDCR). To simplify the evaluation, we made the following assumption: a video test query either contained no copied material, or just a single copied segment was inserted into non-copied background footage. Using a single copy but of varying size and position simplified the

evaluation while keeping it challenging. Query sets were created for a series of transformations (cf. 4). This allowed for modeling detection as a binary classification task, i.e. calculating P_{Miss} as defined above. While one can imagine applications requiring much less or much more overlap, as an initial compromise, a clip was considered correctly identified if at least 50 % of the reference copy extent overlaps with the asserted copy expressed by a system. For CCD, the *probability* of a false alarm was replaced by a false alarm *rate*, defined as follows:

$$R_{FA} = \frac{fp}{T_{queries}} \tag{7}$$

This means the number of false positives per hour as measured on the query collection; $T_{queries}$ being the total duration of all video queries.

The two ingredients P_{miss} and R_{FA} were combined in the so-called detection cost rate DCR as follows:

$$DCR = C_{miss} P_{Miss} R_{target} + C_{FA} R_{FA} \tag{8}$$

The constants were defined as follows:

$$C_{Miss} = \textit{cost of a missed detection} \tag{9}$$
$$C_{FA} = \textit{cost of a false alarm} \tag{10}$$
$$R_{target} = \textit{a priori target rate (\#/hour)} \tag{11}$$
$$\tag{12}$$

In order to compare the detection costs across a range of values of the cost parameters and target rate, we defined the *normalized detection cost rate* as follows:

$$NDCR = \frac{C_{Miss} P_{Miss} R_{target} + C_{FA} R_{FA}}{C_{Miss} R_{target}} \tag{13}$$

The NDCR was computed for each individual query set, corresponding to a given transformation. (Although NDCR via its parameterization supports a variety of application scenarios, comparison of results across studies beyond TRECVID would require comparable parameter settings [Bailer 2010].)

In the first (pilot) year of the CCD task, a single application scenario was tested with the parameters of

$$R_{target} = 0.5/h, C_{Miss} = 10, C_{FA} = 1 \tag{14}$$

simulating forensic applications which need a high cost for missing a target video. In the subsequent years, two application scenarios were tested reflecting two sides of the spectrum. One scenario "No false alarms" (Nofa) had a very high cost for false alarms (since it created extra work and risk of damaged reputation), whereas the second scenario "Balanced" (Bal) concerned an equal cost for missing a target video vs detecting a non-target video. In 2009 the parameters for Nofa where as follows:

$$R_{target} = 0.5/h, C_{Miss} = 1, C_{FA} = 1000 \tag{15}$$

while for Bal profile:

$$R_{target} = 0.5/h, C_{Miss} = 1, C_{FA} = 1 \tag{16}$$

In 2010 and 2011 only R_{target} was changed to $0.005/h$ to reflect more rare copied videos while the cost parameters C_{Miss} and C_{FA} where the same as in 2009.

The parameterized evaluation measure provided a lot of flexibility in modeling a certain application scenario, but was less intuitive from an IR point of view, since it was not a very common measure nor are the authors are aware of any current trend among IR researchers to adopt it.

The detection effectiveness was measured for each individual transformation. For each run, all results of individual transformations were concatenated in separate files and sorted by decision score. Subsequently, each concatenated file (corresponding to a single transformation across all queries from a given run) was used to compute the probability of a miss error and the false alarm rate (P_{Miss} and R_{FA}) at different operating points, by truncating the list at a range of decision thresholds θ, sweeping from the minimum decision score to the maximum score. As a first step, asserted copies that overlap were logged and removed from consideration. Secondly, the computation of true positives was based on

only one submitted extent per query (as defined by the mapping procedure using the Hungarian solution of the bipartite graph matching problem [Kuhn 1955]). All other submitted extents for this query counted as false alarms. This procedure yielded a list of pairs of increasing P_{Miss} and decreasing R_{FA} values. These data points were used to create a P_{Miss} versus R_{FA} error plot (Detection Error Tradeoff (DET) curve) for a given run and transformation. The two error rates were then combined into a single normalized detection cost rate, NDCR, by assigning costs to miss and false alarm errors as in equation (13).

2.3.2. Location accuracy. As a secondary measure, NIST also evaluated the copy location accuracy. It aimed to assess the accuracy of finding the exact extent of the copy in the reference video, once the system has correctly detected a copy. The asserted and actual extents of the copy in the reference data were compared using precision and recall and these two numbers were combined using the F1 measure. Recall and precision were measured at the optimal operating point where the normalized detection cost was minimal.

2.3.3. Detection processing time. A third measure was the mean time (in seconds) to process a query. Processing time was defined as the full time required to process queries from MPEG video files to result file (including decoding, analysis, features extraction, eventual write/read of intermediate results, eventual loadings of reference subsets, results output). Participants were asked to submit the mean time which is the full processing time normalized by the number of queries.

In the task guidelines participants were asked to also submit the type of operating system they used along with the available resources such as CPU and memory. For a detailed analysis about any correlation between the used resources and the performance of any given system we recommend the reader to check the detailed paper of the submitted approach by that research group.

3. DATA

Two main sources of video were used in the four years of TRECVID copy detection evaluations. Reference video in 2008 and 2009 came from the Netherlands Institute for Sound and Vision, which generously provided news magazine, science news, documentaries, educational programming, and historical video in MPEG-1 for use within TRECVID. This was professionally recorded and edited material, adhering to standards for Dutch public television programming. The non-reference video in 2008 and 2009 came from a collection of BBC rushes - professionally shot but unedited raw material for several BBC programs.

In 2010 and 2011 reference and non-reference video both came from a set of videos in MPEG-4/H.264 that was downloaded from the Internet Archive in 2009. These were limited only in that they were available for use in research under a Creative Commons license and were relatively short. This material, uploaded by a vast variety of independent Internet users, exhibits significant diversity in subject, language, capture device, editing style, etc. It is referred to as the IACC.1 collection. Part of the set was divided randomly into 3 disjoint subsets of approximately equal duration - IACC.1.A, IACC.1.B, and IACC.1.C - and a smaller training set: IACC.1.tv10.training. The remainder was set aside for use as non-reference video.

Whether IACC.1 videos differ more from each other than the Sound and Vision videos differ from the BBC rushes is an interesting question, relevant as one factor in the interpretation of results from 2008-9 versus those from 2010-11. It remains as yet unanswered.

In 2008, the copy detection task used 200 h of Sound and Vision video as test reference data: tv7.sv.devel (50 h, 110 files), tv7.sv.test (50 h, 109 files), and tv8.sv.test (100 h, 219 files). BBC rushes video (35 h) was used as non-reference data. For development data, the copy detection task used the MUSCLE-VCD-2007 data [Law-To et al. 2007b].

The test reference data for 2009 comprised 180 h (400 files) of new Sound and Vision video (tv9.sv.test) plus the 200 h used in 2008. BBC rushes (83 h) served as the source for the non-reference video. The 200 h of Sound and Vision video from 2008 were also used for development in 2009. The overlap of development and test reference data, although unusual, was felt to be acceptable since queries were chosen at random from the reference data, all the reference data sets were from the same source, and nearly doubling the size of the test reference data was felt to be a valuable step toward more realism.

In 2010 the reference data for testing comprised 400 h (11 200 files): IACC.1.A and IACC.1.tv10.training. The non-reference data was not identified other than to say it was drawn from

Internet Archive videos available under Creative Commons licenses - a set comprising 4000 h, 12 480 files, with durations between 10 min and 30 min. For development, the reference video and copy detection queries used in TRECVID 2009 were reused: tv9.sv.test, tv7.sv.devel, tv7.sv.test.

For testing in 2011, the reference data was identical to that used in 2010. The non-reference data was not identified to participating researchers other than to say it was drawn from Internet Archive videos available under Creative Commons licenses. Since the query set was large and randomized, reusing the test data allowed comparison of 2011 systems to those from 2010. For development, the reference video and copy detection queries used in TRECVID 2009 were available - as in 2010.

4. QUERIES

The query creation procedure followed during TRECVID 2008-2011 was built on the work done during the first live benchmark initiative for video copy detection that took place during CIVR2007 and supported by the MUSCLE network of excellence. Queries were designed to simulate real world use cases for copied videos either over the Internet or from TV streams. In general, three types of queries were proposed and various types of video and audio transformations were applied to them to simulate real world conditions of copied video such as quality change, camcording or postproductions.

In 2008, only submissions using video-only queries were required while testing on audio-only and audio+video queries was optional. The initial thinking was that videos often contain audio, sometimes the original audio is retained in the copied material, sometimes it is replaced by a new soundtrack. Nevertheless, audio is an important and strong feature for some application scenarios of video copy detection. Since detection of untransformed audio copies was relatively easy [Baluja and Covell 2007], and the primary interest of the TRECVID community was in video analysis, it was decided to model the required task with video-only queries. However, since audio was of importance for practical applications, there were two additional optional tasks using transformed audio-only queries and using transformed audio+video queries. In 2009 participants were required to submit at least two runs using video-only queries and two using audio+video queries reflecting the focus on video analysis but a desire to know how much audio could contribute. Finally, in 2010 and 2011 the importance of the using both audio and video was established and only one sort of query was tested: audio+video.

4.1. Query types

Before queries were generated, 2 video datasets were identified (reference and non-reference) by NIST where copied videos were drawn only from the reference dataset. Participants had no knowledge of the non-reference dataset. Three main types of queries were designed to reflect real world case scenarios:

— Type-1 where the whole query video was selected from a video segment in the reference dataset.
— Type-2 where a segment from a video in the reference dataset was inserted inside a video segment selected from the non-reference dataset.
— Type-3 where the whole query video was selected from a video segment in the non-reference dataset.

Type-1 queries simulated videos generated by camcording movies while type-2 could simulate postproduction of videos in a TV stream. Finally type-3 queries were not copies of any videos and thus the three types together represented to a large extent the types of videos that a real video copy detection system might face.

Three run types were proposed for evaluation based on the modalities used. Those were audio-only, video-only, and audio+video runs. NIST started by generating a set of base queries based on the three above query types (1, 2, & 3) and then audio-only queries are generated by stripping out the video from the base query, video-only queries are generated by stripping out the audio from the base query, while audio+video queries staid identical to the base queries.

We should note here that base queries of type 1 and 2 were checked to make sure they don't include any repeated or self duplicate video segments as the task simulated a single copied video segment per query. And any reported query that included duplicate segments was removed from the evaluation framework. In real-world search engines, near duplicates can affect heavily the retrieval performance as mentioned in [Wu et al. 2007] where a study shows that 27 % of videos retrieved by common queries against Google, Yahoo, etc. were near duplicates.

However, during the course of running the CCD task for 4 years not all modality-based queries were evaluated mainly because the focus was more on the video modality and more importantly how much the audio detection could boost the performance of the audio+video queries. The next step of the query

generation process was to apply some audio and video transformations on the base queries to simulate real-world conditions of copied videos. The next section explains the various types of transformations used during the 4 year evaluation cycle.

4.2. Query Transformations

4.2.1. Audio Transformations. Audio transformations for audio-only queries were generated by Dan Ellis at Columbia University along the same lines as the video-only queries: an audio-only version of the set of base queries was transformed by seven techniques that were intended to be typical of those that would occur in real reuse scenarios: (1) bandwidth limitation (2) other coding-related distortion (e.g., subband quantization noise) (3) variable mixing with unrelated audio content. The chosen 7 transformations were as follows:

— T1 - Same as original audio
— T2 - mp3 compression
— T3 - mp3 compression and multiband companding
— T4 - bandwidth limit and single-band companding
— T5 - mix with speech
— T6 - mix with speech, followed by multiband compress
— T7 - bandpass filter, mix with speech, then compress

4.2.2. Video Transformations. Video transformations for video-only queries were generated by tools developed by the IMEDIA team at INRIA. Two main categories of transformations with random parameters were adopted: individual transformations and combinations of more than one transformation as follows:

— T1 - Camcording: this transformation simulated filming a movie on a screen using different random angles such as camera facing the screen, camera with a small angle and camera with a large angle.
— T2 - Picture in picture type 1: the untransformed base video query was inserted in front of a background video with a change of the scale and position parameters (Fig. 1).
— T3 - Insertion of patterns: different patterns were inserted randomly such as captions, subtitles, logo, and sliding captions (Fig. 2).
— T4 - Strong re-encoding: the resolution and bitrate of the video was changed and at the end the video was re-encoded in MPEG-1 (Fig. 3).
— T5 - Change of gamma: the gamma value for each color was changed randomly between 0.3 to 1.8 (Fig. 4).
— T6 - Decrease in quality: this included choosing randomly 3 transformations from the following: Blur (value from 1 to 3), change of gamma (T5), frame dropping (frequency value from 0.05 to 0.2), contrast (value from 0.7 to 1.3), compression (T4), ratio change (letterbox value from 0.75 to 0.9), and white noise (value from 0.1 to 0.2) (Fig. 5).
— T7 - Decrease in quality: same as T6 except that 5 random transformations were chosen.
— T8 - Post Production: this included choosing randomly 3 transformations from the following: Crop (number of pixels from 5 to 20), Shift (deltaX and deltaY values from -50 to 50), Contrast change, caption (text insertion in different colors), flip (vertical mirroring), Insertion of patterns(T3), Picture in Picture type 2 (the original video is in the background) (Fig. 6).
— T9 - Post Production: same as T8 except that 5 random transformations were chosen.
— T10 - Randomly chose 1 transformation from each of (T1 to T5), Decrease in quality(T6 to T7), and Post production categories(T8 to T9).

Across the 4 year evaluation cycle all transformations were applied except T7 and T9 which were removed after 2008 at the suggestion of participating researchers who felt the distortions created by the transformations were so great that the resulting copies were not likely of realistic utility. In total, each year NIST generated a set of 201 base original untransformed queries. From those the audio and video were stripped out to generate an original 201 audio-only and 201 video-only query sets. Each of the 201 audio-only queries was transformed by the 7 audio transforms yielding total of 1407 audio-only queries and each of the video-only queries was transformed by the 10 video transforms yielding total of 2010 video-only queries. To create the audio+video queries a script was given to participants to run and generate 14 070 queries such that each original base query was transformed 7x10 times to include all the combinations of audio and video transforms.

Fig. 1: Sample of video transformation “Picture In Picture” type 1 - based on material provided for research and copyrighted by the BBC

Fig. 2: Sample of video transformation “Insertion of Patterns” - based on material provided for research and copyrighted by the BBC

Fig. 3: Sample of video transformation “Re-Encoding” - based on material provided for research and copyrighted by the BBC

Fig. 4: Sample of video transformation “change of Gamma”- based on material provided for research and copyrighted by the BBC

Fig. 5: Sample of video transformation decrease of quality “Ratio”- based on material provided for research and copyrighted by the BBC

Fig. 6: Sample of video post production transformation “text insertion”- based on material provided for research and copyrighted by the BBC

5. APPROACHES

This section summarizes by year the approaches taken by most participating teams and then in more detail those employed by one or more top teams (those who achieved minimum NDCR across the majority of transformations) across the four years. The descriptions merely serve as pointers to published work of TRECVID CCD participants. A more comprehensive introduction into system architectures for content based copy detection is provided in [Liu et al. 2013]

5.1. 2008

Generally techniques evaluated in 2008 can be divided into transformation-specific or more generic techniques. The most used features were scale-invariant feature transform (SIFT) descriptors, block-based features and edge histograms.

5.1.1. INRIA-LEAR. Most components of the INRIA-LEAR system were derived from their image search engine. They used uniform sampling to sample query frames and a different asymmetric sampling strategy to sample the dataset videos keyframes. After extracting keyframes they used Hessian-affine invariant region detectors to extract regions of interest and then they used SIFT descriptors to represent them. K-means was used to generate a visual vocabulary by quantizing the SIFT descriptor space. To further filter out non-matching descriptors, they applied hamming embedding to encode the descriptors via binary signatures which were used together with the hamming distance and geometric consistency constraints to generate a candidate set of video segments for each video query. For details see [Douze et al. 2008] and [Douze et al. 2010].

5.2. 2009

In 2009 different approaches included processing speed optimization using graphical processing unit (GPU) based local feature extraction, fusion of frame fingerprints such as SIFT descriptors, block-based features, and global features. There were some transformation-specific approaches and the combination of audio and video used linear combinations or binary fusion methods.

5.2.1. AT&T. The AT&T team started by segmenting the videos of both reference and queries into shots using a shot boundary detection algorithm and took the first frame of each shot as a keyframe. For the query videos they tried first to detect some transformations (to reduce their effect) by presenting such as letterbox and picture in picture (PIP) using edge detection image profiles and pixel intensity variance techniques. Also to remove the gamma change and white noise they applied equalization and blurring and finally as local features were not invariant to mirroring, they generated a flipped version from all queries. In total for each query keyframe, they generated nine versions of it after removing some transformations effects. For the reference keyframes, they generated a strong encoded and half-scale resolution versions for each keyframe to be able to compare with PIP queries and strong encoded query keyframes. After preprocessing the keyframes, they calculated the SIFT features on each and used Local Sensitive Hashing (LSH) to index and search the SIFT features in a scaled and robust way. As a second step of filtering the matched keyframes based on SIFT and LSH they used Random SAmpling Consensus (RANSAC) to refine the matched keyframes and gave each matched keyframe a score. Finally they merged the results of the matched keyframes across the various versions of query keyframes they generated and merged the result at a video-level to rank the matched reference videos for each query before they normalized their scores. For details see [Liu et al. 2009] and [Liu et al. 2010].

5.2.2. Computer Research Institute of Montreal (CRIM). The participants at CRIM submitted video-only, audio-only, and audio+video runs. For video-only runs they started by segmenting the video files into shots to extract from each shot representative keyframes. From each keyframe they found regions of interest using difference of Gaussians (DOG) point detector and calculate SIFT descriptors on them. Each descriptor was composed of 16 gradient histograms from 16 different regions. They defined a vocabulary set to represent each region where each SIFT descriptor was projected into the vocabulary space and its nearest neighbors were found. A binary code was given for each vocabulary cluster to represent if it was close enough to the descriptor gradient histogram using the L1 distance. All descriptor clusters were represented in a hierarchical structure to speed computations. A group of local matches between keyframes was extracted and Bag of Visterms (BOV) was used for representation. Then a Latent Dirichlet Allocation (LDA) generative model was applied to provide a discrete discriminant analysis over matches. To further filter descriptors they used bag-of-words (BOW) models to eliminate the more common local descriptors which were not discriminative enough such as straight lines or cor-

ners found in many images. Finally, RANSAC was used in the temporal domain to estimate the time shift and dilation between time codes of detected keyframe link and the time range of copied segments was found from the shot boundaries from which the keyframes belong to. For audio-only runs they computed audio fingerprints of the audio queries using two features. The first feature was created by calculating the energy differences in consecutive sub-bands. Fingerprint matching was done by moving the query over the test dataset audio fingerprints and counting the total matches (count/sec) for each alignment of the query with the test. The second feature mapped each frame of the test to the closest frame of the query. To measure the closeness they computed 12 cepstral coefficients and normalized energy and its delta coefficients. The distance between the test frame and query frame was defined as the sum of the absolute difference between the corresponding cepstral parameters. To speedup computations they used GPUs. Finally to submit audio+video runs they merged the output of audio-only and video-only runs. When there was an overlap between the results they took the start and end of the copied segment from audio-only run and used a weighted addition of the two scores to compute the confidence value. If there was no overlap between audio and video results they selected the test result with the highest weighted score. For details see [Héritier et al. 2009] and [Gupta et al. 2012].

5.2.3. MCG-ICT. Researchers in the Multimedia Computing Group at the Institute of Computing Technology of the Chinese Academy of Sciences in Beijing decomposed the transformations into four main categories and developed for each category a separate module to solve the challenge. The four modules were global block gradient histogram technique to capture global features, local feature extraction module that used spatial context of salient points based on a new interest point detector called Harris-Hessian, a specialized module for detecting picture in picture queries using edge detection, and a module for detecting flipped queries which used flipped detected features to match videos. Each module worked in parallel on the queries and used consistency time check method to enhance the localization of the detection. They applied two methods of fusion: hierarchical and non-hierarchical to fuse the results of the four modules. Finally to speed up the overall system they used GPUs to calculate the local features which helped to reduce greatly the processing time up to 20 times. For details see [Zhang et al. 2009] and [Xie et al. 2012].

5.3. 2010

Some submissions used only the video modality (e.g., IBM, Nanjing University, National Taiwan Normal University, Univ. of Chile, City University of Hong Kong) while audio modality helped to reduce the false alarm rate for picture in picture video transformations. Most teams fused audio and video at the decision level. Queries with short copied segments tended to be missed. The most popular features used were SIFT, speeded up robust features (SURF), direction-adaptive residual transforms (DART), color, texture, and edge histograms for video and Mel-frequency cepstrum coefficients (MFCCs) and weighted advanced stability feature (WASF) for audio features. Bag of visual words based techniques were the most popular approaches reported.

5.3.1. PKU-IDM. At Peking University they started by preprocessing the queries and reference videos by splitting the video and audio components. Visual keyframes were uniformly sampled at rate of 3 frames per second while audio frames were obtained by dividing the audio signal into uniform durations of 60 ms with 40 ms overlap. Additional preprocessing was done to detect PIP queries using hough transform to detect parallel lines then extract the background, foreground and original keyframes. After extracting the keyframes, four basic detectors were employed upon 2 local visual features, one global visual feature and one audio feature. The local feature detectors used were SIFT and SURF. SIFT was calculated and quantized using k-means along with its scale, orientation, and spatial information. Such integrated information was then stored in an inverted index that was searched by queries after their SIFT features are extracted. They proposed a global image feature based on the relationship between the DCT (Discrete Cosine Transform) coefficients of adjacent image blocks. First the keyframes were normalized to a 64x64 pixel image and converted to YUV color space keeping the Y channel only. Then the Y-channel image was divided into 64 blocks with size of 8x8 pixels each and 2-D DCT was applied over each block to obtain a coefficient matrix. The energies of the first 4 subbands of each block were computed by summing the absolute values of each subband and finally a 256-bit DCT feature was computed by comparing the energies of each band between adjacent blocks and hamming distance was used as the distance metric. The WASF (Weighted ASF) was used as the audio feature which extends the MPEG-7 descriptor Audio Spectrum Flatness (ASF). Both global and

audio features were indexed by LSH. Each detector returned top 20 matched reference keyframes from which a Sequential Pyramid Matching (SPM) was applied on the temporal video space to accumulate the weighted similarities from multiple levels. Finally, a result level fusion was utilized to fuse the detection results from different detectors. For details see [Li et al. 2010] and [Tian et al. 2011].

5.4. 2011

Teams focused more on enhancing their features rather than efficiency. Major features used were SIFT, SURF, LBP (Local Binary Patterns), color histograms, MFCC, and Energy Difference Fingerprint (EDF) / Cepstral (CEPS) audio features and BOW models. picture in picture detectors were utilized for query preprocessing. Many teams borrowed techniques from image recognition based on global features and new teams started to use audio features.

5.4.1. PKU-IDM. The PKU system was almost the same as in 2010 using three complementary features except that they replaced the two local features by only dense color SIFT (DCSIFT) where they used dense sampling to extract feature points and apply BOW by k-means to quantize DCSIFT descriptors together with their spatial information. DCT and WASF remained the same features for global image and audio information. Temporal Pyramid Matching (TPM) was used to fuse the individual detector frame level matches into a video level matches. In order to make their system more efficient and faster they used a cascade architecture where they first used the WASF feature individually to determine if the query is a copy or not. If the query was not a copy it was passed to the second layer which was the DCT detector and only declared as a copy if the DCT found a match, otherwise it was finally passed to the DCSIFT as the final layer. For details see [Jian et al. 2011] and [Jiang et al. 2012].

5.4.2. CRIM. In 2011 the CRIM team decided to apply their audio detection method of nearest neighbor matching for audio queries they used in 2009 on the 2011 video-only queries. They used two methods to calculate video fingerprints. First method divided the frame into 16 sub-squares and for each square calculated the average red-green-blue (RGB) pixel value. These values went through local temporal normalization in a window of 10 frames and the top seven values based on maximum deviation from the mean were selected for coarse quantization between zero and five. The second method used the 16 normalized values per frame as described above but without quantization. To find the test segments they used the count of matching fingerprints of the aligned segments as a confidence value for each possible test segment match. The second method they used to search for test segments employed nearest neighbor by computing the absolute sum between test and query frame fingerprints for each color RGB space separately then summed to get the final count. For audio-only detection they enhanced the features by using MFCC, equalized MFCC and Gaussian MFCC. They first did a fast search using energy difference fingerprints to narrow the results to few choices which were then rescored using the nearest neighbor fingerprints derived from cepstral features and its first differences. They used a voice activity detector to skip the silent segments in the matching process. To submit their audio+video runs they combined the video-only and audio-only results such that they gave higher priority for query results that have overlapping test segments from both modalities compared to results from only one modality. If both modalities gave non-overlapping results they only submitted the results of one of them to reduce the NDCR. For details see [Gupta et al. 2011] and [Gupta et al. 2012].

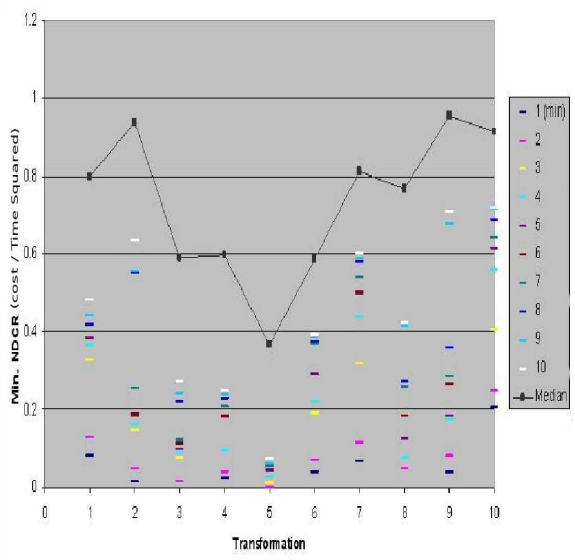
6. EVALUATION

6.1. Evolution of the system task

The basic task was enhanced with various conditions as the evaluation matured - in an effort to learn more about system success and failure and better simulate some real world copy detection needs. In the first year, 2008, participating researchers could build systems to detect copies using video-only (V), audio-only(A), or audiovisual (AV) methods. There were in fact 48 V submissions, 1 A, and 6 AV.

In 2009, participants were required to submit V and AV runs (output of a system against the test data in a given experimental configuration); “A” runs were optional. Submissions were as follows: 53 V, 12 A, 42 AV. Since different applications place different requirements on systems, two different application targets or “profiles” were introduced in 2009. The Nofa profile required the system to avoid false alarms if at all possible; the cost metric set the cost of a false alarm to be very high. The Bal profile set the cost of a false alarm and the cost of not detecting a copy to be equal. Systems were also required to set a decision score threshold - a cutoff point in their ranked output for each query above which the user should believe the query was a copy and below which it was not.

Fig. 7: TRECVID 2008: Min NDCR (Top 10) by video transformation for video-only runs



By 2010, since using of audio and video seemed in general to outperform audio-only and video-only approaches, it was decided only AV submissions would be accepted. Seventy-eight AV runs were submitted; 41 using the Bal profile and 37 using the Nofa. Under the same rules in 2011, 73 runs were submitted, including 41 using the Bal profile and 32 using Nofa.

Due to the fact that the testing data used between 2008-9 are different than the testing data used between 2010-11 we can not conduct cross year analysis from 2008 to 2011. Instead we tried to compare systems in 2010-11 because reference and non-reference datasets were the same and both had systems submitting runs in both application profiles.

As discussed in the design section, participants were asked to submit with their runs some information about used computing resources such as platform,cpu and memory as those can obviously affect measures such as the speed of detection. Across the full 4 years we found that 55 % of submitted runs used Windows operating system while 45 % used Linux and only 3 % used Mac. In regard to memory resources, 90 % of runs used from 1GB to 256GB of memory with average of 25GB while 10 % used memory between 4MB to 512 MB with average of 300MB. Finally, a broad range of CPU models were used by systems with the majority using multiprocessors ranging between 10 to 200 processors and few (about 4 %) actually used GPUs (Graphical Processing Units).

The following sections address some of the results in light of a number of research questions. The questions vary some from year to year as a function of the experimental design and participation.

6.2. 2008

How did systems perform on each of the 3 measures (by transformation)? Figures 7, 8, and 9 present the best results separately for the three main measures of detection (minimum NDCR), localization (F1), and speed (processing time), respectively. The figures of the F1 and cost measures show there was a noticeable spread among the top 10 performance for almost all of the transformations. The one exception was transformation 5 (change of gamma) which also achieved the minimum cost among all transformations. Regarding the processing time, the top 10 achieved maximum about 20 s. We should note here that we are reporting the time measurements that participating research groups submitted to us and have no access to the uncertainties involved in those measurements. Figure 10 shows the percentage of submitted items that were false alarms for the top runs. Some systems achieved very low false alarm rates (reaching 0).

How are the main measures correlated? Figures 11, 12, and 13 show the relationship between the three main measures. Note that the vertical and horizontal axes in transformation T1 is exactly the same in all the other transformations. Figure 11 plots the relationship between minimum NDCR and

Fig. 8: TRECVID 2008: F1 (Top 10) by video transformation for video-only runs

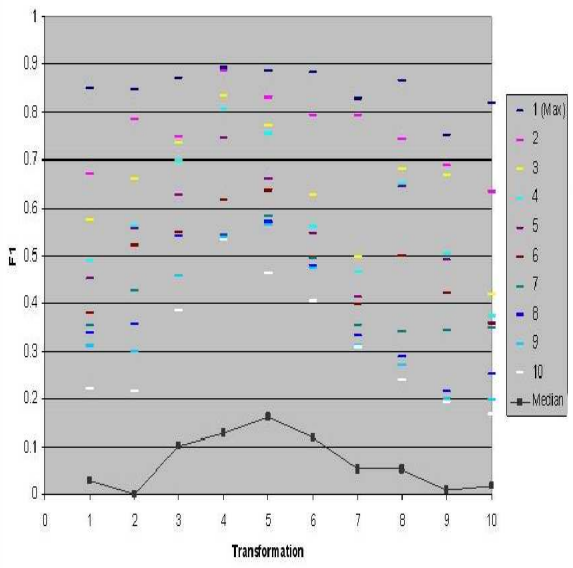
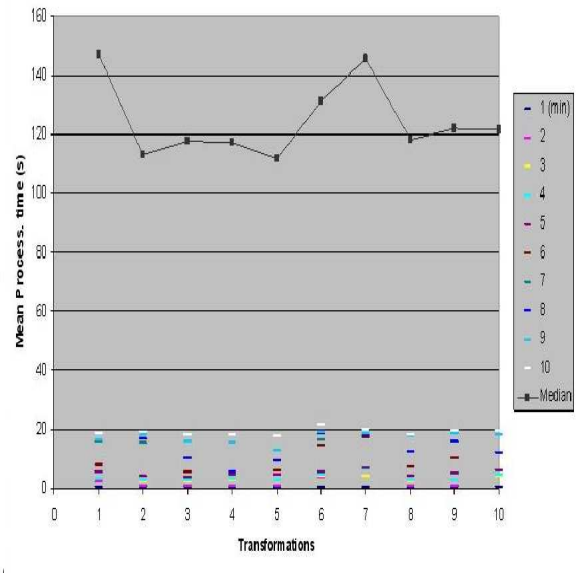
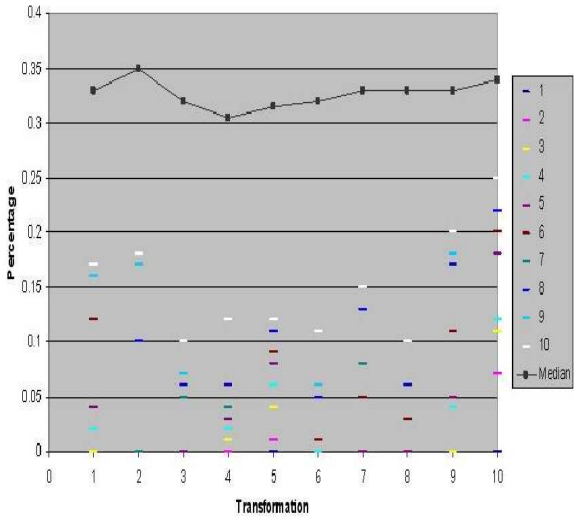


Fig. 9: TRECVID 2008: Processing time (Top 10) by video transformation for video-only runs



F1 for each video transformation. There appears to be little correlation between systems that were good in separating copies from non-copies (low NDCR) and those also good in localization. Transformation 10 probably made it hardest to detect copies. This may be justified by the fact that transformation 10 is a combination of 5 transformations. Similarly, Figure 12 graphs F1 versus processing time. Increasing processing time did not enhance localization. Only a few systems achieved high localization in small processing time. Figure 13 compares minimum NDCR against processing time. Increasing the processing time did not reduce the cost or make the systems stronger. Few good systems were fast with low cost. In general these patterns of correlation were similar to what was observed between 2009-2011 for both Bal and Nofa application profiles: For the systems as evaluated, increasing the processing time

Fig. 10: TRECVID 2008: False alarms (Top 10) by video transformation for video-only runs



didn't help in either enhancing localization or reducing detection cost for the majority of systems. Top systems which were good in detection were also good in localization.

How did video-only compare in detection to audio+video? Figure 14 presents the best minimum NDCR scores for the audio+video queries for each combination of the audio (AT1, AT2, ...) and video transformation (VT1, VT2, ...). For purposes of rough comparison, it also shows the scores for the best video-only queries. As the number of submitted audio+video runs was too limited, we cannot draw general conclusions. However, the relative effect of audio transformations seems similar across video transformations; it seems that using audio (when no speech is mixed in) helped to decrease the cost across transformations compared to using only video (except in video transformation 5). We should note here that after obbserving the effects of audio transformations across the 4 years it seems that audio transformations 5,6, & 7 (mix with external speech) had a consistent effect of making detection harder on a+v runs.

6.3. 2009

Results presented here are for each query type (audio-only, video-only and audio+video) separately. For each query type we present the results of the two application profiles (balanced and no false alarms) based on the submitted actual run threshold and based on the optimum threshold. Note that systems chose one actual threshold per submitted run to be used in the evaluation of all queries (i.e., transformations). In contrast, the optimum threshold was determined separately for each transformation, post-submisssion by NIST using ground truth.

How did the different modalities compare in detection? Comparing the top runs' NDCR scores per transformation for the three query types we found that for the audio-only queries, systems achieved good detection (less than 0.1 NDCR) across all transformations in the two profiles and in the actual as well as optimum results. This might be due to the fact that the audio detection techniques were more mature and advanced than video detection. Video-only queries achieved a worse performance than the audio-only as NDCR scores ranging across transformations vary a lot and reached above 0.9. This indicates that systems had difficulties with some transformations compared to audio-only scores. Top scores on video+audio queries were consistently much better than on video-only queries across all transformation combinations; they didn't exceed the 0.1 NDCR in the two profiles using both the actual and optimum thresholds. This again indicates that the audio feature helped in video copy detection.

For the balanced profile, how did the different modalities compare in detection? Figures 15 through 17 show the performance of the top 10 runs for the three query types for the balanced profile. It is clear that there was a difference between the optimum median and the actual median (medians are across

Fig. 11: TRECVID 2008: Relationship between F1 and cost across video transformations

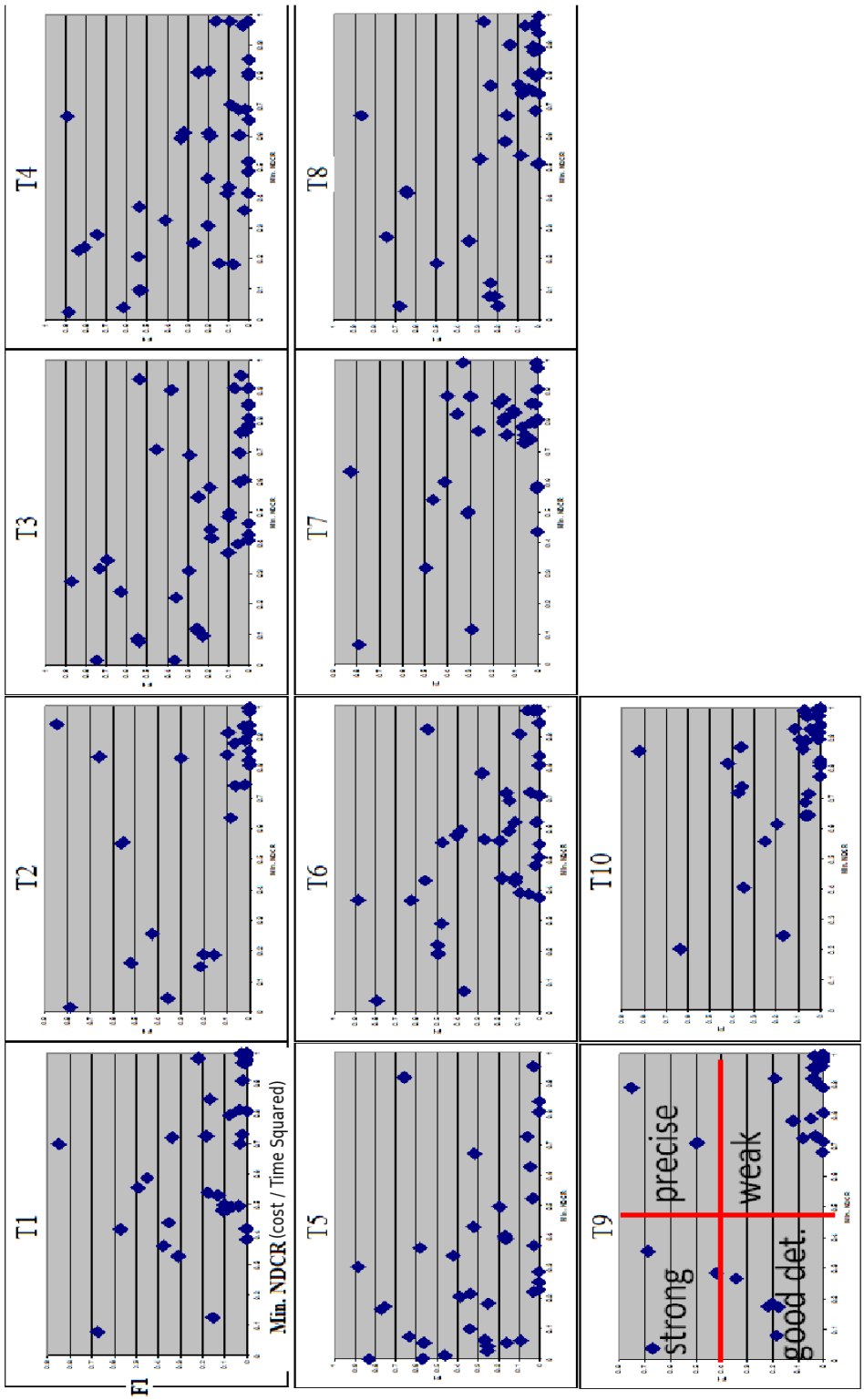


Fig. 12: TRECVID 2008: Relationship between F1 and processing time across video transformations

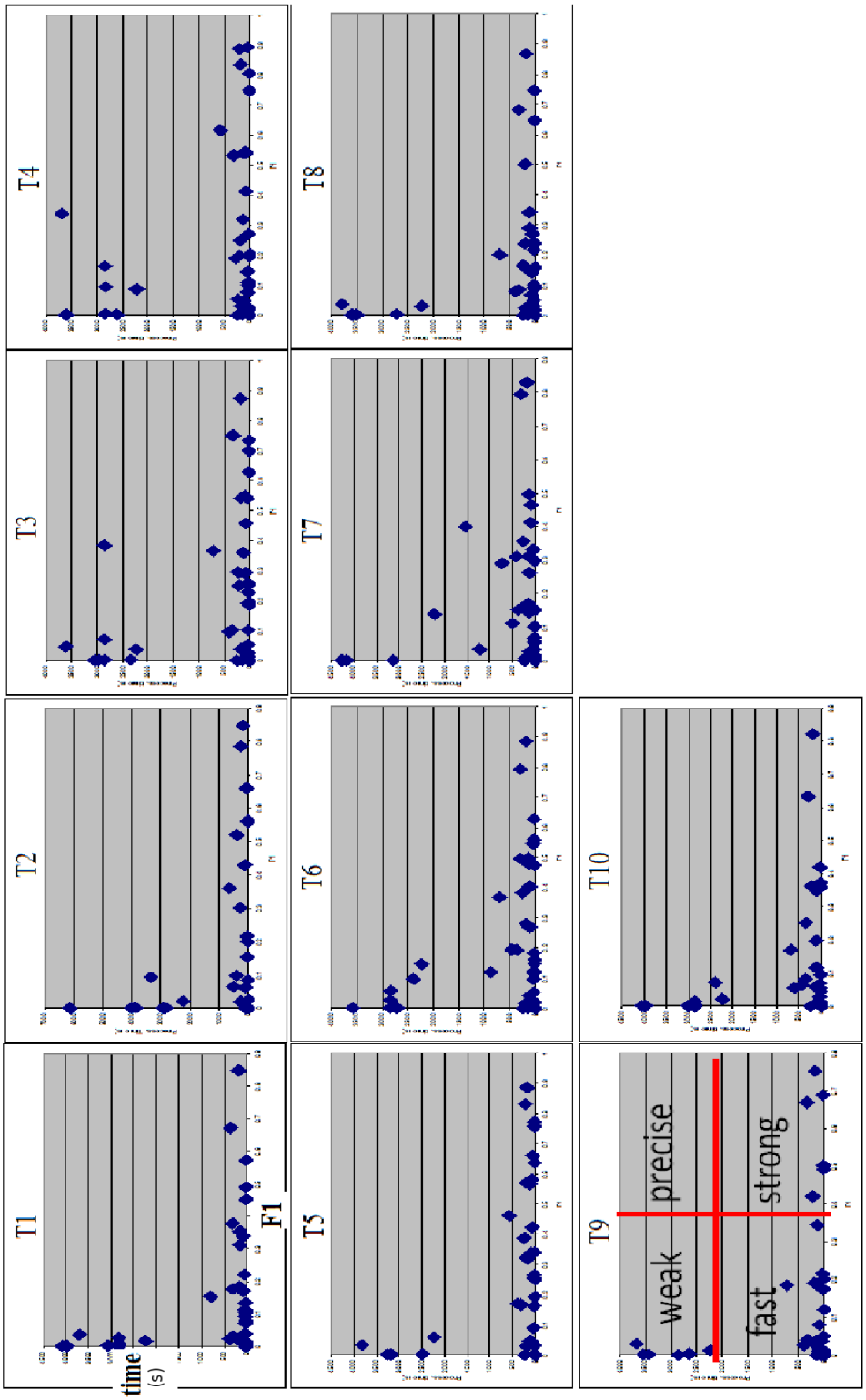


Fig. 13: TRECVID 2008: Relationship between processing time and cost across video transformations

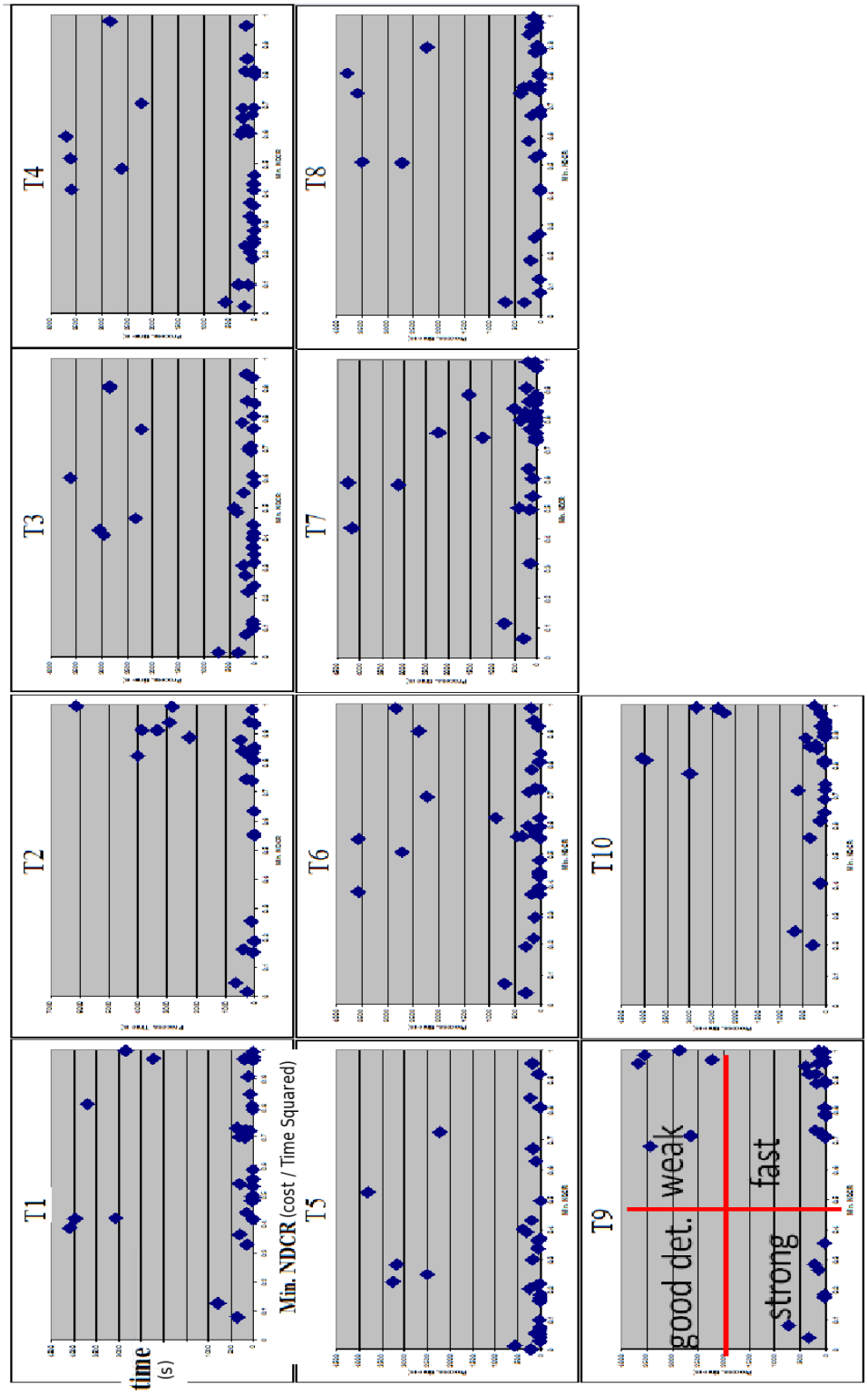


Fig. 14: TRECVID 2008: Audio+video runs vs. video-only

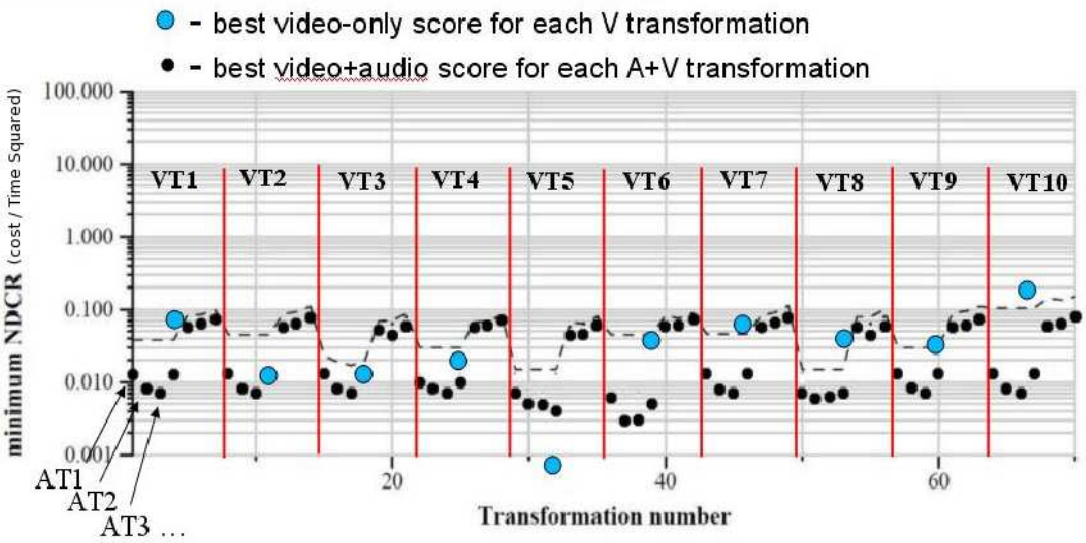
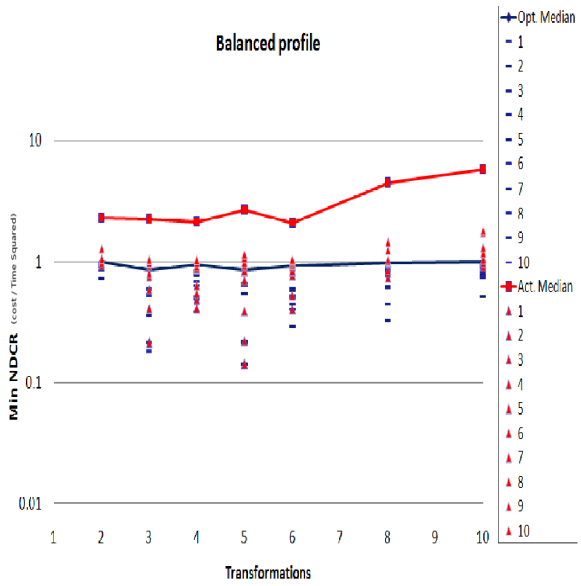


Fig. 15: TRECVID 2009: Video-only top 10 detection performance based on balanced profile



all runs in all evaluations) which indicates that choosing the run threshold was not a trivial task for systems. It can also be noted that this difference was smaller in audio and video+audio compared to video-only. Comparing top audio-only to top video+audio runs it can be shown that video+audio was better and almost didn't exceed detection score of 1.0 as in some transformations in audio-only runs. Similarly, comparing top video+audio runs to top video-only runs we can see that video-only actual median scores are consistently higher than 1.0 compared to video+audio. This shows that video+audio helped both audio-only modality and video-only as well. In other words using both modalities improve detection compared to only using one modality.

For the balanced profile, how did the different modalities compare in localization? Comparing the actual and optimum results of the top 10 runs based on localization performance for the three query types

Fig. 16: TRECVID 2009: Audio-only top 10 detection performance based on balanced profile

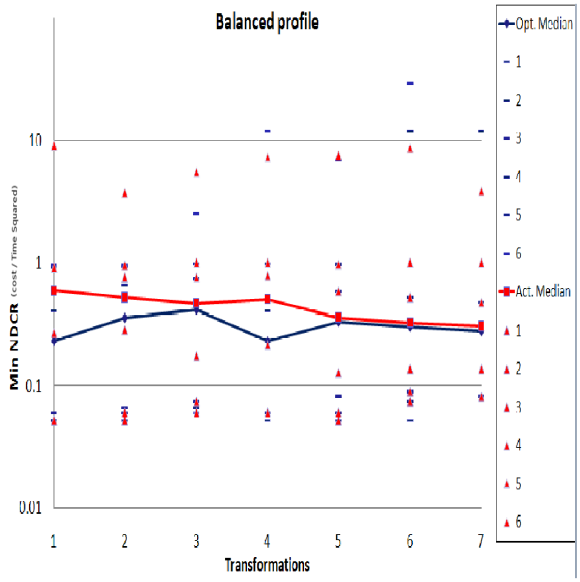
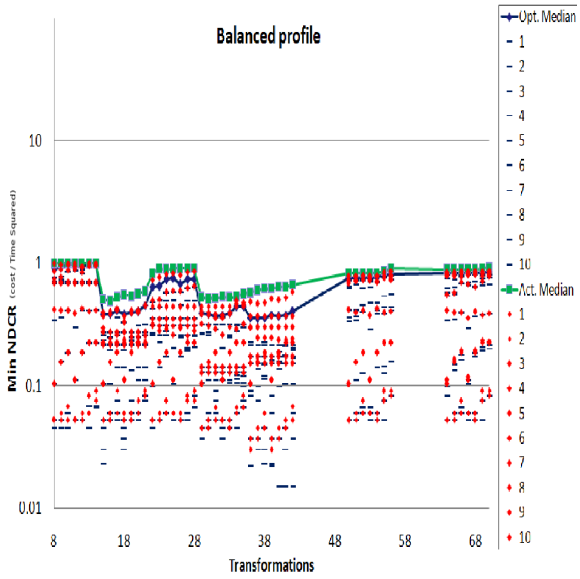


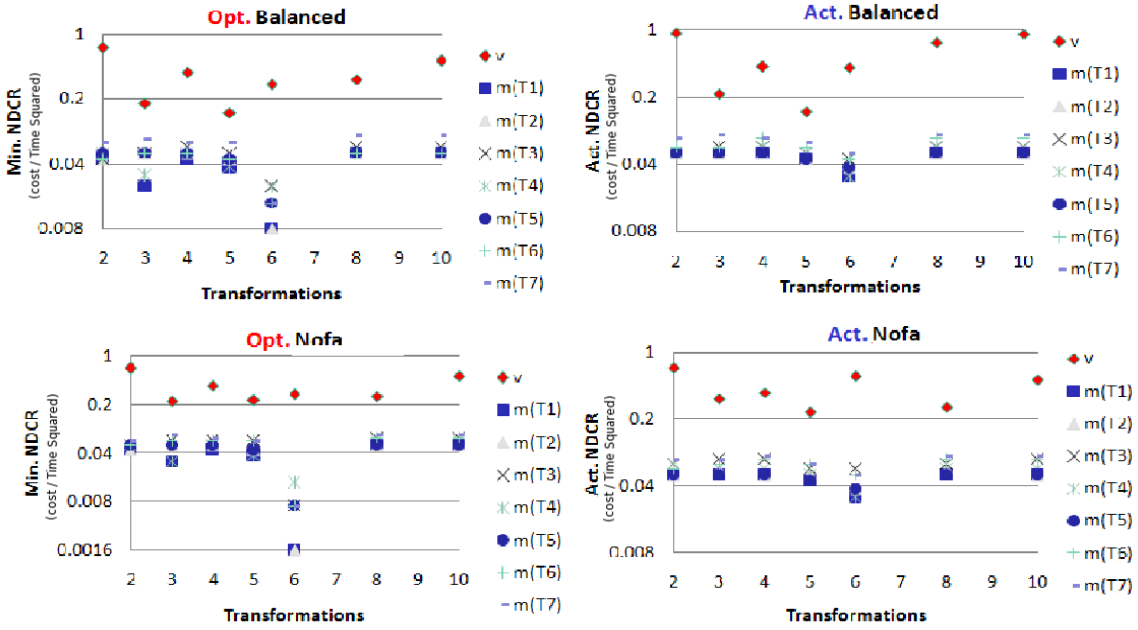
Fig. 17: TRECVID 2009: Video+audio top 10 detection performance based on balanced profile



shows that audio-only median localization was much more accurate than video-only or video+audio queries and its actual and optimal curves almost matched although only 6 runs were submitted for each profile.

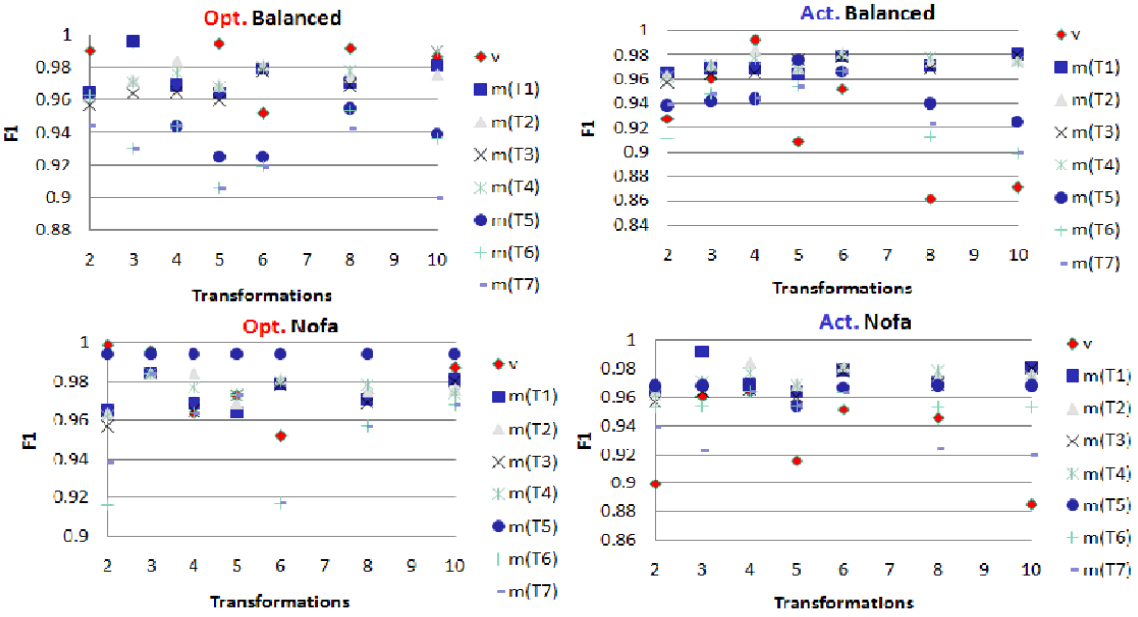
For the balanced profile, how did the different modalities compare in processing speed? Speed comparison among the top 10 shows that in general video+audio achieved the fastest processing time, followed by video-only then audio-only. For some transformations processing time was less than 5 s in video+audio runs, while the median values among the three query types were generally above 100 s. The max time reached about 1200 s in audio-only runs and this was primarily due to that fact that only 6 runs were submitted as audio-only and one of those runs used very slow cross correlation algorithm to calculate their results.

Fig. 18: TRECVID 2009: Video-only (v) vs video+audio (m) detection performance



The a+v runs enhanced the detection performance across all transformations

Fig. 19: TRECVID 2009: Video-only (v) vs video+audio (m) localization performance



The a+v runs helped in the majority of transformations to enhance localization

For the no-false-alarms profile, how did the different modalities compare on the main measures for top runs? The detection performance for the Nofa runs in general seemed to be worse than for the balanced profile. The optimum median across the transformations for the video-only and video+audio queries almost matched the NDCR of 1.0, which suggests that for many systems it would have been better for them just to reject all queries as copies. The localization of the audio-only queries for the Nofa runs was still relatively the most accurate compared to video-only and video+audio. The speed of the three query types for Nofa was comparable to that for the balanced profile.

What was the effect of adding audio as an information source? We compared the best runs of video-only to the best runs of video+audio to show the effect of adding audio as a clue. Figure 18 shows for each video transformation the best performance (in red diamond) and the best performance of the 7 audio transformations when applied to those video transformation (in purple). Using audio has decreased the detection cost across all transformations. The same experiment was done in Figure 19 based on localization performance. Although the same strong effect on detection cost was not seen, using audio has increased localization performance across the majority of the transformations.

Summarizing our observations for TRECVID 2009 for this task we find that determining the optimal operating point (threshold) was critical and required score normalization across queries. It had a huge impact on NDCR scores (especially for video-only runs) as illustrated by the large difference between actual and optimal results. Comparing the application profiles, there was a larger spread in NDCR for Nofa profile compared to balanced. Comparing modality types, audio-only detection outperforms video-only - probably because the audio techniques are more mature or easier. However, the combination of both audio and video improved upon using audio-only and video-only. Video-only systems in general were slightly faster than others and yielded the best localization results (although audio-only have a higher median). Few systems performed well in all three measures, thus there was still a room for systems to improve their accuracy, speed and performance. In general, there was a limited interest in audio-only queries (only 6 runs are submitted).

6.4. 2010

How was detection performance for both profiles? The detection performance, optimal and actual, among best runs for both profiles across all transformations is shown in Figures 20 to 23. Those figures plot the best NDCR score for each audio + video transformation (T1-T70) along the horizontal axis. Each set of a video transformation (V_{T1} - V_{T10}) is shown on the graph where all 7 audio transformations are applied such that for example T1 represents V_{T1} + A_{T1} where A_{T1} is the first audio transformation and V_{T1} is the first video transformation. Similarly T2 represents V_{T1} + A_{T2} where A_{T2} is the second audio transformation, and T8 represents V_{T2} + A_{T1} where V_{T2} is the second video transformation and A_{T1} is the first audio transformation. When two or more runs share the same top score for a transformation, we plot each run score. In those figures run names were removed to allow for better figure quality.

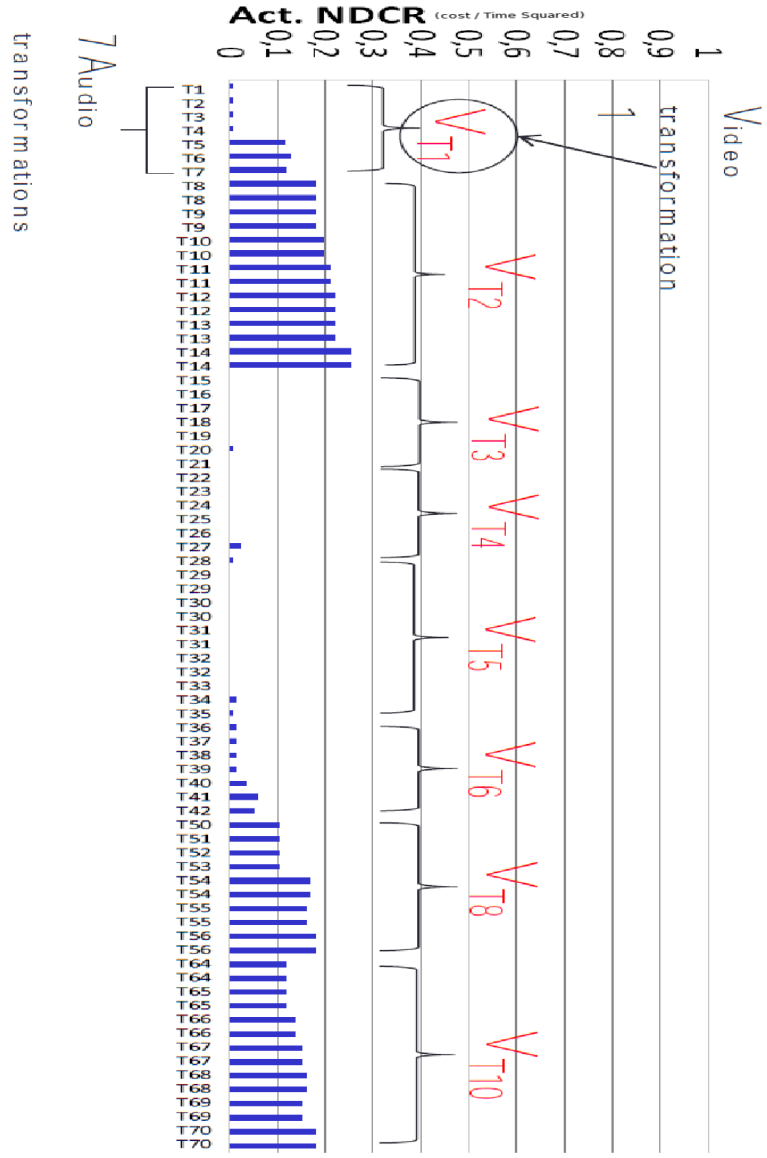
For the balanced profile a noticeable difference can be seen between the actual (Fig. 20) and optimal (Fig. 23) results while for Nofa profile the difference was very small. Optimal performance on transformations 3, 4, and 5 was clearly better than on the other transformations. This was not the case for the actual detection performance, likely due to the fact that transformations 3, 4 and 5 (insertion of patterns, re-encoding, and change of gamma) were simpler than the others, which combined multiple transformations.

A comparison of detection of the top 10 runs only for both profiles is shown in Figures 24 and 25. The gap between the actual median line and optimal median line shows that there was still space for more system improvements. This gap in the Nofa profile was much bigger compared to that under the balanced profile.

How was performance on localization and speed? The top 10 runs performed almost equally in localization. We notice that the optimal median was better than the actual median for most of the transformations except for the first two video transformations. In terms of speed, even though the best runs could detect copies in seconds, the majority of other systems were still far from real-time detection.

In summary, few systems achieved high localization in short processing time, and most systems which increased the processing time did not gain much in both localization and detection. On the other hand, systems that were good in detection were also good in localization. These observations are true for both application profiles. There was still substantial room for improvement available for the balanced profile indicated by difference between actual and optimal results and difference across

Fig. 20: TRECVID 2010: Top run for each transformation, based on Actual NDCR score in balanced profile



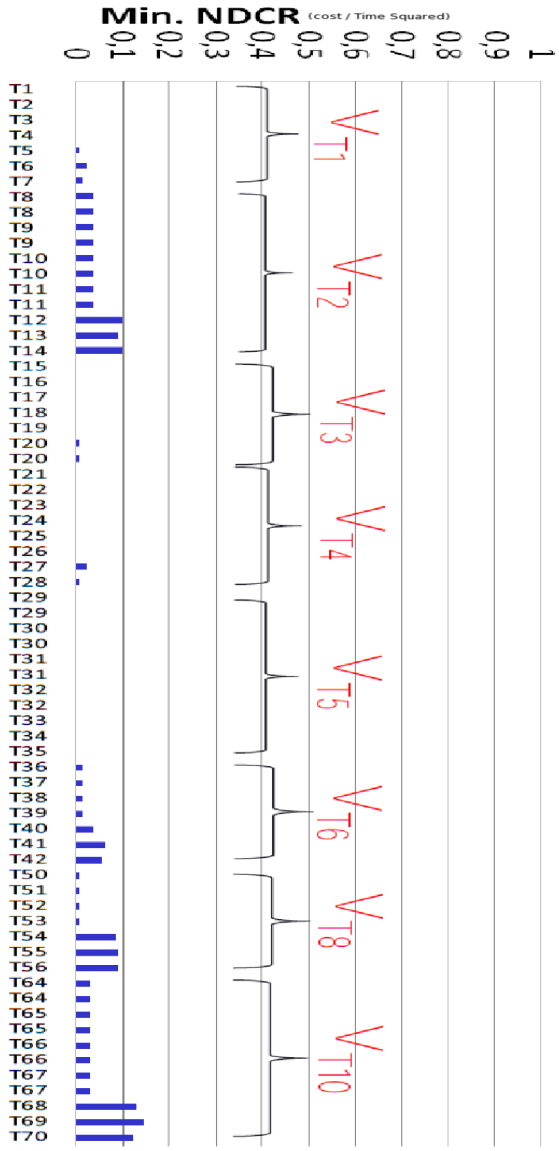
top runs. Determining the optimal threshold was still a major hurdle. Most of the systems were still far from real-time detection while good detecting systems were also good in localization. Complex transformations (mixture with external audio or combinations of video transformations) were more difficult. Camcording was a difficult transformation for some systems.

6.5. 2011

The reference and non-reference data for 2011 was the same as that used in 2010 and so comparison of results between 2010 and 2011 is possible. The query sets were not identical but were large in size and automatically created using randomization procedures. The transformations were the same as in 2010.

How well did the best systems perform in detection? The 2011 detection performance by the best runs for both profiles across all transformations is shown in Figures 26 to 29 (generated similarly to 2010 figures 20 to 23) . For the balanced profile the actual and optimal results were very close, while for

Fig. 21: TRECVID 2010: Top run for each transformation, based on Optimal NDCR score in balanced profile

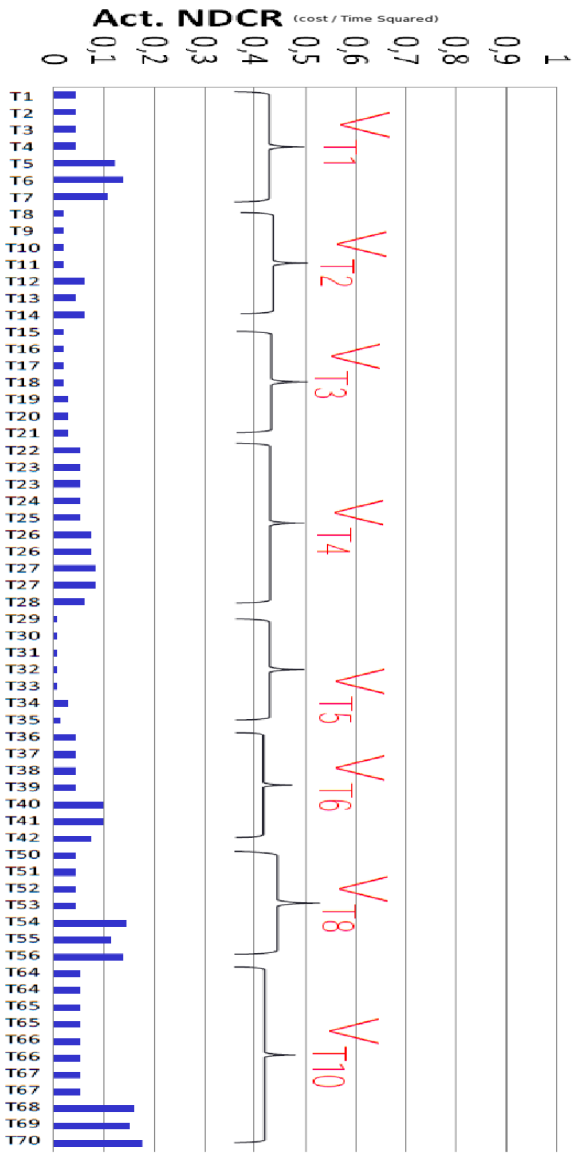


the Nofa profile the larger difference may indicate that the Nofa profile was still harder for systems. A comparison of detection in only the top 10 runs, based on mean performance across all transformations for both profiles, also showed a gap between the actual median line and optimal median line. Again, this gap in the Nofa profile was much bigger compared to the balanced profile. Looking at detection by transformation, transformations 3,4 and 5 achieved the best performance, which was likely due to the fact that those transformations (insertion of patterns, re-encoding, and change of gamma) are simpler than the others (2,6,8, and 10) including Picture in Picture and combined transformations.

How well did the best systems perform in localization and speed? A comparison based on the localization shows that both profiles achieved very high performance. The optimal and actual scores were almost aligned with very small gaps. In terms of speed the best runs could detect copies in few seconds, but still the majority of other systems were far from real-time detection.

How did 2011 performance compare to 2010? In general the detection performance in 2011 was better than 2010 (lower NDCR values). Both application profiles achieved actual median scores in 2011 better

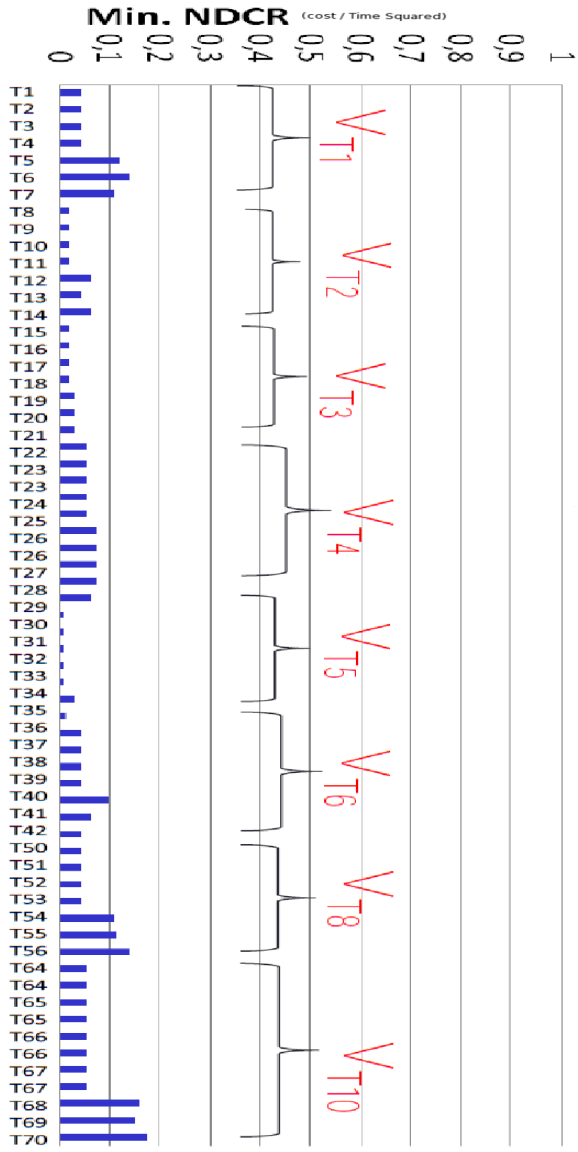
Fig. 22: TRECVID 2010: Top run for each transformation, based on Actual NDCR score in Nofa profile



than 2010. A comparison based on the localization shows that both profiles in 2011 achieved better than 2010. In regard to processing speed, systems were slower in 2011 and with higher median scores compared to 2010. It appears that 2011 systems focused more on enhancing the detection performance but at the cost of less speed.

Finally, we can draw some general observations from the 2011 results. The top actual balanced detection scores were very near to top optimal balanced detection scores. At the same time, in general top balanced detection results were better than no false alarm. Similar to previous year, the gap between actual and optimal medians for no false alarm detection results was bigger than the balanced actual and optimal gap. Good detecting systems were also good in localization as in previous years. Audio transformations 5, 6 & 7 still seem to be the hardest, while video transformations 3, 4, 5 & 6 seem to be the easiest. Finally, NDCR and F1 seemed to be approaching a ceiling.

Fig. 23: TRECVID 2010: Top run for each transformation, based on Optimal NDCR score in Nofa profile

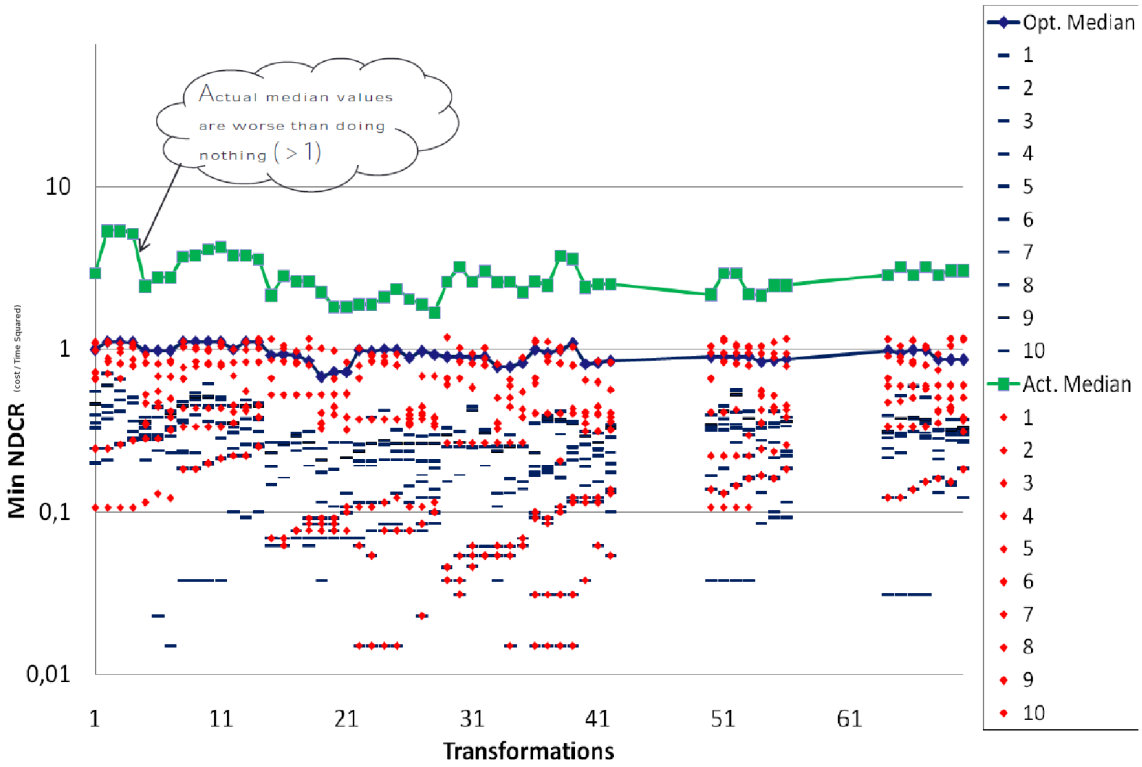


7. DISCUSSION

The TRECVID CCD track has established a community of CCD teams over the years - teams with an existing track record in image copy detection or audio copy detection and newcomers at TRECVID for which the CCD task proved to be an attractive entry level task. The CCD challenge has been instrumental in providing teams with results on unseen data and the different datasets provide rich resources for further experimentation. Reflecting on the four years of the task, a number of thoughts emerge.

Audio: The evaluation framework did not really change over the 4 years, but as documented in section 6.1, the task did evolve gradually over the years with respect to what information channels were optional versus required - audio, video, or audio+video. At the start, there was the usual TRECVID focus on exploitation of the visual information with a secondary curiosity about how much the audio and video channels were each contributing to the solutions. There was also the thought that an audio-only task would provide a very hard baseline to beat, which could have discouraged participants interested

Fig. 24: TRECVID 2010: Top 10 runs NDCR score in balanced profile



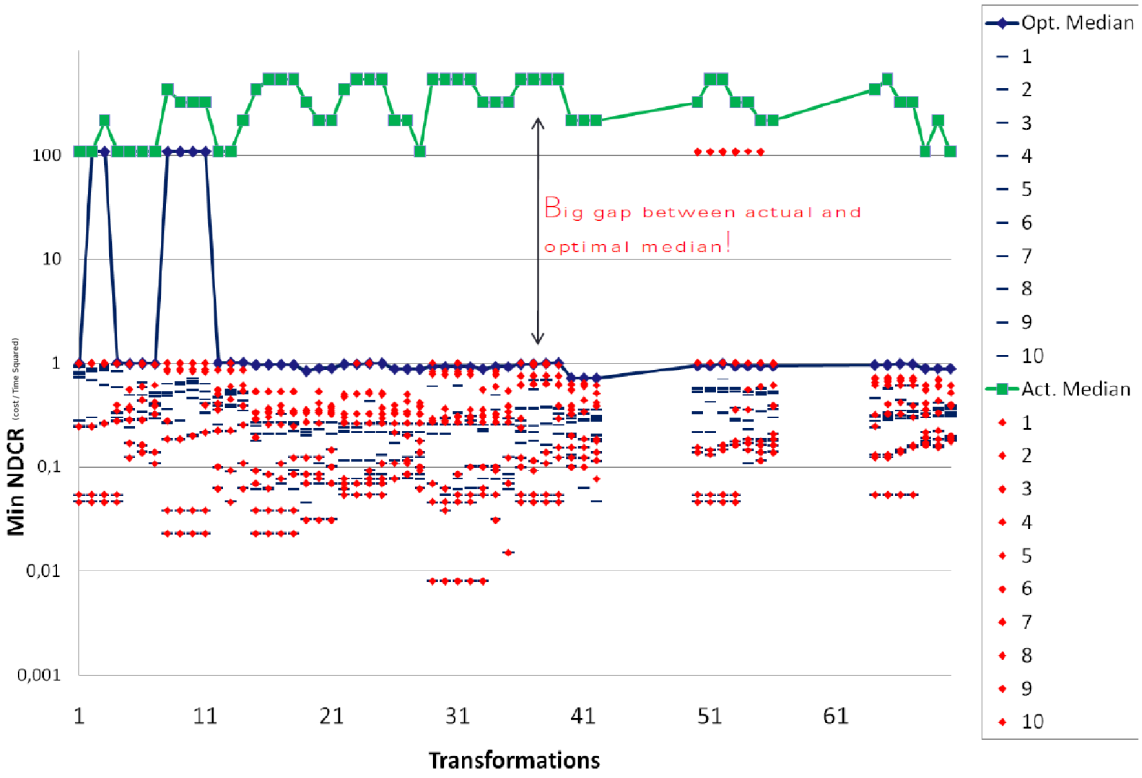
in video-based copy detection. Over the years, it appeared that indeed the use of the audio proved very effective. This was not surprising since the audio copy detection matured much earlier (in the late 1990's) in the context of the development of digital rights management (DRM) technology for digital audio. The decision to include audio as an independent challenge, with a set of audio transformations, attracted several leading teams from the audio copy detection community, eventually leading to cross-fertilization between the communities as audio teams started to experiment with video fingerprints and vice versa.

Although our conclusion in 2009 was that audio modality was easier for systems and achieved better scores than video-only, [Barrios 2013] in more analysis found that in 2010 and 2011 comparing the average performance across all submissions that there is no clear winner between audio-only and video-only. In the end, it was clear that the best, most realistic approaches would use both audio and video, as reflected in the task requirements for 2010 and 2011.

Application profiles: The rich structure of the task (multiple metrics, application profiles and transformations) had pros and cons. On the one hand it provided participants with detailed insights into the performance of their system. On the other hand, it is difficult to compare systems and system variants, since a single point of reference was missing. Still it was useful to include results for different application profiles (Nofa and Bal) in the official evaluation results. Teams have different settings in mind for eventual application and the parameterized cost-based framework accommodated two quite different application scenarios. Of course, the framework enabled experiments with different parameter settings, but in reality, teams usually only invested time in the required runs. Summarizing, the choice of various experimental conditions was always a trade-off. Providing more experimental conditions complicated the task, but enabled organizers to put the research focus of the community on various important questions, such as the relative contributions of modalities or the trade off between false alarms and missed detections.

Transformations: Comparing the transformations' degree of difficulty, it has been shown that mixture with speech audio transformations T5, T6, & T7 were the most difficult for the audio modality while T6–T10 video transformations were the most difficult in the video modality, as each was a com-

Fig. 25: TRECVID 2010: Top 10 runs NDCR score in Nofa profile



bination of three to five different video transformations. How realistic were the tested transformations is an open question. We believe that each transformation on its own was a realistic one but which ones (and how many) are the most combined in real-world copied videos is not very clear.

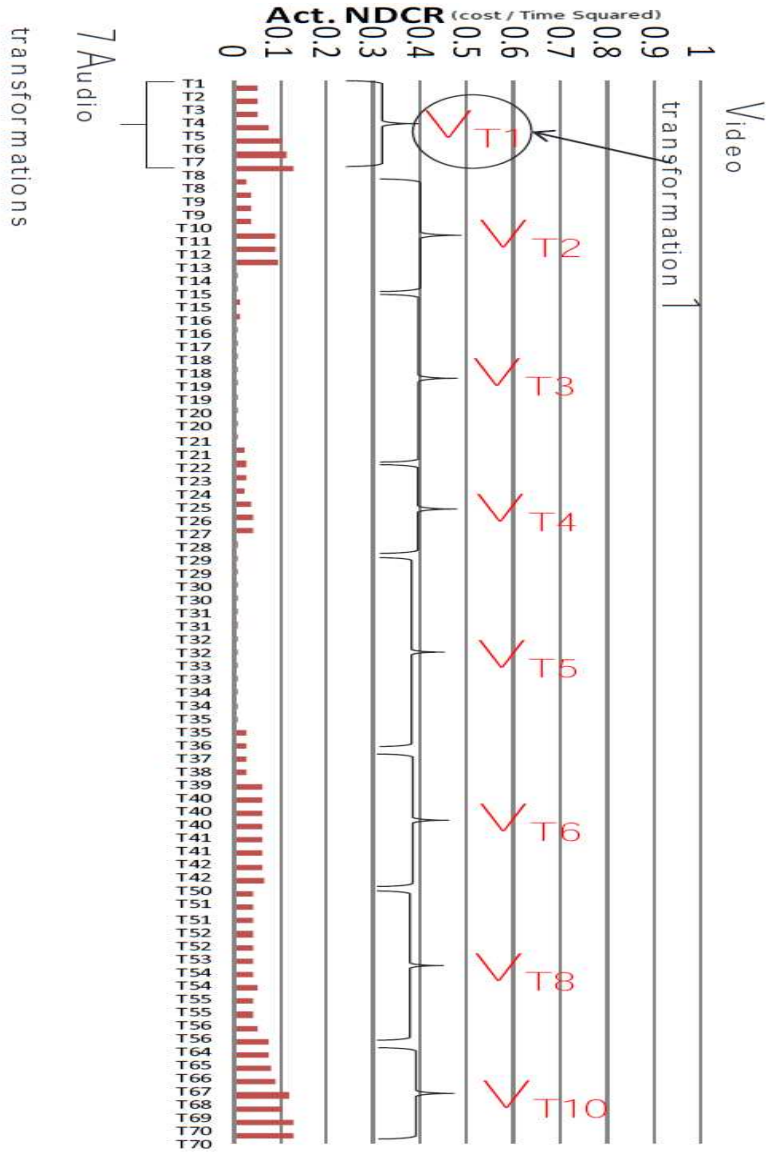
Measures: Comparing the three measures, we think that measuring location accuracy perhaps did not contribute sufficient additional information to merit its status as metric. The analysis of interaction between detection effectiveness and its efficiency was a useful analysis, showing that effective systems could be fast at the same time and that many ineffective systems were slow at the same time.

Synthetic queries: Overall, the idea of using synthetic queries proved to be a very useful model for evaluation. Given the limitations that real samples of copied copyrighted content are difficult to obtain and the fact that it was almost impossible to get clearance for redistribution, working with synthetic copies was an efficient way to create a test collection on a scale that is necessary to have sufficient sample sizes for the reliable estimation of the error rates. At the same time, it is difficult to generalize TRECVID CCD results to real-world video collections on video sharing portals. The CC_WEB_VIDEO collection shows the potential of a methodology to create test collections that might be fit for purpose even if copyright issues preclude redistribution.

Detection cost framework: Working with the detection cost framework and actual and optimum values across a range of video and audio transformations has given CCD participants a very rich insight into the qualities and weaknesses of their systems. It allowed us to build in a prior probability of finding copies, appropriate for the target application but different from the test data. The detection framework also forced teams to normalize detection scores suitable for a simple score threshold classifier. This challenge proved to be difficult for many teams. The difficulty is related to the fact that systems had to submit only one global threshold to be applied on all queries (since the applied transformation for a query is unknown). However, the detection cost framework calculated the optimal threshold per transformation and thus it became hard (specially in NoFA profile) for systems to achieve a similar performance to the optimal scores.

Other metrics: Of the two additional metrics, detection processing time proved to be more important than location accuracy. Participants did not specifically train their systems for location accuracy, since

Fig. 26: TRECVID 2011: Top runs based on Actual NDCR score in balanced profile

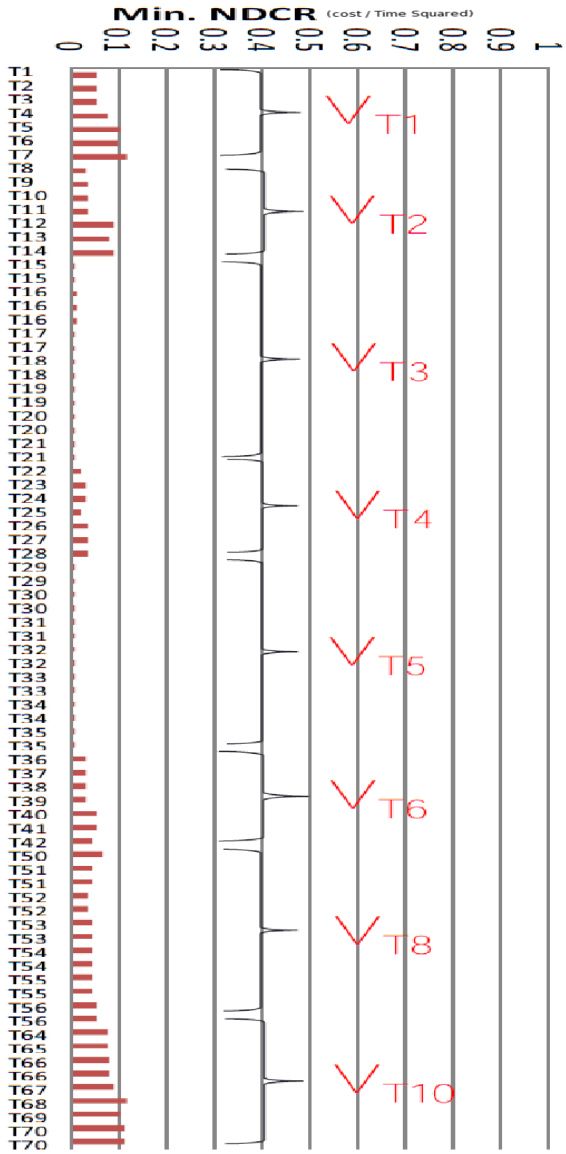


the mere detection of reference material is sufficient for most application scenarios. However we saw some teams that tried to optimize both measures by using GPUs framework to enhance their processing speed and using other techniques such as time sequence consistency check to enhance their localization of copied frames (for more details see [Zhang et al. 2009] and [Xie et al. 2012]).

Architecture: There was some parallel of CCD systems evolution with the early years of the TRECVID high level feature task. In the beginning some transformation specific approaches were investigated, but over the years, more generic techniques became the norm. It is beyond the scope of this paper to do a comprehensive analysis of the system architectures of CCD participants in relation to their results. We encourage the reader to access the publications of the participating teams.

Keyframe selection: An aspect of interest for the sake of scalability is keyframe selection. Different teams applied different keyframe extraction methods but we noticed that among top teams there are systems that only chose to extract 1 keyframe per shot while others used dense sampling. Thus it seems that keyframe rate is not a main factor in systems performance but is one among many others such as

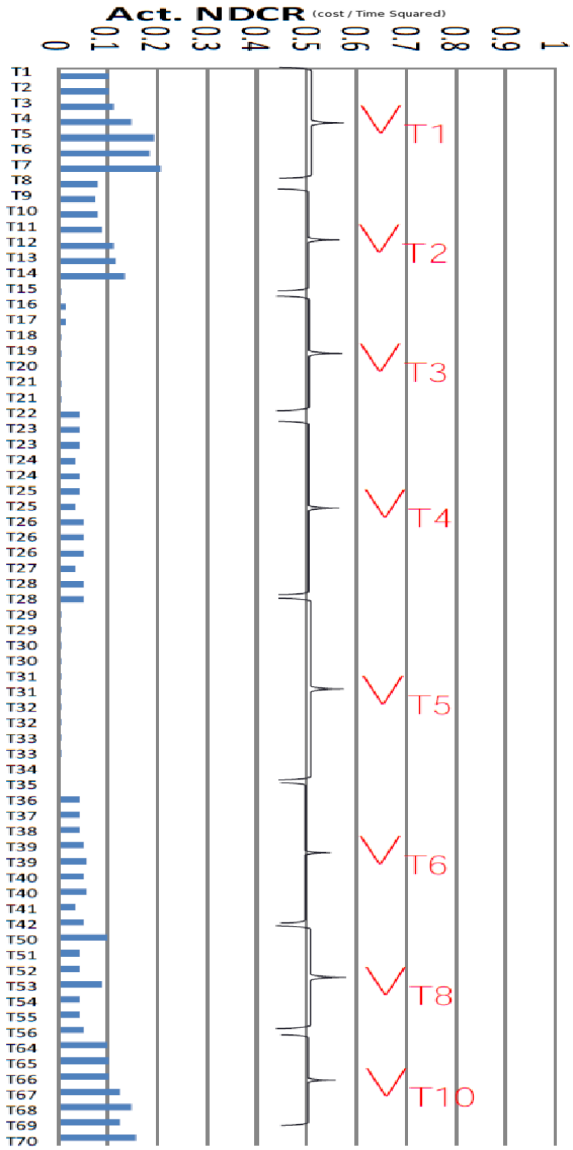
Fig. 27: TRECVID 2011: Top runs based on Optimal NDCR score in balanced profile



features, classifiers, modalities and fusion strategies. The main trends in the hundreds of experiments that have been carried out over these years are the application of Bag of Visual Words approaches; combination of different feature representations, composite classifier pipelines for improved efficiency, fusing audio and video scores at the decision level. Content based video copy detection are becoming more and more mature. Still, challenges remain to develop compact robust video signatures that allow fast search at the petabyte level.

Overall the legacy of CCD at TRECVID is rich. The CCD task has created several public test collections which have sufficient size to do meaningful experiments on video copy detection. There are several evaluation test collections available at http://trecvid.nist.gov/trecvid.data.html. In addition, the evaluation tools and query transformation scripts are available at http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/copy.detection.

Fig. 28: TRECVID 2011: Top runs based on Actual NDCR score in Nofa profile

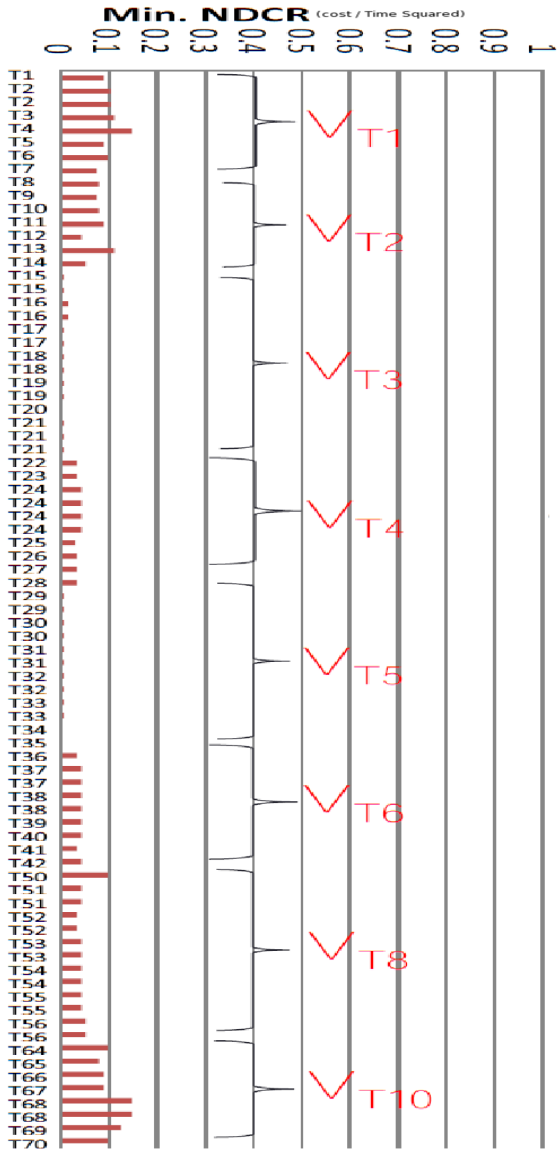


ACKNOWLEDGMENTS

We thank Alexis Joly and other members of the INRIA Imedia group that actively participated in shaping the evaluation plan for the initial TRECVID 2008 CCD track, incorporating the experiences gained in the preceding MUSCLE challenge. INRIA also made their video transformation tools to create queries available for TRECVID. In addition we thank Jonathan Fiscus for active support in creating the NDCR-based evaluation scheme. We thank Dan Ellis for creating the audio-based transformations. The comments and suggestions of three anonymous reviewers and those of Juan Manuel Barrios helped improve the paper; the authors of course retain responsibility for its final form.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Fig. 29: TRECVID 2011: Top runs based on Optimal NDCR score in Nofa profile



REFERENCES

A. Albiol, M. Full, A. Albiol, and L. Torres. 2004. Detection of TV Commercials. In *Proceedings of International Conference on Acoustics, Speech, and Signal processing (ICASSP '04), IEEE*, Vol. 3. Montreal, Quebec, Canada, 541–544.

Werner Bailer. 2010. Evaluating Detection of Near Duplicate Video Segments.. In *CIVR* (2010-07-07), Shipeng Li, Xinbo Gao, and Nicu Sebe (Eds.). ACM, 197–204. http://dblp.uni-trier.de/db/conf/civr/civr2010.html#Bailer10

S. Baluja and M. Covell. 2007. Audio Fingerprinting: Combining Computer Vision & Data Stream Processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 07)*. II213II216.

Juan Manuel Barrios. 2013. *Content-Based Video Copy Detection*. Ph.D. Dissertation. Department of Computer Science, University of Chile.

Jason C Breen. 2007. YouTube or YouLose? Can YouTube Survive a Copyright Infringement Lawsuit. *bepress Legal Series* (2007), 1950.

Chih-Yi Chiu, Cheng-Hung Li, Hsiang-An Wang, Chu-Song Chen, and Lee-Feng Chien. 2006. A Time Warping Based Approach for Video Copy Detection. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, Vol. 3. 228–231. DOI:http://dx.doi.org/10.1109/ICPR.2006.187

Matthijs Douze, Adrien Gaidon, Herve Jegou, Marcin Marszlak, and Cordelia Schmid. 2008. INRIA-IMEDIA TRECVID 2008: Video Copy Detection. http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/inria-lear.pdf. (2008).

Matthijs Douze, Herve Jegou, and Cordelia Schmid. 2010. An Image-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering. *IEEE Transactions on Multimedia* 12, 4 (2010), 257–266.

Pieter Eendebak, Wessel Kraaij, Stephan Raaijmakers, Elena Ranguelova, Ork de Rooij, Andrew Thean, and Marcel Worring. 2008. Visual Tools for Assisting Child Pornography Investigators. In *Proceedings of the NEM Summit 2008*.

Jonathan G. Fiscus and George R. Doddington. 2002. *Topic Detection and Tracking Evaluation Overview*. Kluwer Academic Publishers, Norwell, MA, USA, 17–31. http://dl.acm.org/citation.cfm?id=772260.772263

VishwaNath Gupta, Gilles Boulianne, and Patrick Cardinal. 2012. CRIMs Content-based Audio Copy Detection System for TRECVID 2009. *Multimedia Tools and Applications* 60, 2 (2012), 371–387. DOI:http://dx.doi.org/10.1007/s11042-010-0608-x

Vishwa Gupta, Parisa Darvish Zadeh Varcheie, Langis Gagnon, and Gilles Boulianne. 2011. CRIM AT TRECVID 2011: CONTENT-BASED COPY DETECTION USING NEAREST NEIGHBOR MAPPING. http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/crim.ccd.pdf. (2011).

Vishwa Gupta, Parisa Darvish Zadeh Varcheie, Langis Gagnon, and Gilles Boulianne. 2012. Content-based Video Copy Detection using Nearest-Neighbor Mapping. In *ISSPA*. 918–923.

Maguelonne Héritier, Vishwa Gupta, Langis Gagnon, Gilles Boulianne, Samuel Foucher, and Patrick Cardinal. 2009. CRIMs Content-Based Copy Detection System for TRECVID. http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/crim.pdf. (2009).

Zi Huang, Heng Tao Shen, Jie Shao, Bin Cui, and Xiaofang Zhou. 2010. Practical Online Near-Duplicate Subsequence Detection for Continuous Video Streams. *Multimedia, IEEE Transactions on* 12, 5 (2010), 386–398. DOI:http://dx.doi.org/10.1109/TMM.2010.2050737

Piotr Indyk, Giridharan Iyengar, and Narayanan Shivakumar. 1999. *Finding Pirated Video Sequences on the Internet*. Technical Report. Computer Science Department, Stanford University.

Menglin Jian, Shu Fang, Yonghong Tian, Tiejun Huang, and Wen Gao. 2011. PKU-IDM@TRECVid 2011 CBCD: Content-Based Copy Detection with Cascade of Multimodal Features and Temporal Pyramid Matching. http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/pku-ccd.pdf. (2011).

Menglin Jiang, YongHong Tian, and Tiejun Huang. 2012. Video Copy Detection Using a Soft Cascade of Multimodal Features. In *ICME*. 374–379.

A. Joly, O. Buisson, and C. Frelicot. 2007. Content-Based Copy Retrieval using Distortion-Based Probabilistic Similarity Search. *Multimedia, IEEE Transactions on* 9, 2 (2007), 293–306. DOI:http://dx.doi.org/10.1109/TMM.2006.886278

Yan Ke and Rahul Sukthankar. 2004. Efficient Near-duplicate Detection and Sub-image Retrieval. In *In ACM Multimedia*. 869–876.

Lyndon Kennedy and Shih-Fu Chang. 2008. Internet Image Archaeology: Automatically Tracing the Manipulation History of Photographs on the Web. In *Proceedings of the 16th ACM International Conference on Multimedia (MM '08)*. ACM, New York, NY, USA, 349–358. DOI:http://dx.doi.org/10.1145/1459359.1459406

Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistic Quarterly* 2 (1955), 83–97.

Julien Law-To, Li Chen, Alexis Joly, Ivan Laptev, Olivier Buisson, Valérie Gouet-Brunet, Nozha Boujemaa, and Fred Stentiford. 2007a. Video Copy Detection: a Comparative Study. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007, Amsterdam, The Netherlands, July 9-11, 2007*, Nicu Sebe and Marcel Worring (Eds.). ACM, 371–378.

J. Law-To, A. Joly, and N. Boujemaa. 2007b. Muscle-VCD-2007: A Live Benchmark for Video Copy Detection. (2007). http://www-rocq.inria.fr/imedia/civr-bench/.

Yuanning Li, Luntian Mou, Menglin Jiang, Chi Su, Xiaoyu Fang, Mengren Qian, Yonghong Tian, Yaowei Wang, Tiejun Huang, and Wen Gao. 2010. PKU-INM @ TRECVid 2010: Copy Detection with Visual-Audio Feature Fusion and Sequential Pyramid Matching. http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/pku-idm-ccd.pdf. (2010).

Jiajun Liu, Zi Huang, Hongyun Cai, Heng Tao Shen, Chong Wah Ngo, and Wei Wang. 2013. Near-Duplicate Video Retrieval: Current Research and Future Trends. *Comput. Surveys* 45, 4 (August 2013).

Zhu Liu, Tao Liu, David C. Gibbon, and Behzad Shahraray. 2010. Effective and Scalable Video Copy Detection. In *MIR'10: Proceedings of the International Conference on Multimedia Information Retrieval*. Philadelphia, Pennsylvania, USA.

Zhu Liu, Tao Liu, and Behzad Shahraray. 2009. AT&T Research at TRECVID 2009 Content-based Copy Detection. http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/att.pdf. (2009).

Yan Meng, E. Chang, and Beitao Li. 2003. Enhancing DPF for Near-replica Image Recognition. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, Vol. 2. II–416–23 vol.2. DOI:http://dx.doi.org/10.1109/CVPR.2003.1211498

D. A. Sadlier, S. Marlow, D. N. O'Connor, and N. Murphy. 2002. Automatic TV Advertisement Detection from MPEG Bitstream. *Pattern Recognition Society* 35, 12 (December 2002), 2–15.

YongHong Tian, Menglin Jiang, Luntian Mou, Xiaoyu Fang, and Tiejun Huang. 2011. A Multimodal Video Copy Detection Approach with Sequential Pyramid Matching. In *ICIP*. 3629–3632.

Y.H. Wan, Q.L. Yuan, S.M. Ji, L.M. He, and Y.L. Wang. 2008. A Survey of the Image Copy Detection. In *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*. 738–743. DOI:http://dx.doi.org/10.1109/ICCIS.2008.4670942

Xiao Wu, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. Practical Elimination of Near-Duplicates from Web Video Search. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA '07)*. ACM, New York, NY, USA, 218–227. DOI:http://dx.doi.org/10.1145/1291233.1291280

Hongtao Xie, Ke Gao, Yongdong Zhang, Jintao Li, Yizhi Liu, and Huamin Ren. 2012. Effective and Efficient Image Copy Detection Based on GPU. In *Trends and Topics in Computer Vision*, KiriakosN. Kutulakos (Ed.). Lecture Notes in Computer Science, Vol. 6554. Springer Berlin Heidelberg, 338–349. DOI:http://dx.doi.org/10.1007/978-3-642-35740-4_26

Yong-Dong Zhang, Ke Gao, Xiao Wu, Hon-Ttao Xie, Wei Zhang, and Zhen-Dong Mao. 2009. TRECVID 2009 of MCG-ICT-CAS. http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/mcg-ict-cas.pdf. (2009).

Received August 2013; revised February 2014; accepted April 2014