TRECVID: Evaluating the Effectiveness of Information Retrieval Tasks on Digital Video

Alan F. Smeaton
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland.
Alan.Smeaton@dcu.ie

Paul Over
Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD
20899-8940, USA.
over@nist.gov

Wessel Kraaij
Department of Data Interpretation
TNO TPD
PO BOX 155
2600 AD Delft, the Netherlands.
kraaij@tpd.tno.nl

ABSTRACT

TRECVID is an annual exercise which encourages research in information retrieval from digital video by providing a large video test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. TRECVID benchmarking covers both interactive and manual searching by end users, as well as the benchmarking of some supporting technologies including shot boundary detection, extraction of some semantic features, and the automatic segmentation of TV news broadcasts into non-overlapping news stories. TRECVID has a broad range of over 40 participating groups from across the world and as it is now (2004) in its 4th annual cycle it is opportune to stand back and look at the lessons we have learned from the cumulative activity. In this paper we shall present a brief and high-level overview of the TRECVID activity covering the data, the benchmarked tasks, the overall results obtained by groups to date and an overview of the approaches taken by selective groups in some tasks. While progress from one year to the next cannot be measured directly because of the changing nature of the video data we have been using, we shall present a summary of the lessons we have learned from TRECVID and include some pointers on what we feel are the most important of these lessons.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Video Evaluation

General Terms

Measurement, Design, Standardization

MM'04, October 10-16, 2004, New York, New York, USA.

1. TRECVID – THE PROCESS

TRECVID is an annual benchmarking exercise which encourages research in video information retrieval by providing a large video test collection, a set of topics or queries, uniform methods for scoring performance, and a forum for organizations interested in comparing their results. In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic video segmentation, video indexing, and content-based retrieval of digital video. Beginning in 2003, this track became an independent evaluation (TRECVID) with a 2-day workshop taking place just before TREC. The history of the TREC video track and the emergence of TRECVID can be found in [2] and in this paper we concentrate on the 2003 cycle of TRECVID.

The operation of TRECVID requires the organisers, NIST, to acquire and distribute a large collection of digital video to all participating groups using HDDs. In 2003, the video data consisted of approximately 120 hours (241 $\times$ 30-minute programs) of ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998. In addition, we used 13 hours of C-SPAN programming (30 $\times$ 10- or 20-minute programs) mostly from 2001 consisting of various government committee meetings, discussions of public affairs, lectures, news conferences, public hearings, etc. This collection of video was divided into a training set used to develop and tune systems, and a search set used as the target collection for a number of search topics, as well as for the detection of a number of pre-defined features.

In addition to the video, in MPEG-1, participants were given access to a common shot boundary definition generated by the CLIPS-IMAG group and a set of keyframes extracted from those shots. Having an agreed set of shot bounds means that there is a common basis which defines the unit of retrieval, and the unit identified in other tasks carried out within TRECVID. Groups were also provided with the output of an automatic speech recognition system provided by LIMSI [1] and a closed-captions-based transcript.

Within TRECVID 2003 there were 4 specific tasks that groups were invited to participate in. *Shot boundary detection*, run on a subset of about 5 hours video, required groups to identify the boundaries between different camera shots, both hard cuts and gradual transitions. *Story bound segmen-*

tation required groups to segment TV news broadcasts into non-overlapping news stories. *Feature extraction* involved identifying shots with some of a set of 17 pre-defined features namely outdoors, news subject face, people, building, road, vegetation, animal other than a human, female speech, car/truck/bus, aircraft, news subject monologue, non-studio setting, sporting event, weather news, zoom-in, physical violence, and person X, namely Madeleine Albright. The final task required groups to perform *search* using each of 25 topics or search statements against a subset of about half of the 130 hours of video. Topics were multimedia expressions of an information need and included text and one or more sample images or video clips illustrating the topic. Topics were formed by NIST staff who viewed the video content and then formulated topic descriptions of an information need and who then assessed the shots submitted by each group in terms of relevance. There are two tasks that this models as follows. In *automatic search* the topic definition is turned into a system query which is used as input into the video retrieval system and the ranking (of shots) generated as a result is evaluated. In *interactive search* the scenario modeled is where a user is given a fixed timeframe (15 minutes max.) to locate relevant shots and is allowed to run as many interactive searches and do as much browsing as s/he wishes. Both search modes generate ranked lists of shots which are returned to NIST and once this human-assessed ground truth was available then we were able to evaluate the retrieval performance of each participating group using measures based around precision and recall.

At the time of writing we are in the middle of a two-year TRECVID cycle and this year we are using more of the same kind of data (broadcast TV news) and from the same source, as in 2003. This provides some stability for groups who participated in 2003 and allows them to enhance their systems developed for 2003 or to explore some other aspect of retrieval or feature extraction.

One of the main obstacles to the successful development of features is the availability of accurately marked-up data and to address this a group from IBM Research coordinated a collaborative annotation of approximately 60 hours of video content using over 130 different semantic labels. In this effort, researchers from 20 groups worldwide manually annotated this video data over several weeks in order to provide a source of training data for feature extraction. With this approach we have enabled groups to compare the performance of their feature extraction techniques where a common training set has been used. The features task is partly motivated by the idea that automatic annotation of video with semantic concepts has the potential to improve efficiency and or effectiveness of the search task though increased retrieval effectiveness as a result of the use of features in retrieval is still an open research issue.

2. TRECVID2003 : WHAT GROUPS DID AND LEARNED

Eleven groups completed either interactive or manual searching or both and it is interesting to examine what lessons have been learned from across the sites. Most sites ran more than one variation of their own system within their own participation and that allowed them to focus on examining some aspect of retrieval, but because there was no common search system across sites, and because user groups at each site had different backgrounds and experience levels and were working under different conditions, cross-site comparisons are still not really possible.

Of those who performed interactive searching, some groups used many test users in their experiments (UNC used 36 and MediaMill used 44) while others used fewer (Oulu used 8, CMU used 5, Imperial College used 4, Indiana used 1) and it was generally felt that the more users the better the experiments. Most groups explored variations of their system which compared an ASR text-based retrieval only against retrieval based on text plus features, The best example of this is the UNC group who found that a feature-only system is weaker than a text-only system which in turn is weaker than a text-plus-features system, across the set of 25 topics.

For some groups the question they explored was how to calculate the relative weighting between ASR-based retrieval and feature-based retrieval, and how to weight the contribution of different features. CMU explored fixed vs. per-query weighting schemes for determining how to combine the contributions from different features, including text, in retrieval. The Lowlands group also examined ways of combining ASR text retrieval using a language model, with feature-based retrieval using generative mixture models, while the IBM group termed this multimodal retrieval and, like CMU, examined query-dependent weighted combinations of ASR-based and content (feature) based retrieval. The submissions from the group from Fudan University allowed them to examine combined retrieval based on different approaches, namely text-based combined with several feature and colour-based approaches. While there is no definite consensus from these investigations it seems that the effectiveness of combining individual approaches to retrieval in different ways depends on the nature of the topic, with some topic types doing well for certain types of overall retrieval, i.e. constituent retrieval approaches used in particular combinations.

For other groups the main topic of interest was the interface they developed and how effective it was in supporting user browsing through shots. Those groups who did perform extensive interactive experimentation tended to follow the principles for proper experimentation which were encouraged by the track and that is a positive development. The group from DCU is one such group which concentrated on interactive experimentation as is the CMU group who allowed users to use a sophisticated interface tool to manipulate sets of shots and support advanced browsing and visualisation. The group from Imperial College also used an advanced visualisation technique with keyframes displayed as thumbnails whose respective distance from the center of the screen is in proportion to their dissimilarity to the query. The group from UNC similarly evaluated their retrieval system from the perspective of how well it supports user browsing.

In general it was found that the performance of interactive retrieval was better than the performance of manual retrieval which is not a surprise and the confirmation that different types of topic are better supported with different types of overall retrieval seems to have spawned further work in this area since last year.

The results from other tracks are less confusing than the results from search. For shot boundary detection, the results of the best systems are excellent and for TV news story segmentation there was a great spread in performances with the

best being quite usable, but far from perfect. For feature detection, the performance of the best-performing approaches is variable with some features such as news subject monologue, weather news and sporting event yielding quite good performance while the detection of builldings and physical violence was quite poor. Some of this could be put down to the community only slowly starting the use the output of the collaborative annotation mentioned earlier and in fact many sites used their own data for training which doesn't make cross-site comparisons easy, but this should improve for 2004.

3. TRECVID – LESSONS LEARNED

There are several lessons we have learned and contributions we believe TRECVID has made to the research community and as we come to the end of this fourth annual cycle of TRECVID it is worth reflecting on some of these. The most direct and visible contributions are the kick-starting and the synchronisation of the actual benchmarking activity. Coordinating a quite complex activity among 24 research groups in 2003 (and over 40 in 2004) in a compressed timeframe while maintaining the balance between freedom to allow sites do what interests them while still contributing to the collective, has really tested the willingness of the track participants. Yet one of the greatest contributions that TRECVID has made and continues to make is that more than one corpus of video data, with topics and ground truth for shot bound detection, story bound detection, some feature detection, and searching for video shots, is now available within the research community. This has been collected and is made available to TRECVID participants under very reasonable access conditions and has enabled new, and established, research groups to work on common datasets and as has been shown in the main TREC activity, this does act as a catalyst for research in the field. The difficulties associated with doing this should not be underestimated since issues of copyright ownership and access are real hurdles to more widespread research in this area. Indeed we are seeing more and more usage of this dataset in non-TRECVID experiments appearing in publications and a good example of is this conference itself which has several papers based on TRECVID data.

Except in the area of evaluating the search task, we have progressed the development of new evaluation measures for different video retrieval tasks. For example, in shot bound detection task the track participants have developed and used a variation of precision and recall called frame-precision and frame-recall which accounts for inexact overlaps between a detected gradual shot transition (a fade for example) and the ground truth for that transition.

Within-site comparisons of different techniques for the search task are being carried out by most groups as part of their TRECVID activities but a major amount of *cross-site comparisons* of different search techniques and tools has still not been realized. This means that a given site may be able to make comparisons among variations of its own search techniques, but the bigger questions of how different search techniques developed by different groups compare, has yet to be addressed thoroughly. There is some amount of this when, for example, a researcher identifies a technique used by someone else, incorporates it in his/her system in the next year and tests its effectiveness. This happens across the annual cycles in TREC tracks as new ideas which prove that the improve performance are picked up by others and we expect this to occur in TRECVID, especially where there is stability in the dataset across a number of years like we have now with broadcast TV news. In time we may achieve more formal cross-site comparisons by testing against common baseline systems used across different sites and in 2004 we are seeing some cross-site usage of the same system. The same problem does not arise for the other TRECVID tasks, so shot boundary detection, story bound detection and feature detection results from different groups and sites can be compared directly.

While having stable video data over a couple of cycles of TRECVID does bring advantages it also has a downside and one issue with the TRECVID video data is that by necessity the video data we have used has been very genre-specific. In the first and second years of the TREC video track we used data (NIST documentaries and Prelinger archives from the Open Archive project) simply because it was the only video data that was available to us. In 2003 and 2004 we are using broadcast TV news from CNN and ABC, and this genre strongly influences the kind of search tasks and feature extraction tasks that we cover. Ultimately there is no such thing as the ideal video collection with the right amount of every kind of video we could search, so achieving this is impossible. We are aware of the restriction that the video data imposes and it is our intention to broaden the video data type in future (post-2004) iterations of TRECVID.

Another issue which has been somewhat frustrating for some has been that the findings and the impact of the interactive search task in TRECVID has not yet been as significant as was hoped. By its very nature, effective searching through large video archives will need to involve far more visual browsing than in many other kinds of information seeking. Currently TRECVID participants in the interactive search task build systems for interactive use and we then measure interactive search by putting a time limit on the user searches and we calculate measures such as precision and recall, ignoring issues of user satisfaction with the accomplished task. We are not alone here in that this is a characteristic of most information retrieval research, though it is changing. In TRECVID we require that the elapsed search time be reported for each manual and interactive search but we have yet to see any correlation between retrieval effectiveness and time on search. We do encourage groups to collect all sorts of user satisfaction information through user questionnaires etc., but this information is very expensive to gather and to sift through. Furthermore, it is not of interest to all groups and to date has not yielded much insight and so we dont require user satisfaction information to be gathered, which is part of the low entry barrier to participation in TRECVID. While some groups in TRECVID have looked at other measures besides precision and recall to measure user searching, there is a lot more work to be done to move us towards an agreed framework for measuring real user satisfaction.

To date in TRECVID we have not really addressed video retrieval from the video itself and there has been too much dependence on the closed captions and ASR text within the search task. There are many topics, even in TRECVID, for which the ASR text does not offer any help, such as the search for a locomotive approaching the viewer, or aerial shots containing one or more buildings and one or more roads. Some TRECVID groups have analysed topics into

those which for which the ASR is helpful and those for which the ASR is of no help in retrieval. This falls short of automatically classifying topics into these two types and indeed subsequent to TRECVID in 2003 there is work ongoing by some group(s) on automatically classifying queries to video databases and then using the classification to determine the retrieval strategy to be taken. This highlights the need for more development of searching based on visual aspects. Also, there are other video types such as CCTV footage and personal (home) video for which ASR will not be available. In these cases we need to develop techniques which do retrieval from the video however it would have been wrong to have forced the development and evaluation of visual-only retrieval systems in TRECVID where the genre of the video data was broadcast TV news and to impose non-ASR retrieval on this would have been unnatural and created unrealistic retrieval scenarios. As it stands the retrieval scenario we simulate is one where a searcher is asked to find as many *shots* from an archive of broadcast TV news which match some retrieval criterion and for which there are sample images or short video clips available. While such a scenario is conceivable it is not commonplace, but it is what we had to do given the data we had at the time.

There are many opportunities where tracks within the main TREC activity could usefully be combined with the TRECVID activity to create interesting hybrids. For example, techniques and measures developed in the interactive (text) TREC searching track could be tried in TRECVID; the novelty detection task in TREC is currently applied to text documents, but could be applied to video, such as TV news stories; the question-answering track currently searches through text archives but could equally pose questions to a video archive; document summarization techniques developed for text documents may have application in summarising video of TV news, drama, sports events, etc. These each represent exciting possible developments.

Several of the groups in TRECVID have treated video data as a series of still images and not as a medium of its own with object and camera motion, etc. Part of the reason for this is that there are not many groups who have the resources to handle motion video as more than a series of still images and TRECVID may be doing a disservice to the community by providing "standard keyframes to be used as in this way we may be encouraging the treatment of video as a series of still images.

Another major challenge for TRECVID is designing the evaluation in such a way as to maximize what we learn from the exercise as a whole while allowing each site the independence to do what it is interested in, bearing in mind that participation in TRECVID is a voluntary exercise and funded from within participants' own resources of research income. Within-site comparisons which take place need to use good experimental design techniques in order to maximize what they learn from their experiments and TRECVID does provide guidance in this area. Small, voluntary multi-site experiments such as the common annotation forum in 2003 are an example of multi-site activities and as these will occur more often we need to get a balance between allowing a site freedom and independence, and marching in lockstep with the rest so that the track as a whole can learn.

Finally, as is always the case with TREC activities, TRECVID does suffer from a drop-out rate where groups sign up to take part but under-estimate the scale of resources or development needed in order to achieve the completion of some task(s). A rule of thumb in TREC is that approximately 25% of those groups who apply to take part will drop out. In TRECVID the dropout rates have been higher (17 complete from 29 sign-ons in 2002, 24 complete from 44 sign-ons in 2003) and this may be due to the difficulty of the task. We have tried to address this by making all submissions to the feature detection task available in time for that data to be used as part of any other groups search task. This data, like the shot boundary data and any other data exchanged among participants, is exchanged in MPEG-7 in order to promote the take-up of standards.

4. CONCLUSION

It is difficult to summarise the impact of TRECVID in a few short paragraphs because it is still very much a "work in progress" and ultimately history will judge its true contribution. What we can say is that it has growing popularity within the research community, is attracting attention from academic and corporate research teams, seems to be making progress with helping to advance the area of video retrieval and from a scientific perspective has raised many interesting issues to do with retrieval. It has also helped to overcome one of the major bottlenecks in doing video retrieval research, namely legitimate access to a large video corpus which can be shared by different teams. While there are issues which make TRECVID less than ideal we believe that it will continue to have a substantial impact on research into video information retrieval.

Acknowledgements

On behalf of the entire TRECVID track, the authors acknowledge the substantial contribution made by Jean-Luc Gauvain of LIMSI and Georges Quenot of CLIPS-IMAG who have provided automatic speech recognition and shot boundary/keyframe data for the track participants for a number of years. We would also like to thank all the participants and other contributors on the mailing list whose enthusiasm, patience, and sustained hard work continue to amaze.

5. REFERENCES

[1] J.L. Gauvain, L. Lamel and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89-108,2002.

[2] A.F. Smeaton, W. Kraaij, and Paul Over. TREC Video Retrieval Evaluation: A Case Study and Status Report. In *Proceedings of RIAO'2004, Coupling approaches, coupling media and coupling languages for information retrieval*, Avignon, France, April 2004.