The TREC2001 Video Track: Information Retrieval on Digital Video Information

Alan F. Smeaton[1], Paul Over[2], Cash J. Costello[3], Arjen P. de Vries[4], David Doermann[5], Alexander Hauptmann[6], Mark E. Rorvig[7], John R. Smith[8], and Lide Wu[9]

[1]Centre for Digital Video Processing, Dublin City University, Dublin, 9, Ireland.
[2]National Institute for Standards and Technology, Gaithersburg, Md., USA.
[3] Johns Hopkins University Applied Physics Laboratory, Laurel, Md., USA.
[4]CWI, Amsterdam, The Netherlands.
[5] Laboratory for Language and Media Processing, University of Maryland, College Park, MD. USA
[6]School of Computer Science, Carnegie Mellon University, USA
[7]School of Library Information Sciences, University of North Texas, Tx., USA
[8]IBM T. J. Watson Research Center, Hawthorne, NY, USA.
[9]Dept. of Computer Science, Fudan University, Shanghai, China..

The development of techniques to support content-based access to archives of digital video information has recently started to receive much attention from the research community. During 2001, the annual TREC activity, which has been benchmarking the performance of information retrieval techniques on a range of media for 10 years, included a „track“ or activity which allowed investigation into approaches to support searching through a video library. This paper is not intended to provide a comprehensive picture of the different approaches taken by the TREC2001 video track participants but instead we give an overview of the TREC video search task and a thumbnail sketch of the approaches taken by different groups. The reason for writing this paper is to highlight the message from the TREC video track that there are now a variety of approaches available for searching and browsing through digital video archives, that these approaches do work, are scalable to larger archives and can yield useful retrieval performance for users. This has important implications in making digital libraries of video information attainable.

1. Introduction

The technical challenges associated with generation, storage and transmission of digital video information have received much attention over the last few years and we are now at the stage where we can regard these engineering problems as having made significant progress. This now allows us to create large libraries of digital video information and with that comes the associated challenge of developing effective, efficient and scalable approaches to searching and browsing through video digital libraries.

TREC is an annual activity which has been ongoing for the last decade and which has been benchmarking the retrieval effectiveness of a variety of information retrieval

M. Agosti and C. Thanos (Eds.): ECDL 2002, LNCS 2458, pp. 266-275, 2002.

tasks. This has included retrieval on text documents, documents in a variety of natural languages, spoken audio, web documents, documents corrupted by an OCR process, and so on. In 2001, TREC included a „track" or activity line which explored different approaches to searching through a collection of digital video information. The goal of the TREC2001 video track was to promote progress in content-based retrieval from digital video by using open, metrics-based evaluation and using publicly available video.

The TREC2001 video track had 12 participating groups, 5 from US, 2 from Asia and 5 from Europe and was divided into two distinct tasks namely shot boundary detection and searching. Shot boundary detection is the task of automatically determining the boundaries between different camera shots which is usually used as a fundamental component of video structuring and further details of the shot boundary detection task can be found in [1]. The searching task involved running queries against the video collection and what made the queries particularly interesting and challenging was that they were true multimedia queries as they all had video clips, images, or audio clips as part of the query, in addition to a text description. Participating groups used a variety of techniques to match these multimedia queries against the video dataset, some running fully automated techniques and others involving users in interactive search experiments. 11 hours of MPEG-1 data was collected and distributed as well as 74 topics or queries.

The rest of this paper is organised as follows. In the next section we give an introduction to the search task, covering the video data used, the topics and how they were formed, the evaluation mechanism and the evaluation metrics adopted. In section 3, each of the main groups who participated in the search task give an overview of the approach that they have taken in the search task. Section 4 includes a brief summary and comparison across the approaches as well as including some indicative evaluation results in order to allow the reader to gauge the absolute performance levels of the video retrieval systems. A concluding section assesses the contribution that the TREC2001 video track has made.

2. The TREC2001 Video Track

Like most of the TREC activities, the video track in TREC2001 was coordinated by the National Institute for Standards and Technology (NIST) though participating groups contributed significant amounts of work towards the definition and running of the track. The search tasks in the video track were extensions of their text analogues from previous TRECs. Participating groups were asked to index a test collection of video data and were asked to return lists of shots from the videos in the test collection which met the information need for a set of topics. The boundaries for the units of video to be retrieved were supposed to be shots and were not predefined and each system made its own independent judgment of what frame sequences constituted a relevant shot.

Participants were free to use whatever indexing and retrieval techniques they wished though the search task was divided into two distinct classes, one for interactive retrieval which involved some human in the search loop, and one for automatic retrieval where the retrieved shots were determined completely automatically. This distinction arose because the search task was designed to replicate the situation where a user uses a video information retrieval system to satisfy an information need, sometimes using interactive retrieval, sometimes completely automated. Another feature of the search task, which also reflects its real world nature, is that topics are either „known item“ or „general“. In the case of known item retrieval, the user knows that there is at least one relevant shot in the test collection and the task is to find those shots known to satisfy the information need, while the case of general searching reflects the situation where the user does not know whether or not there are shots in the collection which satisfy the information need.

Although the track decided early on that it should work with more than text recognised from spoken audio, systems were allowed to use transcripts created by automatic speech recognition (ASR) and any group which did this had to submit a run without the ASR or one using only ASR as a baseline. Three groups used ASR.

The test collection for the search task consisted of 85 video programmes representing over 11 hours of video, encoded in MPEG-1 and totalling over 6 Gbytes in size. The content came from the OpenVideo project [2], the NIST organisation itself, and the BBC who provided some stock footage. Further details of the collection can be found on the web pages for the video track [3]. The videos are mostly of a documentary nature but vary in their age, production style, and quality. The only manually created information that search systems were allowed to use was that which was already as part of the test collection, namely the existing transcripts associated with the NIST files and the existing descriptions associated with the BBC material, though most groups did not use this information.

The search topics were designed as multimedia descriptions of an information need, such as someone searching an archive of video might have in the course of collecting material to include in a larger video or to answer questions. While today this may be done largely by searching associated descriptive text created by a human when the video material was added to the archive, the track's scenario envisioned allowing the searcher to use a combination of other media in describing his or her information need. How one might do this naturally and effectively is an open question. Thus topics in the TREC2001 video track contained not only text but possibly examples (including video, audio, images) which represent the searcher's information need. The topics expressed a very wide variety of needs for video clips: of a particular object or class of objects, of an activity/event or class of activities/events, of a particular person, of a kind of landscape, on a particular subject, using a particular camera technique, answering a factual question, etc.

For a number of practical reasons, the topics were created by the participants which is an example of the significant contribution to running the track made by those participants. Each group was asked to formulate several topics they could imagine being used by someone searching a video archive. NIST submitted topics as well, did some selection and pruning, and negotiated revisions. All the topics were pooled and all systems were expected to run on all of these if possible.

All topics contained a text description of the user information need and examples in other media were optional. There were indicators of the appropriate processing (automatic, manual or either) and finally, if the topic was a hunt for one or more known-items, then the list of known-items was included. If examples to illustrate the information need were included then these were to come from outside the test data.

74 topics were produced in this manner and Table 1 gives a summary of the use of example media in those topics.

Table 1. Distribution of other media in topics

Number of topics	74
No. topics with image examples / Avg. number of images	26 / 2.0
No. topics with audio examples / Avg. number of audio	10 / 4.3
No. topics with video examples / Avg. number of videos	51 / 2.4

In the case of the known-item search submissions, these were evaluated by NIST but the evaluation of known item retrieval turned out to be more difficult than anticipated. One reason for this was because each group was able to define the start/stop boundaries of the shots they returned we had to use a parameterised matching procedure between known item and submitted results. Matching a submitted item to a known-item defined with the topic was a function of the length of the known-item, the length of the submitted item, the length of the intersection, and two variables which measured the amount of desired overlap among these. Evaluations were run with different settings of these overlaps. The measures calculated for the evaluation of known-item searching were precision and recall with the ground truth or relevant video clips from the collection being provided by the participants who formulated the topics. The number of known-items across the topics varied from 1 to 60 with a mean of 5.63, so the upper bound on precision in a result set of 100 items was quite low.

Submissions for the general search topics were evaluated by retired information analysts at NIST. They were instructed to familiarize themselves with the topic material and then judge each submitted clip relevant if it contained material which met the need expressed in the topic as they understood it, even if there was non-relevant material present, otherwise they were told to judge the clip as not relevant. They used web-based software developed at NIST to allow them to (re)play the video, audio, and image examples included in the topic as well as the submitted clips. A second set of relevance judgments of the submitted materials was then performed and overall, the two assessors agreed 84.6% of the time. The measure calculated for the evaluation general searching was precision but we have also calculated a partial recall score.

The detailed performance scores from the 8 groups who submitted a total of 21 runs are available online at http://www-nlpir.nist.gov/projects/trecvid/results.html but before we address retrieval performance, the next section will give a thumbnail sketch of the different approaches to video indexing and retrieval taken by the TREC2001 video track participants.

3. Participants in the TREC2001 Video Track Search Task

Of the 12 groups who took part in the TREC2001 video track, most completed the shot boundary detection task and 8 completed the search task and the approaches that each of these groups have taken is described here. Further descriptions on all of the participants work can be found in their papers in the TREC2001 proceedings [4]

3.1 Carnegie Mellon University

The CMU Informedia Digital Video library's standard processing modules were used for the TREC2001 Video evaluations. Among the processing features that were utilized in Video TREC were: shot detection using simple color histogram differences, keyframe extraction, speech recognition using the Sphinx speech recognizer with a 64000 word vocabulary, face detection, video OCR, and image search based on color histogram features in different color spaces and textures.

The Informedia interface was used in the interactive track with only minor modifications, most of which involved user preference settings. For example, users found they wanted to see as many shot results for each query as could fit on the screen, while geographic maps were irrelevant. The main modification was the addition of multiple image search engines, which allowed a user to switch between image retrieval approaches, when nothing relevant could be found using a given image retrieval approach.

For the automatic track, Informedia image retrieval was modified to process I-frames instead of merely keyframes for the image retrieval. We also added a speaker identification component, which determined whether a given segment of audio might have originated from the same speaker as the query audio. Post-mortem analysis of the results showed that image retrieval and video OCR had the largest impact on performance.

3.2 Dublin City University (Ireland)

The group from Dublin City University explored interactive search and retrieval from digital video by employing more than 30 users to perform the search tasks under controlled, timed conditions. In the Físchlár system developed at DCU [5], several keyframe browser interfaces have been developed and the task DCU performed was to evaluate the relative effectiveness of three different keyframe browsers. One of these keyframe browsers was based on a timeline of groups of related keyframes, a second browser interface simply played the keyframes on screen as a kind of slideshow, and the final browser interface was a 4-level hierarchical browser which allowed dynamic navigation through the keyframe sets. In the DCU experiments, 30 users (either final year undergraduates or research students) were employed to spend between 5 and 10 minutes on each topic, and each volunteer did interactive searching on 12 topics using one of the 3 different browsers per topic in round robin fashion. This gave the DCU group the opportunity to compare the relative performances of the three keyframe browser interfaces.

3.3 Fudan University (China)

The group from Fudan University tried 17 topics, including people searching, video text searching, camera motion etc. In order to do the search they also developed several feature extracting modules. These are qualitative camera motion analysis module, face detection and recognition module, video text detection and recognition module, and a speaker recognition and speaker clustering module. In addition they also used the speech SDK from Microsoft to get transcripts. Based on the above feature extraction modules, the Fudan retrieval system consists of two parts. One is the off-line indexing sub-system and the other is on-line searching sub-system.

For the face detection and recognition modules, face detection consists of skin-color based segmentation, and motion and shape filtering; face recognition uses a new optimal discrimination criterion to get features for recognition [6]. For the video text detection and recognition module, the group used vertical edge based methods to detect text blocks and an improved logical level technique to binarize the text blocks. The recognition was done by commercial software after binarization..

3.4 IBM Research[1]

The IBM Research team developed a system for automatic and interactive content-based retrieval of video using visual features and statistical models. The system used IBM CueVideo for computing automatic shot boundary detection results and selecting key-frames. The system indexed the key-frames of the video shots using MPEG-7 visual descriptors based on color histograms, color composition, texture and edge histograms. The MPEG-7 visual descriptors were used for answering automatic searches using content-based retrieval techniques. The system also used statistical models for classifying events (fire, smoke, launch), scenes (greenery, land, outdoors, rock, sand, sky, water), and objects (airplane, boat, rocket, vehicle, faces). The classifiers were used to generate labels and corresponding confidence scores for each shot. The features and models were then used together for answering interactive searches where the user constructed query/filter pipelines that cascaded content-based and model-based searches. This allowed integration of multiple searches using different methods for each topic, for example, to retrieve „shots that have similar color to this image, have label 'outdoors' and show a 'boat.'"

The IBM team also developed a system based on automatic speech recognition (ASR) and text indexing. The speech-based system was used as a baseline for the content-based/model-based system. The overall results showed that the content-based/model-based system performed relatively well compared to the speech-based system and to other systems. In some cases the speech-based system provided better results, for example, to retrieve „clips that deal with floods." In other cases, the content-based/model-based system provided better results, for example, to retrieve „shots showing grasslands." In two cases, the best result was obtained by combining speech-based and content-based/model-based methods, for example, to retrieve „clips of Perseus high altitude plane." The results show promise in particular for the approach based on statistical modeling for video content classification. The overall results show that significant improvements are still needed in retrieval effectiveness in general to develop usable systems. The NIST video retrieval benchmark is helping to accelerate the necessary technology development.

[1] The IBM Research Team consisted of members from IBM T. J. Watson Research Center and IBM Almaden Research Center.

3.5 Johns Hopkins University

The JHU/APL research group developed an automatic retrieval system for the TREC2001 video track that relied on the image content of the digital video frames. Each keyframe in the video collection was indexed by its color histogram and image texture features. The texture measures were calculated using a descriptor proposed by Manjunath [7]. Ignoring audio clips or text descriptions, the query representation consisted of the image and video portions of the information need. A weighted distance between the image features of the query representation and the keyframes in the index served as a similarity measure. The shots that were retrieved for a particular query minimized this distance measure.

3.6 Lowlands Group (Netherlands)

A 'joint venture' between research institutes and universities in the Netherlands approached the challenge offered by the Video Track as the 'Lowlands Team'[2]. The group submitted pure automatic as well as 'interactive' runs, investigating the influence of human interaction on retrieval results.

The visual automatic system heuristically selected a set of filters based on specialized detectors, by analyzing the query text with WordNet; e.g., the face detector is associated to categories 'person, human, individual'. The retrieval system included a face detector, a camera motion detector (pan, tilt, zoom), a monologue detector, and a detector for text found in the keyframes using OCR. The filtered results are ranked with query example images or keyframes from example videos. A transcript-based automatic system used speech transcripts provided by CMU in a retrieval model based on language models. A trivial combination of these two automatic systems has also been tried.

The first interactive run investigated whether better articulated queries are helpful; e.g., Lunar Rover scenes are characterized by 'a black sky', and the Starwars scene by 'shiny gold'. A second interactive run studied whether a user could improve, with limited effort, the results by combining the four other approaches.

A (somewhat disappointing) lesson from the retrieval results was that the transcript-only run outperformed all other approaches, including the interactive runs.

3.7 University of Maryland

The University of Maryland, working with visiting researchers from the University of Oulu, extended methods used for image retrieval based on the spatial correlation for colors by using a novel color content method, the Temporal Color Correlogram, to capture the spatio-temporal relationship of colors in a video shot using co-occurrence statistics. The temporal correlogram is an extension of HSV color correlogram, and

[2] The Lowlands Team consisted of the database group of the CWI, the multimedia group of TNO, the vision group of the University of Amsterdam, and the language technology group of University of Twente.

computes an autocorrelation of the quantized HSV color values from a set of frame samples taken from a video shot.

To implement the approach the video material was segmented to create shots using VideoLogger video editing software from Virage and our own MERIT system. From each shot, the first frame was selected as a representative key frame, and the static image color correlogram was obtained. In order to calculate the temporal correlogram non-exhaustively and to keep the number of samples in equal for varying shot lengths, each shot was sampled evenly with a respective sampling delay so that the number of sample frames did not exceed 40. After segmentation, shot features were fed into our CMRS retrieval system and queries were defined using either example videos or example images depending on the respective VideoTREC topic specification. VideoTREC result submission contained retrieval results of two system configurations. The first configuration was obtained using the temporal color correlogram for the retrieval topics that contained video examples in the topic definition and the second configuration used the color correlogram for topics that contained example images in their definition.

3.8 University of North Texas

The University of North Texas team extracted frames from the collection at regular five-second intervals. These frames were then run through a keyframe extraction process, which removed the redundance of highly similar frames and ensured the presence of frames outside the prescribed normal distribution limits. The resulting keyframes were placed into UNT's Brighton Image Searcher application, which is based on mathematical measures that correspond to primitive image features. Two members of the team independently used this application to attempt to retrieve relevant keyframes for 13 of the original search topics. For each topic, the two people performing the searches selected a keyframe that appeared to answer the question. The chosen keyframe was then used as an exemplar to find keyframes similar to it. Precision scores were better than expected due to the human judgment presence.

4. Summary and Analysis of Approaches

The brief review of the approaches to video indexing and retrieval taken by track participants shows those approaches to be very varied indeed. Some sites ran interactive searching with real users (DCU) while others did their query processing entirely automatically (JHU). Some used automatic speech recognition transcripts (CMU, IBM, Lowlands) while others based their retrieval entirely on the visual aspects of video (UNT, UMd). Some groups used many automatically extracted video features as part of their retrieval (CMU, IBM, Fudan, Lowlands) while others used only a limited set of identified features (UMd, UNT). Some groups were experienced in the video indexing field and were able to leverage upon previous experience and background in working with video (IBM, CMU) while for other groups, this was their first real experience of doing video indexing and retrieval (JHU, UNT).

As might be expected for the first running of an evaluation framework still very much under construction, the results are probably most useful for small-scale comparisons - within-topic and between closely related system variants. Plausible cross-system comparison will have to wait on better consistency in topic formulation, agreement on better measures, larger numbers of comparable data points. We expect some of the participants will do further investigation and analysis of their own TREC2001 video track results and such analysis may give further insights which will be of benefit to those participants.

In terms of performance results, overall the absolute performance figures were very mixed. In the known item search tasks the mean average precision for the best 2 interactive runs (1 site) was a little over 0.6, across ~31 topics, while another group submitted two runs over the same topics and scored a consistent 0.23. Scores for comparable automatic runs ranged from 0.002 to 0.609. The use of averages may be misleading, particularly given the large number of topics for which any given system found no relevant clips. For the general search tasks the results were generally even poorer with mean partial average precision scores (based on half the collection) ranging from 0.03 to 0.23 for interactive runs on 12 topics and from 0.02 to 0.11 for automatic runs on 28 topics. The multiplicity of factors makes success as well as failure analysis a real challenge. Ongoing examination will try to explain differences in performance, but it may be that the first running of any TREC track will always be the one which irons out the difficulties and throws up the unforseen problems and that was certainly true here.

5. Conclusions and Contribution of the TREC2001 Video Track

The TREC2001 video track revealed that there are still a lot of issues to be addressed successfully when it comes to evaluating the performance of retrieval on digital video information. It was very encouraging to see interest from the community who specialise in evaluation of interactive retrieval, in what was achieved in the video track.

Overall, the track was successful with more participants than expected and the promise of even more groups this year (2002). However the real impact of the track was not in the measurement of the effectiveness of one approach to retrieval from digital video libraries compared to another approach but was the fact that we have now shown that there are several groups working in this area worldwide who have the capability and the systems to support real information retrieval on significant volumes of digital video content. As an indication of what our field is now capable of and of the potential we have for future development, the TREC2001 video track was a wonderful advertisement. There have also been many lessons learned from the track, for example the technical issues related to defining frame numbers in video which are consistent across the decoders used by different participants.

One of the interesting questions thrown up by the general search task was to do with the complexity of the topics and the relationship between the text and non-textual parts of the topic where topics had image/audio/video examples. Often it was not clear that all of the example was exemplary, but there was no way to indicate, even to a human, what aspects of the example to emphasize or ignore. We're not sure what to do about this but it may be that by making the topics more focussed, as we are planning this year, this issue may disappear.

For this year we will use a new dataset which is greater in size, and more challenging in nature – at the time of writing it appears that the TREC2002 video track will have over 20 participating groups and that we will repeat the searching task with a more focussed set of topics, some with multimedia topic descriptions.

We are also expecting to have a variety of detection tasks such as the occurrence and number of faces, identifying text in the image and then submitting it for OCR, categorising the audio as either speech, audio or silence, and so on. The search task will be as before, namely emulating the scenario where a user approaches a video retrieval system with some information need which is satisfied by the retrieval of some number of video clips from the video archive and the evaluation will, as before, be done in terms of precision and recall.

Authors' Note: The authors wish to extend our sympathies to the family and friends of our co-author, Mark E. Rorvig, who passed away shortly before this paper was submitted. We thank Diane Jenkins from UNT for helping us to clarify some of the contributions from University of North Texas.

References

1. Smeaton, A.F., Over, P. and Taban R. The TREC-2001 Video Track Report, in NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) (available at http://trec.nist.gov/pubs/trec10/t10_proceedings.html)
2. The OpenVideo Project. Available at http://www.open-video.org/ (last visited 30 April 2002).
3. The TREC Video Track. Available at http://www-nlpir.nist.gov/projects/t01v/t01v.html (last visited 30 April 2002)
4. Proceedings of the Tenth Text REtrieval Conference (TREC-2001), Gaithersburg, Maryland, November 13-16, 2001 Available at http://trec.nist.gov/pubs.html (last visited 30 April 2002).
5. Lee, H. *et al.*: Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video. In *Proceedings of the Fourth European Conference on Digital Libraries (ECDL)*, Lisbon, Portugal, Springer-Verlag LNCS 1923, J. Borbinha and T. Baker (Eds), pp.206-218, September 2000.
6. Yuefei Guo, Lide Wu.,„A novel optimal discriminant principal in high dimensional spaces“, *Proc. International Conference on Development and Learning*, June 2002, MIT
7. Manjunath, B. Wu, P. Newsam, S. and Shin, H. A Texture Descriptor for Browsing and Similarity Retrieval.' *Journal of Signal Processing: Image Communication*, 16(1), pp. 33-43, September 2000.