

TRECVID: The Utility of a Content-based Video Retrieval Evaluation

Alexander G. Hauptmann

Dept. of Computer Science and Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA, 15213-3890

ABSTRACT

TRECVID, an annual retrieval evaluation benchmark organized by NIST, encourages research in information retrieval from digital video. 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 semantic features, and the automatic segmentation of TV news broadcasts. Evaluations done in the context of the TRECVID benchmarks show that generally, speech transcripts and annotations provide the single most important clue for successful retrieval. However, automatically finding the individual images is still a tremendous and unsolved challenge. The evaluations repeatedly found that none of the multimedia analysis and retrieval techniques provide a significant benefit over retrieval using only textual information such as from automatic speech recognition transcripts or closed captions. In interactive systems, we do find significant differences among the top systems, indicating that interfaces can make a huge difference for effective video/image search. For interactive tasks efficient interfaces require few key clicks, but display large numbers of images for visual inspection by the user. The text search finds the right context region in the video in general, but to select specific relevant images we need good interfaces to easily browse the storyboard pictures. In general, TRECVID has motivated the video retrieval community to be honest about what we don't know how to do well (sometimes through painful failures), and has focused us to work on the actual task of video retrieval, as opposed to flashy demos based on technological capabilities.

Keywords: Video analysis performance evaluation, measurement methodology, retrieval benchmarks, content-based video search.

1. Introduction

This paper reviews several years of evaluations of video retrieval. The task involves the search and retrieval of shots from MPEG digitized video recordings using a combination of automatic speech, image and video analysis and information retrieval technologies. These video evaluations have provided an infrastructure for the development and evaluation of video analysis technology and a common forum for the exchange of knowledge between the video analysis and information retrieval research communities. The video track provides an objective testbed where this technology is applied to realistic video collections and objectively evaluated. The evaluations are grouped into interactive (with a human in the loop) and non-interactive (where the human merely enters the query into the system) submissions. Interactive approaches have substantially outperformed all non-interactive approaches, with most systems relying heavily on the user's ability to refine queries and reject spurious answers.

More people use video and television as information sources, yet video retrieval still lags far behind text retrieval in terms of effectiveness. In recent years, video retrieval research has been a very active field, and many different approaches to video retrieval have been proposed both in the non-interactive setting [1–3] and the interactive setting [3–5]. The National Institute of Standards and Technology has sponsored the Text REtrieval Conference (TREC) since 1992 as a means of encouraging research in information retrieval from large test collections. In 2001, the TREC Video Track began with the goal to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. The corpora have ranged from documentaries to advertising films to broadcast news, with international participation growing from 12 to 24 companies and academic institutions from 2001 to 2003, now referred to as “TRECVID” [9]. A number of tasks ranging from shot detection to story segmentation to semantic feature extraction to information retrieval are defined in the TRECVID forum.

The search tasks in the video track are extensions of their text analogues from previous TREC evaluations on pure text documents. Participating groups are asked to index a test collection of video data and to return lists of shots from the videos in the test collection, which meet the information need for a set of topics. The search topics are designed as multimedia descriptions of an information need, which a person searching an archive of video might have in the process of collecting material to create a new video product or to answer questions. Thus topics in the TREC video track contain not only text, but possibly examples (including video, audio, and images) which represent the searcher's information need. Comparable to documents, shots are the boundaries for the units of video to be retrieved. A shot is defined as a single continuous camera operation without an editor's cut, fade or dissolve. [15, 9].

TRECVID brings together an international community of researchers in metrics-based evaluations on a public corpus, and shared metadata. A recent report on future directions in multimedia research notes that "repeatable experiments using published benchmarks are required for the field to progress" [11]. TRECVID is such a benchmark, allowing for repeatable experiments using published data and evaluation metrics.

In the absence of this benchmark, individual independent research can implicitly or explicitly have greatly reduced significance and impact. Consider the topic of finding shots of players throwing a baseball on the baseball field. With a small corpus, the color green in the right locations is a strong predictor of success, and all TRECVID participants may well score highly for this topic, reporting a strong color cue as the reason for the success. As the corpus grows, e.g., to other broadcast news stories, then the color green by itself also brings in many irrelevant shots, which can be rejected if a "sports shot" detector is also in place, so success now requires color plus sports. As the corpus grows to also include soccer shots, then additional cleverness is required to retrieve only the baseball shots. Shared discussions amongst the TRECVID participants help modify easy or misleading topics to ones with greater significance and impact.

Another key advantage of TRECVID is the elimination of a great deal of mundane experimental setup work, including digitizing video, running processing like automatic speech recognition against the video, and creating a common shot reference. While groups are free to run additional processing, much of the work that would be done redundantly across the groups is done once for the benefit of all. For 2001 through 2004, TRECVID has supplied the corpus for testing as MPEG-1 videos, so that no individual research group needs to expend the effort to digitize materials. The speech recognition group at LIMSI has consistently provided an automatic speech recognition (ASR) transcript for TRECVID participants [7]. Since 2002, NIST has been distributing a common shot reference provided by CLIPS-IMAG [16] so that answers for information retrieval topics can be expressed as ordered lists of shots from the reference for easier subsequent evaluation.

Outside of TRECVID, an individual researcher may eliminate portions of the corpus or retrieval topics that do not work well with the techniques under investigation. For example, a "George W. Bush" detector might be reported at 90% recall and precision, but only if the researcher reports results that consider only a few hours of video where Bush is always appearing outside with a red tie. On the same videos, perhaps the "Colin Powell" detector didn't work and it goes unreported. TRECVID as a benchmark protects against such selective reporting, because each year the search results are based on performance across 24-25 topics where each topic must be addressed, and the corpus is fixed and publicly available: if a group ignores a portion of the corpus because they cannot deal with it well, their search results will suffer as they will miss the relevant shots within that portion.

Along with saving labor on setting up experiments, TRECVID allows participants to pool efforts in conducting and evaluating experiments. NIST as the evaluator for the TRECVID topics uses the submitted answer sets as the means to approximate a truth set for retrieval topics. Rather than manually grade each of the 32,318 shots in the common shot reference for TRECVID 2003 across each of the topics, instead NIST makes use of a pooling procedure to approximate the truth as is done in other TREC forums. NIST grades the top-ranked shots from the submitted runs according to this pooling procedure: all shots are taken down to some fixed depth (in ranked order) from the submissions for a given topic, merging the resulting lists and creating a list of unique shots. These are judged manually to some depth based on available assessor time and number of correct shots found. NIST then evaluates each submission to its full depth based on the results of assessing the merged subsets.

One final benefit of TRECVID with respect to video search and retrieval is that the task is defined as part of the benchmark to allow for repeated experiments with comparisons of relative differences between groups and approaches. The individual researcher is saved from struggling to derive tasks representative of real users without bias from the researcher's background, so that the tasks don't become tailor-fitted to one particular strategy.

TRECVID offers a number of advantages for conducting research into information retrieval from video. It brings together an international community of researchers to address the area based on a public corpus, shared metadata, and

metrics-based evaluations. A recent report on future directions in multimedia research notes that “repeatable experiments using published benchmarks are required for the field to progress” [ROWE]. TRECVID is such a benchmark, allowing for repeatable experiments using published data and evaluation metrics.

In the absence of this benchmark, individual independent research can implicitly or explicitly have greatly reduced significance and impact. An experiment can be based on too little data, or data with obscure characteristics, e.g., an ideal “eagle” detector with demonstrated 90% recall and precision against a set of photos. When run on broader data, the results from such individualized research may plummet to 40% or worse, so that in our example the “eagle” detector turns out to be a “blue sky in this particular area of the image” detector. TRECVID corpora have been growing with each annual run, and should obscure features still turn out to be the key components to successful topic retrieval from the larger corpus, the results will show that. Consider the topic of finding shots of players throwing a baseball on the baseball field. With a small corpus, the color green in the right locations is a strong predictor of success, and all TRECVID participants may well score highly for this topic, reporting a strong color cue as the reason for the success. As the corpus grows, e.g., to other broadcast news stories, then the color green by itself also brings in many irrelevant shots, which can be rejected if a “sports shot” detector is also in place, so success now requires color plus sports. As the corpus grows to also include soccer shots, then additional cleverness is required to retrieve only the baseball shots. Shared discussions amongst the TRECVID participants help migrate easy topics or those with misleading solution characteristics to ones with greater significance and impact.

Another key advantage of TRECVID is the elimination of a great deal of mundane experimental setup work, including digitizing video, running processing like automatic speech recognition against the video, and creating a common shot reference. While groups are free to run additional processing, much of the work that would be done redundantly across the groups is done once for the benefit of all. For 2001 through 2004, TRECVID has supplied the corpus for testing as MPEG-1 videos, so that no individual research group needs to expend the effort to digitize materials. LIMSI has consistently provided an ASR transcript for TRECVID participants [GAUVAIN]. NIST has made use of a common shot reference provided by CLIPS-IMAG so that answers for information retrieval topics can be expressed as ordered lists of shots from the reference for easier subsequent evaluation. With a common baseline, results across research groups can be more easily interpreted, e.g., a search run is required for TRECVID 2004 using just the text of the topics and the LIMSI ASR, to better gauge the contributions of visual and other features beyond just ASR text.

Outside of TRECVID, an individual researcher may eliminate portions of the corpus or retrieval topics that do not work well with the techniques under investigation. For example, a “George W. Bush” detector might be reported at 90% recall and precision, but only if the researcher reports results for a few hours of video where Bush is always appearing outside with a red tie. On the same videos, perhaps the “Colin Powell” detector didn’t work and it goes unreported. TRECVID as a benchmark protects against such selective reporting, because each year the search results are based on performance across 24-25 topics where each topic must be addressed, and the corpus is fixed and publicly available: if a group ignores a portion of the corpus because they cannot deal with it well, their search results will suffer as they will miss the relevant shots within that portion.

Along with saving labor on setting up experiments, TRECVID allows participants to pool efforts in conducting and evaluating experiments. With IBM’s lead, TRECVID participants in 2003 collaborated on manually annotating 60 hours of ABC and CNN broadcast news shots, with features such as beach, bridge, and baseball [LIN]. This manual shot assessment was then used by many TRECVID participants as truth data for machine learning algorithms. NIST as the evaluator for the TRECVID topics uses the submitted answer sets as the means to approximate a truth set for retrieval topics. Rather than manually grade each of the 32,318 shots in the common shot reference for TRECVID 2003 across each of the topics, instead NIST makes use of a pooling procedure to approximate the truth as is done in other TREC forums. NIST grades the top-ranked shots from the submitted runs according to this pooling procedure: all shots are taken down to some fixed depth (in ranked order) from the submissions for a given topic, merging the resulting lists and creating a list of unique shots. These are judged manually to some depth based on available assessor time and number of correct shots found. NIST then evaluates each submission to its full depth based on the results of assessing the merged subsets.

One final benefit of TRECVID with respect to video search and retrieval is that the task is defined as part of the benchmark to allow for repeated experiments with comparisons of relative differences between groups and approaches. The individual researcher is saved from struggling to derive tasks representative of real users without bias from the researcher’s background, so that the tasks don’t become tailor-fitted to one particular strategy. The TRECVID search task is defined as follows: given a multimedia statement of information need (topic) and the common shot reference, return a ranked list of up to N shots from the reference which best satisfy the need, with the query formulator (non-interactive) or interactive user having no prior knowledge of the search test collection or topics: N=100 for 2002,

N=1000 for 2003 and 2004. The topics are defined by NIST to reflect many of the sorts of queries real users pose, based on query logs against video corpora like the BBC Archives and other empirical data [ENSER, NIST]. The topics include requests for specific items or people, specific facts, instances of categories, and instances of activities. Mean average precision is used to compare the relative merits of the retrieval systems.

1.1. The TRECVID Evaluation Corpora

TREC Video Retrieval Evaluations (TRECVID) is an independent evaluation forum devoted to research in content-based retrieval of digital video [6]. Its goal is to encourage research in information retrieval from large amounts of videos by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. TRECVID focuses on the single shot as the unit of information retrieval rather than the scene or story/segment/movie.

TRECVID began in 2001 with an 11 hour test corpus of primarily documentary video. Participants agreed that a larger corpus and standard way of submitting and evaluating answers via a common shot reference would improve evaluation and significance, and so for 2002, NIST defined 25 topics to find within a search test collection of 40.12 hours of video from the Prelinger Archives [30] and the Open Video archives [31]. The material consisted of advertising, educational, industrial, and amateur films produced between the 1910s and 1970s, spanning a wide spectrum of quality and attributes, e.g., silent films and animated cartoons were part of the corpus. The search test collection was delineated into 14,524 shots, which became the common shot reference.

The TRECVID test corpora for 2003 and 2004 consisted of broadcast news from ABC, CNN, and C-SPAN (for 2003), with 32,318 reference shots in the test video corpus for 2003 and 33,367 reference shots in 2004. The nontrivial size of the corpus, its definitions of sets of information needs (topics), and human-determined truth for the features and topics provide a starting point for scientifically valid evaluations and comparisons. For 2003, NIST defined 25 search topics and made use of a common shot reference: 32,318 shots representing 55.91 hours of video: ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998, along with six hours of C-SPAN programming from mostly 2001. TRECVID 2002 worked with older video materials with digitization artifacts, noise introduced through aging or inadequate preservation of the source video, and other noise representative of a video archive spanning the last century's first 75 years. News was used in TRECVID 2003, with a new set of 74 broadcast news hours used for TRECVID 2004 so that investments in annotating and truthing the news genre could be reused and iteratively improved.

The 2005 task TRECVID collection is

The TRECVID search task is defined as follows: given a multimedia statement of information need (topic) and the common shot reference, return a ranked list of up to N shots from the reference which best satisfy the need, with the query formulator (non-interactive) or interactive user having no prior knowledge of the search test collection or topics: N=100 for 2002, N=1000 for 2003 and 2004. The topics are defined by NIST to reflect many of the sorts of queries real users pose, based on query logs against video corpora like the BBC Archives and other empirical data [6, 9]. The topics include requests for specific items or people, specific facts, instances of categories, and instances of activities. Mean average precision is used to compare the relative merits of the retrieval systems.

[define MAP] Finally, we should warn against reading too much into small changes in mean average precision. For a given query, there are a total of N_r items in the collection that are relevant to this query. Assume that the system only retrieves k relevant items and they are ranked as r_1, r_2, \dots, r_k . The mean average precision is computed as follows:

$$MAP = \left\{ \sum_{i=1}^k i / r_i \right\} / N_r$$

If you assume an evaluation with 25 queries, where one query happens to have only two relevant shots. If system A only finds result r_1 at rank 500 for this query, and system B, which is otherwise identical to system A, through a random shuffle of the results, happens to find that same result r_1 at rank 10, this single shift in one relevance item could affect the MAP by .001. Thus reporting MAP improvements in the third decimal can be very misleading, since any improvements might be due to only one result in one query, which could very well be due to chance.

2. Evaluation of what is an effective technique

What makes a difference in video retrieval? We examine the question by looking at the TRECVID 2003 and TRECVID 2004 official results, which evaluated video retrieval systems in a variety of conditions and on a variety of topics. Our initial goal was to examine the rankings of the evaluated systems, analyze the descriptions and establish which approaches were more effective than others. However, a closer look at the differences in evaluation scores between the systems, quickly reveals that many evaluation score differences are very small, hinting that any conclusions drawn from the score difference alone may be inconclusive. We began to investigate which differences were statistically significant when comparing results in the TRECVID evaluation. Much to our surprise, the significant differences were relatively sparse.

2.1. REER

As a first approach, we examined the TRECVID evaluation results in the retrieval experiment error rate, as suggested by Voorhees and Buckley [8]. Retrieval experiment error rate (REER) is motivated by a desire to evaluate the reliability of evaluation results in a retrieval experiment like TRECVID. REER is defined as the probability of making opposite effectiveness judgments about two systems over two sets of topics based on a common evaluation metric, like Mean Average Precision (MAP). If we make an effectiveness statement about two retrieval systems (or two submission runs in the TRECVID setting) based on one system's higher MAP over a set of topics, then REER is the likelihood that the effectiveness judgment will be reversed, i.e. an experiment error, if we compare the two systems on another set of topics. Intuitively, if two video retrieval systems is equally effective, we would expect to observe that one system performs better than the other only half of them time, i.e. REER is 0.5, if we compare two systems over different sets of topics. Therefore, only when REER is much lower than 0.5 can we have much confidence that one retrieval system is significantly more effective than the other. By calculating the REER of retrieval experiments, we can obtain better insights into the reliability of the score difference between systems instead of simply assuming that System X is better than System Y merely because the MAP of System X is minimally larger. REER has been shown empirically [8] and theoretically [9] to be reversely correlated with the score difference. In other words, the larger the evaluation metric difference between two systems, the lower the REER, and thus we can conclude that one system is more effective than the other.

REER can be estimated in the frequentist manner [8] by counting how often retrieval experiment errors occur, but this method requires large amounts of retrieval experiment results, and an extra curve fitting step is needed in order to extrapolate a prediction of the REER for larger topic set sizes. Instead, we estimate REER by directly following the theoretical analysis in [9] and estimate REER in the following equation,

$$\text{REER} = 2\Phi\left(\frac{-(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{|T|}}}\right) \left(1 - \Phi\left(\frac{-(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{|T|}}}\right)\right)$$

where μ_X , μ_Y and σ_X^2 , σ_Y^2 are the means and variances of MAP probability distribution of the two systems X and Y, respectively, T is the set of topics in the retrieval experiment, Φ is the standard normal distribution function.

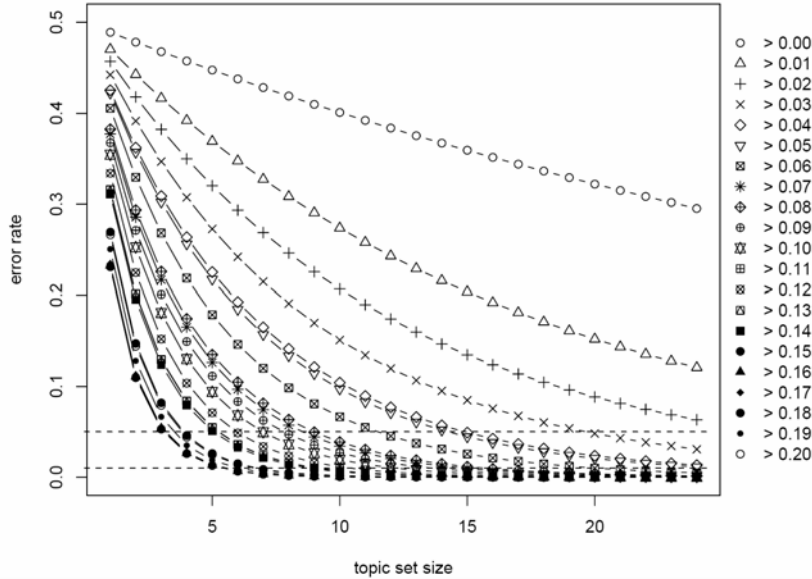


Fig. 1. Retrieval Experiment Error Rate (REER) curves estimated from all search submission runs in TRECVID 2003 and 2004. Each REER curve represents different MAP difference as a function of the topic set size. The topmost curve stands for the MAP difference greater than 0 but less than 0.01, and the second curve stands for the difference greater than 0.01 but less than 0.02, and so on. Two horizontal dashed lines are drawn at the REER levels of 0.01 and 0.05, respectively.

2.2. ANOVA

As an alternative method, we apply an Analysis of Variance (ANOVA) approach to determine how well TRECVID evaluation results can be explained by topics and systems. Instead of applying multiple t-tests and suffering from the multiple testing problems, where random differences appear significant if enough experiments are performed, the Newman-Keuls test is used to estimate is the pairwise MAP difference between two systems is statistically significant. We use a standard Analysis of Variance (ANOVA) repeated measurements design [10] to analyze the data for statistical significant differences. ANOVA models the average precision scores $Y_{i,j}$ of System j for Query Topic i as a combination of effects in the following formula,

$$Y_{i,j} = M + t_i + r_j + e_{i,j} \quad (2)$$

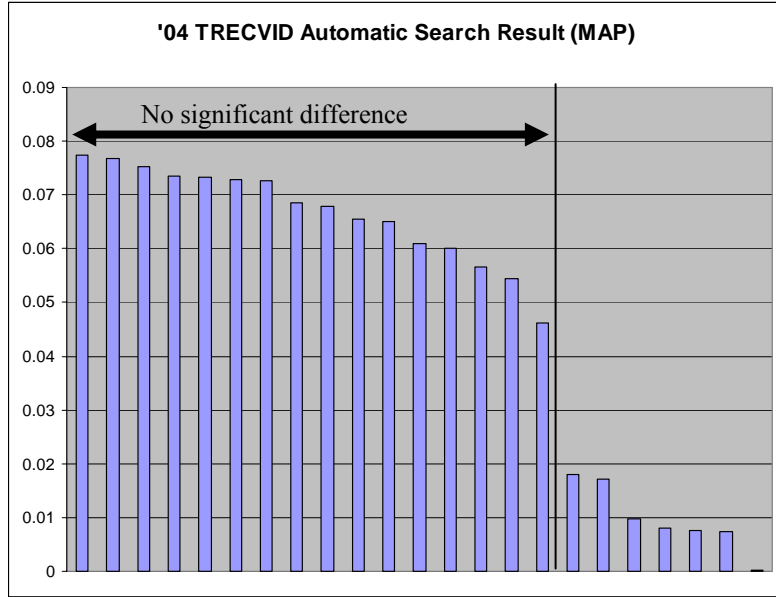
where M is the global mean for all subjects and topics, t_i is Topic i mean for all systems, r_j is System j mean average precision for all topics, and $e_{i,j}$ is the error term, which “explains” the rest of the $Y_{i,j}$ score as due to random measurement noise.

ANOVA then allows us to compute the probability that this model can explain the data. The resulting confidence probability p asserts the rejection of the null hypothesis, i.e. that all data comes from the same distribution according to the model.

2.2.1. The Newman-Keuls Test of Pairwise Significance

For TRECVID data, we generally find that there is a significant effect due to topic and system differences overall, but we also want to find out which pairs of differences are significant, and which are not. The method we used for this is the Neumann-Keuls post-hoc test of pairwise significance. While neither the most conservative or generous test, Newman-Keuls has the advantage that it takes the number of pairwise comparisons into account when computing the significance and adjusts the significance criterion. The reasoning is that if you make many pairwise comparisons on randomly selected data, some will seem to be significant, and the Neumann-Keuls test raises the bar for each additional comparison. This avoids a situation where several hundred t-test are performed at the $p < 0.05$ significance level, and some appear significant due to random sampling effects. Alternative (and in many ways comparable) tests would be Tukey’s test or Scheffe’s test. The Neumann-Keuls test first arranges all means in descending order. According to the

statistic, different cells now have different “critical differences”, depending on the mean square error, the degrees of freedom and a so-called r value. The r value is obtained from the difference in the number of comparison between compared cells.



3. Conclusions

Content based video retrieval is hard, because there is so much variety in video imagery. This heterogeneity prevents a successful image feature similarity matching. Many standard vision techniques also fail, e.g. background subtraction requires a stable, empty background with a stationary camera, which is rarely present in broadcast video. Image only and non-text based searches appear to perform much worse than text transcript searches. Generally, image similarity search works well for images that are almost identical to the query image, but poorly for images with the same content but different composition. Manually selecting good image examples as queries can help, if the collection is well understood. Usually, queries from external images will not provide useful results. [20, 22]. There are still many issues open for debate. Which are the “right” low-level image, motion and audio feature sets to use? Which semantic features are relevant to a query? How can results be combined? The answer to these questions appears to be domain specific, in that it depends on the collection to be searched and the specific query under consideration. The best performing systems in 2002 [21] and 2003 [26,29] heavily exploited the available development set to “learn” the weights for combination of results of each query. But one must wonder if it is realistic that one gets to “practice” a query before it is actually issued.

There are several lessons to be learned from this analysis of the data. The first one is, of course, that one should not believe all the hype surrounding effective techniques in video retrieval. Too often small differences are interpreted as substantial, even though they may just reflect uncertainty in measurement. Both the retrieval experiment error rate metric and ANOVA analysis give a strongly consistent interpretation of the results, and 0.05 MAP difference between two retrievals is the minimal value to have a significant difference. Our data provides consistent evidence, across two years, that there are no clearly distinguished effective techniques for either manual or automatic video retrieval. Perhaps the relatively small number of topics is to blame, compared to the standard text retrieval evaluations, 25 and 23 search topics per year makes it very difficult to ascertain significant differences. If we take the risk of over-generalizing results in Figure 1 and continue the REER curves, we could justify 0.02 MAP difference at the error rate level of 0.01 if we conduct retrieval experiments with 50 topics, but this will pose a significant burden on the TRECVID organizers. What is disappointing about our analysis is that we repeatedly find that none of the multimedia analysis and retrieval techniques provide a significant benefit over retrieval using only textual information such as ASR transcripts or closed captions. This is actually consistent with findings in the earlier TRECVID evaluations in 2001 and 2002, where the best systems were based exclusively on retrieval using automatic speech recognition. However, we should also point out that it is not the case that “nothing works” here. In interactive systems, we do find significant differences among the top systems, indicating that interfaces can make a huge difference for effective video search. Not surprisingly, from

comparisons of our own data, we find that expert users significantly outperform novice users, and visual only systems that do not exploit broadcast news speech transcripts are significantly inferior to systems that exploit all available knowledge. While in 2003, there were big, significant gaps between the top systems, that difference shrunk in the 2004 TRECVID interactive submissions, indicating that the knowledge about effective interactive search systems is more broadly disseminated.

An explicit goal of the TRECVID evaluations is to help chart the progress of automatic feature classifiers like face, people, outdoors, and cityscape, showing that perhaps these classifiers will reach the level of maturity needed for their use as effective filters for video retrieval. The annual results of TRECVID experiments can chart not only the progress of these video shot classifiers, but also indicate under what conditions they hold utility for video retrieval. It is an ongoing research issue as to how to best represent the high level semantics of a video shot given current techniques for automatic lower-level feature extraction. While rich feature spaces can be created to create a correspondence between lower level features and human perception, the resulting high dimensional space is then not well suited for fast, interactive access through indexing [2]. The solution that has worked for TRECVID news and documentary corpora is in leveraging from transcripts of the narrative text to identify a good set of candidate shots. Then, through temporal browsing and relevance feedback, interactively use the relevant shots already located to find additional ones. An ongoing area of research will be to see the level of success achievable in the absence of narrative when visual features and non-speech aural features will remain as the only metadata for information retrieval.

ACKNOWLEDGMENTS

This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037.

REFERENCES

1. Proceedings of the TREC Video Retrieval Evaluation 2004. In: Proceedings of the TREC Video Retrieval Evaluation 2004. (2004)
2. NIST, Digital Video Retrieval at NIST: TREC Video Retrieval Evaluation, 2001-2004, <http://www-nlpir.nist.gov/projects/trecvid/>.
3. NIST: Guidelines for the TRECVID 2004 evaluation. Webpage (2004) <http://www-nlpir.nist.gov/projects/tv2004/tv2004.html>.
4. Voorhees, E.M., Buckley, C.: The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press (2002) 316–323
5. Lin, W.H., Hauptmann, A.: Revisiting the effect of topic set size on retrieval error. SIGIR (2005)
6. Myers, J.L.: Fundamentals of Experimental Design. Allyn and Bacon, Boston, MA (1972)
7. Boldareva, L., de Vries, A., and Hiemstra, D. Monitoring User-System Performance in Interactive Retrieval Tasks. Proc. RIAO 2004 (Avignon, France, April 2004), pp. 474-483.
8. Enser, P.G.B. and Sandom, C.J. Retrieval of Archival Moving Imagery - CBIR Outside the Frame? In Image and Video Retrieval (CIVR 2002 Proceedings), Lecture Notes in Computer Science 2383, Springer-Verlag, Berlin, 206-214.
9. A.F. Smeaton, P. Over, C. Costello, A. P. de Vries, D. S. Doermann, A. G. Hauptmann, M. E. Rorvig, J. R. Smith, and L. Wu: The TREC2001 Video Track: Information Retrieval on Digital Video. ECDL 2002: 266-275, 2002
10. NIST TREC 2002, Results of the Video Track, http://trec.nist.gov/pubs/trec10/appendices/video_results.html
11. The Internet Archive Movie Archive Home Page. (2002) URL: www.archive.org/movies
12. Open Video Digital Library, <http://www.open-video.org/>
13. Smeaton, A. F., Over, P., and Kraaij, W. 2004. TRECVID: evaluating the effectiveness of information retrieval tasks on digital video. In Proceedings of the 12th Annual ACM international Conference on Multimedia (New York, NY, USA, October 10 - 16, 2004). MULTIMEDIA '04. ACM Press, New York, NY, 652-655. DOI=<http://doi.acm.org/10.1145/1027527.1027678>
14. 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.

15. Lin, C.-Y., Tseng, B. L., and Smith, J. R. Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets. In Proc. NIST TRECVID (Gaithersburg, MD, Nov. 2003), <http://www-nlpir.nist.gov/projects/tvpubs/papers/ibm.final.paper.pdf>.
16. Henning Mueller, Stephane Marchand-Maillet, Thierry Pun, The Truth about Corel - Evaluation in Image Retrieval, in Image and Video Retrieval : International Conference Proceedings, CIVR 2002, London, UK, July 18-19, 2002, Lecture Notes in Computer Science, Volume 2383, p.38+, M.S. Lew, N. Sebe, J.P. Eakins (Eds.):