TRECVID: Benchmarking 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

Many research groups worldwide are now investigating techniques which can support information retrieval on archives of digital video and as groups move on to implement these techniques they inevitably try to evaluate the performance of their techniques in practical situations. The difficulty with doing this is that there is no test collection or any environment in which the effectiveness of video IR or video IR sub-tasks, can be evaluated and compared. The annual series of TREC exercises has, for over a decade, been benchmarking the effectiveness of systems in carrying out various information retrieval tasks on text and audio and has contributed to a huge improvement in many of these. Two years ago, a track was introduced which covers shot boundary detection, feature extraction and searching through archives of digital video. In this paper we present a summary of the activities in the TREC Video track in 2002 where 17 teams from across the world took part.

1. Introduction

TREC is an annual exercise which has been running for 11 years and benchmarks the effectiveness of systems on various information retrieval tasks. TREC has world-wide participation and in the 2002 running, 93 groups took part in a variety of specialist "tracks" or activities. The TREC philosophy has always been to facilitate open, metrics-based evaluation and over the last few years TREC has run tracks on searching web documents, cross-lingual information retrieval, retrieval from spoken documents, retrieval from text documents in languages besides English such as Spanish and Chinese, question-answering, retrieval from text documents which have been corrupted by an OCR process, and others.

TREC is coordinated by the National Institute for Standards and Technology (NIST) in Gaithersburg, Md., USA, though groups taking part in TREC are funded from other sources or are self-funded. The *modus operandi* for TREC is that NIST

will gather and distribute data (web pages, text documents, spoken radio news broadcasts, etc.) to participants who have signed up for a TREC track. Participants then install this data on their own IR system – depending on what the track is about – and on a given date, NIST will distribute a set of perhaps 50 topics or descriptions of an information need. Each group will then run each topic on the data using their own system and will send back the top-ranked X documents, where X could be as large as 1000, depending on the task. NIST then pools the top-ranked N (e.g. 100 or 200) documents from each submitted result from across all participating groups for each of the topics and then this pool of "documents" is manually evaluated for relevance to the topic in question. This establishes a ground truth of relevant documents for each topic and once this ground truth is available, the performance of each group' submitted results can be measured against this ground truth using measures such as precision and recall. The ground truth produced by pooling for the TREC text document collections has also been demonstrated to be useful for evaluating systems which did not contribute to the pool.

In 2001 a track on video information retrieval was introduced into TREC, covering three tasks, namely shot boundary detection, high-level feature detection, and searching. The goal of the track was to promote research in content-based retrieval from digital video by beginning the development of a laboratory-style evaluation using tasks which, although abstracted to make them manageable, modeled realworld tasks. In 2002 this track had 17 participating teams, up from 12 in the previous year, and used 73 hours of video material, up from 11 hours the previous year.

Acquiring realistically useful video data for research purposes is a notoriously difficult task because of copyright considerations and in TREC 2002 we used video data mainly from the Internet Archive Movie [1]. This consisted of advertising, educational, industrial and amateur films from 1930 to 1970 which was produced by corporations, non-profit organisations, trade groups, etc. This data is not ideal in that it is noisy and some of the 170+ videos have a sepia tint, but it does represent real archive data which people do want to search among. The 73.3 hours of data used was partitioned into 4.85 hours used for the shot boundary test, 23.26 hours used for training of feature detectors, 5.07 hours of feature testing and 40.12 hours used for the search testing. For TREC 2002 the test collections for feature extraction and search were segmented into shots by one of the participating groups (CLIPS-IMAG) and results for these two tasks were reported in terms of this common set of shot definitions.

In this paper we give a brief review of the achievements of the TREC2002 video track and for each of the three tasks we describe the task, the data used, a summary of the approaches taken by participants and the results obtained by those participants. Each of these three tasks is addressed in each of the following sections.

2. The Shot Boundary Detection Task

Work on algorithms for automatically recognizing and characterizing shot boundaries has been going on for some time with good results for many sorts of data and especially for abrupt transitions between shots. The shot boundary test collection for

the 2002 TREC task comprised 18 videos totaling 4 hours and 51 minutes. The total size of the collection is 2.88 gigabytes and the videos contained a total of 545,068 frames and 2,090 shot transitions.

Reference data corresponding to the shot transitions was created by a student at NIST whose task was to identify all transitions and assign each to one of the following categories:

- 1466 hard cuts (70.1%) or no transitions, i.e., last frame of one shot followed immediately by the first frame of the next shot, with no fade or other combination;
- 511 dissolves (24.4%) where the shot transition takes place as the first shot fades out while the second shot fades in;
- 63 fades out/in to black and back (3.0%) where the shot transition takes place as the first shot fades out and then the second fades in;
- 50 other (2.4%) i.e. everything not in the previous categories e.g., diagonal wipes.

Gradual transitions are generally harder to recognize than abrupt ones and the proportion of gradual transitions to hard cuts in this collection is about twice that reported in [2] or [3]. Participating groups in this task were allowed up to 10 submissions each and these were compared automatically to the shot boundary reference data created manually at NIST. Detection performance for cuts and for gradual transitions was measured by precision and recall and results are shown in Figures 1 and 2. Some systems (e.g., CLIPS) demonstrate good control of the precision/recall tradeoff.

As illustrated, performance on gradual transitions lags, as expected, behind that on abrupt transitions and the numbers in parentheses give the number of runs submitted by each group. It can be seen that some groups (e.g., CLIPS and RMIT) seem to have good control of precision-recall tradeoff. Almost all of the groups who took part in this task used some form of frame-frame comparison but different groups varied how frames to be compared were selected. Further details of the approaches taken can be seen in the papers submitted by each group to the TREC proceedings [4]. The shot boundary detection task is included in the evaluation in part as an introductory problem, the output of which is needed for higher-level tasks such as search. Groups can participate for the first time in this task, develop their infrastructure, and move on to more complicated tasks the next year.

3. The Feature Extraction Task

The automatic extraction of high-level features from video is itself an interesting task but if it serves to help in video navigation and searching then its importance increases. The objective of the feature extraction task was to begin work on benchmarking feature extraction effectiveness and to allow the exchange of feature detection output among participants. The task is as follows: given a small dataset of just over 5 hours of video (1,848 shots) with common shot bounds, locate up to 1,000 shots which contain each of 10 binary features where the frequency of these features varied from

rare (e.g. monologue) to everywhere (e.g. speech or instrumental sound) in the dataset. The feature set chosen was suggested in on-line discussions by track participants and was deliberately kept small so as to be manageable in this first iteration of the task.

The ten features and their definitions are:

- **Outdoors**, a shot containing a recognizably outdoor location;
- **Indoors**, a shot containing an indoor location;
- **Face,** a shot containing at least one human face with nose, mouth and both eyes visible;
- **People,** shot containing a group of two or more humans, each at least partially visible;
- **Cityscape,** shot containing a recognizably city or urban or suburban setting;
- **Landscape,** shot containing a natural inland setting;
- **Text Overlay,** shot with superimposed text, large enough to read;
- **Speech,** shot with human voice uttering words;
- **Instrumental Sound,** shot with sound produced by one or more musical instruments, including percussion;
- **Monologue,** shot during which a single person is at least partially visible and speaker for a long time without interruption by another speaker;

All submitted results from all groups who took part in this task were assessed manually to create reference data and performance was measured using precision and recall. In the case of some features (speech, instrumental sound) the number of shots in the dataset containing that feature exceeded the submitted result set (1,000) and this created an artificial upper bound on possible precision scores. In general, the size of test set was small in relation to the size of the result set. Still, almost all systems at or above the median, preformed better than a baseline created by evaluating 100,000 randomly created results for each feature.

The results of group submissions are presented in Figure 3. These results show average precision for each of the 10 features, for each group which achieved the median result or above and these results vary enormously in their dispersion among features, as well as in their mean. In general, though, some of the results are disappointingly poor. For some like instrumental sound and speech, performance is reasonable but for others such as detection of cityscape or landscape, people, indoors, text overlay or monologue the performance is poor. This could be attributed to insufficient effort made by groups, operational errors or the difficulty of the task itself. The most likely explanation is that groups who did take part in this task underestimated the complexity of the task and the resources necessary to achieve good performance.

For the dozen or so groups which did this task, most hand-labeled some part of the training data and used a machine learning approach, such as a support vector machine on either the video or the audio track. It is to be expected that for whichever of these features are run again in a future TREC Video track, performance will be much improved.

4. The Search Task

The third and final task in the 2002 TREC Video track was the search task, and this took two forms. The very difficult task of fully automatic topic-to-query translation was set aside for a future TREC video track and so two more modest forms of searching were supported. In the "manual" search task, a human, expert in the search system interface, was allowed to interpret each topic and to create one optimal query which was run once, and the results submitted to NIST for assessment. In the "interactive" search task, groups were allowed much more flexibility in using real users to formulate queries, search, browse, re-formulate and re-query, etc., for as long as was necessary.

The data to be searched in the search task consisted of 176 videos with 14,524 shots and was chosen because it represented an established archive of publicly available material that one could imagine being searched. The topics in the both of the search tasks were designed as multimedia descriptions of an information need, such as somebody searching a large archive of video might have in the course of collecting material to include in a larger video. Twenty-five topics were created by NIST and each contained a text description of the information need plus some examples in other media such as video clips or images. Of the 25 topics, 22 had video examples (average 2.7), 8 had images (average 1.9) and others had audio. The topics requested either specific or generic people (George Washington or football players), specific or generic things (golden gate bridge or sailboats), locations (overhead views of cities), activities (rocket taking off) or combinations (people spending leisure time at the beach).

The task in both search tasks was to return up to 100 shots from the collection (of over 14,000 shots) which might be relevant to the topic, using pre-defined and agreed shot boundaries. To help groups develop more sophisticated video retrieval systems, several groups (CLIPS, DCU, IBM, MediaMill, LIMSI and MS Research Asia) ran their detectors or speech recognition systems on this search data set and made their outputs available to other groups, marked up in MPEG-7. This contributed enormously to making the track a more integrated and cooperative effort.

Results submitted from each group had their top 50 shots pooled and then manually judged for relevance by assessors from NIST;subsequent judgment of the remaining shots in each result set found few additional relevant shots except for topics which already had many.. As with other TREC tasks, once the assessments had been made and this reference data available, evaluation of the performance in terms of precision and recall was possible. Results in terms of mean average precision for the top ten manual runs are presented in Figure 4 (for manual runs) and in Figure 5 (for interactive search runs. Beneath the averages across the 25 topics there was a large amount of variability by topic. Groups who submitted runs in the interactive search task also logged the time spent by their users on each topic and the mean average precision versus mean elapsed time spent searching, showed no correlation between search time, and performance. Time spent in interactive searching varied from an average of about 1 minute per topic for one group, up to almost 30 minutes per topic for another group.

The performance of interactive searching is, as expected, better on average than the performance of manual searching. In absolute terms, the performance of the search systems is quite good but could, of course, be improved. For a first real iteration of the search task on a sizeable data set, some groups have performed quite well and the spread of performances across different groups is quite good.

Of the dozen or so groups who took part in the search task, a true kaleidoscope of approaches was represented. Some groups used interactive video browsing systems with sets of real users carrying out searches under controlled environments; many groups used the automatic speech recognised transcript as a fundamental part of their search system; one group used an image retrieval system on video keyframes, in an interactive framework. Further details of the approaches taken can be seen in the papers submitted by each group to the TREC proceedings [4]. The jury is still out on two important search issues. The reliable usefulness of features in search generally or in specific situations has yet to be demonstrated. Similarly, the proper role and usefulness of non-text elements in the topic is not yet clear. Matching the text of the topic against text derived automatically from the video's audio track usually delivered better overall results than searches based on just the visual elements or a combination of the text and the visual elements in a topic. But for topics requesting a particular camera motion (e.g. locomotive approaching the viewer) text from ASR would be unlikely to help. It is too early to draw convincing conclusions about these two issues.

5. Conclusions

Evaluation of the effectiveness of different approaches to information retrieval from digital video archives is something that is rapidly becoming or crucial importance as the more and more techniques are developed and are being tested. It is crucial to have common testbeds and evaluation metrics in order to allow comparisons across systems and this is the primary motivation behind the TREC Video track. A similar approach to evaluation of image retrieval and image browsing can be found in the Viper project at the University of Geneva [5].

While there may be some disappointments associated with the TREC Video Track activity in 2002 in terms of the overall quality of the results for feature extraction especially, the track has been very successful in demonstrating a collaborative, shared, and effective evaluation of shot boundary detection, feature extraction, and video searching on a common dataset and using common and agreed metrics. This can be regarded as a real achievement by the track and the success of the collaborative aspect of the track in terms of open exchange of derived features in MPEG-7 format has shown that really effective video navigation depends on having a range of features which can be accurately extracted from video and automatic feature extraction will form an important part of future TREC video track activities. In terms of open questions, there are many, such as how the limitations of the dataset influence the conclusions that we can reach and what can be said about the balance between precision and recall ? Details of the types of errors that are being made by different feature classifiers and by different approaches to search will appear in the follow-up analysis of the different groups which will be reported by those groups elsewhere, but

at this early stage of the TREC video track evaluation, it is too early to draw any really meaningful conclusions.

In terms of future activities, the TREC Video track will become an independent 1-2 day workshop (TRECVID 2003) taking place at NIST just before the main TREC conference. It will continue with a larger dataset and more groups participating. For 2003 we will have 120 hours of 1998 news video and more of the same in 2004. It is expected that the three basic tasks of segmentation, feature extraction, and searching, will also continue, probably with some more features and with 50 topics instead of 25 The guidelines for 2003 are currently being developed. Data from the previous video tracks is available to researchers. The latest information about the TREC video retrieval evaluation effort is available from the website: http://www-nlpir.nist.gov/projects/trecvid/.

Authors' note: An extended version of this paper containing a more detailed analysis of the results and brief descriptions of the approaches taken by the participating groups, appeared in the proceedings of the TREC 2002 Conference.

References

1. The Internet Archive Movie Archive. http://www.archive.org/movies
2. Boreczky, J.S., and Rowe, L.A. (1996) Comparison of video shot boundary detection techniques. In I.K. Sethe and R.C. Jain (Eds.) *Storage and Retrieval for Still Image and Video databases IV*, *Proc. SPIE 2670*, pp.170-179., San Jose, Calif. USA.
3. Ford, R.M. (1999). A quantitative Comparison of Shot Boundary Detection Metrics. In: M.M. Yueng, B.-L. Yeo and C.A. Bouman (Eds.) .) *Storage and Retrieval for Still Image and Video databases IV*, *Proc. SPIE 3656*, pp.666-676, San Jose, Calif. USA.
4. The TREC 2002 Proceedings: http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
5. Müller, W., Marchand-Maillet, S., Müller, H. and Pun, T. Towards a fair benchmark for image browsers, In *SPIE Photonics East, Voice, Video, and Data Communications*, Boston, MA, USA, November 5-8 2000.

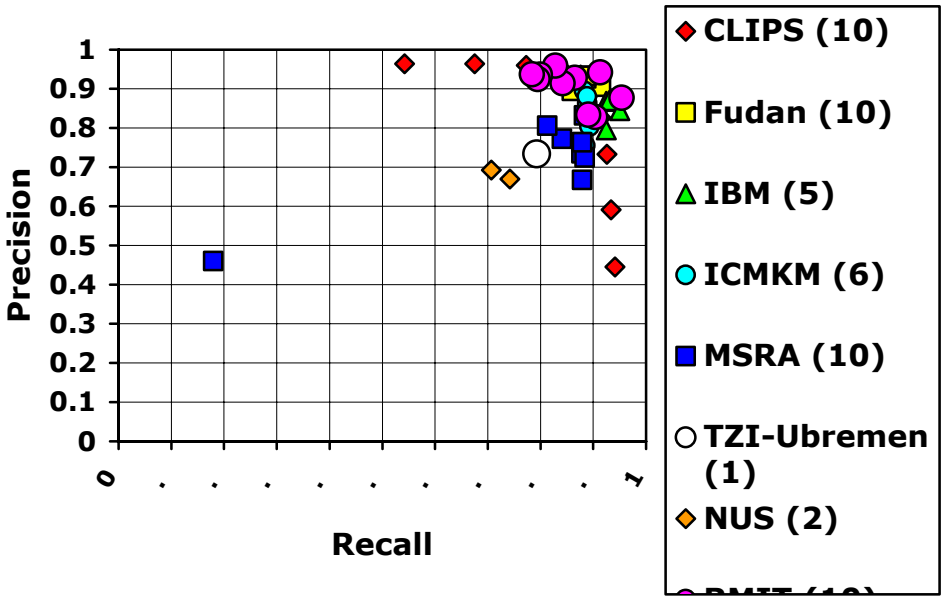

Figure 1: Precision and Recall for Hard Cuts

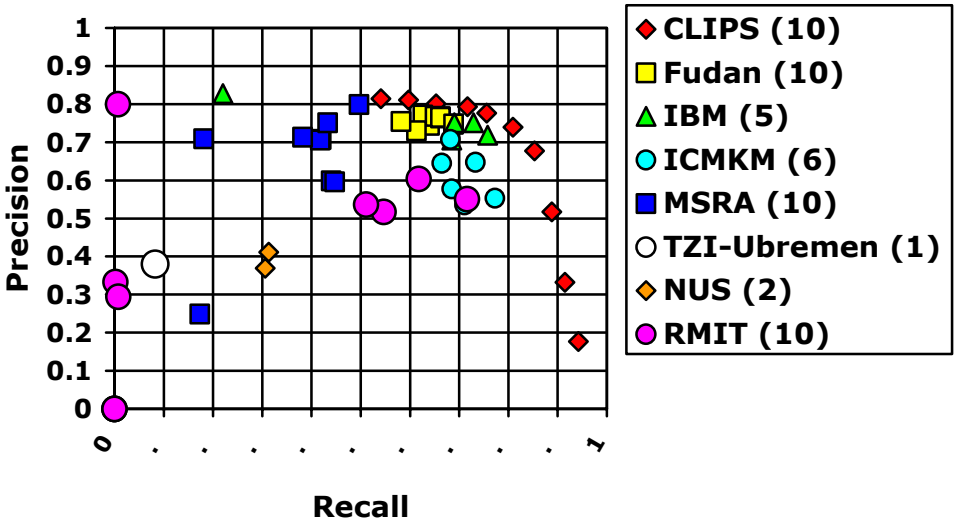

Figure 2: Precision and Recall for Gradual Transitions

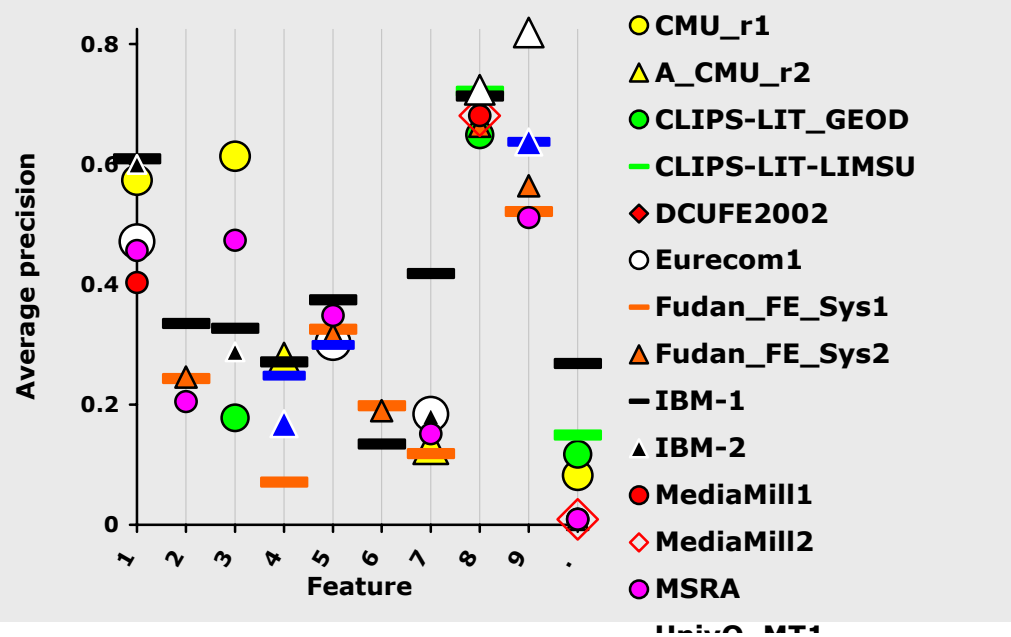

Figure 3: Average Precision by Feature for All Runs at the Median or Above

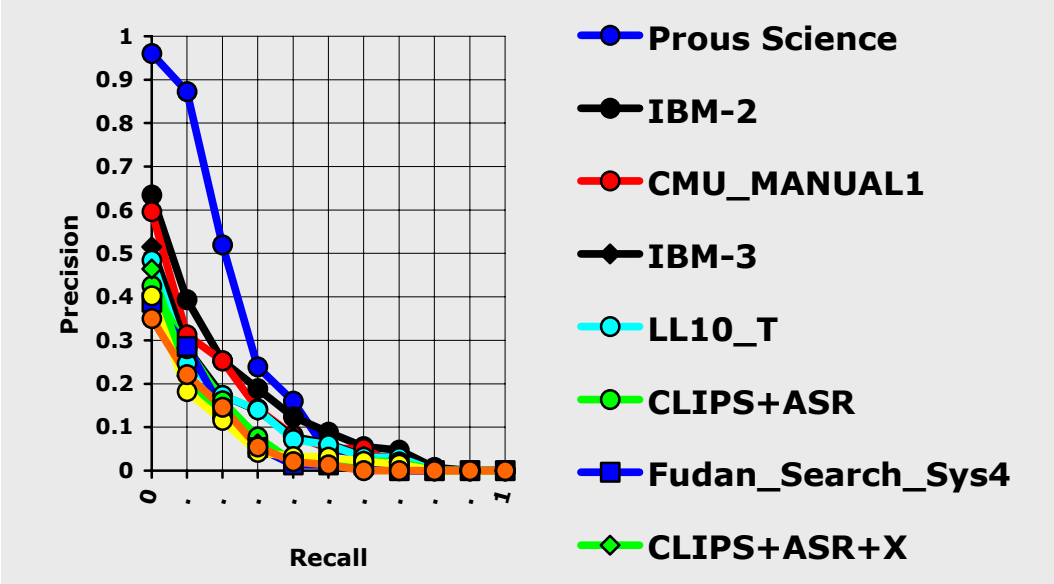

Figure 4: Mean Average Precision for top ten runs for the manual search task

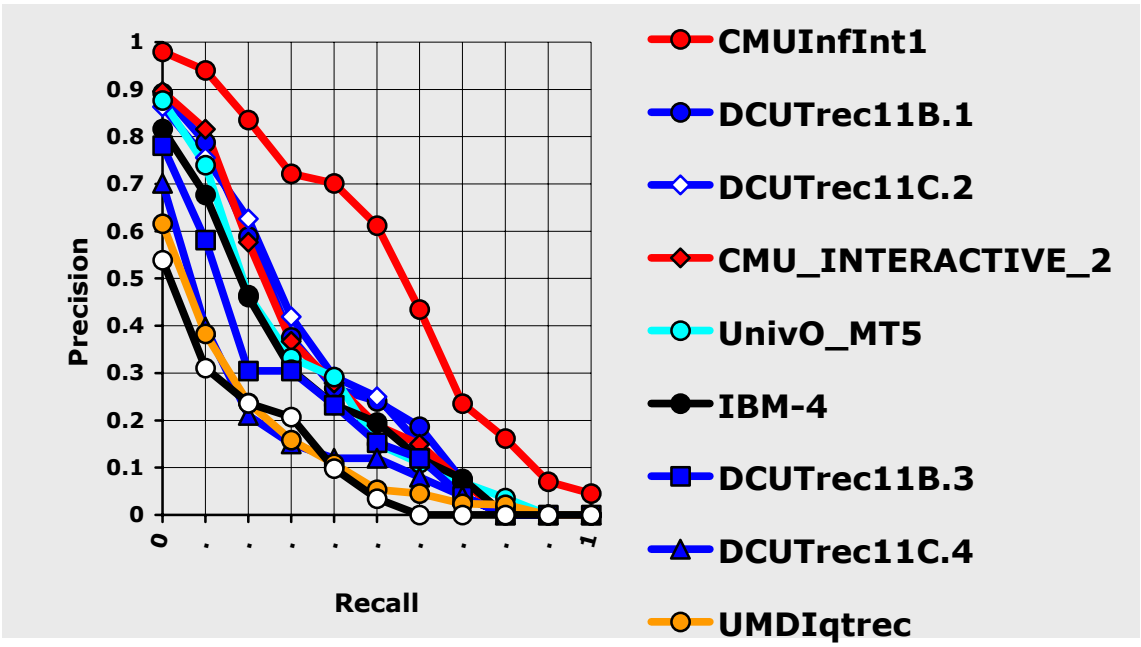

Figure 5: Mean Average Precision for top ten runs for the interactive search task