The relative effectiveness of concept-based versus content-based video retrieval

Meng Yang
Open Video Project, University of North Carolina at Chapel Hill
100 Manning Hall, CB # 3360
Chapel Hill, NC 27599-3360
10-919-9663589
yangm@ils.unc.edu

Barbara M. Wildemuth
Open Video Project, University of North Carolina at Chapel Hill
100 Manning Hall, CB # 3360
Chapel Hill, NC 27599-3360
10-919-9628072
wildem@ils.unc.edu

Gary Marchionini
Open Video Project, University of North Carolina at Chapel Hill
100 Manning Hall, CB # 3360
Chapel Hill, NC 27599-3360
10-919-9663611
march@ils.unc.edu

ABSTRACT

Three video search systems were compared in the interactive search task at the TRECVID 2003 workshop: a *text-only* system, which searched video shots through transcripts; a *features-only* system, which searched video shots through 16 video content features (e.g., airplanes and people); and a *combined* system, which searched through both transcripts and content features. 36 participants each completed 12 video search tasks. The hypothesis that the combined system would perform better than both the text-only and the features-only systems was not supported, and large topic effects were found. Further analysis showed that concept-based video retrieval worked best for *specific* topics, whereas the hybrid retrieval techniques which combine both concept- and content-based video retrieval showed some advantage when searching for *generic* topics. The results have implications for topic/task analysis for video retrieval research, and also for the implementation of hybrid video retrieval systems.

Categories and Subject Descriptors

H. 3. 1. [**Information Storage and Retrieval**]: Content analysis and indexing --- *indexing methods.*

General Terms

Measurement, Performance, Human Factors

Keywords

TRECVID, hybrid video retrieval, user study

1. INTRODUCTION

To effectively retrieve the "relevant" videos or video segments, video indexing plays a crucial role. Traditionally, images and videos are catalogued and indexed manually, an approach called *concept-based* video indexing [1, 2]. In this approach, linguistic cues are used to represent, index and, thus, retrieve the non-linguistic audiovisual content. Because this approach has been used for many years, users have become accustomed to using words to express their queries for multimedia information. However, it is difficult for users to use only words to represent the multimodal video information they have in mind. Thus, multimodal querying and retrieval methods that combine textual, visual, and spatial information are needed to help users form their queries and retrieve videos. That is the research direction of the current *content-based* video indexing approaches. Videos can be indexed based on audiovisual features—either low-level features such as color, texture and shape, or higher-level semantic features such as people, objects, events, and settings. The performance of content-based video retrieval techniques is still a research question, particularly in relation to traditional concept-based indexing and retrieval methods.

Concept-based video indexing methods have "high expressive power" [3], which can easily communicate with users; but at the same time, it involves information loss during the media transformation process and also requires more intensive human labor. Content-based video indexing methods can be automated, and so have the potential to be cheaper and quicker, and they also have the potential to satisfy users' audiovisual information needs; however, they also have the limitation of the "semantic gap" [4] between users' queries and the content features that can be consistently detected and indexed. In short, each type of method has its benefits and deficiencies. To construct a more effective video retrieval system, these two methods could be combined in a "hybrid video retrieval" system [2]. Such a hybrid system may be more effective than either the concept-based system or the content-based system.

Although a few studies (e.g., [5, 6]) have explored user performance on content-based and on hybrid video retrieval systems, additional studies are needed to investigate the applicability of these innovative technologies: do they really help users search and retrieve videos? And if so, do they perform better on some topics than on others? The current study addressed these questions.

2. STUDY METHODS

The current study was conducted within the context of the interactive search task at the TREC Video Retrieval Evaluation (TRECVID) 2003 workshop [7]. We compared the effectiveness of three types of news video search systems: a

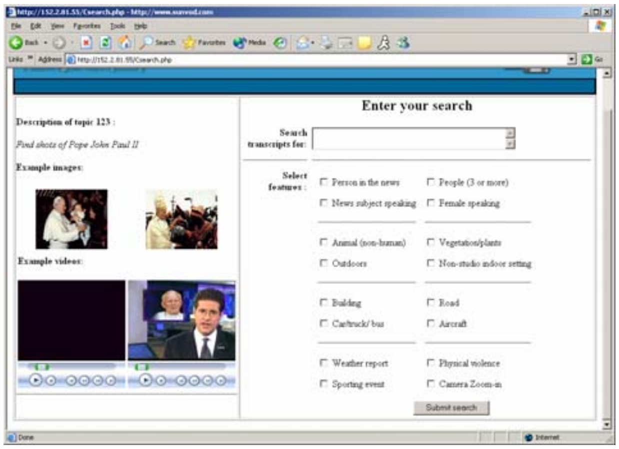
Conference'04, Oct. 10-16, 2004, New York, NY, USA.

text-only/concept-based system, a features-only/content-based system, and a combined/hybrid system. It was assumed that the performance of the combined system would be superior to that of either the text-only or the features-only system, and the performance of the features-only system would be the worst among these three systems.

2.1. The Retrieval Systems

The content included in the three systems was comprised of television news videos. The video collection contained approximately 133 hours of ABC/CNN news, and C-SPAN videos. The goal of the users' searches was to retrieve particular shots from this video collection.

Three types of system were compared: a *text-only* system, based on the video transcripts, with the results ranked based on the relevance score computed by MySQL's full-text search capability; a *features-only* system, based on 16 video content features (e.g., airplanes and people) defined by participants in TRECVID, ranked based on the average feature score for each shot across all features included in the search; and a *combined* system, which searched through both transcripts and content features (see Figure 1). In the combined system, the results were ranked by taking into account both transcript and feature scores, weighting the transcript-based scores twice as heavily as the aggregated feature scores.


http://152.2.81.55/Csearch.php - http://www.surveod.com
File Edit View Favorites Tools Help
Back Search Favorites Media
Links Address http://152.2.81.55/Csearch.php
Go
Done
Internet

Enter your search

Description of topic 123 :

Find shots of Pope John Paul II

Example images:

Example videos:

Search transcripts for:

Select features :

☐ Person in the news ☐ People (3 or more)
☐ News subject speaking ☐ Female speaking
☐ Animal (non-human) ☐ Vegetation/plants
☐ Outdoors ☐ Non-studio indoor setting
☐ Building ☐ Road
☐ Car/truck/ bus ☐ Aircraft
☐ Weather report ☐ Physical violence
☐ Sporting event ☐ Camera Zoom-in

Submit search

Figure 1. Search entry screen for combined search system

The search results were displayed in an agile-view mode [8], which contained four types of views: a horizontal display of the results list, a vertical display of the results list, a display of the shots immediately before and after a selected shot, and a view of additional keyframes provided for TRECVID.

2.2. Study Methods

Thirty-six study participants were recruited from among university students, faculty and staff by posting flyers in several buildings on campus, as well as email announcements within several departments. The TRECVID protocol provided 24 topics, including generic (e.g. roads with vehicles) and specific (e.g., Mercedes logo) topics, and topics related to people (e.g., Yassa Arafat), events (e.g., rocket/missile takeoff), and things (e.g., helicopter). A within-subjects research design was used. Each participant searched half (12) of the available topics, i.e., four topics with each search system. For each topic, the participant had at most 15 minutes to complete the search. The orders of the systems and the topics were counter-balanced among the 36 subjects, who were paid $20 each for their participation.

2.2. Outcome Measures and Data Analysis

Performance on the three systems was compared in terms of the precision and recall achieved by each participant on each topic. Precision was calculated as the proportion of the user-selected items judged relevant by the TRECVID assessors. Recall was calculated as the proportion of the relevant shots (as judged by the TRECVID assessors) retrieved by each study participant.

The three systems were compared with one-way and two-way analysis of variance (ANOVA) and post hoc analyses. In addition, the effects of topic on performance were of interest, and were evaluated by examining the interaction effects between topic and system.

3. RESULTS

Precision and recall for each of the three systems are summarized in Table 1. There was not a statistically significant difference between the systems, in terms of the precision achieved by study participants (F (2, 336) = 0.70, p=0.4968). The differences in recall were statistically significant (F (2, 429) = 11.89, p<0.0001), and post hoc Bonferroni t tests indicated that the recall achieved with the features-only system was lower than that achieved on the other two systems.

Table 1. Summary of performance, by system

	Precision		Recall	
	Mean	s.d.	Mean	s.d.
Combined	0.82	0.27	0.09	0.12
Text-only	0.81	0.31	0.11	0.16
Features-only	0.77	0.31	0.04	0.09

Of particular interest in the current study was the effect of topic on performance, so a further two-way ANOVA was conducted on all three outcome measures. For precision, the topic main effect was significant (F = 6.24, p<0.0001), but the system main effect and the interaction effect were not. For recall, both the system main effect (F = 27.88, p<0.0001) and the topic main effect (F = 18.78, p<0.0001) were statistically significant, as well as the system and topic interaction effect (F = 4.65, p<0.0001).

Post-hoc analysis for the topic-system interaction was difficult to interpret, since there were 24 topics. It is clear that the three systems performed differently for different topics. For instance, although the overall performance of the text-only and combined systems were superior, the text-only system performed best on only 10 topics, and the combined system performed best on 5 topics. The features-only system performed best on 8 topics, in spite of its worst average performance among the three systems.

Based on these analyses, it can be seen that topic differences play an important role in the performance of the different search systems. The text-only system performed best on such topics as roads with lots of vehicles, a mug or cup of coffee, and Pope John Paul II. The combined system performed best on such topics as a helicopter in flight or on the ground, the Mercedes (car) logo, and the Sphinx. The features-only system performed best on such topics as finding shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at, shots of the Tomb of the Unknown Soldier at Arlington National Cemetery, and shots of cats.

Table 2. A faceted classification of the subjects of pictures (Shatford, 1986, p. 49)

	Iconography (Specifics)	**Pre-iconography (Generics)**	**Iconology (Abstracts)**
Who?	Individually named person, group, thing	Kind of person or thing	Mythical or fictitious being
What?	Individually named event, action	Kind of event, action, condition	Emotion or abstraction
Where?	Individually named geographical location	Kind of place: geographical, architectural	Place symbolized
When?	Linear time: date or period	Cyclical time: season, time of day	Emotion, abstraction symbolized by time

To explore the effects of topics further, the 24 topics were each classified as one of two types: generic or specific, based on the Panofsky/Shatford mode/facet matrix [9] (see Table 2). *Specific* topics specify an individually named person, thing, object, event, action, geographic location, etc. *Generic* topics specify a kind of person, thing, object, event, action, geographic location, etc. Two members of the group individually assigned the category to each topic, and discussed the inconsistencies to reach a final agreement. Nine topics were classified as specific and 15 topics as generic; no abstract topics were found.

Table 3. Recall for the three systems on generic and specific topics.

	Com-bined	**Text-only**	**Features-only**	**All systems**
Generic	0.063	0.047	0.033	0.048
Specific	0.146	0.203	0.048	0.132
All topics	0.09	0.11	0.04	

A two-way ANOVA was conducted for each outcome variable. For precision, only the topic type was statistically significant ($F = 12.61$, $p=0.0004$), with the mean precision for specific topics being 0.89 and the mean precision for generic topics being 0.76. For recall (see Table 3), the topic type main effect ($F = 54.85$, $p<0.0001$) and the system main effect ($F = 14.02$, $p<0.0001$) were statistically significant, as well as the interaction effect ($F = 12.45$, $p<0.0001$). The interaction effects are illustrated in Figure 2. It appears that the systems that incorporate searching of the transcripts perform better on specific topics than on generic topics, while the features-only system performs equally on both types of topics.

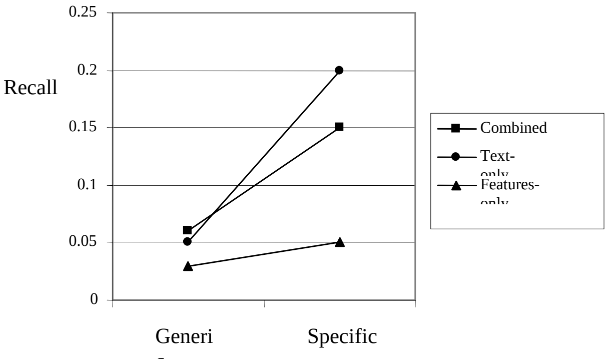


Figure 2. Recall, by system and topic type

Further post-hoc analysis was conducted on the system effect on each type of topic. For *generic* topics only, the combined system performed at the same level as the text-only system on recall, and there was also no statistically-significant difference between the recall of the text-only and features-only systems. Only the difference between the combined system and the features-only system was statistically significant. For *specific* topics only, recall in the text-only system and the combined system were equivalent, but both systems performed significantly better than the features-only system.

4. DISCUSSION

The results described in this study have many useful implications for video retrieval research, especially for the applicability of content-based video retrieval research, and implementations of the hybrid video retrieval research.

First, the hypothesis that the combined or hybrid system would perform better than both the concept-only (i.e., text-only) and content-only (i.e., features-only) systems was not supported. Generally speaking, the performance of the text-only system and the combined system were equivalent, which was consistent with the Dublin City 2002 TRECVID results [5]. However, topic type interacted with system performance: there was a difference between *generic* and *specific* topics.

For *generic* topics, all three systems performed similarly (combined and text-only systems were the same, and text-only and features-only systems were the same). However, for *specific* topics, the two systems incorporating a concept-based approach—the text-only system and the combined system—performed better than the content-based (i.e., the features-only) system. This finding is quite reasonable, since concept-based indexing and retrieval methods have high expressive power [3], and thus might provide more accurate or precise results for more specific topics, such as proper nouns. Additionally, the features-only system did not always perform the worst, and it actually outperformed the other systems on some of the topics, most of which were generic, and many of which matched with one or more of the 16 features the system provided. For instance, the topic "baseball batter swinging" matched with the feature "sporting event", and the topic "roads with vehicles" matched with the features "road" and "car/truck/bus". This actually demonstrates that content-based video retrieval techniques have great potential to help users to formulate their queries and to find the results they want, but only in situations where there is a clear connection between the topic and the available features.

The topic effects discovered in this study have implications for the design of video retrieval systems. *Specific* topics performed significantly better than *generic* topics, so how to improve search performance on generic topics appears to be a major challenge. Part of the challenge is due to the multiple ways that a generic topic might be expressed in natural language; for example, the topic "roads with vehicles" might be expressed with the words, "highways," "interstates," "streets," or "avenues" for the first portion of the topic. Specific topics, e.g., "Yasser Arafat," will likely vary less in how they are expressed in natural language. Since content-based video retrieval techniques have some advantages on generic topic searches, the further development of these techniques is worth pursuing. In addition, conducting some user task/query analysis (e.g. [10]) as the basis for selecting features for automatic identification might be a good way to improve video retrieval efficacy.

Secondly, the study results also provide some suggestions on how to design a hybrid or combined video retrieval system. For instance, one open question is whether to combine the content-based retrieval techniques with the concept-based linguistic techniques during query formulation or as a method for relevance feedback. The systems designed by [6] and [11] were examples of the latter case. They used concept-based retrieval to initiate a search, and content-based retrieval (e.g., color and texture) to find "more like this" results. The combined system implemented in this study used a different "hybrid" mode than [11]. Here, the content-based techniques were not used to provide relevance feedback, but to help in query formulation, as in [5]. Although the relevance feedback approach in [6] and [11] showed great potential, the query formulation approach might also have some advantages, especially for generic topics, if the efficacy of the content-based video retrieval techniques is further improved.

5. FUTURE WORK

There are limitations of the work presented here. For instance, the limited number of features users could choose and the imperfection of the feature extraction results might have influenced their performance on the features-only and combined systems; the search algorithms for all of these three systems need to be further improved; and the browsing interfaces might not be very efficient for users to navigate through different views. The retrieval system and the study design will be improved in future TRECVID participations. Additionally, the TRECVID 2003 workshop used news videos and defined a limited number of search topics, thus the results presented here might not be generalized to other types of videos and other types of topics. It is hoped that other researchers will conduct similar user studies to test our results.

6. ACKNOWLEDGMENTS

Our thanks to NSF IIS 0099638 for funding this research, to NIST for providing video and features data and evaluating the submitted results, and also to our study participants,.

7. REFERENCES

[1] Rasmussen, E. M. Indexing images. *Annual Review of Information Science and Technology, 32*(1997), 167-196

[2] Enser, P. Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science 26,4*(Aug. 2000). 199-210

[3] Eakins, J. P., and Graham, M. E. Content-based image retrieval: a report to the JISC Technology Applications Programme. Institute for Image Data Research, University of Northumbria at Newcastle. 1999. Available at http://www.unn.ac.uk/iidr/report.html .

[4] Lew, M., Sebe, N. and Eakins, J. Challenges of image and video retrieval. International Conference on Image and Video Retrieval, 2002 LNCS 2383, 1-6

[5] Browne, P., Czirjek, C., Gurrin, C., Jarina, R., Lee, H., Marlow, S., Donald, K. M., Murphy, N., O'Connor, N. E.,, Smeaton, A. F. and Ye, J. Dublin City University Video Track Experiments for TREC 2002. Notebook of TREC video track, 2002. 217-226.

[6] Browne, P., Czirjek, C., Gaughan, G., Gurrin, C., Jones, G.J.F., Lee, H., Marlow, S., McDonald, K., Murphy, N., O'Connor, N. E., O'Hare, N., Smeaton, A.F., Ye, J. Dublin City University Video Track Experiments for TREC 2003. Notebook of TREC video track, 2003, 51-62

[7] Guidelines for the TRECVID 2003 evaluation. (2003, May 27). http://www.nlpir.nist.gov/projects/tv2003/tv2003.html.

[8] Wildemuth, B. M., Yang, M., Hughes, A., Gruss, R., Geisler, G., and Marchionini, G. Access via Features versus Access via Transcripts: User Performance and Satisfaction. Proceedings of TRECVID 2003 workshop notebook. 2003, 189-198

[9] Shatford, S. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloguing & Classification Quarterly,1986. 5(3)*. 39-61

[10] Murdock, V. and Croft, W. B. Task-Orientation in Question Answering. *Proceedings of the Twenty Fifth International ACM SIGIR Conference, Tampere, Finland, August 2002*.355-356

[11] Hauptman, A., Baron, R. V., Chen, M., Christel, M., Duygulu, P., Huang, C., Jin, R., Lin, W., Ng, T., Moraveji, N., Papernick, N., Snoek, C., Tzanetakis, G., Yang, J., Yan, R., and Wactlar, H. Informedia at TRECVID 2003 : Analyzing and Searching Broadcast News Video. Proceedings of TRECVID 2003 workshop notebook. 2003, 21-34.