Evaluation Campaigns and TRECVid

Alan F. Smeaton
Centre for Digital Video Proc.
& Adaptive Information Cluster
Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

Paul Over
Information Access Division
Information Technology Lab.
National Institute of Standards
and Technology
Gaithersburg,
MD. 20899, USA
over@nist.gov

Wessel Kraaij
TNO Information and
Communication Technology,
PO BOX 5050 2600 GB Delft
The Netherlands
wessel.kraaij@tno.nl

ABSTRACT

The TREC Video Retrieval Evaluation (TRECVid) is an international benchmarking activity to encourage research in video information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations [1] interested in comparing their results. TRECVid completed its fifth annual cycle at the end of 2005 and in 2006 TRECVid will involve almost 70 research organizations, universities and other consortia. Throughout its existence, TRECVid has benchmarked both interactive and automatic/manual searching for shots from within a video corpus, automatic detection of a variety of semantic and low-level video features, shot boundary detection and the detection of story boundaries in broadcast TV news. This paper will give an introduction to information retrieval (IR) evaluation from both a user and a system perspective, highlighting that system evaluation is by far the most prevalent type of evaluation carried out. We also include a summary of TRECVid as an example of a system evaluation benchmarking campaign and this allows us to discuss whether such campaigns are a good thing or a bad thing. There are arguments for and against these campaigns and we present some of them in the paper concluding that on balance they have had a very positive impact on research progress.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: [Evaluation / methodology]

[1]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.

MIR'06, October 26–27, 2006, Santa Barbara, California, USA.

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Evaluation, Benchmarking, Video Retrieval

1. INTRODUCTION

Evaluation campaigns which benchmark IR tasks have become very popular in recent years for a variety of reasons. They are attractive to researchers because they allow comparison of their work with others in an open, metrics-based environment. They provide shared data, common evaluation metrics and often also offer collaboration and sharing of resources. They are also attractive to funding agencies and outsiders because they can act as a showcase for research results.

Analysis, indexing and retrieval of video shots takes place each year within the TRECVid evaluation campaign and this paper presents an overview of TRECVid and its activities. We begin, in section 2, with an introduction to evaluation in IR, covering both user evaluation and system evaluation. In section 3 we present a catalog of evaluation campaigns in the general area of IR and video analysis. Sections 4 and 5 give a retrospective overview of the TRECVid campaign with attention to the evolution of the evaluation and participating systems, open issues, etc. In section 6 we discuss whether evaluation benchmarking campaigns like TRECVid, Text Retrieval Conferences (TREC) and others are good or bad. We present a series of arguments for each case and leave the reader to conclude that on balance they have had a positive impact on research progress.

2. USER EVALUATION AND SYSTEM EVALUATION OF IR

In the early 1960s, the Cranfield College of Aeronautics wanted to test indexing techniques for text abstracts. They created test queries on a static document collection of some hundreds of documents and each document was judged as either relevant or not relevant to each of a set of user queries. Based on the combination of documents, user queries and relevance judgments, the researchers were able to evaluate different indexing and retrieval strategies using measures such as precision and recall, which are well-known and still used now. That experiment was the first experimental IR

evaluation, and the empirical approach to evaluating IR tasks continues today.

When we build an IR system we build it to serve one part or function in an overall information seeking task. We use a search tool, which is what an IR system is, to retrieve documents or images or video clips in response to a specific, formulated search request, but that search request is just one stage of our overall information need. When we use an IR system, we are engaging in information seeking. It follows that what we should evaluate are things like user satisfaction and the goodness of fit of the system we are using for task completion. But we can't do this because it would involve testing with a significant number of real users every time we want to do such an evaluation. That is prohibitively expensive to do every time we think we've discovered a new indexing or retrieval algorithm or we want to modify and evaluate an existing one. Such evaluations are termed user evaluations, performed from an information science viewpoint and are not common. Instead what we do is system evaluation, which is evaluation more from a computer science viewpoint and that is what is prevalent in IR research [17]. This is summarized below

System evaluation tests the quality of an IR system; processes a high volume of queries; has no user involvement and simulates an end-user; is cheap and very popular and is a highly controlled environment. *User evaluation* tests the quality of IR system and its interface; (usually) processes a low volume of queries; has direct user involvement in the evaluation and is an artificial test;

The development of empirical IR research continues to use test collections of documents, queries and relevance assessments and has been based on system rather than user evaluation, though a small amount of the latter is carried out. As digital document collections (including texts, web pages, images, videos, music, and others) for personal and for work-related use have exploded in size, IR research came under increasing pressure to make IR evaluations realistic. The approach of manual judgments of relevance carried out in individual laboratories or by individual researchers meant that evaluations on collections of the order of thousands of documents was simply not credible as people started to use collections of millions and then billions of documents. The sheer effort, and cost, of creating a dataset which could be used for evaluation and which was credible remains beyond the resources of almost all research groups and so over the last several years we have seen the emergence of benchmarking evaluation campaigns which we discuss in the next section.

3. BENCHMARKING EVALUATION CAMPAIGNS

Following the realization that benchmarking IR tasks needed to scale up in size in order to be realistic, the Text Retrieval Conference (TREC) initiative began in 1991 as a reaction to small collection sizes and the need for a more coordinated evaluation among researchers. This was run by NIST and funded by the Disruptive Technology Office (DTO). It set out initially to benchmark the ad hoc search and retrieval operation on text documents and over the intervening decade and a half spawned over a dozen IR-related tasks including cross-language, filtering, web data, interactive, high accuracy, blog data, novelty detection, video data, enterprise data, genomic data, legal data, spam data, question-answering and others. 2005 was the 14th TREC workshop and 117 research groups participated. One of the evaluation campaigns which started as a track within TREC but spawned off as an independent activity after 2 years is the video data track, known as TRECVid and we shall give further details on TRECVid in the next section of this paper.

The operation of TREC and all its tracks was established from the start and has followed the same formula, basically

- Acquire data and distribute it to participants;
- Formulate a set of search topics and release these to participants en bloc;
- Allow about 4 weeks before accepting submissions of the top-1000 ranked documents per search topic;
- Pool submissions to eliminate duplicates and use manual assessors to make binary relevance judgments;
- Calculate Precision, Recall and other derived measures for submitted runs and distribute results;
- Host workshop in NIST in November;
- Make plans, and repeat the process ... for the next 16 years !

The approach in TREC has always been metrics-based — focusing on evaluation of search performance — with measurement typically being some variants of Precision and Recall. Following the success of TREC and its many tracks, many similar evaluation campaigns have been launched in the IR domain. In particular in the video/image area there are evaluation campaigns for basic video/image analysis as well as for retrieval. In all cases these are not competitions with "winners" and "losers" but they are more correctly titled "evaluation campaigns" where interested parties can benchmark their techniques against others and normally they culminate in a workshop where results are presented and discussed. TRECVid is one such evaluation campaign and we shall see details of that in section 4.

The Cross Lingual Evaluation Forum (CLEF) [12] is in its 7th iteration in 2006 and has 74 groups participating using a total of 12 different languages. CLEF tests aspects of mono- and cross-lingual IR through a variety of 8 different tracks including mono-, bi- and multi-lingual document retrieval on news, mono- and cross-lingual retrieval on structured scientific data, interactive cross-lingual retrieval and question-answering, cross-lingual image retrieval, and so on. CLEF is funded by the EU through the DELOS network.

NTCIR [9] is like CLEF, except it addresses Asian languages (Chinese, Korean and Japanese), and it is not as big. 2005 was the 6th running of NTCIR and it follows the TREC model quite faithfully. It covers multi-lingual, bi-lingual and single language retrieval on three Asian languages as well as question-answering.

INEX [6] is the Initiative for the Evaluation of XML Retrieval and 2006 is the 5th running of the cycle with 80 participating groups. INEX addresses IR which exploits available structural information (XML elements) to yield more focused retrieval and may retrieve a mixture of paragraphs, sections, etc. The collection used in 2006 is 659,300 Wikipedia articles from 113,483 categories with an average

of 161 XML nodes each. Unlike the other evaluation campaigns, and to keep costs down, participants in INEX must create candidate topics in order to gain access to the document collection. The main task in INEX is ad hoc retrieval plus tasks in natural language queries, heterogeneous documents, interactive, document mining and Multimedia [15].

Video Analysis and Content Extraction (VACE) is a US DTO funding program not restricted to US groups which just concluded Phase II with 14 funded participants and began Phase III. VACE addresses the lack of tools to assist human analysts monitor and annotate video for indexing. The video data used in VACE is broadcast TV news, surveillance, UAV, meetings, and ground reconnaissance and the tasks are detection and/or tracking of people, faces, vehicles and text in that data. VACE includes open evaluations with international participation in order to increase progress in problem-solving.

ETISEO [4] is an evaluation campaign that started in 2005, funded by the French government, with 23 participants. The aim to evaluate vision techniques for event detection in video surveillance applications. The video data used is single and multi-view surveillance of areas like airports, car parks, corridors and subways. The ground truth is annotations and classifications of persons, vehicles and groups and the tasks are detection, localization, classification and tracking of physical objects, and event recognition.

FRGC [5], the Face Recognition Grand Challenge, is an evaluation whose goal is to improve performance of face recognition algorithmns by an order of magnitude over the best results in the 2002 Face Recognition Vendor Test. The FRGC has provided data (50,000 recordings), including still and three-dimensional images, as well as computational infrastructure for work on two shared challenge problems and six predefined experiments. Nineteen groups submitted results for the 2005 evaluation.

PETS (Performance Evaluation of Tracking & Surveillance) [10] is in its 7th year in 2006 and is funded by the European Union through the FP6 project ISCAPS. PETS evaluates object detection and tracking for video surveillance, and its evaluation is also metrics based. Data in PETS is multi-view/multi-camera surveillance video using up to 4 cameras and the task is event detection for events such as luggage being left in public places.

The AMI (Augmented Multi-Party Interaction) project [1], funded by the EU, provides a test collection from instrumented meeting rooms, where the instrumentation includes video footage from multiple cameras, and is planning a series of evaluation campaigns. The tasks include 2D multi-person tracking, head tracking, head pose estimation and an estimation of the focus-of-attention (FoA) in meetings as being either a table, documents, a screen, or other people in the meeting. This is based on video analysis of people in the meeting and what is the focus of their gaze.

ImagEval [13] is a new evaluation campaign just launched this year, funded by the French government and now open to other Europeans. There are over a dozen participating groups and the tasks are related to content based image retrieval including recognition of image transformations like rotation, projection, etc., image retrieval based on combining text and image, detection and extraction of text regions from images, detection of certain types of objects in images such as cars, planes, flowers, cats, churches, the Eiffel tower, table, PC or TV, US flag, etc., and (semantic) feature detection - indoor, outdoor, people, night, day, etc.

ARGOS [2] is another evaluation campaign for video content analysis sponsored by the French government and has 10 French participating groups. The set of evaluation tasks have a lot of overlap with TRECVid and includes shot boundary detection, camera motion detection, person identification, video OCR and story boundary detection. The corpus of video used by ARGOS includes broadcast TV news, scientific documentaries and surveillance video.

Finally, we should mention two activities which bring together evaluation activities of others and they are Benchathlon [11] and CLEAR [3]. Benchathlon is a clearinghouse for data, annotations, evaluation measures, tools and architectures for content based image retrieval while CLEAR is a cross-campaign collaboration between VACE and CHIL (Computers in the Human Interaction Loop) concerned with getting consensus and crossover on the evaluation of event classification evaluation from video.

Although these evaluation campaigns span multiple domains and multiple applications, some of which are IR, they have several things in common including the following:

- they are all very metrics-based with agreed evaluation procedures and data formats;
- they are all primarily system evaluations rather than user evaluations;
- they are all are open in terms of participation and make their results, and some also their data, available to others;
- they are all have manual self-annotation of ground truth or centralized assessment of pooled results;
- they all coordinate large volunteer efforts, many with little sponsorship funding;
- they all have growing participation;
- they all have contributed to raising the profile of their application and of evaluation campaigns in general;

We will now look at one specific benchmarking evaluation campaign, TRECVid.

4. THE TRECVID BENCHMARKING EVALUATION CAMPAIGN

The TREC Video Retrieval Evaluations began on a small scale in 2001 as one of the many variations on standard text IR evaluations hatched within the larger TREC effort. The motivation was an interest at NIST in expanding the notion of "information" in IR beyond text and the observation that it was difficult to compare research results in video retrieval because there was no common basis (data, tasks, measures) for scientific comparison. TRECVid's two goals reflected the relatively young nature of the field - promotion of research and progress in video retrieval and in how to usefully benchmark performance. In both areas TRECVid has often opted for freedom for participants in the search for effective approaches over control aimed at finality of results. This is believed appropriate given the difficulty of the research problems addressed and the current maturity of systems.

TRECVid can be compared with more constrained evaluations using larger-scale testing such as the FRGC. In the

context of benchmarking evaluation campaigns it is interesting to compare those in IR and image/video processing mentioned above, with such a "grand challenge". The FRGC is built on the conclusion that there exist "three main contenders for improvements in face recognition" and on the definition of 5 specific conjectures to be tested. The FRGC shares with TRECVid an emphasis on large data sets, shared tasks (experiments) so results are comparable, and shared input/output formats. But the FRGC differs from TRECVid in that the FRGC works with much more data and tests (complete ground truth is given by process of capturing data), more controlled data, focus on a single task, and evaluation only in terms of verification and false accept rates. This makes it quite different to TRECVid.

The annual TRECVid cycle begins more than a year before the target November workshop as NIST works with the sponsors to secure the video to be used and outlines associated tasks and measures. These are presented for discussion at the November workshop a year before they are to be used. They need to reflect interests of the sponsors as well as enough researchers to attract a critical mass of participants. With input from participants and sponsors, a set of guidelines is created and a call for participation is sent out by early February. The various sorts of data required are prepared for distribution in the spring and early summer. Researchers develop their systems, run them on the test data, and submit the output for manual and automatic evaluation at NIST starting in August. Results of the evaluations are returned to the participants in September and October. Participants then write up their work and discuss it at the workshop in mid-November – what worked, what didn't work, and why. The emphasis in this is on learning by exploring. Final analysis and description of the work is completed in the months following the workshop and often include results of new or corrected experiments and discussion at the workshop.

5. TRECVID RETROSPECTIVE

TRECVid 2006 marks the end of 5 years of evaluation, the last 4 of which have worked with TV news. It's appropriate to take a look at what has changed and what has not, in preparation for charting a future course. Here we consider the core elements of the evaluation: tasks, data, and measurements as well as a review of approaches and results.

While the acquisition of data and the support of TRECVid at NIST is funded by DTO and NIST, only two or three participating groups are funded by DTO for their TRECVid research. All other groups find their own funding and participate because the TRECVid tasks fit the group's research agenda and promises sufficient return for their investment. Significant numbers of peer-reviewed publications based on TRECVid research (2002:10, 2003:17, 2004:46, 2005:39) reflect many independent community judgments of the importance and quality of the research particpants are doing – on the foundation provided by TRECVid.

5.1 Tasks

TRECVid is a laboratory, not a user or operational, evaluation of systems but the tasks aim to be abstractions of real user tasks. This link is important to ensure we addresse problems with implications outside the laboratory and because it helps in designing well-motivated rules for the evaluation. Component tasks are also evaluated as part of a "divide and conquer" strategy. The shot boundary determination and search tasks have been evaluated every year. They illustrate two levels of evaluation, each with its own advantages and disadvantages. In between is the high-level feature extraction task. Other tasks have been evaluated where truth data already existed or as pilot projects.

5.1.1 Shot boundary determination

Shots are automatically identifiable basic semantic units that are important in higher level video analysis such as search, browsing, and summarization. Even if TRECVid has demonstrated that the detection of abrupt boundaries (cuts) is largely solved for news video, the shot boundary task continues to provide an opportunity for new participants to overcome basic system and organizational problems before moving on to more complicated TRECVid tasks. It is an important component of higher level tasks.

Shot definition has also come to play an essential role in the TRECVid evaluation infrastructure. The first TRECVid search evaluation used no shared definition of the units of retrieval. This made judging inefficient and comparison of search results fuzzy because each system could retrieve a unique set of segments - many of which nevertheless shared many frames with segments retrieved by other systems. From 2002 onward, a single definition of shots was provided for the development and test data by one of the participants. These "master shots" then serve as the common units of retrieval for the search task and of analysis for the feature detection task added later.

In the shot boundary task we focus the evaluation microscope down onto an important but very narrow problem set - relatively distant from any real user task. In the search task, we zoom out to evaluate a task we can easily imagine as part of a real work context. In zooming in, we can say more about a smaller problem space, but have a hard time generalizing to a real application context. In zooming out we make it easier to draw conclusions about a real task but can say less, because the uncontrolled problem space is much larger. Both sorts of evaluations are needed.

5.1.2 Search

In the search task, the system (with or without a human in the loop) is presented with an as yet unseen multimedia statement of need for video containing certain named or generic objects, people, events, locations, etc. Following practice in TREC, such a statement is called a topic. The topic always contains a short textual description of the need as well as possibly image, video, and audio examples of what is desired. The topics may model an understanding of the need at the beginning of a search, after some successful searching, or as a standing profile.

The system's goal is then to return a ranked list of master shots from the test collection containing video of the sort desired. Ranking was initially foreign to some participants who saw the task as binary classification. But the volumes of data to be processed and the fuzzy nature of the queries mean modern search systems, whether as components or end user applications, must be able to provide information about relative confidence in their results.

Search system builders must find or develop various components and also integrate them. This complexity, especially when a user is included in the loop, requires good experimental designs if one is to draw conclusions about what works

and what doesn't in the presence of so many interacting factors.

5.1.3 High-level feature extraction

A third task, important in its own right and a promising basis for search, was added at the urging of participants in 2003: high-level feature extraction. The features tested have ranged over objects, people, and events with varying degrees of complexity that make some features very similar to topic text descriptions. Unlike topics, feature definitions are known in advance of testing and contain only a short text description. Participants have manually annotated training data for the feature task.

The TRECVid standard for correctness in annotation of feature training data and judging of system output is that of a human - so that examples which are very difficult for systems due to small size, occlusion, etc., are included in the training data and systems that can detect these examples get credit for them - as should be the case in a real system. This differs from some evaluations (e.g. FRGC) in which only a subset of examples that meet specified criteria are considered in the test. We want the TRECVid test collections to be useful long after the workshop in which they are created and even if systems improve dramatically.

Since in video there is no visual correlate of the word as an easily recognizable, reusable semantic feature, one of the primary hypotheses being examined in TRECVid is the idea that, given enough reusable feature detectors, such features might play something like the role words do in text IR. Of course, many additional problems - such as how to decide (automatically) which features to use in executing a given query - remain to be solved [14].

5.1.4 Additional evaluated tasks

TRECVid has addressed additional tasks against news video such as story boundary determination, specialized feature detection and camera motion analysis. Details of these tasks and how systems performed are available in the publications section of the TRECVid website [20].

5.2 Data

Data is the element of the evaluation with the fewest degrees of freedom. While one can ruminate about ideal test collections, in practice one more often takes what one can get – if it can at all be useful – and acquisition of video data from content providers has always been difficult in TRECVid. TRECVid has formally evaluated systems only against produced video but in 2005 and 2006 has explored tasks against unproduced, raw video as well.

5.2.1 Produced video

¿From the 11 hours of video about NIST used for a feasibility test in 2001, TRECVid moved in 2002 to 73 hours of vintage video mainly from the Internet Archive [7] – a real collection still needing a search engine to find video for re-use. Participants downloaded the data themselves.

Then in 2003 TRECVid began working on broadcast news video from a narrow time interval - a new genre, much more consistent in its production values than the earlier data and larger in size. Data set sizes made it necessary to ship the video on hard drives - a method that has worked well with the exception of one year in which groups with back-levels of Windows could not access drives of the size used.

Another important change was the shift to two-year cycles. Within the same genre enough data was secured so that training and test data could be provided in the first year, with the training data annotated and reused in the second year during which only new test data would be provided. This reduced the overhead of system builders adapting to new video, reduced the overhead of training data annotation and maximized its use, and removed a "new genre" factor from influencing results in the second year. TRECVid 2006 will complete the second such two-year cycle. data amounts (training/test in hours) have grown as follows: 2003 (66/67), 2004 (70/0), 2005 (85/85), 2006 (158/0). The video in 2003-2004 was from English-speaking sources. In 2005 and 2006 Chinese- and Arabic-speaking sources were added to the mix. Automatic machine translation was used to get English text from Chinese and Arabic speech.

We have learned that broadcast news video has special characteristics with consequences for the evaluation and systems. It is highly produced, dominated by talking heads, contains lots of duplicate or near duplicate material. Highly produced news video exhibits production conventions that systems will learn but with negative consequences when detectors learned on one news source are applied on another with different production conventions. This a real problem systems need to confront and makes it important that the training data come from multiple sources. There are 8 different sources and 11 different programs in the 2006 test data. A significant number of test data sources did not occur in the training data.

Much of broadcast news footage is visually uninformative - the main information is contained in the reporter's or anchorperson's speech. This makes the TRECVid search task more difficult because the topics ask for *video* of objects, people, events, etc. *not information* about them. Video of a reporter talking about person X does not by itself satisfy a topic asking for video of person X. The search task is designed this way because it models one of two work situations. One is an intelligence analyst looking at open source video, interested in objects, people, events, etc that are visible but not the subject the speech track, in the unintended visual information content about people, infrastructure, etc. The other is a video producer looking for clips to "re-purpose". The original intent often reflected in the speech track is irrelevant. Of course, the speech track (or text from speech) can be very helpful in finding the right neighborhood for browsing and finding the video requested by some topics. But even when speech about X is accompanied by video of X they tend to be offset in time.

Highly produced news video also exhibits lots of duplicate or near duplicate segments - due to repeated commercials, stock footage, previews of coming segments, standard intro and exit graphics, etc. Measuring the frequency of various sorts of duplicates or near duplicates is an unresolved research issue, as is assessing the distorting effect they may have on basic measures such as precision and recall.

5.2.2 Unproduced video - rushes

During 2005 and 2006 TRECVid participants have explored unproduced video - so called "rushes". By its nature this sort of video provides significant new challenges. Rushes are the raw material (extra video, B-rolls footage) used to produce a video. 20 to 40 times as much material may be shot as actually becomes part of the finished

product. Rushes usually have only natural sound. Actors are only sometimes present so very little if any information is encoded in speech. Rushes contain many frames or sequences of frames that are highly repetitive, e.g., many takes of the same scene redone due to errors (e.g. an actor gets his lines wrong, a plane flies overhead introducing extraneous noise, etc.), long segments in which the camera is fixed on a given scene or barely moving, etc. A significant part of the material might qualify as stock footage - reusable shots of people, objects, events, locations, etc. Rushes may share some characteristics with "ground reconnaissance" video.

It is not clear what doable tasks should be set for systems against this unstructured data so in both 2005 and 2006 participants were asked to develop and demonstrate some basic system capabilities to help a person unfamiliar with a large collection of rushes get an idea of what kinds of shots of what sorts of objects, persons, events, locations, etc could be found. The minimal required goals for 2006 are development of a toolkit with the ability to remove/hide redundancy of as many kinds as possible (i.e., summarize at one or more levels) and organize/present non-redundant material according to at least 6 features. The features should be well-motivated from the point of view of some user/task context and cannot all be of one type (e.g. not all cinematographic or camera setting). Groups may add additional functionality as they are able.

Evaluation of such functionality is known to be difficult. So part of the exploration will involve participants designing and performing their own evaluation and presenting the results. No standard keyframes or shot boundaries are provided.

5.3 Measurements

The TRECVid community has not spent significant amounts of time debating the pros and cons of various similar measures. They have profited by battles fought long ago in the text IR community. While choice of a single number (average precision) to describe generalized system performance is as useful (e.g., for optimization, results graphs) as it is restrictive, TRECVid continues the TREC tradition of providing various additional views of system effectiveness for their diagnostic value and better fit for specific applications and analyses.

In its first year TRECVid adopted a large set of shot boundary determination measurements from previous work [21] but soon adopted precision and recall with low threshold for overlap as the main measures. It added frame-precision and frame-recall to gauge separately the degree of overlap in the matches. For search and feature extraction TRECVid adopted the family of precision- and recall-based measures for system effectiveness that have become standard within the TREC retrieval community. Additional measures of user characteristics, behavior, and satisfaction developed by the TREC interactive search track over several years were adopted for use by interactive video search systems.

5.4 Approaches and Results

In what follows we look at approaches and results for the two most difficult, ongoing TRECVid tasks: high-level feature extraction and search.

5.4.1 High-level features

Most TRECVid systems have treated feature detection

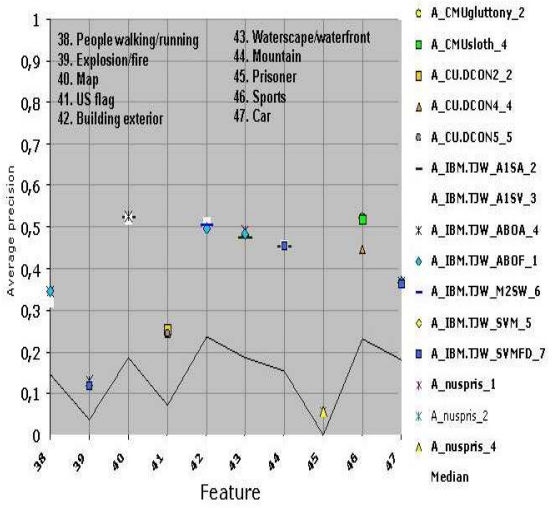


Figure 1: Average precision for top 3 runs by feature

as a supervised pattern classification task based on one key frame for each shot and have converged on generic rather than handcrafted detectors. This is the because of a desire to increase the set of features to many hundreds [8], in which case scalability of learning becomes critical. The TRECVid 2006 feature task recognizes this by requiring submissions for 39 features of which 10 will be evaluated.

Naphade and Smith [19] surveyed successful approaches for detection of semantic features used in TRECVid systems and abstracted a common processing pipeline including feature extraction, feature-based modeling (using e.g., Gaussien mixture models, support vector machines, hidden Markov models, and fuzzy K-nearest neighbors), feature-specific aggregation, cross-feature and cross-media aggregation, cross-concept aggregation, and rule-based filtering. This pipeline may accommodate automatic feature-specific variations [23]. They documented over two dozen different algorithms used in the various processing stages and note a correlation between number of positive training examples and best precision at 100.

Beyond the above generalizations, conclusions about relative effectiveness of various combinations of techniques are generally possible only in the context of a particular group's experiments as described in their site reports on the TRECVid website. In 2005 groups found evidence for the value of local over global fusion, multilingual over monolingual runs, multiple over singe text sources (Carnegie Mellon University), parts-based object representation (Columbia University), various fusion techniques across features and learning approaches (IBM), automatically learned feature-specific combinations of content, style, and context analysis, a larger (101) feature set (University of Amsterdam).

Even though the top 3 runs for each feature are very close to each other in performance as measured by average precision (see Figure *5.4.1*, there are significant differences in the top results — even in runs from the same group. If one sorts all the runs by mean average precision and takes the runs from the top until one has representatives from 10 sites

there are 33 runs. A partial randomization test [18] on the difference in the mean average precision scores shows significant ($p < .01$) differences between runs. Here is a list of how many runs (from the 33) each run is significantly better than. See the publications section of the TRECVid website [20] for details about the algorithms used:

24	A_IBM.TJW_SVMFD_7	1	A_UWAV3_4
22	A_IBM.TJW_ABOA_4	1	B_FD_PCA_LR_2
21	A_IBM.TJW_ABOF_1	1	B_FD_PCA_BC_1
18	A_IBM.TJW_A1SV_3	1	A_nuspris_1
18	A_IBM.TJW_SVM_5	1	A_UWAV2_2
18	A_IBM.TJW_A1SA_2	1	A_UWV1_3
17	A_IBM.TJW_M2SW_6	1	A_UWV3_6
17	A_CU.DCON4_4	1	A_UWV2_5
13	A_CU.DCON3_3	0	A_PicSOM_1
12	A_CU.DCON1_1	0	A_JOAMaxER_5
11	A_CU.DCON5_5	0	A_nuspris_4
8	A_CU.DCON2_2	0	A_tsinghua_6
4	A_CU.DCON6_6	0	B_FD_LPP_BC_3
3	A_CU.DCON7_7	0	A_CMUsloth_4
3	A_CMUgluttony_2	0	A_CMUwrath_5
2	A_UWAV1_1	0	A_CMUavarice_3
		0	A_ICL_NPDE_2

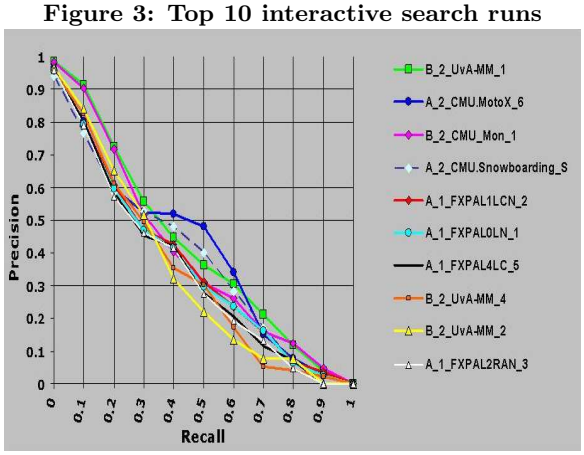
Many questions about detection of high-level features remain for researchers and for evaluation designers but several large ones deserve mention here.

- What are the most useful features for use in modeling a given video genre for a given purpose, e.g., broadcast news for intelligence analysts to search or filter?
- Are there opportunities for improved feature extraction using more than just one keyframe per shot?
- What are the limits on the generalizability of detectors, i.e., how reusable are the detectors, and how can we measure this in an affordable way? Changing data sets is expensive.
- Is it time to settle on an agreed (baseline) architecture and set of components in order to reduce the number of factors affecting results and thus to get more solid evidence for a few important causal relationships?
- Should TRECVID encourage or require groups to work with more than one keyframe per shot?
- How do we assess progress across multiple years and data sets?

5.4.2 Search

Hauptmann and Christel [16] discuss successful approaches to search. They note that, as one might expect for a genre full of talking heads, speech is an important and robust source of evidence in broadcast news and successful systems, used in the form of text. This is true for many topics but not all. Recall that the user being modeled is interested in objects, people, locations and events that were probably not intended as the focus of the original video and so are not being talked about. Successful video seeking in an interactive system may begin with text search or one based on image similarity or concepts but then continues by means of advanced browsing in the temporal domain, via image similarity (including near duplicates), using story boundaries,

Figure 3: Top 10 interactive search runs



and filtering with features at various levels. Experiments have demonstrated humans' considerable abilities to quickly skim, scan, locate the desired material and weed out the undesired. TRECVid interactive searches also make use of positive and negative relevance feedback. For every system, performance varies greatly by topic, as shown in Figure 2. Systems must provide a variety of tools, and users must avail themselves of them in an adaptive way.

The top 10 fully interactive runs clearly outperform their manual and automatic counterparts, as illustrated in Figures 3,4,5. Given the difficulty of the search task, the fact that the top 10 automatic runs in 2005 performed as well as most of the top 10 manually-assisted runs continues to astound. (Shallow precision scores for manual runs suggest the results could in fact be useful so we shouldn't conclude from the overlap of manual and automatic runs that the manual ones were just worse than we thought.)

Beyond these generalizations, drawing conclusions about what techniques work is difficult outside the context of a particular system. Effectiveness varies greatly with topic, collection, and user. Text from the speech remains a strong source of evidence for many topics, but in 2005, working with errorful, misaligned text from machine translation, some groups (e.g. IBM and MediaMill) found their visual-only search performed better than their text-only. In 2005 groups found value in query typing (Carnegie Mellon University), near-duplicate detection (Columbia University), multimodal over text-only search (Helsinki Univ. of Technology), cluster-temporal browsing (Oulu University), enhanced visualizations (FX Palo Alto). More details are available from individual site reports on the TRECVid publications page [20].

There are many open issues for evaluation design and system building. We note some major ones here:

- Can humans decide which concepts will help in executing a query?
- How can we compare interactive systems across sites?
- How do we encourage use of more than one keyframe per shot? Should we require it?
- How many near duplicate sequences are present in the TV news video and what effect does this have on systems, machine learning, and performance measures?

Figure 2: Mean average precision by topic

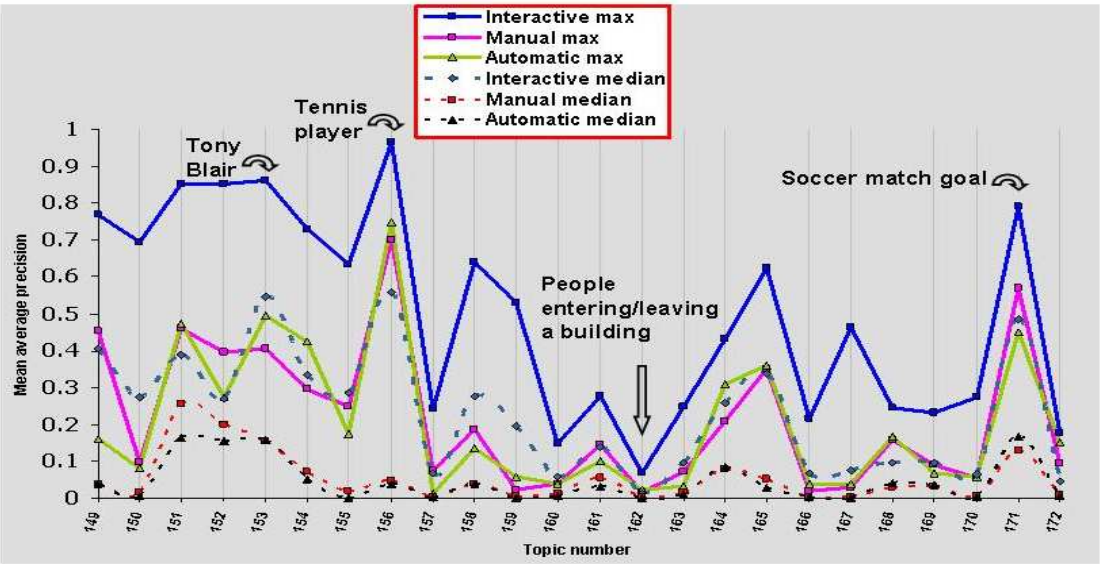


Figure 4: Top 10 manual search runs

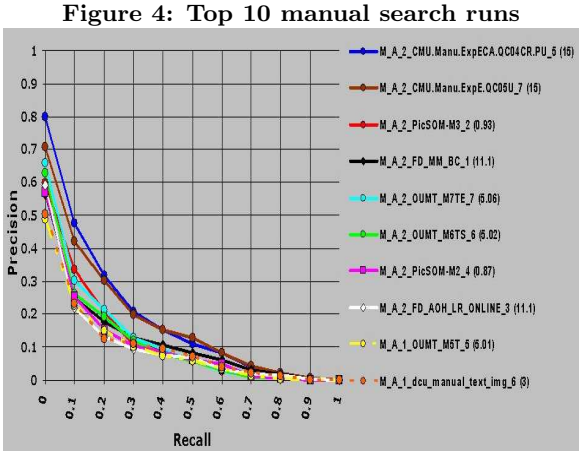
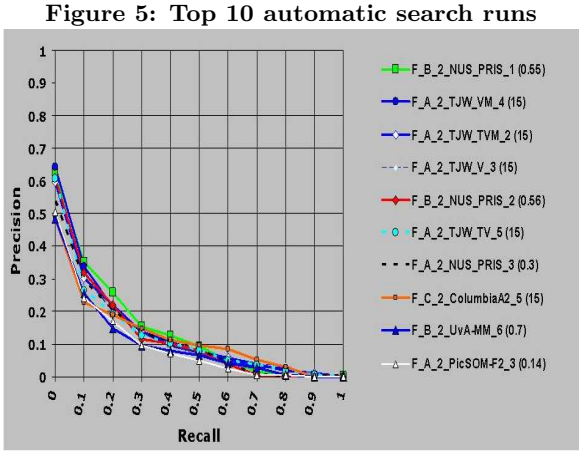


Figure 5: Top 10 automatic search runs



- Should the TRECVid search task be redesigned with fewer degrees of freedom for researchers and more focus on validating a small number of specific hypotheses?
- How do we assess progress across multiple years, data sets, and possibly users ?

6. BENCHMARKING EVALUATION CAMPAIGNS: PROS AND CONS

There are many good things about benchmarking evaluation campaigns, and there are some bad things. Let us examine these in turn, starting with the good things.

- The first, and most obvious good thing about evaluation campaigns is that they can secure, prepare, and distribute data, which is difficult to get. The participants can then use the same data, the same agreed metrics for evaluation and the same ground truth for measurement and this should allow direct comparisons across and within groups. Sometimes, where there are real users involved in the evaluation such as in the TRECVid interactive search task, the human subjects are a variable which cannot be controlled but for the most part comparisons across sites can be direct. Within a campaign, participants also complete the tasks at the same time and this can have benefits of sharing.
- A second, more indirect benefit of evaluation campaigns is that they can create critical mass and motivate donations of data and other resources to the campaign from among the participating groups. Here is a list of major donations to TRECVid 2005:
  - 50 hours of British Broadcasting Corporation rushes (BBC Archive)
  - National Aeronautics and Space Administration video from the Open-Video Project at University of North Carolina at Chapel Hill

  - Master shot reference (Fraunhofer Institute, Berlin)
  - Keyframes for each master shot (Dublin City University)
  - Feature annotation tools (IBM, Carnegie Mellon University)
  - Camera motion annotation tool & output (Joanneum Research, Austria)]
  - Feature annotation (20+ research groups) for 39 features in 50 hours of video
  - Low level feature detection output (Carnegie Mellon University)
  - Story segmentation output (Columbia University)

  These donations really enrich the evaluation and help to progress research in the field. The collaborations and assistance among participants also fosters a community and allows easier breaking into what is a new area for many people, and all this helps to improve overall performance of the tasks being benchmarked.

- By following the known and published guidelines for evaluation, either within or outside a formal evaluation campaign, a research group can perform direct comparisons with the work of others and know that their evaluation methodology is sound and accepted. Deviations from the campaign guidelines and "do-it-yourself" evaluations can introduce unforeseen biases into an experimental methodology.

- Good performance results can be a showcase for funding agencies, for industry and to help to promote a research area. When the collective achievement of participants in an evaluation campaign show good performance figures for a task the outside world can take notice and this kind of positive dissemination of research work can only be of benefit to all.

- Evaluation campaigns can facilitate research groups which want to gradually move into a new area of research. For example, in TRECVid groups can take part in the shot boundary detection task before moving onto search or feature detection.

- Groups can readily learn from each other since they are working on the same problems, data, using the same measures, etc. Approaches that seem to work in one system can be incorporated into other systems and tested to see if they still work. Groups just getting started reach better performance faster.

While these are the positives, there are also some possible negatives as follows.

- The first negative and the one which is thrown at evaluation campaigns most often is that everybody addresses the same research challenges using the same measures and so there is no room for diversity, and no scope for novelty or creativity. Here we disagree and point to the range of new approaches tried out each year in shot boundary detection and search tasks. Novelty and creativity are not stifled but operate within a shared research challenge. Novelty and creativity become even easier in an environment with so much collaboration and data/resource sharing.

- It is true that within evaluation campaigns the evaluation results and papers are usually available publicly but the original data can come with strings attached. This is generally because of copyright restrictions and the cost of purchase from the original owners and this is the case for most of the TRECVid video data where post-campaign, users must purchase the original video data from a supplier.

- There is a belief in some quarters that the agencies who fund evaluation campaigns have a stranglehold on the research directions of those evaluation campaigns and that they can overly-influence the research agenda. This is no more true than saying the same funding agencies have a stranglehold on research direction through the projects that they fund. Funding agencies throughout the world almost always publish their research priorities and strategic objectives and researchers react to these by shaping their research interests into the priorities of the funding agencies. Within the evaluation campaigns it is the participants who finally decide on the tasks to be benchmarked, the metrics to be used, albeit constrained by what is available and achievable by the coordinators. In practice it is the community more than the funders who have the stranglehold and it is the funders who set the restrictions on what their budget can afford.

- A valid criticism of evaluation campaigns is that the data set can both define and restrict the problems to be evaluated. Examples of data defining the tasks are story bound detection and anchorperson detection in TRECVid which were topical because the data over some of the latter years was broadcast TV news video where these tasks were quite important. An example of data restricting problems addressed is the overuse by many groups on keyframes as shot representatives. In TRECVid the organizers provide standard shot boundaries and standard keyframes so that interactive search systems used the same keyframes in their storyboarding and browsing interfaces, the motivation being to reduce the impact of yet another variable on the evaluation results. Yet this is an example of both good (it lowers the entry barrier to participation, and allows better system comparability) and bad (creates a path of least resistance and diverts attention from approaches that work with more of the moving video) [22] so as with many of these issues there is a trade-off.

- A final negative is that the set of problems we could address in future work is constrained by the dataset, that this is true, and that there is nothing we can do about it. But at least as a result of evaluation campaigns and the showcasing of results achieved, data owners and data providers may be more amenable to making their data available to the research community.

7. CONCLUSIONS

Many factors affect the design of evaluation campaigns and they require many choices among competing alternatives. The realization of such designs seldom goes entirely as planned and the evaluations have complex effects on the researchers and their work. No one evaluation type can answer all the questions. A research community needs a variety

of well-designed evaluations focused on high-level and low-level tasks, executable automatically many times or based on human judging carried out at the end of longer development cycles of months against approaches that have already shown real promise.

There is a life-cycle: have a new idea or discover something novel; reason about how to implement it, would it work, does it scale; try it out in-house on some local data; if it appears to work try it out on some data allowing comparison to others - i.e., an evaluation campaign — take part or use its data; if it appears to work then license it, publish it, showcase it. Evaluation campaigns are one stage in the lifecycle of idea-to-product. There is not always an available or appropriate benchmarking and nobody is forced into it, either as part of the annual iterations or to use the archived data afterwards

System-oriented evaluation campaigns like TRECVid have proved to be a fruitful way to concentrate the research efforts of a global community. The quality and importance of the work TRECVid has enabled is reflected in the number peer-reviewed publications and independent funding sources supporting the research. Yet, such campaigns by neccessity put restrictions on possible avenues that are explored and can affect the overall flow of research funds. Is the net effect on research progress positive ?

We think that there are strong indications that this is the case and have cited some of these. Still, this balance has to be evaluated regularly. TRECVid tries to carefully adapt its tasks, data sets, and measures over the years, maintaining a mix of healthy conservatism (recurring tasks, 2 year schedule) and pilot tasks. Also, the TRECVid program is to a large extent influenced by suggestions (e.g., the high-level feature task) from the participating community, which is open to all and continues to grow.

8. ACKNOWLEDGMENTS

Alan Smeaton acknowledges support from Science Foundation Ireland under grant number 03/IN.3/I361 and Wessel Kraaij would like to acknowledge support from the EU project AMI (IST-2002-506811).

9. REFERENCES

[1] AMI: Augmented Multi-Person Interaction. URL:www.amiproject.org/, Last checked 21 June 2006.

[2] ARGOS: Evaluation Campaign for Surveillance Tools of Video Content. URL:www.irit.fr/recherches/SAMOVA/MEMBERS/-JOLY/argos/, Last checked 21 June 2006.

[3] CLEAR'06 Evaluation Campaign and Workshop - Classification of Events, Activities and Relationships. URL:www.clear-evaluation.org/, Last checked 21 June 2006.

[4] ETISEO: Video Understanding Evaluation. URL:www.silogic.fr/etiseo/, June 2006.

[5] Face Recognition Grand Challenge. URL:www.frvt.org/FRGC, 2006.

[6] INEX: INitiative for the Evaluation of XML Retrieval. URL:inex.is.informatik.uni-duisburg.de/, Last checked 21 June 2006.

[7] The Internet Archive Movie Archive home page. URL:www.archive.org/movies, 2006.

[8] Lscom lexicon definitions and annotations. URL:www.ee.columbia.edu/dvmm/lscom, 2006.

[9] NTCIR: NII Test Collection for IR Systems Project. URL:research.nii.ac.jp/ntcir/, Last checked June 2006.

[10] PETS 2006: Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. URL:www.pets2006.net/, Last checked 21 June 2006.

[11] The Benchathlon Network: Home of CBIR Benchmarking. URL:www.benchathlon.net/, Last checked 21 June 2006.

[12] The Cross-Language Evaluation Forum (CLEF). URL:clef.isti.cnr.it/, Last checked 21 June 2006.

[13] The IMAG-EVAL Evaluation Campaign. URL:www.imageval.org/, Last checked 21 June 2006.

[14] M. G. Christel and A. G. Hauptmann. The Use and Utility of High-Level Semantic Features in Video Retrieval. In *Proceedings of the International Conference on Video Retrieval*, pages 134–144, Singapore, 20-22 July 2005.

[15] N. Fuhr and M. Lalmas. Introduction to the Special Issue on INEX. *Information Retrieval*, 8(4):515–519, 2005.

[16] A. G. Hauptmann and M. G. Christel. Successful Approaches in the TREC Video Retrieval Evaluations. In *Proceedings of the 12th ACM International Conference on Multimedia*, pages 668—675, New York, NY, USA, 10-16 October 2004.

[17] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer: the Kluwer International Series on Information Retrieval, 2005.

[18] B. F. J. Manly. *Randomization, Bootstrap, and Monte Carlo Methods in Biology.* Chapman & Hall, London, UK, 2nd edition, 1997.

[19] M. R. Naphade and J. R. Smith. On the Detection of Semantic Concepts at TRECVID. In *Proceedings of the 12th ACM International Conference on Multimedia*, pages 660—667, New York, NY, USA, 10-16 October 2004.

[20] NIST. TREC Video Retrieval Evaluation Publications. URL:www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, 2006.

[21] R. Ruiloba, P. Joly, S. Marchand-Maillet, and G. Quénot. Towards a Standard Protocol for the Evaluation of Video-to-Shots Segmentation Algorithms. In *European Workshop on Content Based Multimedia Indexing*, Toulouse, France, October 1999.

[22] C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, and F. J. Seinstra. On the surplus value of semantic video analysis beyond the key frame. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, July 2005.

[23] C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions, PAMI*, in press, 2006.