VIDEO SEARCHING AND BROWSING USING VIEWFINDER

By

Dan E. Albertson	**Dr. Javed Mostafa**	**John Fieber**
Ph. D. Student	**Associate Professor**	**Ph. D. Candidate**
Information Science	**Information Science**	**Information Science**

School of Library and Information Science
Indiana University

Abstract

Several researchers consisting of students and faculty from the School of Library and Information Science at Indiana University developed a video retrieval system named ViewFinder for the purpose of providing access to video content for a project named the Cultural digital Library Indexing Our Heritage (CLIOH) at Indiana University Purdue University at Indianapolis (IUPUI). For our role in the Text Retrieval Conference (TREC) and its video track, we took the existing system, made notable modifications, and applied it to the video data provided by the conference. After conducting 1 interactive search run, we generated our search results and submitted them to TREC where human judges determined the relevancy of each returned shot and assigned an averaged precision ranking for each topic. From these results we were capable of drawing conclusions of the current system, and how to make it more productive in future versions.

Introduction

With the accumulation of digitized video, groups and individuals are becoming more and more interested in the preservation and organization of such content. Along with this preservation and organization, there is a need for systems that can provide easy and efficient access to this archived video (content). This problem is the focus of ViewFinder, a video retrieval information system.

The main goal of Viewfinder is to have it applied to a project being conducted at Indiana University Purdue University at Indianapolis (IUPUI) named the Cultural digital Library Indexing Our Heritage (CLIOH). This project deals with the preservation of multi-media content of the ancient world (Mayan ruins, etc.). One such form (of content) is video, and this is the current focus of ViewFinder. ViewFinder attempts to provide users with individual keyframes (of shots located within video files) according to the user's information need.

For the purpose of participating in the Text Retrieval Conference (TREC) and its video track, we took the existing system, made notable modifications, and applied it to the video data provided by the conference. We then followed conference procedures and performed 1 interactive ("human in the loop") search run consisting of 25 individual topics (provided by the conference). We then generated results and submitted them to

TREC, where human assessors compared our results with the number of identified relevant shots (in the data set), and assigned an average precision for each topic. You may further explore the average precision formula used by TREC in Vorhees, E. M., and Harman, D. K. (2001). In addition to conducting an interactive search run, our system was developed with use and knowledge of the actual search test collection, known as type-A.

Related Literature Review

In recent years there have been various advances in regards to this research problem. This is possibly due to the large increase of multimedia content (especially video) being digitized and made accessible via the World Wide Web and other multimedia information systems.

Along with this increase in video content, there is an increase in people who choose to search for such content. Spink, Goodrum, and Hurson (2001) concluded from a study on Excite query logs between the years 1997 and 1999 that queries for video content increased over 100%. In fact, video queries counted for 0.7% of overall queries in 1997 and counted for 1.6% in 1999 (Spink et al., 2001). Spink et al. (2001) go on to further conclude that "video searching became more frequent during this period with the expansion of video material on the Web."

The findings above suggest that it is very important for us (as system developers) to explore how to better provide easier and more efficient access to video content. Cruz and James (1999) provide insight into this problem by focusing on aspects such as user query generation coupled with the user interface design. They go on the detail their system named Delaunay (Cruz and James, 1999). They express the importance of users having the capability for "pre- and post-query refinement" (Cruz and James, 1999). Furthermore, Cruz and James (1999) also stress the importance of accommodating the search interface to both novice and expert users of multimedia retrieval systems. One example (of their system) is that novice users have the option of a Search Assistant which may assist in "pre-query refinement"(Cruz and James, 1999).

Spink et al. (2001) also pay close attention to query generation of the user. They claim that, "Web users generally search for multimedia information as they search for textual information" (Spink et al., 2001). Also, Spink et al. (2001), find that multimedia queries contain more search terms (mean of 2.4) that that of general (non-multimedia) Web queries (mean of 1.91).

Spink et al. (2001) further discovered that the term "video" is the most commonly used term when users query for video content. This brings them to suggest that other useful searching features such as the incorporation of searching with file extensions (.mpeg, .avi, .mov, .wav, etc) would be very helpful in user query formation (Spink et al., 2001).

While these search features may prove to be helpful in future multimedia systems, they are still primarily focused on textual searches. Other research has attempted to increase video/image retrieval system performance by actually moving away from textual searching and incorporating content based searching. Zhou and Huang (2002) describe a retrieval system where contextual information (color, shape, texture, etc. described as "low-level features") is combined with user's keywords (or "high-level semantic concepts"). They go on to present that searching on contextual information alone is usually not sufficient in generating relevant results, however, would serve the purpose in thesaurus updating (or adaptation of keywords to images and vice versa) (Zhou and Huang, 2002).

Other studies take into account the contextual based information, but emphasize the user's criteria in image relevancy as opposed to just system evaluation. Hoi and Rasmussen (2002) claim that "image information retrieval systems require not only the technical aspects of image databases but also user-centered aspects of retrieval because the success or failure of a digital image database system ultimately depends on whether or not it can really meet user needs." They go on to further claim that the majority of research is still primarily focused on system performance and data indexing rather than the theoretical background of the design of image retrieval systems and the assessment of data indexing (Hoi and Rasmussen, 2002).

Problem

As mentioned in the earlier sections of this paper, our problem focuses on providing easy and efficient access to video content where large archived (video) data exists. The video data provided by TREC proved to be sufficient in exploring these research problems with ViewFinder. Moreover, the total size of the video collection consisted of 68.45 hours (of MPEG-1) including 40.12 hours for the search test collection, 23.26 hours for the feature development collection, and 5.07 hours for the feature test collection.

To conduct system tests and the TREC tasks (after making some notable modifications) we applied the existing system to the data provided by the conference (through the Internet Archive). Some of the modifications we made consist of a reformulation of (system) queries, switch from a MySQL database to Oracle, display adjustments (for the individual keyframes), incorporated a textual keyword search feature, and adapted the search attributes (for suitable interaction with the Internet Archive video metadata).

Due to time constraints, the database resulted into a very basic structure. The only metadata used is that which is provided from Internet Archive which consists of title, description, and descriptors for each individual video file. In our database we had 1 table and created a separate field (column) for each of the above features.

Although, this proved to be sufficient data to develop a working prototype in order to participate in the video track, we initially assumed that it wouldn't serve the purpose of a practical video retrieval system. Moreover, although TREC requests search returns be individual video shots (located within a video file), our database system only included metadata corresponding to an individual video file. We would eventually encounter the problem of not being capable of distinguishing shots located within the same video file from one another.

This prevented us from being capable of determining any relevancy ranking between the individual shots located within the same video file. As a result, once any relevant (or matching) video(s) are identified, the system returns the keyframes to the user in a sequential fashion. For example, if the system matches the user's query to 2 video files that (both) contain 50 shots a piece, the user would be presented with a shot order such as: shot 1 from video 1, shot 1 from video 2, shot 2 from video 1, shot 2 from video 2 … up until the final shot. In the latter sections of this paper we discuss future improvements of ViewFinder, that we feel will eliminate these problems for upcoming TREC conferences.

Methodology

For our interactive ("human in the loop") search run, we allowed the user to conduct the evaluation of the returned results. Moreover, it was up to the user whether or not to reformulate the query and continue or stop the search topic and settle on results.

While using the ViewFinder system, the user is given several options of searching techniques (or search features). One of the features consists of a keyword search. This keyword search allows the user to type in the keywords (which they wish to use) and compare them to the description (field) for each individual video file. If there are any matches between the keyword and any video description, the keyframes corresponding to matching video(s) are returned to the user.

In addition, the user is presented with several (video) attributes in which they are allowed to browse. These attributes are presented to the user in a series of drop down menus. One example is that the user can select "title" in the "search options" drop down menu, and retrieve all the video titles in the collection. The user can then select a title (by clicking on it and highlighting it) and click on the "search" button, which will run a query for that particular title. This will prompt the system to return the keyframes associated with that video title.

The same operation can be conducted with a "descriptors" option in the drop down search menu. However, unlike the "title" search (which will only return results for one individual video title) it is possible for the descriptors search to return shots from several different video files (if the same descriptors overlap for multiple videos).

After any of the above searches are performed, it was up to the user to evaluate the search results and decide whether or not to reformulate the query and continue searching. In addition, we also made no attempt to restrict the user in any query formulation/reformation.

Another option of ViewFinder is to utilize the "expand" search of ViewFinder. This is found in the drop down menus located directly below each individual keyframe. A user can “expand” a search on any keyframe they would like. This “expand” feature will take the descriptors associated with that particular keyframe, and compare it with the descriptors from all the other video files, and return any matches.

Since ViewFinder can only display up to 9 individual keyframes at one time, the user is still capable of browsing all the video clips returned (in the case of there being more than 9 matching keyframe). Utilizing the “More Clips” and “Back” buttons located on the interface allows for such browsing. The “More Clips” button becomes initialized after more than 9 keyframes are returned by a query, and the “Back Button” is initialized after the “More Clips” button is clicked (and the user is on a page other than the first).

In regards to time restriction for individual topic searches, we didn’t place any on the user. The user finished the search when they felt they exhausted all relevant video shots. After the user decided to end each of the search topics, they would select the “finish” button. The finish button prints out the top 100 (or less) to the Java console where results can be gathered and formatted.

Results

A human assessor (or assessors) from NIST manually judged relevancy of each result shot (for each search topic). After concluding on the number of relevant shots returned as compared to the total number of relevant shots identified in the data set, an averaged precision was assigned to each search topic performed. You can read more on the averaged precision formulation in Vorhees, E. M., and Harman, D. K. (2001).

We conducted 1 interactive search run where we finished (or attempted to answer) all 25 topics. In the end, our results can be summed up several different ways. First, our mean averaged precision for the 25 search topics was 0.05488. We had a range of 0.2451 with a minimum score of 0.0000 (on topics 75 and 86) and a maximum of 0.2451 (on topic 93). Ranking among other systems participating (in the topic) includes a range from 1st (best) to a tie for 35th (worst). Moreover, our average ranking for the 25 topics is 18.64 out of an average of 36.88 participating runs. However, these comparisons among our results and other system results may not be appropriate. The reason is that other search runs varied from an interactive style and manual style and from type-A and type-B. Our search run was the only one that used an interactive task coupled with type-A system development. Furthermore, the majority of search runs were performed using a manual task, and those utilizing an interactive run used a type-B system.

Conclusions

After reviewing the results, we were initially correct in assuming that the lack of metadata for each individual shot greatly inhibited ViewFinder’s searching performance. In future video tracks we plan on populating a database with metadata for each individual shot, and provide a more robust way of searching for specific information needs. For example, instead of limiting our search attributes to title, description, and descriptors alone, we plan on adding attributes for keywords, any subject(s), notable people, important landmarks, presence of animals, landscapes/cityscapes (just to name a few).

Moreover, we also hope to make use of content-based image retrieval. This will allow us to build a more diverse search strategy by allowing users to search for shots/keyframes with similar shapes, patterns, and colors. Not only will this enable us to generate more metadata that can be utilized by the system, but also it will allow the system to set up a sense of "context" surrounding each keyframe or keyword.

References

Brown, P., et al. (2001). Dublin City University video track experiments for TREC 2001. Paper presented at The Tenth Annual Text REtrieval Conference, Gaithersburg, MD.

Choi, Y. & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing & Management, 38*(5), 695-726.

Cruz. I. F., & James, K. M. (1999). User interface for distributed multimedia database querying with mediator supported refinement. *Proceedings of the International Database Engineering and Applications Symposium, Montreal, Canada,* 433 – 441.

Hassan, I, & Zhang, J. (2001). Image search engine feature analysis. *Online Information Review 25* (2), 103 - 114.

Smeaton. A. F., Over, P., & Taban, R. (2001). The TREC-2001 Video Track Report. *NIST Special Publications 500- 250: The Tenth Annual Text REtrieval Conference, Gaithersburg, MD,* 52 – 60.

Spink, A., Goodrum, A., & Hurson, A. R. (2001). Multimedia web queries: Implications for design. *Proceedings of the International Conference on Information Technology: Coding and Computing, Las Vegas, NV,* 589 – 593.

Vorhees, E. M., & Harman, D. K. (Eds.). Common Evaluation Measures. (2001). *NIST Special Publication 500-250: The Tenth Text REtrieval Conference,Gaithersburg, MD*, A14 – A23.

Zhou, X. S., & Huang, T. S. (2002). Unifying keywords and visual contents in image retrieval. *IEEE Multimedia, 9*(2), 23 - 33.