TRECVID 2004: Video Search Experiments at IUB

Dan Albertson and Javed Mostafa, Ph. D.
Laboratory for Applied Informatics Research
Indiana University, Bloomington
1320 E. 10th St. LI 011
Bloomington, IN 47405
812-855-2018
daalbert@indiana.edu jm@indiana.edu

The experiments presented in this paper explore topics surrounding video information retrieval (IR). This paper will discuss in detail our participation at TRECVID 2004. A video retrieval system named ViewFinder was developed to search and browse the TRECVID 2004 test data, and both manual and interactive search experiments were carried out. Each of the performed search experiments were in agreement with the task definitions and conference guidelines developed by TRECVID coordinators. This paper will present our approach for TRECVID participation which includes the development of ViewFinder and other supportive tools, and the experimental designs of our search runs. Results for each experimental search run are also presented.

1. INTRODUCTION

Several researchers from the Laboratory for Applied Informatics Research (LAIR) at Indiana University, Bloomington developed and tested a video retrieval system named ViewFinder for the purposes of participating in the 2004 Text REtrieval Conference's Video Workshop (TRECVID 2004). Researchers from LAIR have been active participants in the TRECVID forum since 2002.

ViewFinder is an ongoing research project exploring user interaction and user-interfaces for video information retrieval (IR). The first phase of this research initiative included developing a version of ViewFinder to effectively search and browse TRECVID 2004 data. After common TRECVID evaluation, other questions surrounding user interaction and interface features are being tested accordingly. In order to adapt ViewFinder to the TRECVID data, several major tasks had to be completed prior to designing and conducting the search experiments.

An in depth discussion regarding system development, search experiments, and search results will follow.

2. SYSTEM DEVELOPMENT

2.1 Data and Keyword Indexing

The test data for TRECVID 2004 included 64 hours of CNN Headline News and ABC World News Tonight video and approximately 33,000 keyframes[1]. The original video was broadcast throughout October, November, and December of 1998 [1].

A variety of textual information accompanied this visual data. One example of the textual information corresponded to the individual video files. Similar information was also issued for TRECVID 2003 and was formatted in XML. This information was extracted using Java's XML API and indexed using JDBC. The extracted information was used to populate our *Video Table* (see Table 1 for database structure). The *Video Table* as shown in Table 1 is identical to the *Video Table* developed for last year's version of ViewFinder [2].

The common shot boundary reference provided by TRECVID was used to populate the *Shot Table* [1]. The shot boundary reference was also issued in XML format and included specific information about individual shots. Since this information was also issued in XML, it was extracted and indexed in a similar fashion to the video information described above. The table structure resulting from the

[1] This number denotes one RKF keyframe per shot.

common shot boundary reference is listed under *Shot Table*.

Table 1: Database structure of ViewFinder.

Table Name	Attributes
Video Table	video_id, video_url, video_use, video_source, video_date, num_of_shots
Shot Table	video_id, video_filename, video_start_time, video_duration, shot_id, shot_start_time, shot_duration, image_url, time_of_shot
Keyword Table	video_id, shot_id, keyword, weight, freq_per_shot, freq_per_video, freq_per_dataset
Unique Terms Table	video_id, keyword, num_of_shots, idf

The automatic speech recognition (ASR) output was next to be indexed. We chose to index the ASR output formatted within the shot boundary information as donated by Jean-Luc Gauvain's Speech Processing Group at LIMSI-CNRS [3]. This ASR output was located within the mp7 directory of the ASR dataset and was in XML format. This choice of ASR output allowed us to easily associate all keywords with shot identifiers. In addition, using this version of ASR output allowed us to overcome many of the limitations of our TRECVID 2003 system such as correctly associating ASR terms with shot IDs, eliminating inconsistent timing intervals, and loss of terms.

A similar approach was used in handling the ASR output, i.e. Java's XML and JDBC APIs was used for parsing and indexing. Considering the XML structure of the ASR output was very similar to the structure issued for the common shot boundary reference, tools for extracting shot boundary information was easily modified to capture keyword information.

The ASR output was first used to populate the *Keyword Table*. The *Keyword Table* contains all the extracted terms along with corresponding video and shot IDs. In the case that any keyword appeared multiple times in the same shot, the redundant keyword(s) were not indexed. To keep track of these redundant keywords, a frequency per shot integer, or the number of times a given keyword appeared in one shot, was counted and indexed along with each keyword. Redundant keywords within a video file were indexed and can be distinguished by different shot IDs.

The ASR output was also used to populate a table of unique terms. The terms indexed in the *Unique Terms Table* are terms that are unique[2] for each video not the entire dataset. Indexed along with each unique term was the number of shots that each term appeared in for a particular video.

After entering all terms into the *Keyword* and *Unique Terms* tables, relevancy weights were applied to each. An *idf* weight was calculated for each record in the *Unique Terms* table. TF•IDF weightings were adjusted to meet the structure of video[3] and applied to each term from the ASR outputs. These TF•IDF weights were used to determine levels of relevance for shots. This approach was also performed for the TRECVID 2003 version of ViewFinder.

2.2 User Interface Features

The graphical features of the ViewFinder user-interface were developed using Java Swing. The client side of ViewFinder is a Java Applet and is accessible over the web using any Java 1.4 + enabled browser.

ViewFinder's interface is made up of two primary panels that include a results panel and a search panel (*See Figure 1 for a screenshot of current ViewFinder interface*). The results panel, located on the left hand side of the interface, has several different utilities. First, it displays keyframes for the shots returned after a search or exploratory browsing. This enables users to visually browse search results or within specific video files. The results panel can display up to eight search results on a page which are ordered according to relevance.

[2] "Unique" in this context refers to terms being indexed once for a video.

[3] IDF was computed using the number of shots per video and the number of shots where the terms appeared. TF represented the number of times the word appeared in a shot.

Moreover, top-left to bottom-right is the descending order of relevance for results after a search has been performed. The keyframes generated and issued by TRECVID were converted to thumbnails for use in the ViewFinder interface.

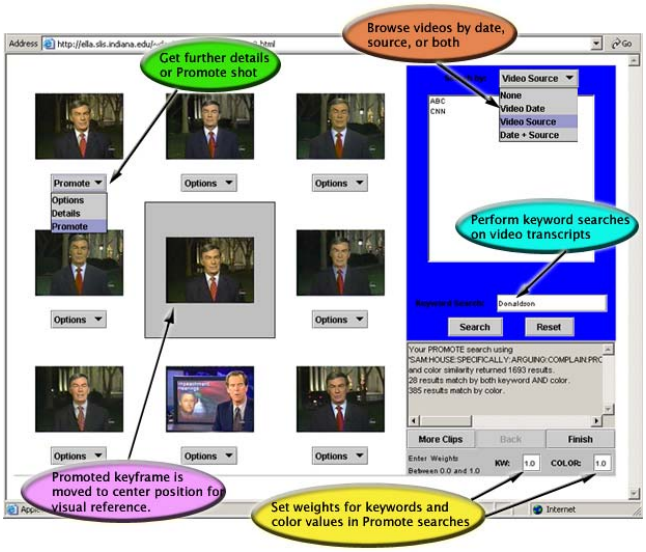

Figure 1: Screenshot of ViewFinder with reference captions.

The results panel also contains the *Promote* and *Details* functions. These functions can be selected from the drop down menus located directly below each of the eight search results and each menu corresponds to the keyframe located directly above it.

The *Promote* function executes a new search. A *Promote* search performs similarly to a *More like This* search feature offered by several popular search engines. Information associated with the selected (i.e. promoted) shot is used to formulate a query. The latest[4] version of ViewFinder has been implemented to include both keyword and color information in query formulation for a *Promote* search. Users can weight each of these characteristics and add additional search terms of their choosing for refining search results[5]. When a user chooses to *Promote* a specific shot ViewFinder will:

1. move the keyframe associated with the promoted shot to middle display position for visual reference
2. retrieve significant[6] keywords indexed for the promoted shot and gather terms entered by the user
3. retrieve other shots with matching keywords[7]
4. retrieve results with similar color characteristics
5. normalize keyword and color similarity values
6. compute similarity scores for color and keyword based on user weightings
7. compute overall similarity score for returned shots
8. sort and return results to user

The *Details* function will retrieve and present additional information corresponding to that particular shot. Users are capable of viewing any shot's source, date, unique IDs, and the full sized keyframe. These details are displayed in a separate pop-up window.

The search panel on the right-hand side of the interface has several features for searching and browsing the TRECVID data. A keyword search has been implemented. The entered keywords are compared against the ASR outputs and shots with matching keywords are retrieved. Similar to what was described for the *Promote* search, if a query contains more than one keyword the system will perform an 'OR' search and the overall relevancy computation is the same to what was described for the *Promote* search.

ViewFinder does provide some limited browsing features. Certain video headings including video date, video source, and the combination of date and source can be selected from the menu at the top of searching panel. Once a heading is selected, a

[4] The latest version employs a color search as part of the Promote function whereas the version evaluated through TRECVID 2004 only incorporated textual searching.
[5] This feature was also not implemented in the TRECVID 2004 evaluation version of ViewFinder.

[6] Significant keywords have a TF•IDF weight that exceeds a predetermined threshold.
[7] *Promote* is an 'OR' search where shots that contain any of the promoted keywords are returned. Shots with two or more matching keywords have their *tf·idf* values combined and an overall relevancy score for that particular shot is calculated.

matching set of choices are retrieved and presented in the list box located directly below the drop down menu. Users then select one of the choices and click the search button and results are retrieved and displayed.

The *More Clips* button, *Back* button, and feedback field are additional features of the search panel. The *More Clips* and *Back* button are necessary for allowing users to review all the retrieved shots. The feedback field displays information corresponding to the last performed search such as the query used, number of results, number of results that match by color, etc.

Communication between client side functions and the backend database is achieved through use of Java Servlets and JDBC.

3. EXPERIMENTAL DESIGN

For our TRECVID 2004 experiments, one interactive and one manual search run were conducted. ViewFinder was classified a category 'C' system as it was developed using the approach described above.

Our manual run fulfilled the mandatory/baseline run required by TRECVID. This mandatory run only allowed searches using the ASR output. All 24[8] search topics were completed in sequential order. One subject completed all 24 search topics over two sessions. The project lead and system developer of ViewFinder performed the manual run. Five minutes were allowed for each topic, and the maximum was used in all topics. No restrictions were placed on the subject in regards to predefined searching or query formulation.

The interactive run also only allowed searching via ASR output. All 24 topics were completed for the interactive run. Similar to the manual run, one subject was used to complete all search topics over two testing sessions. The same subject used for the manual experiment also completed all topics for the interactive run. Interactive searches began with the queries formulated for the manual run, or where the manual run left off. A maximum of ten minutes was allocated for each topic in the interactive run. This approach allowed a total of 15 minutes to complete one manual topic and one interactive topic. Durations for each topic in the interactive run ranged from four minutes to ten minutes, and the overall average for topic completion was 8 minutes 48 seconds. After initial use of the manual query, no further restrictions were placed on the subject, and the subject was free to query the system as needed.

[8] Only 23 of the topics were considered in evaluation.

4. RESULTS

Mean averaged precision (MAP), interpolated recall precision, and precision at *n* shots were all measured by NIST assessors. Definitions for each of these measurements can be further explored in the proceedings of TREC-10 [4]. Further analysis including ViewFinder's ranking for each search topic and averaged ranking across the search runs was conducted by LAIR researchers.

Due to the absence of any relevant shots for topic 146, evaluation of the search runs was based on 23 total topics, or topics 125 – 145 and 147 – 148. Our manual search run returned a total of 492 relevant shots out of a total of 1,800, an average of 21 per topic. The number of relevant shots ViewFinder returned for an individual manual topic ranged from 3 for topic 131 to 72 for topic 130. ViewFinder's MAP for the manual run was 0.031. This was compared to the overall MAP for all submitted manual runs[9] - 0.064. This MAP ranked 43rd out of 52 total manual runs.

Other performance measures for the manual run, including interpolated recall precision and the level of precision at *n* shots, are presented in Table 2: Summary of manual search results.

Table 2: Summary of manual search results.

Interpolated Recall Precision		Precision at *n* Shots	
0.0	0.2935	5	0.1043
0.1	0.1015	10	0.0957
0.2	0.0565	15	0.0870
0.3	0.0382	20	0.0913
0.4	0.0204	100	0.0630
0.6	0.0028	500	0.0325
1.0	0.0000	1000	0.0214

[9] This includes systems using different development and training protocols.

ViewFinder's averaged precision for each individual topic in the manual run were also ranked. There were 52 total runs submitted for the manual search task and ViewFinder's averaged precision ranking ranged from 4th best to 45th. The mean ranking of averaged precision was 30th. The range of averaged precision for the manual search run was from 0.001 for topics 131 and 143 to 0.120 for topic 134.

The interactive run returned 410 relevant shots out of the possible 1,800. This equals approximately 18 relevant shots returned per topic. The range of relevant shots returned for each topic was from 1 for topic 143 to 65 for topic 137. The MAP for our interactive run was 0.044 compared to the overall MAP average across all interactive search runs at 0.168. Our MAP score ranked 54th out of a total of 62 submitted interactive runs.

Interpolated recall precision and the level of precision at *n* shots for our interactive search run are presented in Table 3: Summary of interactive search results.

Table 3: Summary of interactive search results.

Interpolated Recall Precision		Precision at *n* Shots	
0.0	0.4826	5	0.2174
0.1	0.1565	10	0.1565
0.2	0.0728	15	0.1420
0.3	0.0535	20	0.1304
0.4	0.0323	100	0.0809
0.6	0.0093	500	0.0317
1.0	0.0000	1000	0.0178

ViewFinder's averaged precision ranking for each individual topic of the interactive run ranged from 29th to 57th out of the 62 total submitted interactive runs with an average of 47. Averaged precision scores ranged from 0.001 for topics 131 and 143 to 0.170 for topic 135.

5. CONCLUSIONS AND FUTURE WORK

The search experiments conducted for TRECVID 2004 included one manual and one interactive run based solely on searching the ASR output. Several conclusions can be developed after reviewing the results. Many of the conclusions drawn from this year's TRECVID experiments are the same as those drawn after TRECVID 2003 [2]. The conclusions, as listed below, pertain only to system improvements of ViewFinder; however, improvements to future user experiments are also being addressed by LAIR researchers. Our tentative conclusions include:

- *Tf·idf* weighting of the ASR output is believed to be effective when retrieving video; however, improvements can be made when applying to ASR output.
- ViewFinder needs to be more effective in limiting the overall number of returned search results.
- Users should be capable of searching the video collection using all Boolean operators as opposed to limiting queries to 'OR' searches.
- The search and browse functions of ViewFinder should be interconnected.
- Content-based searching needs to be further integrated into ViewFinder.

6. ACKNOWLEDGMENTS

The authors would like to thank the following individuals and organizations for their assistance in this project: Arvind Gopu from the High Performance Computing Support – Indiana University for improving techniques of indexing color similarity results, Stephanie Burks and the Research and Technical Services – Indiana University, TRECVID organizers, NIST, and LIMSI.

7. REFERENCES

[1] *TRECVID 2004 Guidelines*. Retrieved October 31, 2004 from http://www-nlpir.nist.gov/projects/tv2004/

[2] Albertson, D., Mostafa, J., & Fieber, J. (2003). Video searching and browsing using ViewFinder: Participation and assessment in TRECVID-2003. *Proceedings of the Text REtrieval Conference Video Workshop TRECVID-2003, November, 17 - 18, Gaithersburg, MD*.

[3] Gauvain, J. L., Lamel, L., & Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication, 37*(1-2), 89 – 108.

[4] 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.