Físchlár @ TREC2003: System Description

Cathal Gurrin Hyowon Lee Alan F. Smeaton

Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland

{cgurrin, hlee, asmeaton}
@computing.DCU.ie

ABSTRACT

In this paper we give an outline of the Físchlár system developed to enable participation in the interactive searching task within TRECVID 2003. TRECVID is an annual benchmarking exercise which measures the effectiveness of various video information retrieval tasks, including interactive retrieval. The accompanying video provides a usage scenario for our TRECVID2003 system which contains a series of screen snapshots and user interactions highlighting how a user utilises the system in order to perform video shot retrieval.

Categories and Subject Descriptors

H.5.1 [**Information Systems**]: Information Interfaces and Presentation – *Multimedia Information Systems; Video*

Keywords

Video Information Retrieval, Video Libraries, Content Browsing

1. INTRODUCTION

Information Retrieval from digital video archives is becoming increasingly important as the amount of video material available to us in digital form, increases. Many research groups worldwide, are developing approaches and systems to enable analysis, indexing, browsing, searching and summarization of digital video with progress being made in this area. In order to allow the field to develop, and to push out the boundaries, the TRECVID initiative has taken place annually since 2001. TRECVID is an evaluation forum where participating groups use the same data, the same user topics and the same evaluation mechanisms to allow video retrieval systems to be compared [1].

In this paper and in the accompanying video, we present a usage scenario for the video retrieval and browsing system we developed for TRECVID 2003. Our system is one of the family of Físchlár systems [2] developed within our research group and supports user querying on text and/or image as well as subsequent video browsing.

2. SYSTEM OVERVIEW AND OUR TRECVID2003 EXPERIMENTS

The Físchlár system developed for TRECVID2003 reads in and analyses video in MPEG-1 format and from that generates an MPEG-7 description of that video for subsequent content-based operations. Internally, XML documents are generated as a result of user interactions and these XML documents are transformed using stylesheets into HTML for rendering on web browsers. As part of TRECVID, a common set of shot boundaries and a common set of keyframes are used so there was no need for us to apply our own techniques to this task.

The video retrieval task which the search task in TRECVID addresses is where a user has a limited amount of time in which to locate as many individual video shots as possible which satisfy some information need. The "topic" is described in a combination of text narrative, illustrative images or short video clips. So, for example, one TRECVID topic was to locate shots of a locomotive approaching the viewer and several still images and video clips illustrating this were included. These represent rich multimedia statements of an information need and allow interactive users, located at each participating site, to each form a reasonably comprehensive understanding of each topic before embarking on interactive searching. In all there were 25 different topics used in TRECVID in 2003 varying from the specific (find shots of the Mercedes logo) to the more vague (find shots of one or more groups of people, a crowd, walking in an urban environment, for example with streets, traffic, and/or buildings).

In video information retrieval, there are two dominant approaches namely the manual annotation of content description and secondly using the text associated with the spoken dialogue as a lever for retrieval. In the first case, manual annotation is expensive, time-consuming and not scalable to huge archives. In the second case, indexing and retrieval based on the words spoken, either using closed caption text or using the output of some speech recognition, is useful only when the dialogue includes a description of what is on-screen which is not always the case [3]. For example, in a shot of a crowd of people in an urban environment it is unlikely that the dialogue would include things like "here is a crowd of people in an urban environment".

To address this, there is much research carried out into automatic feature detection where basic descriptive features such as indoor, outdoor, people, faces, vegetation, cityscape, etc., can be

ACM Multimedia'04, October 10-16, 2004, New York, NY, USA.

automatically detected and used to reduce the search space for searching and browsing. Unfortunately, at the present time such feature detection is not accurate enough to be reliably used on large sized archives.

In our participation in TRECVID in 2003 we used a technique for video shot retrieval, which combined text-based search through the automatic speech recognition [4] with an image-based search, which matched shot keyframes. Automatic speech-recognised (ASR) text was aligned with shot bounds and a user, given a TRECVID topic description (text, images and/or video clips) was allowed to input search terms into a search query box as the first component of their search. Once some relevant video shots had been identified by the user and saved for submission, the user was allowed to add some shots (strictly speaking, some keyframes taken from video shots) to the query and the query became a combination of text and image(s). Our rationale for doing this was that we were interested in measuring the relative contributions of image retrieval and of text retrieval, in the task of video shot retrieval.

The search test collection used in TRECVID2003 consisted of over 60 hours of CNN, ABC and C-SPAN TV news data. This contains 32,318 shots from the common shot boundary reference provided by the CLIPS-IMAG group. TV news programs are very well structured and follow a fixed form with anchorperson shots being the most common. As such they do not generally provide a good data source for shot retrieval but the TRECVID2003 collection included TV commercials and footage illustrating news stories and this provided the real target for shot retrieval.

In our work we developed two versions of the system, one that supported text-only querying of the video collection (using the ASR text) and one that supported text and image querying. The latter system also incorporated a feedback mechanism whereby a user could select shots that they felt were important and would want to retrieve more shots like this.

Prior to indexing, the ASR transcript for each shot was processed to remove stopwords (words that occur too frequently to aid the search process, “the”, “of”, “and” etc.) and then stemmed using Porter’s algorithm. The ranking algorithm we chose to employ for searching the transcripts was the popular BM25 algorithm [5].

The TREC2003 image search engine performed retrieval based on an index of all keyframes from shots in the collection as well as the representative query images. Three colour-based features and one edge-based feature were used to represent each image [6]. Image-to-Image dissimilarity comparison was computed using an absolute distance measure for each of the four visual features. These four scores were normalised based on ratio averaging.

For any query consisting of both image and textual elements, the method of combining image and textual evidence relies heavily on the user making a judgment of which element is more important for any given query. The weight of both elements was ultimately controlled directly by the user using five radio buttons which allowed the user to determine the weighting between text and image evidence on a five point scale, regulated in steps of 0.25 in size.

Within our interface, an “**Add to Query**” button below each keyframe was used to allow the user to add the shot’s content (keyframe) as part of his/her subsequent query. When the user encounters a shot or shots within the task that she thinks would help the retrieval task, she can click this button to add the shot into the Query panel and then click on the “**Search”** button to submit a new query.

For the search task, we used two variations of the above-described system. A total of sixteen test users took part in our interactive experiments and in each case, for each search, the duration was limited to 7 minutes in order to keep the pressure on users to complete the task as quickly as possible. We used groups of users to try different systems in different combinations in such an order to eliminate both topic and system bias.

The evaluation of our system(s) in terms of performance is reported elsewhere [6]. In summary, we found great variability in the performances of our systems depending on the topic with overall performance on some topics quite good and others quite poor. In terms of systems, there was little difference in overall performances when image retrieval was used and when it was not but further analysis of this data is required in order to determine further conclusions.

In the accompanying video we present the system in use by showing user progress during a search for forest fires. This is not one of the TRECVID2003 topics but serves to illustrate the performance of image retrieval combined with text retrieval.

3. ACKNOWLEDGMENTS

This material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA. The support of the Informatics Directorate of Enterprise Ireland is gratefully acknowledged.

4. REFERENCES

[1] TRECVID – *TREC Video Retrieval Evaluation*. Available online: http://www-nlpir.nist.gov/projects/t01v/t01v.html

[2] Smeaton, A.F., Lee, H., and McDonald, K. Experiences of creating four video library collections with the Físchlár system. *Journal of Digital Libraries: Special Issue on Digital Libraries as Experienced by the Editors of the Journal*, 2004.

[3] Smeaton, A.F. Indexing, browsing and searching of digital video. *Annual Review of Information Science and Technology, 38*, Ch. 8 (2004), 371-407.

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

[5] Walker, S., Robertson, S., Boughanem, M., Jones, G., and Jones, K.S. Okapi at TREC-6 Automatic ad hoc, VLC, routing,filtering and QSDR. In *Proceedings of the 6th Text Retrieval Confrerence (TREC6)* (Gaithersburg, MD, November 19-21, 1997). NIST Special Publication 500-240, Gaithersburg, MD, 1997.

[6] Browne, P., Czirjek, C., Gaughan, G., Gurrin, C., Jones, G., Lee, H., Marlow, S., McDonald, K., Murphy, N., O'Connor, N., O'Hare, N., Smeaton, A.F., and Ye, J. In *Proceedings of the TRECVID 2003 Workshop*, Text REtrieval Conference (TREC) (Gaithersburg, MD, November 17-18, 2003). NIST, Gaithersburg, MD, 2004.