EXPLOITING MULTIPLE MODALITIES FOR INTERACTIVE VIDEO RETRIEVAL

Michael G. Christel, Chang Huang, Neema Moraveji, and Norman Papernick

Carnegie Mellon University
Pittsburgh, PA 15213
{christel, liz+, nmoravej, norm+}@cs.cmu.edu

ABSTRACT

Aural and visual cues can be automatically extracted from video and used to index its contents. This paper explores the relative merits of the cues extracted from the different modalities for locating relevant shots in video, specifically reporting on the indexing and interface strategies used to retrieve information from the Video TREC 2002 and 2003 data sets, and the evaluation of the interactive search runs. For the documentary and news material in these sets, automated speech recognition produces rich textual descriptions derived from the narrative, with visual descriptions and depictions offering additional browsing functionality. Through speech and visual processing, storyboard interfaces with query-based filtering provide an effective interactive retrieval interface. Examples drawn from the Video TREC 2002 and 2003 search topics and results using these topics illustrate the utility of multiple-document storyboards and other interfaces incorporating the results of multimodal processing.

1. INTRODUCTION

Information retrieval from video presents unique challenges, given the aural and visual complexity of the source material and its temporal characteristics. Traditionally, searching through a video collection has implied tedious rewinding and fast-forwarding through linear broadcasts until the relevant content is found, often armed with no more than a video title as an access point. Browsing strategies to more efficiently review and locate information across different video genres have been examined [4], with the conclusion that "news can fall equally into both the informational audio-centric and informational video-centric categories, and can take advantage of a combination of the different indices for effective browsing" [4, p. 176]. Considering news video, the narrative describes a great deal of the information, which can then be effectively indexed as text through automatic speech recognition (ASR). News video also contains a great deal of visual content: a few thousand hours of news broadcasts contains over 2,000,000 shots, where a shot is a contiguous set of frames captured by a single camera. Video collections of news and other genres are gradually becoming digital, underscoring the need for continued work into efficient, effective information retrieval interfaces for digital video, and the consideration of how the aural and visual modalities can best serve information retrieval needs.

In 2001, the TREC Video Track (TREC-V) began with the goal to investigate content-based retrieval from digital video. The focus was on the shot as the unit of information retrieval, rather than the scene or story segment. Digital video has been decomposed into shots by a number of research projects and commercial systems in the past (see [3] for review), and shot boundary detection remains one of the tasks addressed by TREC-V [5]. This paper examines user-driven information retrieval from video, focusing on the lessons learned from the Carnegie Mellon interactive search runs for the past two TREC-V evaluations.

2. TREC-V 2002 AND 2003

For TREC-V 2002, NIST defined 25 information needs, i.e., "topics", to search for within a search test collection of 40.12 hours of MPEG-1 video from the Prelinger Archives and the Open Video archives [6]. The material consisted of advertising, educational, industrial, and amateur films produced between the 1910s and 1970s, spanning a wide spectrum of quality and attributes, e.g., silent films and animated cartoons were part of the corpus. The search test collection was delineated into 14,524 shots, which became the common shot reference used by all participants in the query tasks, allowing for easier assessment.

For 2003, NIST likewise defined 24 interactive search topics and made use of a common shot reference: 32,318 shots representing 55.91 hours of video: ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998, along with six hours of C-SPAN programming from mostly 2001 [5].

The TREC-V interactive task is defined as follows: given a multimedia statement of information need (topic) and the common shot boundary reference, return a ranked

list of up to N shots from the reference which best satisfy the need, where N=100 for 2002, N=1000 for 2003, with the user having no prior knowledge of the search test collection. The topics reflect many of the sorts of queries real users pose, including requests for specific items or people, specific facts, instances of categories, and instances of activities [5]. Mean average precision is used to compare the relative merits of the interactive systems.

3. TREC-V INTERACTIVE SEARCH RESULTS

For TREC-V 2002, our "CMU" interface was designed to present a visually rich set of thumbnail images to the user, tailored for expert control over the number, scale, and attributes of the images [2]. Armed with this interface, an expert user completely familiar with the video retrieval system and its features produced the best scoring interactive run. The results are shown in Figure 1, where 8 research groups submitted 14 interactive query runs.

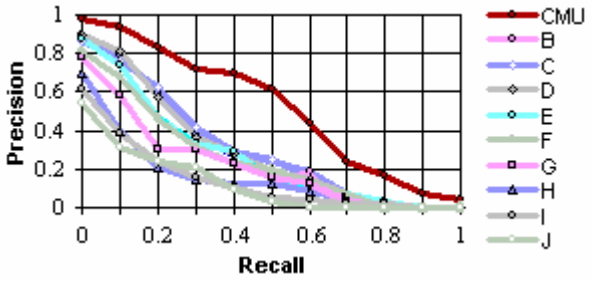

Figure 1. Precision-recall curve for top 10 of 14 interactive search runs in TREC-V 2002.

For TREC-V 2003, 11 groups submitted 37 interactive runs, including 2 runs by our "CMU" group. "CMU2002" used the same system as was used the prior year, but this time by 5 other researchers in the group each addressing a subset of the search topics. "CMU2003" was a new 2003 interface used by the same expert user from 2002 across all 24 interactive search topics. The mean average precision results are shown for the top 20 interactive runs in Figure 2.

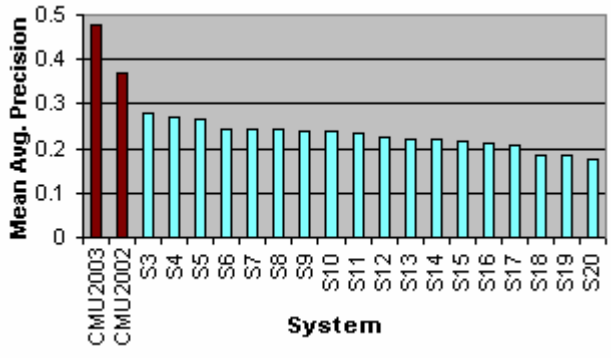

Figure 2. Mean average precision for top 20 of 37 interactive runs in TREC-V 2003.

4. LEVERAGING QUERY CONTEXT

The primary interface component leading to the success of both the 2002 and 2003 CMU systems was a storyboard spanning multiple documents, useful for presenting visual answers with some temporal context [2]. An examination into the construction of the storyboard illustrates one way in which aural and visual cues can be combined into an effective interface component.

For news and documentaries, ASR has been shown to be useful to locate relevant shots for a given topic [1, 2, 7]. The narrative contributes a great deal of detail pertaining to the corpus, detail that through ASR can be captured and indexed as text and accessed through traditional text retrieval algorithms. Audio cues like silence points and speaker changes, coupled with visual cues like black frames, hard cuts, and domain-specific transitions like from a news anchor shot to a commercial/advertisement, can be used to partition the video data into story units, i.e., segments [7]. Text from ASR, coupled with additional text recognized from superimposed video text common in news broadcasts, is then used to describe and index these segments.

The shots for segments returned by a query are presented in a single storyboard, i.e., an ordered set of keyframes presented simultaneously, one keyframe per shot. Without further filtering, most queries would overwhelm the user with too many images. Through the use of query context, the cardinality of the image set can be greatly reduced. The search engine for text queries makes use of the Okapi formula, with a few coefficients tuned to the short texts typical of documentary and news documents as shown in the following equation:

$$Sim(Q,D) = \sum_{qw \in Q} \left\{ \frac{tf(qw,D)\log\left(\frac{N - df(qw) + 0.5}{df(qw) + 0.5}\right)}{0.8 + 1.2\frac{|D|}{avg\_dl} + tf(qw,D)} \right\}$$

where for query Q, $tf(qw,D)$ is the term frequency of word qw in document D, $df(qw)$ is the document frequency for the word qw and avg_dl is the average document length for all the documents in the collection. Through automatic speech alignment, the user's expressed information need in the query can be pinpointed to particular points in the narrative, as well as to superimposed video text. The multiple-document storyboard can be set to show only the shots containing matching words. This strategy of selecting a single thumbnail image to represent a video document based on query context resulted in more efficient information retrieval with greater user satisfaction in past studies [7]. Here, the idea is extended to collapse a set of thumbnails spanning multiple documents to a smaller set of only the shots containing matches to a given query. For example, a query about trains and locomotives returns 77 segments

consisting of 4208 shots for TREC-V 2003; the multi-document storyboard showing only matching shots holds 141 shots, as shown in part in Fig. 3. Shots are ordered by segment relevance to the query, and then temporal order within each segment.

Figure 3. Top of storyboard for train query: match context reduces shot count from 4208 to 141.

These query-based multi-document storyboards leveraged greatly from the audio narrative through synchronized ASR text, as well as the superimposed video text extracted from the visual modality. The difficult automation problem of accurately assigning visual feature classes like indoors, outdoors, and vehicles to the shots was not employed at query formation time, since the ASR accuracy is likely at 70% or better while many feature classifiers have mean average precision of 30% or worse. The more accurate modality (speech), and the known IR methods for dealing with text query were hence used to provide initial results interfaces like that shown in Figure 3 to the user. The user then could visually browse tens or hundreds of thumbnails and select those of relevance to the topic at hand. The 2003 CMU system introduced an improvement to show the thumbnail at the mouse position at full resolution to allow the user to quickly inspect finer visual details that might be lost at smaller thumbnail sizes.

5. ASR WITH VISUAL FEATURE FILTERING

Sometimes query context did not reduce the storyboard shot count enough to allow for easy inspection. In such cases, the visual features were used to interactively filter the shot set down to manageable size. For example, consider the user looking for people with streets or buildings in view. A query on streets and urban settings might produce a result set of 100 segments and 475 matching shots to the urban query, as illustrated in Fig. 4. Through the use of dynamic query sliders, the user can apply a restriction to leave only shots that are definitely not indoors in view, slide the outdoors filter to keep outdoors shots, and then start sliding the people filter to keep shots with people. These filters work against automatic feature classification that assign a [0,1] confidence score. By collapsing the remaining shots to the top of the storyboard, the user can then easily inspect the smaller set, and even with imprecise visual filters the user can interactively select the third, seventh, and other shots as reasonable for the given query. In this way, visual features under user control allow imprecise filtering to be employed in ways that lead the user to relevant shots. These shots can then be used as keys for subsequent color, texture, or QBIC-style color region-based searches that initiate a new storyboard for investigation.

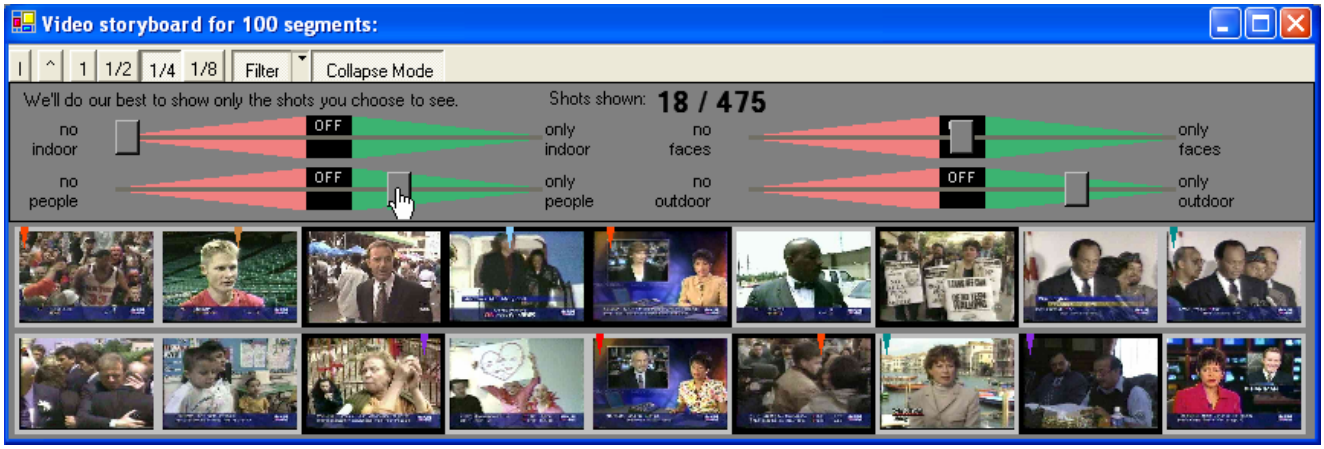

Figure 4. Visual feature filters reduce storyboard complexity: as sliders move, storyboard repaints to show keyframes of shots still meeting the filter restrictions.

6. VISUAL FEATURE BROWSING

As noted by other researchers [1, 3, 4], information in documentaries and news is conveyed in both the audio and video. For some topics, like celebrities and politicians, the narrative mentions names and reasons for being in the news, with the visuals showing the people. Here, ASR and overlay text are ideal indexing strategies to locate relevant shots and display them in storyboards as discussed earlier. For peripheral topics such as people in the street or crowded road traffic, the narrative does not describe the setting adequately to precisely find such material. For those cases, the visual channel holds most or all of the information. Similarly, material more likely to be found in commercials rather than news, such as cups

of coffee or cats, will not be described by the commercial's audio narrative as that is devoted to selling a product. A key improvement in CMU2003 from CMU2002 (see Fig. 2) is the addition of visual feature browsing to launch investigations into such visually oriented topics.

While the state of the art in automatic feature detection is not good enough to precisely pinpoint just the relevant shots (see results of "manual" no-human-in-the-loop search from prior TREC-V runs), the feature classification is good enough to provide a set of images useful for visual inspection. For 2003, the topic on "traffic" was explored by looking at the top-ranking shots classified as having "roads" arranged in a storyboard, and then filtering that set down to the road shots with vehicles and no faces. Genre-specific clues, in this case the difference between commercials and news, were especially useful for a topic like "cats" anticipated to be found in commercial footage. A storyboard of top-ranking commercial shots, as shown in Figure 5, was used to initiate visual browsing. The user quickly found and selected cat images to launch follow-up image-based queries based on shape and color, from which additional cat images were then located, as well as neighboring cat shots within the same commercial.

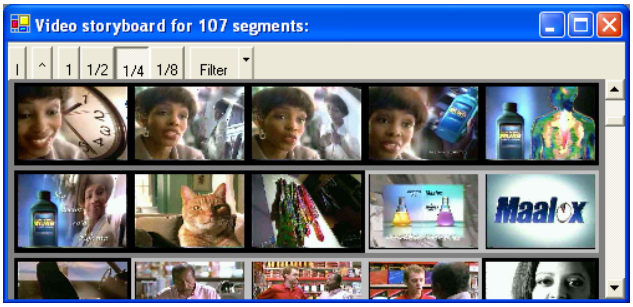

Figure 5. Storyboard of commercial shots, from which the user can visually inspect to find a "cat" shot.

7. CONCLUSIONS AND FUTURE WORK

Text-based features are the most reliable high-level features applicable in news and documentary video retrieval, and for those genres the text is derived primarily from ASR against the spoken narrative. Because of the difference in accuracy of the automatic detectors and in the semantic descriptive power of the derived data, the text is used first to locate a candidate shot set, which the user can subsequently browse and filter based on less accurate visual features. For cases where ASR text is absent (silent films) or poor quality (commercials), the visual modality can be used to provide browsing interfaces to allow the user to locate an image key that can launch a visual-based query leading to more focused candidate sets with additional relevant shots.

Future work includes looking at genres of video where the visual modality must be leveraged completely (as in surveillance video with no scene text), and from that work learning ways in which the visual modality can be better intertwined with the use of ASR text at query formulation time. Early successes are likely to be domain- and genre-specific: for example, we can apply an anchor filter at query time to remove anchors from initial storyboards of news results as anchor shots are nearly always irrelevant. Domain-specific feature classifiers like anchor shots and commercial shots for TREC-V and scoring and activity shots in other venues like specific sports have greater accuracy than more generic high-level features such as indoors and people. Domain-specific visual features are also easier to map to information topics in the given domain. Other obvious areas for future work are utilizing the non-speech aural cues for more than just segmentation, developing a taxonomy of visual features applicable to information needs in a given domain, and using multimodal cues in parallel rather than one filtering the other in sequence. Overall, the investigation into the benefits of aural and visual cues for interactive video retrieval has just begun, with potential for better utilizing the perception and intellect of the human user in achieving faster, better, more satisfying performance levels.

8. ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation (NSF) under Cooperative Agreement No. IRI-9817496 and IIS-0121641. This work is also supported in part by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406. Informedia researchers (please see http://www.informedia.cs.cmu.edu) contributed significant work resulting in the interfaces discussed here.

9. REFERENCES

[1] B. Adams et al., "IBM Research TREC-2002 Video Retrieval System," *TREC 2002 Proc.* (Gaithersburg, MD, Nov. 2002), http://trec.nist.gov/pubs/trec11/papers/ibm.smith.vid.pdf.

[2] M. Christel et al., "Enhanced Access to Digital Video through Visually Rich Interfaces," *Proc. ICME* (Baltimore, MD, July 2003), IEEE CS Press, pp. III-21 - III-24.

[3] H. Lee et al., "Designing the User Interface for the Físchlár Digital Video Library," *J. Digital Info 2*(4), May 2002.

[4] F. Li et al., "Browsing Digital Video," *Proc. ACM CHI '00* (The Hague, Netherlands, April 2000), ACM Press, 169-176.

[5] NIST, *TREC Video Retrieval Evaluation*, http://www-nlpir.nist.gov/projects/trecvid/.

[6] A.F. Smeaton and P. Over, "The TREC-2002 Video Track Report," *TREC 2002 Proc.* (NIST, Gaithersburg, MD, Nov. 2002), http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

[7] H. Wactlar et al., "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library," *IEEE Computer*, *32*, 2, pp. 66-73, 1999.