TRECVID 2009 –Goals, Tasks, Data, Evaluation Mechanisms and Metrics

Paul Over {over@nist.gov}
George Awad {gawad@nist.gov}
Jon Fiscus {jfiscus@nist.gov}
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Martial Michel {martial.michel@nist.gov}
Systems Plus
One Research Court, Suite 360
Rockville, MD 20850

Alan F. Smeaton {Alan.Smeaton@dcu.ie}
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Glasnevin, Dublin 9, Ireland

Wessel Kraaij {wessel.kraaij@.tno.nl}
TNO Information and Communication Technology
Delft, the Netherlands
Radboud University Nijmegen
Nijmegen, the Netherlands

April 8, 2010

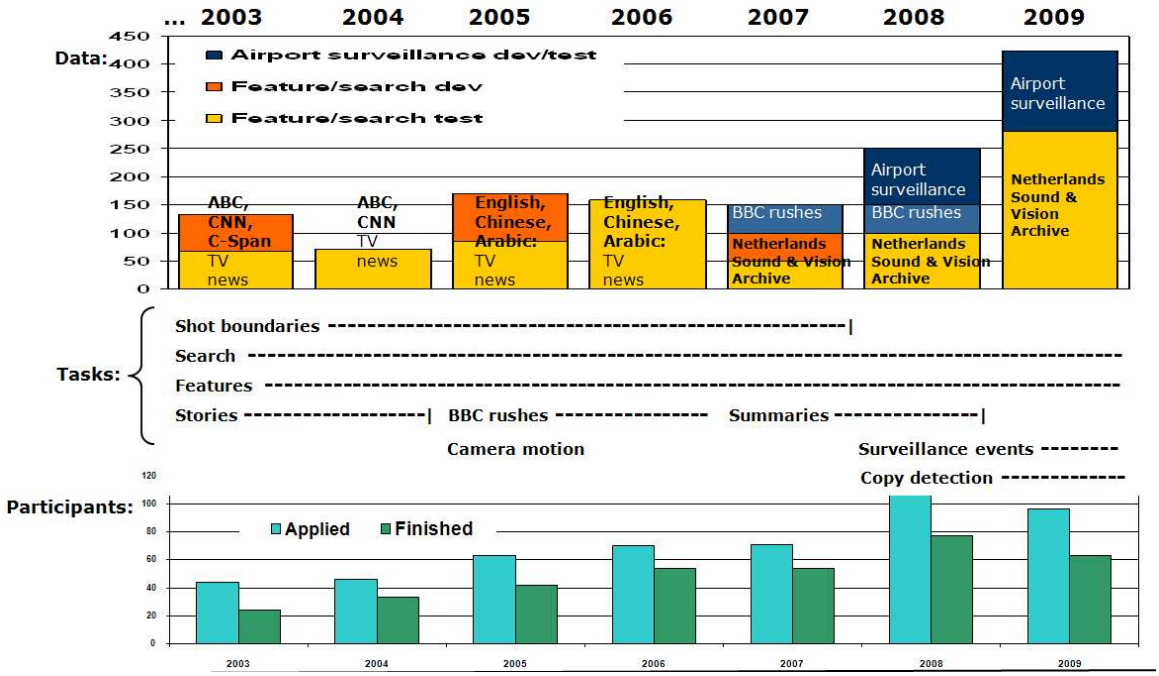
1 Introduction

The TREC Video Retrieval Evaluation (TRECVID) 2009 was a TREC-style video analysis and retrieval evaluation, the goal of which was to promote progress in content-based exploitation of digital video via open, metrics-based evaluation. Over the last 9 years TRECVID has yielded a better understanding of how systems can effectively accomplish such processing and how one can reliably benchmark their performance. TRECVID is funded by the Intelligence Advanced Research Projects Activity (IARPA), the US Department of Homeland Security (DHS), and the US National Institute of Standards and Technology (NIST).

63 teams (see Table 1) from various research organizations — 28 from Europe, 24 from Asia, 10 from North America, and 1 from Africa — completed one or more of four tasks: high-level feature extraction, search (fully automatic, manually assisted, or interactive), copy detection, or surveillance event detection.

In 2009, TRECVID was in the third year of a 3-year cycle using data for feature extraction and search, which is related to the broadcast TV news used in 2003-2006 but significantly different. Test data for the search and feature tasks was about 280 hours of (MPEG-1) TV news magazine, science news, news

Figure 1: Evolution of TRECVID



reports, documentaries, educational programming, and archival video almost entirely in Dutch from the Netherlands Institute for Sound and Vision. About 100 hours of video was available for search/feature system development. The combined 380 hours were used in the copy detection task. About 100 hours of airport surveillance video from the Image Library for Intelligent Detection Systems for Multi-Camera Tracking Training (i-LIDS MCTTR) provided by the UK Home Office was made available for training data in the 2009 surveillance event detection task. Systems were tested on about 15 hours of a new 50-hour test set from the same source.

Results were scored by NIST for almost all tasks against human judgments. Feature and search submissions were evaluated based on partial manual judgments of the pooled submissions. Copy detection submissions were evaluated at NIST based on ground truth created automatically using tools donated by the INRIA-IMEDIA group. NIST evaluated the surveillance event detection results using ground truth created manually under contract by the Linguistic Data Consortium

This paper is an introduction to the evaluation framework — the tasks, data, and measures for the workshop — as well as to the results and the technical approaches taken. For detailed information about the approaches and results, the reader should see the various site reports on the TRECVID website (trecvid.nist.gov).

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.

2 Data

2.1 Video

Sound and Vision data

The Netherlands Institute for Sound and Vision generously provided 400 hours of TV news magazine, science news, news reports, documentaries, educational programming, and archival video in MPEG-1 format for use within TRECVID. TRECVID 2007 used approximately 100 hours of this data — half for devel-

opment and half for evaluation of feature extraction and search systems. TRECVID 2008 used an additional 100 hours for testing. In 2009 all the 2007 data was available for system development and all the 2008 test data plus 180 hours of previously unused video were used for testing.

The collections for the search and feature tasks were drawn randomly so as to be balanced across the various TV program sources. The development data comprised 110 files and 64.3 GB, the test data 419 files and 179 GB.

The entire feature/search collection was automatically divided into shots by Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin. These shots served as the predefined units of evaluation for the feature extraction and search tasks. The feature/search test collection contained 93 902 reference shots.

Roeland Ordelman and Marijn Huijbregts at the University of Twente provided the output of an automatic speech recognition system run on the Sound and Vision data. Christof Monz of Queen Mary, University London contributed machine translation (Dutch to English) for the Sound and Vision video based on the University of Twente's automatic speech recognition (ASR). The LIMSI Spoken Language Processing Group produced a speech transcription for the TRECVID 2007-2009 Sound and Vision data using its recently developed Dutch recognizer.

BBC Archive data

The BBC Archive provided rushes video that was used in the copy detection task as non-reference video.

i-LIDS surveillance video

The development data consisted of the full 100 hours data set used for the 2008 Event Detection (Rose, Fiscus, Over, Garofolo, & Michel, 2009) evaluation.

The video for the evaluation corpus came from the 45 hour Home Office Scientific Development Branch's (HOSDB) Image Library for Intelligent Detection Systems (iLIDS) Multi Camera Tracking Training (MCTTR) data set. The evaluation systems processed the full data set however systems were scored on a 4-day subset of recordings.

3 High-level feature extraction

A potentially important asset to help video search/navigation is the ability to automatically identify the occurrence of various semantic features such as "Indoor/Outdoor","People", "Speech" etc., which occur frequently in video information. The ability to detect features is an interesting challenge by itself but takes on added importance as a reusable, extensible basis for query formation and search. The feature extraction task has the following objectives:

- to continue work on a benchmark for evaluating the effectiveness of detection methods for various semantic concepts
- to allow exchange of feature detection output for use in the TRECVID search test set prior to the search task results submission date, so that a greater number of participants could explore innovative ways of leveraging those detectors in answering the search task queries in their own systems.

The feature extraction task was as follows. Given a standard set of shot boundaries for the feature extraction test collection and a list of feature definitions, participants were asked to return for each feature in the full set of features, at most the top 2 000 video shots from the standard set, ranked according to the highest possibility of detecting the presence of the feature. The presence of each feature was assumed to be binary, i.e., it was either present or absent in the given standard video shot. If the feature was true for some frame (sequence) within the shot, then it was true for the shot. This is a simplification adopted for the benefits it afforded in pooling of results and approximating the basis for calculating recall.

The 20 features test in 2009 comprised 10 from the 2008 test set with moderate frequencies and 10 new features decided upon with input from the TRECVID community. Some feature definitions were enhanced for greater clarity, so it is important that the TRECVID feature descriptions be used and not the LSCOM descriptions.

Work at Northeastern University (Yilmaz & Aslam, 2006) has resulted in methods for estimating standard system performance measures using relatively small samples of the usual judgment sets so that larger numbers of features can be evaluated using the same amount of judging effort. Tests on past data showed the new measure (inferred average precision) to be a good estimator of mean average preci-

Figure 2: infAP by run (cat. A) - top

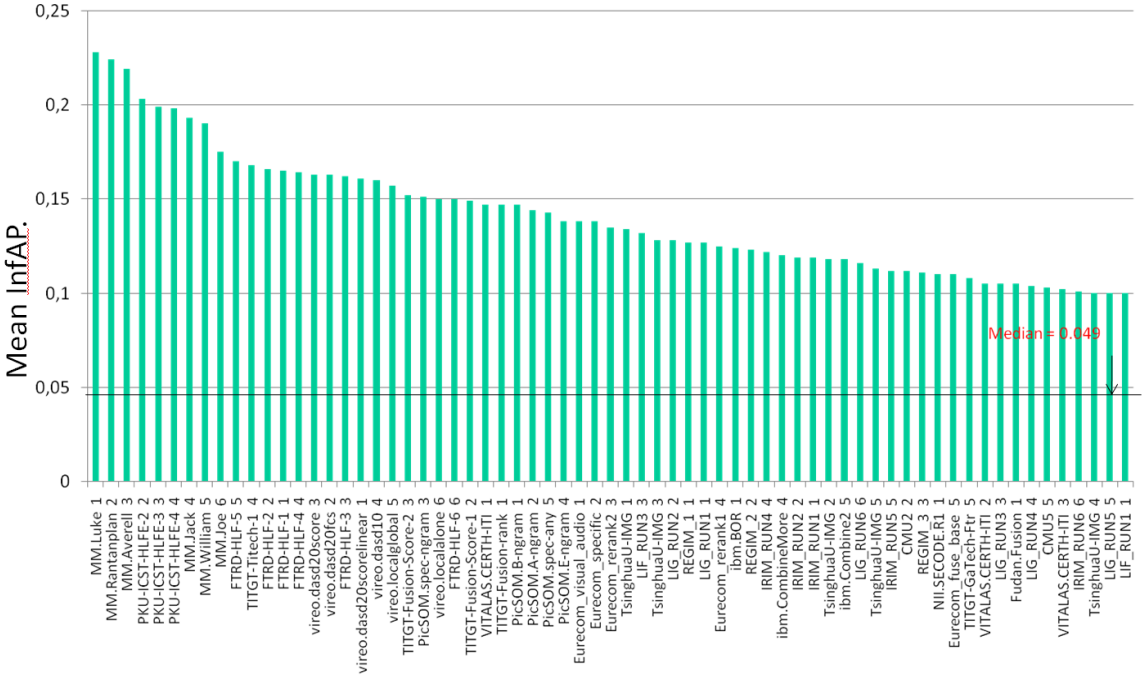


Figure 3: infAP by run (cat. A) - middle

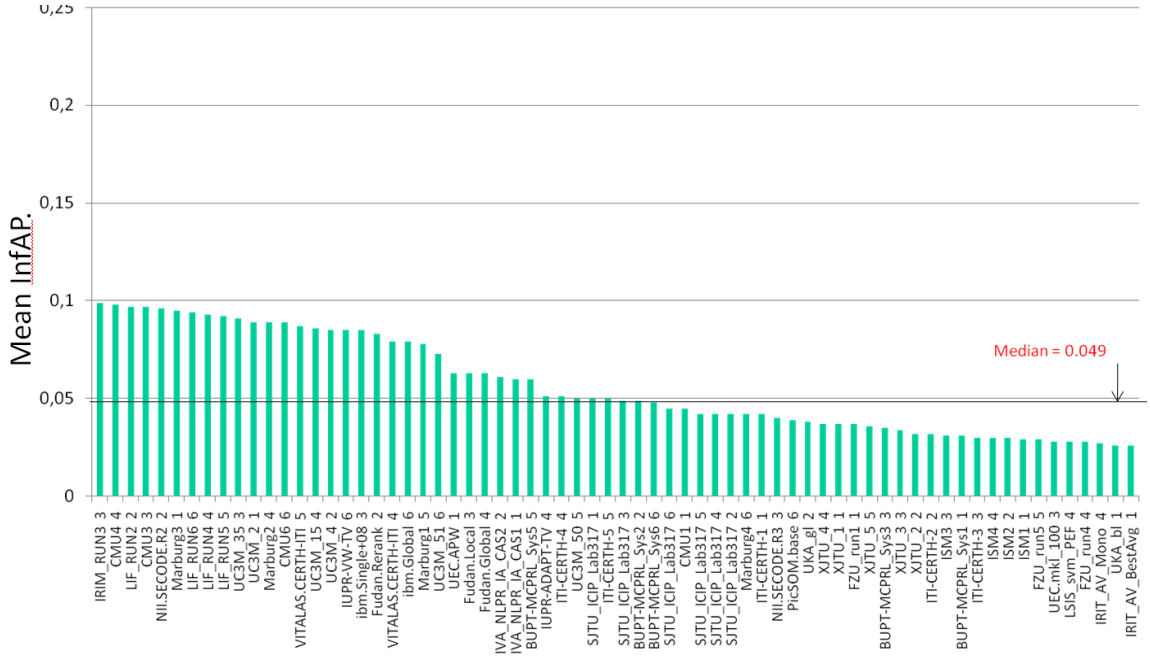
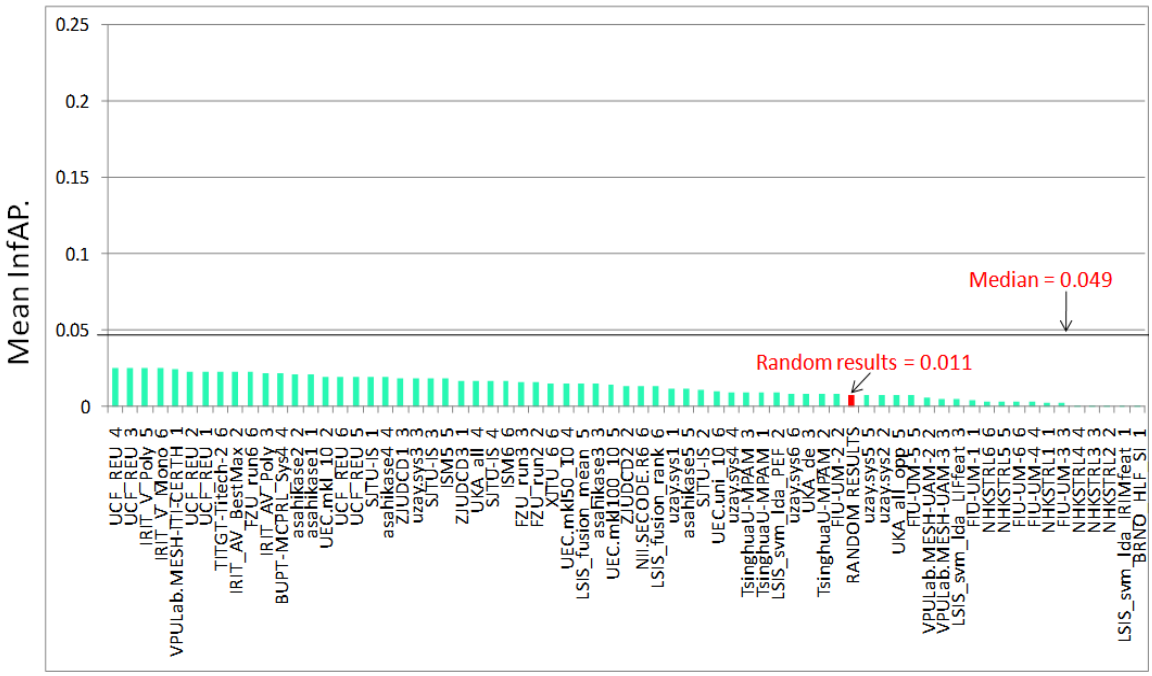


Figure 4: infAP by run (cat. A) - bottom



sion (Over, Ianeva, Kraaij, & Smeaton, 2006). As a result, it was decided to use a 50% sample of the usual feature task judgment set, calculate inferred average precision instead of average precision, and evaluate 20 features from each group.

Features were defined in terms a human judge could understand. Some participating groups made their feature detection output available to participants in the search task which really helped in the search task and contributed to the collaborative nature of TRECVID.

The features to be detected in 2009 were as follows and are numbered 1-20. All were evaluated. Those marked with an asterisk were also tested in 2008. [1] * Classroom, [2] Chair, [3] Infant, [4] Traffic intersection, [5] Doorway, [6] * Airplane-flying, [7] Person-playing-a-musical-instrument, [8] * Bus, [9] Person-playing-soccer, [10] * Cityscape, [11] Person-riding-a-bicycle, [12] * Telephone, [13] Person-eating, [14] * Demonstration-Or-Protest, [15] * Hand, [16] People-dancing, [17] * Nighttime, [18] * Boat-Ship, [19] Female-human-face-closeup, [20] * Singing.

The full definitions provided to system developers and NIST assessors are listed with the detailed feature runs at the back of the notebook and in Appendix B in this paper.

3.1 Data

As mentioned earlier, the feature test collection contained 419 files/videos but seven test files were ignored in the testing due to problems displaying shots from these long files (BG_36684, BG_37970, BG_38162, BG_8887, BG_37942, BG_8650, and BG_38653) in the assessment system. Removing these files left 412 files and 93 902 shots. Testing feature extraction and search on the same data offered the opportunity to assess the quality of features being used in search.

Georges Quénot and Stéphane Ayache of LIG (Laboratoire d'Informatique de Grenoble) again organized a collaborative annotation of the 10 new features in the TRECVID 2009 search/feature development data using an active learning scheme designed to improve the efficiency of the process (Ayache & Quénot, 2008).

The Multimedia Computing Group at the Chinese Academy of Sciences together with the National University of Singapore provided full annotation for 20 features of the 2009 training data.

3.2 Training conditions

In order to help isolate system development as a factor in system performance each feature extraction task submission, search task submission, or donation of extracted features declared its type as one of the following:

A - system trained only on common TRECVID development collection data, the common annotation of such data, and any truth data created at NIST for earlier topics and test data, which is publicly available. or on the former plus additional video (annotations)

C - system is not of type A

There continued to be special interest in how well feature/search systems trained on one sort of data generalize to another related, but different type of data with little or no new training data. The available training data contained some that is specific to the Sound and Vision video and some that was not. Therefore two additional training categories were introduced:

a - same as A but no training data (shared or private) specific to any Sound and Vision data has been used in the construction or running of the system.

c - same as C but no training data (shared or private) specific to any Sound and Vision data has been used in the construction or running of the system.

Groups were encouraged to submit at least one pair of runs from their allowable total that helps the community understand how well systems trained on non-Sound-and-Vision data generalize to Sound-and-Vision data.

3.3 Evaluation

Each group was allowed to submit up to 6 runs and in fact 42 groups submitted a total of 222 runs.

For each feature, all submissions down to a depth of at least 70 (average 100, maximum 170) result items (shots) were pooled, removing duplicate shots, randomized and then sampled to yield a random 50% subset of shots to judge. Human judges (assessors) were presented with the pools - one assessor per feature - and they judged each shot by watching the associated video and listening to the audio. The maximum result set depth judged and pooling and judging information for each feature is listed in Table 3. In all, 68 270 shots were judged.

3.4 Measures

The *trec_eval* software, a tool used in the main TREC activity since it started in 1991, was used to calculate inferred average precisions (infAP) for each result. Recall and precision are estimated by doubling the actual values found in the 50% sample used for judging. Since all runs provided results for all evaluated features, runs are best compared in terms of the mean inferred average precision across all 20 evaluated features. Within-feature comparisons are subject to greater influence by the sampling used.

3.5 Results

Figures 2, 3, and 4 show the results of category A, while Figures 5, 6, and 7 show the results of the other categories performance. The graphs show the median values in each category together with a random baseline result for category A. A small number of runs are below the random generated result. Still category A runs are the most popular type and achieve top recorded performances.

For the random baseline the value of infAP for a run and feature is the mean of infAP for 10000 randomly constructed result sets. Each result set contained, in a randomized ranking, the number hits likely to occur in a set of 2000 randomly selected shots given the actual density of hits for the feature in the judged pools.

Performance varies greatly by feature. Figure 8 shows how many unique instances were found for each tested feature. One feature (Doorway) exceeded 1% hits from the total tested shots percentage, On the other hand, features that had lowest hits were "Infant", "bus", "Airplane-flying", "Person-playing-soccer" and "Demonstration-or-protest". It can also be shown that features such as "Female human face closeup" received hits very near to the 1%. Figure 9 show the frequency of hits for the common 10 features between TV2008 and TV2009. In general TV2009 hits have increased across all features (except the hand feature). However, we also have to take into consideration that these results are based on using TV2008 and TV2009 testing dataset.

Figure 10 shows the performance of the top 10 teams across the 20 features. The behavior varies generally across features. For example some features reflect a large spread between the scores of the top 10

Figure 9: Frequency of features common to 2008/2009

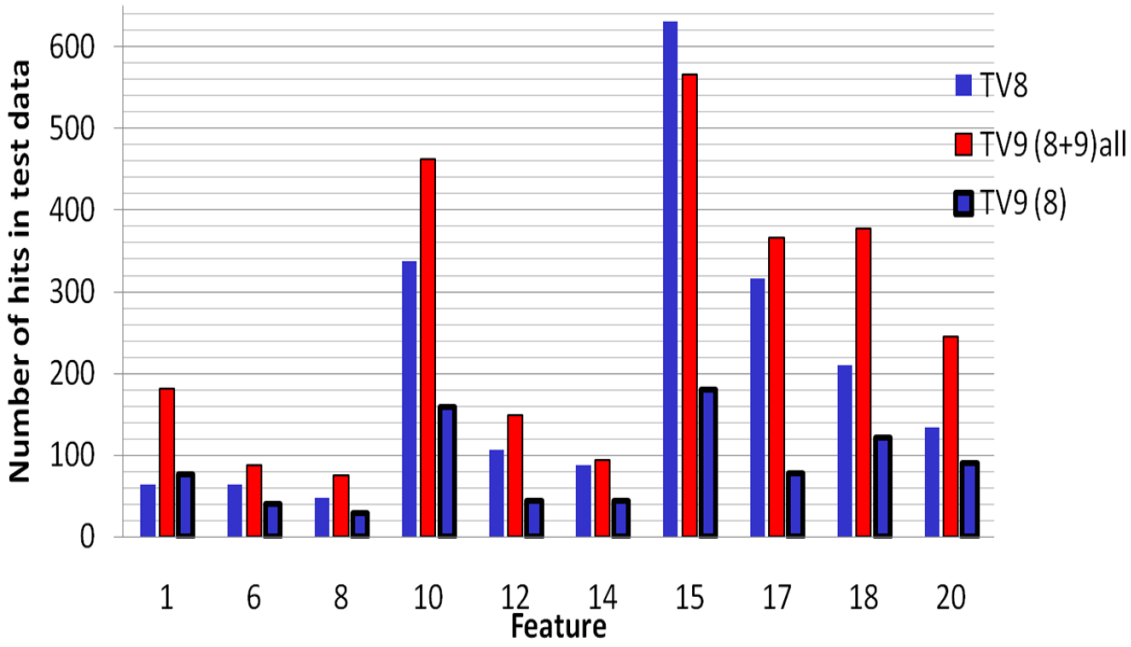


Figure 5: Effectiveness of category a runs

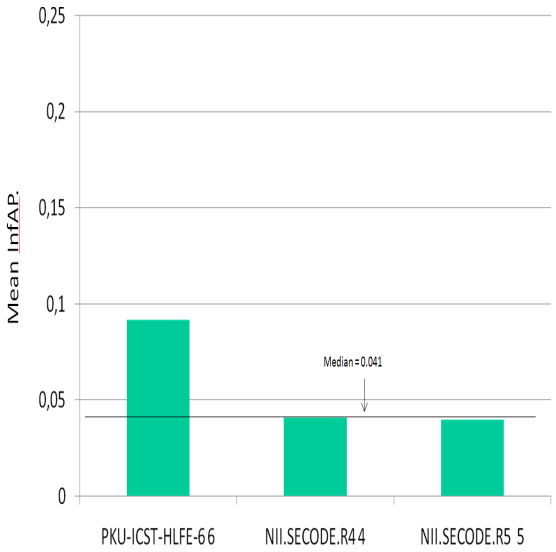


Figure 6: Effectiveness of category C runs

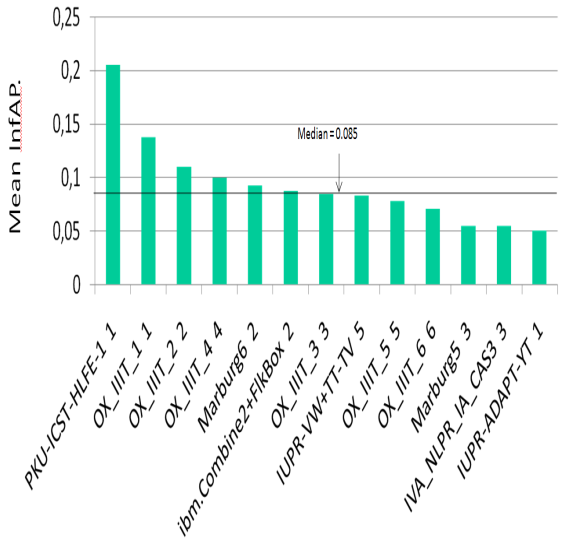


Figure 7: Effectiveness of category c runs

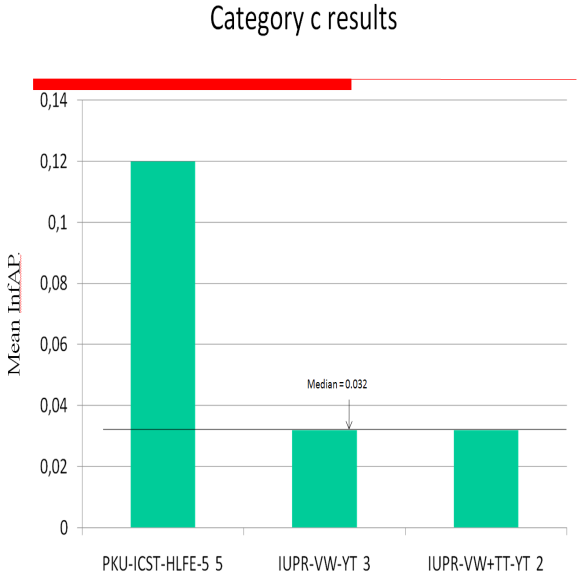
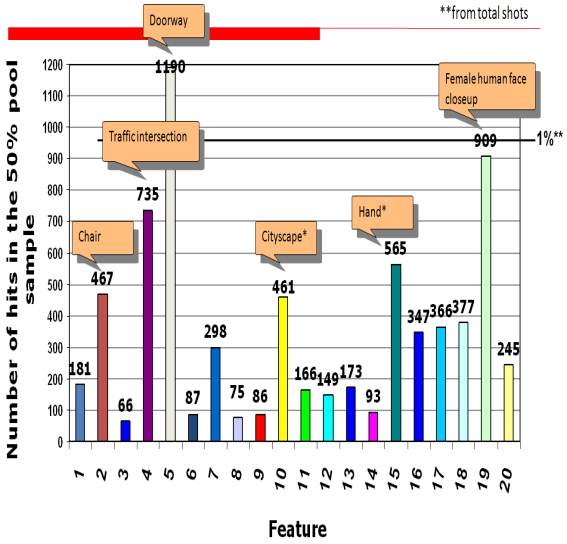


Figure 8: Frequencies of shots with each feature



such as feature "Telephone","Hand","Night-time", "Person-riding-bicycle","Person-playing-soccer", "Female-human-face-closeup", "People-dancing", "Bus" and "Infant". This indicates that there is still room for further improvement, while other features had a tight spread of scores among the top 10 such as feature "Chair", "Classroom", "Doorway", and "Person-eating". In general, the median scores ranged between 0.004 (feature "Infant") and 0.134 (feature Person-playing-soccer). As a general observation, feature "Infant" has the biggest spread across the top 10 and at the same time the minimum median score across all systems, which demonstrates how difficult this feature is for the systems to detect. Also, it can be shown on the graph that the median curve of all features is above the random baseline run generated by NIST.

Figure 11 shows a weak positive correlation between number of hits possible for a feature and the median or maximum score for that feature. To test if there are significant differences between the systems performance, we applied a randomization test (Manly, 1997) on the top 10 runs for each run category as shown in Figures 12 through 16. The left half indicates the sorted top 10 runs, while the right half indicates the order by which the runs are significant according to the randomization test. Figures 13 and 16 apply the randomization test to runs that used Sound and Vision data versus runs that did not use Sound and Vision data for training across same teams.

Based on the submitted site reports, some general observations can be made. Experiments involved efficiency improvements (e.g. graphics processing units (GPU)), audio and motion analysis, using more than one keyframe per shot and temporal context information, comparing fusion strategies together with merging many different representations, and finally automatic extraction of Flickr training data although fewer runs were submitted using external training data. Readers should see the notebook papers posted on the TRECVID website (trecvid.nist.gov) for details about each participant's experiments and results.

Figure 10: Top 10 runs (infAP) by feature

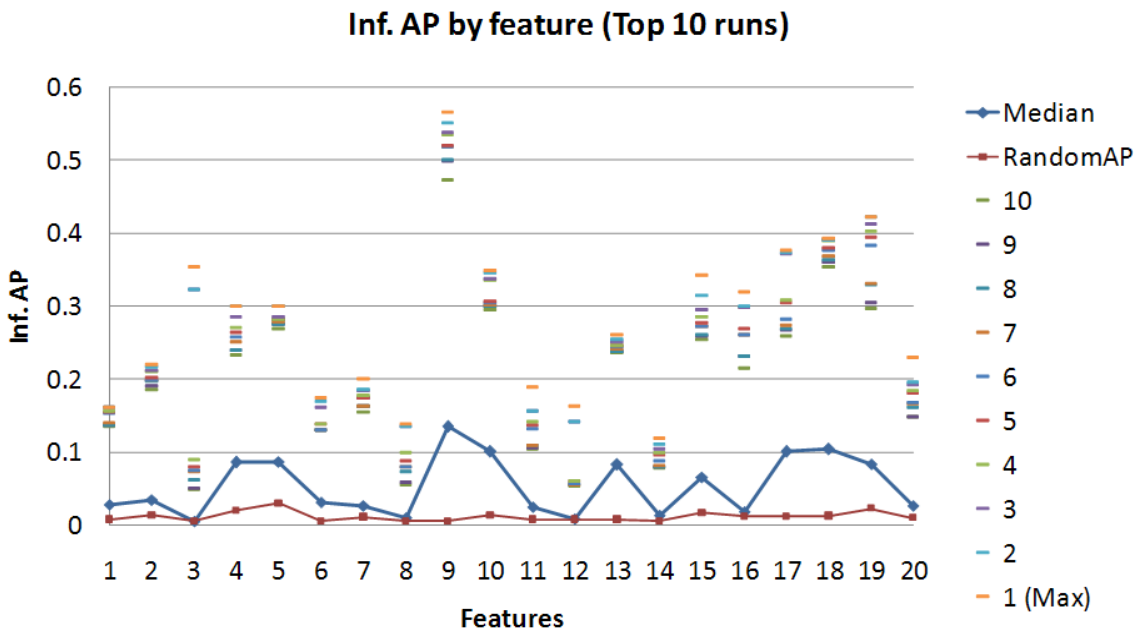


Figure 11: Effectiveness versus number of hits

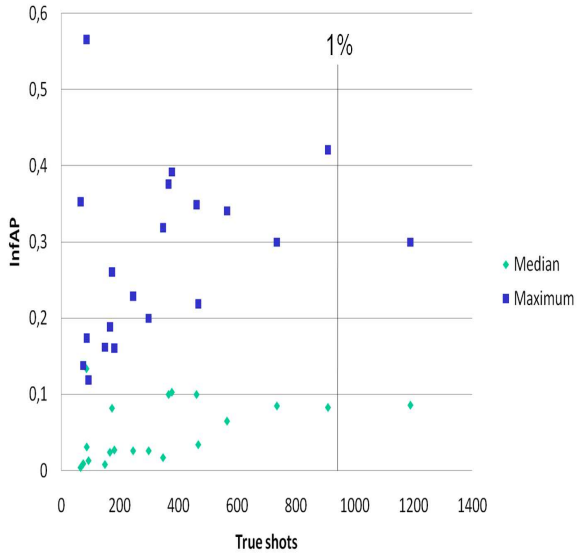


Figure 12: Significant differences among top A-category runs

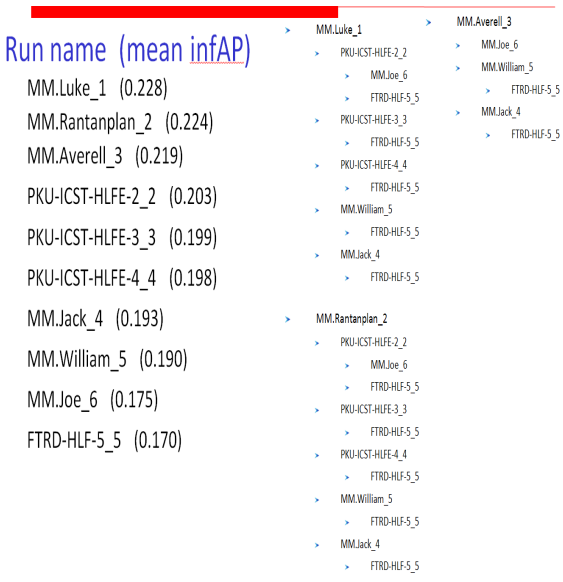


Figure 13: Significant differences among top A/a-category runs

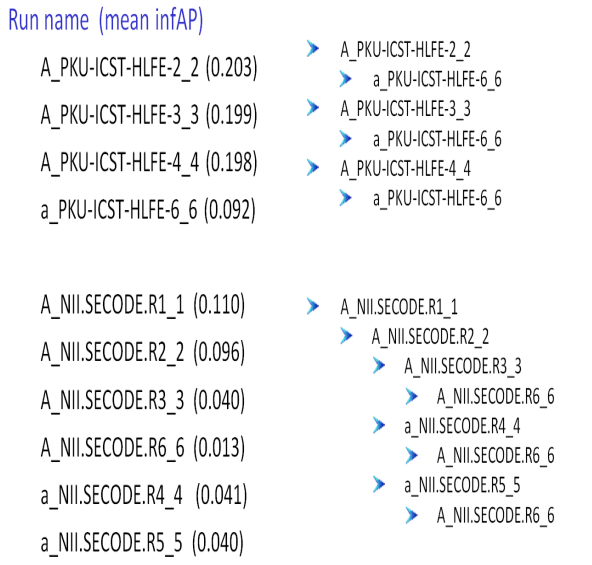


Figure 14: Significant differences among top C-category runs

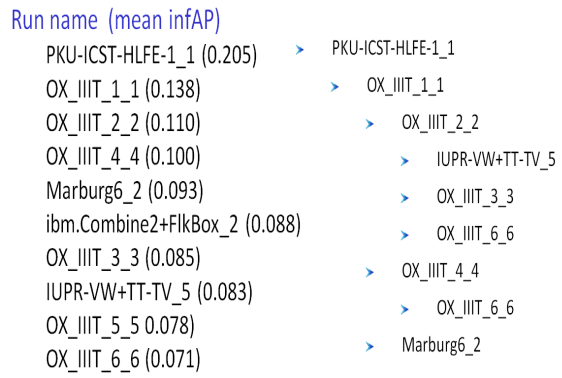


Figure 15: Significant differences among top c-category runs

Figure 16: Significant differences among top C/c-category runs

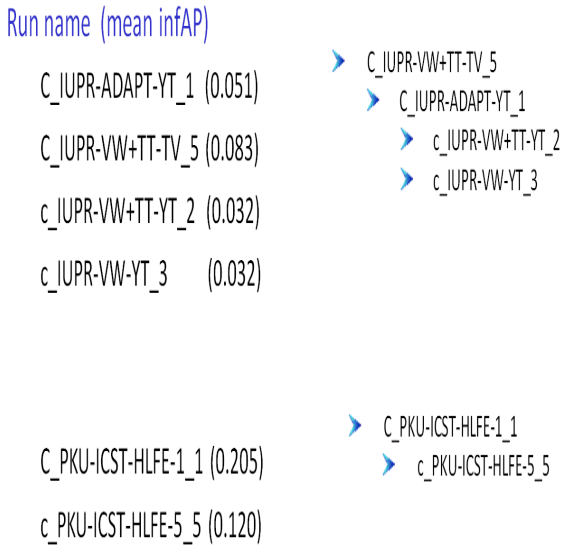


Table 1: Participants and tasks

Task				Location	Participants
−−	FE	SE	−−	Europe	Aristotle University of Thessaloniki - VITALAS
−−	FE	−−	CD	Asia	Asahikasei Co.
−−	−−	−−	CD	N.Amer.	AT&T Labs - Research
ED	FE	SE	CD	Asia	Beijing University of Posts and Telecom - BUPT-MCPRL
ED	∗∗	SE	−−	Asia	Beijing University of Posts and Telecom - RIS
∗∗	FE	∗∗	∗∗	Europe	Brno University of Technology
ED	FE	∗∗	−−	N.Amer.	Carnegie Mellon University
−−	FE	SE	−−	Europe	Centre for Research and Technology Hellas
∗∗	FE	−−	∗∗	Asia	Chinese Academy of Sciences-IVA NLPR IA CAS
−−	−−	SE	CD	Asia	Chinese Academy of Sciences-MCG-ICT-CAS
∗∗	FE	SE	CD	Asia	City University of Hong Kong - Columbia University
−−	−−	−−	CD	N.Amer.	Computer Research Institute of Montreal
−−	FE	−−	∗∗	N.Amer.	Florida International University
−−	FE	−−	−−	Europe	France Telecom Research & Development - Beijing
−−	FE	−−	CD	Asia	Fudan University
∗∗	FE	−−	∗∗	Asia	Fuzhou University
∗∗	FE	∗∗	∗∗	Europe	GDR ISIS - IRIM consortium
−−	FE	SE	−−	Europe	Helsinki University of Technology TKK
−−	∗∗	SE	∗∗	Europe	Hungarian Academy of Sciences
∗∗	FE	∗∗	CD	N.Amer.	IBM Watson Research Center
∗∗	FE	−−	−−	Europe	Institut EURECOM
−−	−−	−−	CD	Europe	Istanbul Technical University
−−	FE	−−	−−	Europe	IUPR - DFKI
∗∗	∗∗	−−	CD	Europe	JOANNEUM RESEARCH Forschungsgesellschaft mbH-JRS
−−	−−	SE	−−	N.Amer.	KB Video Retrieval
−−	∗∗	SE	−−	Asia	Kobe University
−−	FE	∗∗	−−	Europe	Laboratoire d'Informatique de Grenoble
−−	FE	−−	−−	Europe	Laboratoire d'Informatique Fondamentale de Marseille
∗∗	FE	SE	−−	Africa	Laboratoire REGIM
−−	FE	−−	−−	Europe	LSIS, Université Sud Toulon Var
−−	−−	−−	CD	Asia	Nanjing University
∗∗	FE	SE	CD	Asia	National Institute of Informatics
ED	∗∗	∗∗	∗∗	N.Amer.	NEC Laboratories America, Inc. and UIUC
ED	FE	∗∗	∗∗	Asia	NHK Science and Technical Research Laboratories
−−	FE	−−	−−	Europe	Oxford/IIIT
∗∗	FE	SE	∗∗	Asia	Peking University-PKU-ICST
ED	∗∗	−−	∗∗	Asia	Peking University-PKU-IDM
ED	FE	∗∗	−−	Asia	Shanghai Jiao Tong University-IICIP
−−	FE	−−	−−	Asia	Shanghai Jiao Tong University-IS
ED	−−	−−	−−	N.Amer.	Simon Fraser University
−−	−−	−−	CD	Europe	Telefonica I+D
−−	FE	−−	−−	Asia	The Institute of Statistical Mathematics
−−	∗∗	SE	−−	Europe	The Open University
−−	−−	−−	CD	Europe	TNO
ED	FE	−−	−−	Asia	Tokyo Institute of Technology
ED	−−	−−	∗∗	Asia	Toshiba Corporation
∗∗	FE	∗∗	CD	Asia	Tsinghua University-IMG
−−	FE	−−	CD	Asia	Tsinghua University-MPAM
ED	FE	−−	CD	Europe	TUBITAK UZAY
∗∗	FE	∗∗	∗∗	Europe	Universidad Autónoma de Madrid
−−	FE	−−	−−	Europe	Universidad Carlos III de Madrid
∗∗	FE	SE	−−	Europe	University of Amsterdam
−−	−−	−−	CD	Europe	University of Brescia
−−	FE	−−	−−	N.Amer.	University of Central Florida
∗∗	FE	∗∗	−−	Asia	University of Electro-Communications

Task legend. CD:copy detection; ED:event detection; FE:feature detection; SE:search; ∗∗:no runs submitted

Table 2: Participants and tasks (continued)

Task				Location	Participants
−−	**	SE	**	Europe	University of Glasgow
−−	FE	−−	−−	Europe	University of Karlsruhe (TH)
−−	FE	**	−−	Europe	University of Marburg
**	−−	−−	CD	N.Amer.	University of Ottawa
−−	−−	SE	−−	Europe	University of Surrey
−−	FE	**	−−	Europe	UPS - IRIT - SAMoVA
−−	FE	**	CD	Asia	Xi'an Jiaotong University
−−	FE	SE	−−	Asia	Zhejiang University

Task legend. CD:copy detection; ED:event detection; FE:feature detection; SE:search; **:no runs submitted

Table 3: Feature pooling and judging statistics

Feature number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number true	% judged that were true
1	406941	59136	14.5	80	3348	5.7	181	5.4
2	404807	58184	14.4	80	3497	6.0	467	13.4
3	387861	63400	16.3	70	3369	5.3	66	2.0
4	404429	52530	13.0	120	3454	6.6	735	21.3
5	405015	52793	13.0	100	3299	6.2	1190	36.1
6	403666	51866	12.8	120	3317	6.4	87	2.6
7	411765	55195	13.4	90	3394	6.1	298	8.8
8	413014	62055	15.0	80	3330	5.4	75	2.3
9	406066	55626	13.7	110	3358	6.0	86	2.6
10	410959	45891	11.2	120	3460	7.5	461	13.3
11	398981	59252	14.9	100	3387	5.7	166	4.9
12	399963	65656	16.4	70	3402	5.2	149	4.4
13	399973	70608	17.7	90	3398	4.8	173	5.1
14	396976	57275	14.4	100	3364	5.9	93	2.8
15	413236	55297	13.4	110	3324	6.0	565	17.0
16	403629	59882	14.8	80	3377	5.6	347	10.3
17	411349	38875	9.5	170	3383	8.7	366	10.8
18	417686	48846	11.7	130	3389	6.9	377	11.1
19	415069	49565	11.9	100	3436	6.9	909	26.5
20	406484	62973	15.5	90	3488	5.5	245	7.0

Figure 19: Search runs by type

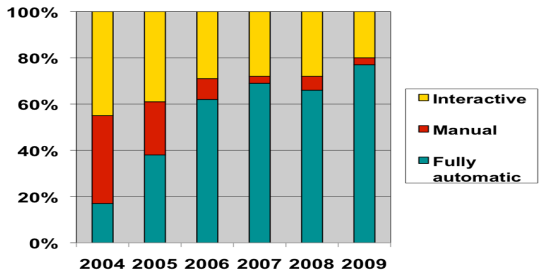
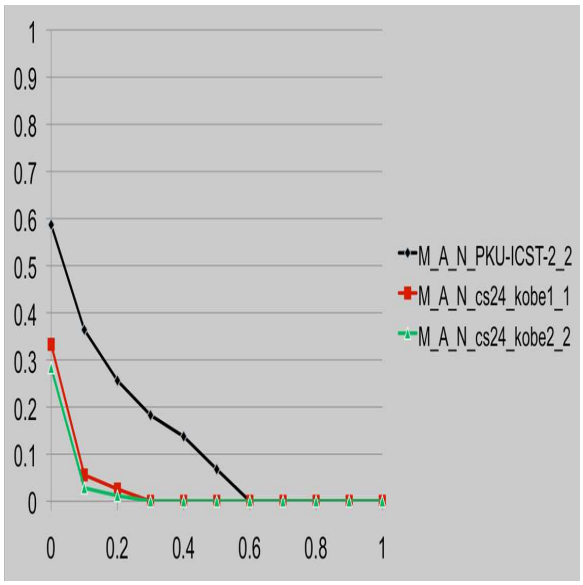


Figure 20: Manual search runs



4 Search

The search task in TRECVID was an extension of its text-only analogue. Video search systems were presented with multimedia topics — formatted descriptions of an information need — and were asked to return a list of up to 1 000 shots from the videos in the search test collection which met the need. The list was to be prioritized based on likelihood of relevance to the need expressed by the topic. A "high precision" option was added in which participants returned only the top 10 shots for each topic.

Figure 21: Top 10 normal automatic search runs

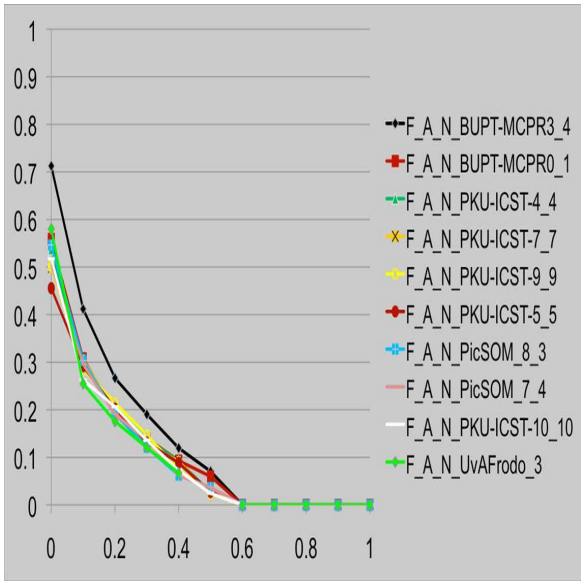


Figure 22: Top 10 normal interactive search runs

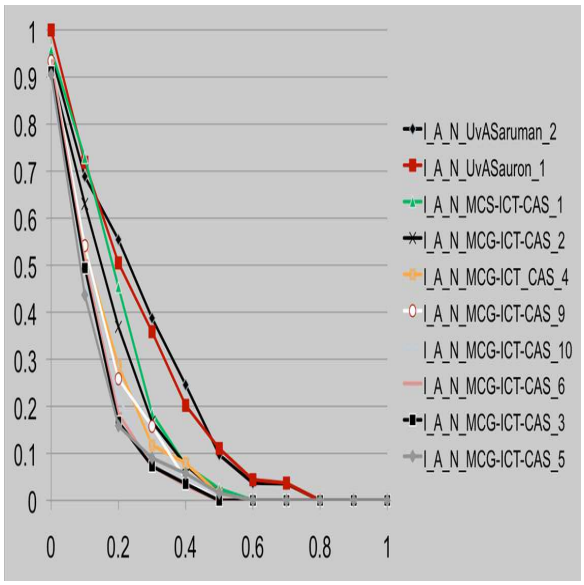


Figure 17: AP by topic

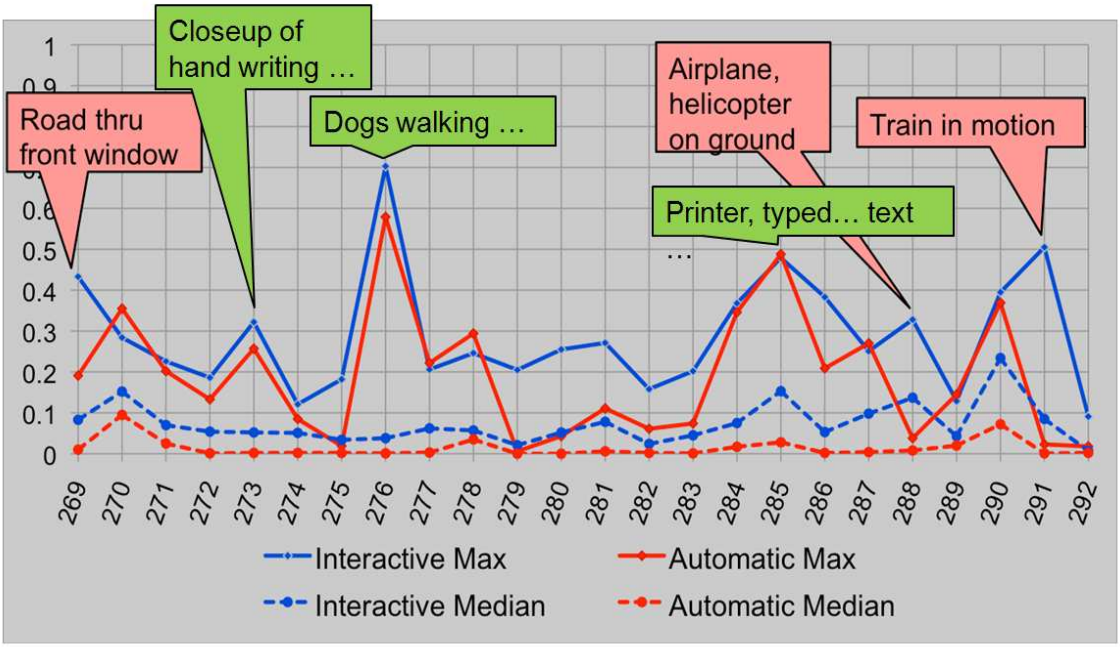


Figure 18: Hits in the test set by topic

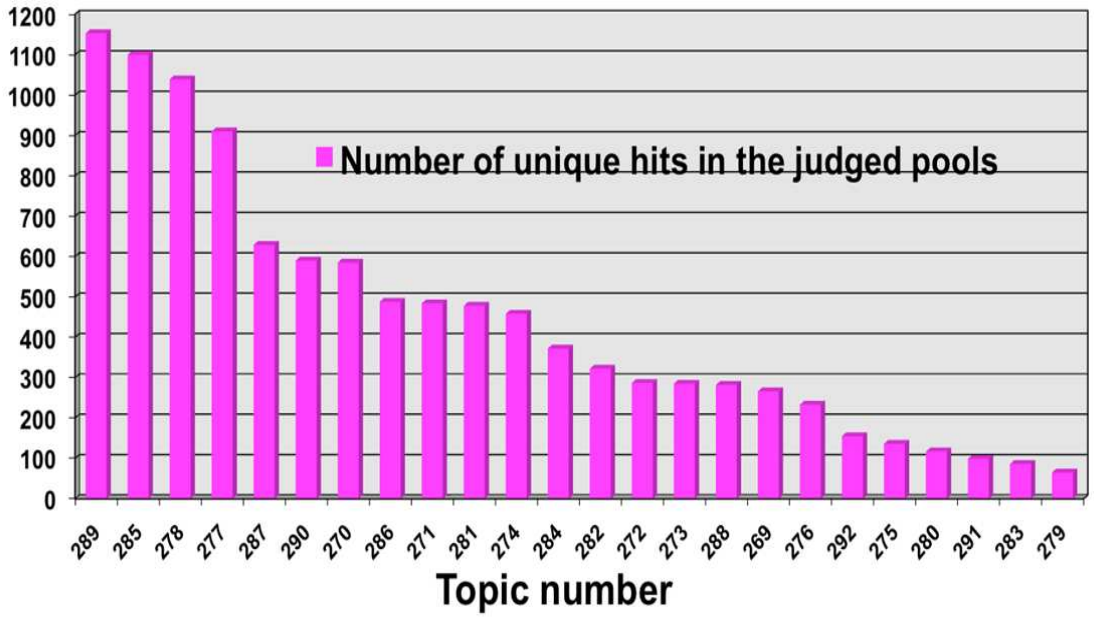


Figure 23: Randomization test results for high precision search

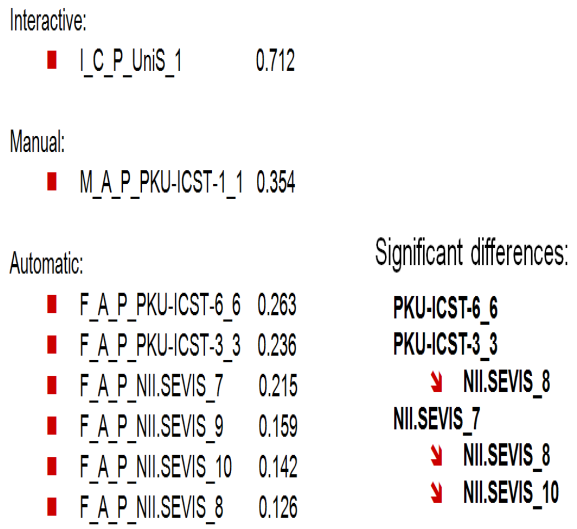


Figure 24: Randomization test results for top 10 normal automatic search

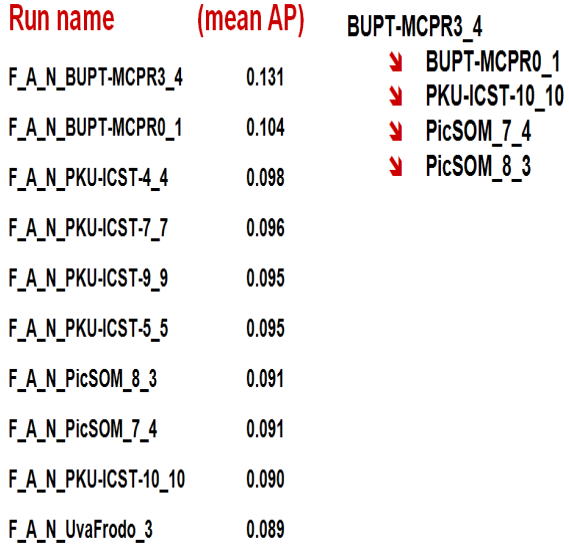


Figure 25: Randomization test results for top 10 normal interactive search

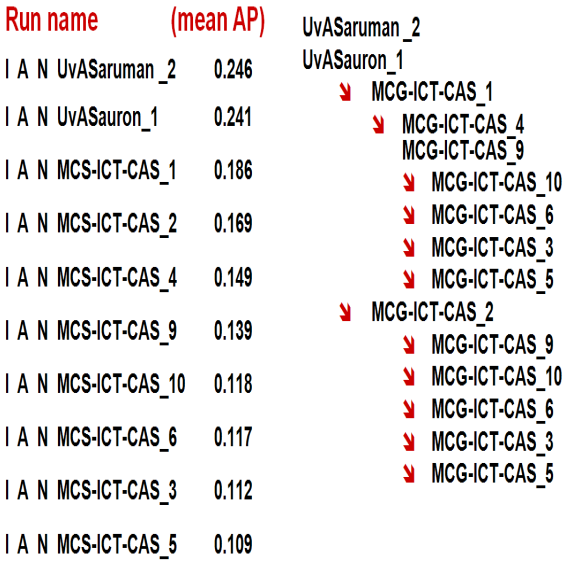
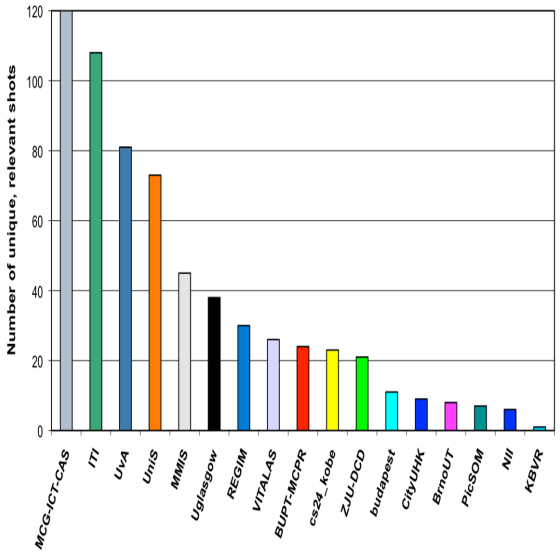


Figure 26: Unique relevant by team



4.1 Interactive, manually assisted, and automatic search

As was mentioned earlier, three search modes were allowed: fully interactive, manually assisted, and fully automatic. A big problem in video searching is that topics are complex and designating the intended meaning and interrelationships between the various pieces — text, images, video clips, and audio clips — is a complex one and the examples of video, audio, etc. do not always represent the information need exclusively and exhaustively. Understanding what an image is of/about is famously complicated (Shatford, 1986).

The definition of the manual mode for the search task allowed a human, expert in the search system interface, to interpret the topic and create an optimal query in an attempt to make the problem less intractable. The cost of the manual mode in terms of allowing comparative evaluation is the conflation of searcher and system effects. However if a single searcher is used for all manual searches within a given research group, comparison of searches within that group is still possible. At this stage in the research, the ability of a team to compare variants of their own system is arguably more important than the ability to compare across teams, where results are more likely to be confounded by other factors hard to control (e.g. different training resources, different low-level research emphases, etc.).

The advantage of using visual information over just textual information from the video (e.g. derived from speech) has been demonstrated with respect to the Sound and Vision data, so no text-only baseline was required in 2009.

4.2 Data

As mentioned earlier, the feature test collection contained 419 files/videos but seven test files were ignored in the testing due to problems displaying shots from these long files (BG_36684, BG_37970, BG_38162, BG_8887, BG_37942, BG_8650, and BG_38653) in the assessment system. Removing these files left 412 files and 93 902 shots.

Search submissions were categorized by the submitted group based on what sort(s) of training data were used - just as in the high-level feature extraction task, described in Section (3.2).

4.3 Topics

Because the topics have a huge effect on the results, the topic creation process deserves special attention here. Ideally, topics would have been created by real users against the same collection used to test the systems, but such queries are not available.

Alternatively, interested parties familiar in a general way with the content covered by a test collection could have formulated questions which were then checked against the test collection to see that they were indeed relevant. This is not practical either because it pre-supposed the existence of the sort of very effective video search tool which participants are working to develop.

What was left was to work backwards from the test collection with a number of goals in mind. Rather than attempt to create a representative sample, NIST has in the past tried to get an approximately equal number of each of the basic types (generic/specific and person/thing/event), though in 2006 generic topics dominated over specific ones. The 2009 topics are all generic due to the diversity of the collection and the resulting difficulty finding enough examples of named people, objects, events, or places. Generic topics may be more dependent on the visual information than the specific, which usually score high on text-based (baseline) search performance. Also, the 2009 topics reflect a deliberate emphasis on events.

Another important consideration was the estimated number of relevant shots and their distribution across the videos. The goals here were as follows:

- For almost all topics, there should be multiple shots that meet the need.
- If possible, relevant shots for a topic should come from more than one video.
- As the search task is already very difficult, we don't want to make the topics too difficult.

NIST developed 24 multimedia topics for use in testing the search systems. The topics express the need for video (not just information) concerning people, things, events, etc. and combinations of the former. The topics were designed to reflect many of the various sorts of queries real users pose: requests for video with specific people or types of people, specific objects or instances of object types, specific activities or instances of activity (Enser & Sandom, 2002).

The topics were constructed based on a review of the test collection for relevant shots. The topic creation process was designed to eliminate or reduce tun-

Table 5: Search type statistics

Search type	'04	'05	'06	'07	'08	'09
%Fully automatic	17	38	62	69	66	77
%Manually assisted	38	23	9	3	7	3
%Interactive	45	39	29	28	27	20

ing of the topic text or examples to the test collection. Potential topic targets were identified while watching the test videos with the sound off. Non-text examples were chosen without reference to the relevant shots found. When more examples were found than were to be used, the subset used was chosen at random. The topics are listed in Appendix A. A rough classification of topic types for TRECVID 2009 based on Armitage & Enser, 1996, is provided in Table 6.

4.4 Evaluation

Groups were allowed to submit a total of up to 10 runs of any types in the search task. In fact 19 groups submitted a total of 123 runs — 25 interactive runs, 4 manual ones, and 94 fully automatic ones. Of the 123 runs, high precision runs formed a very small subset: 1 manual, 1 interactive, and 6 automatic.

All submitted runs from each participating group contributed to the evaluation pools. For each topic, all submissions down to a depth of at least 40 (average 76, maximum 100) result items (shots) were pooled, duplicate shots were removed and randomized. Human judges (assessors) were presented the pools — one assessor per topic — and they judged each shot by watching the associated video and listening to the audio. The maximum result set depth judged and pooling and judging information for each topic are listed in Table 4 for details.

4.5 Measures

The *trec_eval* program was used to calculate recall, precision, and average precision based on the judgment pools. Following (Webber, Moffat, Zobel, & Sakai, 2008), average precision was used to predict precision at 10 for the high precision subtask.

4.6 Results

Participation in the search task was concentrated as usual in the automatic runs though interactive experiments continue despite the attendant complexities (see Figure 19). The high-precision subtask attracted very few groups and so will not be discussed in detail here.

The test collection for 2009 was similar to those for 2007 and 2008 except in size. It was the largest ever - almost twice as large as in 2008. For 4 topics the number of hits found in the judged pools (and doubled to compensate for the 50% sample) ranged from 900 - 1150 as shown in Figure 18.

Results for the top runs are shown as precision/recall curves in Figures 22, 21, and 20.

Simple rankings of runs based on a score such as inferred average precision do not provide any information about which differences might be due primarily to chance rather than technical approach. This question is particularly important when the absolute values of the scores are quite small. Randomization tests provide evidence for which difference between runs are real. Results of such tests on the top runs for the high precision subtask, normal automatic search, and normal interactive search are shown in Figures 23, 24, and 25 respectively. There are significant differences but many runs tend to form groups within which the test finds none.

Results by topic (Figure 17) show the usual large amount of variation with best results significantly better than the median in most cases. For about one third of the topics in 2009 the best automatic system came close to or exceeded the effectiveness of the best interactive system as measured by inferred average precision.

4.7 Approaches

The Beijing University of Posts and Telecom (BUPT-MCPRL) team submitted automatic runs, using high-level features/concepts, and visual, example-based retrieval. They weighted the combination as multimodal fusion, then included face scores. Their 10 runs were various combinations of the above; use of Weight Distribution based on Semantic Similarity (WDSS) yielded top automatic run performance.

The Brno University of Technology automatic runs were based on transformed local image features (points, edges, homogeneous regions), i.e. scale-invariant feature transforms (SIFT). They used face detection and global features, and then color layout and texture features. They were similar to submissions in previous years..

The Budapest Academy of Sciences team (Hungarian Academy of Sciences) employed linear combina-

Table 4: Search pooling and judging statistics

Topic number	Total submitted	Unique submitted	% total that were unique	Max. result depth pooled	Number judged	% unique that were judged	Number relevant	% judged that were relevant
269	104392	28458	27.3	100	4456	15.7	266	6.0
270	103600	27214	26.3	100	4329	15.9	585	13.5
271	103704	28629	27.6	80	3595	12.6	484	13.5
272	103425	35147	34.0	70	4154	11.8	287	6.9
273	103843	34734	33.4	60	3614	10.4	285	7.9
274	104254	35653	34.2	90	4911	13.8	458	9.3
275	103457	36045	34.8	50	3198	8.9	136	4.3
276	104844	33512	32.0	80	4357	13.0	233	5.3
277	104262	34340	32.9	90	4860	14.2	910	18.7
278	103781	28629	27.6	100	4527	15.8	1039	23.0
279	104643	41228	39.4	40	2859	6.9	65	2.3
280	102233	40631	39.7	100	6070	14.9	117	1.9
281	102348	30741	30.0	100	4928	16.0	478	9.7
282	104710	40493	38.7	50	3306	8.2	322	9.7
283	103924	38914	37.4	40	2677	6.9	86	3.2
284	104095	29773	28.6	60	3073	10.3	372	12.1
285	104851	28057	26.8	100	4621	16.5	1100	23.8
286	104335	38764	37.2	100	5842	15.1	488	8.4
287	104142	33374	32.0	60	3357	10.1	629	18.7
288	103964	29003	27.9	70	3371	11.6	282	8.4
289	103845	33331	32.1	50	2761	8.3	1153	41.8
290	104646	24382	23.3	100	3949	16.2	590	14.9
291	102735	31755	30.9	80	4134	13.0	99	2.4
292	104071	33284	32.0	60	3337	10.0	155	4.6

tions of ASR text, image similarity of representative frames, face detector output for topics involving people, weight of high level feature classifiers considered relevant by text based similarity to the topic, motion information extracted from videos where relevant to topic, plus some shot contexts (neighbor shots).

The Centre for Research and Technology Hellas ITI/CERTH Thessaloniki conducted experiments in interactive search, combining retrieval functionalities in various modalities (i.e. textual, visual, and concept search) with a user interface supporting interactive search over all queries submitted.

The Chinese Academy of Sciences (MCG-ICT-CAS) interactive search runs used a "VideoMap" system with a map-based display interface, giving a global view of similarity relationships throughout the whole video collection. The system incorporated multiple modality feedback strategies, including the visual-based feedback, concept-based feedback, and community-based feedback.

The City University of Hong Kong and Columbia University team submitted automatic search runs. In previous years they focused on concept-based search, using various techniques to determine which concepts to use, include Flickr usage. In 2009 they also factored in visual query examples and addressed a combination of multiple search modalities. Multimodal search fusion yielded 10% improvement.

At the Helsinki University of Technology the team submitted automatic runs that combined text from automatic speech recognition and machine translation in search and concept-based retrieval. If none of the concept models could be matched with the query, they used content-based retrieval based on the video and image examples instead. A portfolio of 10 runs was submitted with text, visual similarity,their own concepts, and donated (MediaMill and CU-VIREO374) concepts individually, and in combinations.

KB Video Retrieval (David Etter) worked on au-

tomatic search with a focus on query expansion by adding terms (text) and images, using Wikipedia titles and images as a source

Kobe University also explored a form of detailed relevance feedback in their submission. They used rough set theory to combine the evidence provided by positive and negative shot examples and in particular they determined the characteristics that distinguish positive and negative feedback, thus making the most of positive and negative examples.

The Laboratoire REGIM combined text search (against automatic speech recognition transcripts) and visual search (color, texture, shape) from keyframes

The National Institute of Informatics submitted automatic runs only. They, trained an support vector machine (SVM) concept detector for each query, also used k nearest neighbor matching on visual features, concept selection using visual features, and concept selection using text descriptions

Peking University (PKU-ICST) submitted automatic and manual search runs with list of in-house multimodal variations including weighted combination of visual-based, concept-based, audio features, and faces for some topics. Two retrieval approaches - pairwise similarity and learning-based ranking - gave good performance

The Open University team fielded 8 automatic search submissions based on determining the distance from a query image to a pre-indexed collection of images to build a list of results ordered by visual similarity. They experimented with four metric measures (Euclidian, Manhattan, Canberra and Squared Chord) and two data normalizations.

The MediaMill team from the University of Amsterdam adapted an approach in their interactive search submissions of helping searchers to find good retrieval strategies in their search. This was supported through a process or active zooming as different retrieval strategies were used, and also incorporating a type of latent or implied relevance feedback which involved passive sampling of the searchers' browsing behaviour in order to provide good (and bad) documents for the relevance feedback process.

The University of Glasgow submitted automatic runs based on MPEG-7 features, concepts, and bag-of-words derived from SIFT features. They investigated estimating topic distribution using the Latent Dirichlet Allocation (LDA) with run variants to explore this median performance.

The VITALAS project is a large European Union project and the submission from this large team was coordinated by CWI Amsterdam and the Aristotle University, Thessaloniki, Greece. The work involved a detailed study of the search performance of a set of novice and a different set of professional searchers, examining how they searched, and performed, in interactive searching.

At the time this overview was created no details about the approaches taken by the following teams were available: Beijing University of Posts and Telecom (PRIS), University of Surrey, and Zhejiang University.

5 Copy detection

As used here, a copy is 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. Detecting copies is important for copyright control, business intelligence and advertisement tracking, law enforcement investigations, etc. Content-based copy detection offers an alternative to watermarking. In TV2008, the TRECVID copy detection pilot task was carried out in collaboration with members of the IMEDIA team at INRIA and built on the Video Copy Detection Evaluation Showcase at CIVR 2007.

Based on feedback from last year's participants, some modifications were designed for TV2009 task. First, 3 transformations were dropped to make the queries more realistic and not too extreme. Second, systems were required to submit runs for two required tasks (video-only queries and video + audio queries) and one optional (audio-only queries). Third, two application profiles were required to be simulated for each submitted query type. One requires a balanced cost for misses and false alarms. The other (Nofa) requires no false alarms (i.e., sets a very high cost for false alarms). Fourth, systems were required to submit a decision score threshold believed to correspond to the best performance for the run.

The required system task was as follows: given a test collection of videos and a set of 1407 queries (video-only segments), determine for each query the place, if any, that some part of the query occurs, with possible transformations, in the test collection. Two thirds of the queries contained copies.

A set of 7 possible transformations was selected to reflect actually occurring transformations and applied to each of 201 untransformed (base) queries using tools developed by IMEDIA to include some randomization at various decision points in the construction of the query set. For each query, the tools took a segment from the test collection, optionally transformed it, embedded it in some video segment which did not occur in the test collection, and then finally applied one or more transformations to the entire query segment. One third of the queries contained no test segment; another third were composed entirely of the test segment. Video transformations included, picture-in-picture (T2), insertion of patterns (T3), reencoding (T4), change of gamma (T5), decreasing the quality (T6), and post production alterations (T8), and randomly choosing 3 transformations (T10). The video transformations used were documented in detail as part of the TRECVID Guidelines.

Since detection of untransformed audio copies is relatively easy, and the primary interest of the TRECVID community is in video analysis, it was decided to model the required copy detection tasks with video-only and video+audio queries. However, since audio is of importance for practical applications, there was one additional optional task: a task using transformed audio-only queries.

1407 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 201 base queries was transformed by seven techniques that were intended to be typical of those that would occur in real reuse scenarios: doing nothing (T1), mp3 compression (T2), mp3 compression and multiband companding (T3), bandwidth limit and single-band companding (T4), mix with speech (T5), mix with speech then multiband compress (T6), and bandpass filter, mix with speech and compression (T7).

A script to construct 9849 audio + video queries was provided by NIST. These queries comprised all the combinations of transformed audio(7) and transformed video (7) from a given base audio+video query (201). In this way participants could study the effectiveness of their systems for individual audio and video transformations and their combinations.

5.1 Data

All of the 2007 and 2008 Sound and Vision data (400 hours) and 2009 (180 hours) were used as a source for reference video in testing and development. The 2007 and 2008 BBC rushes video (53 hours) and 2009 BBC rushes data (30 hours) was used as a source for non-reference video.

5.2 Evaluation

In total in 2009, 20 participant teams submitted 107 runs for evaluation. 53 runs were submitted for video-only evaluation, 12 runs for audio-only and 42 runs for mixed (audio+video). Copy detection submissions were evaluated separately for each transformation, according to:

- How many queries they find the reference data for or correctly tell us there is none to find
- When a copy is detected, how accurately the run locates the reference data in the test data.

- How much elapsed time is required for query processing

5.3 Measures (per transformation)

- Minimal Normalized Detection Cost Rate: a cost-weighted combination of the probability of missing a true copy and the false alarm rate. For TRECVID 2009 the cost model assumed copies are very rare (e.g. 0.5/hr) then two application profiles were required. The "Balanced" profile in which misses and false alarms are assigned a cost of 1, and the "Nofa" profile in which a false alarm is assigned a cost of 1000 times the cost of a miss. Other realistic scenarios were of course possible. Normalized minimal detection cost rate (minNDCR) reduced in 2009 to two terms involving two variables: the number of a misses (false negatives: FN) and the number of false alarms (false positives: FP). The total length of queries in hours was 36.49 for audio and for video queries, 255.40 for audio + video queries. For example for the video-only queries with 7 possible transformations, under the "Nofa" profile:

  $$minNDCR = 0.007 * FN + 384.6 * FP$$

  For the same queries under the "Balanced" profile:

  $$minNDCR = 0.007 * FN + 0.38 * FP$$

- Copy location accuracy: mean F1 score combining the precision and recall of the asserted copy location versus the ground truth location

- Copy detection processing time: mean processing time (s)

Finally, the submitted run threshold were used to calculate the actual NDCR and F1 and those results were compared to the minNDCR and F1 using the optimal threshold calculated by the DET curve.

5.4 Results

Results presented here are for each query type (audio-only, video-only and audio+video) separately. For each query type we will present the results of the two application profiles (balanced and no false alarms) based on the optimum threshold and based on the submitted actual run threshold.

Figure 27: Top "video-only" runs based on Actual DET score in balanced profile

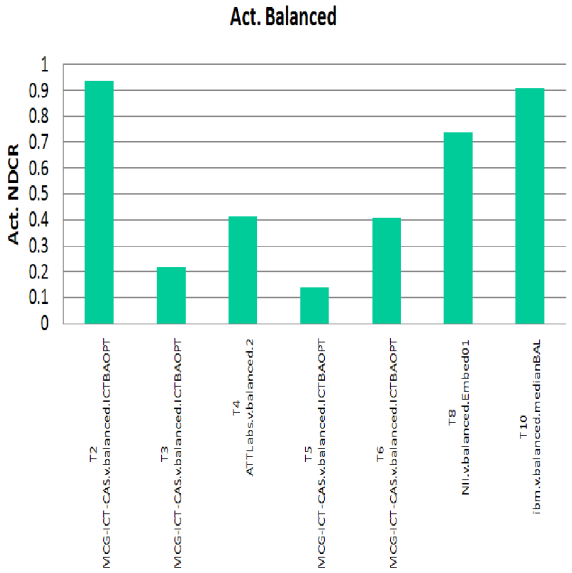


Figure 28: Top "video-only" runs based on Actual DET score in NoFA profile

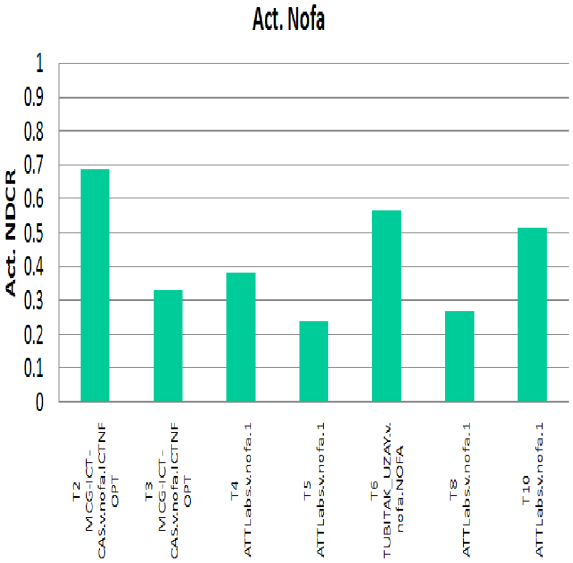


Figure 31: Top "video+audio" runs based on Actual DET score in balanced profile

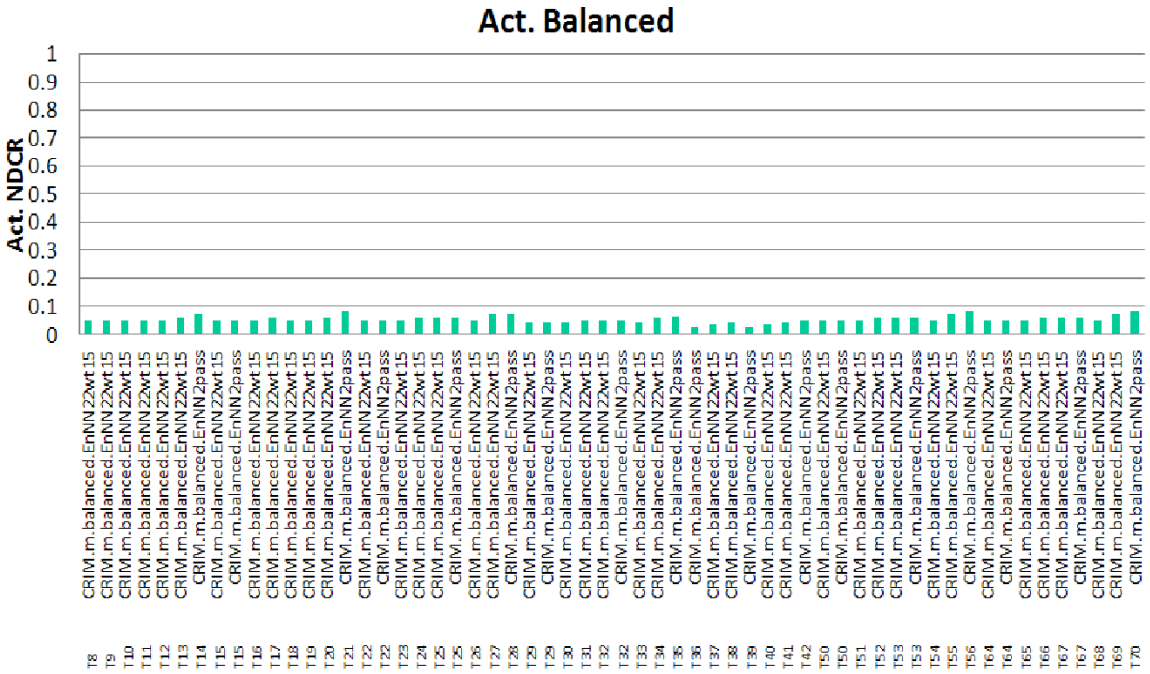


Figure 32: Top "video+audio" runs based on Actual DET score in NoFA profile

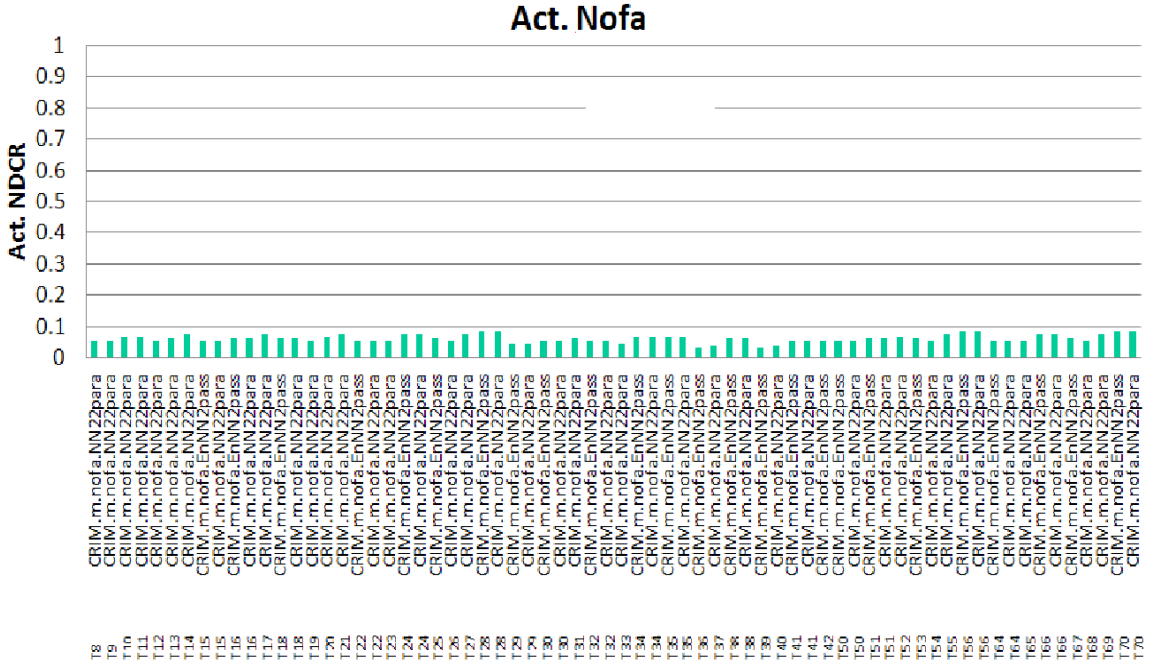


Figure 33: Top "video+audio" runs based on Optimum DET score in balanced profile

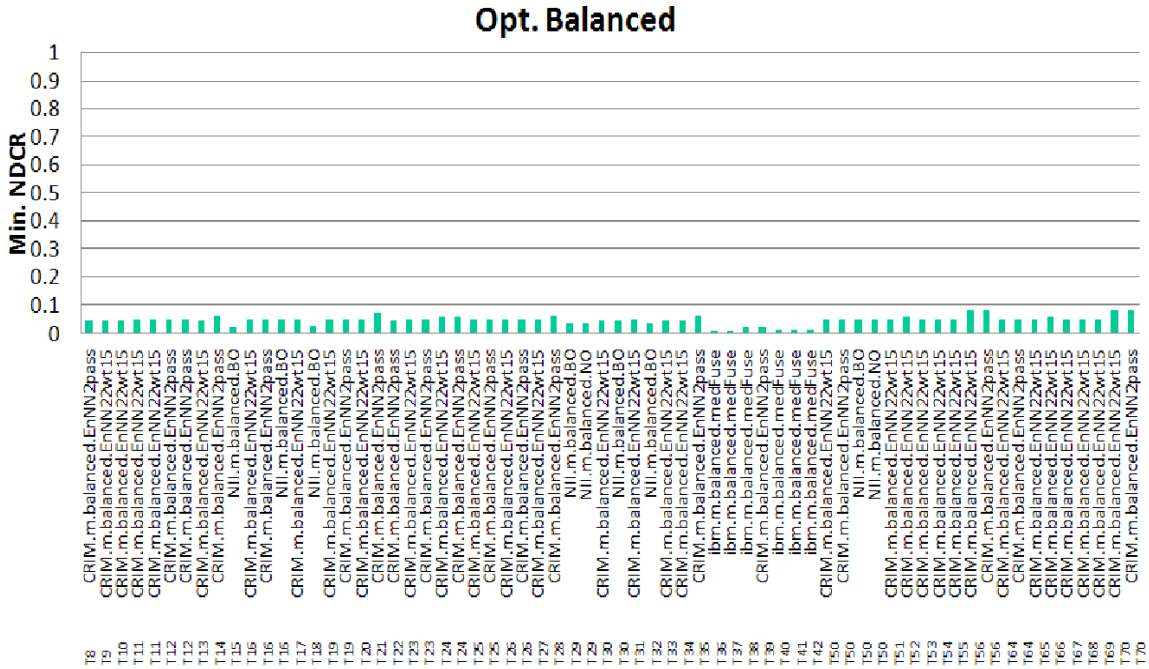


Figure 34: Top "video+audio" runs based on Optimum DET score in NoFA profile

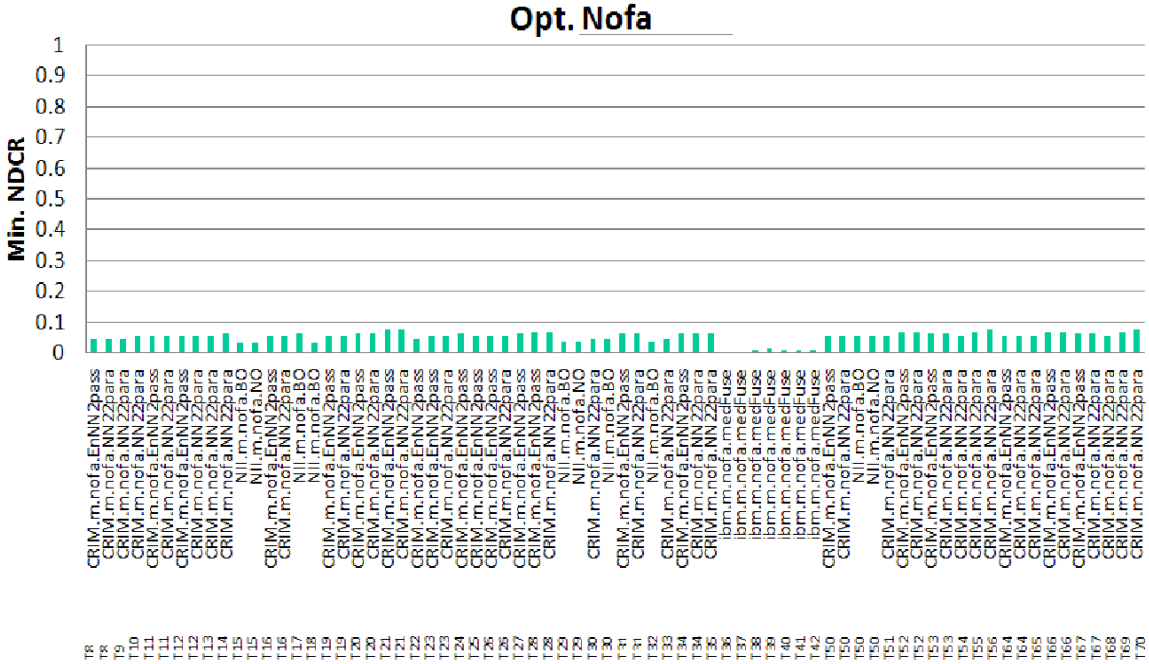


Figure 29: Top "video-only" runs based on Optimum DET score in balanced profile

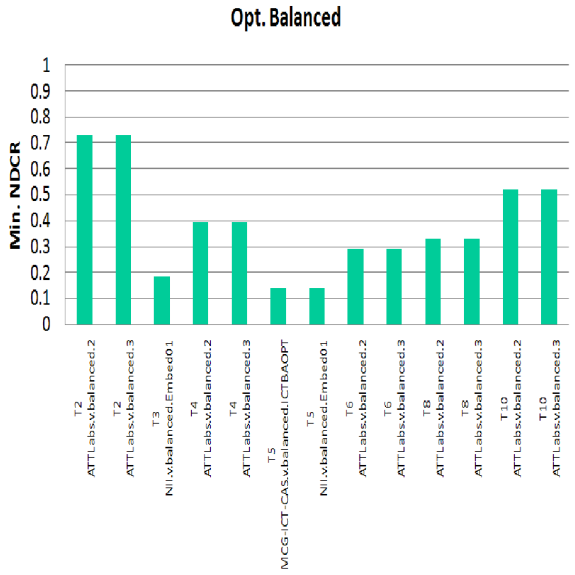


Figure 30: Top "video-only" runs based on Optimum DET score in NoFA profile

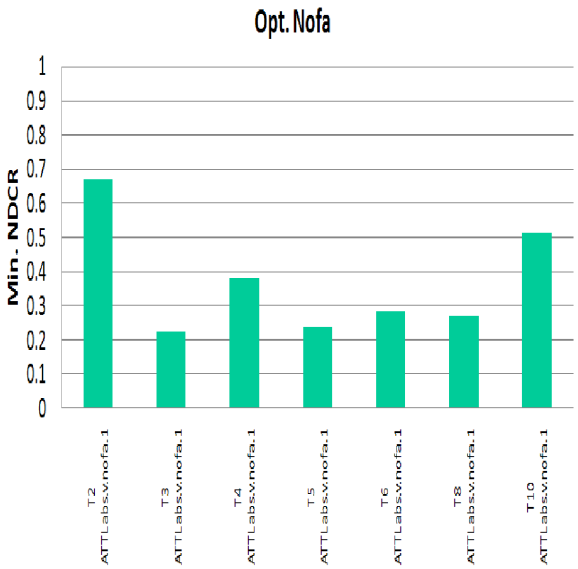
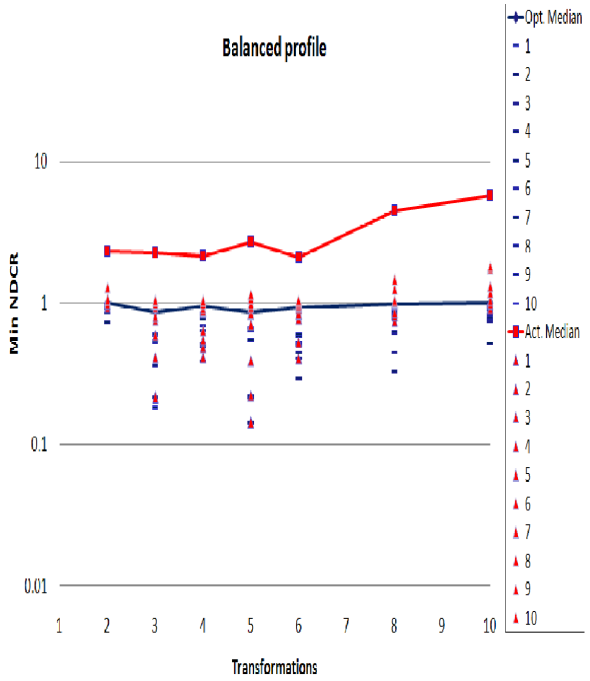


Figure 35: video only top 10 detection performance based on balanced profile



Comparing the top runs' DET scores per transformation for the three query types we found that for the audio-only queries, systems achieved good detection (less than 0.1 DET cost rate) 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 are much more mature and more advanced than video detection. Video-only queries (Figures 27 to 30) achieved a worse performance than the audio-only as DET scores range across transformations vary a lot and reached above 0.9. This indicates that systems have difficulties with some transformations compared to audio-only scores. Video+audio queries (Figures 31 to 34) top scores were much better than video-only across all transformation combinations as it didn't exceed the 0.1 DET cost rates in the two profiles using both the actual and optimum thresholds. This again indicates that the audio feature helps in video copy detection.

To visualize the difference between the actual vs optimal results, Figures 35 through 37 show the performance of the top 10 runs for the three query types for the balanced profile. It is clear that there is a difference between the optimum median and the ac-

Figure 36: audio-only top 10 detection performance based on balanced profile

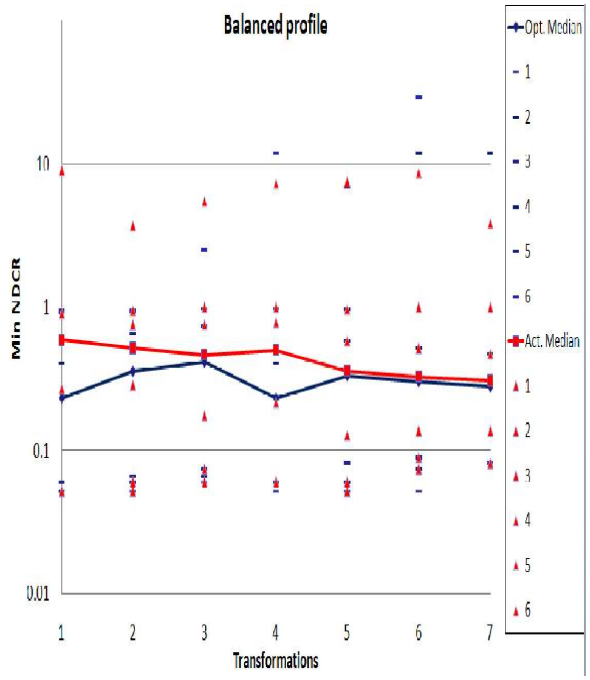


Figure 37: video+audio top 10 detection performance based on balanced profile

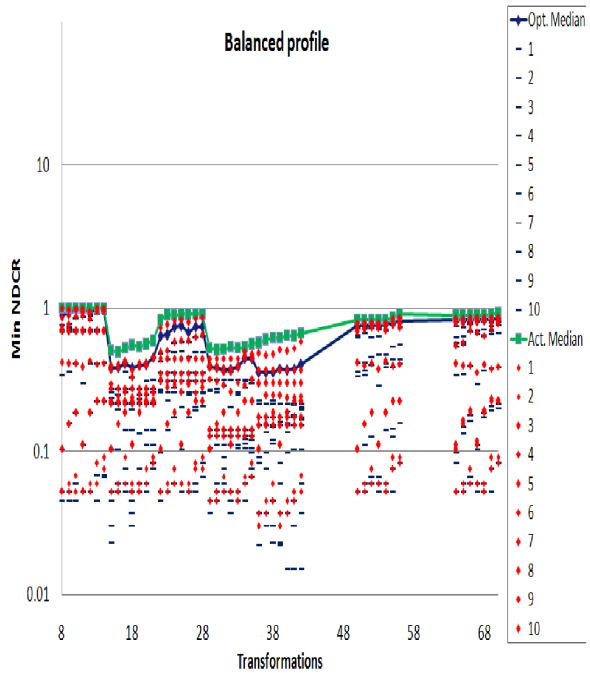
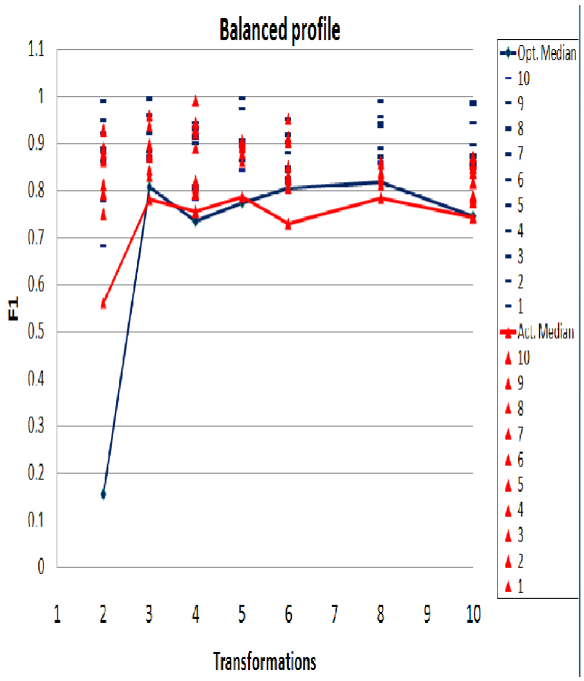


Figure 38: video-only top 10 localization performance based on balanced profile



tual median which indicates that choosing the run threshold is not a trivial task for systems. It can also be noted that this difference is smaller in audio and video+audio compared to video-only.

The same comparison between actual and optimum based on localization performance of the top 10 runs for the three query types is shown in Figures 38 through 40. It is clear that the audio-only median localization is much accurate than video-only or video+audio queries although only 6 runs were submitted for each profile.

Efficiency comparison among the top 10 is shown in Figures 41 through 43. 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 seconds, while the median values among the three query types were generally above 100 seconds.

The evaluation of the three main measures (detection, localization and efficiency) for the top 10 runs based on Nofa profile is shown in Figures 44 through 52. The detection performance for the Nofa runs in general seems to be more difficult than the balanced profile. It can be shown that the optimum median

Figure 39: audio-only top 10 localization performance based on balanced profile

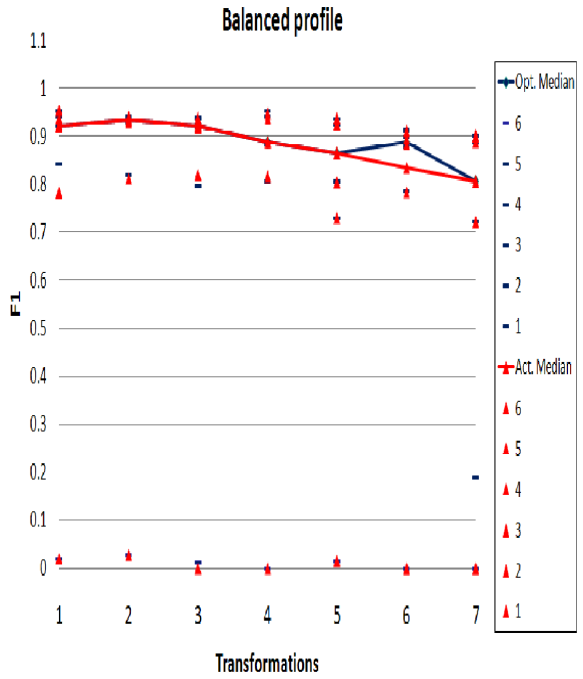


Figure 40: video+audio top 10 localization performance based on balanced profile

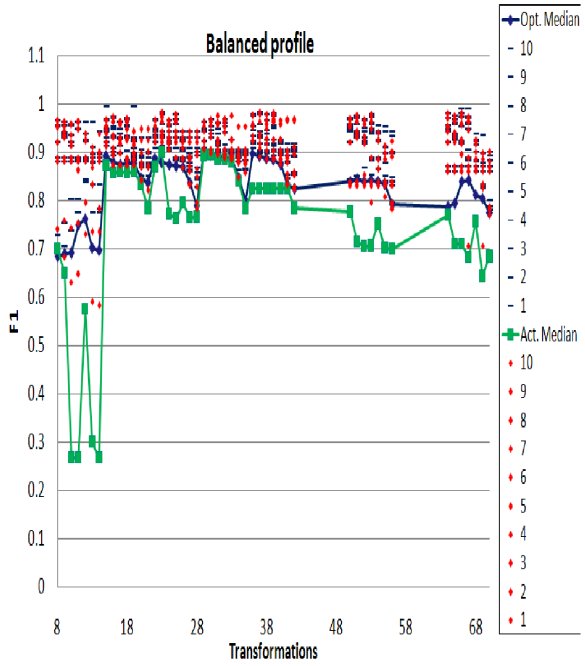


Figure 41: video-only top 10 efficiency performance based on balanced profile

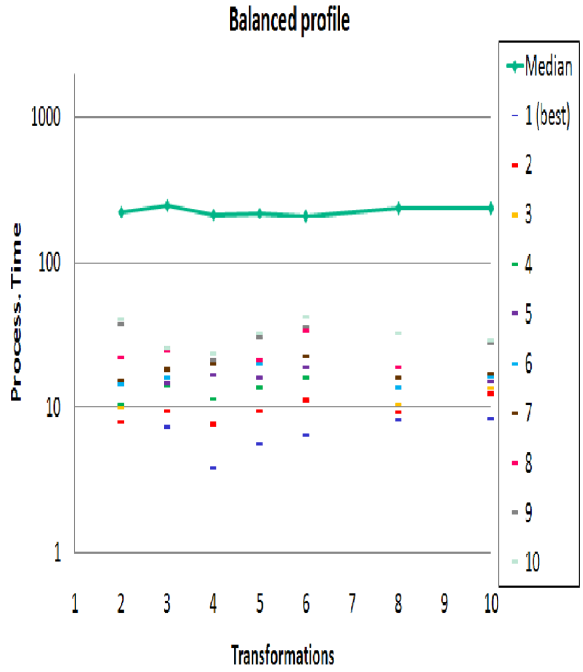


Figure 42: audio-only top 10 efficiency performance based on balanced profile

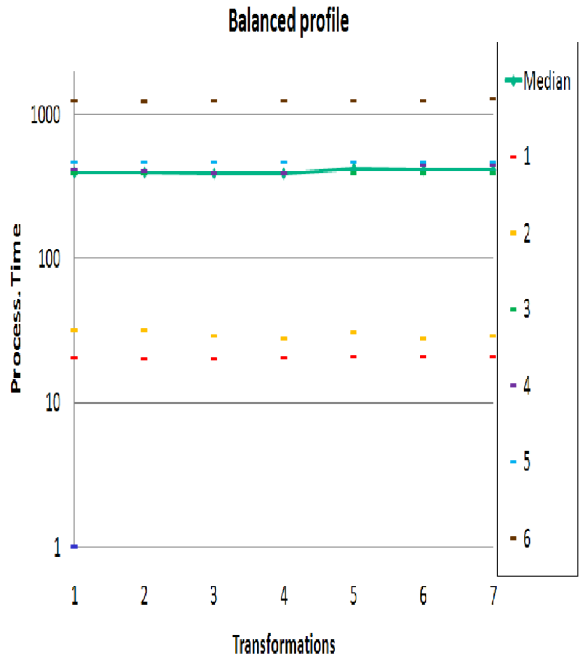


Figure 43: video+audio top 10 efficiency performance based on balanced profile

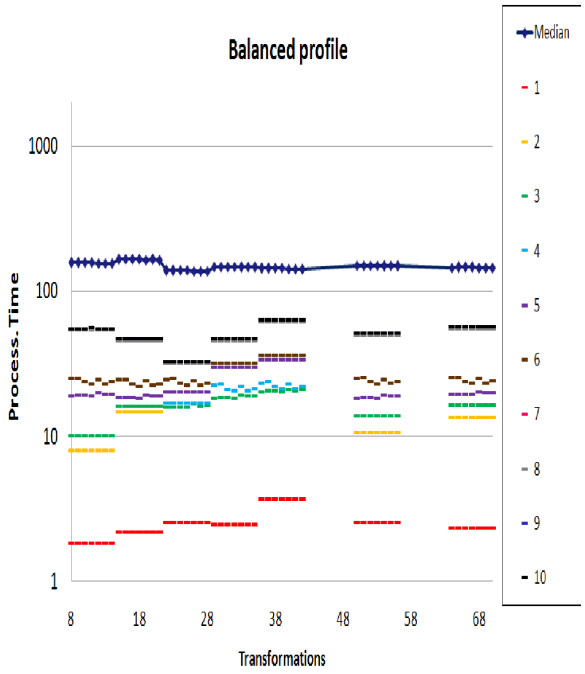


Figure 44: video-only top 10 detection performance based on Nofa profile

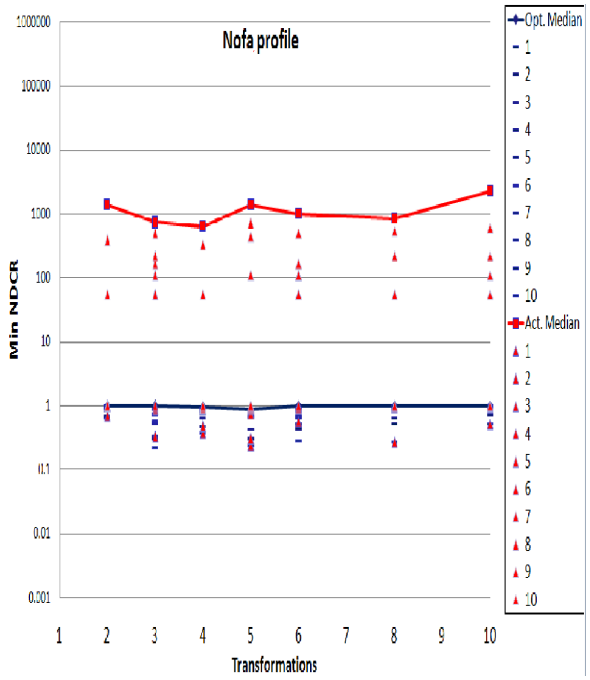


Figure 45: audio-only top 10 detection performance based on Nofa profile

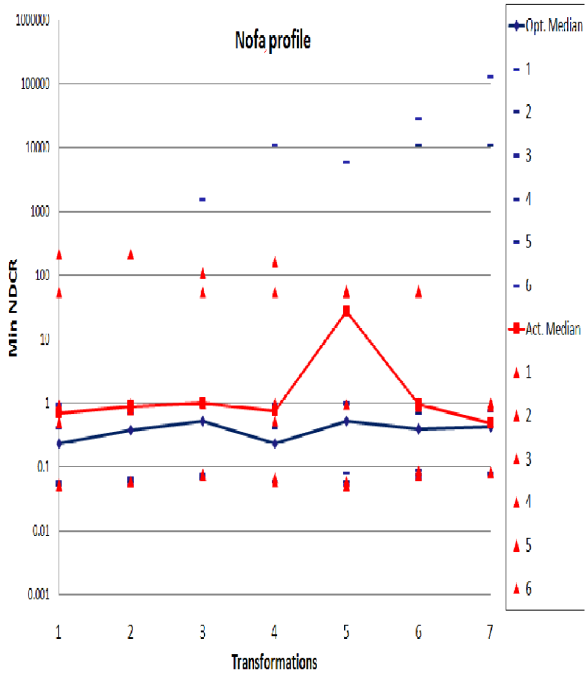


Figure 46: video+audio top 10 detection performance based on Nofa profile

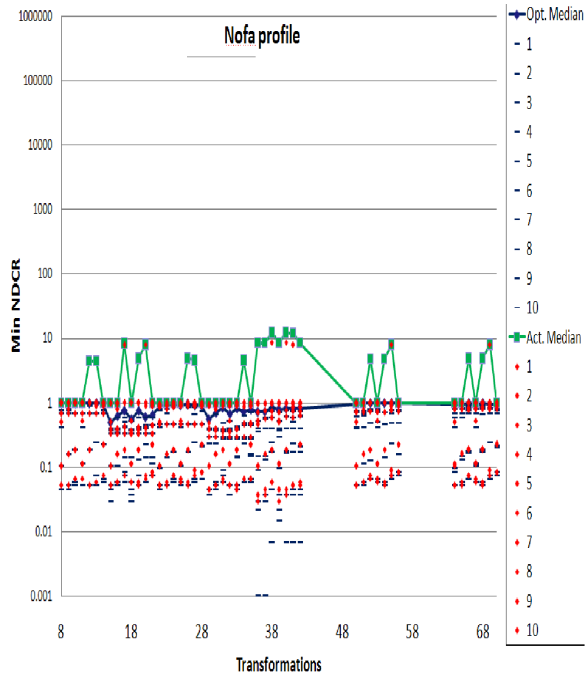


Figure 47: video-only top 10 localization performance based on Nofa profile

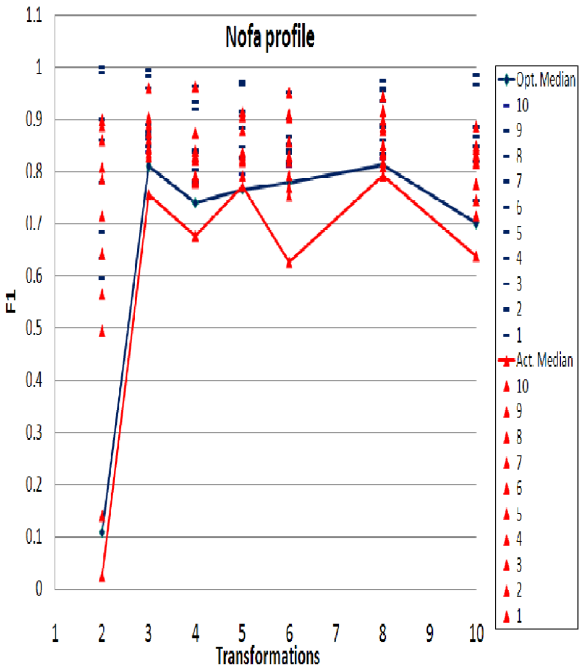


Figure 48: audio-only top 10 localization performance based on Nofa profile

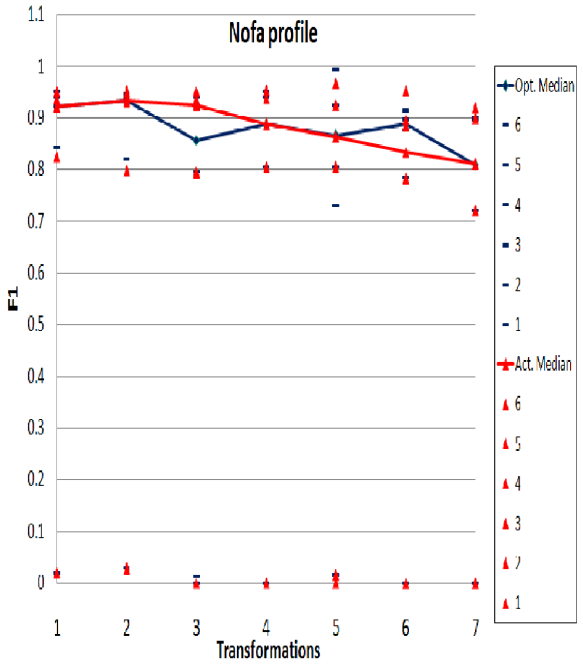


Figure 49: "video+audio" top 10 localization performance based on Nofa profile

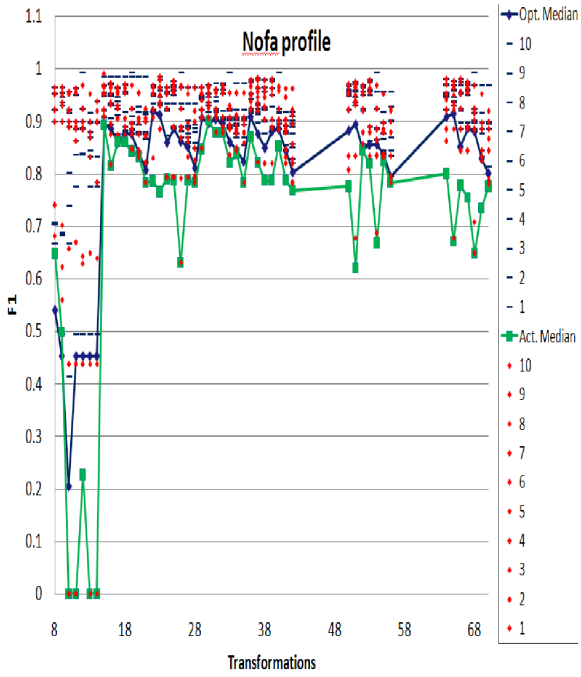


Figure 50: video-only top 10 efficiency performance based on Nofa profile

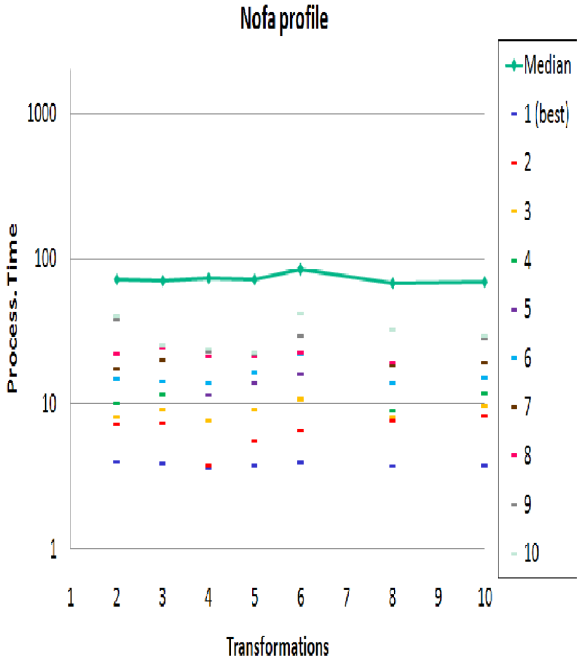


Figure 51: audio-only top 10 efficiency performance based on Nofa profile

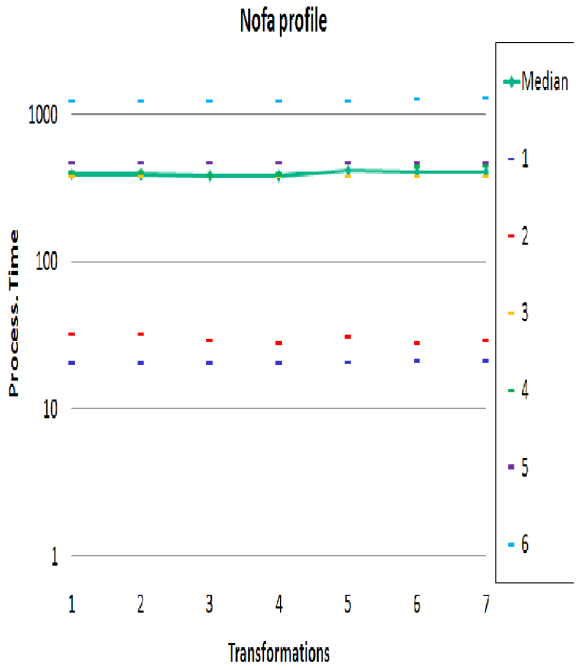
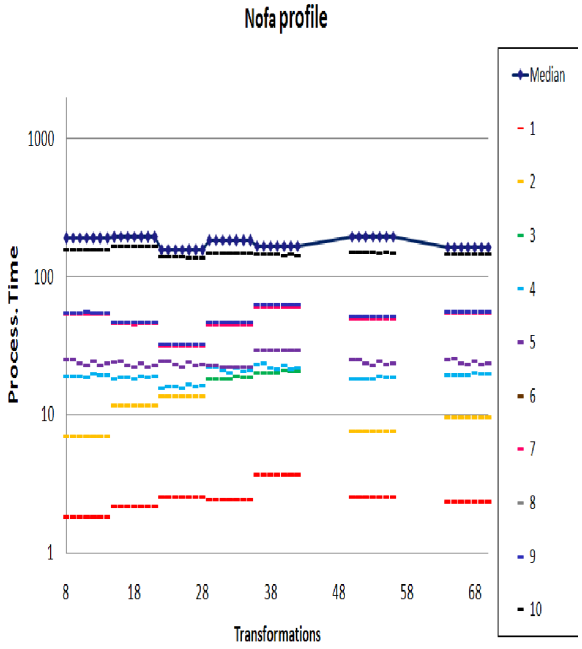


Figure 52: video+audio top 10 efficiency performance based on Nofa profile



curve across the transformations for the video-only and video+audio queries almost matches the DET score, 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 efficiency of the three query types for Nofa were comparable to the balanced profile.

The relationship between the three main measures for both profiles across all transformations for video-only runs is illustrated by Figures 53 through 58. Basically, increasing processing time didn't enhance the localization or detection, while few systems achieved high localization and detection in small time. Also, generally systems that are good in detection (low NDCR) are also good in localization. Those observations are alike in both profiles.

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 59 shows for each video transformation the best performance (in red) and the best performance of the 7 audio transformations when applied on those video transformation (in purple). It is clear that using audio has decreased the detection cost across all transformations. The same experiment was done in Figure 60 based on localization performance. Although the same strong effect on detection cost is not seen, using audio has increased localization performance across the majority of the transformations. To summarize our observations for TV2009 for this task we can conclude that determining the optimal operating point (threshold) is critical and requires score normalization across queries. It has a huge impact on NDCR scores (especially for video-only runs) and is 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 improves upon using audio-only and video-only. Video-only systems in general are slightly faster than others and yield the best localization results (although audio-only have a higher median). Few systems performed well in all three measures, thus there is still a room for improvement for systems to improve their accuracy, speed and performance. In general, there is a limited attraction for audio-only queries (only 6

Figure 53: video-only localization vs Process.time based on balanced profile

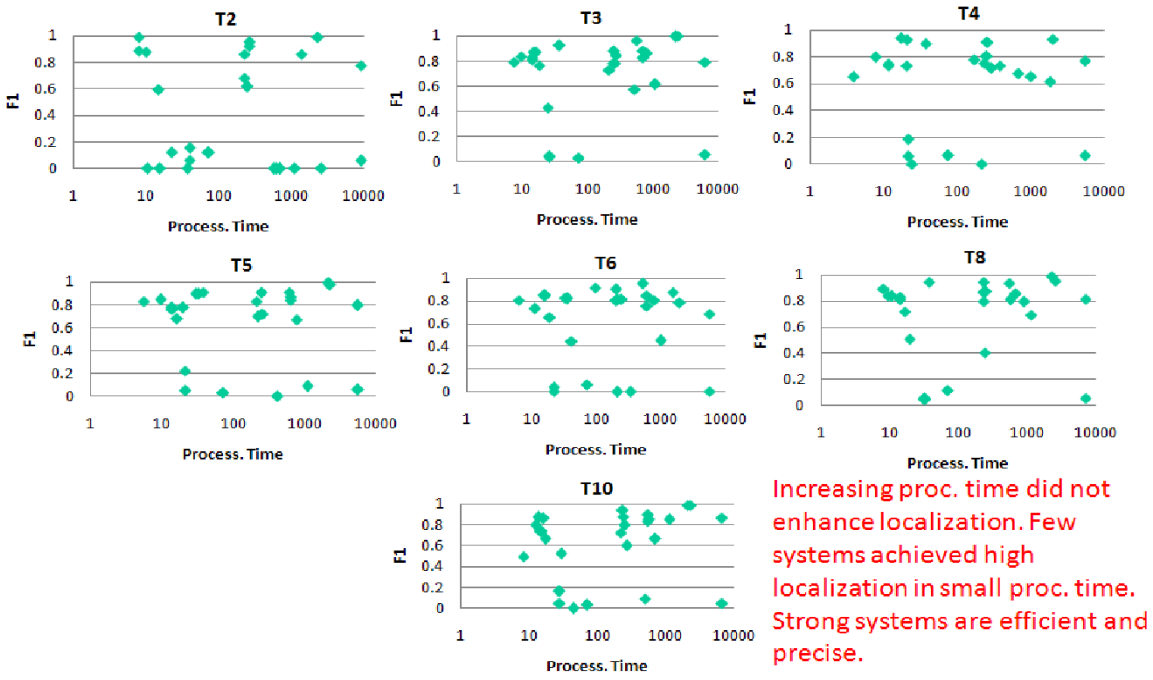


Figure 54: video-only localization vs detection based on balanced profile

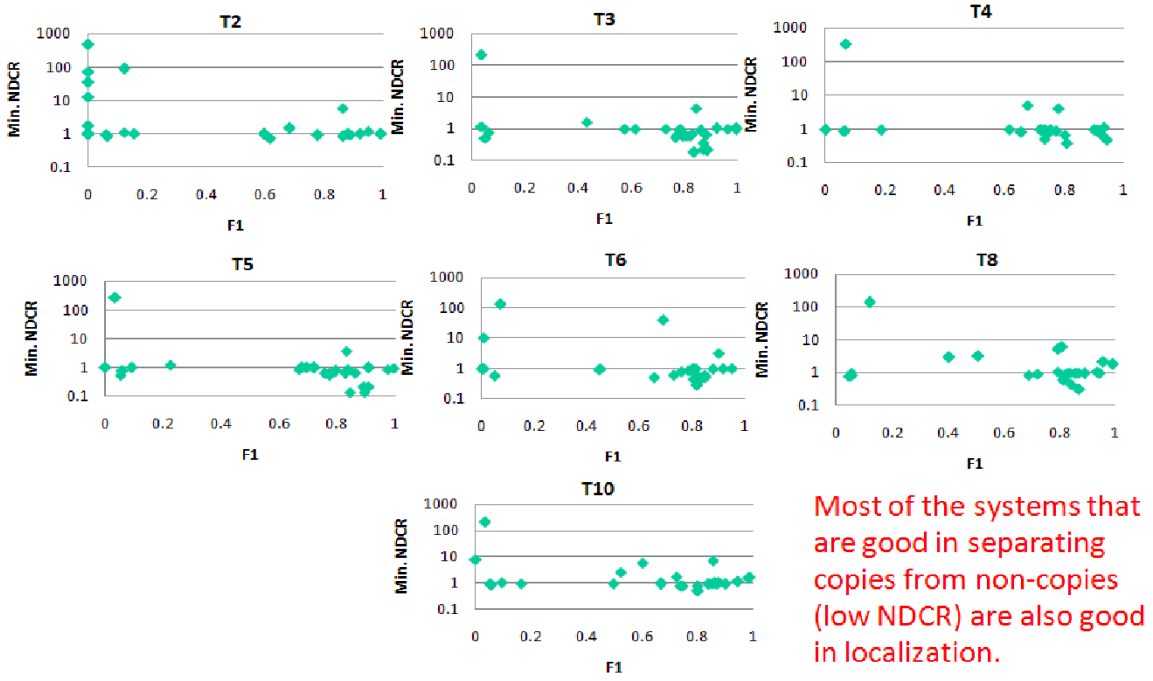


Figure 55: video-only detection vs Process.time based on balanced profile

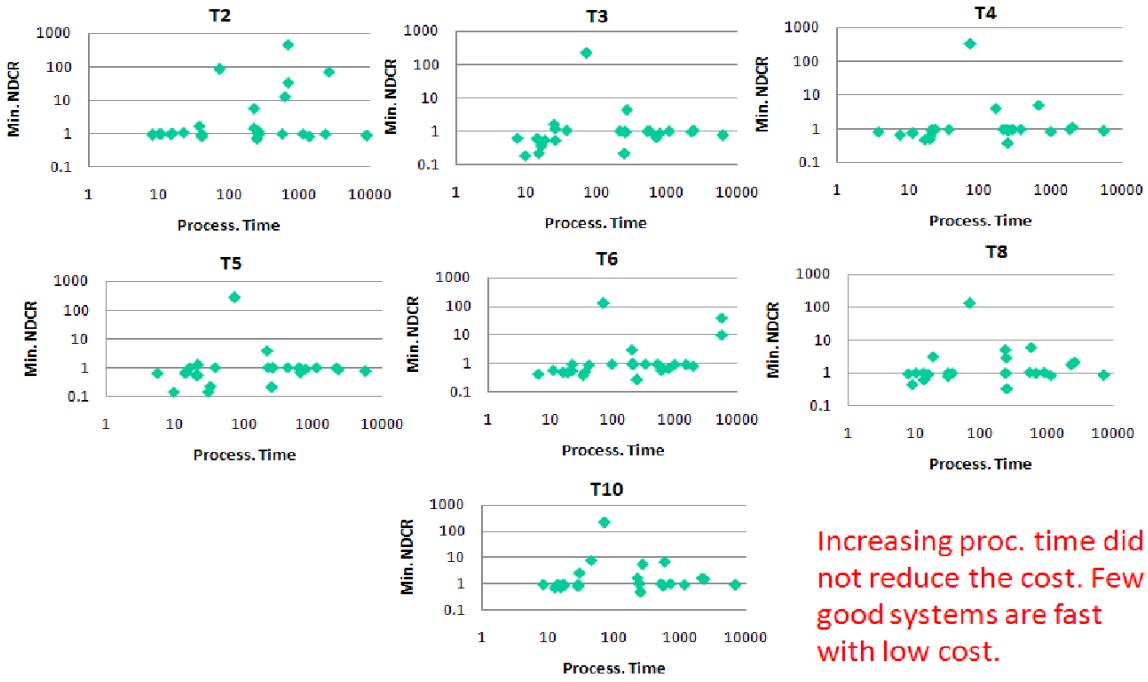


Figure 56: video-only localization vs Process.time based on Nofa profile

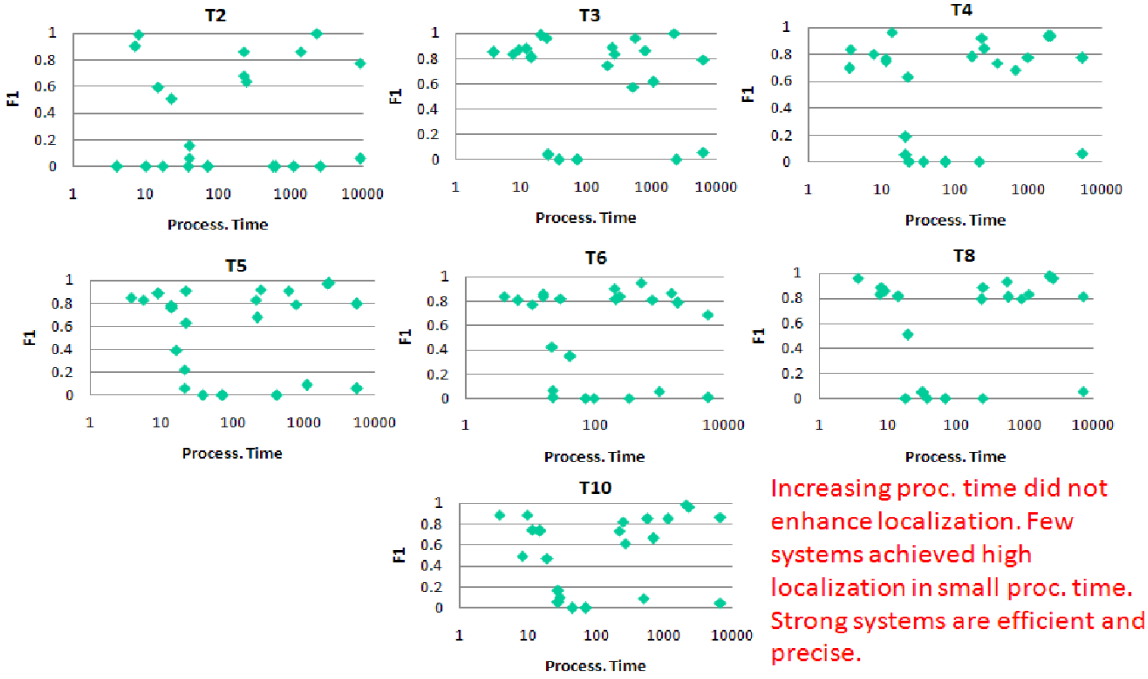


Figure 57: video-only localization vs detection based on Nofa profile

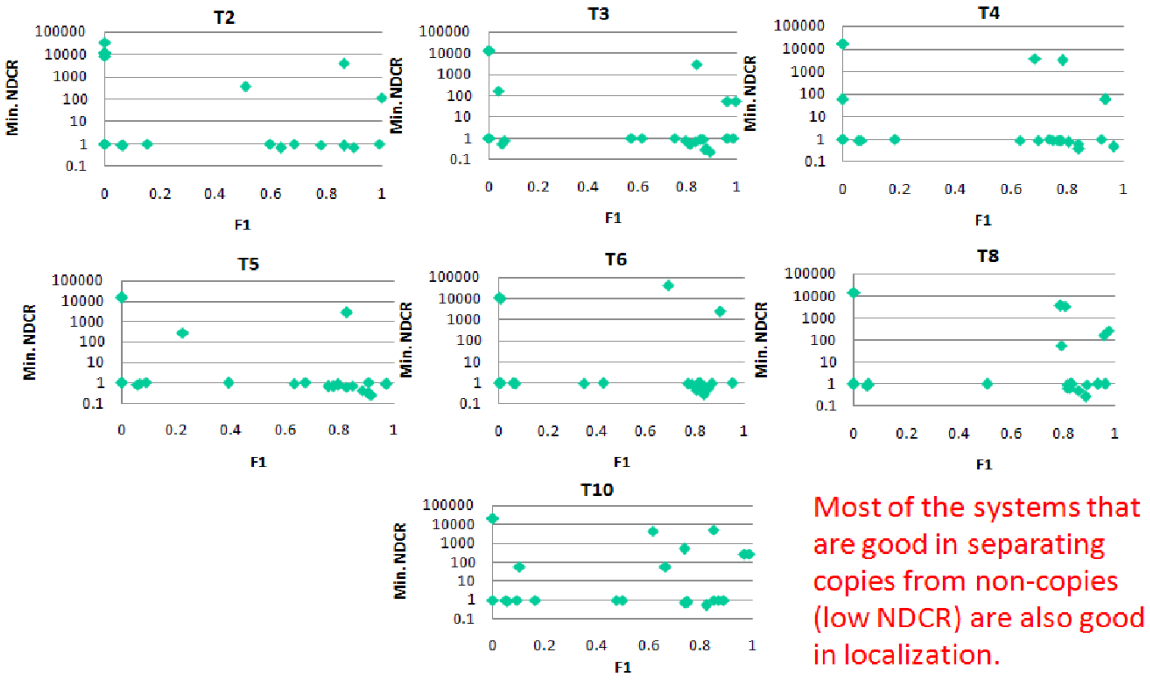


Figure 58: video-only detection vs Process.time based on Nofa profile

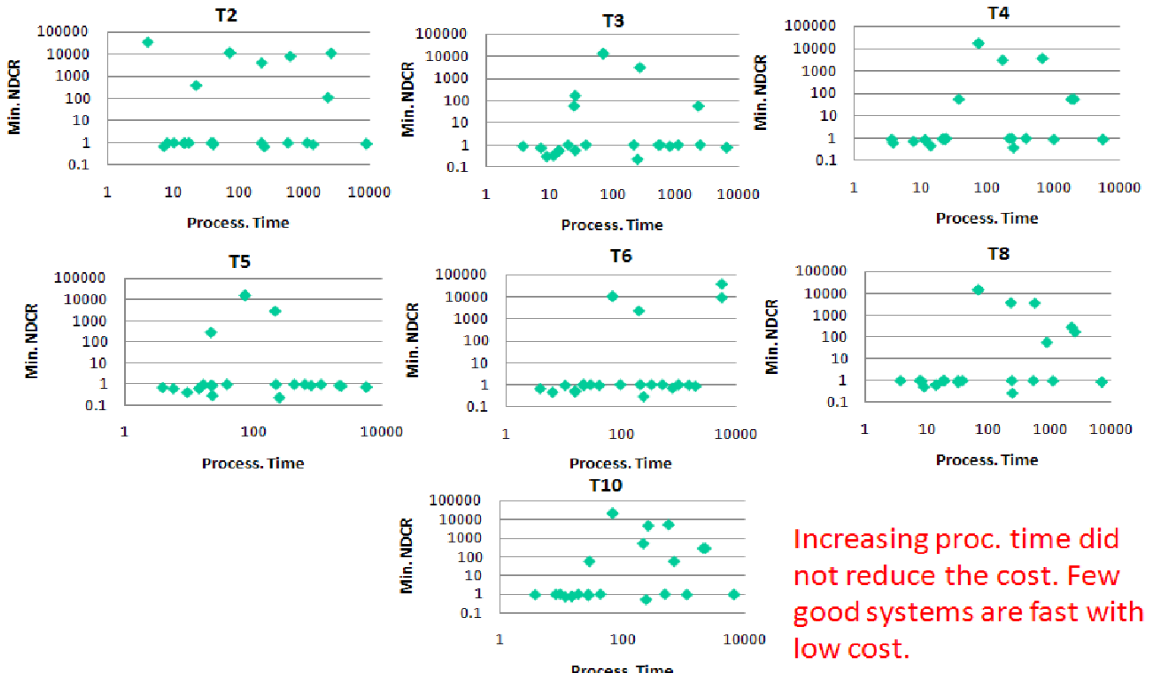
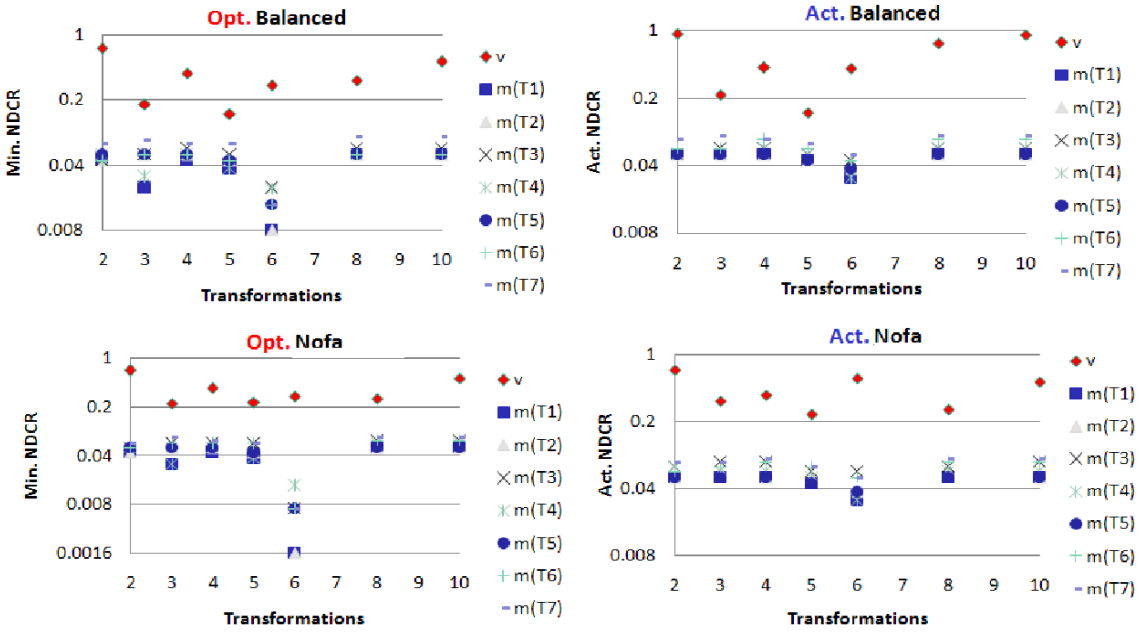
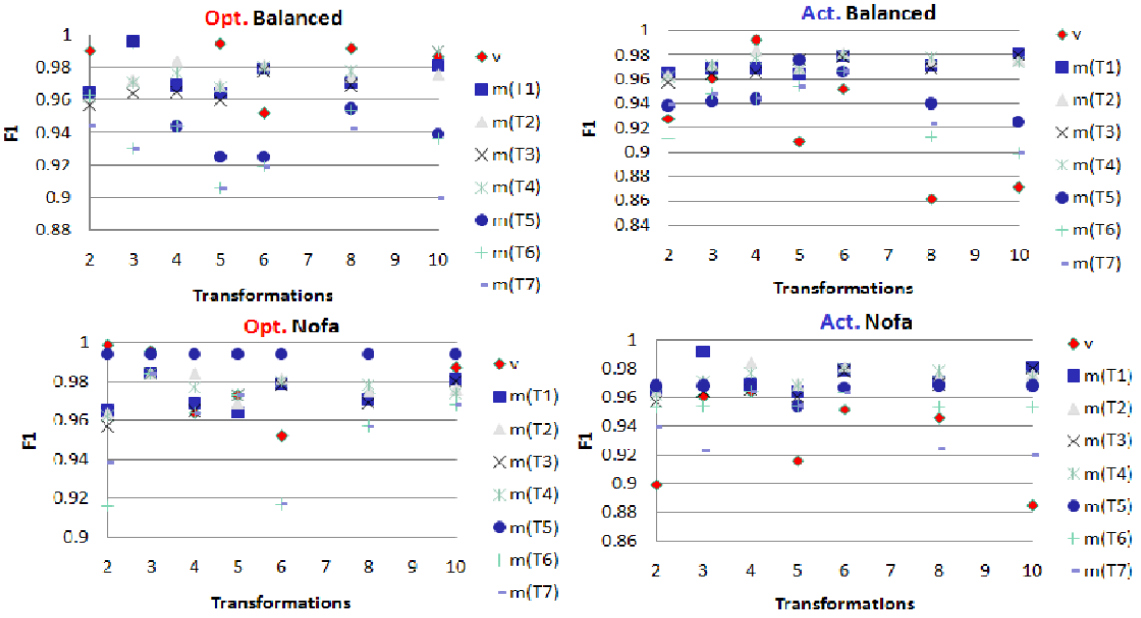


Figure 59: video-only vs video+audio detection performance



The m runs highly enhanced the detection accuracy across all transformations

Figure 60: video-only vs video+audio localization performance



The m runs helped in the majority of transformations to enhance localization

runs are submitted).

Based on site reports of their experiments, different approaches included speed optimization using GPU-based local feature extraction, fusion of frame fingerprints such as SIFT descriptors, black-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. Readers should see the notebook papers posted on the TRECVID website (trecvid.nist.gov) for details about each participant's experiments and results.

6 Surveillance event detection

The 2009 Surveillance Event Detection evaluation was the second evaluation focused on event detection in the surveillance video domain. The first such evaluation was conducted as part of the 2008 TRECVID conference series (Rose et al., 2009).

The goal of the evaluation track is to support the development of technologies to detect visual events (people engaged in particular activities) in a large collection of video data. It was designed to move computer vision technology towards robustness and scalability while increasing core competency in detecting human activities within video. The approach used was to employ real surveillance data, orders of magnitude larger than previous computer vision tests, and consisting of multiple, synchronized camera views.

The 2009 evaluation supported the same two evaluation tasks as the 2008 evaluation, retrospective event detection and freestyle analysis, and the same set of 10 events, found in Appendix C were used.

Retrospective event detection is defined as follows: given a set of video sequences, detect as many event observations as possible in each sequence. For this evaluation, a single-camera condition was used as the required condition (multiple-camera input was allowed as a contrastive condition). Furthermore, systems could perform multiple passes over the video prior to outputting a list of putative events observations (i.e., the task was retrospective). Eleven teams participated in the retrospective task. 75 event runs where submitted for evaluation.

For freestyle analysis, participants were asked to define tasks pertinent to the airport video surveillance domain that could be implemented on the data set. Freestyle submissions were to include rationale, clear definitions of the task, performance measures, reference annotations, and a baseline system implementation. No sites participated in the freestyle task.

While the evaluation tasks did not change, the data used for the evaluation did change in the following ways:

1. The 2008 Event Detection development and evaluation data sets were both designated as 2009 development resources thus expanded the development material to 100 camera-hours.

2. A new evaluation test set was prepared for 2009.

3. The event annotation procedure was changed for 2009.

6.1 Event Annotation

For this evaluation, we define an event to be an observable state change, either in the movement or interaction of people with other people or objects. As such, the evidence for an event depends directly on what can be seen in the video and does not require higher level inference. The annotation guidelines were developed to express the requirements for each event.

To determine if the observed action is a taggable event, a *reasonable interpretation rule* was used. The rule was, "if according to a reasonable interpretation of the video the event must have occurred, then it is a taggable event". Importantly, the annotation guidelines were designed to capture events that can be detected by human observers, such that the ground truth would contain observations that would be relevant to an operator/analyst. In what follows we distinguish between event types (e.g., parcel passed from one person to another), event instance (an example of an event type that takes place at a specific time and place), and an event observation (event instance captured by a specific camera).

Experiments conducted during the 2008 evaluation showed that humans annotators missed a large number of observations when they looked for five events simultaneously (Rose et al., 2009). Experiments also showed that multiple annotation passes, followed by a senior annotator review could reduce the number of missed events and reduce false alarms. The LDC was able to perform 3 independent annotation passes looking for 3 events simultaneously and then resolved differences (via an adjudication process similar to last year's process) within the same cost/time constraints as annotating a single pass over 50 hours of data.

The ElevatorNoEntry event was annotated as a separate, single pass because potential instances of people interacting with the elevator are easy to spot.

The videos were annotated using the Video Performance Evaluation Resource (ViPER) tool. Events were represented in ViPER format using an annotation schema that specified each event observation's time interval.

6.2 Data

The development data consisted of the full 100 hours data set used for the 2008 Event Detection (Rose et al., 2009) evaluation.

The video for the evaluation corpus came from the 45 hour Home Office Scientific Development Branch's

(HOSDB) Image Library for Intelligent Detection Systems (iLIDS) Multi Camera Tracking Training (MCTTR) data set. The evaluation systems processed the full data set however systems were scored on a 4-day subset of recordings.

Both data sets were collected in the same busy airport environment with the same video cameras. The entire video corpus was distributed as MPEG-2 in de-interlaced, Phase Alternating Line (PAL) format (resolution 720 x 576), 25 frames/sec, either via hard drive or internet download.

6.3 Evaluation

Sites submitted system outputs for the detection of any 3 of 10 possible events (Appendix C). Outputs included the temporal extent as well as a decision score (indicating the strength of evidence supporting the observation's existence) and detection decision (yes/no) for each event observation. Developers were advised to target a low miss, high false alarm scenario via the scoring metrics in order to maximize the number of event observations.

A dry run was carried out for one day of collection from the development data in order to test system's ability to generate compliant system outputs capable of being scored by the evaluation infrastructure. A formal evaluation was carried out for four of the twelve days of collection (approx. 15 camera hours). Groups were allowed to submit multiple runs with contrastive conditions.

6.4 Measures of Performance

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, this task used the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance as described in the evaluation plan (Fiscus, Rose, & Michel, 2009). NDCR is a weighted linear combination of the system's Missed Detection Probability and False Alarm Rate (measured per unit time).

$NDCR = P_{miss} + \beta \times R_{FA}$,
where
$P_{Miss} = N_{misses}/N_{Ref}$
$R_{FA} = N_{spurious}/N_{CamHrs}$
$\beta = \frac{Cost_{FA}}{Cost_{Miss} \times R_{Target}}$
$C_{Miss} = 10; C_{FA} = 1; R_{Target} = 20/hour$

NDCR is normalized to have the range of [0,∞] where 0 would be for perfect performance, 1 would be the cost of a system that provides no output, and ∞ is possible because false alarms are included in the measure.

The inclusion of decision scores in the system output permits the computation of Decision Error Tradeoff (DET) curves. DET curves plot P_{miss} vs. R_{FA} for all thresholds applied to the system's decision scores. These plots graphically show the tradeoff between the two error types for the system.

6.5 Results

The NDCRs for the submitted event runs can be found in Figure 61. The figure contains two NDCR values for each submission: the Actual NDCR which is the NDCR based on the binary decisions produced by the system and the Minimum NDCR which is the lowest NDCR possible based on the decision scores produced by the system. The difference between the actual and minimum NDCRs indicates how well the system-identified decision score threshold (via the binary decisions) was tuned to the NDCR function.

The lowest NDCRs were achieved for the ElevatorNoEntry and the OpposingFlow events. Both these two events involve a single person exhibiting a behavior at a specific location. The Actual NDCRs for Toshiba, Carnegie Mellon University, and Peking University's ElevatorNoEntry runs where 0.333, 0.340, and 0.342 respectively. For the OpposingFlow event, the lowest Actual NDCRs where 0.002 and 0.037 for Shanghai Jiao Tong University and Toshiba respectively.

The Actual NDCRs for the rest of the events, which can be ocurred at any location in any camera view, where much higher ranging from 0.971 for the National Hong Kong Science and Technology Research Lab's PersonRuns run to as high as 8.87.

Figure 62 contains a single DET curve for each event. The curve selected for each event was the run with the lowest Minimum NDCR. The DET curves tell the same story as the Actual NDCRs: there's a clear difference in performance for the ElevatorNoEntry and OpposingFlow versus the rest of the events.

7 Summing up and moving on

This overview of TRECVID 2009 has provided basic information on the goals, data, evaluation mechanisms and metrics used as well summary information about technical approaches and results. Further details about each particular group's approach and performance for each task

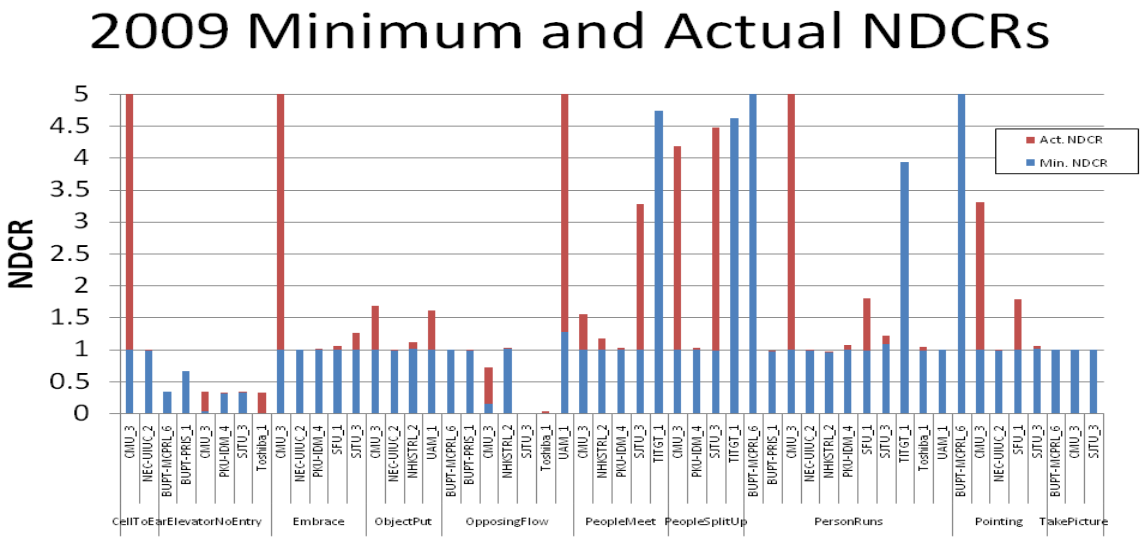


Figure 61: Minimum and Actual NDCRs for submitted runs.

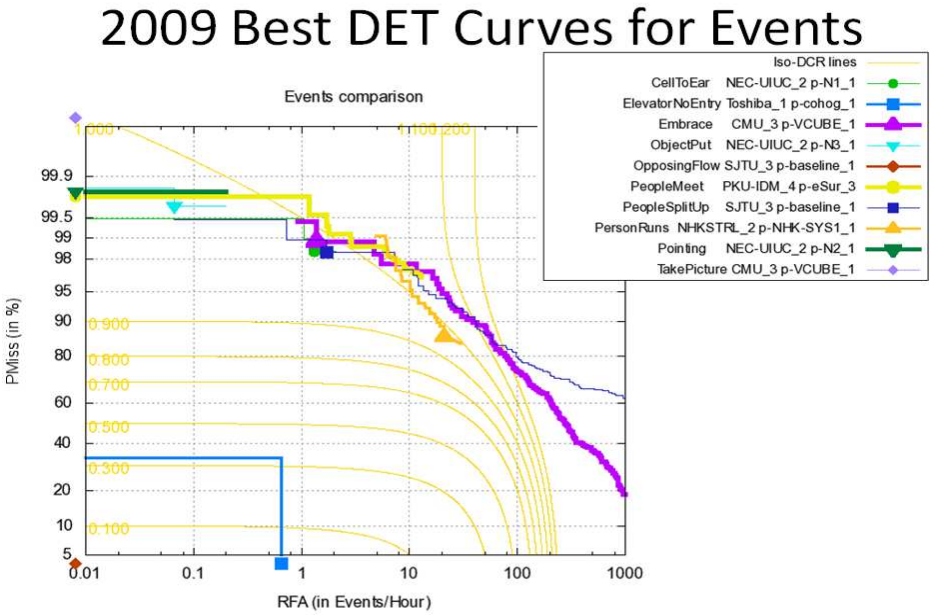


Figure 62: DET Curves for Event Runs with the lowest Minimum NDCR.

can be found in that group's notebook paper in the TRECVID publications webpage: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

8 Authors' note

TRECVID would not happen without support from IARPA and NIST and the research community is very grateful for this. Alan Smeaton is funded by Science Foundation Ireland under grant 07/CA/I1147. The surveillance event detection evaluation was funded in part by the Department of Homeland Security Science and Technology Directorate Predictive Screening Project. Beyond that, various individuals and groups deserve special thanks.

Georges Quét and Stéphane Ayache again organized a collaborative annotation; participants annotated 10 new features NII and the Laboratoire d'Informatique de Grenoble mirrored the video data. Christian Petersohn at the Fraunhofer (Heinrich Hertz) Institute in Berlin provided the master shot reference.

Roeland Ordelman and Marijn Huijbregts at the University of Twente donated the output of their ASR system run on the Sound and Vision data. The LIMSI Spoken Language Processing Group contributed a speech transcription of the 2007-2009 Sound and Vision data using their Dutch recognizer. Christof Monz of Queen Mary, University London contributed MT (Dutch to English) for the Sound and Vision video.

For the copy detection task we reused the video transformation tools created by Laurent Joyeux for INRIA in 2008. Dan Ellis at Columbia University devised and applied the audio transformations to produce the audio-only queries for copy detection.

The MediaMill team at the University of Amsterdam donated detector scores for 64 concepts trained over the last 3 years. The Columbia University and City University Hong Kong team donated detector scores for the TRECVID 2009 test data using the CU-VIREO374 models.

Finally, we want to thank all the participants and other contributors on the mailing list for their enthusiasm and diligence.

9 Appendix A: Topics

The text descriptions of the topics are listed below followed in brackets by the associated number of image examples (I), video examples (V), and relevant shots (R) found during manual assessment of the pooled runs.

Table 6: 2009 Topic types

	Named			Generic		
Topic	Person, thing	Event	Place	Person, thing	Event	Place
269				X	X	X
270				X		X
271				X		
272				X	X	
273				X	X	
274				X		
275				X	X	
276				X	X	
277				X	X	
278				X		X
279				X	X	
280				X		
281				X	X	
282				X		
283				X	X	
284				X		X
285				X		
286				X	X	
287				X		X
288				X		
289				X	X	
290				X		
291				X	X	
292				X	X	

269 Find shots of a road taken from a moving vehicle through the front window.(I/0, V/8, R/266).

270 Find shots of a crowd of people, outdoors, filling more than half of the frame area.(I/1, V/6, R/588).

271 Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible (I/3, V/6, R/484).

272 Find shots of a person talking on a telephone.(I/2, V/4, R/287).

273 Find shots of a closeup of a hand, writing, drawing, coloring, or painting(I/2, V/6, R/285).

274 Find shots of exactly two people sitting at a table.(I/3, V/5, R/458).

275 Find shots of one or more people, each walking up one or more steps.(I/2, V/5, R/136).

276 Find shots of one or more dogs, walking, running, or jumping.(I/3, V/6, R/233).

277 Find shots of a person talking behind a microphone.(I/2, V/7, R/910).

278 Find shots of a building entrance.(I/3, V/7, R/1039).

279 Find shots of people shaking hands.(I/4, V/6, R/65).

280 Find shots of a microscope.(I/4, V/5, R/117).

281 Find shots of two more people, each singing and/or playing a musical instrument.(I/4, V/6, R/478).

282 Find shots of a person pointing.(I/5, V/6, R/322).

283 Find shots of a person playing a piano.(I/4, V/4, R/86).

284 Find shots of a street scene at night.(I/4, V/6, R/372).

285 Find shots of printed, typed, or handwritten text, filling more than half of the frame area.(I/4, V/8, R/1100).

286 Find shots of something burning with flames visible.(I/4, V/6, R/488).

287 Find shots of one or more people, each at a table or desk with a computer visible.(I/1, V/6, R/629).

288 Find shots of an airplane or helicopter on the ground, seen from outside.(I/5, V/7, R/282).

289 Find shots of one or more people, each sitting in a chair, talking.(I/0, V/6, R/1153).

290 Find shots of one or more ships or boats, in the water.(I/4, V/6, R/590).

291 Find shots of a train in motion, seen from outside.(I/2, V/7, R/99).

292 Find shots with the camera zooming in on a person's face.(I/0, V/6, R/155).

10 Appendix B: Features

The features labeled with an asterisk were also among the twenty tested in 2008.

1* Classroom: a school- or university-style classroom scene. One or more students must be visible. A teacher and teaching aids (e.g. blackboard) may or may not be visible

2 Chair: a seat with four legs and a back for one person

3 Infant: a very small child, crawling, lying down, or being held, with no evidence it can walk

4 Traffic intersection: crossing of two roads or paths with some human and/or vehicular traffic visible

5 Doorway: an opening you can walk through into a room or building

6* Airplane-flying: external view of a heavier than air, fixed-wing aircraft in flight - gliders included. NOT balloons, helicopters, missiles, and rockets

7 Person-playing-a-musical-instrument: both player and instrument visible

8* Bus: external view of a large motor vehicle on tires used to carry many passengers on streets, usually along a fixed route. NOT vans and SUVs

9 Person-playing-soccer: need not be teams or on a dedicated soccer field

10* Cityscape: a view of a large urban setting, showing skylines and building tops. NOT just street-level views of urban life

11 Person-riding-a-bicycle: a bicycle has two wheels; while riding, both feet are off the ground and the bicycle wheels are in motion

12* Telephone: any kinds of telephone, but more than just a headset must be visible.

13 Person-eating: putting food or drink in his/her mouth

14* Demonstration-Or-Protest: an outdoor, public exhibition of disapproval carried out by multiple people, who may or may not be walking, holding banners or signs

15* Hand: a close-up view of one or more human hands, where the hand is the primary focus of the shot.

16 People-dancing: one or more, not necessarily with each other

17* Nighttime: a shot that takes place outdoors at night. NOT sporting events under lights

18* Boat-Ship: exterior view of a boat or ship in the water, e.g. canoe, rowboat, kayak, hydrofoil, hovercraft, aircraft carrier, submarine, etc.

19 Female-human-face-closeup: closeup of a female human's face (face must clearly fill more than 1/2 of height or width of a frame but can be from any angle and need not be completely visible)

20* Singing: one or more people singing - singer(s) visible and audible, solo or accompanied, amateur or professional

11 Appendix C: SED Events

1. CellToEar: someone puts a cell phone to his/her ear.
2. ElevatorNoEntry: elevator doors open with a person waiting in front of them, but the person does not get in before the doors close.
3. Embrace: someone puts one or both arms at least part way around another person.
4. ObjectPut: someone drops or puts down an object.
5. OpposingFlow: someone moves through a controlled access door opposite to the normal flow of traffic.
6. PeopleMeet: one or more people walk up to one or more other people, stop, and some communication occurs.
7. PeopleSplitUp: for two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame.
8. PersonRuns: someone runs.
9. Pointing: someone points.
10. TakePicture: someone takes a picture.

References

Armitage, L. H., & Enser, P. G. B. (1996). *Information Need in the Visual Document Domain: Report on Project RDD/G/235 to the British Library Research and Innovation Centre.* School of Information Management, University of Brighton.

Ayache, S., & Quénot, G. (2008). Video Corpus Annotation Using Active Learning,. In *Proceedings of the 30th european conference on information retrieval (ecir'08)* (pp. 187–198). Glasgow, UK.

Enser, P. G. B., & Sandom, C. J. (2002). Retrieval of Archival Moving Imagery — CBIR Outside the Frame. In M. S. Lew, N. Sebe, & J. P. Eakins (Eds.), *Image and Video Retrieval, International Conference, CIVR 2002, London, UK, July 18-19, 2002, Proceedings* (Vol. 2383). Springer.

Fiscus, J., Rose, R. T., & Michel, M. (2009). *2009 TRECVid Event Detection evaluation plan.* `http://www.itl.nist.gov/iad/mig//tests/trecvid/2009/doc/EventDet09-EvalPlan-v03.htm`.

Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology* (2nd ed.). London, UK: Chapman & Hall.

Over, P., Ianeva, T., Kraaij, W., & Smeaton, A. F. (2006). *TRECVID 2006 Overview.* `www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6overview.pdf`.

Rose, T., Fiscus, J., Over, P., Garofolo, J., & Michel, M. (2009). The TRECVid 2008 Event Detection Evaluation. In *IEEE Workshop on Applications of Computer Vision (WACV).* IEEE.

Shatford, S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging and Classification Quarterly, 6*(3), 39—61.

Webber, W., Moffat, A., Zobel, J., & Sakai, T. (2008). Precision-at-ten considered redundant. In *Sigir '08: Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval* (pp. 695–696). New York, NY, USA: ACM.

Yilmaz, E., & Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the Fifteenth*

ACM International Conference on Information and Knowledge Management (CIKM). Arlington, VA, USA.

Table 7: 2009 Participants not submitting runs

ED	FE	SE	CD	Location	Participants
**	−−	−−	**	Europe	Chemnitz University of Technology
**	**	**	**	N.Amer.	CompuSensor Technology Corporation
**	**	**	**	N.Amer.	Computational Analysis and Network Enterprise Solutions
**	−−	**	**	S.Amer.	Digital Image Processing Laboratory
**	−−	**	−−	Europe	Dublin City University
**	**	**	**	Europe	ETIS Laboratory
**	**	**	**	N.Amer.	Florida Atlantic Unviersity
−−	−−	−−	**	Europe	Fraunhofer Institute for Telecommunications HHI
−−	−−	**	−−	N.Amer.	FX Palo Alto Laboratory
−−	−−	−−	**	Europe	Hellenic Open University
−−	**	−−	−−	Asia	Information and Communications University
**	**	−−	−−	Europe	IRISA/INRIA Rennes
**	−−	−−	−−	Europe	JOANNEUM RESEARCH FmbH-SCOVIS
**	**	−−	**	Asia	KDDI R&D Laboratories, Inc.
−−	**	**	**	Europe	Laboratoire d'Intégration des Systèmes et des Technologies
**	**	−−	−−	Europe	Laboratrio de Visã Computacional da UFCG
**	**	−−	**	Europe	LIP6 - Lab. d'Informatique de Paris Uni. P. & M. Curie
−−	−−	−−	**	Asia	National Chung Cheng University
**	**	**	−−	Asia	National Taiwan University
−−	**	−−	−−	Asia	Osaka City University
**	**	**	−−	Austral.	RMIT
−−	**	**	**	Asia	Shandong University
−−	−−	−−	**	Europe	Tampere University of Technology
**	−−	−−	−−	Asia	Tianjin University
**	−−	−−	−−	Europe	Trackers by Federal University of Paraná
−−	**	−−	−−	Asia	Tsinghua University-THEEIE
−−	−−	**	−−	N.Amer.	University of Alabama
−−	−−	**	−−	Europe	University of Alicante
−−	**	**	−−	Asia	University of Malaya
**	**	−−	−−	N.Amer.	University of Memphis
−−	**	−−	**	Europe	University of Sheffield
**	**	**	**	Asia	Wuhan University, China

Task legend. CD: Copy detection; ED: event detection; FE: Feature extraction; SE: Search; ∗∗: Group applied but didn't submit any runs