# Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience

Alan F. Smeaton

Adaptive Information Cluster,
Center for Digital Video Processing,
Dublin City University,
Glasnevin, Dublin 9, Ireland.
Alan.Smeaton@dcu.ie

**Abstract.** Information Retrieval is a supporting technique which underpins a broad range of content-based applications including retrieval, filtering, summarisation, browsing, classification, clustering, automatic linking, and others. Multimedia information retrieval (MMIR) represents those applications when applied to multimedia information such as image, video, music, etc. In this presentation and extended abstract we are primarily concerned with MMIR as applied to information in digital video format. We begin with a brief overview of large scale evaluations of IR tasks in areas such as text, image and music, just to illustrate that this phenomenon is not just restricted to MMIR on video. The main contribution, however, is a set of pointers and a summarisation of the work done as part of TRECVid, the annual benchmarking exercise for video retrieval tasks.

## 1 Introduction

The broad area of Multimedia Information Retrieval (MMIR) represents an intersection of work across several disciplines; by the very definitions of multimedia and of information retrieval this is inescapable and we now draw on computer science, engineering, information science, mathematics, statistics, human-computer interaction, networking, even hardware, and others in order to make progress in MMIR. However, the two areas which contribute most to current work in video retrieval and also in image retrieval are image/video processing and information retrieval. For MMIR, these two are mutually re-enforcing as in order to get usable descriptions of large amounts of video and image content we need to be able to analyse that video and image information, automatically; similarly, in order to be able to build effective content-based retrieval systems[1] for such information we need to use techniques developed in information retrieval research. Given the

[1] We define content-based retrieval systems to be those that support searching, browsing, summarisation, abstracting, (hyper-)linking, categorisation, clustering, and filtering . . . in fact any operations which work directly on image or video content.

origins of MMIR, it is worth looking at how its parent disciplines have come to regard large-scale evaluations.

For many years the main challenges driving the development of technology to support video in digital form were to be able to capture video digitally, to store it efficiently, to be able to package it on portable media, to send it over networks efficiently and then to render it on devices so people could then see it. Each of these tasks required the same essential ingredient in order to make them feasible for all but toy examples, namely effective *data compression.* In the early and mid 1980s vendors developed image processing software without due regard to the benefits of using standard formats for encoding images and users of image processing software had a real headache as proprietary formats predominated and interoperability across image processing software was only a pipe dream. We can see the remnants of that situation even today with the proliferation of image coding standards still available.

As computers became more powerful and manipulation of video in digital format on personal computers started to loom into the realms of possibility, the world became much more conscious of the benefits of using a standard format for encoding media. As the standardisation of encoding formats became a common goal in video processing, we became conscious of the importance of benchmarking and evaluation of video processing techniques, including compression, on standard datasets. A good example of this in video compression is the fact that the *mother and child* and the *table tennis* videos are so well-known and well-used as benchmarks for compression, that the people appearing in them are famous.

## 2 Evaluation Benchmarking Activities

In (text) information retrieval, the advantages of conducting retrieval experiments on a common dataset which was also being used by others has always been a central part of the IR discipline and test collections consisting of documents, queries or topics, and relevance assessments for those topics have been around since the earliest days. These tended to be small in size and not universally available but there has always been reasonable comparability among researchers in terms of empirical research work. About 15 years we saw a scale-up on the size of the collections and more importantly a concerted effort at providing relevance assessments on these large collections with the introduction of the first of the TREC (Text REtrieval Conference) exercises [1]. These have continued annually since 1991 and have spun off many "tracks" or variations including filtering, cross-lingual, non-English, web, high precision, high accuracy, genomic, question-answering, large scale and others. TREC, and its collections, evaluation measures and the sizes of its collections are now established in IR as the baseline for empirical text-based IR.

While TREC has had a huge impact on text information retrieval, its success in attracting researchers to the common benchmark has led to repeats of the basic approach for other media and IR tasks.

The *INitiative for the Evaluation of XML Retrieval* (INEX) [2] is an annual benchmarking exercise culminating in a workshop which examines retrieval performance from highly-structured XML documents. The retrieval of XML elements using a variety of techniques tried in INEX has shown net improvements in retrieval precision over its first few years. The music information retrieval community who target content-based music retrieval based on melody, etc. are launching a TREC-like evaluation in 2005 called "MIREX", the *Music Information Retrieval Evaluation eXchange* [3]. For the community of researchers who target content-based image retrieval the problem of creating a TREC-like evaluation is more difficult because so many CBIR researchers have been doing their own thing for so long, and because the subjectivity of relevance judgments in image retrieval is probably moreso than for any other media. However, a good starting point for evaluation of CBIR can be found in the Benchathlon effort [4].

## 3 TRECVid: Evaluation of Video Retrieval Tasks

In 2001 a new TREC track was introduced on the subject of video retrieval. Unlike previous tracks, the video track had more than one task and included shot boundary detection, feature detection and search as tasks. For the search task, the track followed the usual TREC mode of operation of gathering and distributing (video) data, formulating search topics, accepting search submissions from participating groups, pooling submitted results together and performing evaluation by comparing submissions against a ground truth of relevant shots derived from manual assessment of pooled submissions. The additional, non-search tasks had a similar model for operation involving many of the phases used in search.

In the first year the track had a small community of participants and this grew in the second and subsequent years. At the time of writing there are 63 groups signed up for participation in 2005 and the data for 2005 (development and testing) is over 200 hours of video.

In 2003 TRECVid separated from TREC because it was sufficiently different to the main TREC activity, and it had enough participation to be independent, although the TREC and TRECVid workshops are co-located each year.

The results obtained by participants in TRECVid each year are published in the TRECVid proceedings available online [5] and TRECVid activities have been published in a host of other places [6]. Overviews of the TRECVid activity have also been published previously [7], [8] and rather than repeat that here I will simply summarise the achievements of each of the tasks to date.

### 3.1 Shot Boundary Detection

The shot boundary detection task basically involves automatically determining both hard and gradual transitions between shots. Within TRECVid shots are the basic unit of information for search and feature detection and a common shot boundary detection is made available for these other tasks but for the SBD task a collection of approximately 5 hours is used, with manual ground truth established. The task is regarded as being one of the easier of the TRECVid tasks and has proven to be popular, and a good task with which to enter the area of video retrieval for groups wishing to break into it.

In 2004 we added performance speed as one of the metrics and this has revealed large differences in execution speed (from 1/30 to 3x real time) for approximately the same levels of performance. In general we believe that hard cut detection seems more or less “solved but there is still room for improvement in the detection of gradual transitions. Another question we are planning to address in 2005 is how well do the approaches transfer to other sources/types of video besides broadcast TV news ?

### 3.2 Story Segmentation

The task here is to use audio/video and the automatic speech recognition (ASR) transcript (including transcription errors) to segment a broadcast into individual news stories. This task is a more elaborate version of the task already tried with transcript-only in the Topic Detection and Tracking activity [9]. In TRECVid we have seen a wide range of approaches and of system complexity with the combination of AV and ASR giving only a small gain for segmentation over ASR only. Interestingly, most approaches are generic and not attuned to the peculiarities of the TV broadcasters we used (ABC and CNN).

Although this task ran for only 2 years (2003 and 2004) and the results improved in the second year, the overall results obtained show there is still further room for improvement.

### 3.3 Feature Detection

The automatic detection of features is potentially one of the real enablers for video retrieval. Being able to pre-process video to automatically detect a range of mid- and high-level semantic features would make video retrieval, and post-retrieval result clustering a powerful tool. The difficulty in achieving this is partly determining what makes a good set of features to target for automatic detection (in terms of complementarity and achievability), as well as then realising those detections. Because of its importance, feature detection has been present in each of the years of TRECVid.

The usual mode of operation for feature detection is to take a baseline of annotated video, usually manually annotated, and to train some kind of classifier on this data. Support Vector machines (SVMs) and other machine learning approaches have proved to be popular and feature-neutral systems are particularly attractive since they can be re-applied to new features without a lot of re-engineering needing just more and more training data. the supply of training data for this task has proved a very large obstacle though we are fortunate that in 2003 IBM coordiated an effort among TRECVid participants to annotate a corpus of video data which was then made available to all participants. In 2005 we are repeating this annotation exercise and when we sough volunteers to assist with this we had more than enough volunteer effort to annotate the whole set of training data (80 hours), twice, which is exactly what we are doing ! This training data should be an invaluable resource for those developing feature detectors.

### 3.4 Searching

In TRECVid, search involves matching a multimedia expression of an information need against a corpus of video ad retrieving a ranked list of shots which satisfy the information need. the topic is expressed as a combination of text, example image(s) and example video clip(s). There are 3 types of searching facilitated in TRECVid which vary in the amount of user intervention allowed in the searching process. In the *interactive search task* the user is given a maximum of 15 minutes to use a search tool to find relevant shots; in the *manual search task* the user is allowed to formulate the topic as a query, once, and this is then submitted to the system which produces a ranked list of shots; in the *automatic search task* there is no user intervention at all and the topic is submitted to the system verbatim.

After 4 years of the search task we are genuinely surprised by the amount of variation and creativity that participants introduce into the process. There are many interesting shot browsing interfaces which use the keyframes provided by TRECVid yet the text search which is run against the automatic speech transcripts (ASR) continues to be the single most important modality for video retrieval, being far more important than retrieval based on visual features. There is some use of high-level feature concepts which are made available from the feature detection task (outdoors, people, water, . . .), and a lot of use of low-level features (color, edges, texture, . . .) to enable query by visual similarity against the keyframes. Many groups have managed to use browsing in the temporal neighborhood of an already found shot, leveraging the fact that a story about flooding will likely have several shots of flood waters and thus local context browsing is useful for shots about flood waters. Some groups also use outside resources to enhance their text search, resources such as Google and WordNet and while most groups use positive relevance feedback, not so many use negative relevance feedback.

At the end of a 2-year cycle of evaluation of shot retrieval from TV news we can say that the evaluation has stabilized but in 2004 we did not get any giant leap forward in systems which suggests we have reached a plateau. Yet against this we continue attract more and more groups, each of which brings in new approaches and techniques the best of which are then picked up by others. Furthermore we continue to gain insights and improvements, even in the better systems. Most impressive of all are the range of demonstrations of these systems which when viewed as a collective are at, or near to, the state of te art in video retrieval

### 3.5 TRECVid in 2005

TRECVid 2005 is likely to see growth in the number of participants (63 groups signed up for 2005 whereas just over 30 completed at least 1 task in 2004), in the amount of data (nearly 220 hours of video), and in the complexity of the video (166 hrs of English, Chinese and Arabic news from Nov 2004 with ASR and machine translation to English as well as 50 hours of "rushes" from BBC and further data from NASA educational programs). The specific tasks to be included in 2005 are:

- Shot boundary determination, which is the same as previously except on the NASA data
- Low-level feature extraction (mostly camera motion)
- High-level feature extraction using 10 semantic features, some of which are a repeat from 2004
- Search (interactive, manual, and automatic) on the broadcast TV news in 3 languages, plus a new pilot task on BBC rushes

The search tasks in 2005 are by far the most challenging to date. The noise introduced by using far more broadcast sources as opposed to the CNN/ABC pair used previously, as well as the noise introduced by the machine translation of Chinese and Arabic to English will pose real challenges, especially given the way text has continually shown to be the most important of modes for retrieval. The other challenge is in the task of providing search on the BBC rushes data. Rushes are the raw video clips which are then edited and post-produced into TV programs and the rushes contain lots and lots of redundancy both in the repeated re-filming of a given shot as well as the lingering of the camera before and after the "action" component of the shot. Most interestingly, there is generally no dialogue in rushes video, so there is no ASR and no basis for using text, meaning that the search on rushes will have to be video-only. Definitely, this will be challenging !

## 4 Conclusion

In this paper we have provided a short taster of the work undertaken each year as part of the TRECVid evaluation exercise as well as some pointers to other benchmarking exercises in related fields in IR. TRECVid continues to grow from strength to strength and is probably the single most influential activity in the area of video retrieval. This is something we have to be cautious about, however, and we must be aware of the danger of the field wrapping itself around TRECVid and its tasks rather than having TRECVid as something which should support the development of the field of video information retrieval.

## Acknowledgments

The author wishes to acknowledge the considerable input of Paul Over (NIST) and Wessel Kraaij (TNO) who, along with the author, coordinate TRECVid each year. The author is partly-supported by Science Foundation Ireland under grant 03/IN.3/I361.

## References

1. The TREC Website. Available at http://trec.nist.gov Last visited April 2005.
2. Kazai, G., Lalmas, M. and de Vries, A.P. The Overlap Problem in Content-Oriented XML Retrieval Evaluation. In: *Proceedings of the 27th annual international conference on Research and development in information retrieval*, SIGIR '04, pp. 72–79, Sheffield, United Kingdom, ACM Press, 2004.
3. MIREX 2005: 2nd Annual Music Information Retrieval Evaluation eXchange. Available at *http://www.music-ir.org/evaluation/MIREX/* Last visited April 2005.
4. The Benchathlon Network: Home of CBIR Benchmarking Available at *http://www.benchathlon.net/* Last visited April 2005.
5. TREC Video Retrieval Evaluation Online Proceedings. Available at *http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html* Last visited April 2005.
6. TREC Video Retrieval Evaluation Bibliography. Available at *http://www-nlpir.nist.gov/projects/t01v/trecvid.bibliography.txt* Last visited April 2005.
7. Smeaton, A.F., Kraaij, W. and Over, P. The TREC Video Retrieval Evaluation (TRECVID): A Case Study and Status Report. In *Proceedings of RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, Avignon, France, 26-28 April 2004.
8. Kraaij, W., Smeaton, A. F. , Over, P. and Arlandis, J. TRECVID 2004 - An Overview. In *Proceedings of TRECVid 2004*, National Institute for Standards and Technology http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/tv4overview.pdf Last visited April 2005.
9. J.G. Fiscus and G.R. Doddington. Topic Detection and Tracking Evaluation Overview. In: *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Academic Publishers, Norwell, MA, USA, pp. 17–31, 2002.